Simultaneous Treatment of Random and Systematic Errors in the Historical Radiosonde Temperature Archive

Joshua M. Browning[1] and Amanda S. Hering[1]

August 27, 2014

Abstract

The historical radiosonde temperatures, and indeed any large and lengthy observational dataset, must be quality controlled before it can be used properly. Most research on quality control for such data focuses on the identification and removal of either systematic errors (homogenization) or random errors without considering an optimal process for treatment of both. Additionally, little has been done to evaluate homogenization methods applied to sub-daily data, and no research exists on using robust estimators in homogenization procedures. In this paper, we simulate realistic radiosonde temperature data and contaminate it with both systematic and random errors. We then evaluate (1) the performance of several homogenization algorithms and (2) the sequence in which the random and systematic errors are identified and corrected. In our simulations we find that the Standard Normal Homogeneity Test (SNHT) performs better than the robust counterpart that we introduce, and it is better than several other more modern alternatives. Moreover, we find that systematic errors present in the data lead to poorer performance of random error removal algorithms, but the presence of random errors in the data are not as detrimental to homogenization algorithms.

Some keywords: Outlier Detection; Change Point Detection; Homogenization; Temperature Radiosonde Data

Short title: Simultaneous Random and Systematic Error Detection

[1]Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO 80401, USA. 303.384.2462, E-mail: {jbrownin, ahering}@mines.edu

1 Introduction

Any large dataset whose observations reach far back in time may require treatment for both systematic and random errors. Datasets such as the International Surface Temperature Initiative (ISTI) global land surface databank [17] with over 32,000 stations, and the Integrated Global Radiosonde Archive (IGRA) housed at the National Climatic Data Center (NCDC) [4] are examples of such large datasets. Systematic errors can occur when the station location changes; the area surrounding the station becomes urbanized; or the instrumentation is changed. Random errors can occur due to faulty data transmission; sporadic instrumentation problems; keystroke entries; or errors in data management. **[Cite Figure 1 here with examples of systematic and random errors.]** It is important to treat both sources of errors in large historical datasets as robustly and automatically as possible. In most published research, methods for handling systematic and random errors are treated separately, and opinions among climate and weather scientists differ in terms of which type of error should be handled first. The purpose of this study is to shed light on the order in which systematic and random error methods should be applied to such large datasets when considering both sources of error simultaneously. In addition, robust estimators in homogenization algorithms when random errors are present have not yet been considered, so these will be proposed and investigated as well.

In this paper, we will focus on the Upper Air Database (UADB) housed at the National Center for Atmospheric Research (NCAR). This archive differs from the IGRA archive in that it contains some different stations, and many of the records are older. Since the radiosonde data are the only measured values of the upper atmosphere, it is a very important resource for studies in climate change [5, 6] and for use as an input to global reanalysis datasets [9, 10]. Currently over 2,000 station locations exist, and atmospheric variables are collected at standard pressure levels as the radiosonde rises through the atmosphere. In large datasets such as these, the error detection methods must be automated since the archives are so large that making visual inspections of every station is not feasible.

Many methods have been developed to homogenize radiosonde data, but most are not tested on simulated data [6, **?**, 7, 13, 14, 19]. However, a study was recently conducted by the European

Cooperation in Science and Technology to compare many different homogenization methods [19]. A large, realistic dataset with known change points was simulated, and then researchers were asked to test their homogenization algorithm on the dataset. As the researchers did not have knowledge of the true change point locations, this test provided a way to compare the performance of these methods.

However, most homogenization techniques are designed for monthly or annual time series. Some of these techniques rely on optimizing an objective function over all possible change point configurations [18, 12, 16, 15] but some are too computationally expensive for daily data. Additionally, some methods may only locate proposed change points and not correct for the difference in means, which is necessary for homogenization. In this paper, we compare the Standard Normal Homogeneity Test (SNHT) [1], the PELT algorithm [12], binary segmentation [18], and we propose a robust version of the SNHT.

Automated random error detection methods have not been investigated as thoroughly. Recently in [2], several random error detection methods were proposed and evaluated via simulated datasets. The authors found that the optimal error detection algorithm required two steps: first scanning for observations that were too many standard deviations from the global mean and secondly scanning for observations that were too many standard deviations away from their local mean. Robust estimators of mean and standard deviation were used in both cases to mitigate the influence of errors.

However, to our knowledge, no research has been done to date describing which method should be applied first to a dataset containing both types of errors. While it may seem intuitive that homogenization should occur before random error detection, this hypothesis has not been tested formally. We do a simulation study in which data is contaminated with both known random errors and with change points so that we can evaluate the performance of and the sequence in which different quality control algorithms are applied.

We shall henceforth refer to the choice of performing random error detection or systematic error detection first as the "sequence of the quality control methods" or simply the "sequencing." We study the influence of both alternatives via simulated data where we know the truth and

contaminate it with errors and change points. In Section 2, we discuss the details of our data simulation and contamination. Section 3 describes the homogenization algorithms we test and their comparative performance on the simulated, contaminated data. Section 4 describes the random error detection algorithm along with the results of the sequencing study. Finally, we conclude in Section 5.

