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

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

January 14, 2016

Abstract

The historical radiosonde temperature archive, 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 or random errors without considering an optimal process for treatment of both. Additionally, little has been done to evaluate homogenization methods that identify and correct systematic errors when 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, (2) the influence of removing seasonality, and (3) the sequence in which the random and systematic errors are identified and corrected. We introduce a robust Standard Normal Homogeneity Test (SNHT) and find in simulations that it performs better than the traditional SNHT, and it is better than several other 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 is not as detrimental to the robust SNHT homogenization algorithm.

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

Short title: Simultaneous Random and Systematic Error Detection

[1]Department of 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 [25] with over 32,000 stations, and the Integrated Global Radiosonde Archive (IGRA) housed at the National Climatic Data Center (NCDC) [6] 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. To illustrate, Figure 1 plots the temperature recorded by radiosondes at the 50 mb pressure level at Station 70219 (Bethel, Alaska, USA); this time series appears to have random and possibly systematic 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 presented separately, and opinions among climate and weather scientists differ in terms of which type of error should be handled first. We use the term *homogenization algorithms* to refer to the process of identifying and correcting systematic errors while *quality control* (QC) methods are used to identify and remove random errors. 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 both sources of error are present. In addition, homogenization algorithms that use estimators that are robust against random errors have not yet been considered, so these are proposed and investigated as well.

In this paper, we 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 direct measured values of the upper atmosphere, it is an important resource for studies in climate change [8, 9] and for use as an input to global reanalysis datasets [16, 17]. Currently over 2,000 station locations exist, and atmospheric variables are collected at standard pressure levels as the radiosonde rises through the atmosphere. In datasets such as these, error detection methods must be automated since the archives are so large that visual inspections of every station are

not feasible. In addition, the very old historical records have nothing against which they can be externally validated. Reference datasets back to 1958 exist and can be used as in [11, 12], but the older the launches are, the more difficult it is to find historically accurate observations for comparison.

Many methods have been developed to homogenize radiosonde data, but most are not tested on simulated data with known contamination errors [9, 11, 12, 21, 22, 29]. However, a study was recently conducted by the European Cooperation in Science and Technology to compare many different homogenization methods [29]. A single 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 experiment provided a way to compare the performance of these methods. Another approach has been to generate realistic radiosonde temperatures using the Hadley Centre's atmospheric model HadAM3, add random Gaussian noise, and then introduce realistic systematic errors to compare homogenization methods [27, 28], but the computational requirements are so great that only a few realizations can be simulated.

Most homogenization techniques are designed for monthly or annual time series, but radiosonde observations occur, on average, twice daily. Some of these techniques rely on optimizing an objective function over all possible points at which a systematic error could occur [20, 23, 24, 26]. Referred to in the statistical literature as *change point detection* methods, many of these approaches are too computationally expensive for daily data and are not designed to work with nonstationary time series. For example, a strong seasonal and daily trend is present in radiosonde temperatures, making them ill-suited for application of these methods. Additionally, some methods may only locate proposed change points and may not correct for the difference in means, which is a necessary homogenization step. In this paper, we compare the Standard Normal Homogeneity Test (SNHT) [1], the PELT algorithm [20], and binary segmentation [26] when applied to data with a seasonal trend and for which a seasonal trend has been removed. We also propose a robust version of the SNHT.

Automated random error detection methods for radiosonde temperature data have not been

investigated as thoroughly [6, 7, 21]. Models such as [15] for the entire vertical column using pairs of locations could be adapted for random error detection, but recently several random error detection methods, corresponding to the climatological check step of [7], for a given location and pressure level are proposed and evaluated with simulated datasets [2]. The authors find that the optimal error detection algorithm requires two steps: first scanning for observations that are too many standard deviations from the global mean and secondly scanning for observations that are too many standard deviations from their local mean. Robust estimators of mean and standard deviation are used in both cases to mitigate the influence of errors, and a robust, asymmetric estimate of standard deviation is introduced to account for skewness in temperature distributions.

However, to our knowledge, no research has been done to date describing which type of error should be handled first when a dataset contains both types of errors. We do a simulation study in which data is contaminated with both known random errors and with known change points so that we can evaluate (1) the performance of our robust SNHT compared to other modern alternatives; (2) the effect of strong seasonality on homogenization methods; and (3) the sequence in which different quality control algorithms are applied. We henceforth refer to the choice of performing random error detection or systematic error detection first as "the sequence of the methods" or simply "the sequence." We first describe the details of our data simulation and contamination. Then, the performance of the homogenization algorithms are evaluated, and the results from the sequencing study are given. Finally, we conclude with a case study of this method applied to a real dataset and offer some overall recommendations.

2 Simulation Method

Observational data cannot be used to evaluate the performance of homogenization and QC methods directly since we cannot know exactly where true change points and random 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, and our approach bases the simulation on actual data. 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 radiosonde temperature data. GAMs are flexible, non-parametric models that allow the response variable to be a linear combination of smoothed functions of the input variables [13]. In our case, we model temperature (for a fixed location and pressure level) as a function of hour of day, day of year, and year. We model the annual trend with a linear term to capture long term increases or decreases in the series. Thus, the model we fit is

$$t_i = \beta_0 + s_1(h_i) + s_2(d_i) + \beta_1 y_i + \epsilon_i, \tag{1}$$

where t_i is the temperature at a given station and pressure level; h_i, d_i, and y_i are the hour, day, and year of the i-th observation, respectively; β_0 is the intercept; β_1 is the coefficient for the long term trend; and $s_1(\cdot)$ and $s_2(\cdot)$ are cubic regression splines.

Typically the error term, ϵ_i, 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 as shown by [2] for radiosonde temperatures. Thus, we use a skew-t distribution for the errors of this model, which has 4 parameters, ξ, σ, α, and ν, which are useful in controlling the first four moments of the distribution [3]. This distribution is very flexible and can handle skewed and heavy-tailed data.

In addition, we expect there to be temporal correlation in the error terms. However, since we have already included hourly and seasonal terms in the model, most of this autocorrelation is already explained, so an AR(1) time series model is 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 guaranteed to be equally spaced in time.

Launches are scheduled globally at 0 and 12 UTC; however, many deviations from this pattern are observed, especially in the historic record. 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, $\phi(h)$, in hours, we must use only those observations that are h time steps apart:

$$\widehat{\phi}(h) = \frac{1}{|\mathcal{P}_h|} \sum_{(\hat{\epsilon}_i, \hat{\epsilon}_j) \in \mathcal{P}_h} \frac{(\hat{\epsilon}_i - \bar{\epsilon})(\hat{\epsilon}_j - \bar{\epsilon})}{s^2_{\hat{\epsilon}}}, \tag{2}$$

where $\mathcal{P}_h$ is the set of all pairs of residuals that are h hours apart (or within some window); $|\mathcal{P}_h|$ is the number of pairs of residuals in the set $\mathcal{P}_h$; $\hat{\epsilon}_i$ is the observed residual from Equation (1); $\bar{\epsilon}$ is the average of the residuals; and $s^2_{\hat{\epsilon}}$ is the standard deviation of the residuals. For an AR(1) model, we need only estimate $\phi(\cdot)$ at $h = 12$ hours, and we use a window of 5% of 12 hours, or 0.6 hours.

2.2 Data Simulation

The data simulation procedure has seven steps:

1. Fit Equation (1) to observed radiosonde temperature data at a given location and pressure level. Then, fit a skew-t distribution to the error terms, and model the autocorrelation with Equation (2).

2. We choose a fixed time period and assume that two observations occur for each day 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.

3. We use the GAM model fit in step 1 to determine the expected value of temperature at the simulated time, denoted $\hat{t}_i$.

4. To simulate the noise in the observations, we randomly draw values δ_i from a skew-t distribution with parameters as fit in step 1.

5. We wish to introduce autocorrelation in these δ_i. Thus, we simulate an AR(1) model via

$$\epsilon_i = \widehat{\phi}(12)^{\Delta_{i-1}/12}\epsilon_{i-1} + \delta_i,$$

where ϵ_i is the simulated noise in the model at time i, and Δ_{i-1} is the time difference, in hours, between the $(i-1)$th and ith observation. Note that the ith term in this series will depend on all of the previous $i-1$ values. To ensure the correct correlation structure, we simulate 1,000 more values than needed and discard the first 1,000.

