Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The emergence of regional record-breaking summer heat predictability

Emily Gordon

emily.gordon@auckland.ac.nz

University of Auckland

Noah Diffenbaugh

Stanford University

Article

Keywords:

Posted Date: September 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10796829/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Record-breaking heat continues to be one of the deadliest climate-related hazards, increasing the urgency to reliably predict these events years in advance to allow adequate preparation and adaptation. We demonstrate a machine learning method for predicting future record-breaking summers within 5- and 10- year time horizons, and demonstrate its ability to outperform a dynamical model at predicting historical extreme events. The convolutional neural networks in this study ingest GMT and SST variability to predict whether a record-breaking summer will occur within a future time horizon. As expected, the CNNs's confidence in predicting future record heat increases with global warming; however, predictions also rely on SST conditions that more predictably lead to local warming. The combined predictability from global warming and SST variability results in the emergence of record-heat predictability in the historical record. This emergence has already occurred in regions with a high warming signal, but also in midlatitude regions that are influenced by large-scale teleconnections. We estimate the plausible range of regional predictability emergence, finding that in most land regions, skillfully predicting the onset of a record-breaking summer in a 5-year time frame will be possible within 10–30 years if current warming rates continue.

Introduction

Extreme events remain the most devastating impacts of anthropogenic climate change. The increase in frequency and intensity of extreme heat, precipitation and sustained drought accelerates the urgency to predict these events both in the short-term, but also months to years in advance (1–3). These longer timescales can provide crucial time to mobilise resources in preparation of the coming extreme (4). Extreme event prediction has however proven difficult on these timescales as the events themselves result from a combination of climate and weather processes, some predictable and some unpredictable (5, 6).

Among the sources of predictability of extreme events, it is still debated how to attribute the predictability that derives from the warming trend since, to first order, we know extreme heat will increase with climate change. One approach is to estimate and remove the forced response, and investigate the residual prediction skill that derives from internal variability (7, 8). This approach allows for the quantification of how well a predictive model captures behaviour that diverges from a trend. However, if the extreme event is defined before trend removal, then the trend removal may decrease the event amplitude to the extent that the predictability of the extreme would be attributed to forcing, even if it was, in part, driven by internal variability (9). Conversely, if we define extremes only after removing the forced response and investigate their predictability (10), this definition may miss recent extreme events that contain a high contribution from forcing so would not be considered "extreme" due to internal variability alone. In either case, the estimation of the externally forced trend is itself fraught because quantifying the forced response at regional scales in the historical era is difficult (11, 12) and trend estimations thus introduce biases to the estimation of the predictable variability signal (13).

If we instead consider that external forcing amplifies warm phases of internal variability, it is apparent that there is some level of warming where the combination of internal variability and external forcing together may result in a predictable extreme that would not be predictable if we were to consider only their individual contributions. This idea was recently proposed by (14), who found that convolutional neural networks trained to predict abrupt summertime warming from sea surface temperature (SST) variability and global mean temperature (GMT) become more likely to predict summertime warming at higher GMT, but there exists an intermediate level where predictions of summertime warming still required favourable SST variability conditions. The level of warming where extreme event predictability "emerges" will be locally dependent on both the warming signal and the fraction of variability that derives from low frequency variability.

In this paper, we build on the past work of (14) and demonstrate the regional emergence of extreme event predictability. We train machine learning models to take the combined information of recent SST variability and GMT anomaly, and predict whether a record hot summer will occur in the next 5- or 10-year period (Fig. 1a). We compare these predictions with an initialized hindcast ensemble, demonstrating that, in many regions, the machine learning model is able to outperform the predictions from a fully coupled dynamical model. Crucially, the machine learning model's record hot summer predictions become more frequent at higher GMT anomaly while remaining distinct from a GMT-only prediction, implying that the inclusion of SST variability as an input results in the early detection of predictable extreme summertime temperatures.

Results

Recent SST variability improves predictions of future record hot summers

We first demonstrate that it is possible to predict future record hot summers in the observational record with a convolutional neural network. We examine the CNN's predictions of the likelihood that summer temperature in the next 5 years will exceed the hottest summer in 1960–1990 in the historical era using ERA5 reanalysis which is out-of-sample data. For an example grid point over Eastern Europe (40-50N, 30-40E) the CNN is able to distinguish when the combined effect of recent SST variability and mean warming combines to make a future extreme predictable (Fig. 1b), as the CNN's confidence in a future extreme summer (black line) increases around the year 2000. The CNN's prediction is earlier than the event itself, likely because the CNN detected the hot but not record-breaking summer in 1999 (red line).

While predicting future summers that exceed a historical baseline can be useful for future climate change preparedness, the historical threshold can also be made less extreme due to "shifting baselines" which obscure heatwave statistics and potentially mislead public perception of heat events (15–17). We therefore also examine the CNN's ability to predict future summers that exceed any prior summer from 1960 onwards (Fig. 1c). We again use Eastern Europe (30-40E, 40-50N) as an illustrative example of the CNN's predictions. Here, we find that the CNN predicts that a record-breaking summer will occur around

the early 2000s. However, in this case it then correctly detects that this new record (red dashed line) will not be broken in the 2010s and instead only begins to again predict a new summertime temperature record around 2023, i.e. the CNN predicts that a summer in 2024–2028 will exceed the prior summertime temperature record set in 2010.

We train an individual CNN ensemble at each 10°x10° land surface area and compare the average skill of three best performing CNNs against the MPI-ESM1.2 LR initialized large ensemble using the Brier skill score (Fig. 2a). The CNNs's event prediction skill is significantly improved over the initialized hindcast ensemble for a large fraction of the land surface, particularly Northern Africa and the Mediterranean region, East and Southeast Asia, and parts of Australia and South America. Notably, the CNNs perform very poorly over central North America because it predicts enhanced extreme summer temperatures from around 2015 which have not occurred in observations (18), whereas the hindcast ensemble captures this lack of warming. This is in contrast to western and central Europe, where the CNNs capture the enhanced warming trend that is itself largely absent from GCM projections (19).

We also examine the skill of the CNNs at predicting future summers that break *all* previous summertime temperature records at each 10°x10° land surface area and compare the average skill of three best performing CNNs against the initialized large ensemble (Fig. 2b). While there are skill improvements over the initialized ensemble, the improvements are only statistically significant over Northern Africa, Western Europe and East Asia. As events that break all prior records are less common, it is not surprising that the CNNs and hindcasts are largely indistinguishable over the small sample size, however we do note that the CNNs demonstrate skill in predicting the onset of record-breaking summers in the observational record.

We find some skill at predicting future record-breaking summers within five years, so we extend the forecast horizon to 10 years, i.e., the CNNs must predict whether a record-breaking summer will occur in the next 10 years. We calculate the CNNs's skill in predicting events that exceed the 1960–1990 maximum summer (Fig. 2c) and predicting events that exceed any prior summer (Fig. 2d). In both cases, the CNNs show a larger skill enhancement over the MPI hindcast than the respective 5-year experiments, implying that the CNNs are able to learn the precursors that lead to future record-breaking summers. The MPI hindcast ensemble uses the same historical forcing as in the ensemble members used to train the CNNs, meaning that the CNNs and MPI ensemble should approximately predict the same response to forcing. However, in the MPI hindcast ensemble (and many initialized ensembles) the predictable internal variability signal in the free-running hindcast is too weak, particularly at longer lead times (20). We therefore suggest that the CNNs's enhanced performance at longer lead-times is due to detection of long-lasting predictable signals that the free-running model simulates poorly.