2 Simulation Method

Observational data cannot be used to evaluate the performance of quality control (QC) and homogenization methods directly since we cannot know exactly where true change points and errors occur. Therefore, a rigorous simulation study is developed in order to accurately compare methods and their sequence. Evaluation of methodology via simulation is commonplace in the statistics literature, but this approach combines actual data with the simulation method. In order for this simulation study to validate methods for radiosonde data, it is crucial that we simulate data that is similar in structure to true radiosonde data.

2.1 Modeling Radiosonde Data

In order to capture seasonal and hourly trends, we fit a Generalized Additive Model (GAM) to the radiosonde temperature data. [**At what location and pressure level?**] GAMs are flexible, non-parametric models that allow the response variable to be a linear combination of smoothed functions of the input variables [2]. In our case, we model temperature (for a fixed location and pressure level) to be a function of hour of day, day of year, and year. We model the annual trend with a linear term to capture overall increases or decreases in the series. Thus, the model we fit is

$$t_i = s_1(h_i) + s_2(d_i) + \beta y_i + \epsilon_i, \tag{1}$$

where t_i is the temperature at a fixed station; h_i, d_i and y_i are the hour, day, and year of the i-th observation, respectively; β is the estimated coefficient for year; and $s_1(\cdot)$ and $s_2(\cdot)$ are cubic regression splines.

Typically the error term in Equation (1) would be modeled as normal with some unknown variance, but the distribution of the error terms could be skewed or have heavier tails than a normal distribution. Thus, we use a skew-t distribution for the errors of this model, which has 4 parameters, μ, σ, α, ν, each of which is related to one of the first four moments of the distribution. [**Needs an Azzalini citation.**] Thus, the distribution is very flexible and can model skewed and heavy-tailed data.

Additionally, we expect there to be temporal correlation in the error terms. However, since we have already included hourly and seasonal terms in the model, we expect most of this auto-correlation to be explained, so an AR(1) time series model should be sufficient to account for the remaining structure in the residuals. This model assumes that each error term has some fixed correlation with the error one time step in the past, and thus can be estimated by simply computing the correlation between t_i and t_{i+1} when the observations are equally spaced.

However, for radiosonde data, observations are not equally spaced in time. Launches are scheduled globally at 0 and 12 UTC, but especially in the historic record, many deviations from this pattern are observed. Most observations are within an hour or two of the scheduled launches, but in some instances, no launches occur on a given day, and on others, more than two radiosondes are launched. Thus, to estimate the lag-h autocorrelation, we must use use only those observations that are h time steps apart:

$$\phi(h) = \sum_{(\epsilon_i, \epsilon_j) \in \mathcal{P}_h} \frac{(\epsilon_i - \bar{\epsilon_i})(\epsilon_j - \bar{\epsilon_j})}{\sqrt{s_{\epsilon_i} s_{\epsilon_j}}}, \tag{2}$$

where $\phi(h)$ is the autocorrelation at time lag h; $\mathcal{P}_h$ is the set of all pairs of observations that are h units apart (and within some threshold); and ϵ_i is the error from Equation (1). For an AR(1) model, we need only estimate $\phi(h)$ at $h = 1$ day, and we use a window of 5% of 1 day, or 1.2 hours. [**Note to self: When we rerun simulations, use ϕ from a 12 hour estimate.**]

2.2 Data Simulation

We choose a fixed time period and assume that two observations for each day occur within that time period: one in the morning and one in the evening. The time of each morning (evening) observation is simulated by sampling a time from the morning (evening) subset of the observed data. This process is done to ensure that variability in the simulated hour of observation is comparable with that of the observed data.

Next, we use the GAM model fit based on Equation (1) to determine the expected value of temperature at the simulated time. To simulate the noise in this observation, we randomly draw values e_i from a skew-t distribution with parameters as fit previously. From these e_i, we construct the AR(1) model with

$$\epsilon_i = \hat{\phi}(1)^{\Delta_{i-1}}\epsilon_{i-1} + e_i,$$

where ϵ_i is the simulated noise in the model at time i, and Δ_{i-1} is the time difference, in days **[hours?]**, between the $(i-1)$th and the ith observation. As the kth term in this series will only depend on the previous $k-1$ values, we simulate 1,000 more values than we need and discard the first 1,000.

Lastly, we contaminate this data with systematic and random errors. Random errors are generated by sampling 1, 2, 5 or 10% of the observations and adding or subtracting 4 to 6 times the standard deviation, σ, of the simulated series. Systematic errors are generated by sampling 1, 2, or 3 change points and then drawing a break size from a $N(0, 0.04\sigma^2)$. The break size is then added to all observations after the change point. Both the contaminated and uncontaminated datasets are stored for future comparison purposes.

