An Improved Ensemble of Land Surface Air Temperatures Since 1880 Using Revised Pair-Wise Homogenization Algorithms Accounting for Autocorrelation

DUO CHAN,[a,b] GEOFFREY GEBBIE,[b] AND PETER HUYBERS[c]

[a] *School of Ocean and Earth Science, University of Southampton, Southampton, United Kingdom*
[b] *Department of Physical Oceanography, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts*
[c] *Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts*

(Manuscript received 5 June 2023, in final form 26 December 2023, accepted 16 January 2024)

ABSTRACT: Land surface air temperatures (LSAT) inferred from weather station data differ among major research groups. The estimate by NOAA's monthly Global Historical Climatology Network (GHCNm) averages 0.02°C cooler between 1880 and 1940 than Berkeley Earth's and 0.14°C cooler than the Climate Research Unit estimates. Such systematic offsets can arise from differences in how poorly documented changes in measurement characteristics are detected and adjusted. Building upon an existing pairwise homogenization algorithm used in generating the fourth version of NOAA's GHCNm(V4), PHA_0, we propose two revisions to account for autocorrelation in climate variables. One version, PHA_1, makes minimal modification to PHA_0 by extending the threshold used in breakpoint detection to be a function of LSAT autocorrelation. The other version, PHA_2, uses penalized likelihood to detect breakpoints through optimizing a model-selection problem globally. To facilitate efficient optimization for series with more than 1000 time steps, a multiparent genetic algorithm is proposed for PHA_2. Tests on synthetic data generated by adding breakpoints to CMIP6 simulations and realizations from a Gaussian process indicate that PHA_1 and PHA_2 both similarly outperform PHA_0 in recovering accurate climatic trends. Applied to unhomogenized GHCNmV4, both revised algorithms detect breakpoints that correspond with available station metadata. Uncertainties are estimated by perturbing algorithmic parameters, and an ensemble is constructed by pooling 50 PHA_1- and 50 PHA_2-based members. The continental-mean warming in this new ensemble is consistent with that of Berkeley Earth, despite using different homogenization approaches. Relative to unhomogenized data, our homogenization increases the 1880–2022 trend by 0.16 [0.12, 0.19]°C century^{-1} (95% confidence interval), leading to continental-mean warming of 1.65 [1.62, 1.69]°C over 2010–22 relative to 1880–1900.

SIGNIFICANCE STATEMENT: Accurately correcting for systematic errors in observational records of land surface air temperature (LSAT) is critical for quantifying historical warming. Existing LSAT estimates are subject to systematic offsets associated with processes including changes in instrumentation and station movement. This study improves a pairwise homogenization algorithm by accounting for the fact that climate signals are correlated over time. The revised algorithms outperform the original in identifying discontinuities and recovering accurate warming trends. Applied to monthly station temperatures, the revised algorithms adjust trends in continental mean LSAT since the 1880s to be 0.16°C century^{-1} greater relative to raw data. Our estimate is most consistent with that from Berkeley Earth and indicates lesser and greater warming than estimates from NOAA and the Met Office, respectively.

KEYWORDS: Climate change; Temperature; Climate records; Bias; Changepoint analysis

1. Introduction

Land surface air temperature (LSAT), as measured by weather stations, is crucial for monitoring long-term climate variations, but is also subject to systematic errors including those associated with changes in instrumentation, movement of stations, and changes in measurement environment (Trewin 2010). The process of detecting discontinuities in records and removing biases to better recover climatic variations is generally called homogenization (Peterson et al. 1998; Costa and Soares 2009; Venema et al. 2012). Various homogenization approaches indicate that temperature observations prior to the 1940s need to be adjusted several tenths of a degree Celsius cooler, thereby increasing the implied warming over the last century (Menne et al. 2018; Rohde et al. 2013b). Despite this agreement in the sign of adjustment, the magnitude of adjustments remains uncertain, leading to continental mean temperatures that differ by up to 0.2°C between 1880 and 1940 among existing estimates (Fig. 1), where temperatures anomalies are specified relative to a 1982–2014 average.

The most commonly applied means of homogenizing LSATs is pairwise station homogenization (Menne and Williams 2009, hereafter MW09). This method, which we refer to as PHA_0, is based on comparing one station with its neighbors. PHA_0 has been carefully tested and routinely used for over a decade (Menne and Williams 2009; Lawrimore et al. 2011; Menne et al. 2018).

We briefly review PHA_0 to establish nomenclature. PHA_0 first identifies neighbors for each station according to distance between stations and correlation coefficient in temperature series. As the second step, a standard normal homogenization test (SNHT; Alexandersson 1986) is performed to each difference temperature series between a station and its neighbors to find potential

Corresponding author: Duo Chan, duo.chan@soton.ac.uk

DOI: 10.1175/JCLI-D-23-0338.1

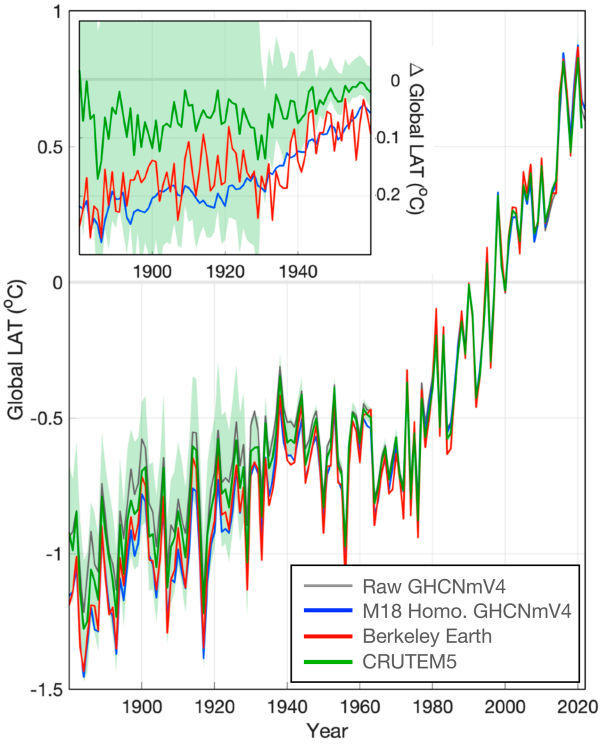

FIG. 1. Continental mean temperature anomalies in existing estimates. Post-1880 temperatures from the raw monthly Global Historical Climatology Network version 4 (GHCNmV4; Menne et al. 2018; gray), homogenized GHCNmV4 (by Menne et al. 2018; blue), Berkeley Earth Temperature (Rohde et al. 2013b; red), and Climate Research Unit Temperature (CRUTEM5; Osborn et al. 2021; green). Anomalies are relative to the mean over 1982–2014. To calculate the global mean series for raw and homogenized GHCNmV4, we first calculate station-wise temperature anomalies using a pairing-and-matching method following Chan et al. (2023). All temperature anomalies are binned to 5° × 5° resolution before averaging over the least common data coverage globally. The green shading shows the 95% c.i. of a 200-member ensemble associated with CRUTEM5, where land temperatures are derived by subtracting HadSST4 (Kennedy et al. 2019) from non-infilled HadCRUT5 (Morice et al. 2021). The upper left panel shows the adjustments to individual datasets relative to the raw GHCNmV4 estimate.

breakpoints. SNHT involves calculating the sum of the squared means of two consecutive segments of a normalized time series:

$$T_0 = \max_{1 \le v < n} [v\bar{z}_1^2 + (n - v)\bar{z}_2^2], \quad (1)$$

where n is the length of the record, v is a time index, and $\bar{z}_1$ and $\bar{z}_2$ are the means over months 1 to v and months $v + 1$ to n, respectively. In contrast to a weighted linear sum of the means that would be invariant to the selection of breakpoint, T_0 is maximized when either $\bar{z}_1$ or $\bar{z}_2$ become large. A null critical value for T_0 is determined by repeatedly realizing T_0 from randomly generated time series. As described further below, these time series have historically been realized as white noise, or devoid of autocorrelation. If the sample value of T_0 exceeds the null critical value, the time series is broken into two segments at the index v that maximizes T_0.

The test is performed iteratively between a splitting phase, where the algorithm tests whether each segment of time series contains any further breakpoints, and a merging phase, where the algorithm combines consecutive segments if the combined time series fail to pass SNHT. After this identification, PHA_0 double-checks each potential breakpoint to confirm that it reflects a break rather than linear trend using a Bayesian information criterion approach (Schwarz 1978).

In the third step, PHA_0 attributes confirmed breakpoints to stations that show the greatest difference with neighbors. In a fourth step, PHA_0 combines attributed breakpoints that are temporally close to one another to account for uncertainties in the timing of identified breakpoints. In the fifth step, an adjustment for each breakpoint is estimated by comparing the station to which a breakpoint is attributed with at least two homogeneous neighbors, and finally, in the sixth step, these adjustments are applied to individual stations relative to values in the last segment.

PHA_0 has been used to homogenize temperatures compiled under the Global Historical Climate Network Monthly version 4 (GHCNmV4; Menne et al. 2018). In addition to central estimates generated using a default combination of PHA_0 parameters, an ensemble generated by perturbing algorithmic parameters in PHA_0 (Williams et al. 2012, hereafter WMW12) is used to quantify uncertainties in GHCNmV4 at global and regional scales. We aim to closely reproduce PHA_0 in MATLAB and comparison against the results from the original FORTRAN software (https://www.ncei.noaa.gov/pub/data/ghcn/v3/software/) indicates consistency.

The paper is organized as follows. Section 2 begins with a test of PHA_0 on a set of simulated temperatures, augmented with random breakpoints, highlighting how the skill of PHA_0 diminishes with increasing autocorrelation in data. In section 3, we introduce two revised algorithms designed to enhance PHA_0 by accounting for autocorrelation. The first, named PHA_1, minimally modifies PHA_0 by adjusting the threshold used in SNHT to be a function of autocorrelation. The second, termed PHA_2, replaces SNHT with a more sophisticated statistical technique known as penalized likelihood (PL; Lund et al. 2023). The skill of both algorithms is assessed in section 4 using simulated temperatures and synthetic data. In section 5 we apply both revised algorithms to station temperatures compiled in GHCNmV4, thereby creating an ensemble of adjusted station temperatures dating back to the 1880s. Finally, in section 6, we discuss our findings, compare our results with existing temperature datasets, and reflect on the implications and future research directions.

2. Applying PHA_0 to perturbed CMIP6 simulations

We evaluate PHA_0 using synthetic cases where we introduced a fixed set of random breakpoints into temperatures simulated by 17 Coupled Model Intercomparison Phase 6 (CMIP6; Eyring et al. 2016) models.[1] We use surface air

[1] Models we use are ACCESS-CM2, CAMS-CSM1-0, CMCC-CM2-SR5, E3SM-1-1, EC-Earth3, EC-Earth3-Veg, EC-Earth3-Veg-LR, FGOALS-f3-L, FGOALS-g3, FIO-ESM-2-0, INM-CM4-8, INM-CM5-0, MIROC6, MRI-ESM2-0, NESM3, NorESM2-LM, and NorESM2-MM.

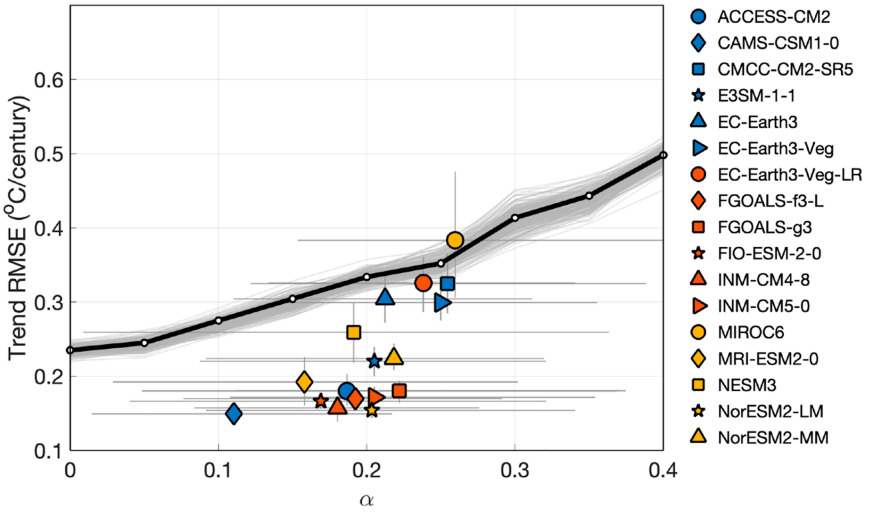

FIG. 2. The skill of the original pairwise station homogenization algorithm (PHA_0) decreases with autocorrelation of climatic signals. The skill of PHA_0 is quantified using the station-wise root-mean-square error (RMSE) of long-term trends over the continental United States after adjustment for 17 CMIP6 models (markers) and synthetic analyses (white dots connected by a black line). RMSE increases with the lag-1 autocorrelation (α) in the difference temperature series between neighbors. The horizontal bar on each marker represents the 95% confidence interval for values of α across individual stations, and the vertical bar is the 95% confidence interval for mean RMSE over all stations. The confidence interval of RMSE is estimated by bootstrapping blocks of 100 stations with replacement.

temperature from the r1i1p1f1 member of each model and concatenate the historical all-forcing experiment from 1970 to 2014 and the SSP5–8.5 experiment from 2015 to 2019. Temperatures are interpolated to the location of U.S. weather stations using a bilinear method to retain the covariance and autocorrelation structures in temperature field. A set of randomly timed breakpoints having random magnitudes are then introduced to each simulation. Appendix A contains details regarding the distribution of breakpoint timing and magnitude.

Breakpoints are identical across models but the skill of PHA_0 in recovering temperature trends, as measured by station-wise root-mean-square error (RMSE), varies by more than a factor of 2 across models (Fig. 2). CAMS-CSM1-0 has the lowest RMSE at 0.15°C century^{-1} (one standard deviation), whereas MIROC6 has the highest RMSE at 0.39°C century^{-1}. We present evidence that the difference in the skill of PHA_0 across models relates to differences in the autocorrelation of temperature. Higher autocorrelation leads to a higher chance of realizing values of T_0 that exceed the critical value by chance. There is a strong correlation across models of 0.75 between the mean lag-1 autocorrelation in the difference temperature series between neighboring stations, referred to as α, and the RMSE between inferred and actual temperature trends (Fig. 2).