6. Then, the simulated ϵ_i is added to $\hat{t}_i$ to construct a simulated series that is similar to real radiosonde temperatures.

7. Lastly, we contaminate this data with systematic and random errors. (a) Random errors are generated by sampling 1, 2, 5, or 10% of the observations and adding or subtracting a random error following a distribution of $N(10\sigma, 1\sigma^2)$, where σ is the standard deviation of the simulated series, estimated by the variance of the observed errors in Equation (1). (b) Systematic errors are generated by sampling 1, 2, or 3 observations uniformly per simulated decade 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 comparison. Figure 2 shows an example of one of the radiosonde temperature datasets that we use as a basis for simulation as well as two different realizations of simulated and contaminated datasets.

We vary several additional factors within our data simulation to understand the effect that each factor has on homogenization algorithms and the sequence in which the algorithms are applied.

Climate Zones: Radiosonde temperature data from different climate zones can be dramatically different, so we analyze data from many different climate zones. In [2], ten representative

stations are chosen and analyzed from ten different climate types, and we simulate data based on models fit to these ten stations.

Pressure Level: Radiosonde temperature data can also vary over pressure level, and so we analyze the pressure levels chosen in [2]: 100 mb, 300 mb, and 850 mb.

Sample Size: For the sequencing study, sample sizes of twice-daily data are simulated for 20, 40, and 80 years. The study comparing homogenization algorithms, however, is much more computationally expensive, so we use sample sizes of 10, 20, and 40 years.

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 such as changes in instrumentation, relocation of a station, or post-processing of data. Methods which detect and/or correct these errors are referred to as homogenization algorithms, and many such techniques have been developed by the meteorological community [1, 5, 10, 11, 22, 23, 24, 29]. 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 [11, 12], the SNHT is applied by combining both metadata and the ERA-40, which reaches back to 1958. However, since many of the UADB records are older than this, we use a simplified SNHT that operates purely on the observed data.

In this section, we compare the abilities of four different homogenization algorithms to detect systematic errors when random errors are also present in the data. We summarize the methods we investigate, namely Binary Segmentation (BinSeg) [26], Pruned Exact Linear Time (PELT) [20], SNHT, and a new robust SNHT. In addition, we study the performance of these algorithms when the seasonal trend is present versus having been modeled and removed. We simulate data as described in Section 2 and introduce change points and random errors, and then we evaluate the ability of the algorithms to detect the known change points.

3.1 Methodology

Two algorithms, BinSeg and PELT, detect the number and location of change points by optimizing a cost function of the form

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

where τ_j is the time that the jth change point occurs; m is the number of change points; $\mathcal{C}$ is a cost function; $y_{(\tau_{j-1}+1):\tau_j}$ is the observed data between the $(j-1)$ and jth change point; and $\beta f(m)$ is a penalty term on the number of change points to prevent overfitting [20]. Note that, for notational convenience, τ_{m+1} is defined to be the time of the last observation. 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 uses a divide-and-conquer algorithm: each observation is considered a candidate change point, and the one which leads to the largest reduction in the cost function is chosen as a change point. This change point 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 number and location of change points is determined for the first two observations only. The optimal number and location of change points for the first three observations is then determined using this information, and more generally the optimal number and location of change points for the first $k+1$ observations is determined by considering the optimal configurations for the first $2, 3, \ldots, k$ observations. PELT is computationally efficient. For our analysis, we used the BinSeg and PELT algorithms implemented in the `changepoint` package in R [19].

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

statistic

$$T_i = \frac{N}{s_i^2}\left((\bar{X}_{L,i} - \bar{X}_i)^2 + (\bar{X}_{R,i} - \bar{X}_i)^2\right), \tag{4}$$

is computed where $\bar{X}_i$ is the mean of $\bar{X}_{L,i}$ and $\bar{X}_{R,i}$, and s_i is the estimated standard deviation over the N days prior and N days following observation i. If there are not N observations both before and after the current observation, no test is performed. If the largest T_i exceeds some threshold at time $i = i^*$, we conclude that a change point occurred at time i^*, and we adjust all observations after time i^* by $\bar{X}_{L,i^*} - \bar{X}_{R,i^*}$. This homogenizes the time series to the older data, which is done here in simulations for convenience. However in practice, it is generally preferable to homogenize to the most recent data, which is considered to be more reliable [5]. Homogenization now proceeds iteratively. T_i is recomputed for all i that are sufficiently far away from the current change points, $i \in \{1, \ldots, n\} \setminus \{i^* - k, \ldots, i^* + k\}$, and the test is performed again until no T_i exceed the threshold, and we use $k = N$. We note that the test statistic in [11] differs from Equation (4) in that a denominator of s_i is used in place of s_i^2; however, we believe this to be a typographical error, as the statistic in [11] is derived from [1]. In [1], $\bar{X}_{L,i}, \bar{X}_{R,i}$, and $\bar{X}_i$ are first standardized, and this is equivalent to our formulation.

The threshold for determining when a change point has occurred is based on the sampling distribution of Equation (4). The authors in [1, 11, 18] generate the distribution of the SNHT with Monte Carlo techniques; however, this can be time-consuming, and we show here that the sampling distribution can be well-approximated with a parametric distribution. By replacing $\bar{X}_i$ in Equation (4) with $(\bar{X}_{L,i} + \bar{X}_{R,i})/2$, Equation (4) can be rewritten as follows:

$$T_i = \frac{N\left(\bar{X}_{L,i} - \bar{X}_{R,i}\right)^2}{2s_i^2}.$$

Under the null hypothesis that the means on the left-hand and right-hand sides of observation i are the same, namely that $\bar{X}_{L,i} \sim N(\mu, \sigma^2/N)$ and $\bar{X}_{R,i} \sim N(\mu, \sigma^2/N)$, the distribution of $(\bar{X}_{L,i} - \bar{X}_{R,i})$ is $N(0, 2\sigma^2/N)$. If the observations, $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{2N})$, are independent and normally distributed, and if s_i^2 is replaced with the true variance σ^2, then clearly T_i follows a chi-squared distribution with 1 degree of freedom.

However, radiosonde temperature data are not likely to be independent, as temperatures follow seasonal trends which induce a strong dependence among the observations. Thus, homogenization will likely be improved if the seasonal components of the data are removed. We use a model-based approach to removing the seasonality; however, we choose a model simple enough so that it does not inadvertently reduce the size of a change point, as a pure linear trend could do. Thus, we use a GAM model of the form

$$t_i = \beta_0 + s_1(d_i) + \epsilon_i, \tag{5}$$

where t_i is the temperature at a given station and pressure level, d_i is the day of the year of the i-th observation, β_0 is the intercept, ϵ_i is the error of the model, and s_1 is a cubic regression spline. We then subtract this fit from the data, giving errors about an estimated seasonal mean. These errors will no longer exhibit a seasonal trend, and the assumption of independence becomes more reasonable. This simple model is chosen because the same adjustment is applied to the observations both before and after t_i, and thus it leaves the numerator of the SNHT statistic nearly unchanged. In addition, with large sample sizes, the sample means in the SNHT statistic will be approximately normal by the Central Limit Theorem even if the observations themselves are not, thereby allowing us to approximate the distribution of the SNHT statistic with the χ^2_1. We use the 99% quantile of this distribution as our threshold. Since the SNHT statistic is applied many times to the same time series, we use the Benjamini-Hochberg adjustment for multiple testing of dependent tests to ensure that the family-wise Type I error rate remains below the 1% threshold [4].

We now propose a robust version of the SNHT wherein we replace the means and standard deviation in Equation (4) with the Huber M-estimator of the mean and standard deviation [14]. These robust estimators of center and spread are computed as follows:

1. First, the estimates of the mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$, are initialized to

$$\hat{\mu} = \text{median}(\mathbf{x})$$
$$\hat{\sigma} = \text{MAD}(\mathbf{x}),$$

   where $\mathbf{x}$ is a vector of the data, and MAD is the median absolute deviation, defined as

$$MAD = \text{median}(|x_i - \text{median}(\mathbf{x})|).$$

2. Then, Winsorized values, y_i, are computed. These are defined as

$$y_i = \begin{cases} \hat{\mu} - k\hat{\sigma} & : x_i \leq \hat{\mu} - k\hat{\sigma} \\ x_i & : \hat{\mu} - k\hat{\sigma} < x_i \leq \hat{\mu} + k\hat{\sigma} \\ \hat{\mu} + k\hat{\sigma} & : x_i > \hat{\mu} + k\hat{\sigma}, \end{cases}$$

   and k is commonly taken to be 1.5, which is what we use.