[**There has been no discussion yet on the actual temperature data that is used to estimate the parameters in the model and that the simulation is based on. Different climate zones/pressure levels, as in Ashley's paper. It would also be nice to see a realization from this model both before and after contamination.**]

3 Homogenization Algorithms

Radiosonde observations are collected over long periods of time, as long as 100 years for some stations, and therefore systematic changes in the mean temperature are not uncommon. These errors can happen for one of many reasons: changes in instrumentation, location of a station, post-processing of data, etc. Methods which detect and/or correct these breaks are referred to as homogenization algorithms, and many such techniques have been developed by the meteorological community. [**References needed here.**] Many of the homogenization algorithms make use of metadata, which document changes in the data collection process and/or compare data from neighboring stations. We do not evaluate such algorithms since we simulate data from one station and pressure level at a time. In [7], SNHT is applied by combining both metadata and the ERA-40, but we use a simplified version that operates purely on the observed data.

In this section, we [**Briefly summarize what will be presented in this section.**]

3.1 Homogenization Methodology

We compare the abilities of four different homogenization algorithms to detect systematic errors when random errors are also present in the data. Two algorithms, Binary Segmentation (BinSeg) [**Cite reference(s).**] and Pruned Exact Linear Time (PELT) [**Cite reference(s).**] detect the number and location of changepoints by optimizing a cost function of the form

$$\sum_{i=1}^{m+1} [\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i})] + \beta f(m), \tag{3}$$

where m is the number of changepoints; $\mathcal{C}$ is a cost function; $y_{(\tau_{i-1}+1):\tau_i}$ is the observed data between the $(i-1)$th and ith changepoint; and $\beta f(m)$ is a penalty term on the number of changepoints to prevent overfitting [12]. Often, $\mathcal{C}$ is chosen to be twice the negative log likelihood, and $f(\cdot)$ is linear.

Optimization of Equation (3) can be done in several ways. BinSeg follows a divide-and-conquer algorithm: each observation is considered a candidate changepoint, and the one which leads to the largest reduction in the cost function is chosen as a changepoint. This changepoint

then segments the data into two groups, and the same procedure is repeated on each segment. If no observations lead to a reduction in the cost function, then the procedure is terminated. BinSeg is known to be computationally efficient but is not guaranteed to reach the global minimum of the cost function.

PELT is another algorithm for optimizing Equation (3), but it computes the exact minimum. It proceeds recursively as follows: first, the optimal changepoint configuration [**What does "optimal cp configuration mean?"**] is determined for observations 1 and 2 only. The optimal configuration for the first three observations is then determined using this information, and more generally the optimal configuration for the first $k+1$ observations is determined by considering the optimal configurations for the first $2, 3, \cdots, k$ observations. PELT is also known to be computationally efficient. For our analysis, we used the BinSeg and PELT algorithms implemented in the `changepoint` package in R [11].

The SNHT test works as follows. For each observation, two means are computed: one for the N days prior to the observation, $\bar{X}_{L,t}$, and one for the N days following, $\bar{X}_{R,t}$. Then, the test statistic

$$T_t = \frac{N}{s_t}\left((\bar{X}_{L,t} - \bar{X}_t)^2 + (\bar{X}_{R,t} - \bar{X}_t)^2\right), \tag{4}$$

is computed where $\bar{X}_t$ is the mean of $\bar{X}_{L,t}$, and $\bar{X}_{R,t}$, and s_t is the estimated standard deviation over the N days prior and N days following observation t. If there are not N observations both before and after the current observation, then no test is performed. If the largest T_t exceeds some threshold, we conclude that a break occurred at time t, and we adjust all observations after time t by $\bar{X}_{L,t} - \bar{X}_{R_t}$. Homogenization now proceeds iteratively. T_t is recomputed for $t = 1, \ldots, n$, and test is performed again until no T_t exceed the threshold. In [7], they recommend a threshold of 100, and we found this value to work well in our simulations. Note that in practice, it is generally preferable to homogenize to the most recent data, as that data is considered to be more reliable.

We propose an alternative estimator that replaces the means and standard deviation in Equation (4) with Tukey's biweight estimator of the mean and standard deviation. [**Cite reference(s).**] This estimator is robust against random errors, which may be present during homogenization. None of the other three homogenization algorithms considered are robust against

random errors when $\mathcal{C}$ is chosen to be twice the negative Gaussian log likelihood. **Does this last statement make sense? Should we consider robust versions of PELT and BinSeg?– (ASH: Only if you think you have time. We can discuss in person.) I think we could maybe implement them in a way similar to the skew-t, i.e. bounding the likelihood. In my current simulations, I'm introducing random errors by adding $N(0, 4^2)$ to $N(0, 6^2)$ observations, maybe larger outliers would show the robust estimators to be better performing?** All of the homogenization algorithms considered have tuning parameters: for PELT and BinSeg we must choose penalty functions and the β constant, and for SNHT and its robust variant we must specify the period N. Thus, in our simulations we vary these tuning parameters to observe their effect on the overall performance.