Predictability of extreme hot summers emerges

It is apparent that, in many regions, the CNNs require a particular combination of climate variability and GMT level to predict the onset of a future record-breaking summer. In some cases, the event would not

be predicted were it not for the increased global mean temperature. We term this phenomenon, where the increase in global mean temperature results in record-breaking events becoming more predictable when they are combined with patterns of variability that predictably lead to local warming, “predictability emergence”. We quantify predictability emergence by finding the GMT anomaly at which the CNNs first predicts that a future record-breaking extreme occurs in ERA5 (Fig. 3a). Note GMT anomaly is defined as the average anomaly of the prior 10 years before a prediction, and is presented as the anomaly from 1850–1899. We illustrate lower GMT emergence with brighter colours to emphasize that these regions are likely to be warming at a faster rate. The CNNs detect that future summers that break the local 1960–1990 record within the next five years have been largely predictable in the observational record––and in many places, these predictions perform better than the initialised hindcast ensemble. This predictability emergence is especially consistent over Eastern Asia, Continental Europe and the Western United States.

To verify that the inclusion of SST variability results in earlier detection of predictable record-breaking summers, we compare the CNNs’s predictions to the GMT of emergence in the initialised ensemble (Fig. 3b) and to the CNNs when it is provided only an average SST climatology (Fig. 3c). Notably, the CNNs detect predictability emergence either at the same time as, or earlier than, both the MPI hindcasts and the CNNs with climatological SST. The results are similar for record-breaking summers within a 10-year forecast horizon (Fig. 3d-f), except the GMT at which predictability emerges is earlier, with the longer forecast horizon increasing the likelihood of a future extreme occurring within the horizon.

We also examine the case of predictability emergence for summers that are hotter than any prior summer within five years (Fig. 3g-3i) and 10 years (Fig. 3j-3l). The predictability of record-breaking summers within five years has only emerged in places of high warming signal, including equatorial Central/South America, Northern Africa, Europe, and Southeast Asia (Fig. 3g). The CNNs are able to skillfully predict these events within 5 years from as early as 0.3˚C GMT anomaly in places that have experienced a high warming signal during the historical period, and within 10 years at around 0.5-1˚C GMT anomaly across much of the Northern Hemisphere mid-latitudes which have experienced a combination of relatively large warming, and are influenced by low frequency oceanic variability (Fig. 3g). We again find that predicting the onset of a record-breaking summer within 10 years occurs earlier than within five years because these events are simply more likely over a 10-year horizon. The exception is from 0-30N in the MPI hindcast ensemble which predicts earlier or similar emergence to the CNNs. We find that the MPI ensemble is able to predict some record-breaking extremes in the historical period due to the skill at short leadtimes: for Northern hemisphere summers, the first summer prediction occurs just 7–9 months after initialisation and therefore the MPI initialised ensemble successfully predicts these record-breaking events when they occur the summer after initialisation, further emphasizing how record-breaking summer predictability has emerged across different methods of prediction. Notably, the CNNs still predict these events more reliably than the initialized hindcast model over Europe, East Asia, the Arabian Peninsula, and Western North America.

Future predictability emergence from historical SSTs

Our analysis to this point has underscored the importance of SST variability in improving predictions of future record-breaking summers and therefore providing predictability emergence. We now quantify the lowest plausible GMT anomaly corresponding to predictability emergence at each grid point. We repeatedly provide the CNNs with all historical SST samples, but each time we increase the GMT anomaly input by 0.05˚C increments starting from − 0.3˚C and increasing to 3.0˚C. We then plot the lowest GMT anomaly at which the CNNs predict that the local 1960–1990 record summer will be broken within the next 5 years (Fig. 4a). Compared with the observed historical emergence in Fig. 3a, it is apparent that at many grid points, it is plausible that a future summer record breaking summer is predictable at the 2016–2025 GMT anomaly value (1.33°C) at almost all grid points if the preceding SST pattern is in the most favourable state for predicting extremes.

Additionally, we calculate the predictability emergence with incrementally increasing GMT anomaly when the CNNs are provided with SST climatology, and plot the GMT where the CNNs first predict a future record-breaking summer (Fig. 4b). If we interpret this GMT anomaly to be the level at which record-breaking events become predictable when SST provides no skill, then this provides an upper estimate on the GMT level at which we expect a future record-breaking summer to be predictable. In the case of predicting an event that exceeds the 1960–1990 maximum within 5 years, we therefore find that predictability has emerged at almost all grid points even with the estimate that SST provides no skill. For those grid points where predictability emergence occurs post-2025, e.g. Northern Canada and Northern Europe, we expect these events could be predictable at GMT anomaly as low as 1.3–1.5˚C GMT anomaly if SST conditions are favourable (Fig. 4a), and approximately 1.5–1.8˚C GMT anomaly if SST conditions do not provide enhanced predictability through this period (Fig. 4b).

As before, we extend our analysis to quantifying the theoretical predictability emergence for predicting summers that break the local 1960–1990 record in the next 10 years (Figs. 4c-4d). These events emerge earlier than the 5-year predictions because they are more likely to occur over a 10 year period with increasing warming. For all grid points except Equatorial Africa, it is possible to predict that a record-breaking summer will occur in the next 10 years, and even under climatological conditions they are predictable at current GMT levels (Fig. 4d).

We also quantify predictability emergence for predicting that a future summer will occur that will exceed the current (at the time of writing) record summer temperature at each grid point within 5 years (Fig. 4e) and 10 years (Fig. 4g). Given SST variability that is favourable for predicting enhanced warming, it is theoretically possible to predict these events within 5 years in Central South America, Central Africa, and Central and East Asia at the current warming level. Extending the lead time to 10 years results in predictability emerging at the current GMT anomaly across most of the global continental regions, excepting equatorial Africa and the maritime continent. In comparing the emergence GMT anomaly when the CNNs are provided with the optimal SST, to the climatological SST (Fig. 4f), we find that predicting a summer that exceeds the current record within the next 5 years from the GMT signal alone

emerges around 1.5–2.1˚C GMT anomaly. This translates to approximately 0.2–0.8˚C of warming above the current level for predicting summers that exceed anything in the observational record within a future 5-year period.

Discussion & Conclusion

A major challenge in decadal prediction is adequately attributing the predictability that is provided by anthropogenic climate change, and thus distinguishing predictable internal variability (21). In this study, we propose a new framework by investigating where the combined effects of internal variability and external forcing result in a predictable record-breaking summer. To demonstrate this concept, we investigate the predictability of future record-breaking summers and quantify "predictability emergence" as the warming level at which a regional summertime temperature record may be predictably exceeded. In this way, we exploit the fact that forcing provides predictability for record-breaking heat, but demonstrate that extreme event predictability is nevertheless tangibly improved with initial condition information. While this is not necessarily aligned with the definition of "predictability" as a predictable deviation from a forced (or slowly warming) timeseries (22), we are able to demonstrate that improved prediction skill is not solely due to forcing or internal variability.

We quantify the earliest predictability emergence that could plausibly occur from observed SST conditions in the historical era (Fig. 4a, 4c, 4e, 4g), as well as for climatological SST (Fig. 4b, 4d, 4f, 4h), and take this difference to be the range over which we expect a predictable signal to become detectable. We find that, even in regions where the predictability of breaking the hottest summer on record within the next 5 years has not yet emerged, it is likely that it will emerge with an additional 0.2–0.8˚C of warming above the current (2016–2025) average. The current estimate of anthropogenic warming is 0.27˚C/decade (23), translating to predictability emergence in the next 10–30 years for all grid points if we continue warming at the current rate. We note a caveat with this estimate: in our calculation of the earliest plausible predictability emergence, we provide the CNNs with SST variability that has occurred in the historical era. If future SST patterns contribute to more regional warming than anything we have observed (24), this would result in a lower GMT level when predictability emerges in the future than that presented here. We find that SST observations from the most recent 20-year period (i.e. with a larger contribution from the warming trend) are not necessarily the SST patterns that provide the most regional warming predictability throughout the historical record (see Figure S3), supporting our current estimate, though this may change if the warming trend continues to steepen, or SST patterns evolve in an unprecedented manner.