3. Updated estimates of $\hat{\mu}$ and $\hat{\sigma}$ are computed as the mean of $\mathbf{y}$ and the standard deviation of $\mathbf{y}$, respectively.

4. Steps 2 and 3 are repeated until $\hat{\mu}$ changes by less than $10^{-6}\hat{\sigma}$.

This definition forces unusually large observations to have little to no influence on the estimators of the mean and standard deviation, and they are robust against random errors, which may be present during homogenization. BinSeg and PELT are not robust against random errors when $\mathcal{C}$ is chosen to be twice the negative Gaussian log likelihood. We have implemented the robust SNHT in an `R` package available on CRAN at `http://cran.r-project.org/web/packages/snht/index.html`.

Evaluation of homogenization algorithms can be done by computing the number of simulated change points in the data that were accurately detected. However, it is unlikely that a homogenization algorithm will detect the exact time of the change point, and thus hit rate is not a very useful metric. Instead, we use efficiency as defined in [5]. Let $\mathbf{x}$, $\mathbf{c}$, and $\mathbf{h}$ be the

original, contaminated, and contaminated and homogenized time series, respectively and let the i-th observation of each set 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 series, where 1 means perfect skill, 0 means no improvement, and negative values indicate degradation is

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

The homogenization algorithm is not designed to locate or correct random errors. Furthermore, random errors in the data introduce variability in the estimate of efficiency, so we remove the random errors in $\mathbf{c}$ and $\mathbf{h}$ before computing the RMSE scores.

We compare the efficiency of all four different homogenization algorithms on simulated datasets. The simulated datasets are either 10, 20, or 40 years long, and change point locations are simulated uniformly at random across the entire time series excluding the first and last year. We simulate either one, two, or three change points per decade. 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 the following tuning parameters to observe their effect on the overall performance:

- PELT: We consider penalties of $\beta = n/2$, n, $2n$, $4n$, and $8n$.
- BinSeg: We use no penalty term and instead restrict the maximum number of change points that can occur, varying from 1 to 10.
- SNHT: This algorithm computes means of seasonal data, so periods which are multiples of a year should be considered. Thus, we use one and two year averaging windows with $N = 365$ or $N = 730$.

- Robust SNHT: We use $N = 365$ and $N = 730$.

In summary, the simulation process is as follows:

1. Simulate and contaminate data as described in Section 2.2.

2. Apply each homogenization algorithm to the contaminated dataset. Counting all of the homogenization algorithms along with each one's unique tuning parameters, we apply a total of 19 different homogenization algorithms. Since each is applied to both the original and detrended data, we have 38 different methods to compare.

3. Store the efficiency of each method.

4. Steps 1-3 are repeated 500 times for each combination of climate zone, pressure level, and sample size.

3.2 Results

Figure 3 depicts a boxplot of the efficiencies measured for each of the different algorithms across all 90,000 simulations, i.e. for 500 simulations at each of 3 pressure levels, 10 stations, 3 sample sizes, and 2 seasonal removal (detrended or not) configurations. The robust version of the SNHT appears to achieve the best efficiency among all homogenization algorithms considered. The BinSeg algorithms perform best when we force the algorithm to choose a small number of change points. However, in practice, we will not know the true number of change points, and the BinSeg algorithm is very sensitive to this choice. The PELT algorithm appears to perform best with a penalty of $n/2$, but its performance is worse than the alternative algorithms. Removing the seasonal trend generally improves the efficiency of the algorithms; however, the improvement is greatest for the algorithms with lower efficiency.

To further understand the performance of these algorithms and to understand their sensitivity to different simulation parameters, we fit a logistic regression model to the simulation results. The response variable is 1 if efficiency is positive and 0 otherwise, and the independent variables we use are the sample size n, the outlier contamination rate, the station, the pressure level, the

homogenization algorithm, and if a seasonal adjustment is applied. We fit 4 different logistic regression models: first one with main effects and then models with k-way interactions, where $k = 2, 3, 4$. Table 1 reports the deviance, a measure of logistic regression model fit, for each model. As the deviance does not decrease substantially after k increases beyond 2, we use the model with only 2-way interaction terms.

Table 2 displays the average fitted efficiency as a function of n, outlier contamination, and homogenization method. Only the most promising methods from Figure 3 are displayed here, and the first set of columns shows the estimated efficiencies when the seasonality is not removed. The second set of columns shows the increase in fitted efficiency when the seasonality is first removed with Equation (5). The first number indicates the fitted efficiency, averaged across all station and pressure level combinations, and the number in parentheses indicates the proportion of station and pressure level combinations where this model attains the highest fitted efficiency. The robust SNHT with seasonal trend removed is the superior model in almost all scenarios. However, the traditional SNHT is occasionally better when the outlier contamination rate is small (0% or 1%). The robust SNHT with a 730 day period performs slightly better on the 40 year datasets but not enough to warrant its use. However, this suggests that it may be optimal if many years of data are available. Almost all of the methods are improved by detrending first, but the SNHT and robust SNHT are not nearly as affected by detrending. The BinSeg method is improved by over 20% for small sample sizes with the gain decreasing as the outlier contamination increases. Based on these results, we use the robust SNHT for the remainder of this paper with $N = 365$ and seasonal trend removed.

Plots of the fitted probability that efficiency is positive are given in Figure 4. Each individual line represents a different simulation configuration (number of years simulated and outlier contamination rate). As seen previously, this plot shows that the SNHT and robust SNHT attain the highest fitted probabilities in all cases. Also, the efficiency generally improves as the sample size increases.

4 Sequencing Study

Many radiosonde temperature datasets have observations collected over long periods of time. Thus, it is possible and likely that both systematic and random errors exist in the data. It is not clear if random errors should be removed from the data prior to systematic errors, or vice versa. Thus, this simulation study investigates the performance of different sequences of these quality control methods.

4.1 Random Error Detection

We follow the random error identification process developed and tested 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 two-sided Huber estimator, which produces a robust measure of the location and robust measures of scale for both the left and right sides of the distribution [14]. The two-sided Huber estimator differs from the estimator described in Section 3.1 only in that it produces two estimates of scale, σ_R and σ_L. The estimate for σ_R (σ_L) is computed using only the data to the right (left) of $\hat{\mu}$.

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 across all years and computes parameter estimates for each one. Their final algorithm first removes observations whose z-scores based on the Global parameter estimates are greater than 6, 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 random error detection; random error detection followed by homogenization; homogenization followed by random error detection followed by homogenization;

and random error detection followed by homogenization followed by random error detection. We refer to these approaches as "Sys-Ran," "Ran-Sys," "Sys-Ran-Sys," and "Ran-Sys-Ran," respectively.

We hypothesize that some random errors may not be detected if the data is not homogenized and that the homogenization procedure may not perform as well if random errors are not first removed. For these reasons, we included the two additional methods "Sys-Ran-Sys" and "Ran-Sys-Ran". In both of these approaches, a homogenization procedure is performed after random error detection. Likewise, a random error detection will be performed after a homogenization algorithm as well.

In summary, the simulation process is as follows:

1. Simulate and contaminate data as described in Section 2.2.

2. Apply each of the four sequencings.

3. Store the true and false positive rate for random error detection as well as the efficiency of the homogenization algorithm. The true positive rate, TPR, is defined as the percent of identified errors that are random errors, and the false positive rate, FPR, is defined as the percent of identified errors that are not random errors.

4. Steps 1-3 are repeated 1,000 times for each climate zone, pressure level, and sample size. Additionally, the analysis is performed with both the robust and traditional SNHT.

4.3 Sequencing Results

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 again fit logistic regression models to the simulation results, where we model each of TPR, FPR, and the probability that efficiency is positive as the response variables. The dependent variables are sample size, outlier contamination rate, station, pressure level, sequencing, and type of homogenization (robust or traditional SNHT).

We begin by fitting five logistic models for each of the three responses, first with only main effects and then models with all k-way interaction terms, where $k = 2, 3, 4, 5$. The deviances are

given in Table 3. For the FPR and efficiency models, we find that the deviance does not decrease much when 3-way interaction terms are included in the model, and thus we use models with 2-way interaction terms. For the TPR model, a decrease in the deviance of 2% when including two-way interactions seems insubstantial after the 63% decrease in deviance from the intercept model to the linear terms model, and thus we use the linear terms model for TPR.

For the TPR model, Table 4 shows the estimated coefficients from the TPR logistic regression model. We find that there is no significant difference between the sequencings "Ran-Sys-Ran," "Sys-Ran," and "Sys-Ran-Sys" even at the significance level of $\alpha = 0.10$, but that these three sequencings are significantly better than the "Ran-Sys" sequencing at even the $\alpha = 0.01$ significance level. This agrees with our initial hypothesis: random error detection will be less accurate if systematic errors have not been removed from the dataset. Moreover, using the robust SNHT leads to a small but significant (at the $\alpha = 0.01$ significance level) improvement in the TPR. Additionally, these effects are plotted in Figure **??**.

Table 5 shows the fitted false positive rates averaged across all stations and pressure levels, and these effects are plotted in Figure 5. The figure clearly depicts that there is not a substantial or significant difference among the four sequencings. In almost all simulations, the "Ran-Sys" sequencing with a traditional SNHT homogenization algorithm attained the lowest FPR; however, FPR is quite low across all models. Due to this fact, and the conclusions from the TPR model, we recommend one of the sequencings "Ran-Sys-Ran", "Sys-Ran", or "Sys-Ran-Sys" if the end goal is to minimize FPR and maximize TPR.

Lastly, results from fitting the efficiency model are shown in Table 6. Note that, in terms of efficiency, the "Ran-Sys" and "Ran-Sys-Ran" models will have identical performance. The largest fitted probability of positive efficiency is almost always obtained with the sequence "Ran-Sys". Additionally, efficiency is optimized by the robust model in the cases of longer time series or higher outlier contamination rates, and the traditional model is preferred with shorter time series and lower outlier contamination rates. These results differ from those in the homogenization study (see Table 2); previously the nonrobust estimator had a lower probability of positive efficiency in all cases. However, the current results are different in that a random error check is performed

prior to the homogenization; in this scenario, the nonrobust SNHT is the best estimator for some configurations. As shown in Figure 6, the difference among the four sequencings and the difference between the robust or traditional SNHT is not significant. Thus, we conclude that the probability of positive efficiencies between the different algorithms are not significantly different but that the "Ran-Sys" and "Ran-Sys-Ran" sequencings are preferred and that the robust SNHT is preferred when many outliers are present or when we have long time series.

In practice, when applying a quality-control algorithm to a historical dataset, we can only use one sequencing and we wish to maximize TPR, minimize FPR, and maximize efficiency simultaneously. The sequencing maximizing both TPR and efficiency is "Ran-Sys-Ran." The difference between the various sequencings for the FPR is very minimal, and thus this sequencing is also a reasonable choice. The robust SNHT leads to only a small improvement in the TPR, and thus the choice between the robust or traditional SNHT should maximize efficiency and thus is based on the length of the time series and assumed outlier contamination percentage.

5 Case Study

In [11], station 70219 (Bethel, Alaska, USA) is analyzed for change points during the period 1958 to 2005. We perform a similar analysis to the temperatures observed at this station from 1951 to 2014, as shown in Figure 1, although we use raw radiosonde temperatures rather than the difference between the observed temperature and a reanalysis. We apply the robust SNHT to this time series followed by the outlier detection algorithm described in Section 4.1 in the Ran-Sys-Ran sequence. As the dataset contains 53 years of data, we use a window of 2 years, $N = 730$.

The algorithm detects a total of 17 change points and 22 outliers. The detected change points have mean shifts ranging from -4.19°C to 2.86°C, and the average absolute shift is 2.17°C. The global random error detection algorithm identifies 17 errors whose test statistics range from 6.05 to 29.41 and average 13.17. The windowed random error detection algorithm detects an additional 5 errors, and their test statistics range from 5.68 to 7.85 and average 6.49. The random error detection algorithm applied after homogenization detects no additional errors.

Figure 7 depicts a plot of the time series before and after the quality control algorithm has been applied. Only a subset of the data is plotted in the middle panel in order to show the effects of homogenization more clearly. In the 10 year period shown, three change points are detected as indicated by the dashed vertical lines. The detected change points shift the mean of the process, and the corrections made in the quality controlled dataset seem to improve the homogeneity of the final product. The results found here are similar to those in [11] who found 13 change points, and when we restrict our analysis to the same time period, we identify 14 change points.

6 Conclusion

In this study we evaluate several different homogenization techniques, and we find that the robust SNHT method with seasonality removed performs the best among those considered. It attains a high efficiency, indicating that this method is reasonably effective at returning the data to its uncontaminated state. It attains higher efficiencies than the BinSeg and PELT algorithms, and it outperforms the non-robust SNHT even when the outlier contamination rate is small. The optimal period may be a function of the size of the dataset, but the 365 day period is optimal for datasets with 40 years or less.

We also evaluate 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 in most cases. Additionally, failing to remove random errors before detecting systematic errors leads to a lower efficiency. Therefore, we must use a sequencing of length three, and we find that "Ran-Sys-Ran" outperforms "Sys-Ran-Sys" by a slight, insignificant amount for both TPR and efficiency.

In the absence of metadata regarding a station's history or for very old historical records, we recommend performing random error detection first followed by the SNHT and then an additional random error detection. The robust SNHT should be used in cases where we have a long time-series or many outliers while the traditional SNHT is preferred otherwise. This three step procedure performs only slightly better than the other three step sequencing, and it substantially

outperforms the two step procedures. Our recommended approach can be applied throughout the archive with minimal human intervention; however, gains in random and systematic error detection may be realized if metadata, multiple pressure levels, or multiple station locations are combined, as in [15]. More work is needed to determine how to best handle both types of errors when such additional information is incorporated.

Acknowledgments

The authors wish to thank the organizers and participants of the SAMSI Surface Temperature Initiative Workshop held at the National Center for Atmospheric Research in Boulder, Colorado on July 8-16, 2014 for helpful discussions regarding the content of this paper.

References

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

[2] A. Anderson, J. M. Browning, J. Comeaux, A. S. Hering, and D. Nychka. A comparison of automated statistical quality control methods for error detection in historical radiosonde temperatures. *Accepted to International Journal of Climatology*, 2015.

[3] A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 367–389, 2003.

[4] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29: 1165–1188, 2001.

[5] P. Domonkos. Measuring performances of homogenization methods. *Quarterly Journal of the Hungarian Meteorological Service*, 117(1): 91–112, 2013.

[6] I. Durre, R. S. Vose, and D. B. Wuertz. Overview of the integrated global radiosonde archive. *Journal of Climate*, 19(1): 53–68, 2006.

[7] I. Durre, R. S. Vose, and D. B. Wuertz. Robust automated quality assurance of radiosonde temperatures. *Journal of Applied Meteorology and Climatology*, 47: 2081–2095, 2008.

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

[9] R. E. Eskridge, O. A. Alduchov, I. V. Chernykh, Z. Panmao, A. C. Polansky, and S. R. Doty. A comprehensive aerological reference data set (CARDS): Rough and systematic errors. *Bulletin of the American Meteorological Society*, 76(10): 1759–1775, 1995.

[10] C. Gruber and L. Haimberger. On the homogeneity of radiosonde wind time series. *Meteorologische Zeitschrift*, 17(5): 631–643, 2008.

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

[12] L. Haimberger, C. Tavolato, and S. Sperka. Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations. *Journal of Climate*, 25(23): 8108–8131, 2012.

[13] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.

[14] P. J. Huber. *Robust Statistics*. Springer, 2011.

[15] R. Ignaccolo, M. Franco-Villoria, and A. Fassò. Modelling collocation uncertainty of 3d atmospheric profiles. *Stochastic Environmental Research and Risk Assessment*, 2014. doi: DOI10.1007/s00477-014-0890-7.