3.2 Results

Evaluation of homogenization algorithms can be done by computing the number of simulated breaks in the data that were accurately detected. However, it is unlikely that a homogenization algorithm will detect the exact time of the break in the data, and thus hit rate is not a very useful metric. Instead, we consider efficiency as defined in [3]. Let $\mathbf{x}$, $\mathbf{c}$, and $\mathbf{h}$ be the original, contaminated, and contaminated and homogenized time series, respectively and let the ith observation be denoted by x_i, c_i, and h_i respectively. The Root Mean Square Error (RMSE) of $\mathbf{h}$ is then defined as follows:

$$\text{RMSE}(\mathbf{h}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h_i - x_i)^2}.$$

Then, the efficiency of the homogenized time series, where 1 means perfect skill and 0 means no improvement is

$$E(\mathbf{h}) = \frac{\text{RMSE}(\mathbf{c}) - \text{RMSE}(\mathbf{h})}{\text{RMSE}(\mathbf{c})}.$$

Also, we remove the random errors in $\mathbf{c}$ and $\mathbf{h}$ before computing the RMSE scores. [**Why are the random errors removed before computing the efficiency? Explain why.**]

We compare the efficiency of all four different homogenization algorithms on simulated datasets. The simulated datasets contain 10 years worth of simulated data with changepoints

introduced at 2.5, 5, and 7.5 years (configuration 1) or 3, 4, and 7 years (configuration 2). Also, we test the following tuning parameters:

- PELT: We consider penalties of $\beta = n/2$, n, $2n$, $4n$, and $8n$.
- BinSeg: Instead of adjusting the penalty term, we restrict the maximum number of changepoints from 1 to 10. We use a small enough penalty so that we would always detect the maximum number of allowed changepoints.
- SNHT: This algorithm computes means of seasonal data, and so periods which are multiples of a year should be considered. Thus, we use one and two year averaging windows, so $N = 365$ or $N = 730$.
- Robust SNHT: We use $N = 365$ and $N = 730$.

We find that the best homogenization algorithm for both configurations is the SNHT with $N = 365$ (see Figure 2). The robust version of the SNHT performs almost as well, but we recommend against using it due to its slight drop in efficiency and large increase in computation time. The BinSeg algorithms perform best when we force the algorithm to choose a small number of changepoints, which is not surprising as we have only simulated three changepoints. Its efficiency degrades when we allow more changepoints. The PELT algorithm performs best with a penalty of $4n$, but its performance is much worse than the alternative algorithms.

In addition, a common concern in climatology is to ensure that the trend is not changed by the homogenization procedure. We examine the estimated trend from the homogenized data, and find that homogenization does not generally bias the trend estimate (see Figure 3). [**You need to explain better here what is being plotted in Figure 3.**] Again, we see that the SNHT and Robust SNHT methods perform very well. Moreover, in all methods of homogenization, the estimate of the trend trend is generally improved as compared to the trend in the contaminated (Raw) data. To conclude, we find that the traditional SNHT estimator outperforms all alternatives. Therefore, we use the SNHT for the remainder of this paper.

4 Sequencing Study

[Needs a paragraph here summarizing the content of this section.]

4.1 Random Error Detection

We follow the error identification process described and validated in [2]. Given that errors are present in the data, traditional methods of computing the mean and standard deviation are known to perform poorly. Thus, the authors use the Huber estimator, which produces a robust measure of the location and measures of scale for both the left and right sides of the distribution [8]. The Huber estimator is computed by initially estimating the center, μ, and scale, σ_R and σ_L with the median and the mean absolute deviation, respectively. Observations that are more than $\hat{\mu} + k\,\hat{\sigma}_R$ are replaced with $\hat{\mu} + k\,\hat{\sigma}_R$, and observations smaller than $\hat{\mu} - k\,\hat{\sigma}_L$ are replaced with $\hat{\mu} - k\,\hat{\sigma}_L$; this process is called Winsorizing. A new set of the mean and standard deviations are computed using this Winsorized data, and the process is repeated until the estimates converge.

Anderson et al. [2] investigate several different strategies for selecting subsets of observations with which to estimate the Huber mean and standard deviations. The *global* set uses all of the observations to estimate the parameters, and the *Hourly Combined* set takes all observations within a 45 day and 12 hour window of each observation and computes parameters estimates for each one. Their final algorithm first removes observations whose z-scores based on the global parameter estimates are greater than 5, and then removes observations whose z-scores based on the Hourly Combined parameter estimates exceed 5.

4.2 Sequencing Simulation

We apply four different sequencings of homogenization and random error identification to the data: homogenization followed by error detection; error detection followed by homogenization; homogenization followed by error detection followed by homogenization; and error detection followed by homogenization followed by error detection. We refer to these approaches as "Sys-Ran," "Ran-Sys," "Sys-Ran-Sys," and "Ran-Sys-Ran," respectively.

