Supplemental information for: The emergence of regional record-breaking summer heat predictability

This document contains the supplemental information for Gordon & Diffenbaugh 2026

Bias correction

We use initialized hindcasts from the MPI-ESM1-2-LR hindcast ensemble. Here we provide further information on the bias correction and anomaly calculation used in the study, along with two alternative bias correction methods to ensure that our results are not qualitatively affected by the choice of bias calculation.

1. The bias correction method used in the study can be expressed as:

$$\hat{Y}_{n,m,t} = Y_{n,m,t} - \frac{1}{M}\frac{1}{N-1}\sum_{m=1}^{M}\sum_{n=1}^{N-1} Y_{n,m,t} \quad (1)$$

Where $\hat{Y}$ refers to the bias-corrected hindcast, and *n,m,t* refer to start date, ensemble member, and lead time respectively. This equation estimates the lead time dependent bias as the mean of all start dates and ensemble members up to the current start date, and is guided by the method in Risbey et al. (2021).

Summertime temperatures are represented as the anomaly from the 1960-1990 mean in the observational dataset and hence we also convert the bias-corrected data to anomalies from the 1960-1990 average summer, requiring the climatology to be estimated from the hindcasts. The anomaly calculation method we adopt can be expressed as:

$$Y'_{n,m,t} = \hat{Y}_{n,m,t} - \left[\frac{1}{M}\frac{1}{30}\sum_{m=1}^{M}\sum_{n=1}^{30} Y_{m,t=0,n} - \frac{1}{M}\frac{1}{N-1}\sum_{m=1}^{M}\sum_{n=1}^{N-1} Y_{m,t=0,n}\right] \quad (2)$$

Where Y' refers to the bias-corrected hindcast anomaly prediction. In this method, assuming that n=1 corresponds to 1961 and n=30 corresponds to 1990, we estimate the climatology to be the average lead zero (t=0) hindcast from 1960 to 1990 across all start dates and ensemble members. The second term on the RHS essentially “bias-corrects” the climatology.

2. Bias correction as in Meehl et. al 2022

Another recommended bias correction method is in Meehl et al. (2022), and recommends bias-correcting using only the prior 15 years of hindcasts. This can be expressed as

$$\hat{Y}_{n,m,t} = Y_{n,m,t} - \frac{1}{M}\frac{1}{N-1}\sum_{m=1}^{M}\sum_{n=N-15}^{N-1} Y_{n,m,t} \quad (3)$$

We then calculate the anomaly of the hindcast from its own 1960-1990 climatology (i.e. keeping with the hybrid method from Meehl et al. (2022) which recommends anomaly calculation from the model’s climatology rather than the observational climatology, but updating it to our prediction task):

$$Y'_{n,m,t} = \hat{Y}_{n,m,t} - \left[\frac{1}{M}\frac{1}{30}\sum_{m=1}^{M}\sum_{n=1}^{30} Y_{m,t=0,n} - \frac{1}{M}\frac{1}{N-1}\sum_{m=1}^{M}\sum_{n=N-15}^{N-1} Y_{m,t=0,n} \right] \quad (4)$$

Similar to before we estimate the 1960-1990 climatology to be the lead zero average hindcast across all start dates and ensemble members from this period. Again, the second term on the RHS bias-corrects this anomaly.

3. Testing the effect of anomaly calculation

To test the effect of the anomaly calculation method, we investigate the use of a leadtime dependent anomaly. Taking the bias-corrected hindcast, $\hat{Y}$, to be as defined in (3), we could define the bias-corrected hindcast anomaly, Y' as:

$$Y'_{n,m,t} = \hat{Y}_{n,m,t} - \left[\frac{1}{M}\frac{1}{30}\sum_{m=1}^{M}\sum_{n=1}^{30} Y_{m,t,n} - \frac{1}{M}\frac{1}{N-1}\sum_{m=1}^{M}\sum_{n=N-15}^{N-1} Y_{m,t,n} \right] \quad (5)$$

i.e., define the bias-corrected anomaly as the difference between the bias-corrected hindcast and the bias-corrected anomaly. Keen readers will note that the final term on the RHS is just the definition of the bias in (3) and hence these terms will cancel out, resulting with the bias-corrected hindcasts being inherently defined as the anomaly from the model's 1960-1990 lead time dependent climatology:

$$Y'_{n,m,t} = Y_{n,m,t} - \frac{1}{M}\frac{1}{30}\sum_{m=1}^{M}\sum_{n=1}^{30} Y_{n,m,t} \quad (6)$$

Where, as above, Y' represents the bias-corrected hindcast as the anomaly from 1960-1990.