The machine learning models in this study demonstrate skill at predicting future record-breaking summers, defined both as breaking the 1960–1990 record, and as breaking any prior record. While these predictions do not exactly correspond to the prediction of a future heatwave, we choose this predictand as one that is correlated with the increased likelihood of an extreme heat event (25), but is still a regional and seasonal average and hence contains a predictable signal on decadal timescales. We expect other climate variables that are influenced by both low frequency variability and climate change to exhibit

predictability emergence but, as we demonstrate with the two definitions in this paper, the concept of predictability emergence depends on the specific extreme event and its definition.

In our experiment design, we train the CNNs on output from a single GCM. However, different models with different climate sensitivities may perform differently in this prediction task, and both the CNNs and hindcast ensemble predictions may be improved by including additional GCMs (26, 27). To avoid introducing GCM differences as a source of uncertainty, we choose to use only one climate model in this study, which enables comparison of the skill of different prediction models that are derived from the same underlying GCM physics. Interestingly, even with only a single GCM we find that the CNNs are able to predict the increase in record-breaking heat in Western Europe in recent years, which is underestimated in the testing members of the MPI large ensemble (Figure S2). It is proposed that recently observed circulation trends which result in extreme heat are not captured in GCM simulations (19, 28). However, it appears that the increase in record-breaking heat in the 2010s is detected by the CNNs, which implies that certain dynamical drivers of extreme heat (whether or not they are driven by external forcing) can be inferred from SST variability. On the other hand, the CNNs are poor over Central North America where warming is below what is predicted in GCMs (Figure S2). We suggest that this is likely due to the CNNs inheriting GCM biases in precipitation and evaporation trends in this region, the cause of which is still debated (18, 29). As we explicitly provide the CNNs with dynamical variability (SST) but no local thermodynamic information, we speculate additional input variables, such as recent precipitation or soil moisture anomaly may improve the CNNs's ability to capture extreme heat that is modulated by moisture near the surface. We leave this as an avenue for future work.

This work presents a new methodology for quantifying the emergence of extreme heat predictability, which implicitly includes the forced response as a source of predictability while still demonstrating that internal variability contributes to the predictability of record-breaking summers. While our methodology does not necessarily align with the definition of predictability as predicting a difference from a trend, we demonstrate that our approach allows the prediction of impactful climate events while performing more skillfully than considering forcing alone. As increases in extreme events outpace the ability of human and natural systems to adapt (30, 31), we demonstrate that we may already have the ability to predict record-breaking summers, and this predictability will continue to emerge as the climate continues to warm.

Materials & Methods

Large Ensemble Data

We use simulation data from the MPI-ESM1.2 LR grand ensemble version 2 (32). This dataset is a 50-member ensemble of the historical era (1850–2014) using CMIP6 historical forcing. Each member is then branched into each of four of the CMIP6 shared socio-economic pathways (SSPs) to simulate the climate response to future forcing (33). We use data from simulation years 1960–2014 of the historical era, and 2015–2100 for four SSPs (SSP1-2.6, SSP2-4.5, SSP3-7.0, SSP5-8.5), with 25 members randomly

allocated for CNN training, 13 members for validation, and 12 members kept independent for testing. This study uses the SST and near-surface temperature (temperature 2 metres above the surface) output from these simulations. SSTs are bilinearly regridded to 4°x4° and then annual means are calculated at each grid point. We calculate the summertime land surface temperatures by first area-weighting the near-surface temperature grid to a 10°x10° grid, only including points where at least 25% of the 10°x10° grid box is over a land surface region. We then calculate the climatological summertime (DJF in the Southern Hemisphere, JJA in the Northern Hemisphere) temperature time series at each of these 10°x10° grid boxes.

Initialized hindcast data

As a baseline against which to compare the CNN predictions, we use the MPI-ESM1.2 LR hindcast ensemble, generated for the decadal climate prediction project (DCPP) as a contribution to CMIP6 (7, 34). We use near-surface temperature for all available start dates (1961–2021), and all available ensemble members (16 for each start date). As with the CNN training data, we area-average the temperature data to 10°x10° regions over land, and calculate the summertime average (DJF in the Southern Hemisphere, JJA in the Northern Hemisphere). To correct the leadtime-dependent bias, we average all start dates and ensemble members up to a specific start date and then subtract this bias from each member of the subsequent hindcast. For example, to bias correct the 1990–1999 hindcast, we only use start dates up to 1989 to calculate the leadtime dependent bias, then for the 1991–2000 hindcast we use all start dates up to 1990 to calculate the bias, etc.. We set a minimum of 20 start dates over which to calculate the leadtime bias, and therefore the forecast evaluation begins in 1980. We further translate the bias-corrected forecasts to anomalies from 1960–1990 climatology by calculating the difference between the zero lead bias for start date *n* and the zero lead bias for all start dates between 1960–1990, and subtracting this from the bias corrected forecasts. While we have chosen this specific method of bias correction and anomaly calculation as we believe it to be the most fair (35), differing choices of bias correction and anomaly calculation do not substantially affect our results. Further detail about the bias correction method and sensitivity plots are included in the supplemental information (Figure S1).

ERA5

We use data from the ERA5 reanalysis (36) as out-of-sample input to the convolutional neural networks to examine whether the predictions made by the CNNs are valid in historical reanalysis. As with the simulation data, we use sea surface temperature and 2-m temperature output from ERA5 over the reanalysis record from 1960 to 2025. ERA5 SSTs are bilinearly regridded to 4°x4° and grid point annual mean temperature is calculated, while the near surface temperature data is coarsened to 10°x10°, with the same 25% land criterion as applied to the GCM data, and climatological average summertime temperature is calculated for each 10°x10° region.

We calculate GMT in ERA5 by calculating the area-weighted mean global annual mean temperature. We then calculate the GMT anomaly from the 1850–1899 estimated global mean temperature by first calculating the GMT anomaly timeseries from 1991–2020 and add the pre-calculated correction factor of 0.88°C for scaling ERA5 GMT to the anomaly from 1850–1899 (37). Note that the CNN's are provided with the GMT anomaly from 1960–1990 during training (see below) but all results presented in this study are re-scaled to the 1850–1899 anomaly.

Convolutional Neural Network

We train CNNs to predict whether a record-breaking summertime temperature will occur over a future time period τ based on recent climate variability (Fig. 1a). The CNN inputs are 10 images of the annual mean SST map from the prior 10 years (one map per year). The average SST map of a particular input period is removed from each of the maps in the period, essentially creating an SST anomaly timeseries that is agnostic of the actual forcing timeseries but that contains information about the recent SST variability. The CNNs are also provided with the average GMT anomaly over the 10-year input period to capture the effect of the climate change trend. The GMT anomaly is calculated individually within each ensemble member as the difference from the 1960–1990 average GMT. The CNNs are also provided with a temperature threshold from which to predict a summertime temperature exceedance. The choice of preprocessing steps is intended to provide the CNNs with the context of recent climate change, but is agnostic of ensemble mean statistics so that the CNNs may be applied to the reanalysis as an "out-of-sample" dataset without need for approximating ensemble statistics for observations.