To further investigate the relationship between α and the performance of PHA_0, we conduct synthetic analyses using spatially and temporally correlated temperatures. Synthetic temperatures are generated from a multivariate Gaussian process with fixed α values across all stations (see appendix A for details). Synthetic ensembles having larger α are systematically associated with higher RMSE, paralleling the trend found across CMIP6 simulations (Fig. 2) and indicating that differences in autocorrelation are the primary explanation for cross-model differences in skill. These results suggest that accounting for autocorrelation in climate signals may improve the skill of PHA_0 in detecting breakpoints and recovering long-term temperature trends. In this study, we test whether two revised algorithms that account for autocorrelation show improved skill.

In addition to a potential need to account for autocorrelation, we note that, on average, only 55% of identified breakpoints in the CMIP6 synthetic analyses are adjusted, reflecting the fact that PHA_0 estimates adjustments only if several neighboring stations that have homogeneous data are present. It follows that running further iteration of the homogenization algorithm can allow for a greater fraction of breakpoints to be adjusted. We expect such an iterative approach to be especially helpful when changes in instrumentation that may be associated with breakpoints are clustered in space and time, which is known to be the case in the U.S. weather network (Menne and Williams 2009; Williams et al. 2012).

3. Revised pairwise homogenization algorithms

We introduce two revised algorithms that develop upon PHA_0 to account for autocorrelation when identifying breakpoints in the difference series between target–neighbor pairs. Later steps to attribute breakpoints to specific stations, combine

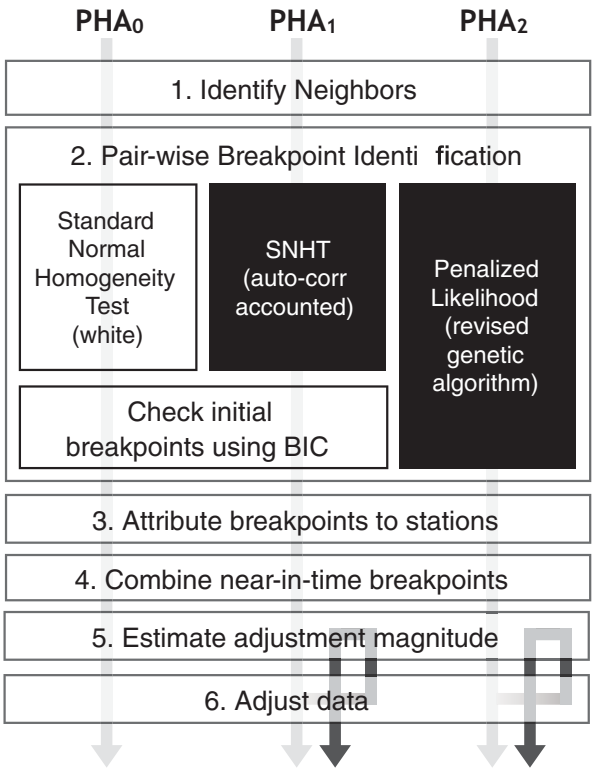

FIG. 3. Schematic of $PHA_{0,1,2}$. PHA_0 steps that are also used by $PHA_{1,2}$ are in white boxes. PHA_1 differs from PHA_0 in the threshold used in the standard normal homogeneous tests (SNHT), and PHA_2 uses penalized likelihood rather than SNHT in pairwise breakpoint identification (black boxes). Both PHA_1 and PHA_2 allow for iterating between estimating and applying adjustments such that breakpoints not adjusted in the first round of estimation can be picked up later (dark arrows).

redundant breakpoints, and estimate the remainder's adjustment are kept the same as in PHA_0 (Fig. 3). In the first revised algorithm, which we call PHA_1, we make a small modification to PHA_0 to adjust the threshold of SNHT according to autocorrelation estimates. The second algorithm, called PHA_2, replaces SNHT with a technique called penalized likelihood (PL; Lund et al. 2023). The main texts focuses on the revision in both algorithms relative to PHA_0, and more details for the full algorithms are in the appendixes.

a. PHA_1

PHA_1 accounts for autocorrelation by making critical values a function of both series length n and lag-1 autocorrelation α. To estimate these critical values, we model temperature difference time series as an order-1 autoregressive process:

$$x_t = \alpha x_{t-1} + \epsilon_t. \tag{2}$$

In Eq. (2), α is the system memory and ϵ is white noise drawn from a standard normal distribution $N(0, 1)$. We explore values of α between 0 and 0.4, a typical range across CMIP6 simulations, and values of n between 5 and 3500. For each combination of α and n, we generate 50 000 random series and normalize each to calculate the SNHT test statistics T_0 following Eq. (1). Higher values of α give greater autocorrelation and increased SNHT critical values above which a breakpoint is provisionally identified. For example, for a series with more than 500 time steps, the 90% threshold of T_0 when α equals 0.3 is more than 1.7 times of the threshold when α equals zero.

We use a sliding window method to estimate α. This approach is useful because breakpoints in a time series can systematically shift subsequent data points, leading to positively biased α if calculated directly from a series containing these disruptions. Estimating α from shorter segments reduces this bias because a shorter segment is less likely to contain breakpoints. Specifically, a window's length is chosen as the smaller of two values: 100 months or one-third of the total length of the time series, and the window moves sequentially from the start to the end of the series. Given that α is assumed to be temporally stationary but α estimates might contain occasional outliers due to breakpoints, the median of all α values calculated across these windows is used as the final estimate. This approach not only accounts for variations within the series but also helps to mitigate the impact of any individual outlier segments. Additionally, to refine the accuracy of α estimation, its value is recalculated at the beginning of a splitting phase by excluding windows overlapping with any time steps marked as breakpoints.

Following the discussion near the end of section 2, PHA_1 iterates between estimation (step 5 in Fig. 3) and implementation of adjustments (step 6 in Fig. 3). Specifically, in the second iteration, we send not-yet-adjusted breakpoints, due to either having insufficient homogeneous neighbors or nonsignificant adjustment estimates, back to adjustment estimation. Adjustments of these breakpoints are estimated against adjusted data after the first iteration. This process can be iterated until no further adjustments are possible.

b. PHA_2

Unlike PHA_0 and PHA_1, which are based on SNHT, PHA_2 uses a penalized likelihood-based approach (Lund et al. 2023) to identify multiple breakpoints by selecting among a set of models containing all possible combinations of breakpoint timing (Fig. 3). The penalized likelihood method gets its name from using a loss function that is defined to maximize data likelihood while penalizing, in this case, higher numbers of breakpoints. Compared with SNHT-based approaches, penalized likelihood shows better skill in identifying breakpoints (Shi et al. 2022b).

The computational cost of penalized likelihood approaches to identifying breakpoints can be large because the number of candidate models goes as 2^n, where n is the number of time steps. It follows that penalized likelihood approaches have generally been applied in the climate sciences to single, short time series including annual global and regional mean surface temperatures (Li and Lund 2012; Beaulieu and Killick 2018; Shi et al. 2022a), single-station annual precipitation (Li and Lund 2012), annual sea ice coverage (Lund et al. 2023), and

the Pacific decadal oscillation index (Beaulieu and Killick 2018).

Application of a penalized likelihood approach to monthly temperatures from the global network of weather stations is computationally challenging because the network contains approximately 28 000 stations that each span more than 1000 months (Menne et al. 2018). Moreover, skillful breakpoint detection involves comparing each station against 20–80 neighbors (Williams et al. 2012). Approximate solutions can be obtained, however, using various genetic algorithm approaches (Killick et al. 2012; Li and Lund 2012). A further issue is that the records we are dealing with have missing data, and, to our knowledge, this issue has not been accounted for in previous applications.

We model an interstation difference series $\Delta\mathbf{T}$ as the summation of differences in climatic variability $\Delta\mathbf{C}$ and differences induced by breakpoints $\Delta\mathbf{D}$. For each monthly time step t, where $t = 1, 2, \ldots, n$, we have

$$\Delta T_t = \Delta C_t + \Delta D_t[+\gamma t]. \tag{3}$$

The term γ_t denotes an optional linear trend, and the full version of model formulation containing this trend term is in appendix C. Similar to Eq. (2), ΔC_t is modeled as an order-1 autoregressive process:

$$\Delta C_t = \alpha \Delta C_{t-1} + \epsilon_t. \tag{4}$$

The choice of an autoregressive order-1 process in modeling temperatures follows wide use in the literature (e.g., Tingley 2012). For purposes of simplicity, we do not distinguish between random observational error and temperature variability that is independent between stations.

To account for autocorrelation, ΔT_t is prewhitened by substituting Eq. (3) for time steps t and $t-1$ into Eq. (4):

$$Y_t = \Delta T_t - \alpha \Delta T_{t-1} = \Delta D_t - \alpha \Delta D_{t-1} + \epsilon_t. \tag{5}$$

In the case where an interval of data is missing, prewhitening uses the closest previous time step having data. To account for the greater expected variance associated with multiple time steps between data points, we first define $k_t - 1$ as the number of missing data in a row before time step t. A prewhitened version of Eq. (5) with stationary variance can be defined as

$$Y_t = \frac{\Delta T_t - \alpha^{k_t} \Delta T_{t-k_t}}{S_{k_t}} = \frac{\Delta D_t - \alpha^{k_t} \Delta D_{t-k_t} + \sum_{i=1}^{k_t} \alpha^{i-1} \epsilon_{t-i+1}}{S_{k_t}}. \tag{6}$$

The summation denotes an accumulated random component over unsampled time steps, and $S_{k_t} = \sqrt{(1-\alpha^{2k_t})/(1-\alpha^2)}$ is a factor used to scale the random component to have the same variance as Eq. (5). For the initial condition, we have $Y_1 = \Delta T_1 / S_{k_1}$, where $S_{k_1} = S_\infty = \sqrt{1/(1-\alpha^2)}$.

The vector $\Delta\mathbf{D}$ defines the offset series segmented by breakpoints. Following Li and Lund (2015), we represent $\Delta\mathbf{D}$ using $\mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is a design matrix with dimensionality $n \times s$. The term s is the number of segments or the number of breakpoints plus one. Entries of $\mathbf{X}$ are zeros and ones, indicating in which segment a time step lies. Vector $\boldsymbol{\beta}$ indicates the mean value in each segment and has the dimensionality $s \times 1$. Given the timings of breakpoints (i.e., $\mathbf{X}$), $\boldsymbol{\beta}$ (and likewise γ) is found using an ordinary least squares fitting.

The loss function of a model fit is

$$L = n \ln(2\pi) + \sum_{t=1}^{n} \ln(e_t) + \sum_{t=1}^{n} \frac{S_{k_t}^2 (Y_t - \hat{Y}_t)^2}{e_t} + [2(s-1) + 3]\ln(n), \tag{7}$$

where time steps with missing values are excluded from the loss calculation. The error term e_t is quantified as $S_{k_t}^2$ times the residual variance of the fitting. The first three terms of Eq. (7) sum to -2 times the log likelihood of data, assuming ϵ_t independently and identically follows a Gaussian distribution. The fourth term represents a penalty formulated according to the Bayesian information criterion (Shi et al. 2022a). In case a trend is also fitted, the penalty becomes $[2(s-1)+4]\ln(n)$. Because $\boldsymbol{\beta}$ (and γ) found using the ordinary least squares minimizes Eq. (7) conditional on the timing of breakpoints $\mathbf{X}$, the problem hence transforms into finding the combination of the timing of breakpoints that minimizes Eq. (7).

Killick et al. (2012) and Li and Lund (2012) proposed specific genetic algorithms for optimization. We implemented both approaches and find that both converge to the minimum loss within 10 s for series shorter than 200 time steps. However, when tested on synthetic series longer than 1000 time steps, which is common for the data we focus on, their solutions generally do not converge within 5 min and, when terminated, contain small-magnitude breaks that are nonoptimal. Such performance is problematic given the large number of differences we seek to evaluate. The existence of small-magnitude false alarms is also unsatisfactory because false alarms in neighboring stations reduce the number of homogeneous neighbors and, thereby, prevents the estimation of adjustments present (Menne and Williams 2009).

To overcome these issues, we introduce a multiparent algorithm that improves speed and converges through breaking long time series into shorter segments and treating each segment independently when generating descendants. More details of the multiparent genetic algorithm are in appendix C. Our updated genetic algorithm can find the optimal solution for a 100-yr time series within 20–60 s, significantly reducing the computation time.

For purposes of estimating a trend, we run the penalized likelihood breakpoint identification algorithm two times, once with and once without a linear trend, and accept the one with lower loss. Note that because PL avoids mis-identifying breakpoints in long-term trends by optimizing globally, there is no need to double check if each of the identified breakpoints represents long-term trends as in PHA_0 and PHA_1 (Fig. 3).