In our initial simulations, we only applied "Sys-Ran" and "Ran-Sys." However, we noticed that many errors were not detected if the data was not first homogenized and that our homogenization procedure did not perform as well if errors were not first removed. Thus, we consider the two additional methods "Sys-Ran-Sys" and "Ran-Sys-Ran" to ensure that a homogenization always occurs prior to error detection and vice versa. [**Needs a better explanation for why these 3 stage sequencings were introduced.**]

In summary, the simulation process is as follows: [**Should move a version of this to the Homogenization Study section.**]

1. Fit Equation (1) to observed radiosonde temperature data, and fit a skew-t distribution to the errors from this model. Compute the autocorrelation of these errors via Eq(2).

2. Using the results from step 1, simulate two observations per day for 57 years, or roughly 20,000 observations. Contaminate 1%, 2%, 5%, or 10% of the observations with random errors that are 4 to 6 times the standard deviation of the simulated data. Contaminate the dataset with 1, 2, or 3 change points and with size 0.25, 0.5, 1.0 or 2.0 times the standard deviation of the simulated data.

3. Apply each sequencing of the quality control process.

4. Store the true and false positive rates [**These rates have not yet been defined.**] for random error detection as well as the efficiency of the homogenization algorithm.

5. Repeat steps 2 through 4 [**XX**] times.

4.3 Sequencing Results

To evaluate the performance of the various sequencings, we examine the true and false positive rates, TPR and FPR, respectively. The percent of error contamination as well as the number of simulated change points in the data can strongly influence TPR and FPR. Thus, we perform a 3-way ANOVA to assess the influence of those two factors as well as the order of the quality control process on TPR and FPR.

In our simulation, we find the false positive rate to generally be very small, with a mean value of 0.017%. [**Is this the mean across all factors?**] No interaction terms are significant in the ANOVA model, and we find that the percent of errors simulated was the only variable that had a significant effect on FPR. Not surprisingly, increased contamination with errors leads to slightly smaller false positive rates.

For TPR, we find significant two-way interaction terms: the sequencing with both the simulated number of breaks and the percent of error contamination. Figure **??** shows interaction plots, and it is clear that, for example, the influence of outlier contamination changes as the sequencing changes. However, the interaction effects are small and, while statistically significant, are not very meaningful in interpretation. However, it is evident that the "Ran-Sys" approach performs poorly compared to the three alternatives. This suggests that it is important to homogenize the data prior to performing error detection procedures.

Using the efficiency metric defined in Section 3.2 as the response, we use a 3-way ANOVA to assess the influence of percent of error contamination, the number of simulated change points, and the sequencing. We find that the only significant interaction term is between the percent error contamination and the number of simulated change points. [**And how does the interaction behave? Describe what you see in the figure.**] Furthermore, the influence of the sequencing is not significant (see Figure 5). [**Sequencing is not significant at all?? Not even as a main effect? It does appear to affect TPR, so we need to figure out what this means in the context of the problem. Does sequencing have no effect at all?**]

5 Conclusion

In this study we have evaluated several different homogenization techniques, and we found that the SNHT method performs well. It attains a high efficiency, indicating that this method is reasonably effective at returning the data to its uncontaminated state. Also, it does not bias our estimate of the trend in the data, and the estimate of the trend has a smaller variability than the other homogenization techniques.

We have also evaluated the effect that the sequence in which the random error detection

and homogenization algorithms are applied have on the final performance of the overall quality control routine. We find that failing to remove systematic errors before searching for random errors leads to a much lower true positive rate of the error removal algorithm. However, the removal of random errors first does not have a large influence on the detection of systematic errors. [**Isn't this contrary to what we said would be intuitive in the introduction? If so, we should comment on that.**]

Thus, we recommend performing data homogenization first followed by random error detection. This two step procedure performs significantly better than its reversal, and it performs similarly to three step procedures. [**We need a table of numbers and standard errors that backs up this statement, not just plots, and it should probably go in Section 4.3 not in the Conclusion section.**] The three step procedures do not perform significantly better than "Sys-Ran." As the three step procedures can be more computationally expensive, especially given the large amount of radiosonde data, we recommend against their use.

References

[1] Hans Alexandersson. A homogeneity test applied to precipitation data. *Journal of climatology*, 6(6):661–675, 1986.

[2] Ashley Anderson, Amanda Hering, Joey Comeaux, and Doug Nychka. A simulation study to compare statistical quality control methods for error detection in historical radiosonde temperatures. *In Preparation*, 2014.

[3] Peter Domonkos. Measuring performances of homogenization methods. *QJ Hung Meteorol Serv*, 117(1):91–112, 2013.

[4] Imke Durre, Russell S Vose, and David B Wuertz. Overview of the integrated global radiosonde archive. *Journal of Climate*, 19(1):53–68, 2006.

[5] William P Elliott and Dian J Gaffen. On the utility of radiosonde humidity archives for climate studies. *Bulletin of the American Meteorological Society*, 72(10):1507–1520, 1991.

[6] Robert E Eskridge, Oleg A Alduchov, Irina V Chernykh, Zhai Panmao, Arthur C Polansky, and Stephen R Doty. A comprehensive aerological reference data set (cards): Rough and systematic errors. *Bulletin of the American Meteorological Society*, 76(10):1759–1775, 1995.