The CNNs's prediction task is to output a confidence, between 0 and 1, of summertime average temperature exceeding the input temperature threshold within the next τ years. We choose two future time periods (τ = 5 years, and τ = 10 years) over which the CNNs must predict whether a record-breaking summer will occur. The CNN architecture is directly adapted from (14), with only two changes: 1) the prior temperature record is added as an additional input, and 2) the hidden layers are updated to a single layer of 1000 nodes, which was found from hyperparameter tuning to perform slightly better for this task. We train an ensemble of 10 CNNs for each land surface grid point, and select the three that perform the best on the validation data (i.e., exhibit the largest difference between CNN accuracy and random guessing) for further analysis. The predictions in this manuscript are the mean of these three best performing CNNs. The CNNs are trained at a learning rate of 0.01 using the Adam optimizer for up to 500 epochs with learning rate decay by a factor 10 if the validation loss does not decrease for 10 consecutive epochs, and early stopping if the validation loss does not then decrease for an additional 10 consecutive epochs. The CNNs are trained using a categorical cross entropy loss function with adaptive sample weighting to account for the imbalance in class sizes at different grid points.

Hindcast predicted confidence

To calculate the initialized hindcasts' confidence in a future extreme summer, we rank each ensemble member for each year within the forecast horizon. We then take the maximum value of each rank over the forecast horizon (τ = 5 years or τ = 10 years) to be the distribution of predicted maximum summer temperature. The hindcasts' confidence in a future summer exceeding some threshold is taken to be the difference between 100% and the percentile of the temperature threshold in the distribution of predicted maximum summer temperature. For example, a 90% confidence prediction implies that the 10th percentile of the forecast distribution exceeds the previous record-summer temperature at least once within a forecast horizon.

CNNs with climatological SST

As an additional baseline to examine the emergence of predictability of record-breaking summers, we examine the CNNs's predictions of record-breaking summers when the full SST variability is replaced with the 1970–2025 climatology, so the CNNs are only provided the GMT anomaly, prior local record temperature value, and climatological SSTs. We define this climatology by finding the average 10-year SST evolution in the historical era. This climatology is calculated by generating the SST input samples for all years in the historical era, then taking the averaging over all historical input samples.

Skill calculation and significance testing

We present CNN prediction skill as the Brier skill score (BSS) which is a metric that compares the Brier score of a categorical forecast (in this case, the CNNs) against the Brier Score of some baseline (in this case, the MPI-ESM1.2 LR hindcast). The Brier score (BS) is akin to mean squared error but for categorical predictions where:

$$BS = \frac{1}{N}\Sigma\left(p_i - o_i\right)^2$$

The observations o_i are 0 for no event and 1 for event, while the p_i are the predicted confidence and may vary from 0 to 1. Brier skill score is given as:

$$BSS = 1 - \frac{BS_{CNN}}{BS_{hindcast}}$$

Therefore, Brier skill scores greater than zero indicate that CNN predictions outperform hindcast predictions.

We test the statistical significance of BSS for the different experiments using the block bootstrap method: we resample the predictions from both the CNNs and hindcast with replacement and a block length of three samples to account for autocorrelation between predictions (38). We calculate the BSS for each resampled set of forecasts and set a significance threshold that 90% of resampled BSS exceed 0 (CNN outperforms hindcasts in 90% of resampled forecasts) This significance level and evaluation method is comparable to e.g. (39).

Declarations

Funding Declaration

Emily Gordon was supported by the University of Auckland. Noah Diffenbaugh was supported by Stanford University.

Author Contribution

EMG and NSD conceived this study and EMG performed the analysis and visualisation. EMG wrote the manuscript draft and both authors contributed to editing of the manuscript.

Acknowledgement

We would like to thank the Stanford Doerr School Center for Computation and the Stanford Research Computing Center for providing computational resources. The authors wish to acknowledge use of the eResearch Infrastructure Platform hosted by the Crown company, Research and Education Advanced Network New Zealand (REANNZ) Ltd., and funded by the Ministry of Business, Innovation & Employment.

Data Availability

MPI-ESM 1.2 LR large ensemble and hindcast data is available to download from Earth System Grid Federation, https://esgf-metagrid.cloud.dkrz.de/search. ERA5 data can be downloaded from Copernicus Climate Data Store https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means?tab=download

References

1. N. S. Diffenbaugh, D. Singh, J. S. Mankin, D. E. Horton, D. L. Swain, D. Touma, A. Charland, Y. Liu, M. Haugen, M. Tsiang, B. Rajaratnam, Quantifying the influence of global warming on unprecedented extreme climate events. *Proc. Natl. Acad. Sci. U. S. A.* 114, 4881–4886 (2017).
2. Y. Kushnir, A. A. Scaife, R. Arritt, G. Balsamo, G. Boer, F. Doblas-Reyes, E. Hawkins, M. Kimoto, R. K. Kolli, A. Kumar, D. Matei, K. Matthes, W. A. Müller, T. O'Kane, J. Perlwitz, S. Power, M. Raphael, A. Shimpo, D. Smith, M. Tuma, B. Wu, Towards operational predictions of the near-term climate. *Nat. Clim. Chang.* 9, 94–101 (2019).
3. C. Cassou, Y. Kushnir, E. Hawkins, A. Pirani, F. Kucharski, I.-S. Kang, N. Caltabiano, Decadal Climate Variability and Predictability: Challenges and Opportunities. *Bull. Am. Meteorol. Soc.* 99, 479–490 (2018).