[16] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. 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.

[17] M. Kanamitsu, W. Ebisuzaki, J. Woollen, S.-K. Yang, J. Hnilo, M. Fiorino, and G. Potter. NCEP-DOE AMIP-II reanalysis (r-2). *Bulletin of the American Meteorological Society*, 83 (11): 1631–1643, 2002.

[18] M. Khaliq and M. J. Ouarda, T. B. On the critical values of the standard normal homogeneity test (snht). *International Journal of Climatology*, 27: 681–687, 2007.

[19] R. Killick and I. A. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3): 1–19, 2014. URL `http://www.jstatsoft.org/v58/i03/`.

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

[21] J. 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.

[22] J. R. Lanzante, S. A. Klein, and D. J. Seidel. Temporal homogenization of monthly radiosonde temperature data. part I: Methodology. *Journal of Climate*, 16(2): 224–240, 2003.

[23] Y. Li and R. Lund. Bayesian multiple changepoint detection using metadata. *(submitted)*, 2014.

[24] Q. Lu, R. Lund, and T. C. Lee. An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1): 299–319, 2010.

[25] J. Rennie, J. Lawrimore, B. Gleason, P. Thorne, C. Morice, M. Menne, C. Williams, W. G. Almeida, J. 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. doi: 10.1002/gdj3.8.

[26] A. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974. doi: 10.2307/2529204.

[27] P. W. Thorne et al. A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *Journal of Geophysical Research*, 116(D12): 1–19, 2011.

[28] H. A. Titchner, P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker. Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. *Journal of Climate*, 22(3): 465–485, 2009.

[29] V. K. Venema et al. Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8(1): 89–115, 2012.

Model Type	Deviance	% Reduction
Intercept Only	1973919	-
Linear Terms	1539236	22.02%
2-Way Interactions	1474259	3.29%
3-Way Interactions	1448078	1.33%

Table 1: Deviance for the efficiency logistic regression models.

# of Years	Chngpt Contam	Estimated Efficiency				Change in Efficiency with Seasonal Model			
		SNHT-365	robust-365	robust-730	BinSeg-2-log(n)	SNHT-365	robust-365	robust-730	BinSeg-2-log(n)
10	1 chngpt/decade	94.3% (0%)	98.4% (30%)	97.8% (0%)	54.0% (0%)	2.5% (0%)	**0.7% (70%)**	0.3% (0%)	28.6% (0%)
	2 chngpts/decade	95.3% (0%)	98.3% (30%)	97.3% (0%)	66.8% (0%)	2.1% (0%)	**0.8% (70%)**	0.4% (0%)	22.6% (0%)
	3 chngpts/decade	96.1% (0%)	98.2% (30%)	96.6% (0%)	77.6% (0%)	1.8% (0%)	**0.8% (70%)**	0.5% (0%)	16.1% (0%)
20	1 chngpt/decade	95.0% (0%)	98.6% (37%)	98.4% (0%)	64.0% (0%)	1.8% (0%)	**0.5% (63%)**	0.0% (0%)	22.7% (0%)
	2 chngpts/decade	95.5% (0%)	98.4% (33%)	97.8% (0%)	73.8% (0%)	1.7% (0%)	**0.6% (67%)**	0.0% (0%)	17.6% (0%)
	3 chngpts/decade	95.9% (0%)	98.1% (30%)	97.0% (0%)	81.7% (0%)	1.5% (0%)	**0.7% (70%)**	0.1% (0%)	12.8% (0%)
40	1 chngpt/decade	96.2% (0%)	99.0% (47%)	99.2% (7%)	80.7% (0%)	0.7% (0%)	**0.2% (47%)**	-0.2% (0%)	11.8% (0%)
	2 chngpts/decade	95.9% (0%)	98.6% (53%)	98.6% (0%)	84.8% (0%)	0.8% (0%)	**0.3% (47%)**	-0.4% (0%)	9.5% (0%)
	3 chngpts/decade	95.6% (0%)	98.0% (53%)	97.7% (0%)	88.1% (0%)	0.9% (0%)	**0.4% (47%)**	-0.6% (0%)	7.5% (0%)

Table 2: The first set of columns show the fitted probability of positive efficiency averaged over all station and pressure level combinations when the seasonality is not removed. Numbers in parentheses indicate the percent of station and pressure level combinations where the given model obtained the highest fitted probability of positive efficiency. The second set of columns shows the change in probability of positive efficiency when the seasonality is fit and removed first. Bolded numbers are the best within each row. For brevity, only those models with the best probability of positive efficiency are included here.

	TPR		FPR		Efficiency	
Model Type	Deviance	% Reduction	Deviance	% Reduction	Deviance	% Reduction
Intercept Only	73733264	-	16318581	-	144476	-
Linear Terms	27283481	63%	5920274	63.72%	131823	8.76%
2-Way Interactions	26704211	2.12%	1118133	81.11%	129102	2.06%
3-Way Interactions	26507711	0.74%	1028341	8.03%	128173	0.72%
4-Way Interactions	26446311	0.23%	1013260	1.47%	127742	0.34%
5-Way Interactions	26431890	0.05%	1010137	0.31%	127618	0.1%

Table 3: Deviance for the sequencing logistic regression models.

Variable	Estimate	Standard Error
(Intercept)	4.2961	0.0019
Seq: Ran-Sys-Ran	4.9604	0.0025
Seq: Sys-Ran	4.9555	0.0025
Seq: Sys-Ran-Sys	4.9555	0.0025
station: 51777	0.0471	0.0015
station: 70133	-0.1686	0.0014
station: 70261	-0.2697	0.0014
station: 72387	-0.0933	0.0015
station: 72456	-0.1471	0.0014
station: 74794	-0.2349	0.0014
station: 82332	-0.2366	0.0014
station: 85543	-0.1904	0.0014
station: 94150	-0.1222	0.0014
pressure: 300	0.1576	0.0008
pressure: 850	0.1903	0.0008
outlier %	-23.6029	0.0118
n	-0.0001	0.0000
robust	0.0025	0.0006

Table 4: Coefficients and standard errors from the TPR regression model. All coefficients are significant at the 0.01 level.

Number of Years	Outlier Contamination	Nonrobust Ran-Sys	Nonrobust Sys-Ran	Robust Ran-Sys	Robust Ran-Sys-Ran
20	0%	**0.116% (60%)**	0.133% (3%)	0.116% (33%)	0.134% (3%)
	1%	**0.105% (63%)**	0.125% (3%)	0.106% (30%)	0.127% (3%)
	2%	**0.095% (67%)**	0.118% (3%)	0.096% (27%)	0.120% (3%)
	5%	**0.070% (87%)**	0.099% (0%)	0.073% (13%)	0.103% (0%)
	10%	**0.044% (97%)**	0.076% (0%)	0.047% (3%)	0.081% (0%)
40	0%	**0.103% (63%)**	0.133% (3%)	0.104% (30%)	0.135% (3%)
	1%	**0.093% (63%)**	0.125% (3%)	0.095% (30%)	0.128% (3%)
	2%	**0.084% (70%)**	0.118% (0%)	0.086% (27%)	0.121% (3%)
	5%	**0.063% (87%)**	0.099% (0%)	0.065% (13%)	0.103% (0%)
	10%	**0.039% (97%)**	0.075% (0%)	0.042% (3%)	0.081% (0%)
80	0%	**0.082% (67%)**	0.133% (0%)	0.085% (33%)	0.138% (0%)
	1%	**0.074% (70%)**	0.125% (0%)	0.077% (30%)	0.130% (0%)
	2%	**0.067% (73%)**	0.118% (0%)	0.070% (27%)	0.123% (0%)
	5%	**0.050% (87%)**	0.098% (0%)	0.052% (13%)	0.104% (0%)
	10%	**0.031% (100%)**	0.075% (0%)	0.033% (0%)	0.082% (0%)

Table 5: Fitted FPR averaged over all station and pressure level combinations. Numbers in parentheses indicate the percent of station and pressure level combinations where the given sequencing obtained the lowest fitted FPR. Bolded numbers are the best within each row. Sequencings that never obtained the lowest fitted FPR were removed for space reasons.

