Benchmarking the performance of pairwise homogenization of surface temperatures in the United States

Claude N. Williams[1], Matthew J. Menne[1], Peter W. Thorne[2]*

[1] NOAA's National Climatic Data Center, Asheville, NC, USA
[2] CICS-NC, North Carolina State University and NOAA's National Climatic Data Center, Asheville, NC, USA

An edited version of this paper was published by AGU. Copyright 2012 American Geophysical Union.

Williams, C. N., Jr., M. J. Menne, and P. Thorne (2012), Benchmarking the performance of pairwise homogenization of surface temperatures in the United States, *J. Geophys. Res.*, doi:10.1029/2011JD016761, in press.

**Corresponding author:*

Matthew J. Menne
NOAA's National Climatic Data Center
151 Patton Avenue
Asheville
NC 28801
USA
Matthew.Menne@noaa.gov

Abstract

Changes in the circumstances behind in situ temperature measurements often lead to biases in individual station records that, collectively, can also bias regional temperature trends. Since these biases are comparable in magnitude to climate change signals, homogeneity "corrections" are necessary to make the records suitable for climate analysis. To quantify the effectiveness of U.S. surface temperature homogenization, a randomized perturbed ensemble of the USHCN pairwise homogenization algorithm was run against a suite of benchmark analogs to real monthly temperature data. Results indicate that all randomized versions of the algorithm consistently produce homogenized data closer to the true climate signal in the presence of widespread systematic errors. When applied to the real-world observations, the randomized ensemble reinforces previous understanding that the two dominant sources of bias in the U.S. temperature records are caused by changes to time of observation (spurious cooling in minimum and maximum) and conversion to electronic resistance thermometers (spurious cooling in maximum and warming in minimum). Error bounds defined by the ensemble output indicate that maximum temperature trends are positive for the past 30, 50 and 100 years, and that these maximums contain pervasive negative biases that cause the unhomogenized (raw) trends to fall below the lower limits of uncertainty. Moreover, because residual bias in the homogenized analogs is one-tailed under biased errors, it is likely that maximum temperature trends have been underestimated in the USHCN. Trends for minimum temperature are also positive over the three periods, but the ensemble error bounds encompass trends from the unhomogenized data.

1. **Introduction**

Despite the relative data richness and substantial efforts to analyze the record over several decades [*Karl et al.,* 1986, 1988; Karl and Williams, 1987; *Quayle et al.*, 1991; *Hubbard and Lin*, 2007; *Menne et al*. 2009], scientific [*Peterson*, 2006; *Pielke et al.*, 2007a,b; *Menne et al.* 2010] and political [*http://science.house.gov/hearing/full-committee-hearing-climate-change*] controversy remains over estimates of the long-term rate of temperature change reported for the conterminous United States. Many changes in instrumentation, observing practice and siting conditions have occurred over time, all of which can alter the bias of surface temperature measurement. Moreover, the nature and timing of these changes is not always known, and overlapping measurements are rarely available during transition periods from one observing system to another. While the importance of this kind of information is well known (e.g., GCOS, 2004), most surface temperature measurements come from networks that are not specifically managed to meet the desired standards for climate. Rather they were designed to meet the needs of agriculture, hydrology, weather forecasting, etc. Consequently, there is a need to undertake statistically-based adjustments after the fact (homogenization) based upon incomplete station history information. Because each observing network site has its own set of unique non-climatic artifacts, identifying breaks and estimating adjustments is subject to some level of uncertainty, and it is unlikely that any single approach will work well for every situation [Venema et al. 2011].

For the United States Historical Climatology Network (USHCN), a fully automated homogenization algorithm has been developed based upon pairwise

neighbor comparisons [*Menne and Williams,* 2009]. This automation allows for the homogenization of the large number of surface temperature records in the network and provides traceability and reproducibility of methods. Nevertheless, creating a comprehensive breakpoint identification and adjustment scheme requires making a number of judgment calls at various decision points in the algorithm [*Thorne et al.,* 2005] no matter how robust the underlying statistical methods might be. Decisions are required for all processing steps from how to define target and reference series to the particular statistical breakpoint tests applied and mechanisms for adjusting each detected break. Seemingly innocuous choices could, in theory, have large impacts upon the final product. A frank assessment of what these parameters are and allowing them to take on a range of reasonable values can reveal the algorithm's sensitivity to these choices [*McCarthy et al.,* 2008, *Titchner et al.*, 2009]. With an automated algorithm, an ensemble of solutions can readily be produced using these values. The resulting ensemble can then provide a measure of the parametric uncertainty of the algorithm.

Although important, this quantification of internal algorithm uncertainty nevertheless has limited value in informing where the output may lie with respect to the true climate signal. Use of internal system statistics such as spatio-temporal consistency of the resulting fields to pick an overall winning configuration among parametric choices has been shown to be potentially misleading in efforts to extract the climate signal [*Sherwood et al.,* 2009].

One way forward is to create a set of plausible analogs which share the likely principal characteristics of the raw data such as spatio-temporal sampling structure, noise and bias characteristics, but where, unlike the real world, the truth is known *a*

priori. Running the algorithm against such a suite allows a quantifiable benchmarking of algorithm strengths and weaknesses. When applied to real-world data, the same algorithm settings can lead to a reappraisal of real-world trends. These kinds of analog cases have been constructed and used for both paleoclimate reconstructions [*Mann and Rutherford*, 2002, *Mann et al*., 2005, *Von Storch et al.*, 2004] and more latterly radiosonde temperatures [*Titchner et al*., 2009, *Thorne et al*., 2011a]. Benchmarking has also been carried out for surface networks much smaller than the conterminous U.S. (CONUS) network for temperature and precipitation [*Venema et al.,* 2011]. Such a paradigm is also commonplace in other scientific areas such as metrology (termed software testing) and has been called for as part of the incipient global surface temperatures initiative [*Thorne et al*., 2011b].

This paper describes initial results from applying such an approach to the USHCN surface temperature record. An initial ensemble of 100 randomized versions of the *Menne and Williams* [2009] algorithm is applied across eight analog datasets and the real world observations, the latter both with and without removal of time of observation biases beforehand. This analysis concentrates solely upon large-scale long-term trend metrics because this is currently of greatest scientific and societal interest. However, the ensembles result in a rich set of data including estimated adjustments at the station level both for the analogs and the real world data.

The paper is organized as follows. The methodology of *Menne and Willams* [2009] is briefly summarized in Section 2. In Section 3 those methodological choices identified as decision points are outlined and the allocation of a sensible range of values is discussed along with the ensemble creation methodology. Section 4 outlines the creation of a set of eight analog worlds – details of which were kept

from the first two authors (the algorithm developers) until they had produced the ensembles. In section 5, ensemble results using the analog worlds are discussed. Section 6 provides a summary of the ensembles produced using real-world data as input in the context of implications for our understanding of real-world changes in surface temperatures in the lower 48 states. Finally, some discussion and conclusions are offered in sections 7 and 8. In addition to provision of extra analyses and information in Supplementary Information, full data including the analogs, ensemble output, and automated dataset creation algorithm code provision will accompany publication of this paper at {ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/algorithm-uncertainty}.

2. The automated pairwise breakpoint identification and adjustment algorithm

To produce the USHCN version 2 monthly temperature data [*Menne et al.*, 2009], apparent shifts in measurement bias in temperature records from USHCN stations were detected and corrected through a relative homogeneity testing scheme (Conrad and Pollack, 1962) based on automated pairwise comparisons of mean monthly maximum and minimum temperature series. In particular, the algorithm seeks to identify and adjust for cases in which there is a shift in one station series relative to many others, the assumption being that a spatially isolated and sustained shift in the mean of the temperature series is an artifact caused by factors other than changes in weather and climate. The specific processing steps are briefly summarized below.

1. First, target-neighbor differences are calculated between each mean monthly maximum and minimum temperature series and a number of corresponding series from surrounding Cooperative Observer stations. Serial monthly

differences are used rather than separate series for each calendar month or season. Since each station also gets treated as a target series, pairwise differences are formed between large fractions of all possible combinations of station series pairs in all localized regions around each station in the network. Although not always possible, the algorithm also pairs stations to ensure that a minimum number of neighbors have data coincident with the target series at any given time.

2. Next, the Standard Normal Homogeneity Test (SNHT; *Alexandersson*, 1986) for undocumented changepoints is used to identify breaks in each paired difference series. A hierarchy of changepoint models is used to distinguish whether the changepoint appears to be a change in mean with no trend [*Alexandersson and Moberg*, 1997], a change in mean within a general trend [*Wang*, 2003], or a change in mean coincident with a change in trend [*Lund and Reeves*, 2002]. A break in any one difference series is temporarily attributed to both station series used to calculate the differences. The result of this step is a matrix of potential changepoint dates for each station series.
3. The matrix of changepoint dates is then "unconfounded" by identifying the station that is a common factor in multiple difference series that share the same changepoint date (see Menne and Williams, 2009 for more detail).
4. After the unconfounding step, breaks in the difference series attributed to a particular station may be assigned to nearly, but not exactly, the same month. This is because identifying the timing of undocumented breaks is subject to some sampling uncertainty and the detected break date in a group of target-neighbor reference series will likely cluster around the true date. To

distinguish cases in which nearby dates represent the same break from those that are separate breaks, a window of uncertainty is estimated as function of the estimated break size (with smaller jumps having wider windows of uncertainties than large jumps). Any cluster of undocumented changepoint dates that falls within overlapping windows of uncertainty is conflated to a single date according to

a. a known change date as documented in the target station's history archive (meaning the discontinuity appears to be documented), or
b. the most common undocumented changepoint date within the uncertainty window (meaning the discontinuity appears to be truly undocumented).

5. Steps 1–4 are necessary simply to identify undocumented changepoints in the USCHN temperature series. In many cases station histories are also available. Where possible, the dates of documented change events are combined with the undocumented breakpoint dates to ensure that any documented change not implicated in step 4 is evaluated as an additional potential break. Adjustments are then determined by calculating multiple pairwise estimates of the step change using overlapping segments from neighboring series that appear to be homogeneous for a minimal period before and after the target breakpoint. The range of pairwise estimates for a particular break is used to determine a confidence interval for the size of the adjustment. When this confidence interval includes zero, an adjustment is not made. Adjustments are treated as seasonally invariant.

3. Identification of algorithm parameters and ensemble settings

Across the above steps we have identified a total of 17 distinct parameters associated with decision points in the algorithm that required a judgment call. These parameters and their allowable values are provided in Table 1, grouped loosely by their role in the processing steps as described below. Different permitted values effectively reflect a range of what the authors consider to be plausible choices based upon fundamental assumptions about data and metadata veracity and spatio-temporal coherency of climate anomalies. Such choices reflect the parametric uncertainty/sensitivity of the pairwise algorithm but do not entail the creation of a suite of completely independent algorithms. Independent algorithms would elucidate structural uncertainty [*Thorne et al., 2005*] and their development and robust evaluation is also strongly encouraged [*Thorne et al., 2011b*].

3.1 Choosing neighbors to test for relative homogeneity

In the default setting of the algorithm, neighbors are selected using both distance from target (key word=NEIGH_CLOSE) and correlation with target (key word=NEIGH_CORR). Specifically, a maximum of the 40 highest correlated among the hundred nearest neighbors are used (key word=NEIGH_FINAL). In the randomized versions, between 80 and 200 of the nearest station series are considered and from these a maximum of 20 to 80 of the highest correlated series are selected. Correlation is calculated using first differences in the default version and an effort is made to ensure that at least seven of the selected neighbors have data coincident with the target at any given time (key word=MIN_STNS). In randomized versions, correlation can also be calculated directly using monthly anomaly series rather than first differences or only the closest neighbors are used

regardless of correlation (NEIGH_CORR). The minimum number of neighbors is also allowed to vary in the randomization, as is the minimum correlation between target and neighbor in versions where both distance and correlation are used (key word=CORR_LIM).

3.2 Resolving breaks in the difference series

Breaks in all difference series are resolved using SNHT with a semi-hierarchical splitting algorithm [*Menne and Williams*, 2005]. In the default setting, a 5% significance level is used (key word=SNHT_THRES) whereas in the randomized versions the value can be 1, 5 or 10%. As described in step 2 above, an evaluation of the nature of the break is also conducted at this stage to determine whether a trend may be present in addition to or instead of a step. In the default setting, the most appropriate model is selected using the Bayesian Information Criterion (BIC; *Schwartz*, 1978; key word=BIC_PENALTY). In the randomized versions, model selection can be evaluated using the Akaike Information Criterion (AIC; *Akaike*, 1973), or not conducted at all.

3.3 Identifying the cause of the break

Attributing the cause of a break requires multiple target-neighbor difference series for a particular target to have coincident breaks. In the default algorithm setting, at least 2 difference series must implicate the target (key word=CONFIRM). In the randomization, this number is allowed to range from 2 to 5. The date of the apparent break is assigned using the most frequent breakpoint date as determined by SNHT or via a metadata event (if available) for those dates that fall within

overlapping windows of uncertainty for the timing (key word=AMPLOC_PCT). Empirical confidence limits are used to quantify the timing uncertainty of a break and limits of 90, 92.5 (default) and 95% may be used. In addition, metadata dates can be used in conjunction with undocumented changepoint detection as in the default (key word=SHF_META), not used at all, or used exclusively without conducting a search for undocumented breaks.

3.4 Estimating the magnitude of the break

Estimating the size of each break in a target series requires calculating the magnitude of a jump in the target-neighbor difference series using neighbors that appear to be homogeneous for some number of months before and after the target break (key word=ADJ_WINDOW). The default value is ±24 months, but in the randomization the number ranges from no minimum at all up to ±120 months. If the target series appears to have successive breaks that are too close in time to adjust (key word=ADJ_MINLEN), then an adjustment is made for the combined effect of the two or more breaks. The minimum interval between adjustable breaks can range from 18 months in the default up to 48 months in the randomized versions.

More than one pairwise estimate of the target break size is required to make an adjustment (key word=ADJ_MINPAIR) and these values are used to quantify the uncertainty in the adjustment. In the default version at least three estimates of break size are required whereas in the randomization, this number may range from 2 to 5. Further, the estimates of break size may be subject to an outlier test where possible (key word=ADJ_FILTER). The default setting uses a variant of the Tukey

outlier test [*Tukey*, 1977], but the randomized versions may also use the Bayesian Information Criterion to determine whether a step-change is justified, both tests, or none at all. Following the outlier test, the median break size (default), the average break size or the average of the inter-quartile range is used as the final breakpoint adjustment (key word=ADJ_EST). Finally, because there may be limitations in the number of neighbors and the length of their homogeneous segments before and after some target breakpoints, there is an option to increase the length of homogeneous segments at neighboring series by merging segments of these neighbors where the confidence limits for the magnitude of a break include zero. This step allows an increase in the number of target break size estimates in data sparse periods or when breaks are clustered throughout a region and in time. The default option is to allow this merging (key word=NS_LOOP), but the randomized versions may or may not do the merging step.

3.5 Creating a set of ensembles

To create the 100 member ensemble a methodology similar to that employed by *Titchner et al.* [2009] was followed. A random number generator was used to seed the value for each tunable parameter in each ensemble member. This ensures that a broad range of plausible solution space is spanned but comes at a cost vis-à-vis potential for systematic investigation. Some keywords are inter-related and any illogical combinations were precluded. The specific settings for each of the 100 ensemble members are tabulated in supplementary information. In addition to the randomized ensemble the operational (default) configuration [*Menne and Williams*, 2009] was also run against the analogs.

4. Creation of analog cases

To ensure plausible geographical data variability and teleconnections across the conterminous U.S., the analog benchmarks were derived from gridded surface temperature output from Global Climate Models (GCMs). Six different climate model runs were downloaded from the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset [*Meehl et al.*, 2007], each of which was sub-sampled in space and time to an observational data mask that matches the U.S. network. GCMs were used as the basis for generating the analogs because they reproduce many of the fundamental surface temperature characteristics. Sensitivity to the choice of model fields is assessed by applying the same error structure to four different model estimates to create one family of analogs whose differences are a function only of the underlying climate model evolution.

The data mask applied to the GCM output replicates the geographic distribution and periods of record for both the USHCN stations as well as the larger U.S. Cooperative Observer (COOP) network (Fig. 1) whose stations are used as neighbors to homogenize the USHCN subset [*Menne et al.,* 2009]. The total number of stations with 8 or more years of temperature records is about 7,200 in the COOP, of which 1218 constitute the USHCN subset. The analogs reproduce this data record for the twentieth century. Data for each analog station record were sampled from the nearest gridbox with no additional interpolation. Because the models have much coarser resolution than the U.S. COOP station density, climatological offsets and random noise were applied to the resampled model data before adding any errors

to the generated station records. This ensures that nearby 'stations' arising from the same GCM gridpoint are not identical and that the analog station series more closely resemble potential differences caused by elevation and other features unique to each local environment.

The efficiency of neighbor-based homogenization algorithms depends largely on the magnitude of the covariance between neighboring station series, which is generally related to station density. If the covariance is too high in the analogs, the test results will be overly pessimistic because the breakpoints will be harder to identify in the analog world than in the real-world and vice-versa. The noise added to the analog series was calibrated to have the approximately the same characteristics as inter-station statistics following homogenization of USHCN under *Menne et al*. [2009]. Specifically, the standard deviation of the inter-station difference series and their AR(1) autocorrelation were assessed (Fig. S1) across the network as a whole. While the real and analog world station covariance structures were designed to be broadly consistent, the homogenization results will reflect any deviation in the covariance between generated series and what occurs in the real COOP network.