4. N. Dunstone, J. Lockwood, B. Solaraju-Murali, K. Reinhardt, E. E. Tsartsali, P. J. Athanasiadis, A. Bellucci, A. Brookshaw, L.-P. Caron, F. J. Doblas-Reyes, B. Früh, N. González-Reviriego, S. Gualdi, L. Hermanson, S. Materia, A. Nicodemou, D. Nicolì, K. Pankatz, A. Paxian, A. Scaife, D. Smith, H. E. Thornton, Towards Useful Decadal Climate Services. *Bull. Am. Meteorol. Soc.* 103, E1705–E1719 (2022).
5. D. R. Easterling, G. A. Meehl, C. Parmesan, S. A. Changnon, T. R. Karl, L. O. Mearns, Climate extremes: observations, modeling, and impacts. *Science* 289, 2068–2074 (2000).
6. K. E. Trenberth, J. T. Fasullo, T. G. Shepherd, Attribution of climate extreme events. *Nat. Clim. Chang.* 5, 725–730 (2015).
7. G. J. Boer, D. M. Smith, C. Cassou, F. Doblas-Reyes, G. Danabasoglu, B. Kirtman, Y. Kushnir, M. Kimoto, G. A. Meehl, R. Msadek, W. A. Mueller, K. E. Taylor, F. Zwiers, M. Rixen, Y. Ruprich-Robert, R. Eade, The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geosci. Model Dev.* 9, 3751–3777 (2016).
8. D. M. Smith, R. Eade, A. A. Scaife, L.-P. Caron, G. Danabasoglu, T. M. DelSole, T. Delworth, F. J. Doblas-Reyes, N. J. Dunstone, L. Hermanson, V. Kharin, M. Kimoto, W. J. Merryfield, T. Mochizuki, W. A. Müller, H. Pohlmann, S. Yeager, X. Yang, Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science* 2, 1–10 (2019).
9. C. Delgado-Torres, M. G. Donat, A. Soret, N. González-Reviriego, P.-A. Bretonnière, A.-C. Ho, N. Pérez-Zanón, M. S. Cabré, F. J. Doblas-Reyes, Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes. *Environ. Res. Lett.* 18, 034031 (2023).
10. E. E. Tsartsali, S. G. Yeager, P. J. Athanasiadis, S. Tibaldi, S. Gualdi, Predictability of temperature extremes in multi-annual forecasts. *Environ. Res. Lett.*, doi: 10.1088/1748-9326/adfa3c (2025).
11. R. C. J. Wills, C. Deser, K. A. McKinnon, A. Phillips, S. Po-Chedley, S. Sippel, A. L. Merrifield, C. Bône, C. Bonfils, G. Camps-Valls, S. Cropper, C. Connolly, S. Duan, H. Durand, A. Feigin, M. A. Fernandez, G. Gastineau, A. Gavrilov, E. Gordon, M. Günther, M. Höver, S. Kravtsov, Y.-N. Kuo, J. Lien, G. D. Madakumbura, N. Mankovich, M. Newman, J. Rader, J.-R. Shi, S.-I. Shin, G. Varando, Forced Component Estimation Statistical Method Intercomparison Project (ForceSMIP). *J. Clim.* 39, 1927–1953 (2026).
12. C. Frankignoul, G. Gastineau, Y.-O. Kwon, Estimation of the SST Response to Anthropogenic and External Forcing and Its Impact on the Atlantic Multidecadal Oscillation and the Pacific Decadal Oscillation. *J. Clim.* 30, 9871–9895 (2017).
13. G. A. Meehl, H. Teng, D. Smith, S. Yeager, W. Merryfield, F. Doblas-Reyes, A. A. Glanville, The effects of bias, drift, and trends in calculating anomalies for evaluating skill of seasonal-to-decadal initialized climate predictions. *Clim. Dyn.*, doi: 10.1007/s00382-022-06272-7 (2022).
14. E. M. Gordon, N. S. Diffenbaugh, Machine learning predictions of summertime warming jumps on decadal timescales. *Environ. Res.: Climate* 5, 025001 (2026).
15. F. C. Moore, N. Obradovich, F. Lehner, P. Baylis, Rapidly declining remarkability of temperature anomalies may obscure public perception of climate change. *Proc. Natl. Acad. Sci. U. S. A.* 116,

4905–4910 (2019).

16. R. J. H. Dunn, C. P. Morice, On the effect of reference periods on trends in percentile-based extreme temperature indices. *Environ. Res. Lett.* 17, 034026 (2022).
17. V. A. Fraser-Leonhardt, G. C. Hegerl, J. D. P. Mollard, Sensitivity of heatwave characteristics to the climatic baseline. *Atmos. Sci. Lett.* 27, e70017 (2026).
18. J. K. Eischeid, M. P. Hoerling, X.-W. Quan, A. Kumar, J. Barsugli, Z. M. Labe, K. E. Kunkel, C. J. Schreck, D. R. Easterling, T. Zhang, J. Uehling, X. Zhang, Why has the summertime central U.s. warming hole not disappeared? *J. Clim.* 36, 1–49 (2023).
19. R. Vautard, J. Cattiaux, T. Happé, J. Singh, R. Bonnet, C. Cassou, D. Coumou, F. D'Andrea, D. Faranda, E. Fischer, A. Ribes, S. Sippel, P. Yiou, Heat extremes in Western Europe increasing faster than simulated due to atmospheric circulation trends. *Nat. Commun.* 14, 6803 (2023).
20. L. Zheng, J. Sillmann, L. Borchert, Decadal Predictions of the Link Between European Hot-Dry Compound Summers and North Atlantic Sea Surface Temperature. *Geophysical Research Letters* 53, e2025GL119820 (2026).
21. G. A. Meehl, J. H. Richter, H. Teng, A. Capotondi, K. Cobb, F. Doblas-Reyes, M. G. Donat, M. H. England, J. C. Fyfe, W. Han, H. Kim, B. P. Kirtman, Y. Kushnir, N. S. Lovenduski, M. E. Mann, W. J. Merryfield, V. Nieves, K. Pegion, N. Rosenbloom, S. C. Sanchez, A. A. Scaife, D. Smith, A. C. Subramanian, L. Sun, D. Thompson, C. C. Ummenhofer, S.-P. Xie, Initialized Earth System prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment* 2, 340–357 (2021).
22. T. DelSole, M. K. Tippett, Predictability in a changing climate. *Clim. Dyn.* 51, 531–545 (2018).
23. P. M. Forster, T. Walsh, C. Smith, W. F. Lamb, R. Lamboll, C. Cassou, M. Hauser, Z. Hausfather, J.-Y. Lee, M. D. Palmer, K. von Schuckmann, A. B. A. Slangen, S. Szopa, B. Trewin, J. Yun, N. P. Gillett, S. Jenkins, H. D. Matthews, K. Raghavan, A. Ribes, J. Rogelj, D. Rosen, X. Zhang, M. Allen, R. M. Andrew, C. Atkinson, R. A. Betts, A. Bombelli, S. N. Burgess, L. Cheng, H. E. Claxton, P. Friedlingstein, T. L. Frölicher, C. M. Domingues, T. Gasser, C. H. Gregory, R. M. Hoesly, D. Huppmann, M. Ishii, C. Kadow, A. Karwat, J. Kennedy, R. E. Killick, M. V. M. Kovilakam, P. B. Krummel, X. Lan, J.-F. Lamarque, A. Liné, B. Martín-Míguez, D. P. Monselesan, C. Morice, J. Mühle, P. Mussak, G. P. Peters, A. Pirani, J. Pongratz, M. Rigby, R. Rohde, A. Savita, S. I. Seneviratne, S. J. Smith, G. Taha, C. Tassone, P. Thorne, C. Wells, L. M. Western, G. R. van der Werf, S. E. Wijffels, M. Zecchetto, J. Zhong, X.-Y. Zhang, V. Masson-Delmotte, P. Zhai, Indicators of Global Climate Change 2025: annual update of key indicators of the state of the climate system and human influence. *Earth Syst. Sci. Data* 18, 3889–3933 (2026).
24. M. J. Alessi, M. A. A. Rugenstein, Surface temperature pattern scenarios suggest higher warming rates than current projections. *Geophys. Res. Lett.* 50, e2023GL105795 (2023).
25. D. Barriopedro, E. M. Fischer, J. Luterbacher, R. M. Trigo, R. García-Herrera, The hot summer of 2010: redrawing the temperature record map of Europe. *Science* 332, 220–224 (2011).
26. C. Delgado-Torres, M. G. Donat, N. Gonzalez-Reviriego, L.-P. Caron, P. J. Athanasiadis, P.-A. Bretonnière, N. J. Dunstone, A.-C. Ho, D. Nicoli, K. Pankatz, A. Paxian, N. Pérez-Zanón, M. S. Cabré, B.

Solaraju-Murali, A. Soret, F. J. Doblas-Reyes, Multi-Model Forecast Quality Assessment of CMIP6 Decadal Predictions. *J. Clim.* 35, 4363–4382 (2022).