4. Applying PHA_1 and PHA_2 to simulations and synthetic data

We assess the skill of our revised algorithms, PHA_1 and PHA_2, relative to PHA_0 using perturbed CMIP6 simulations

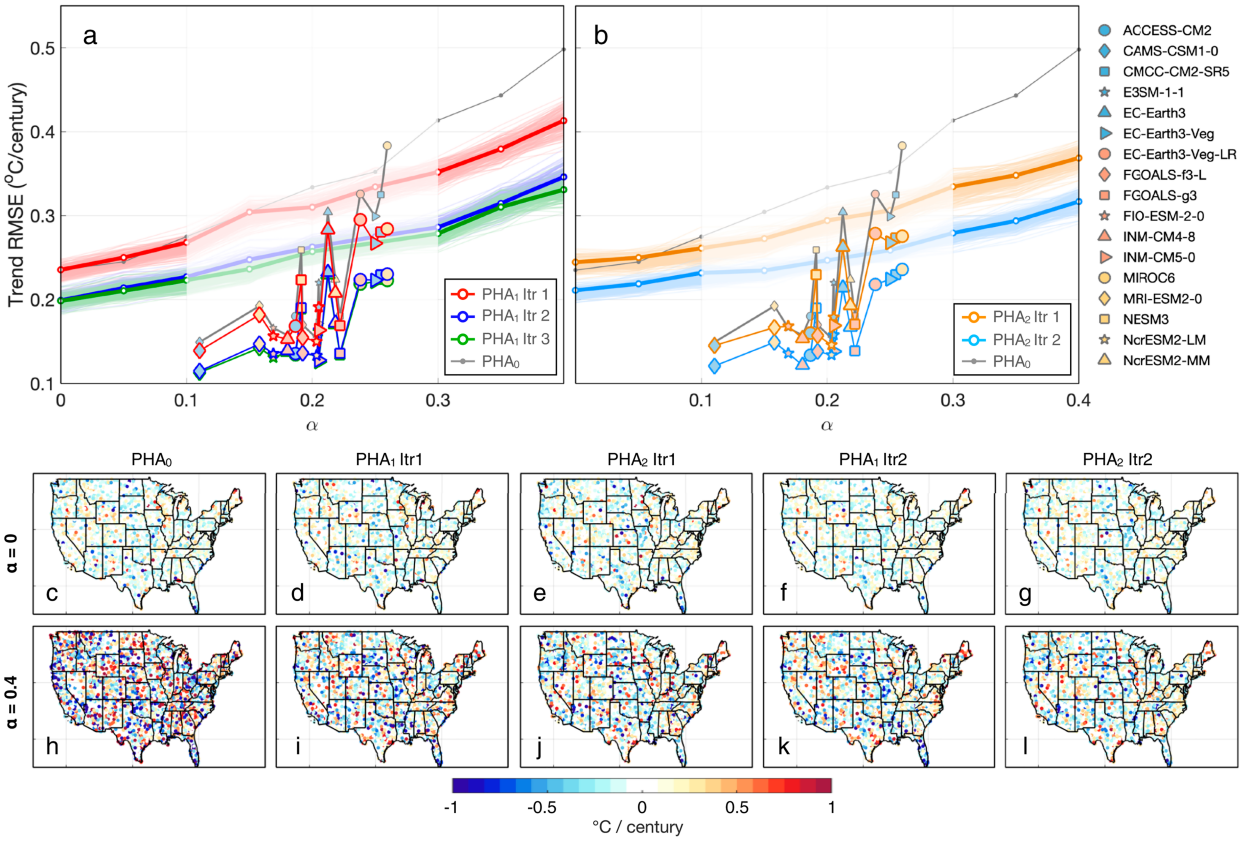

FIG. 4. Skill of the revised pairwise homogenization algorithms in recovering long-term trends. (a) Trend RMSE for the multivariate Gaussian process (MGP; lines) and CMIP6 (markers) ensemble after running PHA_1 for one (red), two (blue), and three (green) iterations. Results from PHA_0 (gray) are included for comparison. (b) As in (a), but for PHA_2, which runs for two iterations (orange for the first and light blue for the second). (c)–(l) Maps of trend error. (from left to right) Results for PHA_0, PHA_1 with one iteration, PHA_2 with one iteration, PHA_1 with two iterations, and PHA_2 with two iterations, respectively. The MGP synthetic case with (c)–(g) $\alpha = 0$ and (h)–(l) $\alpha = 0.4$.

and a synthetic data ensemble generated from a multivariate Gaussian process (MGP). The revisions in both algorithms improve skill. We also show that the reason for improved skill is that the revised algorithms correctly identify more breakpoints while being subject to fewer false alarms, or false alarms that are of small magnitude and, thus, have little effect on long-term trends. Unless otherwise stated, both PHA_1 and PHA_2 are run using a default parameter combination (Williams et al. 2012; see ensemble 1 in Table B2 herein).

a. *The performance of PHA$_1$*

To evaluate the performance of PHA_1, we begin by comparing the root-mean-square error (RMSE) of long-term temperature trends on the MGP-based synthetic ensemble (Fig. 4a). After a single iteration, trend RMSE values in PHA_1 are, on average, 0.32°C century^{-1}, a value that is 0.03°C century^{-1} lower than PHA_0. The reduction in RMSE increases with the strength of the autocorrelation, α, from zero when α is zero to 0.09°C century^{-1} when α is 0.4. Running estimation and adjustment multiple times results in another systematic reduction in RMSE that is less dependent on autocorrelations. The reduction in RMSE is, averaged over α from 0 to 0.4, 0.05°C century^{-1}. A third iteration reduces RMSE by less than 0.01°C century^{-1}.

The improvement in skill shown by PHA_1 is consistent when applied to perturbed CMIP6 simulations (Fig. 4a). When running PHA_1 for one iteration, the reduction in RMSE ranges from 0.01°C century^{-1} in CAMS-CSM1-0 (the model with the lowest α) to 0.09°C century^{-1} in MIROC6, the model with the highest α. The RMSE reduction across CMIP6 models approximately follows a one-to-one relationship with that of the MGP synthetic ensemble, indicating that autocorrelation in temperature variability is a sufficient explanation of differences in skill. Although autocorrelation can be expected to vary across regions, the linear scaling of RMSE with α indicates that the global average is an appropriate metric. Running PHA_1 a second time further reduces RMSE in the CMIP6 ensemble by an average of 0.04°C century^{-1} but there is no significant change after a third iteration. These improvements suggest that PHA_1 may have better breakpoint identification under the regime of high autocorrelation.

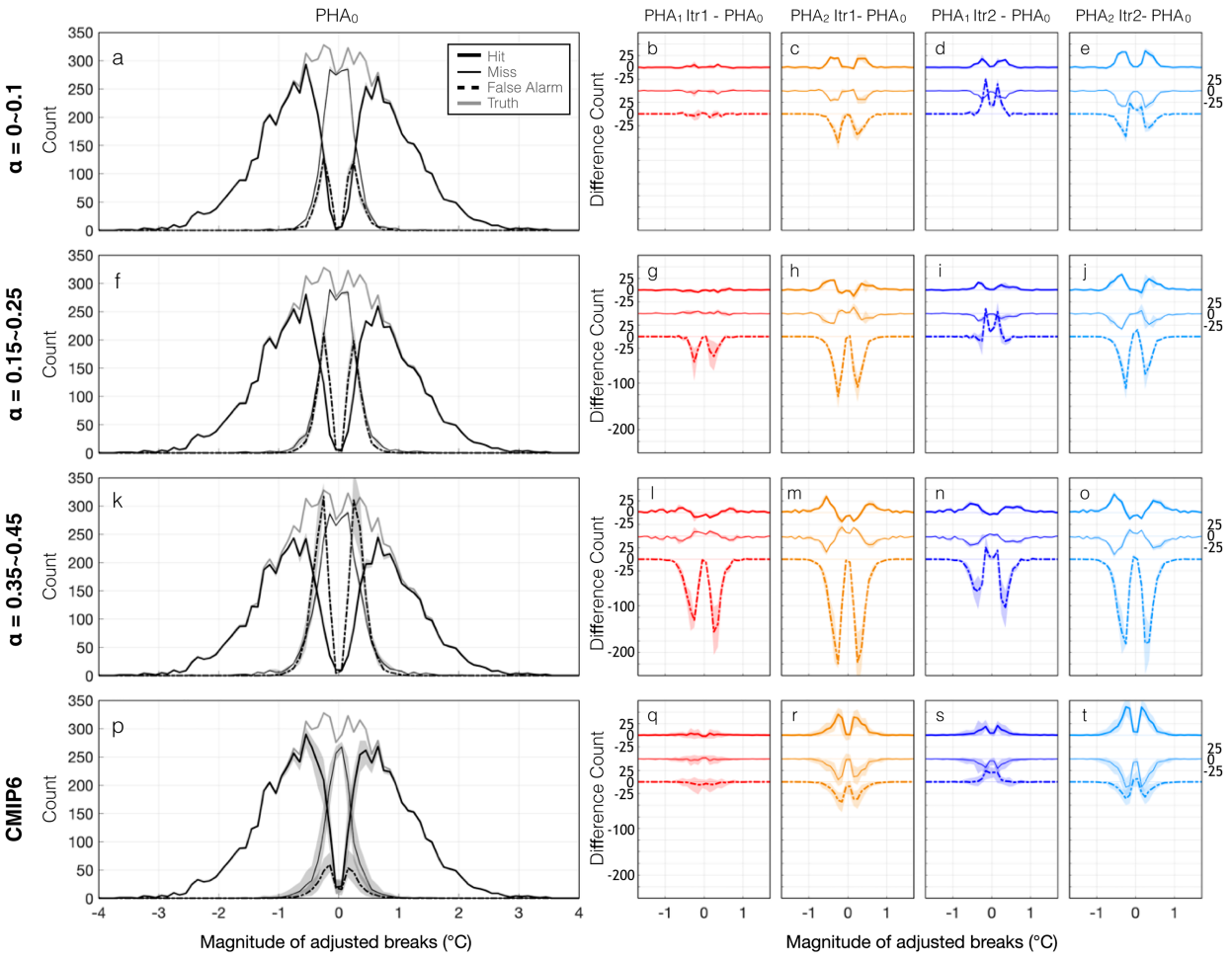

FIG. 5. Histograms of hits, misses, and false alarms. (left) Histograms of hits (thick solid), misses (thin solid), and false alarms (dashed) using PHA_0. (right) The differences between the results of PHA_0 and one iteration of PHA_1 (red) and PHA_2 (orange), or two iterations of PHA_1 (blue) and PHA_2 (light blue). Lines are offset for visibility. (from top to bottom) Rows show results averaged over three synthetic analyses for α = 0–0.1, α = 0.15–0.25, and α = 0.35–0.45, as well as a 17-model CMIP6 ensemble. The shadings indicate the range across MGP ensemble members or CMIP6 models.

To demonstrate this improvement in breakpoint identification, we develop a scoring system by counting the number of hits, misses, and false alarms. Specifically, a hit is if a breakpoint is identified within a 2-yr epoch that centers on the timing of a true breakpoint. Breakpoints identified outside of an epoch are considered false alarms, and epochs not identified to have a breakpoint are misses. Changing the length of this epoch to 1–3 years does not qualitatively change our results.

The improvement associated with running PHA_1 for the first iteration comes mainly from reducing false alarms. As α increases, PHA_0 achieves fewer hits and gives more false alarms (Figs. 5a,f,k). Among the 8142 introduced breaks, the number of hits decreases from 6334 with an α of 0 to 5715 when α is 0.4, whereas false alarms increase from 573 to 1972 (Figs. 5a,f,k). When α is 0, PHA_1 behavior is the same as PHA_0 (Fig. 5b). As α increases, PHA_1 makes fewer false alarms than PHA_0, with 185 fewer when α is 0.2 (Fig. 5g) and 930 fewer when α is 0.4 (Fig. 5l). Such a reduction is consistent with accounting for autocorrelation, whereby a higher T_0 threshold limits SNHT mis-identifying climatic variations as breakpoints. Interestingly, the higher threshold does not lead to a decrease in hits or an increase in misses because, although there are fewer initially identified breakpoints, the subsequent steps in the algorithm ultimately lead to essentially no net change in hits and misses.

The improvements associated with running PHA_1 for a second iteration come mainly from increasing the number of hits. Over all α values examined, PHA_1 with two iterations makes 116 [78, 145] (95% confidence interval, hereinafter c.i.) more hits than PHA_0 (Figs. 5d,i,n). In this case, increasing the hit rate also leads to an increase in the false alarms (Figs. 5d,i,n), although the median absolute magnitude of additional hits is larger at 0.28°C, as compared to 0.14°C for false alarms. As a result, the effect of increasing the hit rate is advantageous for purposes of reducing RMSE. Qualitatively similar decreases in false alarm rates and increases in hit rates are also found

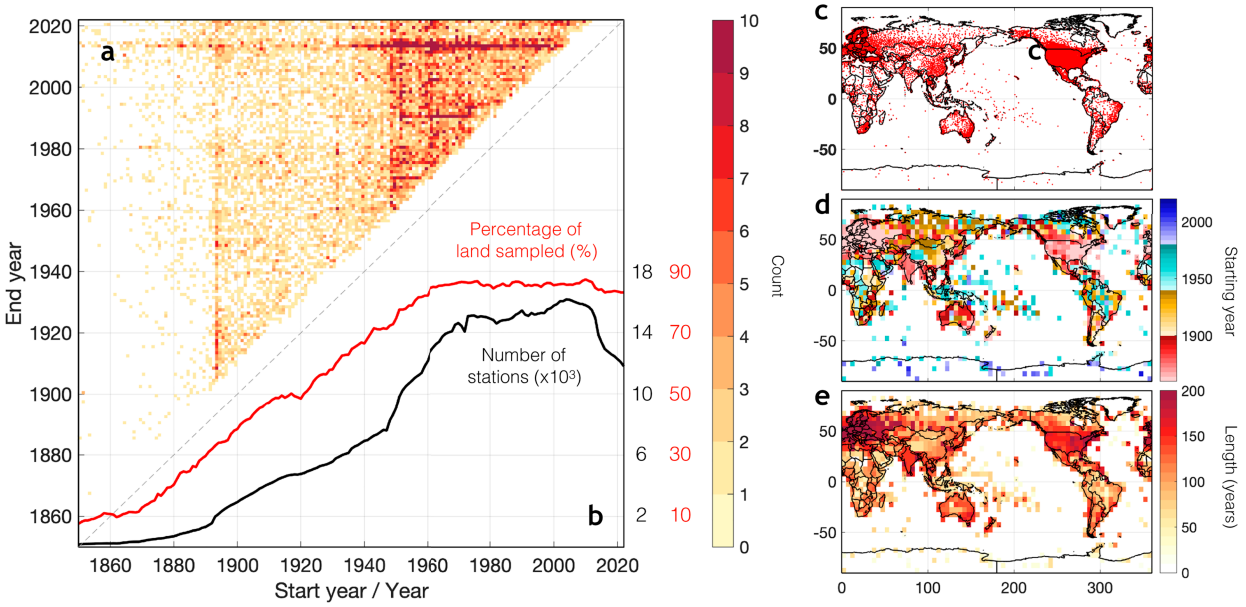

FIG. 6. Statistics of GHCNmV4. (a) Two-dimensional histogram of the starting and ending year of weather stations used in this study. (b) Number of stations as a function of year (black; units are in thousands of stations) and the percentage of land area sampled (red). Percentages are calculated after binning the station coverage to 5° × 5° grids. (c) Locations of 27 868 weather stations used in this study. (d) The earliest sampled year in each 5° × 5° grid box. (e) The length of the sampled period at each grid box in years, i.e., the number of monthly data points divided by 12.

when applying a second iteration of the algorithm to synthetic data developed from CMIP6 simulations (Figs. 5q,s).

b. *The performance of PHA$_2$*

PHA$_2$ produces results that are slightly better than PHA$_1$ with respect to trend RMSE (Fig. 4b). When $\alpha > 0.1$, PHA$_2$ with one iteration decreases trend RMSE by an average of 0.03°C century^{-1} more than PHA$_1$ when applied to the synthetic MGP-based synthetic data. A second iteration of PHA$_2$ decreases trend RMSE relatively less than the second iteration of PHA$_1$, but the overall performance of PHA$_2$ with two iterations is still better than PHA$_1$ by an average of 0.02°C century^{-1}. Application of PHA$_2$ with two iterations to synthetic data derived from CMIP6 simulations also gives results that are slightly better in terms of RMSE than those of PHA$_1$ (Figs. 4a,b).