Four principal break structures were imposed on the analogs by the third author, the nature of which was unknown to the first two authors until the 100 member ensembles were produced for each analog. The imposed errors were specifically designed to test the efficacy of the algorithm's ability to estimate the true long-term trend at the regional scale. The analogs were intended to cover a range of scenarios from overly-simple to arguably too challenging to ascertain the performance of the pairwise homogenization algorithm under a number of

scenarios. Specifically, if a homogenization algorithm cannot cope with a simple error structure then its use on real-world data is problematic. Likewise, creating difficult, but not impossible benchmarks should allow algorithm developers clearer goals for improvements to address the tougher issues that may exist in the real-world.

The details of the errors are provided in Table 2, but in all cases breaks were assigned as seasonally invariant step changes with varying degrees of associated metadata. The real-world situation is undoubtedly more complex; however, a trade off is required in this initial analysis (where the concept is being applied to this particular problem for the first time) between complexity and ability to analyze the results. Future exercises should include more complex (realistic) error structures [e.g. *Venema et al.*, 2011] and global benchmarking [*Thorne et al.,* 2011b] of this algorithm, which is now also used to create the global GHCN v3 product [*Lawrimore et al.,* 2011]. The details of the analogs are described in more detail below in order of the complexity of the error models.

4.1 "Perfect data"

The '*perfect data*' analog was produced solely to test whether the algorithm can do "harm" by identifying numerous false breaks and substantially altering the real-world behavior in the unlikely event that the real-world raw data are perfectly homogeneous. It consists of exactly the same data as that for '*Clustering and sign bias – c20c1*' (Section 4.4) prior to the addition of errors.

4.2 "Big breaks, good metadata"

The '*Big breaks, good metadata'* analog consists of predominantly large breaks with no preferential sign bias and the timing of each break is recorded in associated metadata. While the imposed breaks have a large standard deviation, they are normally distributed around zero, which means that there are a number of very small breaks that may not be considered statistically significant. The average period between breaks is twenty years, but error seeding was random (for all analogs) so there are stations with breaks in closer succession and/or with more than the average of 5 breaks in the record (and vice-versa).

4.3 "Mixed break sizes, some clustering"

In the '*Mixed breaks, some clustering'* analog a more plausible error structure was added. It is known, for example, that the USHCN network experienced at least two pervasive changes that afflicted the majority of the network with a change in observation time and a move from liquid in glass (LiG) thermometers in Stevenson screens to the electronic resistance thermometer known at the Maximum/Minimum Temperature Sensor [MMTS; *Quayle et al.*; 1991; *Menne et al.*, 2009]; that the metadata is far from perfect and may be less complete for the earlier parts of the record; and that not all breaks will be large. In this analog these aspects were added, but the clustering of similar breaks is relatively relaxed in time compared to our current knowledge of the real world data and the number of applied breaks is still arguably lower than the likely frequency of real-world breaks with an average return period of between fifteen and twenty years.

4.4 "Clustering and sign bias" family

The error structure applied to these analogs contains more breaks and some exhibit a much tighter degree of clustering of similar events than in '*Mixed breaks some clustering'* reflecting that the majority of the LiG to MMTS transition happened well within a decade [*Menne et al.*, 2009]. Many breaks have a sign bias leading to a false warming trend in the 'raw' analog world data. This is opposite to the suspected behavior in the real-world where the raw data are apparently negatively (cool) biased [*Menne et al*., 2009]. Whether the sign bias is positive or negative is less important than adding an overall sign preference to ascertain whether the shift between raw and adjusted series is likely to be uncovering a real trend bias in the real-world data or occurring solely by chance.

Four variants with this error structure were produced using different GCMs to address whether algorithm performance depends on the presence of an underlying signal and the phasing and nature of the climate variations arising from natural variability. The first two variants '*clustering and sign bias – c20c1'* and '*clustering and sign bias – c20c2'* both consider output from 20th Century climate simulations but from different models. These are meant to sample any potential impacts of the differences in model physics, phasing and characterization of natural variability and inclusion versus exclusion of specific forcings [*Santer et al.*, 2005, 2006]. Nevertheless, both models have grossly the same multi-decadal characteristics of an accelerating warming trend. If the ability to recover the true trend is impacted by this choice then it is likely that algorithm performance is sensitive to natural variability. The third variant of this family, '*clustering and sign bias – control',* again uses a different model but in this case no external forcings are changed so there is no forced signal component. Comparing this to the first two

analogs in the family permits exploration as to whether algorithm performance is impacted by the presence or absence of a real underlying climate change signal. The final analog '*clustering and sign bias – committed*' allows an assessment of sensitivity to shape of the underlying signal. The c20c forcings runs approximate in their underlying forcing an exponential increase due to greenhouse gas increases, which is reflected in the multi-decadal temperature changes. A committed forcings run involves no increase in forcing but starts with the coupled climate system substantially out of equilibrium. The forced change component in such a run is therefore more akin to a natural logarithm with relatively rapid changes early in the record as faster response components catch up with the now stable forcing, tailing off later on as slower deep ocean responses continue.

4.5 "Very many mainly small breaks"

The final analog '*Very many mainly small breaks*' represents the most pessimistic set of assumptions about the errors. A small percentage of the breaks are large, but most are small. There are breaks on average once every ten years throughout the network and forty percent of the breaks have no metadata recorded. Furthermore, sixty percent of the breaks have a sign tendency associated with them. Despite not explicitly including clustering, this analog is arguably hardest for any dataset algorithm to cope with. First, any breakpoint algorithm is going to have real trouble finding and adjusting for small breaks without greatly inflating the false positive count, yet these breaks constitute real units of red noise that project most strongly onto the trend. Secondly, with so many breaks having a sign bias the failure to detect a substantial proportion of these is likely to lead to biases, on average, in

the adjustments because apparently homogeneous neighbor segments will not always be so and in these cases will yield a systematic adjustment tendency. Lastly, having so many breaks in the data will lead to a much greater preponderance to have intra-station breaks in close proximity. Any algorithm will struggle when the interval between breaks is short relative to the time step regardless of break size because a smaller homogeneous segment population requires a larger test statistic value for significance to be attained.

5. Results against analog cases

The pairwise homogenization algorithm produces a list of breakpoint dates and adjustments for each input series. Although it is possible to evaluate results at the individual station series level, the focus here is on the aggregate, network-wide impacts as reflected in changes to the regional mean value. We present these aggregate results beginning with the simplest analog error structure and moving progressively to the more complex models.

Figure 2 provides a geographic perspective of the trends in the "perfect data" analog both for the raw input data (Fig. 2a) and for the data homogenized by the default version of the algorithm (Fig. 2b). The trends were calculated by interpolating the annual temperature values to a 0.25x0.25 degree grid and then calculating the trend for each grid box as described in *Menne et al.* [2009]. The default version of the algorithm essentially preserves the pattern of trends although there appears to be some minor smoothing of the spatial pattern. Nevertheless, in the case of "perfect data", no version of the pairwise algorithm makes unwarranted adjustments sufficient to move the average CONUS trend away from the true trend,

and the average series produced by the 100 randomized versions of the algorithm are indistinguishable from those based on the raw input data (see Supplementary Information).

In the "Big breaks, perfect metadata" case, the unadjusted input data are characterized by a noisy, heterogeneous field of trends caused by the imposition of random breaks in the network throughout the series. As shown in Fig. 3a, the impact is a mix of trends with positive and negative biases. In this case, the default algorithm comes close to reproducing the true spatial pattern and magnitude of trends (Figs. 3b and 3c), which is expected given that the timing of all breaks is known. Nevertheless, some randomized versions of the algorithm do not make use of the metadata and treat all breaks as undocumented. Further, the use of a significance test when estimating the magnitude of each break means the recovery of the true climate signal from the input data is not necessarily perfect. However, since there is not an overall bias associated with the imposed errors, the randomized versions of the algorithm all produce CONUS average trends that do not deviate substantially from the true background trend (Fig. 4) and there is no sign preference to the potential residual error.

In the "Mixed break sizes, some clustering" analog, errors are clustered in time (between 1915 and 1975 and somewhat more heavily from 1915 to 1945), and a sign preference is present in the errors. In this case, the homogenized trends since 1900 and since 1950 from the ensemble are all greater than the raw input trend (Fig. 5), an indication that the algorithm is accounting for the sign bias in the imposed errors during the periods when the errors are concentrated.

In the "Clustering and sign bias" family of analogs, the imposed errors exhibit an even larger sign preference and are more clustered in time, including nearer to the end of the series, which biases average trends for all periods since 1900. The impact of the sign bias on the raw input trends for the full period can be seen in Fig. 6. Relative to the true values (Fig. 6b) a larger number of trends are too high rather than too low in the unadjusted data (Fig. 6a). Nevertheless, the default version of the pairwise homogenization algorithm comes close to reproducing both the magnitude and pattern of the underlying temperature trends (Fig. 6c) in spite of the sign preference. As shown in Fig. 7, all randomized versions of the algorithm produce homogenized series that bring the CONUS average closer to the true value for all trend periods, with some algorithm configurations, including the default version, yielding results very close to "truth" - moving the trend more than 95% percent towards the true climate signal. In particular, the impact of the pervasive positive errors seeded in 70% of the analog series after 1980 is reduced by all ensemble members. Notably, the potential residual error is essentially one-tailed in this case; there is a low probability of overcompensating for the bias changes by a small amount.

Figure 8 provides a summary overview of the "Clustering and sign bias" family of analogs (and additional time series are provided as Supplementary Information). Because each of these four analogs was seeded with identical errors, any difference in homogenization performance for a particular ensemble member is a function only of the presence or absence of a forced response component and the timing and patterns of natural internal variations simulated by the various underlying models. Results indicate that while the efficiency of individual members

is somewhat dependent on the nature of the underlying climate signal and covariance structure, the relative performance of each member measured by the degree to which the true trend is recovered remains largely unchanged from analog to analog within the family. In other words, the performance of any particular version of the algorithm appears to be largely—but not completely--invariant of underlying climate signal as shown in Fig. 9. Moreover, a comparison of Figures 4, 7, and 8 also suggests that the underlying error structure is a more fundamental consideration in the ability of the algorithm to retrieve the true underlying climate signal rather than the nature of the climate signal itself. In light of this, it may be possible to choose a number of pairwise algorithm configurations that should be expected to be relatively good performers under a wide variety of error characteristics.

Results for the most challenging analog "Very many small breaks with sign bias" are summarized in Figs. 10 and 11. In this case, a large percentage of the breaks are likely below the magnitude that can be efficiently detected by the pairwise (or perhaps any) algorithm. Consequently, the various ensembles produced by the randomized versions of the algorithm do not move the trend far enough towards the true trend value (Fig. 10). Likewise, the geographic distribution of trends (Fig. 11) indicates that the systematic bias caused by the imposed errors are only partially removed by the homogenization algorithm, the consequence of which is a residual mean bias that underestimates the true CONUS trend and a heterogeneous field of trends.

Finally, we note that a 100-member randomization was considered at the outset to be sufficient to explore the sensitivity of the various parameters, especially

since not all of them were expected to have a substantial impact on the results. By way of confirmation, the "clustering and sign bias-C20C1" analog was run through 500 randomizations of the algorithm and the results were compared to the original 100 member ensemble as well as smaller numbers of combinations. As Figs. S6-S10 indicate, the median and interquartile ranges are well represented with 100 members and the worst case scenario implication from this expanded randomization is that the range of the ensemble trends may underestimated by about 25%. However, it is worth noting that the only outlier in the expanded 500 member ensemble not captured by the 100-member ensemble resulted from a particularly conservative set of settings that minimized the impact of the homogenization. More generally, it is the conservative tail, which minimizes adjustments, that is poorly quantified with smaller ensemble sizes rather than the more aggressive tail of the distribution that samples solutions closer to the target truth. In future the potential exists to massively parallelize such dataset creation through citizen scientists and their IT capabilities akin to e.g. climateprediction.net [*Allen,* 1999] if the pairwise homogenization code can be made suitably portable and platform independent. This could also open up new opportunities such as derivation of a neural network algorithm tuning approach either explicitly or through, for example, interfacing with the serious gaming community [*Krotoski,* 2010].

To summarize, based on all analog results we conclude that:

1. In cases where there is no sign bias to the seeded errors, the randomized versions of the algorithm produces results clustered around the true trend.
2. For cases in which there were errors seeded with a sign bias, all randomized versions of the algorithm moved the trend in the correct direction.

3. Rather than overcorrect, the randomized algorithms generally do not correct the trend enough in the presence of errors with a sign bias because of incomplete adjustments that bias the underlying trends. The propensity to under-correct is sensitive to the frequency and magnitude of imparted breaks with more frequent and smaller breaks leading to more incomplete corrections.
4. The algorithm is potentially capable of adjusting data even when pervasive network wide quasi-contemporaneous changes of a similar nature occur.
5. Although algorithm performance is somewhat impacted by natural climate variations and the presence of forced changes, this impact is secondary to that of the error structure imparted on the raw observations. The error structure, which is unknown in the real-world, is the primary limiting factor on algorithm efficiency.

6. Reassessment of likely real-world trends

We applied the same 100 randomized versions of the algorithm (as well as the default version) to the real-world monthly mean maximum and minimum temperature data. Both the raw temperature data and data first corrected for changes in the time of observation [*Karl et al.* 1987] were used as input to the pairwise algorithm. (Note that the pairwise algorithm is run operationally on data already de-biased for changes in the time of observation [*Menne et al.*, 2009]). This correction relies solely upon intra-station statistics and metadata and accounts for the specific systematic impact that changes in the time of observation bias (TOB) have on monthly mean temperatures. Although TOB changes are small at many

locations, there is no *a priori* reason why the pairwise algorithm cannot be applied to data without this step, especially since many changes in time of observation coincide with other station changes that are evaluated by the pairwise algorithm following the TOB adjustment. Running the pairwise algorithm on both raw and TOB-debiased input data provides a useful quasi-independent check on the TOB adjustment itself and is also a test of the skill of the pairwise algorithm independent of the benchmarks.

Trends in monthly maximum and minimum temperatures for three separate sub-periods and for all ensemble members are shown in Figure 12 (time series for the ensemble are provided as Supplementary Information). Unlike for the analog data the true trend in the real data is unknown. While there is considerable spread in the ensemble, all versions of the pairwise algorithm produced adjusted maximum temperature data with trends higher than in the raw data (Figure 12a). The default (operational) version of the algorithm produced trends above the median of all solutions, but not close to the highest. In these respects, results for the real world U.S maximum temperatures have some resemblance to the "clustering and sign bias" family and the "very many mainly small breaks" analog results.

Homogenization results for maximum temperature data with the prior correction for TOB are shown in Fig. 12b. Because the TOB impact is negative for both maximum and minimum temperatures [*Vose et al.,* 2003], mean monthly maximum temperatures adjusted for TOB have higher trends than the raw data [*Menne et al.* 2009]. Nevertheless, even with this prior correction, all versions of the pairwise homogenization algorithm still yield CONUS average maximum temperature trends that exceed the TOB-only adjusted input data for all periods.

We can therefore infer that other network changes (e.g., instrument changes) have artificially reduced maximum temperature trends like the TOB. Not surprisingly, the greatest differences between the input data (raw or de-biased for TOB) and the pairwise adjusted data occur in the post-1979 period, when the widespread installation of the MMTS led to an artificial decrease in maximum temperatures [*Qualye et al.*, 1991; *Hubbard and Lin*, 2006, *Menne et al.*, 2010]. Notably, trends from the operational (default) version for maximum temperatures are almost identical whether TOB adjustments are applied prior to instigation or left to the algorithm to adjust for.

Given that all ensemble members move the maximum temperature trend away from the raw and TOB-adjusted values in the same manner as the analog data with sign bias errors, uncertainty estimates for maximum temperatures can be considered as essentially one-tailed with the raw data forming an absolute lower bound for confidence limits defining the true magnitude of maximum temperature trends. Realistically, the TOB-only corrected data very likely form this lower boundary since all algorithm versions move the raw input trend in the same direction as the TOB- only adjusted data, and all versions further increase the TOB-only adjusted trends when these data are used as input. Regarding the upper bound, it is quite possible that the operational version of the homogenization algorithm is underestimating the true magnitude of U.S. average maximum temperature trends, in agreement with *Menne et al.* [2010] and *Fall et al.* [2011].

For minimum temperatures, the impact of homogenization depends largely on the period over which the trend is calculated (Figure 12c). In the long-term (1900-2010), the randomized algorithms are divided between increasing and

decreasing the U.S. trend relative to the raw value. In contrast, members tend to increase the trend for the period 1950-2010, but reduce it for the period 1979-2010. This suggests that the magnitude of the time of observation bias likely dominates the period after 1950, but not necessarily after 1979. With the prior TOB correction (Figure 12d), all members reduce the 1979-2010 trend, consistent with evidence that the widespread transition from liquid-in-glass thermometers to electronic thermistors led to an artificial increase in minimum temperatures, which likely overwhelms the impact of any TOB changes during this period. Again, the operational algorithm version is broadly similar for all trend periods regardless of whether TOB adjustments are applied first.