27. F. V. Davenport, E. A. Barnes, E. M. Gordon, Combining neural networks and CMIP6 simulations to learn windows of opportunity for skillful prediction of multiyear sea surface temperature variability. *Geophys. Res. Lett.* 51, e2023GL108099 (2024).
28. G. J. van Oldenborgh, S. Drijfhout, A. van Ulden, R. Haarsma, A. Sterl, C. Severijns, W. Hazeleger, H. Dijkstra, Western Europe is warming much faster than expected. *Clim. Past* 5, 1–12 (2009).
29. M. Ting, R. Seager, C. Li, H. Liu, N. Henderson, Future summer drying in the U.s. corn belt and the role of midlatitude storm tracks. *J. Clim.* 34, 1–33 (2021).
30. N. S. Diffenbaugh, Committed acceleration of climate stresses in the coming decades. *Environ. Res.: Climate* 4, 043001 (2025).
31. M. Burke, M. Zahid, M. C. Martins, C. Callahan, R. Lee, T. Avirmed, S. Heft-Neal, M. Kiang, S. Hsiang, D. Lobell, "Are we adapting to climate change?" (w32985, National Bureau of Economic Research, Cambridge, MA, 2024); https://doi.org/10.3386/w32985.
32. D. Olonscheck, L. Suarez-Gutierrez, S. Milinski, G. Beobide-Arsuaga, J. Baehr, F. Fröb, T. Ilyina, C. Kadow, D. Krieger, H. Li, J. Marotzke, É. Plésiat, M. Schupfner, F. Wachsmann, L. Wallberg, K.-H. Wieners, S. Brune, The new Max Planck Institute Grand Ensemble with CMIP6 forcing and high-frequency model output. *J. Adv. Model. Earth Syst.* 15, e2023MS003790 (2023).
33. B. C. O'Neill, C. Tebaldi, D. P. van Vuuren, V. Eyring, P. Friedlingstein, G. Hurtt, R. Knutti, E. Kriegler, J.-F. Lamarque, J. Lowe, G. A. Meehl, R. Moss, K. Riahi, B. M. Sanderson, The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geosci. Model Dev.* 9, 3461–3482 (2016).
34. I. Polkova, S. Brune, C. Kadow, V. Romanova, G. Gollan, J. Baehr, R. Glowienka-Hense, R. J. Greatbatch, A. Hense, S. Illing, A. Köhl, J. Kröger, W. A. Müller, K. Pankatz, D. Stammer, Initialization and ensemble generation for decadal climate predictions: A comparison of different methods. *J. Adv. Model. Earth Syst.* 11, 149–172 (2019).
35. J. S. Risbey, D. T. Squire, A. S. Black, T. DelSole, C. Lepore, R. J. Matear, D. P. Monselesan, T. S. Moore, D. Richardson, A. Schepen, M. K. Tippett, C. R. Tozer, Standard assessments of climate forecast skill can be misleading. *Nat. Commun.* 12, 4346 (2021).
36. H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, Jean-Noël Thépaut, The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* 146, 1999–2049 (2020).
37. Copernicus Climate Change Service (C3S), Climate Indicators, Temperature Indicator (2026). https://climate.copernicus.eu/climate-indicators/temperature.
38. L. Goddard, A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. J. Mason, B. P. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, G. Hegerl, C. A. T. Ferro, D. B.

Stephenson, G. A. Meehl, T. Stockdale, R. Burgman, A. M. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, T. Delworth, A verification framework for interannual-to-decadal predictions experiments. *Clim. Dyn.* 40, 245–272 (2013).

39. D. J. Befort, T. Kruschke, Decadal prediction of the probability of extreme seasons. *Environ. Res. Lett.* 20, 054054 (2025).

Figures

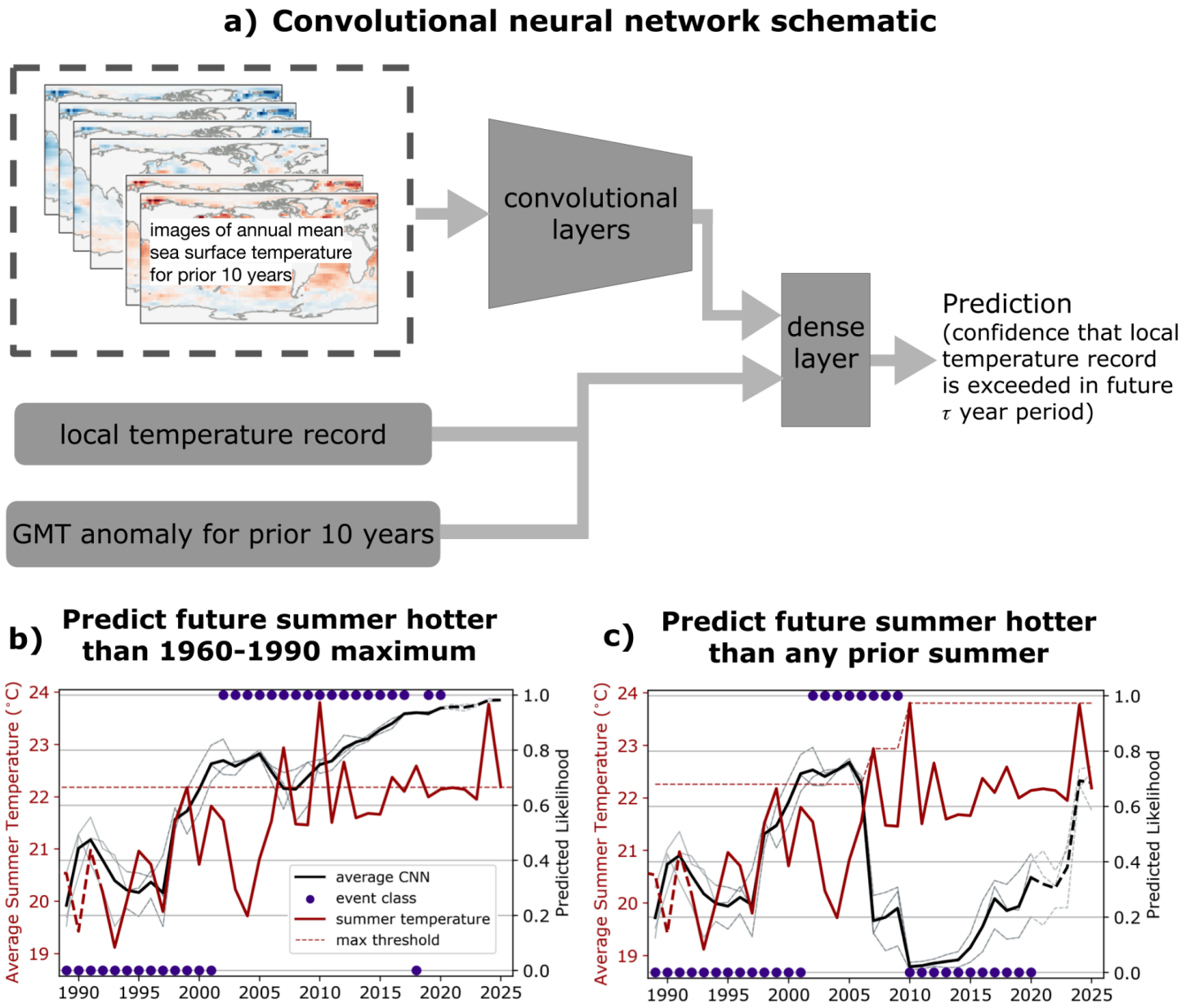


Figure 1

Convolutional neural network schematic and predictions. a) Schematic of the convolutional neural network (CNN) task in this study. Images of the previous 10 years of SST variability are passed through

three sets of convolutions and pooling layers. The output of the convolutions is concatenated with inputs of the current 10-year average GMT anomaly, and the current summertime temperature record. This combined information is passed into the dense layer which outputs a confidence that the inputted temperature record will be exceeded in the next τ years. **b)** CNN's prediction of a future record breaking event occurring within the next five years in Eastern Europe (30E-40E, 40N-50N) (black line) with confidence scale from 0 to 1 on the right y-axis. The event threshold is indicated with the dashed red line, and the ERA5 average summertime temperature is shown in the solid red line, with temperature scale indicated on the left y-axis. The true event class is indicated with indigo dots, with 1 indicating an event occurs within a five year period of that timestep (i.e. average summer temperature in the red solid line will exceed the temperature threshold in the red dashed line with five years of a true event), and 0 indicating no event occurs in the next five years. **c)** As in **b)** but for CNNs predicting an event that exceeds all prior records (indicated by red dashed line).