To demonstrate that the choice of bias correction method does not qualitatively affect our results, we plot the Brier skill score of the CNN for the historical period against the bias-corrected hindcasts in Figure S1. The skill scores for method 2 are represented in Figure S1a)-d), and the skill scores for method 3 are represented in Figure S1e)-h). These figures, compared to Figure 2 in the main text, demonstrates that the choice of bias correction method leaves the results qualitatively unchanged.

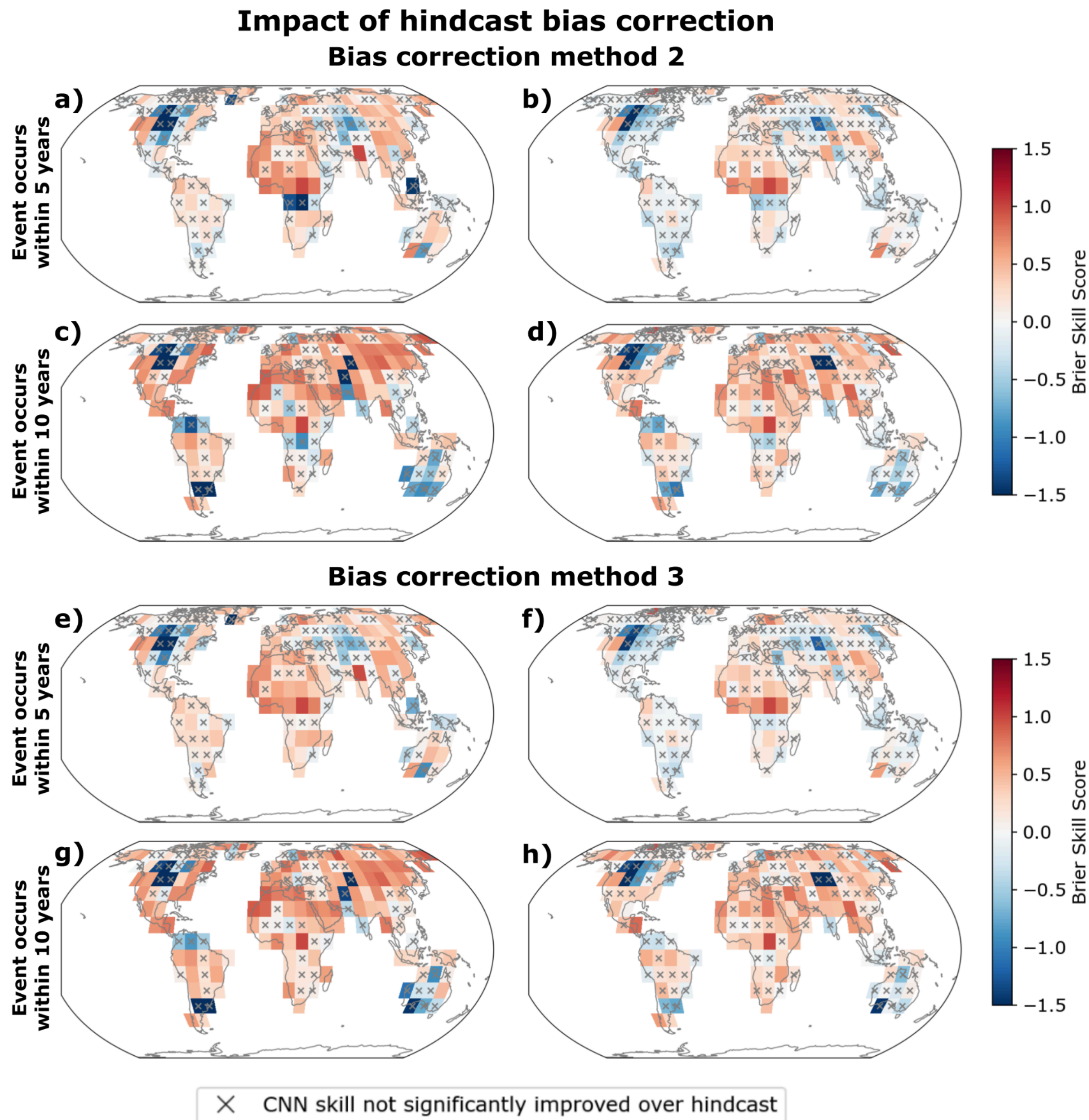

Figure S1 Impact of bias correction a)-d) Brier skill score for the CNN predictions in the historical record compared to the MPI-ESM 1.2 LR hindcasts when bias corrected using method 2 described in the supplemental text. The prediction tasks are: **a)** predict summer that exceeds 1960-1990 maximum within 5 years, **b)** predict summer that exceeds all prior records within 5 years, **c)** predict summer that exceeds 1960-1990 maximum within 10 years, **d)** predict summer that exceeds all prior records within 10 years. **e)-h)** Brier skill score for the CNN predictions in the historical record compared to the MPI-ESM 1.2 LR hindcasts bias corrected using method 3 described in the supplemental text. The prediction tasks are: **e)** predict summer that exceeds 1960-1990 maximum within 5 years, **f)** predict summer that exceeds all prior records within 5 years, **g)** predict summer that exceeds 1960-1990 maximum within 10 years, **h)** predict summer that exceeds all prior records within 10 years.

Impact of observed SST variability in Europe and North America

We test the sensitivity of CNN predictions in North America and Europe to the SST variability input by using SST variability from the 12 left-out testing members from the MPI-ESM 1.2 LR large ensemble. We provide the CNNs with GMT and record temperature thresholds from ERA5, but replace the ERA5 SST inputs with the corresponding dates from the large ensemble for each of the 12 testing members. Figure S2, similar to Figure 1b), demonstrates the CNNs's predictions of a future summer breaking the 1960-1990 