For the longer-term trends (1900 and 1950 onwards), the ensembles encompass the raw data, even when first corrected for TOB. This suggests that there are factors causing breaks with a negative sign bias before 1979 (in addition to the TOB) that are offsetting the largely positive shifts caused by the transition to MMTS afterwards. For example, there may have been a preference for station relocations to cooler sites within the network, that is, away from city centers to more rural locations especially around the middle of the twentieth century [*Hansen et al.* 2001]. Detecting undocumented breaks in the pre-1979 minimum temperature data may also play a role since there appears to be a bias in favor of undocumented negative shifts [*Menne et al.* 2009]. However, short of a more thorough analysis into the cause of the non-TOB related breaks prior to 1979, uncertainty estimates for the long-term minimum temperature trends must include the raw data. For the period after 1979, we can be confident that the TOB-corrected data likely form an absolute *upper* bound for minimum temperature trends with the

lower bound being less than the trends produced by the default version of the algorithm. Likewise for the period after 1950, the raw data are likely an absolute lower bound for trend estimates because of the TOB.

Finally, it is interesting to note that the ensemble ranges for maximum and minimum temperatures overlap across all trend periods (Table 3) when the raw data are used as input. This suggests that much of the observed difference in maximum and minimum temperature trends in the contiguous U.S. is linked to changes in observing practices, and that the true difference is likely much smaller than the network-wide raw data suggest as noted also by *Fall et al.* [2011].

7. Discussion

Over past few decades a great deal of effort has been devoted to collate, prepare and analyze historical surface temperature data. More recently, fully automated homogenization methods have started to emerge that are designed to remove the impacts of artifacts that bias the records of large networks. Moreover, these automated methods have been shown to be capable of achieving comparable skill to manual methods [*Venema et al.,* 2011]. This opens up the possibility of more readily exploring the sensitivities of climate data homogenization to fundamental methodological choices and parametric decisions.

Here we have assessed the sensitivity of USHCN trend estimates to the parametric choices used in the automated pairwise homogenization algorithm [*Menne and Williams*, 2009]. A brute force style, monte-carlo simulation type ensemble has been created by identifying decision points in the algorithm and allowing each to take on a range of values in random combinations. To benchmark

the performance of the algorithm, eight analogs to U.S. temperature record were created that share many of the fundamental characteristics of the observed data, except that, unlike the real world, the underlying climate signal is known. The analog cases and perturbed ensemble build upon the assessment described in *Menne and Williams* [2009] and provide further evidence that pairwise algorithm has a low false alarm for truly homogeneous input data and that it yields unbiased regional trends when network-wide errors are themselves unbiased. In cases where the analog world data contained biased errors clustered in time, all randomized versions of the homogenization algorithm moved the average trend for the conterminous U.S. closer to the true trend, though generally not far enough. This most likely reflects a twofold problem – first, the breaks that are not detected by the algorithm are likely to share the sign bias and thus their impact will not be accounted for; second, the unidentified breaks also may be aliased onto adjustment estimates for the detected breaks leading to biased estimates on average. Nevertheless, the consistency of the analog world results leads to additional confidence in interpreting the homogenization results for the real world data.

When applied to the real-world USHCN observations, the ensemble essentially reaffirms earlier conclusions regarding the pervasive biases in the raw USHCN temperature record. In the case of maximum temperature, there is strong evidence that there are widespread negative (cool) biases that artificially depress the true rate of temperature increase for all periods since 1900. These biases are the sum of time of observation change effects after 1950 as well as other changes, primarily the transition to electronic resistance thermometers beginning in the middle 1980s. Notably, the raw maximum temperature trend for the USHCN is

below the range of confidence limits defined by parametric uncertainty of the algorithm for all trend periods.

Benchmarking results for minimum temperature records appear to be somewhat more complicated, especially for the period before 1950 when parametric uncertainty is large. Since 1950, results are in agreement with earlier studies that the competing biases of changes in observation time (spurious cooling) and installation of electronic resistance thermometers (spurious warming) dominate. This competition among biases leads to raw data that underestimate the true USHCN trends since 1950 and overestimate the trends since 1979. Estimates of parametric uncertainty overlap for trends in maximum and minimum temperatures for all trend periods and suggest that some asymmetry in these trends may be due to residual biases in the adjusted data.

The analog results also revealed that the ensemble was far from equi-probable. Certain ensemble members were consistently worse than others regardless of the error structure or underlying spatio-temporal variations arising from model estimated natural variability. A supposedly equi-probable solution approach such as used in HadSST3 [*Kennedy et al.*, 2011a,b] may, however, be viable. Additional evaluation may allow further optimization and the selection of set of parameters that can produce a more equi-probable solution set. This allows any potential non-linear interdependencies between such uncertain choices to be explicitly represented.

8. Concluding remarks

The benchmarking experiment described here was carried out as a proof of concept. In this way, the sensitivity analysis is limited to parametric (internal) uncertainty sources. We encourage more benchmarking efforts like *Venema et al.* [2011] so that multiple homogenization algorithms can be run against a common set of global analogs as proposed by the surface temperatures initiative [*Thorne et al.*, 2011a]. In particular, the proposed double blind nature would be a distinct advantage, although it should be stressed that our analysis was blind in that the nature of the errors was only made available to the first two authors after the ensembles were completed. Blind studies avoid potential pitfalls associated with tuning an algorithm to perform well under certain, specific error assumptions when in reality the true error structure is unknown. In future it would be useful to consider more complex error structures with, for example, seasonal cycles or local temporary trends in addition to step-like changes.

The creation of an ensemble of pairwise algorithm solutions to assess parametric uncertainty and its application to both the real observations and eight analogs of those observations has served to strengthen our existing understanding of U.S. temperature records. The analogs indicate that the homogenization algorithm does not add spurious trends to the spatial temperature average and adjusts the data in the right direction in the presence of network-wide systematic biases, although not necessarily far enough. The benchmarking reaffirms that the dominant systematic and network-wide biases in the U.S. are caused by changes in time of observation from the mid-twentieth Century onwards (spurious cooling to both maximum and minimum temperatures) and conversion from liquid in glass to electronic resistance thermometers, primarily during the mid-1980s (spurious

cooling in maximum and warming in minimum). Results for the real-world are similar regardless of whether time of observation adjustments are applied in advance or left to the pairwise algorithm to adjust directly, building confidence both in the reality of this effect and the capabilities of the algorithm.

We conclude that raw maximum temperatures are outside the assessed range of plausible trends - the real U.S. trends are very likely greater than the raw data imply. Raw minimum temperatures are not as obviously biased, at least at the centennial timescale. Internal uncertainty for the homogenized maximum and minimum trends over the periods 1900-2010, 1951-2010 and 1979-2010 does not encompass zero so there is high confidence in the conclusion that the conterminous U.S. temperature trends are positive at these time scales. However, the internal algorithm uncertainty for the rate of temperature change indicates that the default settings used of the pairwise algorithm used to produce the USHCN Version 2 adjusted temperature data is likely underestimating maximum temperature trends.

Acknowledgements

Peter Thorne's early work was funded by Government Business Programme project PHEATS and the Joint DECC and Defra Integrated Climate Programme - DECC/Defra (GA01101) whilst employed by the UK Met Office. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Met Office Hadley Centre model output (c) Crown copyright 2005, Data provided by the Met Office Hadley Centre. The authors thank Ken Kunkel, Tom Peterson and three anonymous reviewers for

providing helpful comments on earlier drafts of the paper, as well as Dan Rowlands for helpful discussions about the generation of ensembles.

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Cs`aki), pp. 267–281. Budapest: Akad´emiai Kiad´ o.

Alexandersson, H., (1986), A homogeneity test applied to precipitation data. J. *Climatol.*, 6, 661–675.

Alexandersson, H., and A. Moberg, 1997: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *Int. J. Climatol.*, 17, 25–34.

Allen, M. R., 1999: Do it yourself climate prediction. *Nature*, **401**, 642

Collins, W. D., C. M. Bitz, M. L. Blackmon, G. B. Bonan, C. S. Bretherton, J. A. Carton, P. Chang, S. C. Doney, J. J. Hack, T. B. Henderson, J. T. Kiehl, W. G. Large, D. S. McKenna, B. D. Santer, and R. D. Smith, 2006: The Community Climate System Model Version 3 (CCSM3). *J. Climate*, 19(11), 2122-2143.

Conrad, V., and L. W. Pollack, 1962: Methods in Climatology. Harvard University Press, 459 pp.

Delworth, T. D. et al. (2006), GFDL's CM2 global coupled climate models - Part 1: Formulation and simulation characteristics. *J. Climate*, 19, 643-674.

Fall, S., A. Watts, J. Nielsen-Gammon, E. Jones, D. Niyogi, J. R. Christy, and R. A. Pielke Sr. (2011), Analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends, *J. Geophys. Res.*, 116, D14120, doi:10.1029/2010JD015146.

Global Climate Observing System. 2004. Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC. GCOS-92, WMO/TD No. 1219, World Meteorological Organization, Geneva, Switzerland.

Gordon, H. B., Rotstayn, L. D., McGregor, J. L., Dix, M. R., Kowalczyk, E. A., O'Farrell, S. P., Waterman, L. J., Hirst, A. C., Wilson, S. G., Collier, M. A., Watterson, I. G., and Elliott, T. I. (2002): The CSIRO Mk3 Climate System Model [Electronic publication]. Aspendale: CSIRO Atmospheric Research. (CSIRO Atmospheric Research technical paper; no. 60). 130 pp. (to http://www.cmar.csiro.au/e-print/open/gordon_2002a.pdf)

Hubbard, K. G., and X. Lin, 2006: Reexamination of instrument change effects in the U.S. Historical Climatology Network. *Geophys. Res. Lett.,* **33,** L15710, doi:10.1029/2006GL027069.

Johns, T. C., C. F. Durman, H. T. Banks, M. J. Roberts, A. J. McLaren, J. K. Ridley, C. A. Senior, K. D. Williams, A. Jones, G. J. Rickard, S. Cusack, W. J. Ingram, M. Crucifix, D. M.

H. Sexton, M. M. Joshi, B-W. Dong, H. Spencer, R. S. R. Hill, J. M. Gregory, A.B. Keen, A. K. Pardaens, J. A. Lowe, A. Bodas-Salcedo, S. Stark, and Y. Searl (2006), The new Hadley Centre climate model HadGEM1: Evaluation of coupled simulations, *J. Climate*, 19, 1327-1353.

K-1 model developers (2004), K-1 coupled model (MIROC) description, K-1 technical report, 1, H. Hasumi and S. Emori (eds.), Center for Climate System Research, University of Tokyo, 34pp. (http://www.ccsr.u-tokyo.ac.jp/kyosei/hasumi/MIROC/tech-repo.pdf)

Karl, T. R., and C. N. Williams Jr. (1987), An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Climate Appl. Meteor.,* **26,** 1744–1763.

Karl, T.R., C. N. Williams Jr., P. J. Young, and W. M. Wendland (1986), A model to estimate the time of observation bias associated with monthly mean maximum, minimum, and mean temperature for the United States. *J. Climate Appl. Meteor.,* **25,** 145–160.

Karl, T.R., H. F. Diaz, and G. Kukla (1988), Urbanization: Its detection and effect in the United States climate record. *J. Climate,* **1,** 1099–1123.

Kennedy J.J., N. A. Rayner, R. O. Smith, M. Saunby and D. E. Parker (2011a), Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 1: measurement and sampling errors. *J. Geophys. Res.*, 116, D14103, doi:10.1029/2010JD015218.

Kennedy J.J., N. A. Rayner, R. O. Smith, M. Saunby and D. E. Parker (2011b), Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 2: biases and homogenisation. *J. Geophys. Res.*, 116, D14104, doi:10.1029/2010JD015220.

Krotoski, A, 2010, Serious fun with computer games. *Nature*, 466, 695

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 Dataset, Version 3. *J Geophys. Res.*, 116, D19121, doi:10.1029/2011JD016187.

Lund, R., and J. Reeves (2002), Detection of undocumented changepoints: A revision of the two-phase regression model. *J. Climate*, 15, 2547–2554.

Mann ME, Rutherford S. 2002. Climate reconstruction using 'pseudoproxies'. *Geophys. Res. Letters* , 29, 10, doi: 10.1029/2001GL014554.

Mann ME, S. Rutherford, E. Wahl, C. Ammann, (2005), Testing the fidelity of methods used in proxy-based reconstructions of past climate. *J. Climate*, 18, 4097–4107.

McCarthy M. P., H. A. Titchner, P. W. Thorne,S. F. B. Tett, L. Haimberger, D. E. Parker (2008), Assessing bias and uncertainty in the HadAT adjusted radiosonde climate record. *J. Climate*, 21, 817-832.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007), The WCRP CMIP3 multi-model dataset: A new era in climate change research, *Bul. Amer. Meteor. Soc.,* **88**, 1383-1394.

Menne, M.J., and C. N. Williams Jr. (2005), Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate*, 18, 4271–4286.

Menne, M.J., and C.N. Williams (2009), Homogenization of temperature series via pairwise comparisons. *J. Climate*, 22, 1700-1717.

Menne, M. J., C. N. Williams, R.S. Vose (2009), The US Historical Climatology Network monthly temperature data, version 2, *Bull. Amer. Meteor. Soc.*, 90, 993-1007.

Menne, M. J., C. N. Williams, Jr., and M. A. Palecki (2010), On the reliability of the U.S. surface temperature record. *J. Geophys. Res.*, 115, D11108, doi:10.1029/2009JD013094.

Quayle, R. G., D. R. Easterling, T. R. Karl, and P. Y. Hughes (1991), Effects of recent thermometer changes in the Cooperative Station Network. *Bull. Amer. Meteor. Soc.,* **72,** 1718–1723.

Peterson, T. C. (2006), Examination of potential biases in air temperature caused by poor station locations. *Bull. Amer. Meteor. Soc.,* **87,** 1073–1089.

Santer B.D., T.M.:. Wigley, C. Mears, F.J. Wentz, S.A. Klein, D.J. Seidel, K.E. Taylor, P.W. Thorne, M.F. Wehner, P.J. Gleckler, J.S. Boyle, W.D. Collins, K.W. Dixon, C. Doutriaux , M. Free, Q. Fu, J.E. Hansen, G.S. Jones, R. Ruedy, T.R. Karl, J.R. Lanzante, G.A. Meehl, V. Ramaswamy, G. Russell, G.A. Schmidt (2005), Amplification of surface temperature trends and variability in the tropical atmosphere. *Science*, 309, 1551-1556

Santer, B. D., J. E. Penner, P. W. Thorne (2006), How well can the observed vertical temperature changes be reconciled with out understanding of the causes of these changes? In *Temperature trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences.* T. R. Karl, S. J. Hassol, C. D. Miller and W. L. Murray, Editors. A report by the Climate Change Science Program and Subcommittee on Global Change Research, Washington, DC, 89-118.

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

Sherwood, S. C., H. A. Titchner, P. W. Thorne, M. P. McCarthy (2009), How do we tell which estimates of past climate change are correct? *Int. J. Climatol.*, **29**, 1520-1523.

Thorne, P. W., D. E. Parker, J. R. Christy, C. A. Mears (2005), Uncertainties in climate trends - Lessons from upper-air temperature records. *Bull. Amer. Meteor. Soc.,* 86, 1437-1442

Thorne, P. W., P. Brohan, H. A. Titchner, M. P. McCarthy, S. C. Sherwood, T. C. Peterson, L. Haimberger, D. E. Parker, S. F. B. Tett, B. D. Santer, D. R. Fereday, J. J. Kennedy (2011a), A quantification of the uncertainty in historical tropical tropospheric temperature trends from radiosondes. *J. Geophys. Res.*, 116, D12116, doi:10.1029/2010JD015487.

Thorne, P. W., K. M. Willett, R. J. Allan, S. Bojinski, J. R. Christy, N. Fox, S. Gilbert, I. Jolliffe, J. J. Kennedy, E. Kent, A. Klein Tank, J. Lawrimore, D. E. Parker, N. Rayner, A. Simmons, L. Song, P. A. Stott, B. Trewin (2011b), Creating a Comprehensive Surface Temperature Resource for 21st Century Climate Science. *Bull. Amer. Meteor. Soc*., in press.

Titchner, H. A., P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, D. E. Parker (2009), Critically assessing tropospheric temperature trends from radiosondes using realistic validation experiments. *J. Climate*, 22, 465-485

Trenberth, K.E., P.D. Jones, P. Ambenje, R. Bojariu, D. Easterling, A. Klein Tank, D. Parker, F. Rahimzadeh, J.A. Renwick, M. Rusticucci, B. Soden and P. Zhai, 2007: Observations: Surface and Atmospheric Climate Change. In: *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 235-336.

Tukey, J. W. (1977), Exploratory Data Analysis. Addison-Wesley, 688 pp.

Venema, V.K.C, and co-authors (2011): Benchmarking monthly homogenization algorithms. Submitted to *Climates of the Past*.

von Storch H, Zorita E, Jones JM, Dimitriev Y, Gonzalez-Rouco F, Tett SFB (2004), Reconstructing past climate from noisy data. *Science,* 306, 679–682.

Vose, R. S., C. N. Williams Jr., T. C. Peterson, T. R. Karl, and D. R. Easterling (2003), An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network, Geophys. Res. Lett., 30(20), 2046, doi:10.1029/2003GL018111.

Wang, X. L. (2003), Comments on ''Detection of undocumented changepoints: A revision of the two-phase regression model.'' *J. Climate*, 16, 3383–3385.