The fact that PHA$_2$ shows slightly lower RMSE than PHA$_1$ is consistent with PHA$_2$ tending to identify both more true breakpoints and give fewer false alarms than PHA$_1$ (Fig. 5). Overall, PHA$_2$ identifies 3% more breakpoints than PHA$_1$ and give 60% fewer false alarms (Figs. 5c,h,m). The increase in hits arises partly from the fact that PHA$_2$ does not use adjacent segments to double check if initially identified breakpoints indicate local trends (see step 3 in Fig. 3). PHA$_2$, therefore, keeps potential breaks that otherwise can be excluded as linear trends. The decrease in false alarms reflects that the penalized likelihood method tends to reject small breaks by optimizing globally. These results are consistent with previous findings that penalized likelihood methods generally identify breakpoints more accurately than likelihood ratio tests such as SNHT (Shi et al. 2022b).

5. Analysis of GHCN monthly temperatures

Having established the skill of PHA$_1$ and PHA$_2$ using trials on synthetic data, in this section, we apply them to monthly air temperatures compiled within the Global Historical Climatology Network (GHCNm) version 4 (Menne et al. 2018). GHCNmV4 contains monthly mean temperatures from 27 868 stations (Fig. 6c). The number of stations increases from the 1850s to the 1970s, plateaus from the 1970s to the 2000s, and declines thereafter (Fig. 6b). Records prior to the 1900s are mainly from Europe, the United States, India, coastal Australia, and Japan (Fig. 6d). More than 3000 stations have records longer than 100 years (Figs. 6a,e). Despite the recent drop in total number of stations, the percentage of sampled land area, calculated by counting 5° × 5° grid boxes, remains approximately 85% throughout the past 60 years (Fig. 6b). To perform an initial quality screening, we exclude records having QC flags that identify possible issues including duplication, outlier behavior, spatial inconsistency, and isolation (Menne et al. 2018).

a. *Breakpoint detection and temperature adjustments under the default parameter combination*

Under the default parameter combination (Table B2, ensemble 1), applying PHA$_0$ to the quality-controlled stations leads to identification of 61 500 breakpoints between 1880 and 2023 (black in Fig. 7a). In comparison, NOAA's homogenized

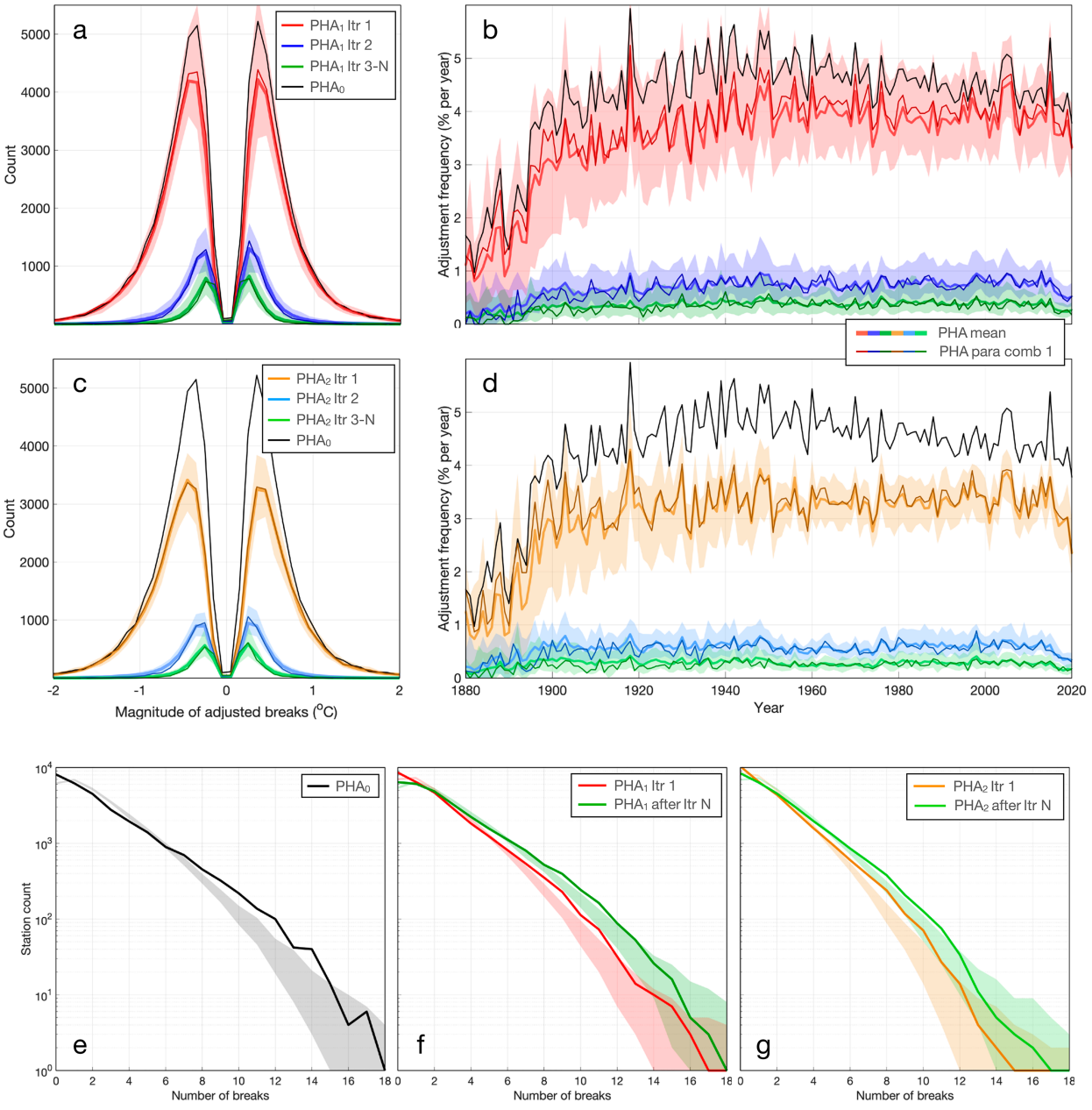

FIG. 7. Adjusted breakpoints in GHCNmV4. (a) Histogram of the magnitude of newly adjusted breakpoints using PHA_1 for the first (red), second (blue), and third-to-the-last (green) iterations. The estimation runs until fewer than 100 breakpoints are newly adjusted, which we call iteration *N*. Shown results are for parameter combination 1 (thin line; our default parameter combination), mean over 50 members (thick line), and range (shading). Results for PHA_0 (thin black) are shown for comparison. (b) As in (a), but for the rate of adjustment as a function of year. (c),(d) As in (a) and (b), but for the PHA_2 ensemble. (e) Histogram of the number of adjustments in each station (black) for PHA_0. Also shown is the 95% c.i. (shading) over a 500-member ensemble generated from binomial distributions, assuming the occurrence of breakpoints within a station is temporally independent. (f) As in (e), but for PHA_1 parameter combination 1 after one (red) and *N* iterations (green). (g) As in (f), but for PHA_2.

GHCNmV4 version (Menne et al. 2018) contains approximately 71 000 breaks from 1880 to 2016. We are unsure as to the origin of the discrepancy in the number of reported breaks, though one possible reason is that we do not use metadata in our PHA analyses. The code for PHA_0 and detailed results are available in order to facilitate intercomparison going forward (see the data availability statement). Compared with PHA_0, PHA_1 makes fewer false alarms in synthetic analysis (Fig. 5). When applied to GHCNmV4, PHA_1 with the same parameter combination makes only

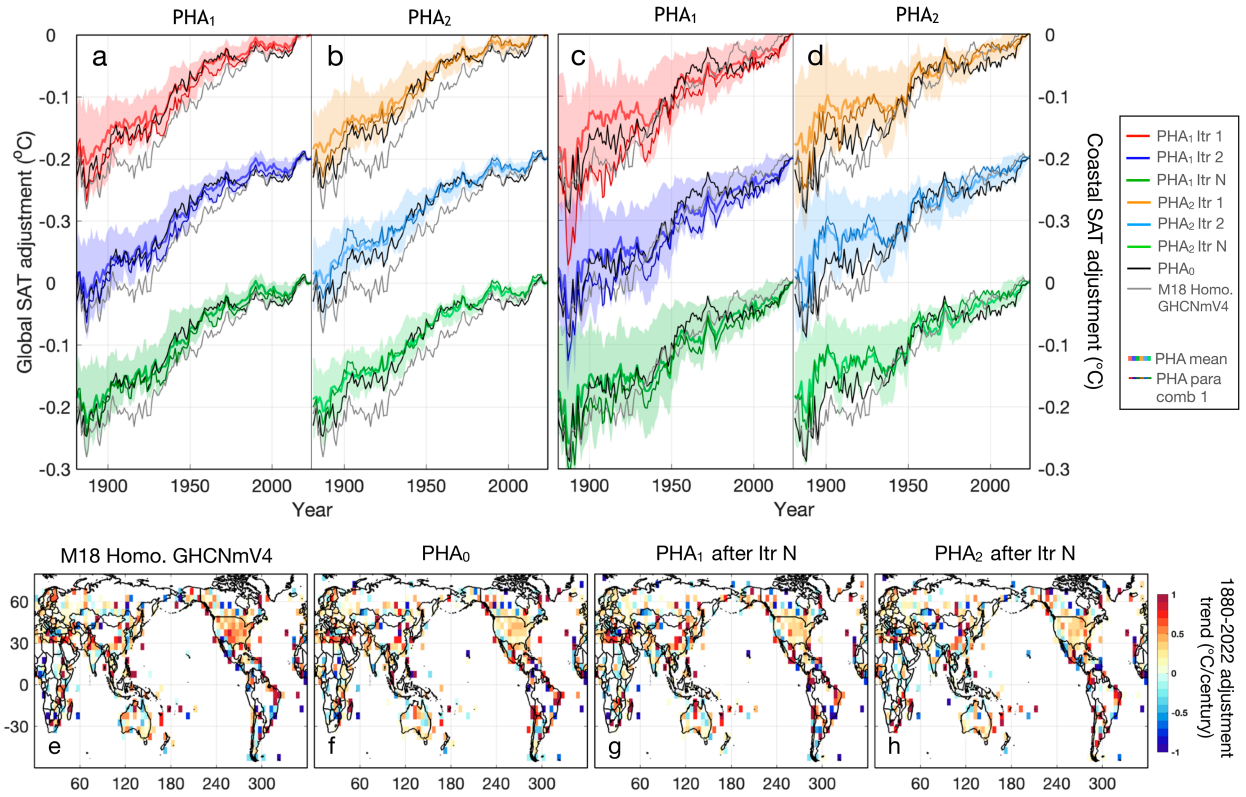

FIG. 8. Adjustments at global and regional scales. (a),(b) Continental-mean adjustments for (a) PHA_1 and (b) PHA_2 after one (upper set of lines and shading), two (middle set) and N iterations (bottom set). Results are for parameter combination 1 (thin), mean over 50 members (thick), and range (shading). Results of PHA_0 (black) and the homogenized GHVNmV4 by Menne et al. (2018, gray) are shown for comparison. (c),(d) As in (a) and (b), but for coastal mean adjustments. (e)–(h) Long-term trend of 1880–2022 adjustments for (e) GHCNmV4 homogenized by Menne et al. (2018), (f) PHA_0, (g) PHA_1 after N iterations, and (h) PHA_2 after N iterations. A trend is calculated if a grid box has at least 10 decades that each have at least 1 month of data. Maps are all computed using parameter combination 1.

54 328 adjustments between 1880 and 2023 after running one iteration of adjustment estimation (red in Fig. 7a). A second iteration of PHA_1 gives an additional of 9964 breaks (blue in Fig. 7a).

In synthetic analyses, we find that running two iterations would be sufficient because fewer than 10 adjustments are made in further iterations. The required number of iterations to maximally address the list of not-yet-adjusted breakpoints should, however, increase with breakpoint frequency and the degree to which breakpoints are concentrated in space and time. In the application to GHCNmV4, which may have a different breakpoint distribution from synthetic analyses, we run the iteration until fewer than 100 breakpoints are further adjusted and denote that number of iterations as N. For PHA_1, 4847 more breakpoints are adjusted in five additional iterations for a total of seven iterations (green in Fig. 7a).

Compared with PHA_0 and PHA_1, PHA_2 has the advantage of suppressing small breaks and giving fewer false alarms (Fig. 5). Applied to GHCNmV4, PHA_2 adjusts the fewest breakpoints, with 44 788, 52 307, and 55 778 after running one, two, and seven iterations of adjustment estimation (Fig. 7c).

Similar to Menne et al. (2018), PHA_0, PHA_1, and PHA_2 detect more negative than positive breakpoints, and the mean of detected breaks is negative in each case (Figs. 7a,c). It follows that continental mean temperature adjustments show positive linear trends for both PHA_1 and PHA_2 over 1880–2022 (Fig. 8a), in this case both equaling 0.17°C century^{-1}. Similar to Fig. 1, continental and coastal-mean series are calculated by initially binning station data into 5° × 5° monthly grids, a conventional approach used in datasets like HadSST4 (Kennedy et al. 2019) and CRUTEM5 (Osborn et al. 2021). Gridded data are then averaged globally, while weighting each grid by the cosine of its latitude. Although PHA_2 adjusts fewer breakpoints than PHA_1, they give highly consistent adjustments for continental mean temperatures. These trends are also qualitatively consistent with the 0.16°C century^{-1} trend found using PHA_0 but are slightly smaller than the 0.19°C century^{-1} trend reported for the GHCNmV4 product homogenized by Menne et al. (2018) (Fig. 8a). Our estimates of global adjustments using PHA_0 and the Menne et al. (2018) homogenization are highly consistent after the late 1980s, although they diverge and show an offset of up to 0.03°C going back in time. One possible explanation is that our homogenization does not rely on metadata, yet more

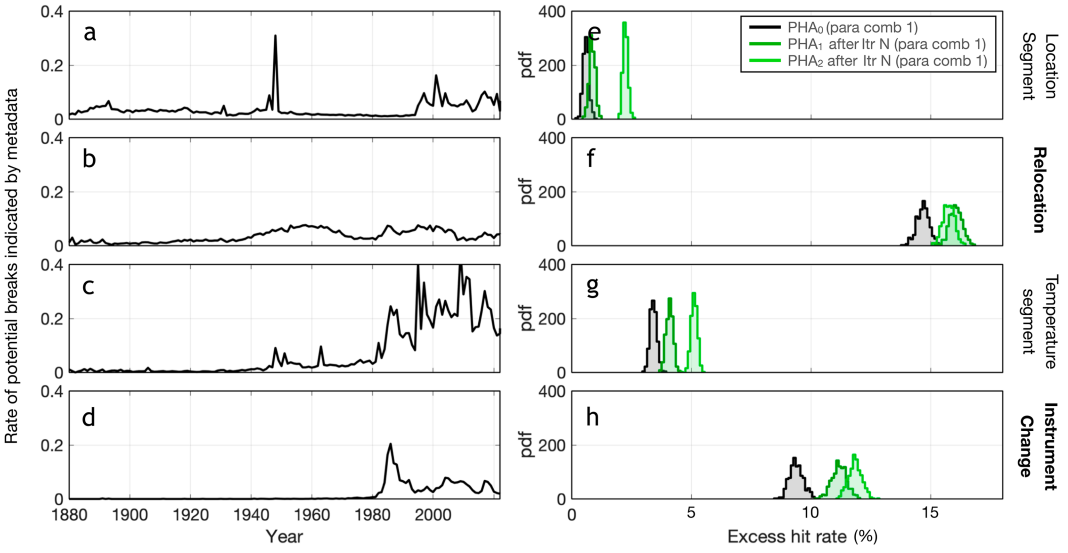

FIG. 9. Comparison with metadata. (a)–(d) Frequency of metadata-suggested potential breaks when using (a) segmented location information, (b) recorded relocation, (c) segmented temperature information, and (d) recorded instrument changes. (e)–(h) Excess hit rates for PHA_0 (black) and for PHA_1 (dark green) and PHA_2 (light green) after estimating adjustments with N iterations. For the default parameter combination, $N = 7$ in both cases. The excess rate is the hit rate relative to a null hypothesis where metadata indicated discontinuities are randomly placed in time (see text for more details).

detailed comparison appears a necessary undertaking in future work.