[7] Leopold Haimberger. Homogenization of radiosonde temperature time series using innovation statistics. *Journal of Climate*, 20(7):1377–1403, 2007.

[8] Peter J Huber. *Robust statistics*. Springer, 2011.

[9] Eugenia Kalnay, M Kanamitsu, R Kistler, W Collins, D Deaven, L Gandin, Mo Iredell, S Saha, G White, J Woollen, et al. The ncep/ncar 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3):437–471, 1996.

[10] Masao Kanamitsu, Wesley Ebisuzaki, Jack Woollen, Shi-Keng Yang, JJ Hnilo, M Fiorino, and GL Potter. Ncep-doe amip-ii reanalysis (r-2). *Bulletin of the American Meteorological Society*, 83(11):1631–1643, 2002.

[11] Rebecca Killick and Idris A. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

[12] Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[13] John R Lanzante. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *International Journal of Climatology*, 16(11):1197–1226, 1996.

[14] John R Lanzante, Stephen A Klein, and Dian J Seidel. Temporal homogenization of monthly radiosonde temperature data. part i: Methodology. *Journal of Climate*, 16(2):224–240, 2003.

[15] Yingbo Li and Robert Lund. Bayesian multiple changepoint detection using metadata. *(submitted)*, 2014.

[16] QiQi Lu, Robert Lund, Thomas CM Lee, et al. An mdl approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1):299–319, 2010.

[17] JJ Rennie, JH Lawrimore, BE Gleason, PW Thorne, CP Morice, MJ Menne, CN Williams, W Gambi Almeida, JR Christy, M Flannery, et al. The international surface temperature initiative global land surface databank: monthly temperature data release description and methods. *Geoscience Data Journal*, 2014.

[18] AJ Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.

[19] Victor KC Venema, Olivier Mestre, Enric Aguilar, Ingeborg Auer, Jose A Guijarro, Peter Domonkos, G Vertacnik, Tamas Szentimrey, Petr Stepanek, P Zahradnicek, et al. Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8(1):89–115, 2012.

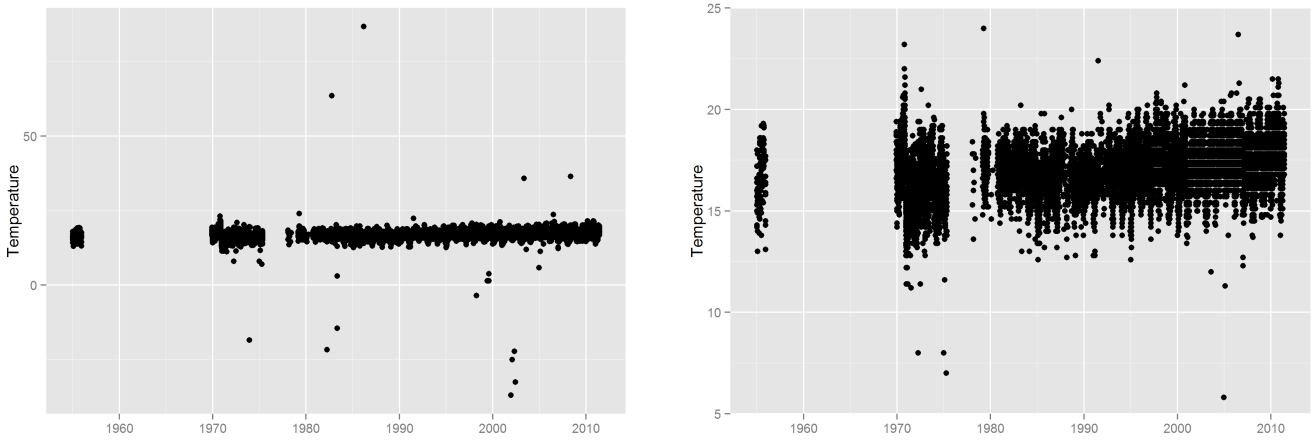

Figure 1: Temperature for Station **???** plotted over time. The second image is the same as the first but with a smaller y-axis window so as to show more detail.