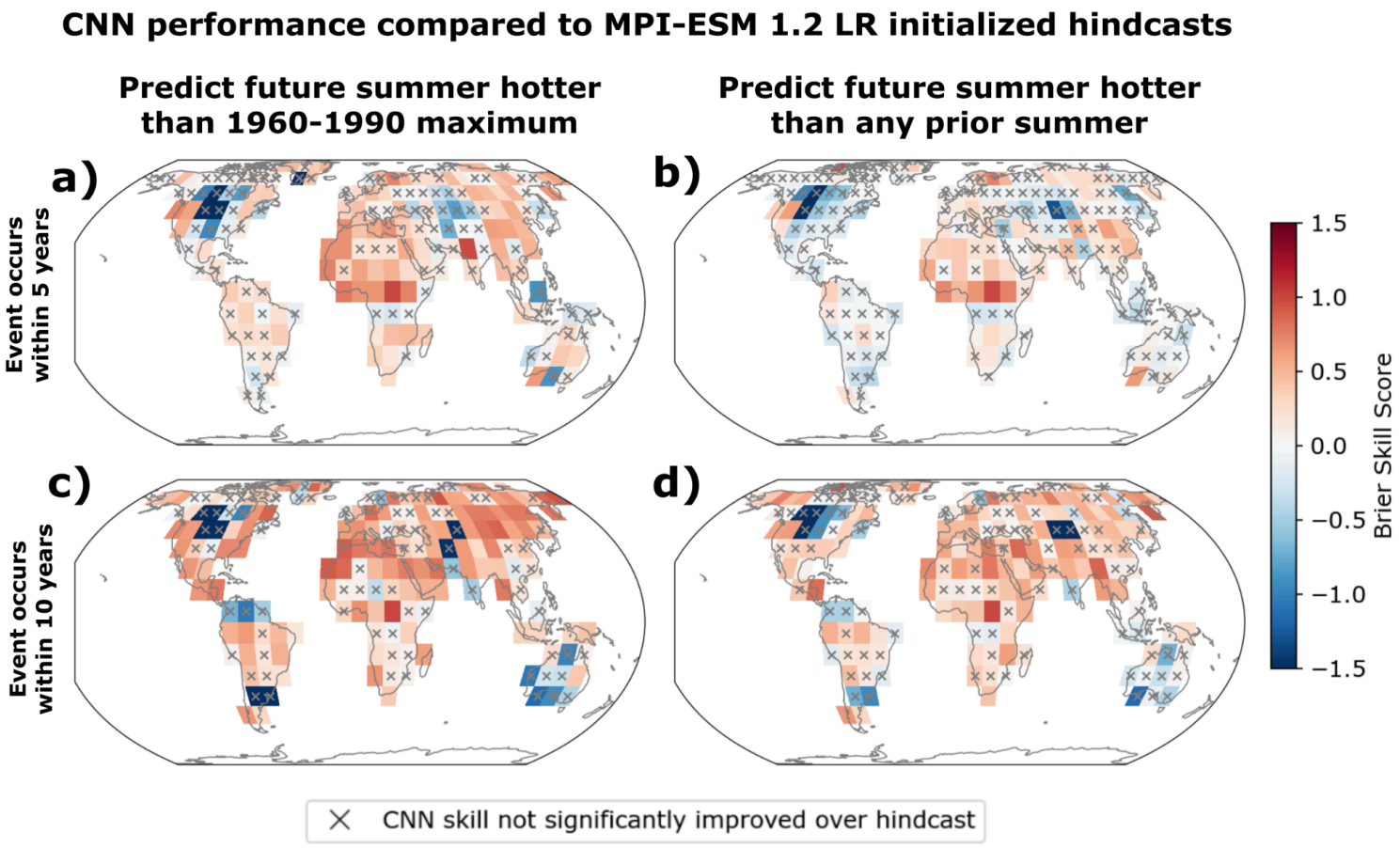


Figure 2

CNN skill at predicting record-breaking summers. a) Brier skill score for CNN prediction of a future summer breaking 1960-1990 record within a future 5 year period with MPI-ESM1.2 LR hindcast as a baseline, where red indicates the BSS is above 0 i.e. CNNs outperform the hindcasts. **b)** Brier skill score (BSS) for CNN prediction of future summer breaking all prior records within a future 5 year period with MPI-ESM1.2 LR hindcast as baseline. **c)** as in Figure 1a) except for CNN predictions of events breaking

the 1960-1990 record within 10 years. **d)** as in Figure 1b) except for CNN predictions of events breaking all records within 10 years.