Despite overall consistency among the different homogenization estimates, distinct features are present at various spatial and temporal subsets. The spatial correlation between century-long adjustment trends (Figs. 8e–h) estimated by Menne et al. (2018) and PHA_0 equals only 0.66. The spatial correlation between PHA_0 and PHA_1 is somewhat larger at 0.76 and between PHA_1 and PHA_2 it equals 0.77. Whereas global temperature adjustments appear consistent, for example, with the RMSE between PHA_1 and PHA_2 annual adjustments being 0.02°C, the difference at the 5° × 5° regional level is also larger, having an average RMSE of 0.26°C. Additional future study at the regional level to identify regions and sources of discrepancies appears worthwhile.

b. *Comparison with station metadata*

The frequency of detecting breakpoints is consistent throughout the twentieth century. For PHA_1, the first iteration adjusts breaks at an average rate of once per 26 years within a given record (thin red curve in Fig. 7b). This rate increases to about once per 20 years after running additional iterations. For PHA_2, the frequency of adjustment is lower at once per 31 years for the first iteration and once per 25 years with additional iterations (Fig. 7d). Some level of breakpoints is expected. For example, the U.S. Historical Climate Network contains measurements from liquid in glass thermometers in Stevenson screens that were replaced by electronic resistance thermometers known as the Maximum-Minimum Temperature Sensor during the mid-to-late 1980s (Menne and Williams 2009; Williams et al. 2012).

To more specifically examine the rate and pattern of breakpoints that are algorithmically identified, we compare detected breakpoints with potential breaks suggested by available station history data compiled under the Historical Observing Metadata Repository (HOMR; https://www.ncei.noaa.gov/access/homr/). For each station, we record the timing when metadata suggest potential changes in temperature measurement technique or location. Four categories are examined: segmented location information, recorded relocation, segmented temperature information, and recorded instrument changes. Station metadata are, however, limited. Among 27 868 GHCNmV4 stations, only about 10 000 stations have metadata indicating at least one potential discontinuity throughout their entire station history, and more than 99% of these stations are from the United States or U.S.-affiliated islands.

The rate at which available metadata indicates potential discontinuities varies with time, and the temporal evolution differs across sources of information. For example, documented relocation rates increase in the late 1930s, drop in the 1970s, and again peak in the 1990s and 2000s (Fig. 9b). Documented instrument changes, however, are rare before they peak in the 1980s (Fig. 9d). We are unaware of whether changes in reported rates among the records with station data reflect changes in the actual rates of relocation and instrumentation change or, instead, the recording of such changes. For this reason, we only focus on the rate at which metadata-indicated discontinuities correspond with identified breakpoints. Specifically, if a metadata-indicated discontinuity lies within a 2-yr epoch of detected breaks, as defined in section 4, we count it as a hit.

We formulate a null hypothesis for hit rate by randomizing metadata adjustments. Specifically, the null is constructed by randomly shuffling the timing of metadata-indicated discontinuities within each station, while keeping both the number of metadata-indicated discontinuities and the location of PHA-detected breaks unchanged. For each randomization, we count the rate of shuffled breakpoints falling into an epoch associated with a PHA-detected break point. Repeating this process 500 times gives a null distribution against which the observed hit rate of PHA is assessed. A hit rate significantly higher than the null distribution, or what we call excess hit rate being greater than zero, indicates that PHA is skillful. Note that in a limit where breakpoints are defined at every time step PHA cannot be skillful because the null results would have a 100% hit rate.

The hit rate of metadata-indicated changes with PHA-identified breakpoints is significantly higher than our null distribution for each category of metadata ($P < 0.001$), indicating the skill of PHA-based methods. Averaging across stations and PHA approaches, the correspondence of metadata-indicated changes with breakpoints is 1%, 15%, 4%, and 11% higher than adjustments with random timing for segmented location information, recorded relocation, segmented temperature information, and recorded instrument changes, respectively (Figs. 9e–h). Moving stations and changing measurement approaches are, apparently, more likely to result in identifiable breakpoints. Both PHA_1 and PHA_2 give higher excess hit rates than PHA_0 for all metadata types, confirming the skill of our revisions to PHA_0. Whereas PHA_2 is better at catching instrumental changes, PHA_1 is slightly better for relocation.

Using a homogenization algorithm appears important for uniform treatment of data, especially given the unequal distribution of metadata across nations and that more than 90% of breakpoints identified by PHA_1 and PHA_2 are not associated with an event indicated by relocation or instrumental changes. In the United States, where most metadata are available, rates of relocation or instrumental change reported in the metadata range from 2% yr^{-1} between 1900 and 1950 to 8% yr^{-1} between 1980 and 2023 (Figs. 9b,d), whereas the ratio of PHA-identified breakpoints between these two intervals remains relatively stable at about 4%–5% yr^{-1} (Figs. 7b,d).

Although the rate of PHA-detected breakpoints is stable in time, stations with one breakpoint are more likely to be associated with another break. To demonstrate this point, we compare the number of breaks per station between GHCN and a null hypothesis assuming the occurrence of breakpoints is independent across time and stations. To construct this null hypothesis, we draw, for each station, a number of breakpoints from a binomial distribution $B(p_B, n_B)$, where the success rate or average percentage of years having breaks is p_B and n_B is the number of years with data. We repeat the process 500 times to obtain a distribution assuming independent breakpoint occurrence. Raw GHCN data homogenized using either PHA_0 or PHA_1 have significantly more stations without breaks, fewer stations with fewer than six breakpoints, and more stations with seven or more breakpoints (Figs. 7e,f). PHA_2 adjusts fewer breakpoints in general but still shows a similar structure in the deviation from binomial processes (Fig. 7g). Possible explanations include certain stations being subject to repeated moves or instrument updates or that some discontinuities detected by PHA_0 are associated with problematic segments that recover later in time, in which case breakpoints tend to appear in pairs.

c. *Uncertainty quantification*

We use an ensemble method to quantify parametric uncertainties in PHA_1 and PHA_2 associated with errors in the timing of identified breakpoints and the magnitude of required adjustments, similar to Williams et al. (2012). That is, in addition to the default parameter combination, we perturb all parameters in the algorithm (Table B1). Note that randomized parameter combinations tend to give higher error rates, often because of conservative breakpoint adjustments that relax the magnitude of trend adjustments toward zero (Williams et al. 2012).

To account for this potential bias, we first run a 300-member randomized parameter ensemble on synthetic data from the multivariate Gaussian process with $\alpha = 0.2$, the median across CMIP6 models, for both PHA_1 and PHA_2. The resulting trend RMSE after running adjustment estimation two times ranges from 0.25° to 2.71°C $century^{-1}$, while the default combination gives an average RMSE of 0.26°C $century^{-1}$. The high error for some combinations is associated with insufficiently adjusting breaks, which is typically associated with SNHT either identifying too many or too few breakpoints in the initial screening. Whereas using too few initial breakpoints naturally results in fewer adjustments, too many initial breakpoints gives insufficient numbers of homogeneous neighbors required for estimating adjustments. We subset the 50 combinations from 300-member PHA_1 and PHA_2 ensemble that give the lowest RMSE (Tables B2 and B3). The fact that both PHA_1 and PHA_2 generate qualitatively similar results in terms of both continental-mean adjustments (Fig. 8a) and excess hit rates (Fig. 9) allows for pooling the two ensembles together to generate a 100-member LSAT ensemble. The trend RMSE of the 100 combinations ranges from 0.25° to 0.30°C $century^{-1}$ when applied to the Multivariate Gaussian synthetic data. When applied to GHCN, adjusted temperatures in each member correspond to the last iteration—iteration N—in which fewer than 100 new adjustments are made.

Applying the trimmed parameter ensemble to GHCNmV4, we detect, respectively, 46 287 [37 374, 64 084] (median and range across 100 parameter combinations), 54 370 [45 050, 75 461], and 58 033 [49 082, 81 125] breakpoints after estimating adjustments for one, two, and N iterations (Figs. 7a,b). For the final adjustments after N iterations, the number of adjusted breaks are 65 916 [56 400, 81 125] for 50 PHA_1 members and 56 168 [49 082, 62 067] for PHA_2 members. The mean magnitude associated with adjusted breakpoints ranges from −0.07° to −0.04°C. Thus, accounting for timing and magnitude uncertainties provides support that the raw GHCNmV4 data underestimate long-term trends in temperature warming at the global scale. Global-average adjustments have 1880–2022 trends ranging from 0.12° to 0.20°C $century^{-1}$ (Figs. 8a,b), and

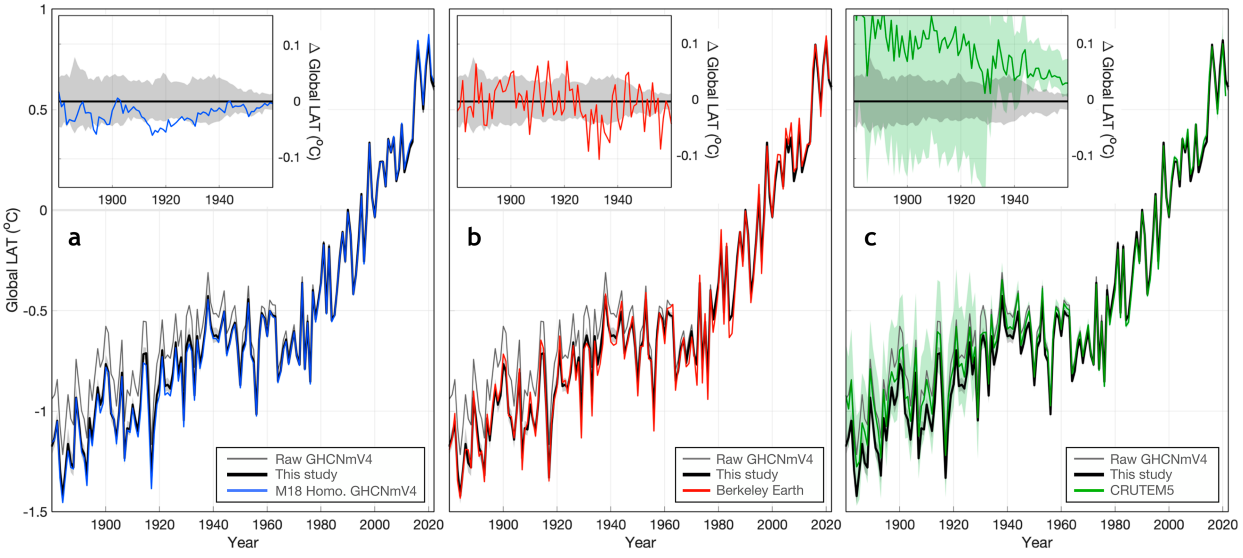

FIG. 10. Comparison of continental mean temperature anomalies with existing estimates. (a) Homogenized temperatures using our revised algorithm (black), homogenized GHCNmV4 by Menne et al. (2018) (blue), and raw GHCNmV4 (gray). Anomalies are relative to the mean over 1982–2014. Shading shows the 95% confidence interval over the 100-member ensemble. Coverage uncertainties are not accounted for. Shown in the inset panel in the upper-left corner is the difference from the central estimate of our adjusted temperatures. (b) As in (a), but for comparison with Berkeley Earth Temperature (red). Berkeley temperature is masked to have the same data coverage as GHCNmV4. (c) As in (a), but for CRUTEM5 (green). The green shading shows the 95% c.i. over a 200-member ensemble derived from subtracting HadSST4 (Kennedy et al. 2019) from non-infilled HadCRUT5 (Morice et al. 2021).

this range is consistent between the PHA_1 ([0.12, 0.20]) and PHA_2 ([0.12, 0.18]) subensembles.

Our ensemble for continental mean temperatures (Figs. 8a,b) is consistent with the previously published homogenized GHCNmV4 dataset (Menne et al. 2018) at the global scale, but the adjustments found in the previously published GHCNmV4 dataset for coastal stations are more negative than our ensemble, especially with respect to the PHA_2 subensemble over the early twentieth century (Figs. 8c,d). In a recent paper, we showed that discrepancies exist between SSTs and LSATs near coastlines during the early 1900s (Chan et al. 2023) and that LSATs could be used to correct SSTs. An implication is that using the previously published homogenized GHCNmV4 dataset leads to an SST trend that is approximately 0.05°C century^{-1} higher than indicated by our LSAT ensemble.

6. Discussion and conclusions

To further improve the detection and adjustment of discontinuities in historical temperature records from weather stations, we propose two revised pairwise homogenization algorithms that account for autocorrelation in time series. One algorithm, PHA_1, involves minor modifications of an existing algorithm, PHA_0 (Menne and Williams 2009), to detect breakpoints in the presence of autocorrelated temperature data. The other algorithm, PHA_2, makes a larger change to PHA involving replacing a standard normal homogeneity test (SNHT) for breakpoints with a penalized likelihood method.