maximum within five years in the observational record (black line), with the record threshold indicated in the red dashed line, the observed average summertime temperature in the red solid line, and the event class (record-breaking event or not) indicated by indigo dots. We include the range of predictions from replacing historical SST variability with the large ensemble SSTs in tan shading. We repeat this analysis for three grid points in North America (Figure S2 a)-c)), and three grid points in Western and Central Europe (Figure S2 d)-f)). The grid points are illustrated by the maps at the start of each row, with each point corresponding to each panel from left to right.

In all cases, the CNNs's predictions fall at the upper end of the range of predictions from the large ensemble after around 2015. Since the only difference is the SST input, this suggests that the SST variability in the observational record is responsible for the CNNs's increased confidence in future record-breaking events in all these regions. While these predictions are generally correct in central and eastern Europe, the predicted record breaking events in the late 2010s-early 2020s in central North America did not materialize. Notably, even the lowest prediction from the large ensemble SSTs exceeds 0.5 confidence after 2015 in Figure Sb, and 2020 in Figure Sc, implying the signal of record-breaking heat is consistent in all ensemble members, and therefore contains a contribution from the forced response in central North America. From this evidence we draw two conclusions, firstly, observed SST in the mid-to-late 2010s was in a phase that supported enhanced summertime warming in North America, and Western Europe, including warming exceeding any sampled in this particular subselection of GCM members. Secondly, the CNNs are able to detect the dynamical drivers of warming, leading to the generally correct trend in Europe, however, they can be misled in places where warming has been modulated by thermodynamic feedbacks that are not captured by the GCM, likely leading to their poor performance over central North America.