Number of Years	Outlier Contamination	Nonrobust Ran-Sys	Robust Ran-Sys	Robust Sys-Ran	Robust Sys-Ran-Sys
20	0%	**91.2% (100%)**	83.9% (0%)	86.0% (0%)	84.5% (0%)
	1%	**91.7% (100%)**	85.9% (0%)	87.1% (0%)	86.3% (0%)
	2%	**92.2% (100%)**	87.6% (0%)	88.0% (0%)	87.9% (0%)
	5%	**93.5% (87%)**	91.8% (0%)	90.6% (0%)	91.8% (13%)
	10%	95.2% (17%)	**96.0% (83%)**	93.7% (0%)	95.8% (0%)
40	0%	**94.4% (70%)**	93.2% (0%)	93.7% (30%)	93.0% (0%)
	1%	**95.0% (67%)**	94.4% (0%)	94.5% (33%)	94.2% (0%)
	2%	**95.5% (60%)**	95.4% (40%)	95.2% (0%)	95.2% (0%)
	5%	96.8% (13%)	**97.5% (87%)**	96.8% (0%)	97.3% (0%)
	10%	98.2% (0%)	**99.1% (100%)**	98.4% (0%)	99.0% (0%)
80	0%	97.6% (0%)	**98.9% (100%)**	98.8% (0%)	98.7% (0%)
	1%	98.1% (0%)	**99.2% (100%)**	99.1% (0%)	99.0% (0%)
	2%	98.5% (0%)	**99.4% (100%)**	99.3% (0%)	99.3% (0%)
	5%	99.2% (0%)	**99.8% (100%)**	99.7% (0%)	99.7% (0%)
	10%	99.8% (0%)	**100.0% (100%)**	99.9% (0%)	99.9% (0%)

Table 6: Fitted probability of positive efficiency averaged over all station and pressure level combinations. Numbers in parentheses indicate the percent of station and pressure level combinations where the given sequencing obtained the highest fitted probability of positive efficiency. Bolded numbers are the highest within each row.

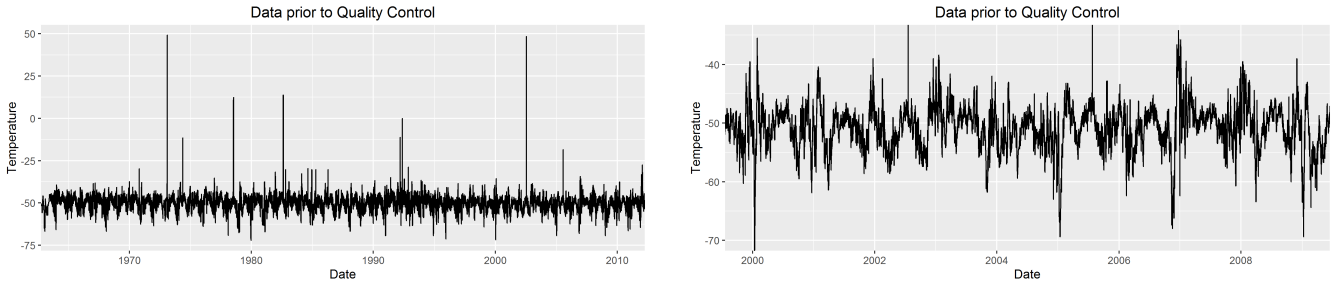

Figure 1: Temperature for Station 70219 plotted over time. The right-hand image is the same as the first but zoomed in to the last ten years to show more detail.

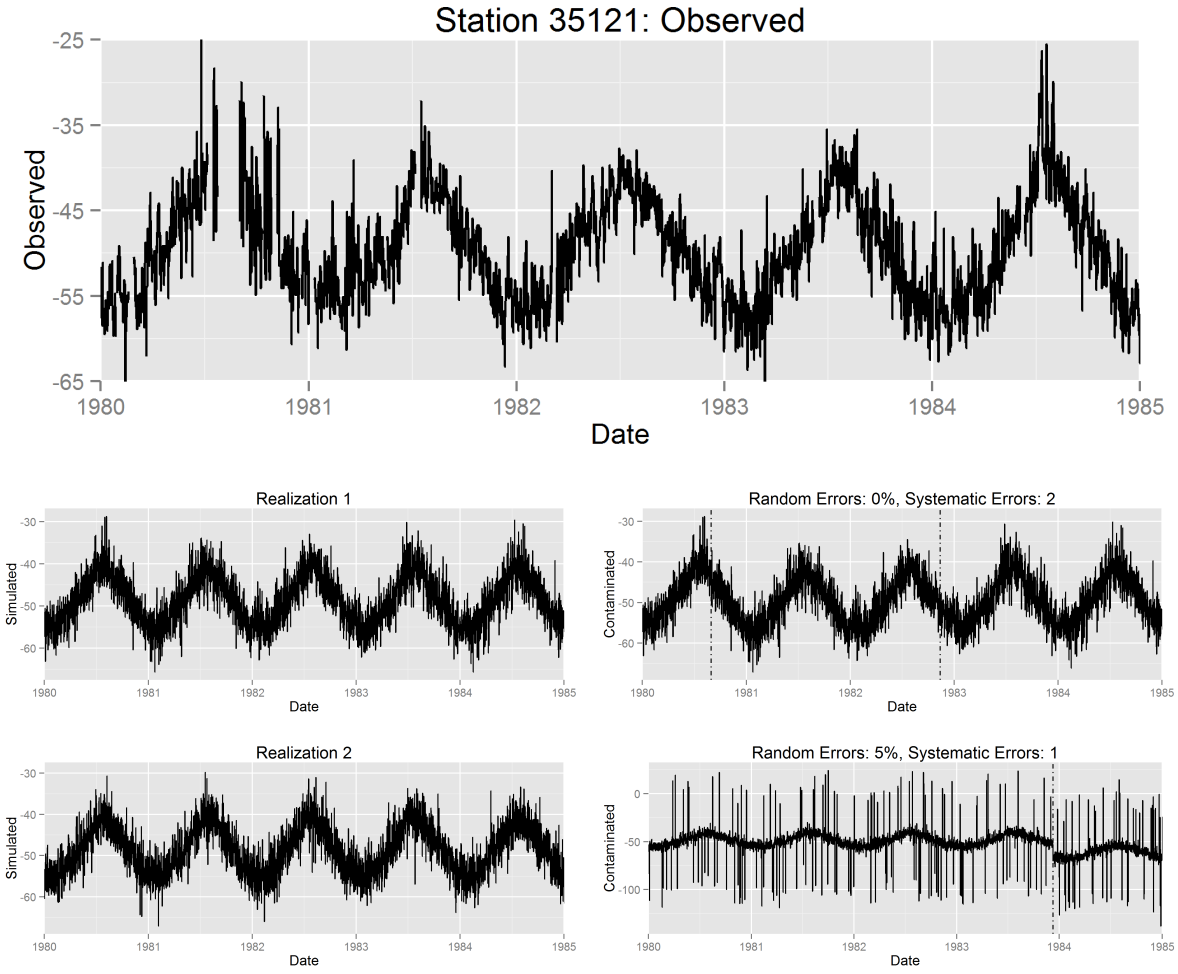

Figure 2: Time series plots of radiosonde temperature data from station 35121. The top plot shows the observed time series, and the following two rows show realizations of simulated datasets. The plots on the left show the simulated data prior to contamination, and the right plots show the data after contamination with a combination of random and systematic errors.