Application to perturbed CMIP6 simulations and synthetic data with different levels of autocorrelation indicates that both PHA_1 and PHA_2 identify more breaks and produce fewer false alarms than PHA_0, implying higher skill in recovering long-term temperature trends. Moreover, PHA_2 surpasses PHA_1 in identifying true breaks and minimizing false alarms, thereby further improving the long-term-trend estimate slightly. That said, PHA_1 offers greater computational efficiency, requiring only about 0.1 s for a 600-time-step series, compared to about 10 s for PHA_2. When applied to the homogenized temperatures in GHCNmV4, both PHA_1 and PHA_2 show highly consistent global LSAT adjustments and comparable skill when evaluated against events recorded in metadata. Given PHA_1's minimal changes to the benchmark PHA_0 and PHA_2's slight accuracy advantage, we integrate both in our ensemble of improved land surface temperature estimates.

Applying PHA_1 and PHA_2 to GHCNmV4 station temperatures increases the 1800–2022 trend in continental mean temperature by 0.16 [0.12, 0.19]°C century^{-1} (mean and 95% c.i.) relative to unhomogenized trends. We estimate that continental mean temperature over 2010–22 was 1.65 [1.62, 1.69]°C (mean and 95% c.i.) warmer than the 1880–1900 average. Uncertainty is quantified using a 100-member ensemble that accounts for model uncertainty through using PHA_1 and PHA_2 approaches and for parametric uncertainties within each method. The code and detailed results of our algorithm are publicly accessible at https://doi.org/10.7910/DVN/AA0OM0.

We compare our continental mean temperatures with three existing estimates (Fig. 10) from NOAA homogenized CHCNmV4 using PHA_0 (Menne et al. 2018), CRUTEM5 (Osborn et al. 2021), and Berkeley Earth (Rohde et al. 2013a). To facilitate direct comparison, we average only over

grid boxes where all products have observations after regridding to the CRUTEM5 5° × 5° resolution. Although similar in most respects, NOAA homogenized GHCNmV4 using PHA_0 shown significantly greater warming between 1880–1900 and 2010–22 than our 100-member ensemble at 1.70°C (Fig. 10a). Such a result is consistent with our previous finding that the global-mean adjustment by Menne et al. (2018), 0.19°C century^{-1}, is on the high end of our PHA ensemble, [0.12, 0.19] (95% c.i.).

CRUTEM5 indicates less warming since the 1880s than our ensemble of only 1.50 [1.27, 1.72]°C (Fig. 10c). This reduced warming may reflect that CRUTEM5 used homogenization efforts by national or regional initiatives, as opposed to a global statistical algorithm (Osborn et al. 2021). Note that CRUTEM5 makes an ensemble characterization of uncertainties publicly available that, in addition to accounting for parametric uncertainty, also accounts for sampling and measurement errors within individual grid boxes and instrumental exposure biases from nonstandard screening (Osborn et al. 2021), leading to a larger 95% confidence interval, particularly prior to the 1930s.

The Berkeley Earth temperature estimate is consistent with our ensemble from 1880 to 1940 (Fig. 10b), indicating a warming of 1.66°C over 2010–22 relative to 1880–1900. Berkeley Earth detects breakpoints using a method similar to steps 1–4 of PHA_0, but rather than explicitly adjusting temperatures, records are split at breakpoints and treated as distinct when calculating temperature anomalies relative to a climatological period (Rohde et al. 2013a).

We suggest that our PHA_1 and PHA_2 ensemble gives the most credible estimate of warming since the 1880s. This credibility is supported by PHA_1 and PHA_2 outperforming PHA_0 in synthetic trials and consistency of our ensemble with the point estimate provided by the partially distinct methodology of Berkeley Earth. It will be useful to integrate PHA_1 and PHA_2 results with ongoing work to combine land and sea surface temperature datasets (e.g., Cowtan et al. 2018; Chan et al. 2023) as well as to infill for missing regions (e.g., Kadow et al. 2020; Meinshausen et al. 2022) in order to obtain global estimates of temperature variability.

Acknowledgments. The authors acknowledge two anonymous reviewers for comments that improved the quality of the paper. G. Gebbie is supported by NSF OCE-82280500. P. Huybers is supported by NSF Grant 2123295. The authors have no conflict of interests to declare.

Data availability statement. All datasets used in this study are available as follows: GHCNmV4 (https://www.ncei.noaa.gov/pub/data/ghcn/v4/; last access, 16 October 2023); HOMR (https://www.ncei.noaa.gov/access/homr/; last access, 27 October 2023); Berkeley Earth Monthly temperature (https://berkeley-earth-temperature.s3.us-west-1.amazonaws.com/Global/Gridded/Complete_TAVG_LatLong1.nc; last access, 11 July 2022); CRUTEM5.0.1.0 (https://www.metoffice.gov.uk/hadobs/crutem5/data/CRUTEM.5.0.1.0/download.html; last access, 11 June 2022); HadSST4.0.1.0 200-member ensemble (https://www.metoffice.gov.uk/hadobs/hadsst4/data/download.html; last access, 20 May 2022); HadCRUT5.0.1.0 200-member ensemble (https://www.metoffice.gov.uk/hadobs/hadcrut5/data/HadCRUT.5.0.2.0/download.html; last access, 11 June 2022). Monthly CMIP6 outputs are from the ESGF portal (https://aims2.llnl.gov/search/cmip6/; last access, 16 August 2021). PHA_1 and PHA_2 code, together with our 200-member ensemble of monthly station LSAT temperature, are at https://doi.org/10.7910/DVN/AA0OM0.

APPENDIX A

Developing Synthetic Data

We develop synthetic data using both CMIP6 simulations and draws from a multivariate Gaussian process. For CMIP6, we interpolate simulated temperatures using a bilinear method to locations of weather stations and add a random number of breakpoints with random timing and random magnitude. The number of breakpoints for a given time series n_b is specified by drawing a random number from a normal with a mean of 3 and standard deviation of one, truncating values to range between 0 and 6, and then rounding. We impose break at an average rate of 3 (50 years)$^{-1}$ and assign magnitudes to each breakpoint that are randomly drawn from $N(-0.05, 1)$. The rate and distribution of breakpoints are comparable to those found by PHA_0 as well as those reported in Menne et al. (2018). Note that the nonzero centered distribution introduces a bias in long-term trends as is generally inferred (Fig. 8).

Synthetic temperatures that are correlated in space and time are generated using an AR-1 multivariate Gaussian process:

$$\mathbf{T}_t = \alpha \mathbf{T}_{t-1} + \boldsymbol{\epsilon}_t. \tag{A1}$$

Vector $\mathbf{T}_t$ represents temperatures at time t in a network of weather stations, for which we choose continental U.S. stations in GHCNmV4. We run Eq. (A1) for 700 time steps and discard the first 100 warm-up steps. Varying the system memory α permits controlling the autocorrelation of generated time series and their differences.

The noise process $\boldsymbol{\epsilon}_t$ follows a multivariate Gaussian distribution:

$$\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{A2}$$

where $\boldsymbol{\Sigma}$ is a covariance matrix generated according to $\Sigma_{ij} = (1 - \alpha)^2 \exp(-|\Delta d|/\tau)$. The variable $|\Delta d|$ is the arc length, in degrees, between stations i and j, and τ is the decorrelation distance, for which we choose 5°, approximately half of the Rossby deformation radius for the midlatitude atmosphere. The variance of the noise innovation is a decreasing function of α such that the expected variance of $\mathbf{T}$ is constant for α between 0 and 1.

The same seeding of random numbers is used for all synthetic experiments, such that identical breaks are introduced to both CMIP6 models and synthetic data generated from a multivariate Gaussian process. Differences between the CMIP6 and multivariate Gaussian process results include that spatial correlation decays more slowly at small distances across the CMIP6 temperatures.

APPENDIX B

Revised Pairwise Station Homogenization Algorithm (PHA_1)

A step-by-step description of PHA_1 is provided for purposes of repeatability. PHA_1 generally follows that of Menne and Williams (2009, hereafter MW09) and Williams et al. (2012, hereafter WMT12). We specifically note below where our approach differs from MW09 and WMT12.

a. Identify neighbors

Neighboring stations are first identified. For each target station, we first identify the nearest "NEIGH CLOSE" (**100**/120/150/200) stations. Numbers in the parentheses denote possible values of the algorithm parameter inside quotation marks, and the number in boldface is our default value, which is also listed as ensemble member 1 in Tables B2 and B3. The distance, "NEIGH DIS", is evaluated using one of the following metrics: **difference correlation (1 diff)**, Pearson's correlation (corr), or physical distance on the sphere (near). Difference correlation is the correlation between month-to-month temperature changes, a metric that helps diminish the effects of abrupt breaks in determining the correlation (Peterson et al. 1998).

Following MW09, seasonal cycles are removed by subtracting the mean temperature over the entire period for each station before evaluating correlations. To guard against small sample sizes giving spurious correlations, stations having fewer than "NUM4COV" (**60**/120/180) in common with the target station are excluded. When evaluating correlations (1 diff and corr), we also exclude stations whose correlations are smaller than "CORR LIM" (**0.1**/0.3/0.5) with the target station. When using spherical distance (near), we still remove seasonal cycles but do not use the "CORR LIM" parameter.

Among the eligible neighboring stations meeting the distance and correlation requirements, the top "NEIGH FINAL" (**40**/60/80) are first selected. Our algorithm then loops over the remaining stations in descending order. If adding a remaining station increases the number of neighbors for any month that has fewer than "MIN STNS" (5/**7**/9) neighbors, the least correlated or the farthest station is replaced. Difference monthly temperature anomalies between the target station and each selected neighbor are calculated.

b. Identify breakpoints from pairwise difference series

For each difference series, we apply an iterative standard normal homogeneity test (SNHT). The test is performed iteratively between a splitting phase, where the algorithm tests whether each segment of time series contains any further breakpoints, and a merging phase, where the algorithm combines consecutive segments and excludes the identified breakpoint in middle if no breakpoints are identified by SNHT in the combined time series. This process repeats until no more breakpoints can be identified or the number of iterations reaches ten. Unlike MW09 and WMT12, which used the 95% confidence level estimated from white noise series, the revised algorithm uses "SNHT THRES" (80%/**90%**/95%) estimated from autocorrelated random series.

To estimate updated SNHT thresholds, we first generate n-sample red noise series using an order-one autoregressive process, $x_t = \alpha x_{t-1} + \epsilon_t$, where α is the memory of the system, for which we loop over 0 to 0.4 at an increment of 0.01, and n is the length of time series ranging between 5 and 3500. For each combination of α and n, 50 000 random series are generated and then normalized to zero mean and unit variance.

For each synthetic series, we calculate lag-1 autocorrelation α and the SNHT test statistics, $T_0 = \max_{1\le v<n}[v\overline{z}_1^2 + (n-v)\overline{z}_2^2]$ (Alexandersson 1986). Here $\overline{z}_1$ and $\overline{z}_2$ are, respectively, the mean over the two periods before and after time step v, and the calculation loops v over 1 to $n-1$ to find the maximum value. For each n value, we calculate the revised "SNHT threshold" as the 80%, 90%, and 95% quantiles of T_0 within 0.1 incremental bins of α.

When performing SNHT using revised thresholds, we first evaluate α for each difference series using a sliding window of 100 months, or one-third of the time series if shorter than 100 months. We take the median value of the α values sampled across the time series. α is updated in every splitting phase of SNHT, and windows overlapping with any detected breakpoints are discarded in the calculation of median values. This method reduces bias in autocorrelation estimates due to artificial discontinuities. The length of the

TABLE B1. Parameters in the revised pairwise homogenization algorithms.

Parameter	Meaning
ADJ EST	Methods to determine adjustments from multiple pairwise estimates
ADJ COMB	Minimum length of data period that can be adjusted
ADJ MINLEN	Minimum number of months on two sides of breakpoints to estimate adjustments
ADJ MINPAIR	Minimum number of nonproblematic neighbors to estimate adjustments
AMPLOC PCT	Confidence window used to conflate breakpoints
CORR LIM	Minimum correlation to be identified as a neighbor
MIN STNS	Minimum number neighbors with coincident data
NEIGH CLOSE	Maximum number of neighboring series to consider
NEIGH DIS	Similarity matrix used for ranking neighbors
NEIGH FINAL	Final (maximum) number of neighbors per station
NUM4COV	Minimum number of overlapping months for evaluating correlation
SNHT THRES	Confidence level of the standard normal homogeneous test (SNHT)

TABLE B2. Parameters for individual members of the PHA_1 ensemble. Member 1 is the default parameter combination, whose values are from Williams et al. (2012). Members 2–50 are from randomly perturbing PHA parameters and then trimmed according to trend RMSE in synthetic analyses.