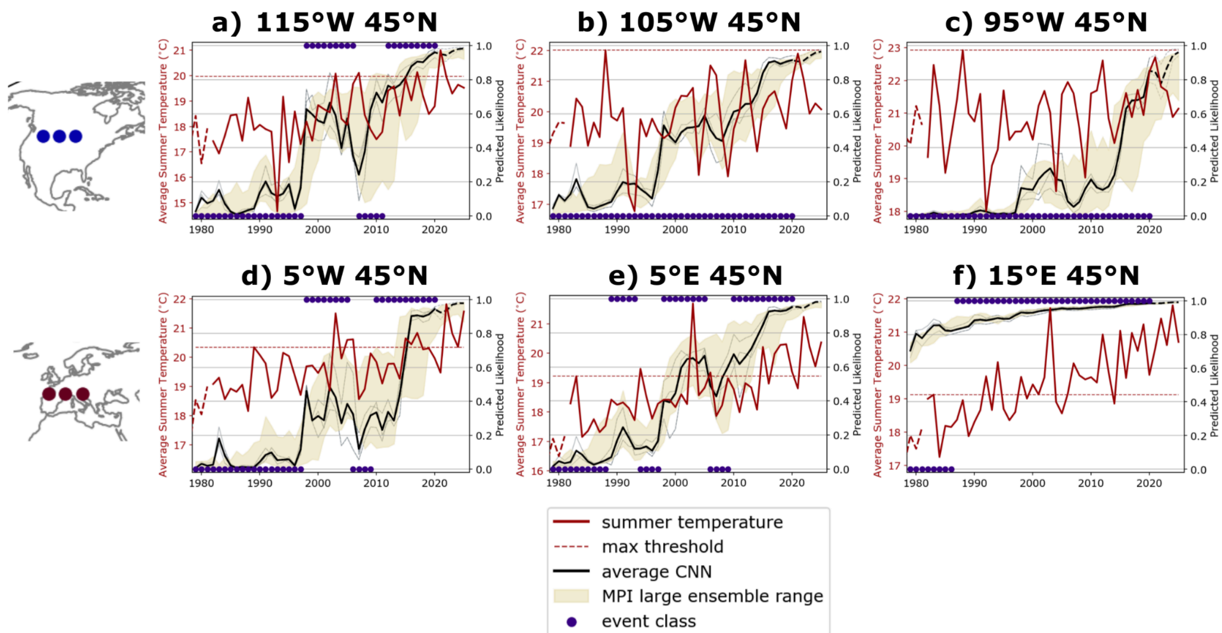

Figure S2 The effect of SST variability on record-breaking heat predictions a) CNN's prediction of a future record breaking event occurring within the next five years in western North America (black line). The event threshold is indicated with the dashed red line, and the ERA5 average summertime temperature is shown in the solid red line. The true event class is indicated with indigo dots, with 1 indicating an event occurs within a five year period of that timestep, and 0 indicating no event occurs in the next five years. The range of CNN prediction confidence garnered when historical SST variability is replaced with SST variability from the large ensemble testing members is shown in tan shading. **b)-f)** as in **a),** but for different grid points as indicated in the panel headings, and visualised by the maps to the left of each row.

SST conditions that are most likely to lead to a prediction of a record breaking summer

We identify the historical SST conditions that are most likely to lead to a prediction of a future record breaking heat event from the analysis in Figure 4. We identify the SST patterns that first lead to predictability emergence at each grid point in Figure 4 and make a histogram of the corresponding time step when that SST pattern occurred in the historical record (Figure S3). Additionally, we identify two more SST patterns that are the next to provide predictability emergence and add their corresponding years to the histogram. From this it is evident that the SST pattern in 2001 is the pattern that most commonly leads to predictability emergence across the globe, followed by the SST pattern around 2022. Note we record the year as the final year before the CNN makes a prediction, so 2001 corresponds to SST variability from 1992-2001, and 2022 corresponds to SST variability from 2013-2022.