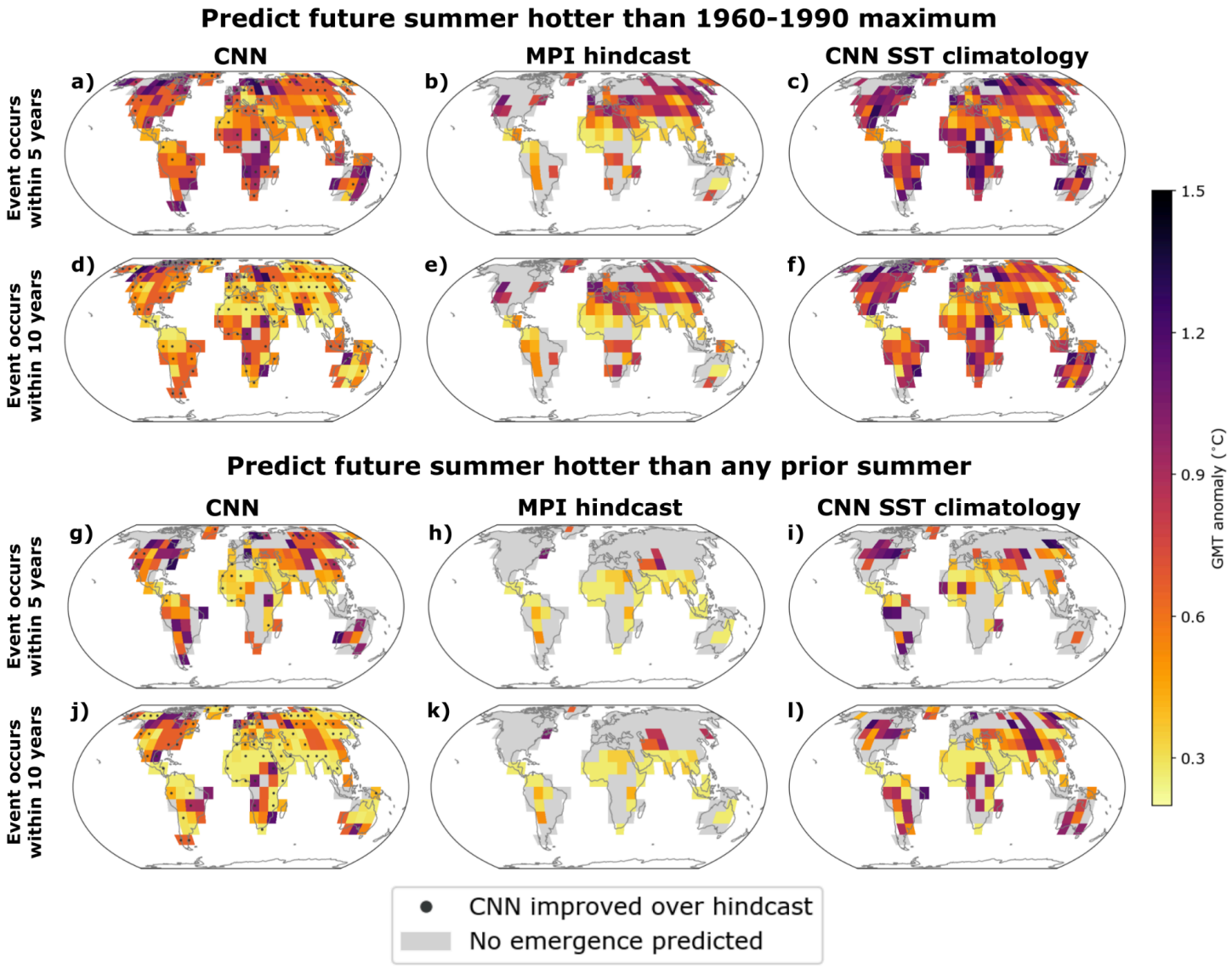


Figure 3

GMT anomaly at which predictability of future record-breaking summers emerges. a) 10-year average GMT anomaly from 1850-1899 when CNNs first predict that a future summer will break the 1960-1990 record within a future 5 year period. Stippling indicates grid boxes where the CNNs's BSS is significantly improved over the MPI hindcast (i.e. un-stippled grid boxes in Figure 1c). Land grid points without data (represented in grey) indicate regions where CNNs never predict a future record-breaking summer in the historical record. **b)** GMT anomaly when the MPI-ESM1.2 LR hindcast first predicts that a record breaking summer will occur in a future 5 year period. **c)** GMT anomaly when the CNNs provided climatological SST first predicts a future record breaking summer will occur in a future 5 year period. **d)** GMT anomaly when CNNs first predict that a future summer will break the 1960-1990 record within a future 5 period. Stippling indicates where CNNs's BSS is better than MPI hindcast (i.e. un-stippled grid boxes in Figure

1e). **e)** GMT anomaly when MPI-ESM1.2 LR hindcast first predicts that a record breaking summer will occur in a future 10 year period. **f)** GMT anomaly when the CNNs provided climatological SST first predicts a future record breaking summer will occur in a future 10 year period.

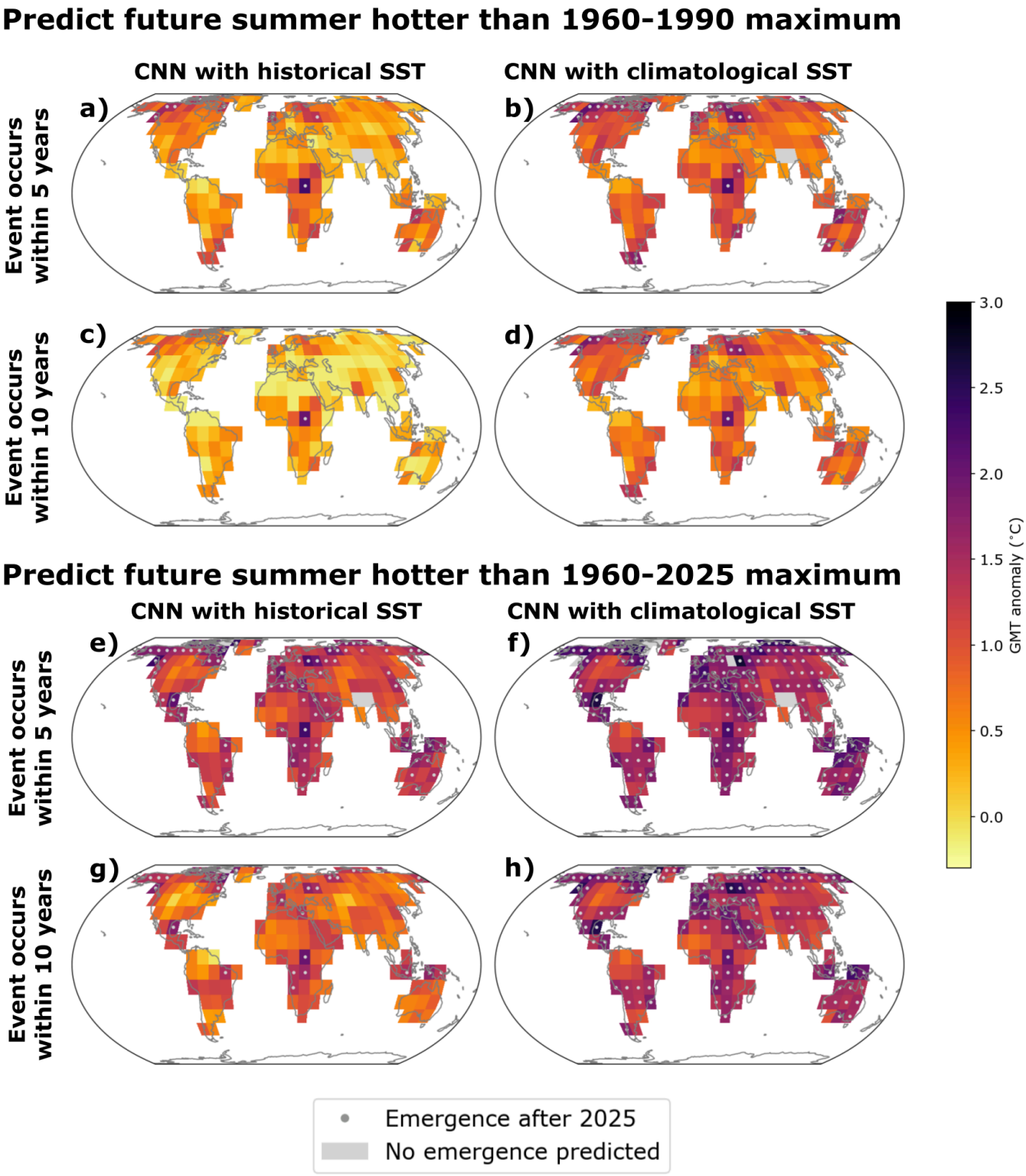


Figure 4

Future predictability emergence from historical SSTs. a) Lowest GMT anomaly at which the CNNs could predict that the average summertime temperature in the next 5 years will exceed the 1960-1990 record when provided with historically plausible SST patterns. **b)** GMT anomaly at which the CNNs predict that the average summertime temperature in the next 5 years will exceed the 1960-1990 when provided climatological SSTs. **c)** As in Figure 4a) but for predicting a recording breaking summer within 10 years. **d)** As in Figure 4b) but for predicting a record-breaking summer in 10 years with climatological SSTs. **e)** Lowest GMT anomaly at which the CNNs could predict that the average summertime temperature in the next 5 years will exceed any summer in the historical record (1960-2025) when provided with historically plausible SSTs. **f)** Lowest GMT anomaly at which the CNNs could predict that the average summer temperature in the next 5 years will exceed any summer in the historical record (1960-2025) when provided with climatological SST. **g)** As in Figure 4e) but for predicting a record-breaking summer within 10 years. **h)** As in Figure 4h) but for predicting a record summer in 10 years with climatological SSTs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PredictabilityEmergenceSupplemental.pdf