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

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

September 17, 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 robust Standard Normal Homogeneity Test (SNHT) 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 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 [21] with over 32,000 stations, and the Integrated Global Radiosonde Archive (IGRA) housed at the National Climatic Data Center (NCDC) [5] 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.] We still need to find a good dataset to reference.** 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 [6, 7] and for use as an input to global reanalysis datasets [13, 14]. 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 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 [7, 9, 17, 18, 23**?**]. However, a study was recently conducted by the European Cooperation in Science and Technology to compare many different homogenization methods [23]. 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 [16, 19, 20, 22] and many of these approaches 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 [16], and binary segmentation [22]; also, 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 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 known 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 method" or simply "the sequence". In

Section 2, we discuss the details of our data simulation and contamination. Section 3 evaluates the homogenization algorithms we used, and Section 4 gives the results from the sequencing study. Finally, some conclusions are offered 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 Modelling 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?] I added a section where I describe this in more detail below. Does that layout make sense?** GAMs are flexible, non-parametric models that allow the response variable to be a linear combination of smoothed functions of the input variables [10]. 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 long term 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 the long term trend; and $s_1(\cdot)$ and $s_2(\cdot)$ are cubic regression splines.

Typically the error term in Equation (1) would be modelled 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 [3] **Is this the right reference? On page 17, it gives relationships between the first four moments and the parameters of the distribution, but I wouldn't say each parameter is related to only one of the moments.**. Thus, the distribution is very flexible and can handle 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 autocorrelation to be explained, so an AR(1) time series model should be **Maybe we should say "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 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, we must use only those observations that are h time steps apart:

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

Why don't we put the hat over the ϕ here? Isn't it an estimate? where $\phi(h)$ is the autocorrelation at time lag h, in hours, $\mathcal{P}_h$ is the set of all pairs of residuals that are h hours apart (or within some window), and $\widehat{\epsilon_i}$ is the error from Eq(1). For an AR(1) model, we need only estimate ϕ at $h = 12$ hours and we used a window of 5% of 12 hours, or 0.6 hours.

2.2 Data Simulation

The data simulation procedure has two parts: first, fitting a model to radiosonde temperature data, and second, simulating data according to this model:

Step 1: Fit the models described in section 2.1.

Step 2.a: 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.

Step 2.b: We use the GAM model fit based on Equation (1) to determine the expected value of temperature at the simulated time.

Step 2.c: To simulate the noise in the observations, we randomly draw values e_i from a skew-t distribution with parameters as fit in step 1.

Step 2.d: We wish to introduce autocorrelation in these e_i. Thus, we simulate an AR(1) model via

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

where ϵ_i is the simulated noise in the model at time i, and Δ_{i-1} is the time difference, in hours, between the $(i-1)$th and ith observation. Note that the kth term in this series will depend on all the previous $k-1$ values. To ensure bias is not introduced, we simulate 1,000 more values than we need and discard the first 1,000.

Step 2.e: 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 a random error following a distribution of $N(10\sigma, 1\sigma^2)$, where σ is the standard deviation of the simulated series. Systematic errors are generated by sampling 1, 2, or 3 change points 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 future comparison purposes.

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, and so we analyze data from many different climate zones. In [2], 10 representative stations are chosen and analyzed from the ten different climate types, and we analyze these 10 stations.

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

Sample Size: For the sequencing study, sample sizes of twice-daily data simulated for 20, 40, and 80 years were used. The study comparing homogenization algorithms, however, was much more computationally expensive, and so we used 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: 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 [1, 4, 8, 9, 18, 19, 20, 23]. 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 [9], 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 compare the abilities of four different homogenization algorithms to detect systematic errors when random errors are also present in the data. We consider Binary Segmentation (BinSeg) [22], Pruned Exact Linear Time (PELT) [16], SNHT, and we propose a robust SNHT. We simulate data as described in section 2 and then introduce changepoints and random errors, and we evaluate the ability of the algorithms to detect the known changepoints.

3.1 Methodology

Two algorithms, BinSeg and PELT, 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 τ_i is the ith changepoint; 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 $(i-1)$ and ith changepoint; and $\beta f(m)$ is a penalty term on the number of changepoints to prevent overfitting [16]. Note that, for notational convenience, τ_{m+1} is defined to be the last observation. Often, $\mathcal{C}$ is chosen to be twice the negative log likelihood, and $f(\cdot)$ is linear.

Optimization of Eq(3) can be done in several ways. BinSeg uses 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 Eq(3), but it computes the exact minimum. It proceeds recursively as follows: first, the optimal number and location of changepoints is determined for observations 1 and 2 only. The optimal number and location of changepoints for the first three observations is then determined using this information, and more generally the optimal number and location of changepoints 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 [15].

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, no test is performed. If the largest T_t exceeds some threshold at time $t = t^*$, 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 all t that are sufficiently far away from the current changepoints, $t \in \{1, \ldots, n\} \setminus \{t^* - k, \ldots, t^* + k\}$, and the test is performed again until no T_t exceed the threshold, and we use $k = N$. In [9], 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 the Huber M-estimator of the mean and standard deviation [11]. These estimators are computed as follows:

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

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

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

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

2. y_i is defined as

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

3. Updated estimates of $\hat{\mu}$ and s 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 $10^{-6}s$.

This definition forces unusually large observations to have little to no influence on the estimators of the mean and standard deviation. 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.

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 [4]. 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 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

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

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, and 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 contain 10, 20, or 40 years worth of simulated data, and changepoints locations are simulated uniformly at random across the entire time series minus the first and last year. We simulate an average of one, two, or three changepoints per decade. 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 $N = 730$.

Figure 2 depicts a boxplot of the efficiencies measured for each of the different algorithms across all 30,000 simulations (1,000 simulations for each pressure level/Station combination). 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 changepoints. However, in practice, we will not know the true number of changepoints, and the BinSeg algorithm is very sensitive to this choice. The PELT algorithm appears to perform best with a penalty of n, but its performance is much worse than the alternative algorithms.

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 used were the sample size n, the outlier contamination rate, the station, the pressure level, and the homogenization algorithm. We fit 5 different logistic regression models: one with linear response terms and models with k-way interactions, where $k = 2, 3, 4, 5$. Table 1 reports the deviance for each model. As the deviance does not appear to decrease after k increases beyond 2, we chose a model with 2-way interaction terms only.

Table 2 displays a table with the average fitted probability as a function of n, outlier contamination, and the homogenization model. 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 attained the highest fitted efficiency. The robust SNHT 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%). Also, a longer period for the robust SNHT should be used when more data is available, as the robust-365 tends to perform best for the 10 and 20 year data but the robust-730 performs best for the 40 year data. Therefore, we use the robust SNHT for the remainder of this paper with $N = 365$.

Plots of the fitted probability that efficiency is positive are given in Figure 3. The center of each errorbar is the mean of the fitted probability over all stations and pressure levels, and the max (min) of the errorbar is the highest (lowest) fitted efficiency over all station and pressure level combinations. As seen previously, this plot shows that the SNHT and robust SNHT attain the highest fitted probabilities in almost all cases. Interestingly, the efficiency seems to improve as the outlier contamination rate increases, at least for $N = 40$.

4 Sequencing Study

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

control methods.

4.1 Random Error Detection

We follow the 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 measures of scale for both the left and right sides of the distribution [12]. 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 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 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 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.

We hypothesize that many random errors will not be detected if the data is not homogenized and that the homogenization procedure will not perform as well if random errors are not first removed. To alleviate this problem, we consider 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 data as described in section 2.2.

2. Apply each sequencing of the quality control process.

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 is defined as the percent of detected errors within the observations that are random errors, and the false positive rate is defined as the percent of detected errors within the observations that are not random errors.

4. Repeat steps 1-3 1,000 times for each climate zone and pressure level.

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 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 n, outlier contamination rate, station, pressure level, and sequencing.

We begin by fitting five logistic models, one with linear terms and models with all k-way interaction terms, where $k = 2, 3, 4, 5$. The deviances obtained are given in tables 3, 5, and 7. We again 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 all three responses.

For the TPR model, we find that there is no significant difference between the sequencings "Ran-Sys-Ran", "Sys-Ran", and "Sys-Ran-Sys." Thus, those three levels are grouped into one level termed "Other." Table 4 shows the fitted TPR averaged across all stations and pressure levels, and in parentheses shows the percent of the time when that model attained the highest fitted TPR. The table shows that the "Ran-Sys" sequencing performs best when the outlier

contamination rate is small and when the sample size is relatively small. However, we are more interested in cases where the outlier contamination is high, and the other sequencings perform best in those scenarios.

For the FPR model, we found no significant difference between the sequencings "Sys-Ran" and "Sys-Ran-Sys", and so those to levels were grouped into one level "Sys-Ran-*". Table 6 shows the fitted false positive rates averaged across all stations and pressure levels. In almost all simulations, "Ran-Sys" attained the lowest FPR; however, FPR is quite low across all models. Due to this fact, and the conclusions from the TPR model, we suggest one of the sequencings "Ran-Sys-Ran", "Sys-Ran", or "Sys-Ran-Sys" if the end goal is only optimizing FPR and TPR.

Lastly, results from fitting the efficiency model are shown in table 8. The largest fitted efficiency is almost always obtained with the sequence "Sys-Ran-Sys". However, as shown in figure 7, the difference between the four sequencings is not statistically significant. Thus, we conclude that the sequencing chosen does not appear to have any effect on the efficiency of the final homogenized data.

5 Conclusion

In this study we have evaluated several different homogenization techniques, and we found that the robust SNHT method performs well. It attains a high efficiency, indicating that this method is reasonably effective at returning the data to its uncontaminated state. It attains higher efficiencies then the BinSeg and PELT algorithms, and it outperforms the SNHT when the outlier contamination rate is larger than 1%. The optimal period appears to be a function of the size of the dataset, and should increase as more data is available.

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 in most cases. However, the removal of random errors first does not have a large influence on the detection of systematic errors. This is contrary to our initial assumptions, and goes against what is typically done in

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

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

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

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

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

[7] 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.

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

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

[10] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[11] P. Huber. *Robust Statistics*. John Wiley, New York, 1981.

[12] P. J. Huber. *Robust statistics*. Springer, 2011.

[13] 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.

[14] 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.

[15] 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/`.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

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

[23] V. K. Venema, O. Mestre, E. Aguilar, I. Auer, J. A. Guijarro, P. Domonkos, G. Vertacnik, T. Szentimrey, P. Stepanek, P. Zahradnicek, et al. Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8(1):89–115, 2012.

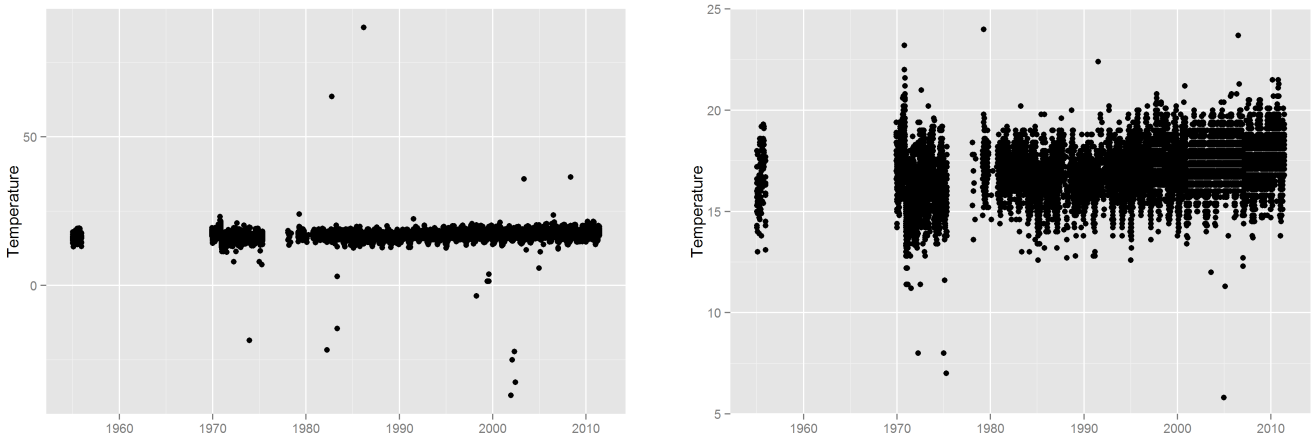

Figure 1: Temperature for Station **???** plotted over time. The second image is the same as the first but with a smaller y-axis window so as to show more detail.

Model Type	Deviance
Intercept Only	672836
Linear Terms	529612
2-Way Interactions	501158
3-Way Interactions	490553
4-Way Interactions	487125
5-Way Interactions	486493

Table 1: Deviance table for the efficiency logistic regression models.

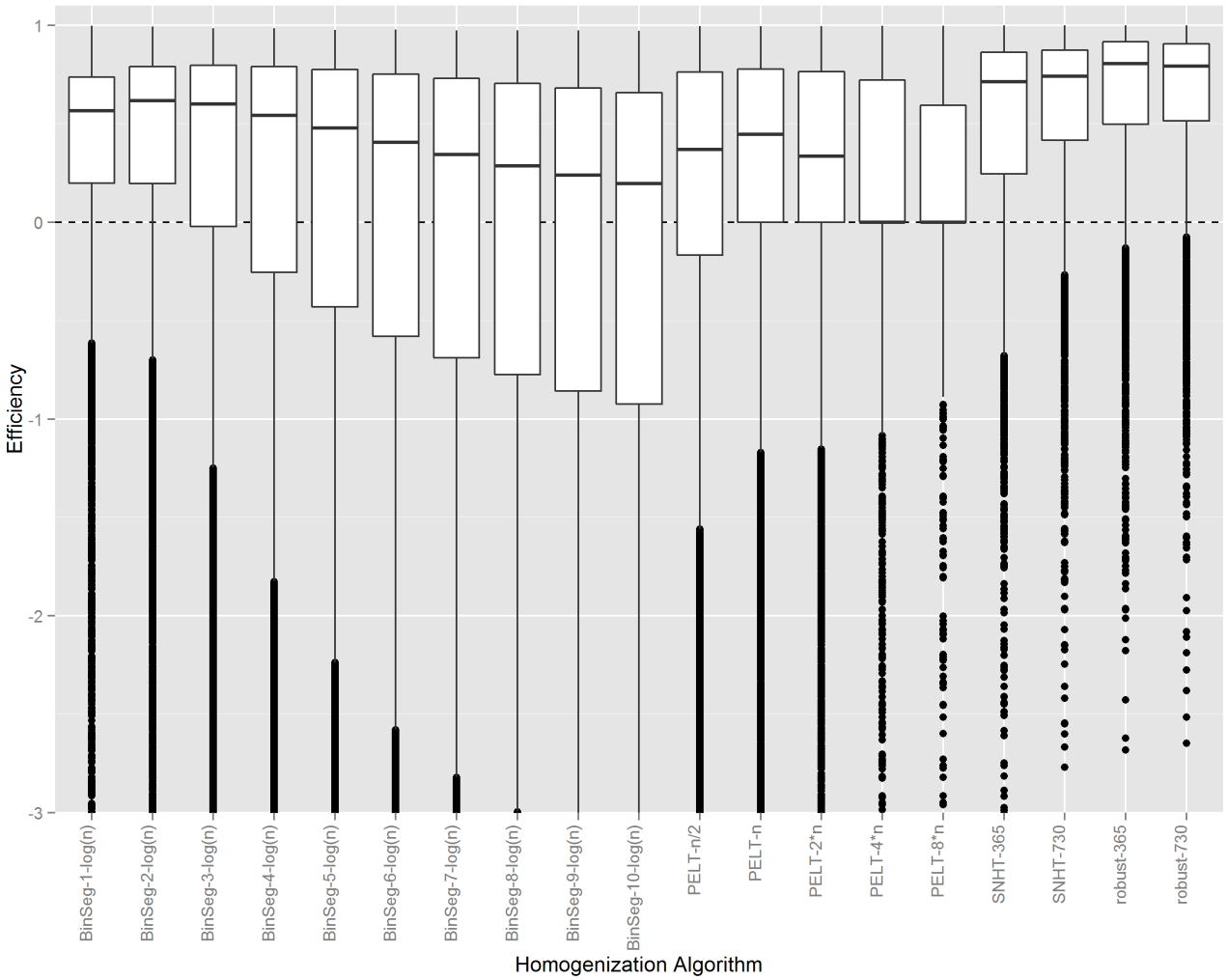

Figure 2: Boxplot of efficiency scores for the various homogenization algorithms. Note that this graph is constrained to the efficiency range of $(-3, 1)$ in order to show more detail. Note that the SNHT and the robust SNHT perform dramatically better than their alternatives.

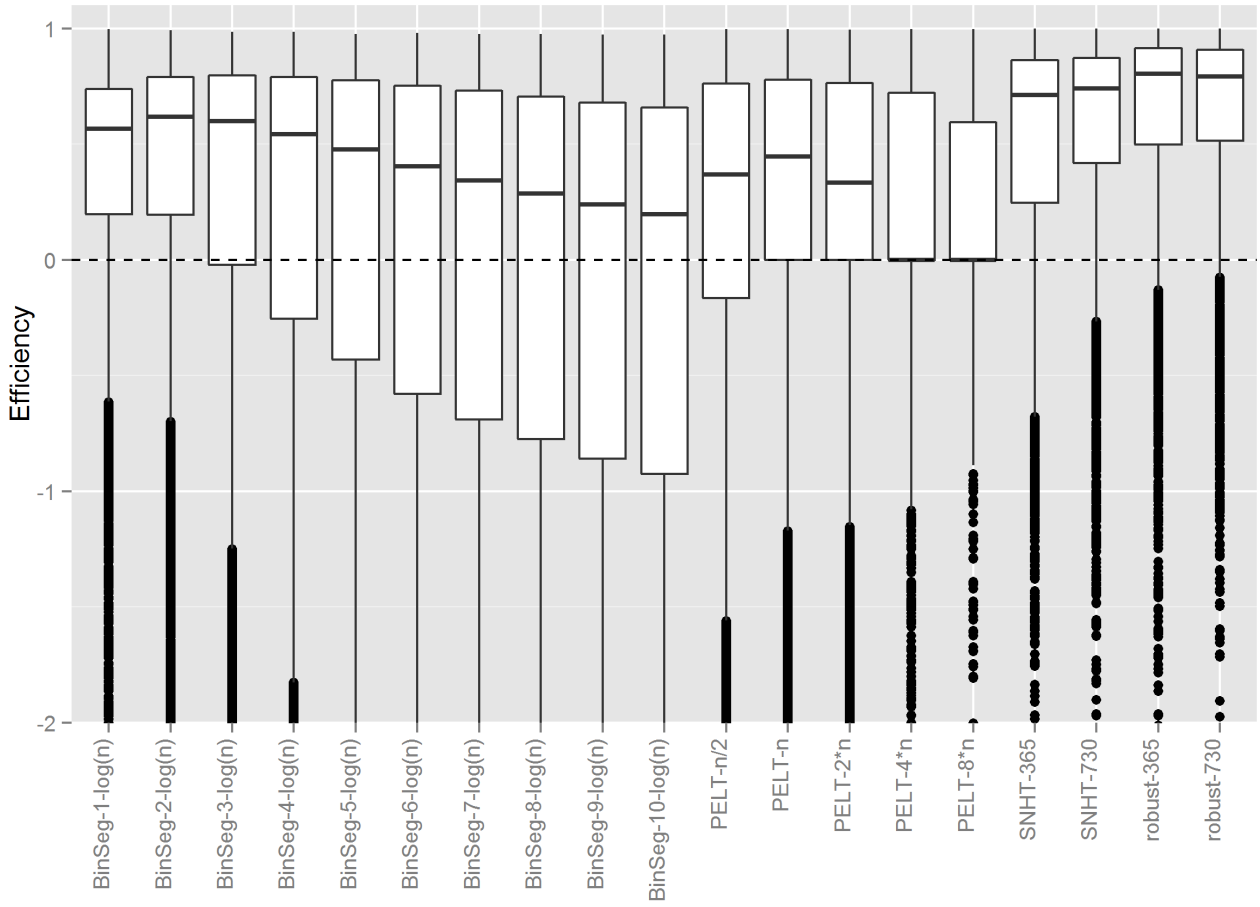

Figure 3: This graph depicts the estimated probability from the efficiency logistic regression model. The middle of each errorbar is the estimated probability averaged over all station and pressure level combinations. The maximum (minimum) of the errorbar is the highest (lowest) efficiency obtained across all station and pressure level combinations. Note that for clarity not all homogenization algorithms are plotted, but the ones left out performed worse than all the plotted algorithms.

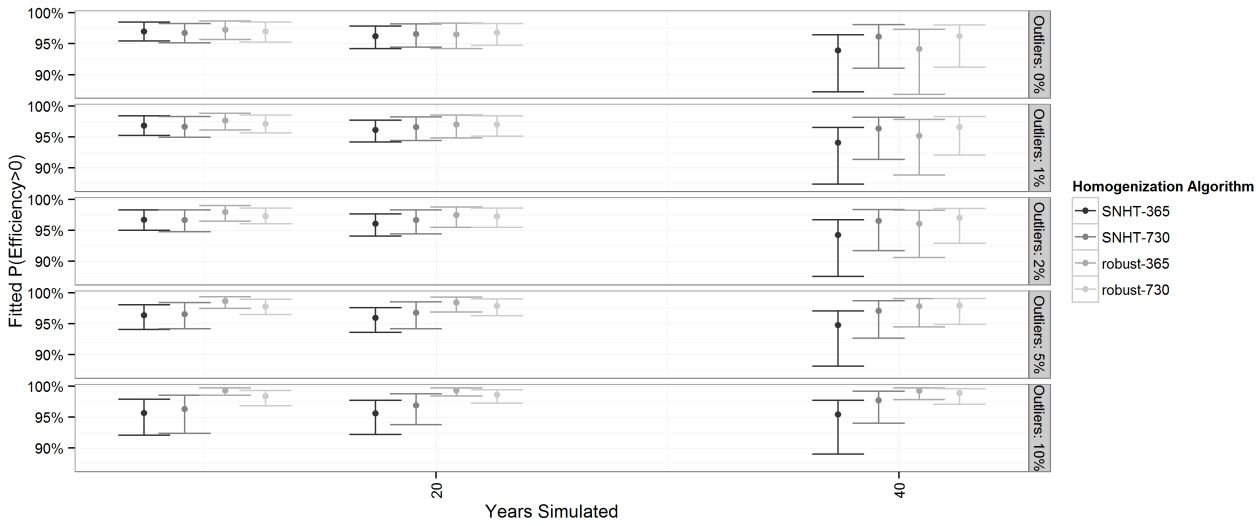

Figure 4: **Not sure if we need both this and the previous one, but I'm not sure which to pick either**. This graph depicts the estimated probability from the logistic regression model, averaged over all stations and pressure levels. It shows that the SNHT and robust SNHT outperform all other algorithms. It also shows that efficiency of these algorithms tend to improve as the outlier contamination rates increase. Note that for clarity not all homogenization algorithms are plotted, but the ones left out performed worse than all the plotted algorithms.