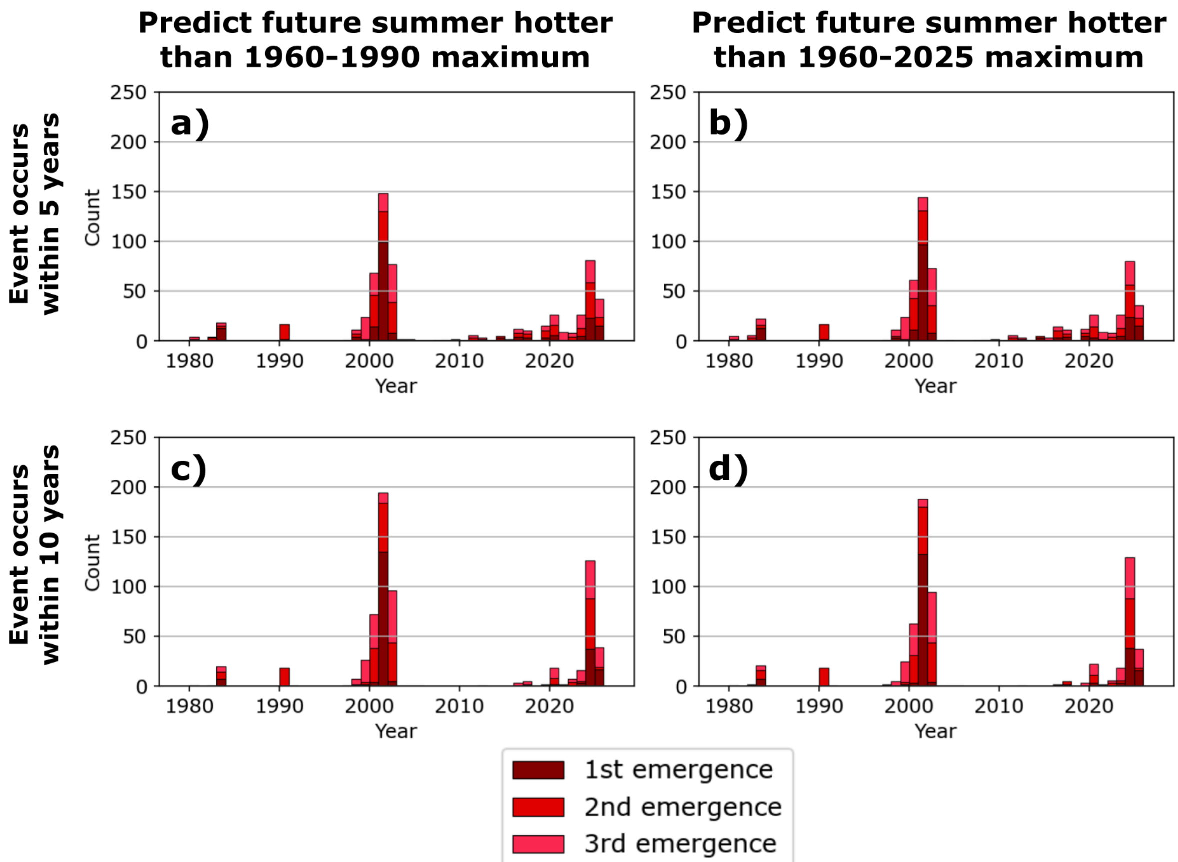

Figure S3 Years of SST emergence patterns. Histogram of the year in the historical record when the SST pattern provided predictability emergence. The darkest bars are the counts for the SST pattern that first corresponded to predictability emergence, with lighter red corresponding to the year where the SST pattern next corresponded to predictability emergence, and the lightest red corresponding to the third SST pattern to provide predictability emergence. **a)** Years corresponding to SST patterns for predicting the 1960-1990 record is broken within 5 years. **b)** Years corresponding to SST patterns for predicting the 1960-2025 record is broken within 5 years. **c)** Years corresponding to SST patterns for predicting the 1960-1990 record is broken within 10 years. **d)** Years corresponding to SST patterns for predicting 1960-2025 record is broken within 10 years.

References

Meehl, Gerald A., Haiyan Teng, Doug Smith, et al. 2022. “The Effects of Bias, Drift, and Trends in Calculating Anomalies for Evaluating Skill of Seasonal-to-Decadal Initialized Climate Predictions.” *Climate Dynamics*, ahead of print, April 30. https://doi.org/10.1007/s00382-022-06272-7.

Risbey, James S., Dougal T. Squire, Amanda S. Black, et al. 2021. “Standard Assessments of Climate Forecast Skill Can Be Misleading.” *Nature Communications* 12 (1): 4346.