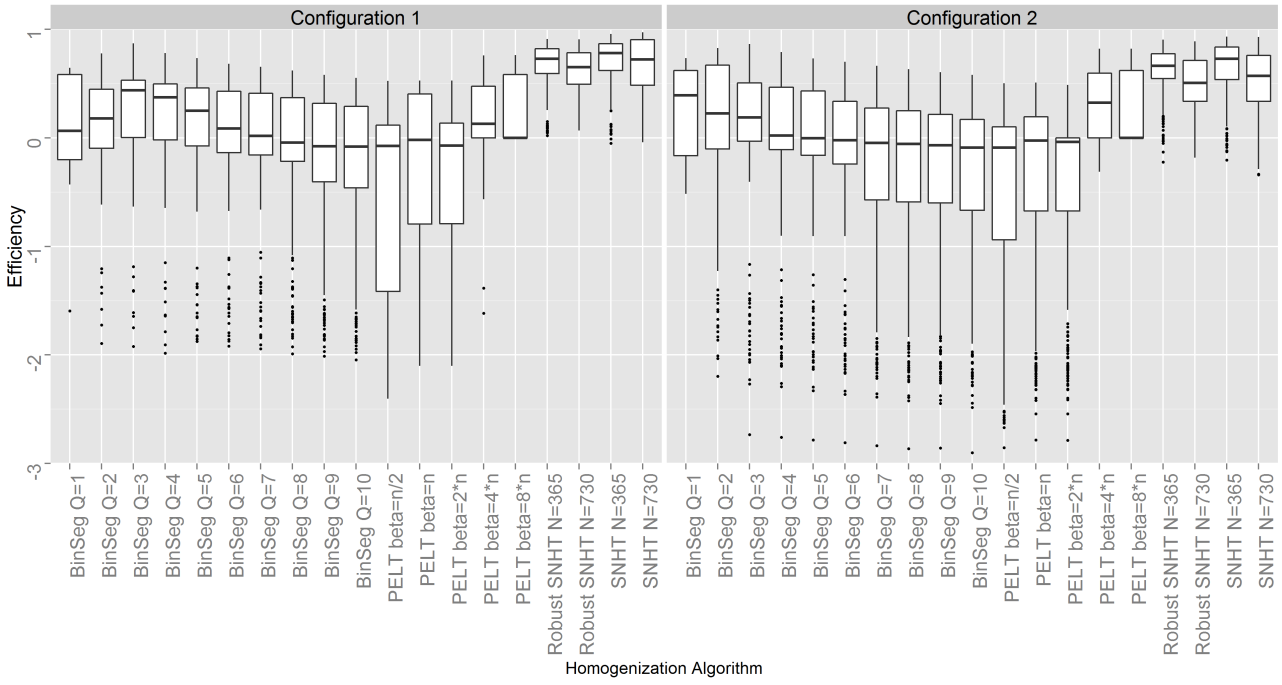

Figure 2: Mean efficiency for the various homogenization algorithms.

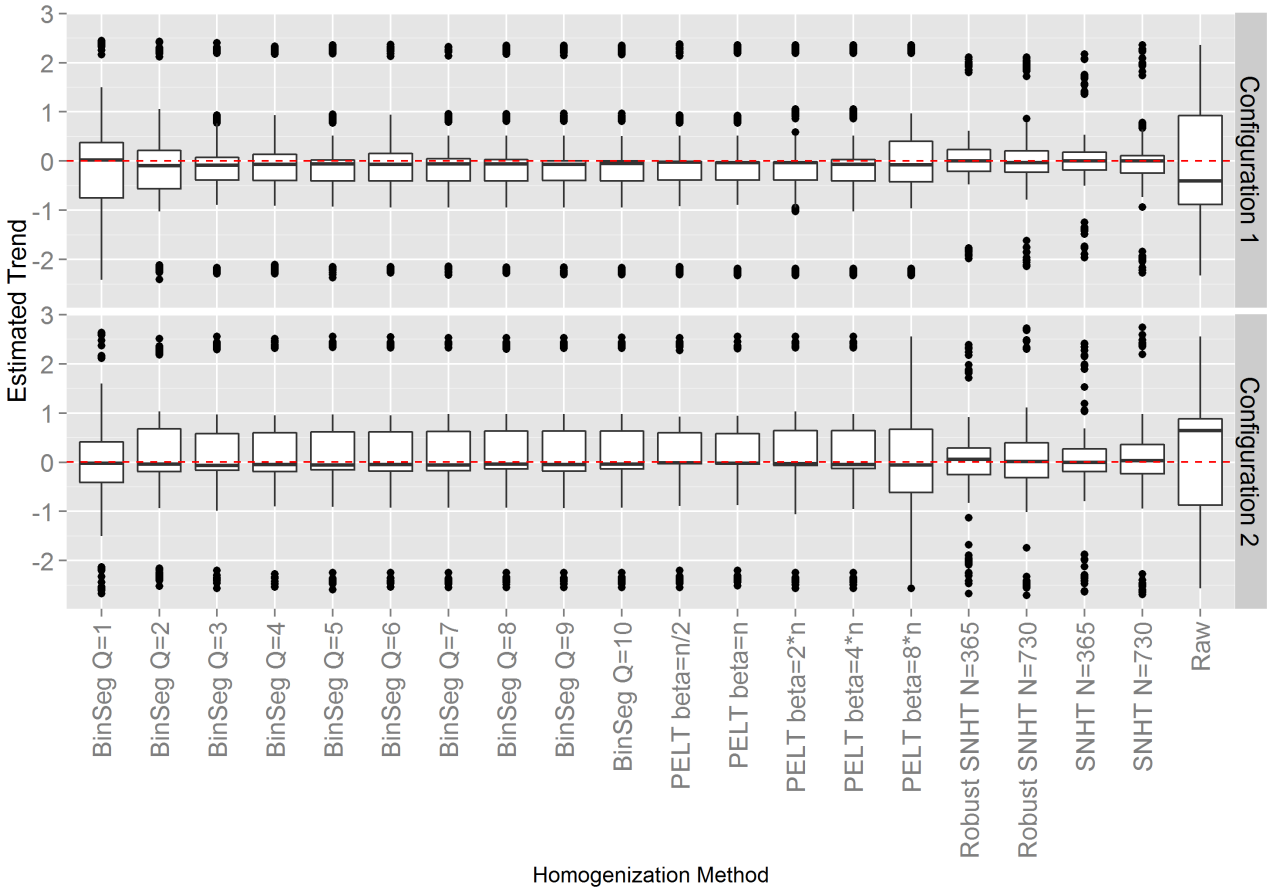

Figure 3: Estimates of trend from homogenized data. "Raw" method indicates no homogenization has been performed, and the red dashed line is the true/simulated trend.