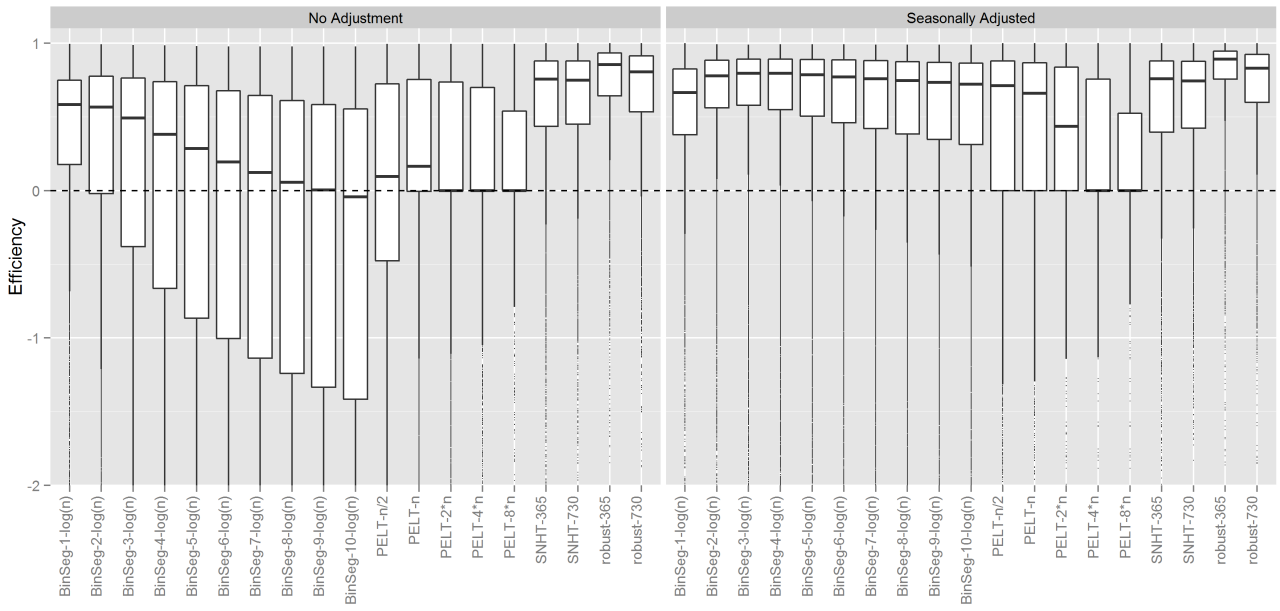

Figure 3: Boxplot of efficiency scores for the various homogenization algorithms. Note that this graph is constrained to the efficiency range of $(-2, 1)$ in order to show more detail. The SNHT and the robust SNHT perform substantially better than their alternatives.

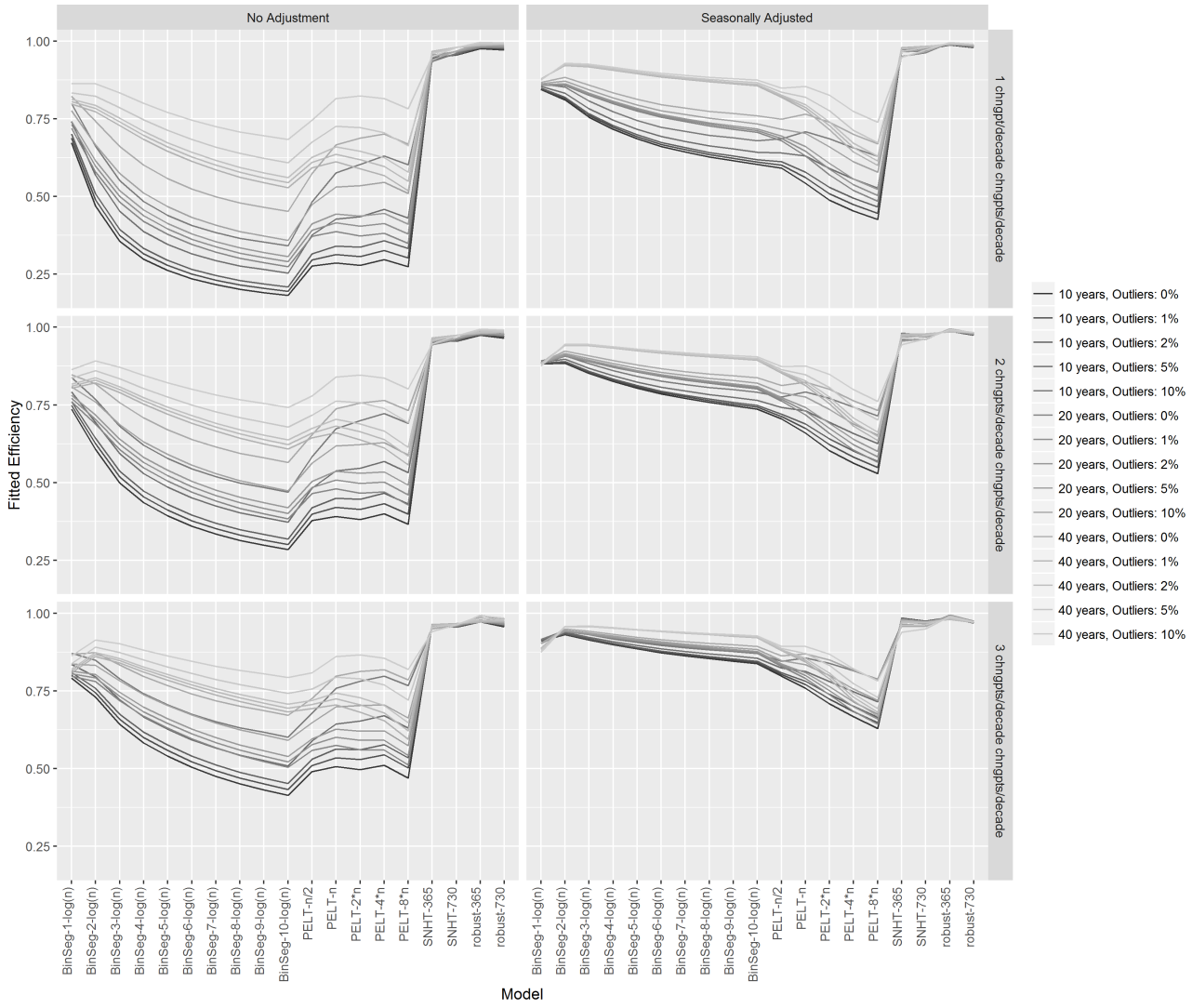

Figure 4: This graph depicts the fitted probability of positive efficiency from the efficiency logistic regression model. Each line represents a different simulation configuration (number of years simulated and outlier contamination rates). There is substantial variability among the different simulations, but the SNHT and robust SNHT models perform better than the alternatives in all simulated configurations, and the probability of positive efficiency improves with sample size.

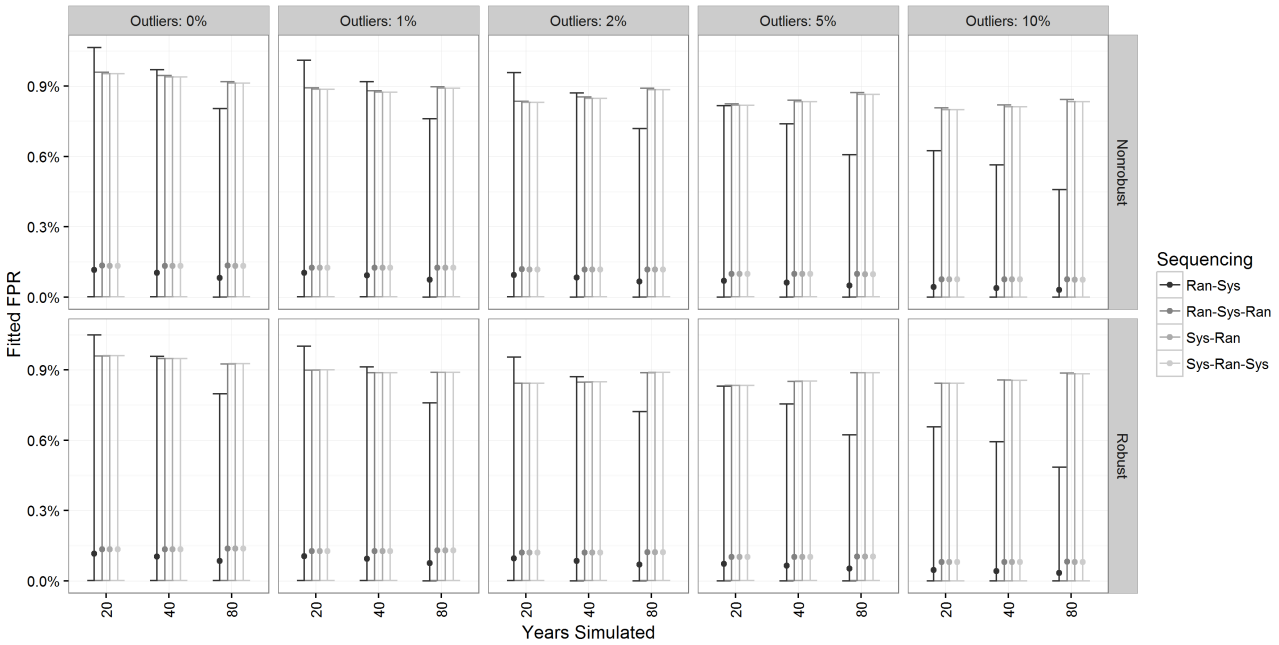

Figure 5: This graph depicts the estimated FPR from the logistic regression model. The dot in each error bar is the estimated FPR averaged over all station and pressure level combinations. The maximum (minimum) of the error bar is the highest (lowest) FPR obtained across all station and pressure level combinations.