Washington, W.M., J.W. Weatherly, G.A. Meehl, A.J. Semtner Jr., T.W. Bettge, A.P. Craig, W.G. Strand Jr., J.M. Arblaster, V.B. Wayland, R. James, Y. Zhang, 2000: Parallel climate model (PCM) control and transient simulations. *Clim. Dyn.*, 16, 755—774.

Table 1 System tunable keywords varied in the creation of the 100 member ensemble. Brackets denote default values as described in *Menne and Williams* [2009]. Ensemble settings are available as part of the supplementary information in .csv format.

Algorithm Step	Key word name	Permitted values	Functional Description
Choosing neighbors	NEIGH_CLOSE	80, [100], 150, 200	Maximum number of neighboring series to consider
	NEIGH_CORR	[1diff], near, corr	Method used for ranking neighbors based on degree of similarity (1diff=calculate correlation using first differences; near=sort by distance only; corr=use anomalies to calculate correlation)
	CORR_LIM	[0.1], 0.5, 0.7	Minimum correlation coefficient with target to quality as a neighbor
	MIN_STNS	5, [7], 9	Minimum number neighbors with coincident data
	NEIGH_FINAL	20, [40], 60, 80	Final (maximum) number of neighbors per target station
Resolving breaks in difference series	SNHT_THRES	1, [5], 10	SNHT significance threshold (in percent)
	BIC_PENALTY	[BIC], AIC, none	Penalty function used to determine the form of the break (BIC=Bayesian Information Criterion; AIC=Akaike Information Criterion; none=no model fitting)
Identify the series causing the break	SHF_META	-1, 0, [1]	Toggle for metadata (-1=only adjust when break coincides with metadata; 0=run without use of metadata; 1=identify undocumented breaks and exploit metadata when availabe)
	AMPLOC_PCT	90, [92], 95	Confidence window table used to coalesce changepoints
	CONFIRM	[2],3,4,5	Number of target-neighbor difference series with coincident breaks required to implicate the target as the source of the break
Estimating the magnitude of the break	ADJ_MINLEN	[18], 24, 36, 48	Minimum length of data period (in months) that can be adjusted
	ADJ_MINPAIR	[2], 3, 4, 5	Minimum number of pairwise estimates of break size required to determine the size of adjustment
	ADJ_OUTLIER	0, [1]	Toggle to test and remove outliers using the Tukey outlier test
	ADJ_WINDOW	0, [24], 60, 120	Minimum number of months before and after a break in the difference series

			necessary to calculate breakpoint size
	ADJ_FILTER	bicf, [conf], both, none	Outlier filtering method for the pairwise break estimates
	ADJ_EST	Aavg, [Medi], Qavg	Method used to determine the adjustement factor from the multiple pairwise estimates
	NS_LOOP	0, [1]	Toggle to merge data segments when the breaksize is statistically insignificant (this loop increases the length of the homogeneous segments available to estimate other breakpoint sizes in data sparse periods)

Table 2. Look up table for gross characteristics of the set of analog worlds. Breaks are added in all cases as seasonally invariant deltas to all points prior to the assigned breakpoint. Breakpoint sizes and locations were allocated by random number generators seeded from system time at time of instigation.

Analog world	**Model**	**Forcings and period in model years**	**Break and metadata structure imparted**
Perfect data	MIROC 3.2 hires [K-1 model developers (2004)]	20th Century forcings, 1900-1999, run 1	No breaks, no metadata
Big breaks, good metadata	GFDL CM2.0 [*Delworth et al.,* 2006]	20th Century forcings, 1861-1960, run1	5 per station on average seeded randomly across the network and through time, with metadata (σ=1.5, avg=0)
Mixed break sizes, some clustering	Same as perfect data	Same as perfect data	70% of stations within 15 years starting 1930, with metadata (σ=0.7, avg=-0.2) 70% of stations within 30 years starting 1945, with metadata (σ=0.4, avg=-0.3) Average one break per station randomly seeded throughout period, with metadata (σ=0.35, avg=0) No metadata, more prevalent early in record, 4 per station on average (σ=0.3, avg=-0.1) 1.5 false metadata events per station, more prevalent later
Clustering and sign bias – c20c1	Same as perfect data	Same as perfect data	70% of stations within 7 years in 1980s, with metadata (σ=0.7, avg=0.35) 70% of stations within 30 years from 1945, with metadata (σ=0.4, avg=-0.2) Average one per station in latter half of record with metadata (σ=0.5, avg=0.8) Average of 2 breaks per station associated with metadata (σ=0.8, avg=0) No metadata, more prevalent early in record, 4 per station on average (σ=0.8, avg=0) Average 2 metadata events not associated with a break.

Clustering and sign bias – c20c2	CSIRO MK3.5 [*Gordon et al.*, 2002]	20th Century forcings, 1871-1970, run 1	Same breaks as "clustering and sign bias c20c1"
Clustering and sign bias - control	UKMO - HadGEM1 [*Johns et al.,* 2006]	No changes in external forcings, 2000-2099	Same breaks as "clustering and sign bias c20c1"
Clustering and sign bias - committed	NCAR CCSM3.0 [*Collins et al.,* 2006]	Stabilisation run 2000-2099	Same breaks as "clustering and sign bias c20c1"
Very many mainly small breaks	NCAR PCM [*Washington et al.*, 2000]	CO_2 1%/yr to $2xCO_2$, 0071-0170,	2 breaks on average per station seeded randomly throughout network and over time with metadata (σ=1, avg=0) 2 breaks on average per station but probability twice as prevalent later in record and sign biased, with metadata (σ=0.25, avg=-0.2) 2 breaks on average per station but probability twice as prevalent later in record, with metadata (σ=0.25, avg=0) 4 breaks per station unassociated with metadata, more prevalent early, slight sign bias (σ=0.2, avg=-0.075)

Table 3. Summary of maximum and minimum homogenized temperature trends (°C/decade) from the 101 member ensemble. Trends for the pre-homogenization input data are also shown.

	Maximum Temperature (raw input)	Minimum Temperature (raw input)	Maximum Temperature (input first corrected for TOB)	Minimum Temperature (input first corrected for TOB)
		1900-2010		
Lowest	0.0148	0.0027	0.0256	0.0142
Highest	0.0548	0.0852	0.0613	0.0989
Median	0.0393	0.0585	0.0504	0.0738
Average	0.0379	0.0568	0.0474	0.0704
Original input data	0.0008	0.0474	0.0212	0.0741
		1950-2010		
Lowest	0.0322	0.1082	0.0583	0.1294
Highest	0.1165	0.1622	0.1289	0.1988
Median	0.0823	0.1402	0.1058	0.1676
Average	0.0808	0.1395	0.1019	0.1681
Original input data	-0.0049	0.1163	0.0473	0.1755
		1979-2010		
Lowest	0.1538	0.1633	0.1970	0.1962
Highest	0.2962	0.2267	0.3070	0.2875
Median	0.2374	0.2130	0.2698	0.2523
Average	0.2327	0.2110	0.2635	0.2523
Original Input Data	0.1059	0.2241	0.1791	0.3081

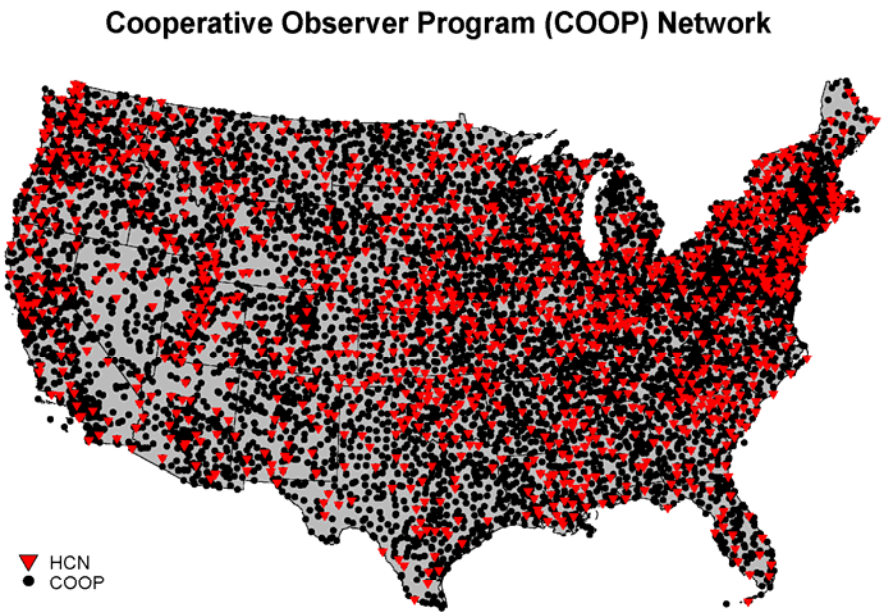

Figure 1. Distribution of COOP stations in the CONUS (black dots) and the U.S. HCN version 2 sites (red triangles).

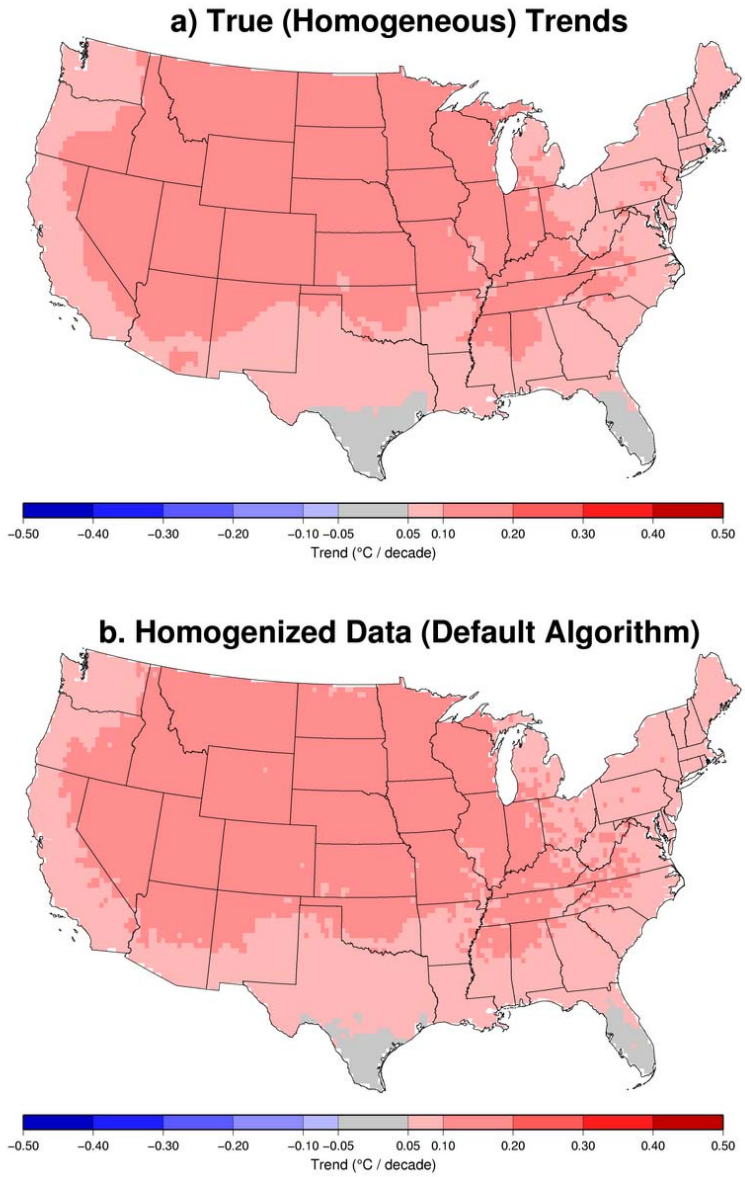

Figure 2. (a) gridded trends for the period 1900-1999 for the "Perfect data" raw input; (b) the homogenized version of the data produced by the default version of the pairwise homogenization algorithm.

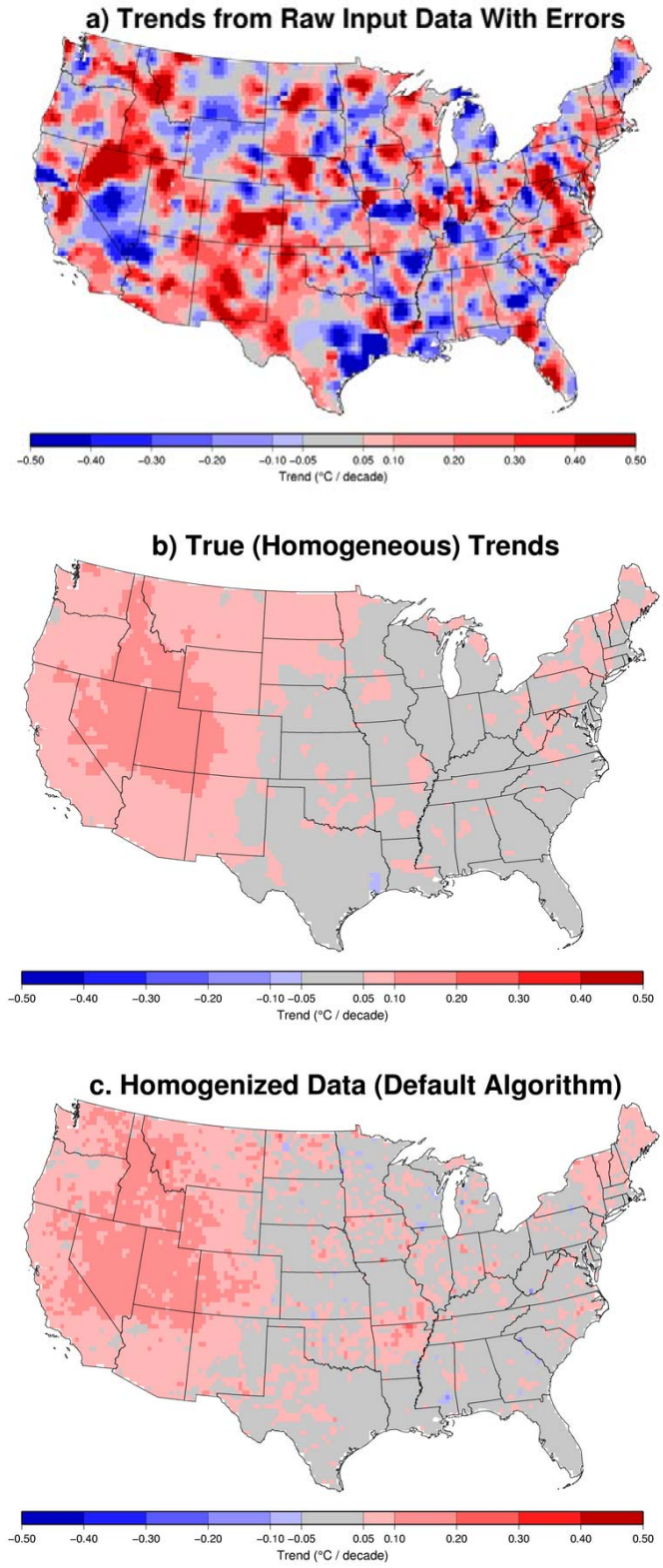

Figure 3. Gridded trends for the period 1900-1999 for the "Big breaks, good metadata" USHCN analog: (a) Raw (unadjusted) input data; (b) True (homogeneous) data; and, (c) data homogenized by the default version of the pairwise algorithm.

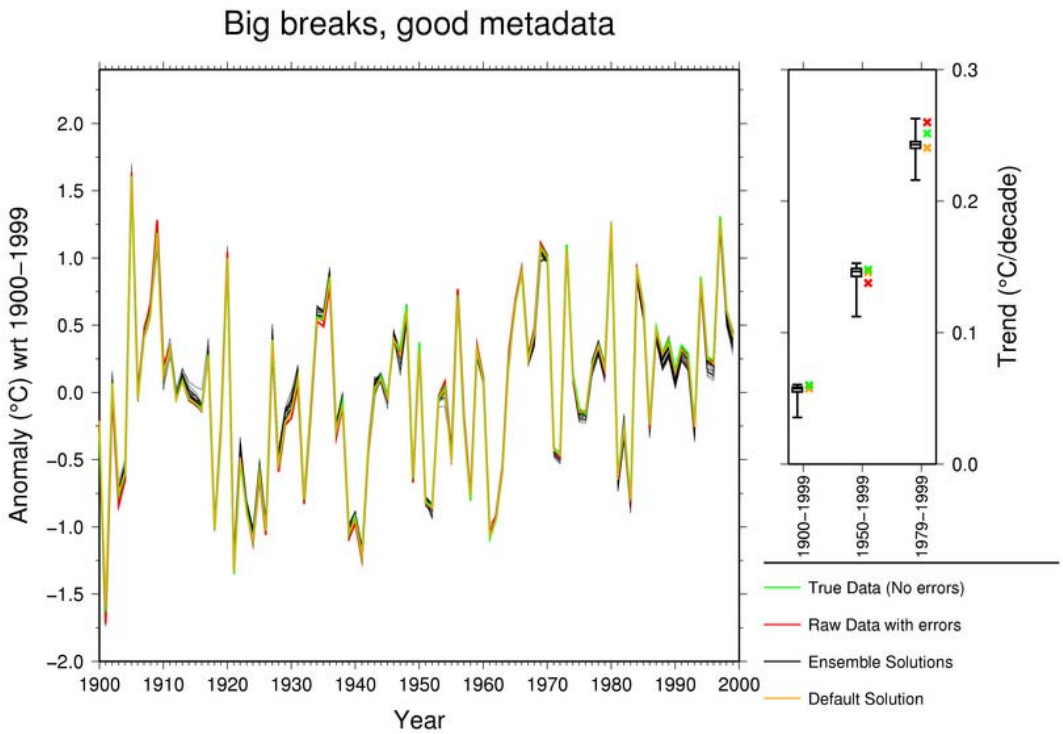

Figure 4. Annual average contiguous U.S. (CONUS) temperature series calculated using the USHCN monthly temperature series from the "Big breaks, good metadata" analog. Spatial averages are based on output from the 100 randomized versions of the pairwise algorithm (in black) as well as from the default version (in orange). CONUS averages for the non-homogenized (raw) input values with the seeded errors are shown in red. Averages based on the true data series without errors are shown in green. Box plots depicting the range of CONUS average trends for the three different summary periods used in *Trenberth et al.* [2007] produced by the 100 randomized versions of the pairwise homogenization algorithm are also shown along with the trends based on the true data, the raw input input data with errors and on the homogenized data produced by the default algorithm. Whiskers denote the full range, boxes the inter-quartile range and horizontal line within the box the median estimator for the 100 member ensemble.