	Ensemble number												
	1	2	3	4	5	6	7	8	9	10	11	12	13
ADJ EST	med	mean	mean	mean	Qavg	mean	mean	mean	Qavg	med	Qavg	mean	med
ADJ COMB	24	24	24	24	24	24	24	24	24	24	24	24	24
ADJ MINLEN	18	24	18	24	24	24	18	18	18	24	18	24	18
ADJ MINPAIR	2	4	2	5	5	5	3	3	4	3	3	2	2
AMPLOC PCT	92.5	92.5	95	95	95	90	90	92.5	95	92.5	95	92.5	90
CORR LIM	0.1	–	–	0.5	–	–	–	–	–	–	0.3	–	–
MIN STNS	7	9	5	5	7	5	7	9	9	5	9	7	9
NEIGH CLOSE	100	100	150	100	150	100	200	120	150	100	100	150	150
NEIGH DIS	1 diff	near	near	1 diff	near	near	near	near	near	near	1 diff	near	near
NEIGH FINAL	40	40	40	40	40	40	40	40	40	40	40	40	40
NUM4COV	60	120	60	60	120	120	60	180	60	180	180	120	120
SNHT THRES	90	95	95	95	95	80	95	90	95	95	95	80	95

	14	15	16	17	18	19	20	21	22	23	24	25	26
ADJ EST	med	Qavg	med	med	med	Qavg	med	Qavg	med	med	Qavg	mean	mean
ADJ COMB	24	24	24	24	24	24	24	18	24	24	18	18	18
ADJ MINLEN	24	24	18	18	24	24	18	18	18	18	18	18	18
ADJ MINPAIR	4	4	4	5	2	4	3	4	3	4	3	5	4
AMPLOC PCT	95	90	90	95	90	92.5	92.5	90	90	90	90	95	92.5
CORR LIM	0.1	–	0.5	–	–	–	0.1	–	–	0.1	–	–	0.3
MIN STNS	9	9	7	9	9	9	7	9	7	5	7	5	9
NEIGH CLOSE	200	150	100	120	200	100	200	120	100	200	200	100	100
NEIGH DIS	1 diff	near	1 diff	near	near	near	1 diff	near	near	1 diff	near	near	1 diff
NEIGH FINAL	40	40	40	40	40	40	40	40	40	40	40	40	40
NUM4COV	120	120	120	180	60	180	60	60	120	60	120	120	60
SNHT THRES	80	90	90	90	90	80	80	95	80	90	90	80	95

	27	28	29	30	31	32	33	34	35	36	37	38	39
ADJ EST	mean	Qavg	mean	med	Qavg	med	Qavg	med	med	med	med	Qavg	Qavg
ADJ COMB	24	18	18	18	24	24	18	24	18	24	24	24	24
ADJ MINLEN	24	18	18	18	24	24	18	18	18	24	24	24	24
ADJ MINPAIR	5	3	2	4	2	4	5	5	2	2	5	2	2
AMPLOC PCT	90	92.5	95	95	95	90	95	92.5	92.5	95	95	95	95
CORR LIM	–	–	0.3	–	0.3	–	–	–	–	–	–	–	0.3
MIN STNS	7	7	9	9	7	7	9	9	5	7	5	9	5
NEIGH CLOSE	100	200	100	200	100	120	120	150	100	120	100	200	150
NEIGH DIS	near	near	1 diff	near	1 diff	near	near	near	near	near	near	near	1 diff
NEIGH FINAL	60	40	40	40	60	60	40	60	40	60	60	60	60
NUM4COV	60	180	60	60	180	60	120	60	60	60	60	60	60
SNHT THRES	90	95	80	80	90	90	90	90	90	80	80	80	80

	40	41	42	43	44	45	46	47	48	49	50
ADJ EST	med	med	Qavg	mean	med	med	Qavg	Qavg	Qavg	mean	Qavg
ADJ COMB	24	24	24	24	24	24	24	24	24	24	24
ADJ MINLEN	18	18	24	18	18	24	24	24	24	18	18
ADJ MINPAIR	2	2	3	4	2	2	5	2	2	2	4
AMPLOC PCT	90	95	95	95	90	95	90	95	92.5	90	95
CORR LIM	–	–	–	0.1	0.3	–	0.1	0.5	–	0.5	0.3
MIN STNS	9	7	5	5	7	5	5	5	5	5	5
NEIGH CLOSE	120	150	100	200	200	120	150	150	100	100	100
NEIGH DIS	near	near	near	corr	1 diff	near	corr	corr	near	1 diff	corr
NEIGH FINAL	60	60	60	40	60	60	40	40	60	60	40
NUM4COV	60	60	60	180	180	180	180	120	180	60	180
SNHT THRES	95	90	80	90	80	80	95	95	80	90	90

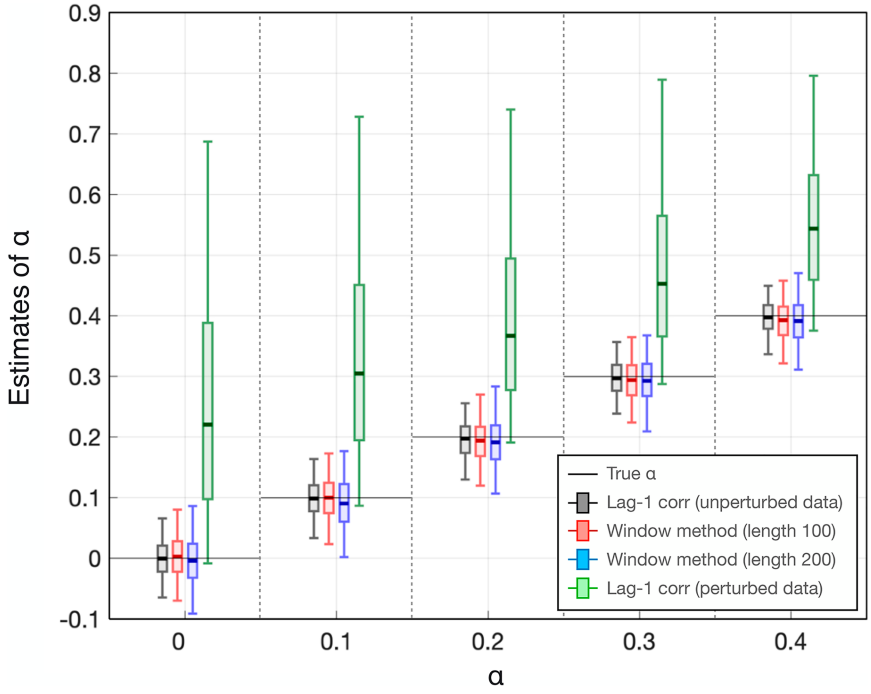

FIG. B1. Skill of the moving-window autocorrelation estimator. Skill is evaluated for α from 0 to 0.4 with an increment of 0.1. For each α, we randomly generate 1000 synthetic series that each has 1000 time steps. Breakpoints with random timing and random amplitudes are then introduced following the approach in appendix A. Whereas lag-1 autocorrelation of the unperturbed series is calculated to indicate an unbiased estimate (black box plot), three estimates of α are obtained from the perturbed series, which are 1) the moving-window method with a window length of 100 time steps (red box), 2) the moving-window method with a window length of 200 time steps, and 3) lag-1 autocorrelation over the entire perturbed series (green box). For the box plot, the thick bar shows the median, the box indicates the interquartile range, and the whiskers show the 95% c.i.

sliding window could influence the estimation of α (Gallagher et al. 2022), but synthetic tests for 1000-step times series and alpha over 0–0.4 indicate that estimates of α are only slightly biased by -0.01 to -0.02 for both window lengths equaling to 100 and 200 time steps (Fig. B1). The slight negative bias arises because we use a 90% threshold in the SNHT test, such that large true variability would be misidentified as breakpoints and excluded from autocorrelation estimates. That said, this bias is small and does not seem to depend strongly on the window length. The fact that consistent results are found in synthetic trials using PHA_2 also indicates that this approach is adequate. SNHT thresholds not explicitly precomputed for a given n and α are estimated using bilinear interpolation.

After identifying potential breaks using SNHT, a check is made as to whether each identified breakpoint reflects breaks or long-term trends using a Bayesian information criterion approach (BIC; Schwarz 1978). Specifically, for a potential breakpoint, i, whose timing is t_i, we take the two segments on which it neighbors and calculate the BIC for seven different models. In addition to the five candidate models tested in MW09, PHA_1 also tests two other models:

$$y_t = \begin{cases} \mu_1 + k_1 t + \epsilon_t & t_{i-1} < t \le t_i \\ \mu_2 + \epsilon_t & t_i < t \le t_{i+1} \end{cases}, \tag{B1}$$

$$y_t = \begin{cases} \mu_1 + \epsilon_t & t_{i-1} < t \le t_i \\ \mu_2 + k_2 t + \epsilon_t & t_i < t \le t_{i+1} \end{cases}. \tag{B2}$$

We fit models using the Theil–Sen estimator (Theil 1950), which uses the median value of slopes between every possible pair of data to obtain a robust fitting that is less affected by outliers. After fitting each model, we calculate BIC following

$$\text{BIC}(p) = -n' \log\left(\frac{\text{SSE}}{n'}\right) + \log(n')p, \tag{B3}$$

where p is the number of parameters in a model, n' is the number of time steps from $t_{i-1} + 1$ to t_{i+1}, and SSE is the sum of squared error for a particular model fit. A breakpoint is confirmed if any model other than a straight line has the lowest BIC. Otherwise, we exclude this point from further analysis. For each confirmed breakpoint, we record estimates of its normalized magnitude, $\hat{m} = (\mu_2 - \mu_1)/\sqrt{\text{SSE}/(n' - 1)}$.

Note that for PHA_2, (Table B3) this step of breakpoint detection is replaced by a genetic algorithm-based penalized likelihood approach (GAPL; details in appendix C).

TABLE B3. As in Table B2, but for parameter combinations for the PHA_2 ensemble.

	Ensemble number												
	1	2	3	4	5	6	7	8	9	10	11	12	13
ADJ EST	med	mean	Qavg	mean	mean	mean	mean	mean	med	mean	med	med	Qavg
ADJ COMB	24	24	24	24	24	24	24	24	24	24	24	24	24
ADJ MINLEN	18	24	24	18	24	18	24	18	24	24	18	24	24
ADJ MINPAIR	2	4	2	3	3	4	2	5	5	3	2	5	4
AMPLOC PCT	92.5	92.5	90	90	90	95	90	92.5	92.5	95	95	92.5	90
CORR LIM	0.1	–	–	–	0.1	0.1	0.3	–	–	–	0.5	–	0.5
MIN STNS	7	7	9	9	7	5	5	9	5	9	5	7	9
NEIGH CLOSE	100	200	100	150	150	200	120	120	100	200	100	100	150
NEIGH DIS	1 diff	near	near	near	1 diff	1 diff	1 diff	near	near	near	1 diff	near	1 diff
NEIGH FINAL	40	40	40	40	40	40	40	40	40	40	40	40	40
NUM4COV	60	120	180	180	60	60	60	120	60	120	60	60	60
	14	15	16	17	18	19	20	21	22	23	24	25	26
ADJ EST	med	Qavg	Qavg	med	med	Qavg	med	med	med	Qavg	med	mean	med
ADJ COMB	24	24	24	24	24	24	24	24	24	24	24	24	24
ADJ MINLEN	24	24	18	24	24	18	24	18	24	24	24	18	24
ADJ MINPAIR	3	5	5	3	3	2	5	3	5	2	5	3	3
AMPLOC PCT	92.5	95	95	92.5	92.5	90	90	95	92.5	90	95	90	92.5
CORR LIM	0.3	0.3	–	0.5	–	0.1	0.1	–	–	–	–	–	–
MIN STNS	9	9	5	5	5	9	5	7	5	5	9	9	5
NEIGH CLOSE	150	120	150	200	100	200	120	120	200	100	100	120	120
NEIGH DIS	1 diff	1 diff	near	1 diff	near	1 diff	1 diff	near	near	near	near	near	near
NEIGH FINAL	40	40	40	40	40	40	40	40	60	60	40	60	60
NUM4COV	180	180	60	180	60	120	120	180	180	180	120	180	60
	27	28	29	30	31	32	33	34	35	36	37	38	39
ADJ EST	Qavg	mean	med	med	mean	med	mean	mean	med	med	mean	mean	Qavg
ADJ COMB	24	24	24	24	24	24	24	24	24	24	18	24	24
ADJ MINLEN	24	18	18	24	24	18	24	18	18	24	18	24	24
ADJ MINPAIR	2	3	4	2	5	2	3	4	3	5	5	5	3
AMPLOC PCT	90	95	90	95	92.5	90	92.5	95	95	92.5	95	95	95
CORR LIM	–	–	–	–	0.3	–	0.3	–	–	–	–	0.3	–
MIN STNS	9	9	7	9	5	5	5	9	9	5	9	9	9
NEIGH CLOSE	200	200	150	150	150	100	150	150	200	100	200	200	150
NEIGH DIS	near	near	near	near	1 diff	near	1 diff	near	near	near	near	1 diff	near
NEIGH FINAL	60	60	60	60	60	60	60	60	60	60	40	60	60
NUM4COV	180	60	120	60	60	60	120	120	180	120	60	120	60
	40	41	42	43	44	45	46	47	48	49	50		
ADJ EST	med	med	med	Qavg	med	mean	mean	mean	Qavg	med	median		
ADJ COMB	24	24	18	24	24	18	18	18	18	18	18		
ADJ MINLEN	24	24	18	18	24	18	18	18	18	18	18		
ADJ MINPAIR	5	4	3	4	5	4	4	3	4	2	3		
AMPLOC PCT	92.5	95	92.5	95	90	92.5	92.5	92.5	95	95	90		
CORR LIM	0.3	0.1	0.5	0.5	0.1	–	–	0.3	–	–	–		
MIN STNS	9	9	7	5	9	7	5	7	5	7	5		
NEIGH CLOSE	200	120	150	120	100	100	120	150	100	150	100		
NEIGH DIS	1 diff	1 diff	1 diff	1 diff	1 diff	near	near	1 diff	near	near	near		
NEIGH FINAL	60	60	40	60	60	40	40	40	40	40	40		
NUM4COV	180	120	180	180	60	180	120	120	120	120	120		

c. Attribute breakpoints to stations

Breakpoints confirmed in a difference series can be due to breaks in either station involved. As a result, we follow PHA_0 to attribute breaks to individual stations using a count-down method. For each station at each time step, we count the number of neighbors with which a target station shows a break. When two breaks involve stations that are mutually targets and neighbors, we exclude one of the target–neighbor pairs to avoid double counting. After forming a list of breakpoint counts, the station and time step with the highest count is associated with a breakpoint and that count is reset to zero. Counts of neighboring stations

that were originally associated with this breakpoint are decreased by one. The procedure is repeated until no count is greater than one, reflecting the fact that we require two neighboring stations at a time step to confirm a target as the source of a break.

The presence of missing data makes the process above more complicated. Here we follow an approach implemented in the FORTRAN code associated with WMT12 but that appears to not yet have been documented in the literature. If a breakpoint in, for example, station 1 occurs at a time step corresponding to missing data in station 2, the pairwise algorithm will identify a breakpoint at the timing of the nearest previous time step with data in station 2. For purposes of tracking, we assign universal IDs (UIDs) to individual breakpoints. If there are missing data following a detected breakpoint in a record or a neighbor from which a difference series is computed, we mark all subsequent time steps with missing data under the same UID. After attributing a breakpoint, all time steps sharing the same UID are decreased by one.

d. Combine near-in-time breakpoints

The timing of breakpoints can be uncertain, and multiple breakpoints can be found in succession when only a single breakpoint exists. To address this issue near-in-time breakpoints are combined to account for timing errors. We follow Menne and Williams (2009), who estimated the timing error by realizing 100-sample random time series with breakpoints of different magnitudes added at the 50th time step. They performed SNHT to each of the synthetic series and calculated the error of the timing of identified breakpoints, which decreases with the magnitude of breaks. Although timing error may also depend on autocorrelation, we keep this estimation for simplicity.

For each station, each attributed breakpoint is assigned with an epoch, whose length is the 90%/**92%**/95% interval of timing error, "AMPLOC PCT". The timing with the most neighbors is first marked as occupied. The epoch of the breakpoint with the second highest number of neighbors is then checked for overlap with occupied timings. If overlapping, the breakpoint is combined with the nearest occupied timing. Otherwise, the timing of this breakpoint is set as occupied, and the process continues until all breakpoints are checked.