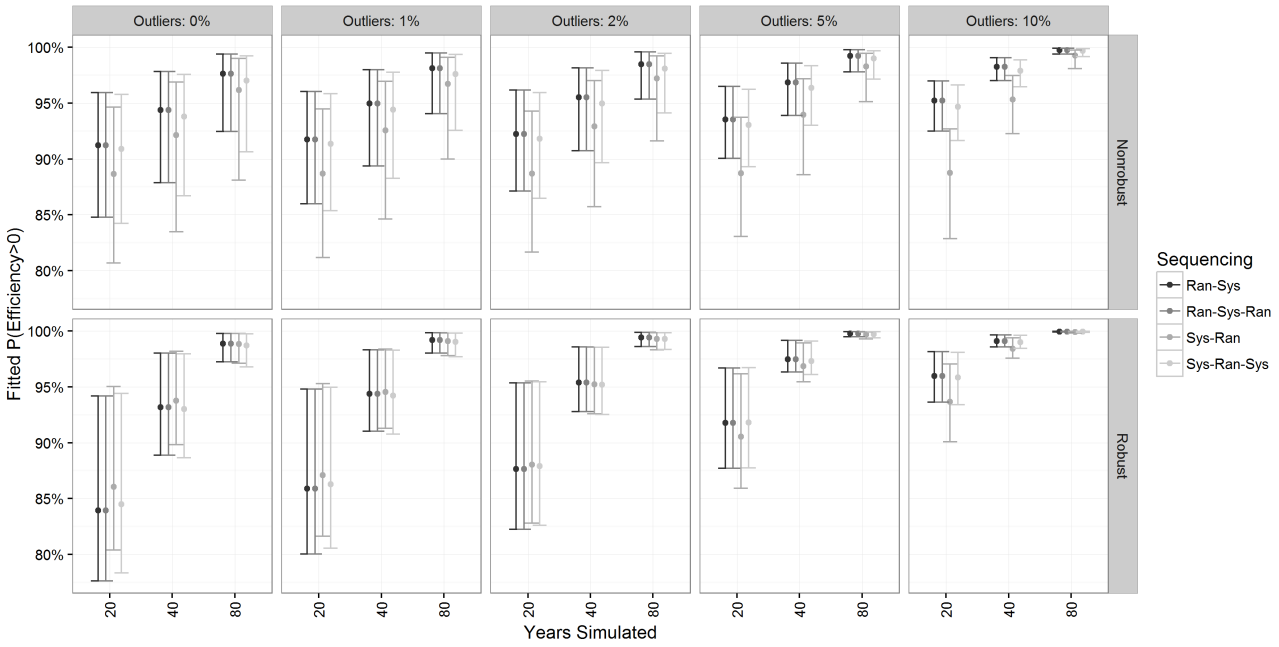

Figure 6: This graph depicts the estimated probability of positive efficiency from the logistic regression model. The middle of each error bar is the estimated probability of a positive efficiency averaged over all station and pressure level combinations. The maximum (minimum) of the error bar is the highest (lowest) probability obtained across all station and pressure level combinations.

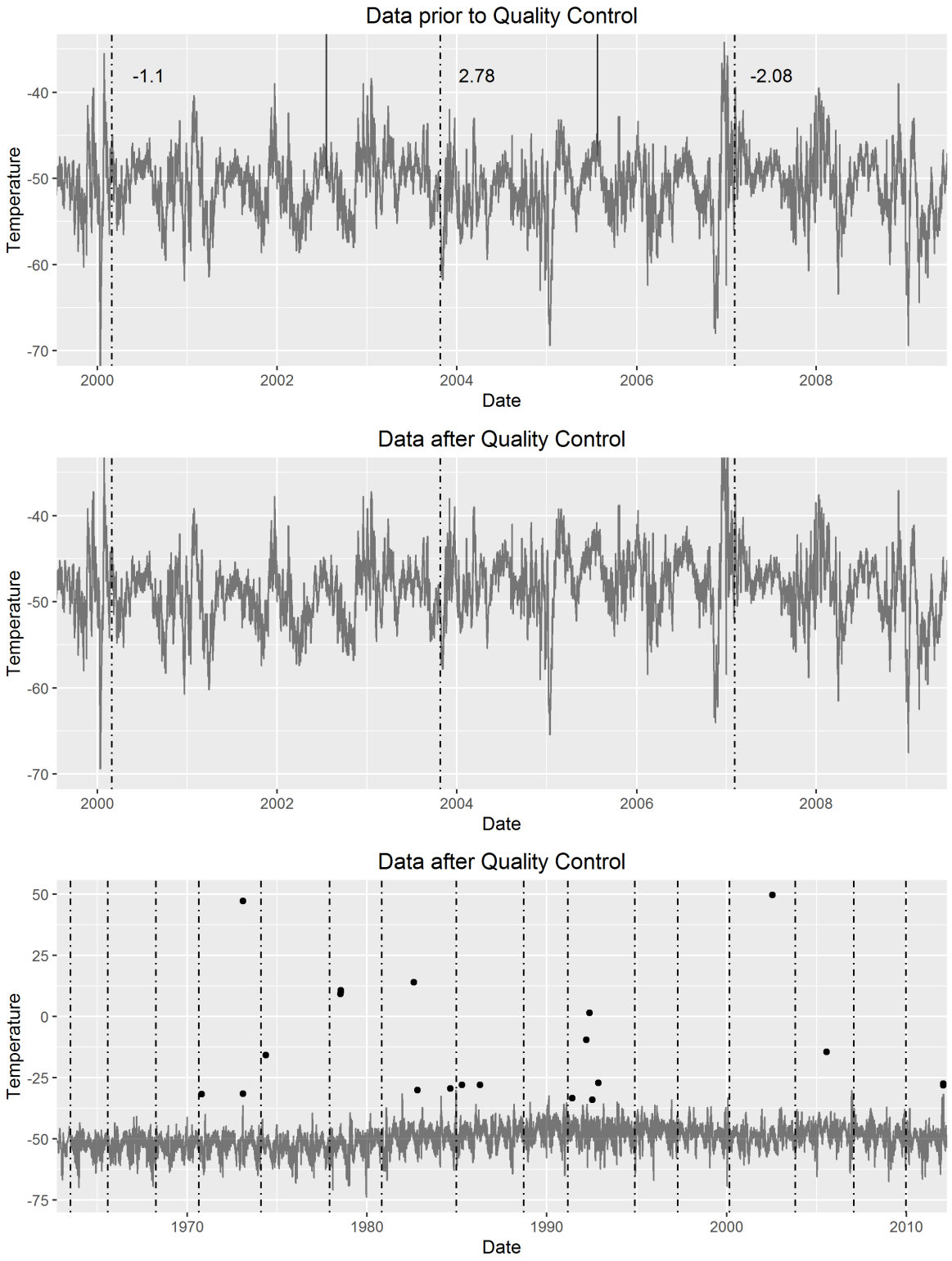

Figure 7: Time series plot of the radiosonde temperature data at station 70219 (Bethel, Alaska, USA). The top plot shows the data prior to the quality control algorithm, and the other two show the data after for a subset (middle) and the entire record (bottom). Dashed vertical lines indicate detected change points, numbers in the top panel indicate the magnitude of the corresponding change point, and dots indicate detected random errors.