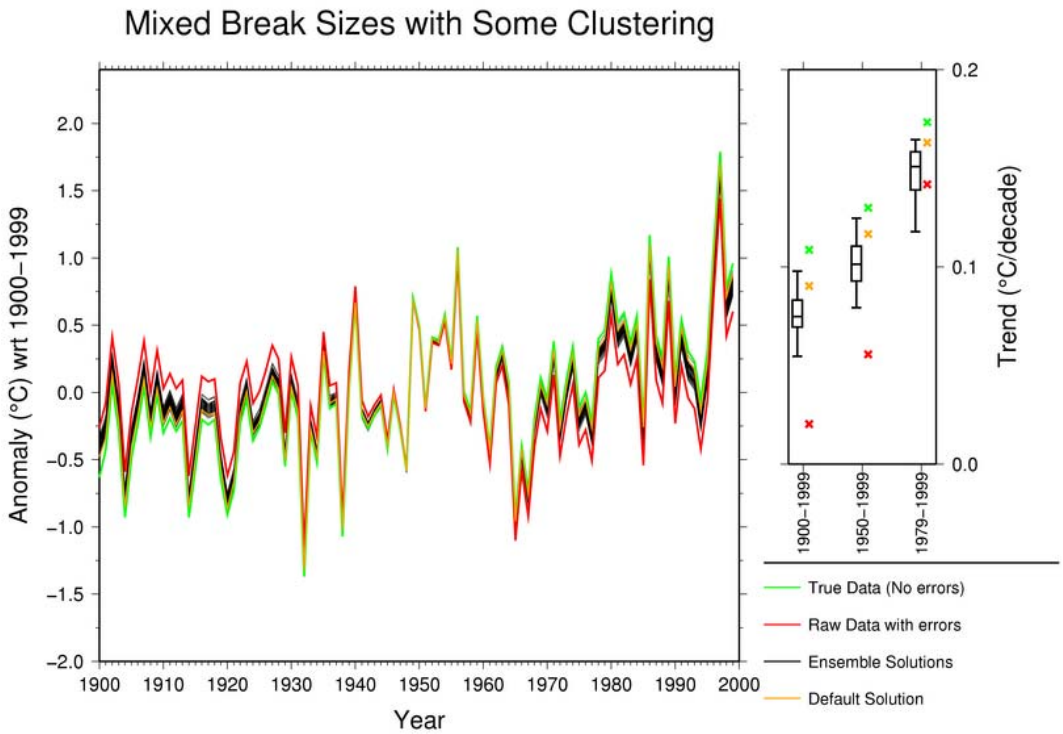

Figure 5. As in Fig. 4, except for the "Mixed Break Sizes with Some Clustering" analog.

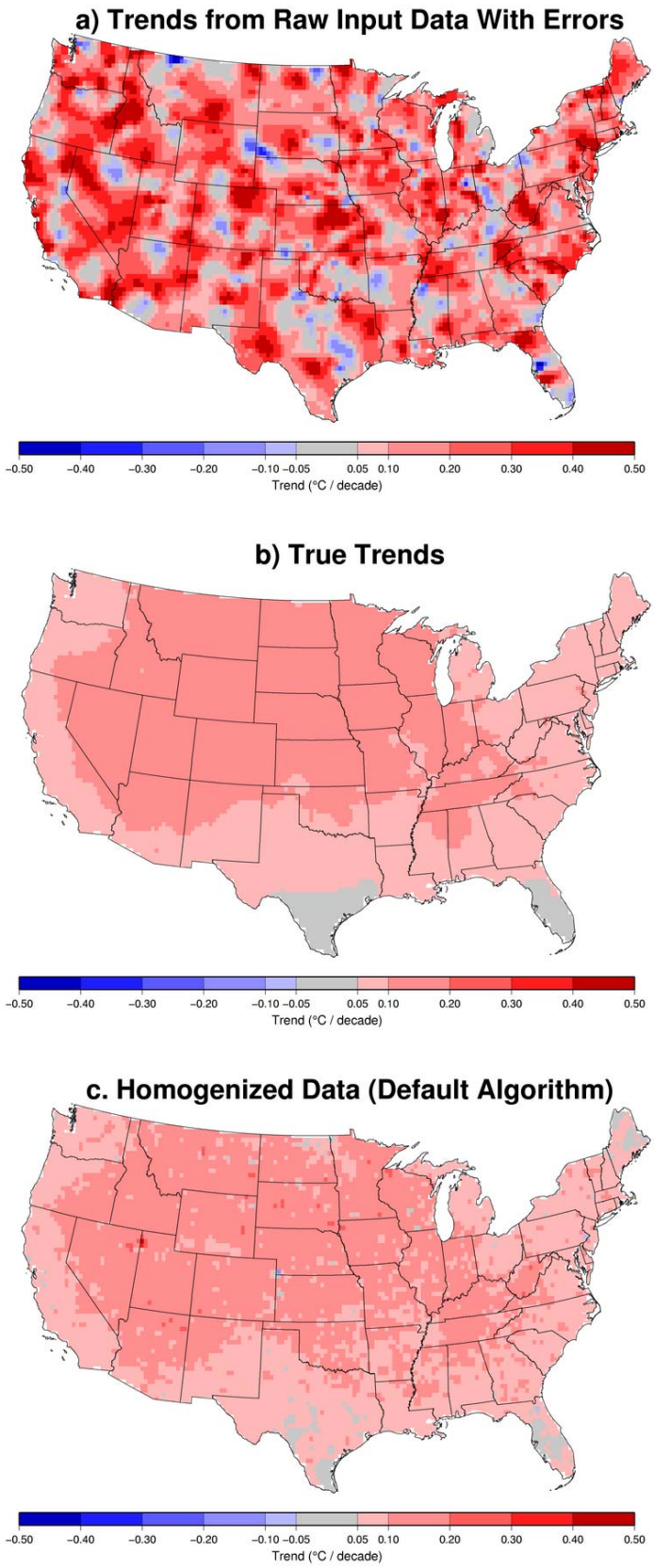

Figure 6. As in Fig. 3 except for the "Clustering and Sign Bias—C20C1" analog.

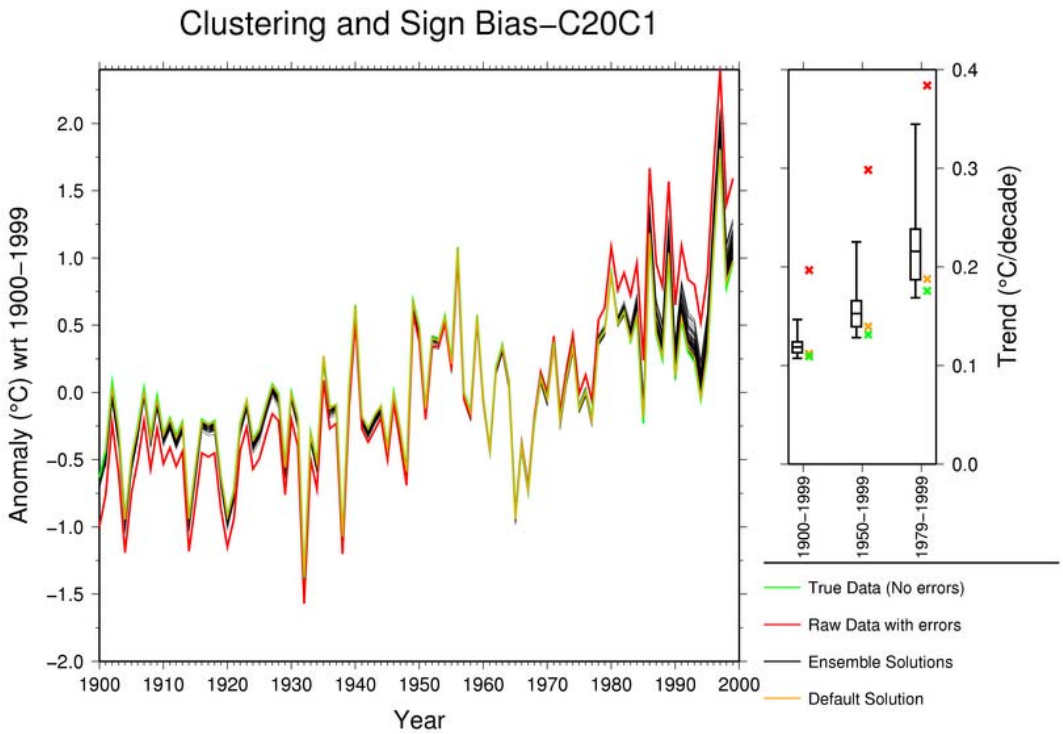

Figure 7. as in Fig. 4 except for the "Clustering and sign bias-c20c1" analog.

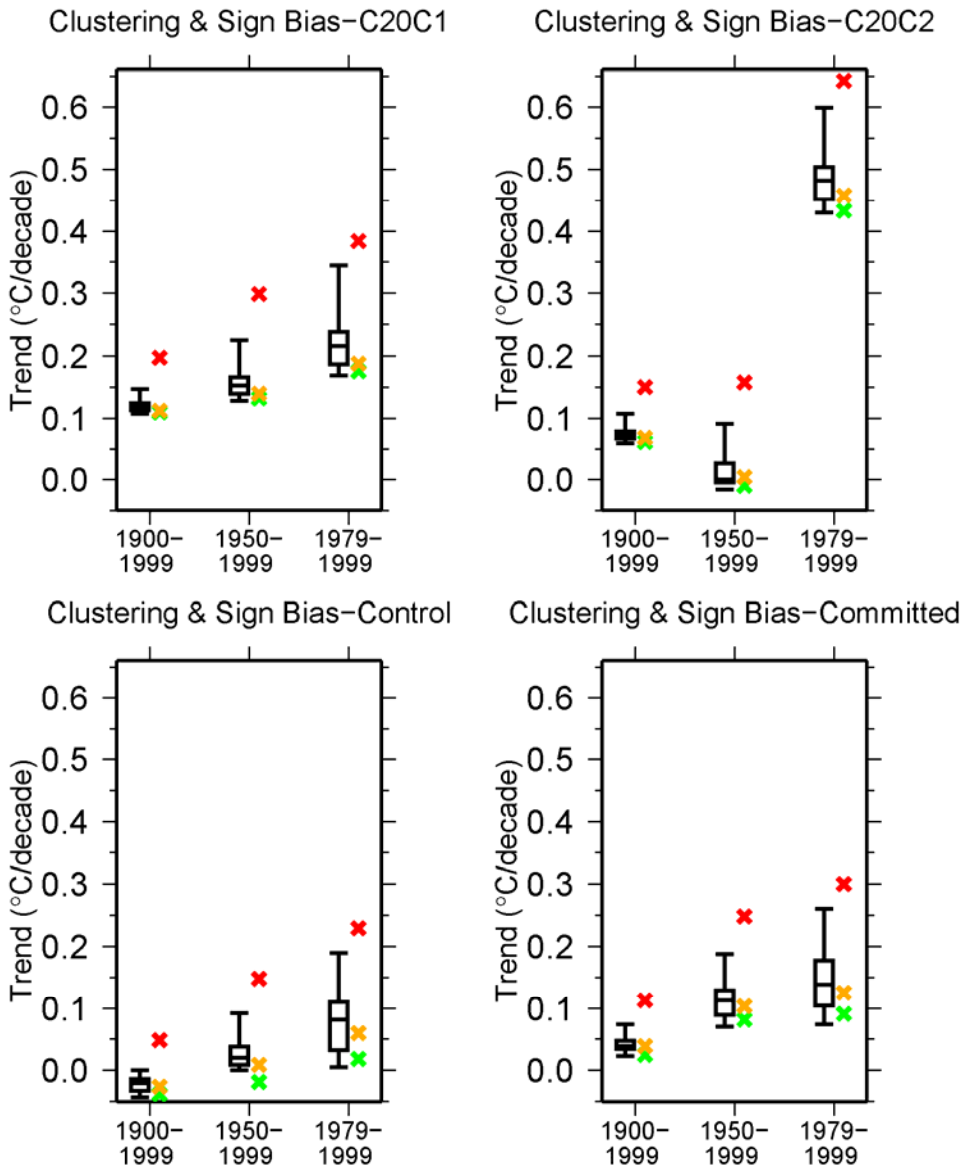

Figure 8. Box plots depicting the range of CONUS average trends for the three different summary periods used in *Trenberth et al.* [2007] produced by the 100 randomized versions of the pairwise homogenization algorithm. The magnitude of the CONUS average trends based on the raw input data are given by the red "X", the magnitude of the true (homogeneous) trends are given by the green "X". The magnitude of trends produced by the default version of the homogenization algorithm is shown by the yellow "X". Whiskers denote the full range, boxes the inter-quartile range and horizontal line within the box the median estimate for the 100 member ensemble.

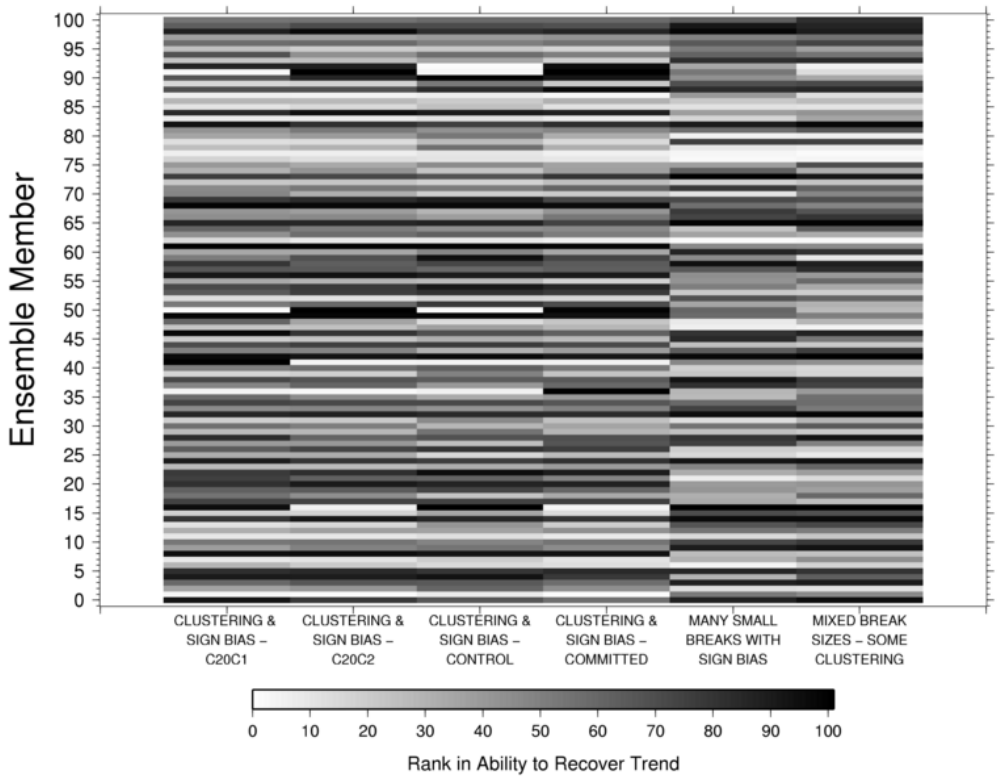

Figure 9. Ranking from 1 (worst) to 101 (best) of the degree to which each version of the algorithm was able to recover the true period-of-record CONUS trend in analog series seeded with errors that have a sign bias. The default version of the algorithm is denoted as ensemble member "0" and members "1-100" are the randomized versions. Dark shades denote high rankings and indicate versions of the algorithm that were the most successful at recovering the CONUS average trend for each particular analog world; light shadings denote low rankings where versions were the least successful in recovering the true trend.