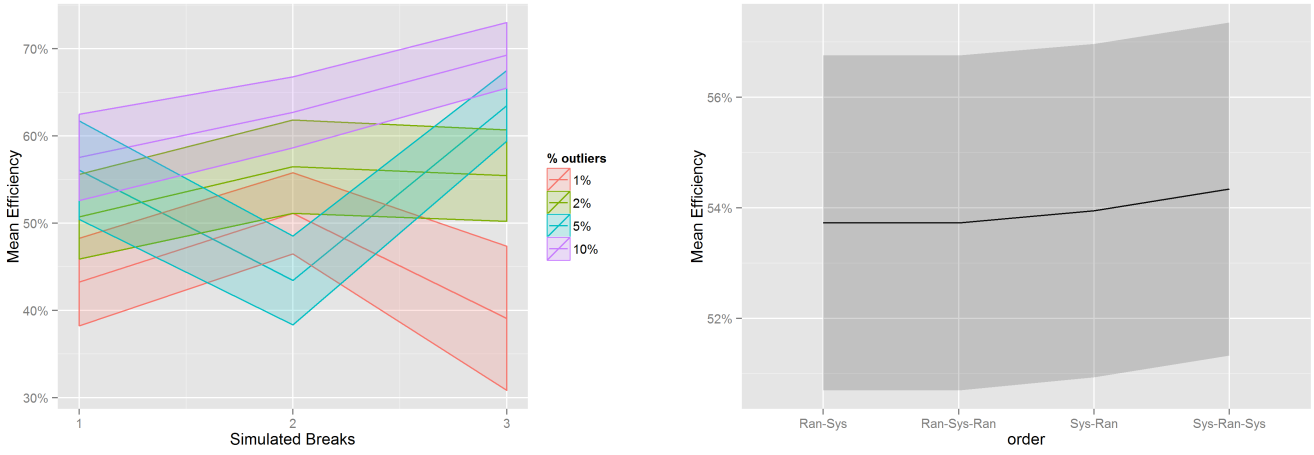

Figure 4: Mean efficiency for the various quality control methods. The bands represent the 2.5% and 97.5% quantiles for all simulations.

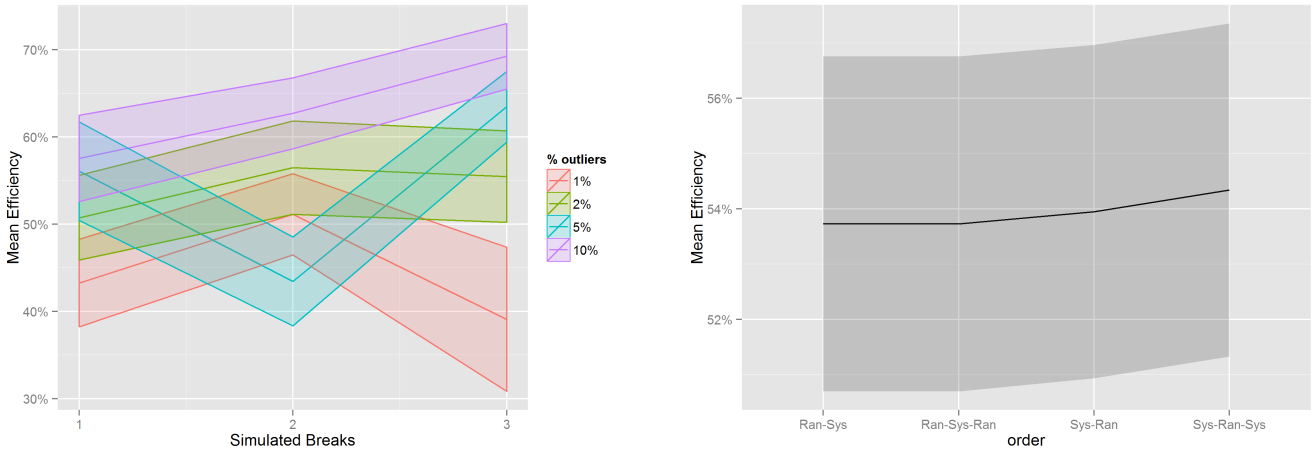

Figure 5: Mean efficiency for the various quality control methods. The bands represent the 2.5% and 97.5% quantiles for all simulations.