Following WMT12, we also combine breakpoints when they are within "ADJ COMB" (18/**24**) months. Specifically, the latter breakpoint is removed along with data in between the two breakpoints.

e. Estimate adjustment magnitudes

Steps 1–4 identify breakpoints in a network of temperature series, and it remains to estimate the adjustment associated with each breakpoint. We estimate the required adjustments for each breakpoint independently. Taking breakpoint i for station S as an example, we first subset the time interval $t_{i-1} + 1$ to t_{i+1}. If a neighbor of station S does not contain any breaks during this interval, we use the corresponding difference series from $t_{i-1} + 1$ to t_{i+1} to estimate the magnitude using the changepoint model in step 3 that has the lowest BIC for this breakpoint. If a neighbor contains breakpoints, but none are within "ADJ MINLEN" (**18**/24/36/48) months before and after the target break, we estimate an adjustment using the difference series from the neighbor's last break before the target and the first break after. Otherwise, no adjustments are estimated. For each breakpoint, looping over all neighbors results in its collection of estimated adjustments.

We then trim the collection of adjustment estimates involving a record and its paired neighbors using a Tukey method (Tukey 1977). The Tukey method is based upon finding the median (Q_2) and the first (Q_1) and third quartiles (Q_3) within a collection and trimming estimates that are smaller than $Q_1 - k(Q_2 - Q_1)$ or larger than $Q_3 + k(Q_3 - Q_2)$, where $k = 1.64$, a value used by WMT12. If more than "ADJ MINPAIR" (**2**/3/4/5) estimates remain, another Tukey method is applied to these remaining estimates. If $Q_1 - k(Q_2 - Q_1)$ and $Q_3 + k(Q_3 - Q_2)$ are of the same sign, we use "ADJ EST" (median/mean/average of the 25% and 75% quartiles) as the adjustment. Otherwise, this breakpoint is discarded for now.

Step 5 is run twice to ensure that all breakpoints are either discarded for now or that an estimated adjustment is specified for each.

f. Adjust and iterate

Step 5 gives a list of estimated adjustments and a list of breakpoints not yet adjusted (discarded for now in the last step). Following PHA_0, our revised PHA algorithms also adjusts estimated breakpoints relative to values in the last segment. After all adjustments estimated in step 5 are made, the adjusted temperatures and the list of breakpoints not yet adjusted are sent back to step 5, and step 5 is then rerun to check whether breakpoints in this remaining list now become adjustable and estimate the magnitude of required adjustments accordingly. In theory, this process can iterate until no more adjustments can be made. In the synthetic analyses, it usually takes two to three iterations to reach that ending point. In the application to GHCNmV4 dataset, this process is iterated between steps 5 and 6 until fewer than 100 breakpoints are adjusted in step 6. Six to eight iterations are usually required before meeting this criterion.

APPENDIX C

Multiparent Genetic Algorithm for Penalized Likelihood

The penalized likelihood method aims at finding the minimum penalized likelihood for a model:

$$\Delta T_t = \Delta C_t + \Delta D_t[+\gamma t], \tag{C1}$$

where ΔT_t denotes interstation difference at time step t. Terms ΔC_t and ΔD_t denote differences associated climatic

variability and breakpoints, respectively. The term γt denotes an optional linear trend.

Assuming ΔC_t follows an order-1 autoregressive process (i.e., $\Delta C_t = \alpha \Delta C_{t-1} + \epsilon_t$), Eq. (C1) is prewhitened to be

$$Y_t = \Delta T_t - \alpha \Delta T_{t-1} = \Delta D_t - \alpha \Delta D_{t-1}[+(\alpha + t - \alpha t)\gamma] + \epsilon_t, \tag{C2}$$

for a case without missing data, and

$$Y_t = \frac{\Delta T_t - \alpha^{k_t} \Delta T_{t-k_t}}{S_{k_t}} = \frac{\Delta D_t - \alpha^{k_t} \Delta D_{t-k_t}[+(\alpha^{k_t} k_t + t - \alpha^{k_t} t)\gamma] + \sum_{i=1}^{k_t} \alpha^{i-1} \epsilon_{t-i+1}}{S_{k_t}}, \tag{C3}$$

when missing data exist. In Eq. (C3), k_t is the time difference between the current time step and the nearest previous one with data. Terms in brackets denote components associated with fitting an optional linear trend.

Let the breakpoint component $\mathbf{\Delta D}$ contain s segments divided by $s - 1$ breakpoints, such that our model contains a total of $2s + 1$ parameters to be determined, comprising $s - 1$ timings of breakpoints, $s - 1$ magnitudes of breakpoints, the mean over the entire record, the autocorrelation (α), and variance of climatic variability (ϵ_t). When a trend is also fit, the number of parameters increases by one. The penalized loss function using Bayesian information criteria is

$$L = n \ln(2\pi) + \sum_{t=1}^{n} \ln(e_t) + \sum_{t=1}^{n} \frac{S_{k_t}^2 (Y_t - \hat{Y}_t)^2}{e_t} + (2s + 1)\ln(n), \tag{C4}$$

which is identical to Eq. (7) in the main text.

To conduct the optimization efficiently, we develop a multiparent genetic algorithm. This algorithm aims at finding the timing of breakpoints and autocorrelation. Conditional on breakpoint timing and autocorrelation, an ordinary least squares approach is used to find the maximum likelihood estimate of other parameters following Eq. (C3). Our genetic algorithm develops upon methods proposed by Killick et al. (2012, hereafter K12) and Li and Lund (2012, hereafter LL12) to allow for fast convergence for time series longer than 1000 time steps. Below we describe the algorithm in four steps.

In step 1, a population is initialized. For each member of the population, we initialize a Boolean vector $\boldsymbol{\eta}$ to represent the timing of breakpoints. This vector has the length of nonmissing data points, and each element of $\boldsymbol{\eta}$ is assigned probability $P = 0.01$ of indicating a breakpoint at the corresponding timing, i.e., η_t = true. An initial guess of autocorrelation is drawn from a uniform distribution $U(-0.99, 0.99)$. Finally, we evaluate loss for individual initial members.

In step 2, descendants are generated. Similar to K12 and LL12, the probability of choosing a member to be a parent is inversely proportional to the rank of loss in an ascending order. Unlike K12 or LL12 that find the parent using the total loss for the entire time series, our multiparent approach breaks long series into 300 time step (25-yr) blocks and finds the parents of each block independently according to the local loss. Local loss is also calculated using Eq. (C4) except n is the length of data within a block. It follows, for example, that a 100-yr descendant series can have at most eight different parents. Such a modification ensures that each descendant has the tendency of inheriting the best fitted segments, hence speeding up convergence.

In step 3, after all parents are determined, timing associated with breakpoints per segment are pooled together and each breakpoint is assigned a probability of $P = 0.5$ of being dropped. Conversely, each time step not indicating a break is assigned a probability of 0.4% of becoming a break, i.e., turning η_t = false into η_t = true. The 0.4% probability corresponds to the occurrence of breakpoints approximately every 20 years—the frequency in GHCN. Breakpoints within four time steps are combined, and an autocorrelation is estimated for each descendant using the window method, as described in section 2a.

LL12 further perturbed the timing of retained breakpoints by 0, +1, or −1, which is useful for fine tuning the optimal timing. Similarly, we perturb the timing using a number drawn from a normal distribution and rounded to the nearest integer. Inspired by simulated annealing (Bertsimas and Tsitsiklis 1993), we specify that the standard error of the normal distribution decrease exponentially, $\sigma = 3 \exp(-N_g/20)$, where N_g is the number of generations in the genetic algorithm. Compared with LL12, our perturbation is larger at the beginning, allowing for greater exploration of the parameter space and generally, in our simulations, speeding up convergence. Simulated annealing also guarantees that in later iterations the location of identified breakpoints are only perturbed slightly for purposes of fine tuning.

In step 4, the loss of each generated descendant is evaluated. Unlike existing approaches, we also evaluate whether removing all or each of the breaks whose magnitude is smaller than half of the residual error further reduces global loss. Whenever removing small breaks improves fitting, we update the timings of breakpoints accordingly, thereby allowing for efficient suppression of small breaks, which speeds up convergence. Whereas dropping small breaks may result in the algorithm converging to a local minimum, small breaks rarely improve the fit given the existence of a penalty term.

The algorithm iterates between steps 2–4 until the best member remains the same for 10 rounds. To further speed up the convergence, we use an island approach similar to K12. Our setup has three subpopulations that each has 150 members. The generation of descendants is within subgroups and after every five rounds, the best 20 members in subgroup j is migrated to replace the worst 20 members in subgroup $j + 1$ before generating descendants.

REFERENCES

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *J. Climatol.*, **6**, 661–675, https://doi.org/10.1002/joc.3370060607.

Beaulieu, C., and R. Killick, 2018: Distinguishing trends and shifts from memory in climate data. *J. Climate*, **31**, 9519–9543, https://doi.org/10.1175/JCLI-D-17-0863.1.

Bertsimas, D., and J. Tsitsiklis, 1993: Simulated annealing. *Stat. Sci.*, **8**, 10–15, https://doi.org/10.1214/ss/1177011077.

Chan, D., G. Gebbie, and P. Huybers, 2023: Global and regional discrepancies between early-twentieth-century coastal air and sea surface temperature detected by a coupled energy-balance analysis. *J. Climate*, **36**, 2205–2220, https://doi.org/10.1175/JCLI-D-22-0569.1.

Costa, A. C., and A. Soares, 2009: Homogenization of climate data: Review and new perspectives using geostatistics. *Math. Geosci.*, **41**, 291–305, https://doi.org/10.1007/s11004-008-9203-3.

Cowtan, K., R. Rohde, and Z. Hausfather, 2018: Evaluating biases in sea surface temperature records using coastal weather stations. *Quart. J. Roy. Meteor. Soc.*, **144**, 670–681, https://doi.org/10.1002/qj.3235.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Gallagher, C., R. Killick, R. Lund, and X. Shi, 2022: Autocovariance estimation in the presence of changepoints. *J. Korean Stat. Soc.*, **51**, 1021–1040, https://doi.org/10.1007/s42952-022-00173-5.

Kadow, C., D. M. Hall, and U. Ulbrich, 2020: Artificial intelligence reconstructs missing climate information. *Nat. Geosci.*, **13**, 408–413, https://doi.org/10.1038/s41561-020-0582-5.

Kennedy, J. J., N. A. Rayner, C. P. Atkinson, and R. E. Killick, 2019: An ensemble data set of sea surface temperature change from 1850: The Met Office Hadley Centre HadSST.4.0.0.0 data set. *J. Geophys. Res. Atmos.*, **124**, 7719–7763, https://doi.org/10.1029/2018JD029867.

Killick, R., P. Fearnhead, and I. A. Eckley, 2012: Optimal detection of changepoints with a linear computational cost. *J. Amer. Stat. Assoc.*, **107**, 1590–1598, https://doi.org/10.1080/01621459.2012.737745.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. Geophys. Res.*, **116**, D19121, https://doi.org/10.1029/2011JD016187.

Li, S., and R. Lund, 2012: Multiple changepoint detection via genetic algorithms. *J. Climate*, **25**, 674–686, https://doi.org/10.1175/2011JCLI4055.1.

Li, Y., and R. Lund, 2015: Multiple changepoint detection using metadata. *J. Climate*, **28**, 4199–4216, https://doi.org/10.1175/JCLI-D-14-00442.1.

Lund, R. B., C. Beaulieu, R. Killick, Q. Lu, and X. Shi, 2023: Good practices and common pitfalls in climate time series changepoint techniques: A review. *J. Climate*, **36**, 8041–8057, https://doi.org/10.1175/JCLI-D-22-0954.1.

Meinshausen, N., S. Sippel, E. M. Fischer, V. Humphrey, R. A. Rohde, I. E. de Vries, and R. Knutti, 2022: New land vs. ocean based global mean temperature reconstructions reveal high consistency except for early 20th century ocean cold anomaly. *2022 Fall Meeting*, Chicago, IL, Amer. Geophys. Union, Abstract GC23C-01.

Menne, M. J., and C. N. Williams, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**, 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

——, ——, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018: The Global Historical Climatology Network monthly temperature dataset, version 4. *J. Climate*, **31**, 9835–9854, https://doi.org/10.1175/JCLI-D-18-0094.1.

Morice, C. P., and Coauthors, 2021: An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data set. *J. Geophys. Res. Atmos.*, **126**, e2019JD032361, https://doi.org/10.1029/2019JD032361.

Osborn, T. J., P. D. Jones, D. H. Lister, C. P. Morice, I. R. Simpson, J. Winn, E. Hogan, and I. C. Harris, 2021: Land surface air temperature variations across the globe updated to 2019: The CRUTEM5 data set. *J. Geophys. Res. Atmos.*, **126**, e2019JD032352, https://doi.org/10.1029/2019JD032352.

Peterson, T. C., and Coauthors, 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol.*, **18**, 1493–1517, https://doi.org/10.1002/(SICI)1097-0088(19981115)18:13%3C1493::AID-JOC329%3E3.0.CO;2-T.

Rohde, R., and Coauthors, 2013a: Berkeley Earth temperature averaging process. *Geoinf. Geostat.*, **1** (2), 1–13, https://doi.org/10.4172/gigs.1000103.

——, and Coauthors, 2013b: A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinf. Geostat.*, **1** (1), 1–7, https://doi.org/10.4172/2327-4581.1000101.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Shi, X., C. Beaulieu, R. Killick, and R. Lund, 2022a: Changepoint detection: An analysis of the Central England temperature series. *J. Climate*, **35**, 6329–6342, https://doi.org/10.1175/JCLI-D-21-0489.1.

——, C. Gallagher, R. Lund, and R. Killick, 2022b: A comparison of single and multiple changepoint techniques for time series data. *Comput. Stat. Data Anal.*, **170**, 107433, https://doi.org/10.1016/j.csda.2022.107433.

Theil, H., 1950: A rank-invariant method of linear and polynomial regression analysis. *Adv. Stud. Theor. Appl. Econom.*, **23**, 345–381, https://doi.org/10.1007/978-94-011-2546-8_20.

Tingley, M. P., 2012: A Bayesian ANOVA scheme for calculating climate anomalies, with applications to the instrumental temperature record. *J. Climate*, **25**, 777–791, https://doi.org/10.1175/JCLI-D-11-00008.1.

Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 490–506, https://doi.org/10.1002/wcc.46.

Tukey, J., 1977: *Exploratory Data Analysis*. Addison-Wesley, 688 pp.

Venema, V. K. C., and Coauthors, 2012: Benchmarking homogenization algorithms for monthly data. *Climate Past*, **8**, 89–115, https://doi.org/10.5194/cp-8-89-2012.

Williams, C. N., M. J. Menne, and P. W. Thorne, 2012: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geophys. Res.*, **117**, D05116, https://doi.org/10.1029/2011JD016761.