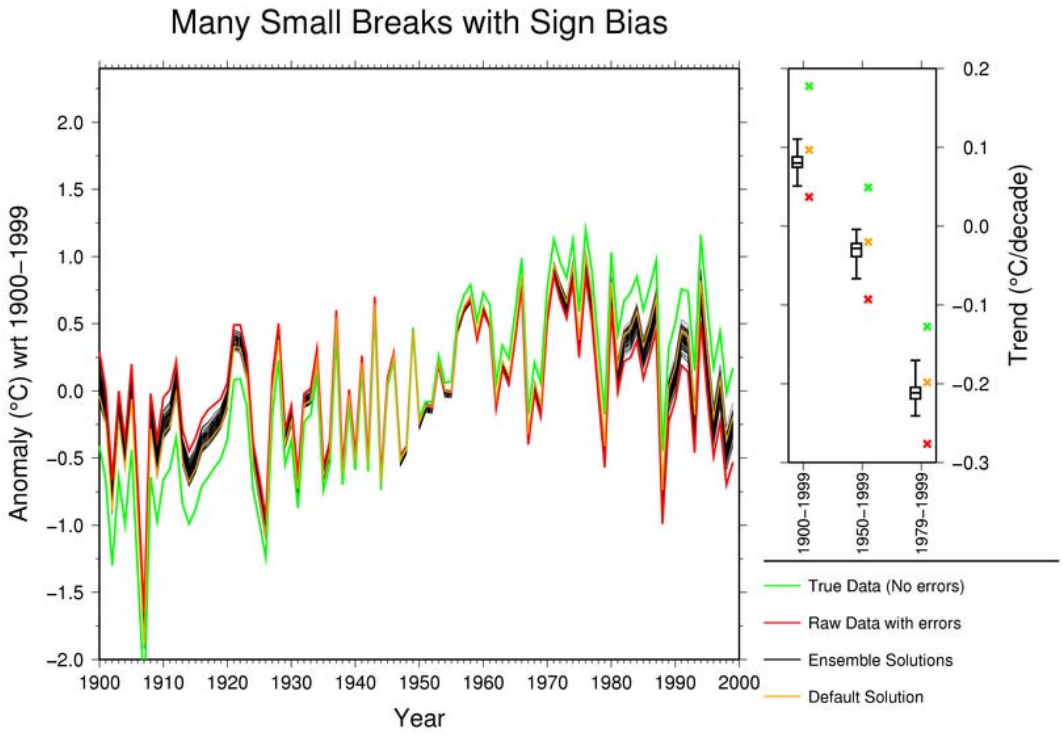

Figure 10. As in Fig. 2, except for the "very many mainly small breaks" analog.

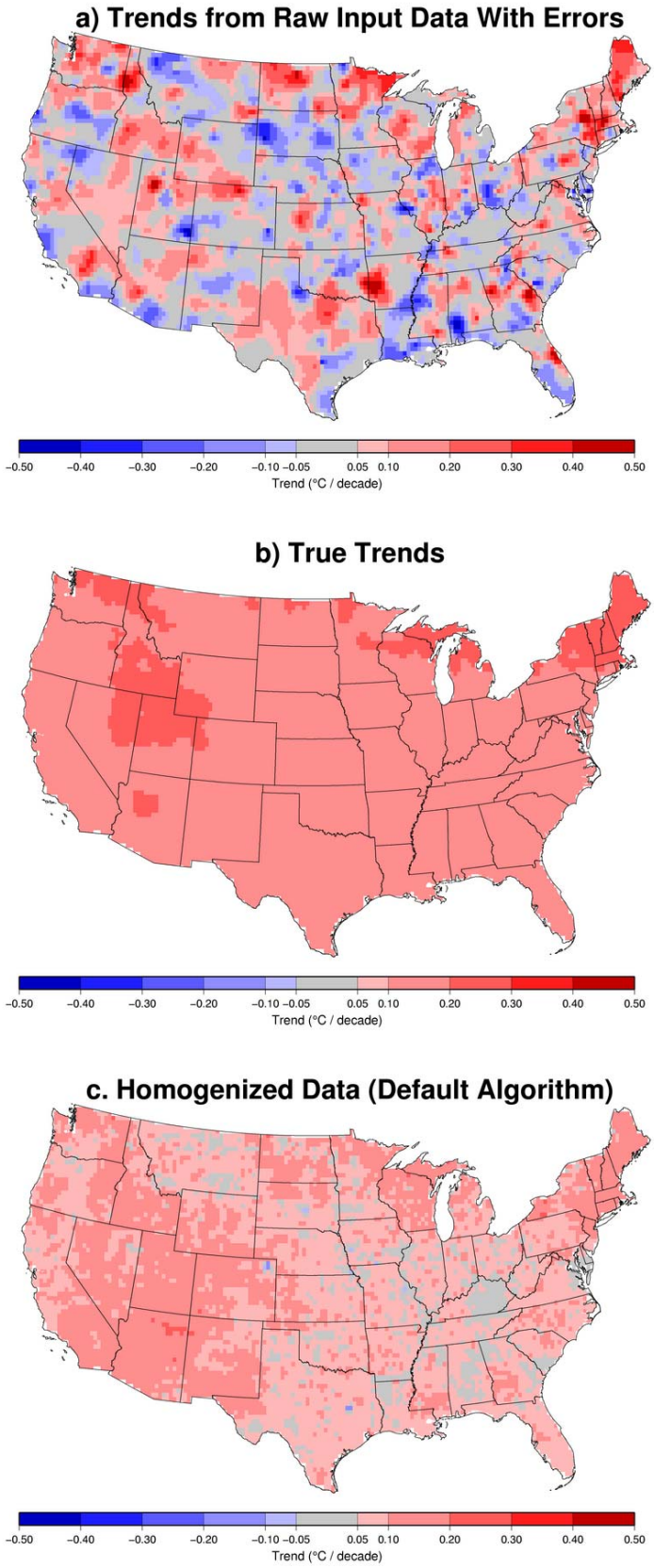

Figure 11. As in Fig. 3 for the "very many, mainly small breaks" analog.

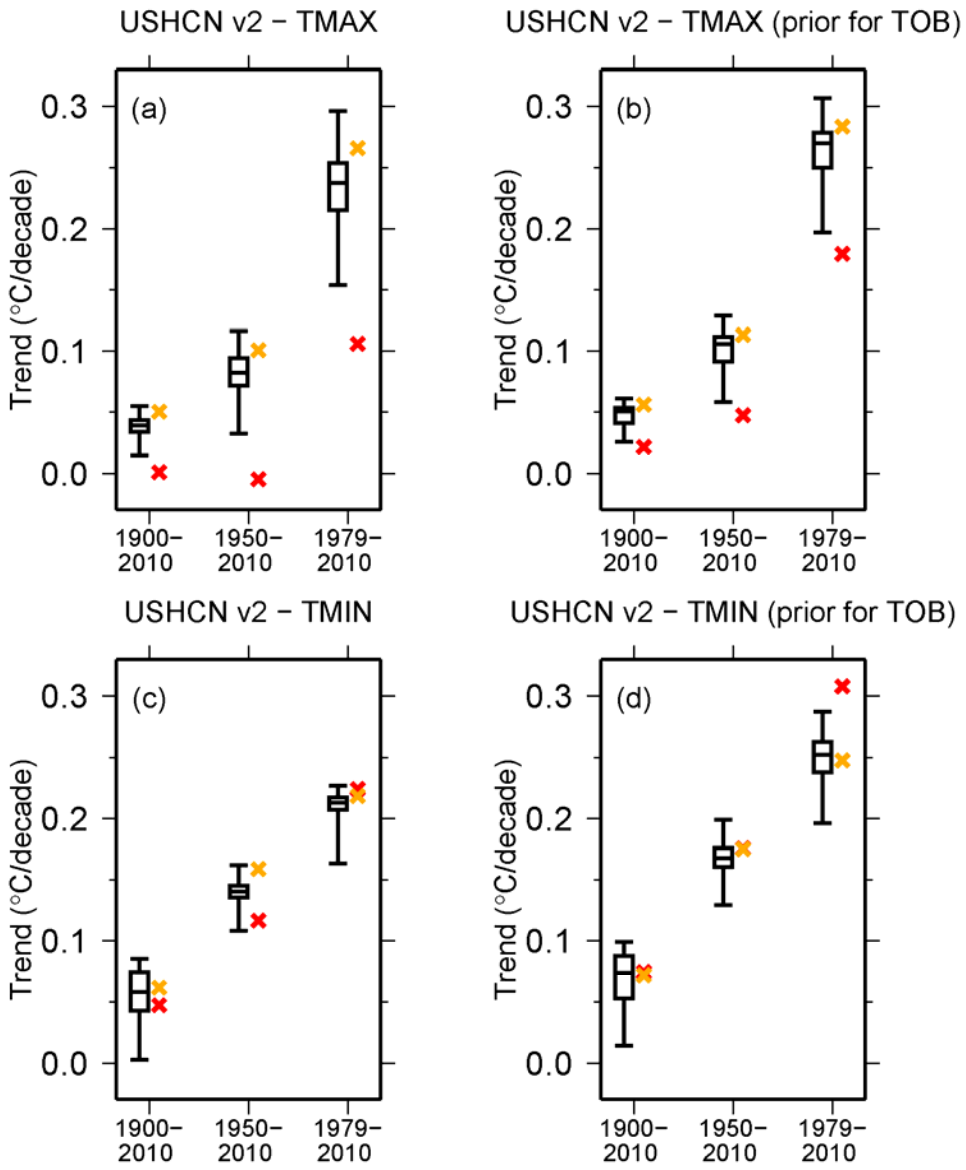

Figure 12. As in Fig. 7, except for the observed USHCN version 2 monthly temperature series using the following as input to the adjustment algorithm: (a) raw monthly maximum temperatures; (b) time of observation adjusted maximum temperatures; (c) raw minimum monthly temperatures; (d) time of observation adjusted minimum temperatures.

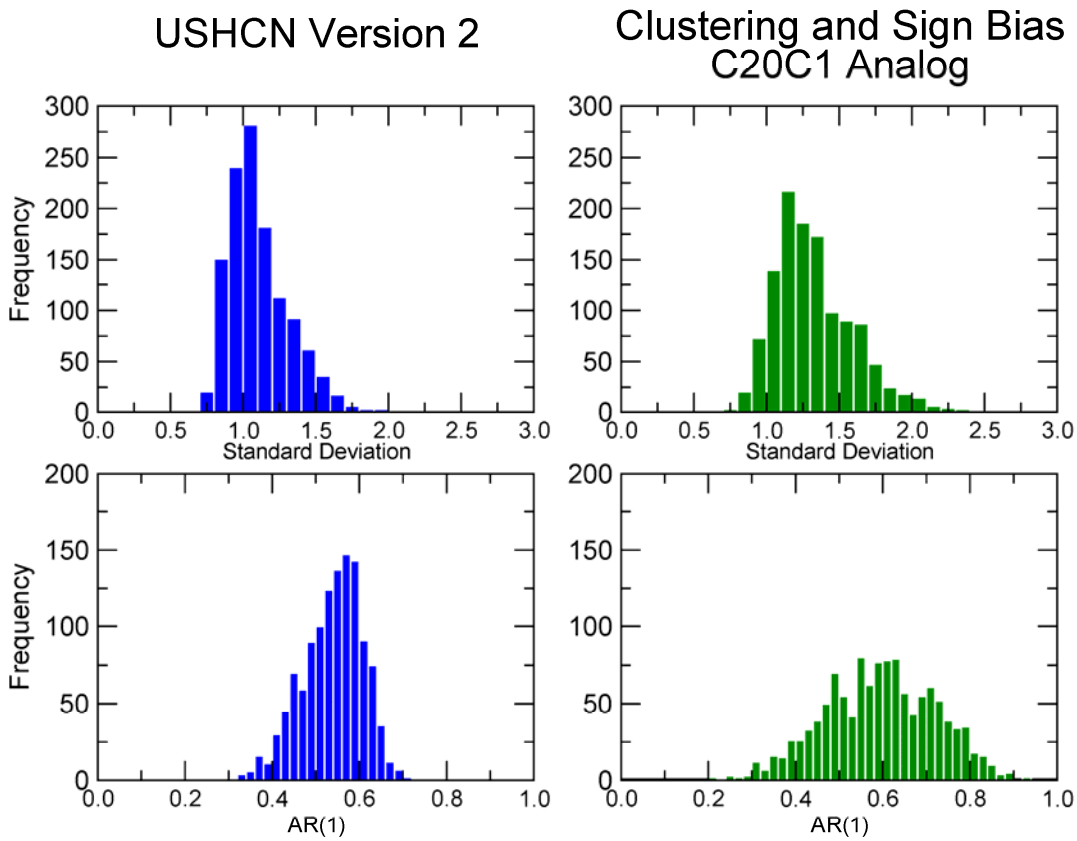

Figure S1. Distribution of the standard deviation and AR(1) autocorrelation in the monthly mean temperature differences at USHCN temperature stations and their 40 most highly correlated Cooperative Observer Network "neighbors". Differences were calculated using the USHCN version 2 homogenized data (blue) and the "Clustering and Sign Bias—C20C1" analog data (green).

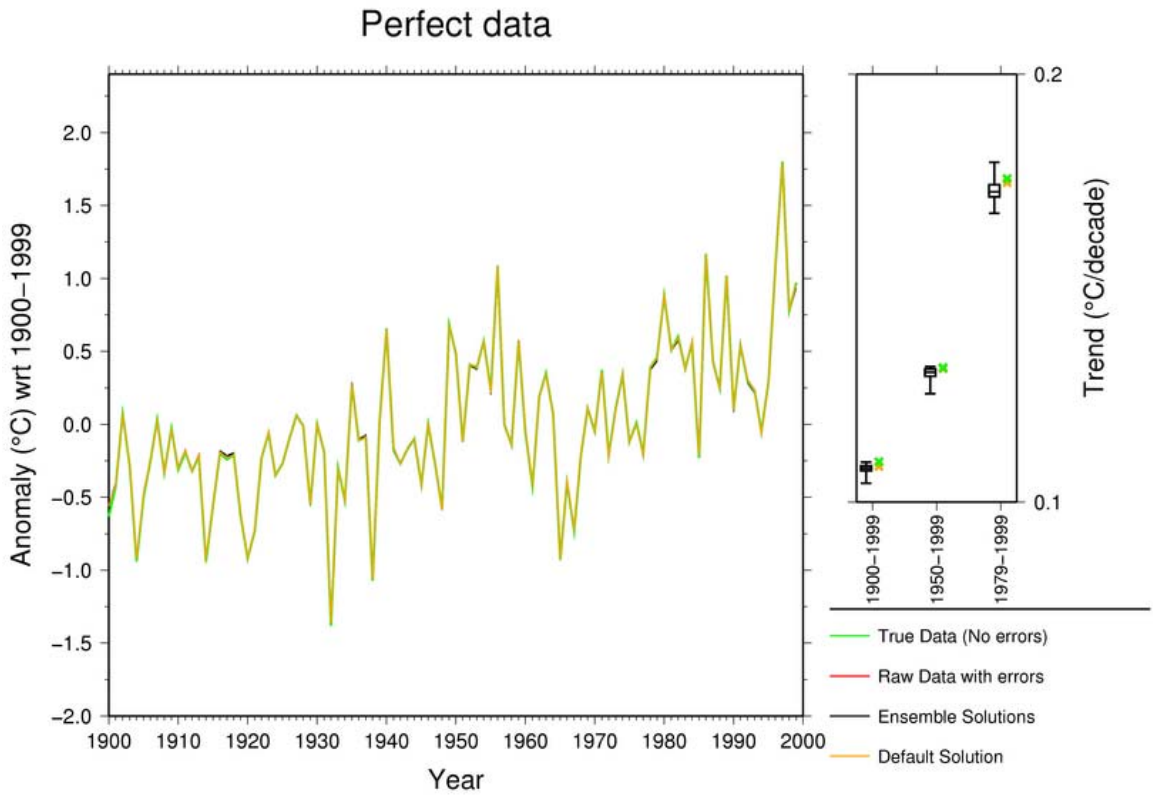

Figure S2. As in Fig. 4, except for the “Perfect data” analog.

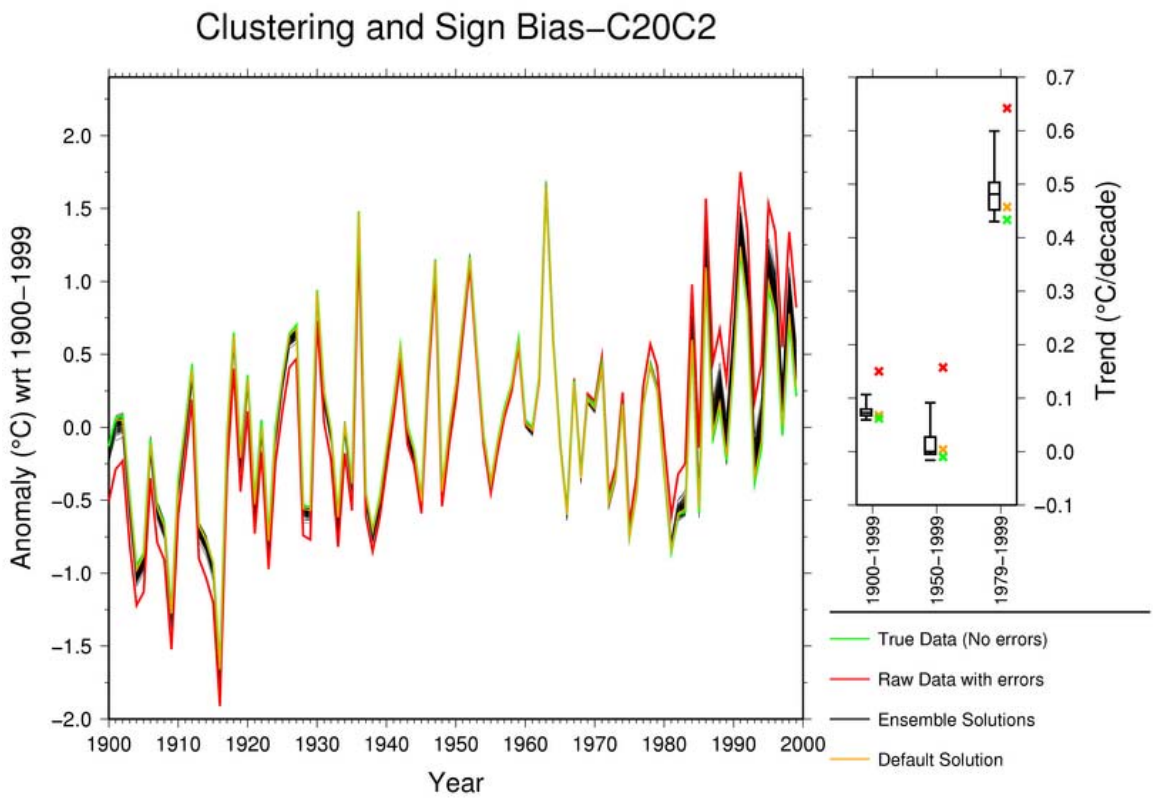

Figure S3. As in Fig. 4, except for the “Clustering and sign bias-C20C2” analog.

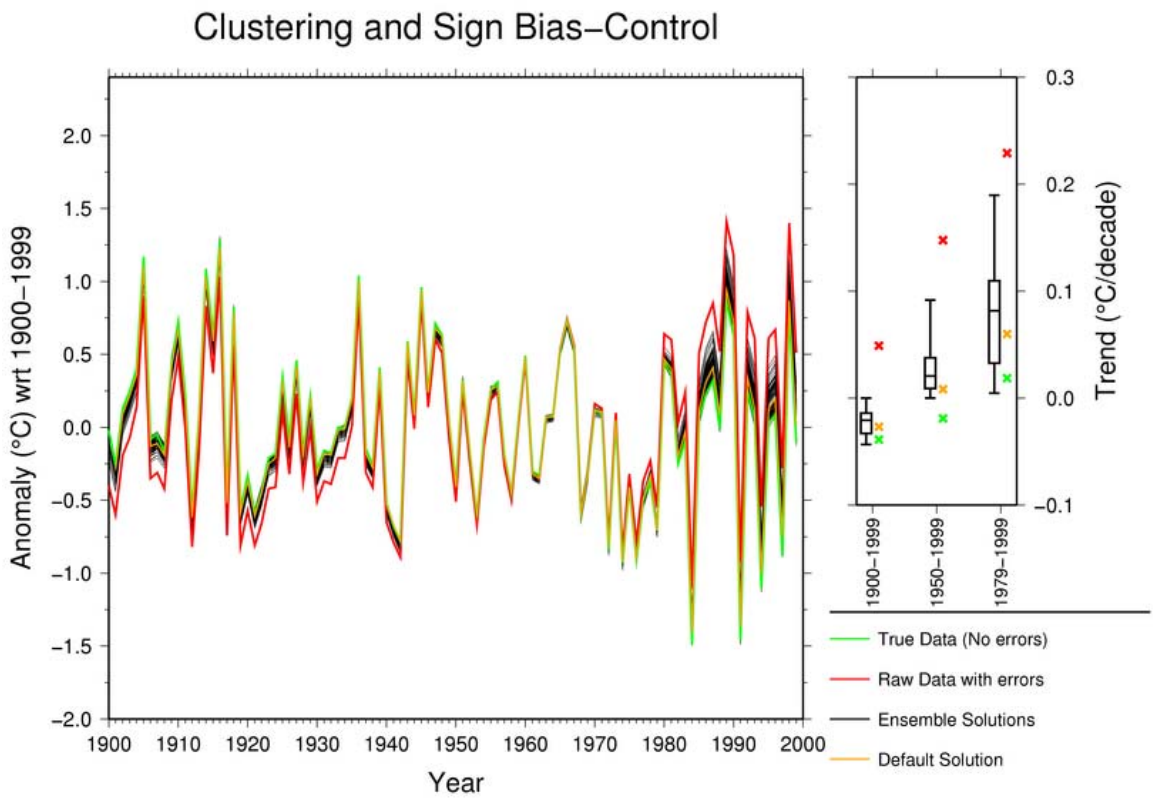

Figure S4. As in Fig. 4, except for the "Clustering and Sign Bias—control" analog.

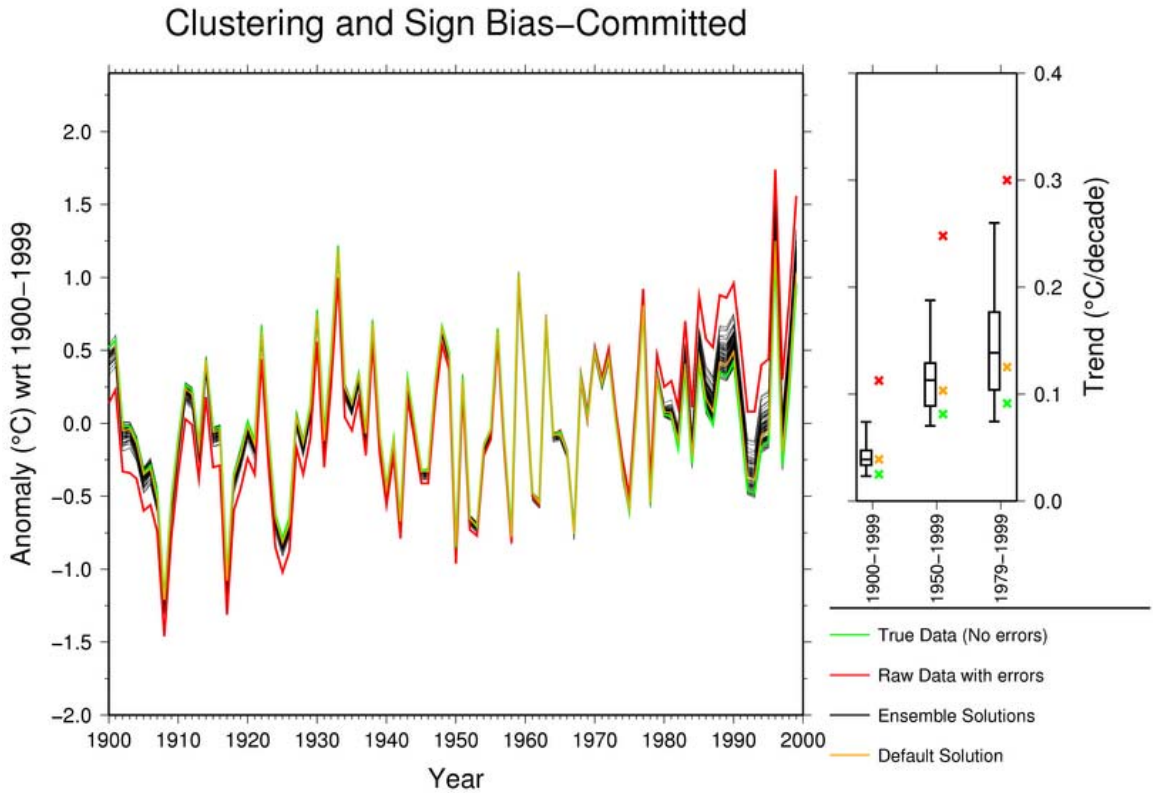

Figure S5. As in Fig. 4, except for the "Clustering and Sign Bias—committed" analog.

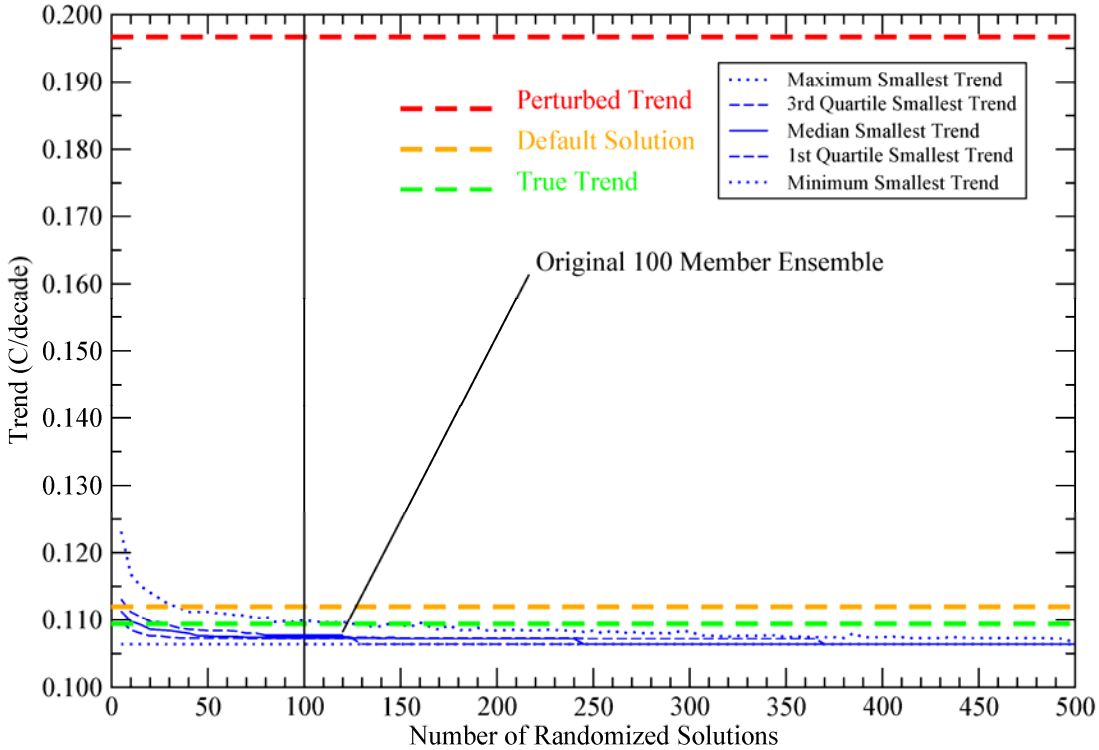

Figure S6. Distribution of the minimum CONUS average trend (1900-1999) for various numbers of ensemble members. The distribution represents the minimum trend produced by varying numbers of the randomized versions of pairwise homogenization algorithm (drawn from a total population of 500). The minimum trend for the 100-member solution shown in Fig. 7 is indicated as the "Original 100 Member Ensemble" value.

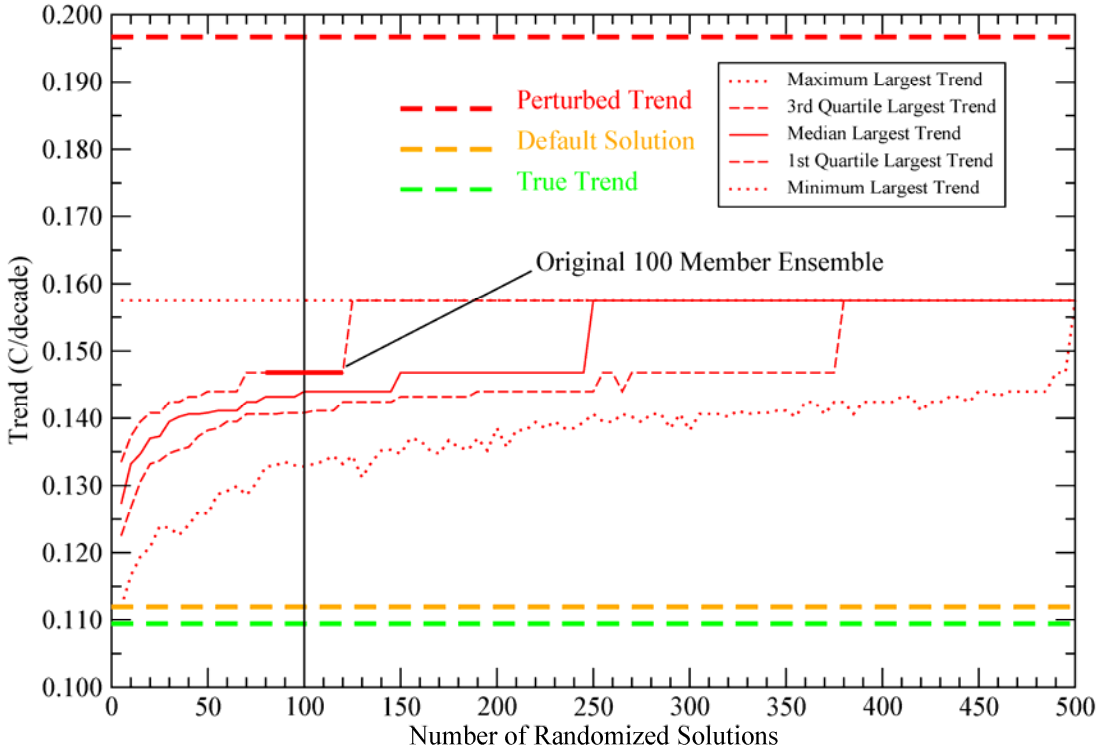

Figure S7. Distribution of the maximum CONUS average trend (1900-1999) for various numbers of ensemble members. The distribution represents the maximum trend produced by varying numbers of the randomized versions of pairwise homogenization algorithm (drawn from a total population of 500). The maximum trend for the 100-member solution shown in Fig. 7 is indicated as the "Original 100 Member Ensemble" value.

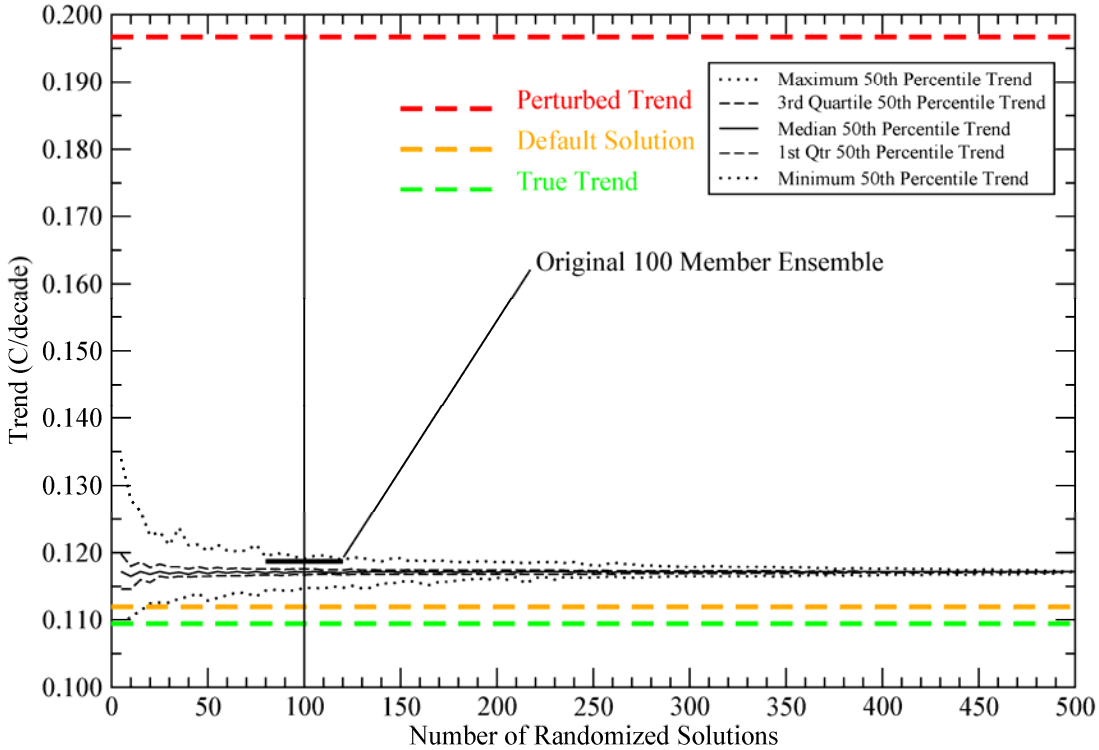

Figure S8. Distribution of the median CONUS average trend (1900-1999) for various numbers of ensemble members. The distribution represents the median trend produced by varying numbers of the randomized versions of pairwise homogenization algorithm (drawn from a total population of 500). The median trend for the 100-member solution shown in Fig. 7 is indicated as the "Original 100 Member Ensemble" value.

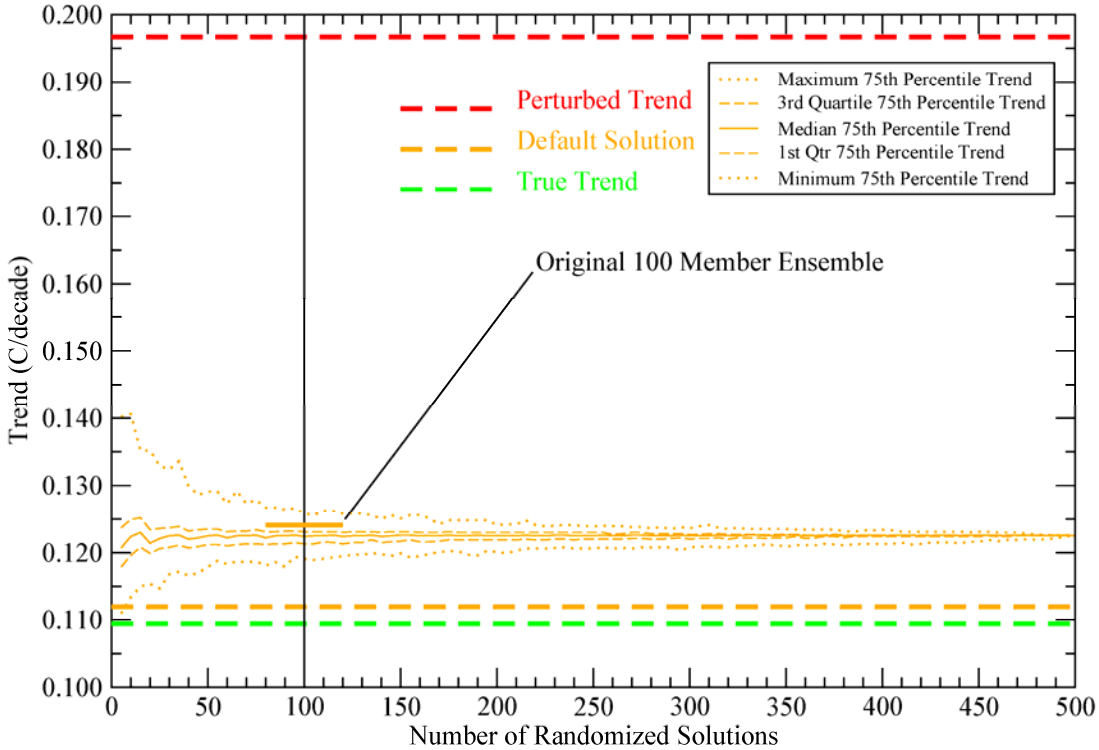

Figure S9. Distribution of the 75th percentile of CONUS average trend (1900-1999) for various numbers of ensemble members. The distribution represents the 75th percentile of the trend produced by varying numbers of the randomized versions of pairwise homogenization algorithm (drawn from a total population of 500). The 75th percentile trend for the 100-member solution shown in Fig. 7 is indicated as the "Original 100 Member Ensemble" value.

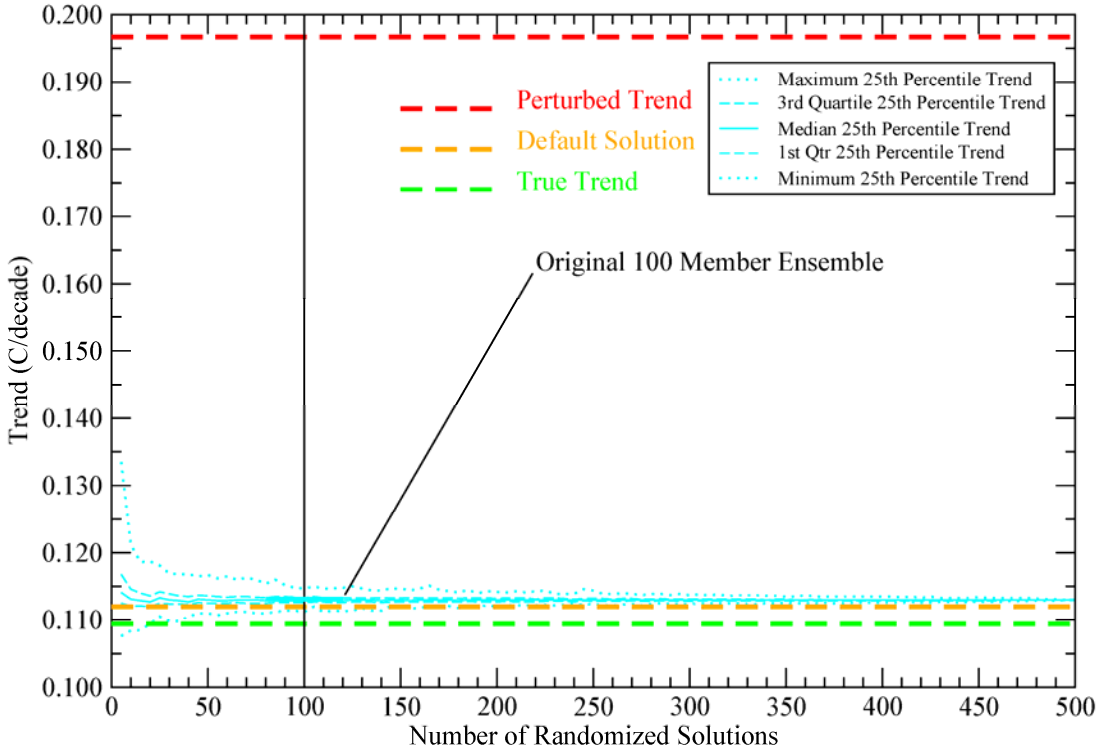

Figure S10. Distribution of the 25th percentile of CONUS average trend (1900-1999) for various numbers of ensemble members. The distribution represents the 25th percentile of the trend produced by varying numbers of the randomized versions of pairwise homogenization algorithm (drawn from a total population of 500). The 25th percentile trend for the 100-member solution shown in Fig. 7 is indicated as the "Original 100 Member Ensemble" value.

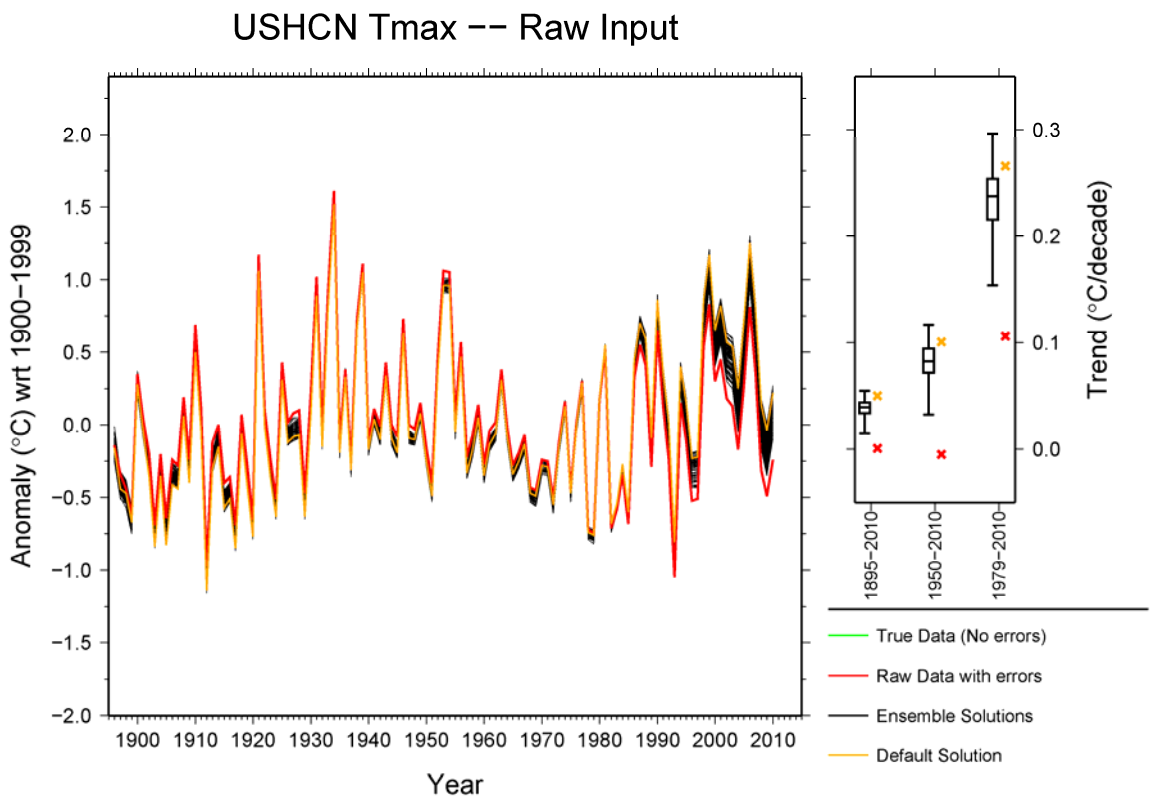

Figure S11. As in Fig. 4, except for the real world U.S. HCN mean monthly maximum temperatures (raw input).

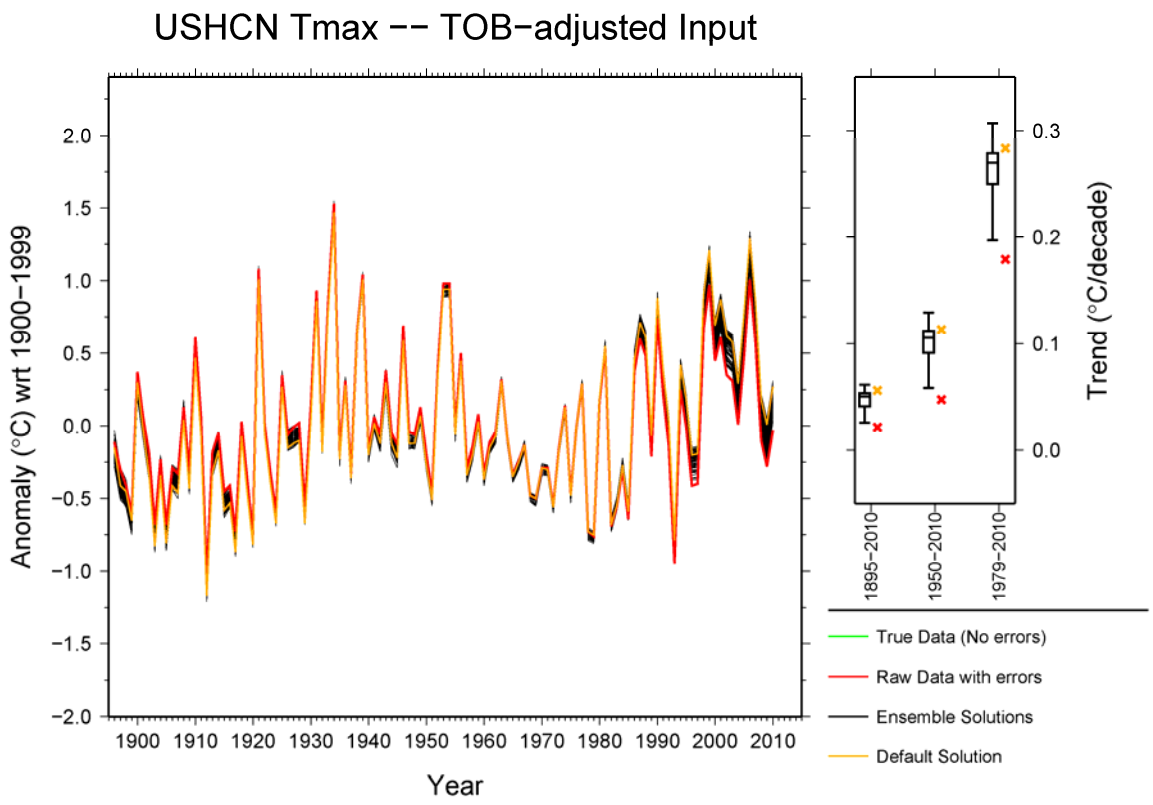

Figure S12. As in Fig. 4, except for the real world U.S. HCN mean monthly maximum temperatures (TOB-adjusted input).

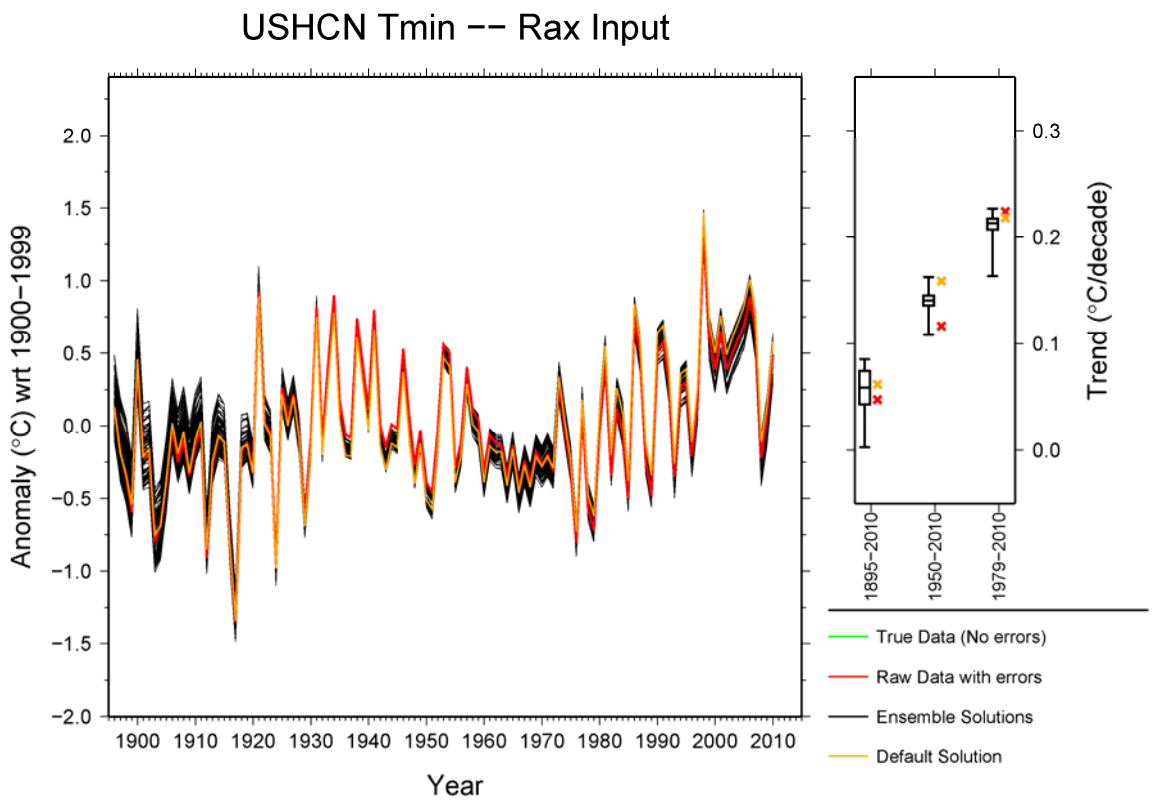

Figure S13. As in Fig. 4, except for the real world U.S. HCN mean monthly minimum temperatures (raw input).

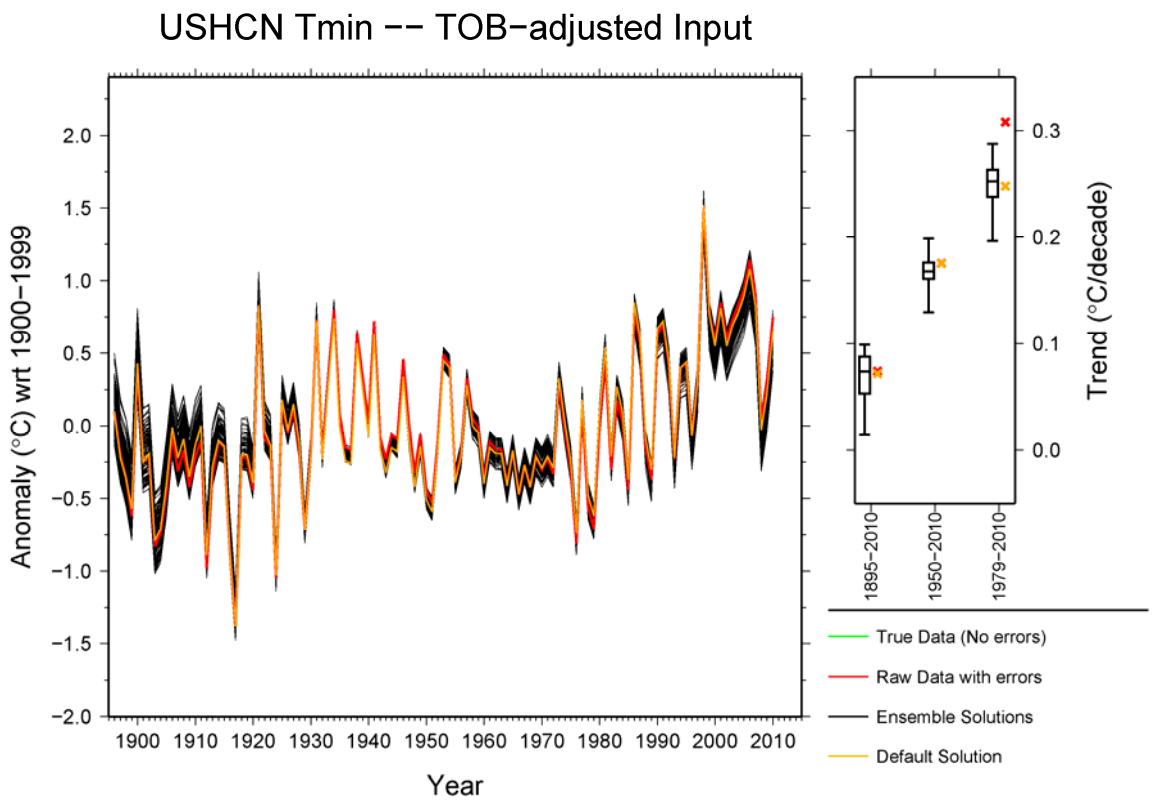

Figure S14. As in Fig. 4, except for the real world U.S. HCN mean monthly minimum temperatures (TOB-adjusted input).