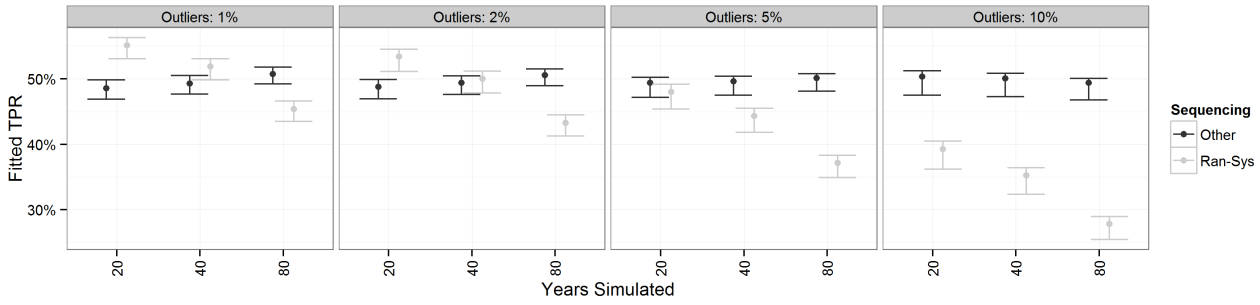

Figure 5: This graph depicts the estimated TPR from the logistic regression model. The middle of each errorbar is the estimated TPR averaged over all station and pressure level combinations. The maximum (minimum) of the errorbar is the highest (lowest) TPR obtained across all station and pressure level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

Outlier Contamination	Number of Years	BinSeg-3-log(n)	PELT-n/2	SNHT-365	SNHT-730	robust-365	robust-730
0%	10	46.8% (0%)	39.5% (0%)	96.9% (20%)	96.7% (3%)	**97.3% (70%)**	97.0% (7%)
1%	10	49.4% (0%)	41.4% (0%)	96.8% (3%)	96.7% (0%)	**97.6% (87%)**	97.1% (10%)
2%	10	52.0% (0%)	43.4% (0%)	96.7% (0%)	96.6% (0%)	**97.9% (97%)**	97.3% (3%)
5%	10	59.8% (0%)	49.4% (0%)	96.4% (0%)	96.5% (0%)	**98.6% (100%)**	97.8% (0%)
10%	10	71.9% (0%)	59.2% (0%)	95.6% (0%)	96.3% (0%)	**99.3% (100%)**	98.3% (0%)
0%	20	65.3% (0%)	50.3% (0%)	96.2% (0%)	96.6% (23%)	96.5% (23%)	**96.8% (53%)**
1%	20	68.2% (0%)	52.7% (0%)	96.1% (0%)	96.6% (0%)	**97.0% (50%)**	**97.0% (50%)**
2%	20	71.0% (0%)	55.2% (0%)	96.1% (0%)	96.6% (0%)	**97.5% (80%)**	97.3% (20%)
5%	20	78.4% (0%)	62.4% (0%)	95.9% (0%)	96.7% (0%)	**98.4% (100%)**	97.8% (0%)
10%	20	87.7% (0%)	73.1% (0%)	95.6% (0%)	96.9% (0%)	**99.3% (100%)**	98.6% (0%)
0%	40	90.1% (10%)	72.4% (0%)	93.9% (0%)	96.2% (43%)	94.2% (0%)	**96.2% (47%)**
1%	40	91.6% (10%)	75.2% (0%)	94.1% (0%)	96.4% (17%)	95.2% (0%)	**96.6% (73%)**
2%	40	92.9% (10%)	77.8% (3%)	94.3% (0%)	96.6% (0%)	96.0% (0%)	**97.0% (87%)**
5%	40	95.8% (10%)	84.5% (7%)	94.7% (0%)	97.1% (0%)	97.8% (27%)	**98.0% (57%)**
10%	40	98.3% (10%)	91.9% (3%)	95.4% (0%)	97.7% (0%)	**99.2% (87%)**	98.9% (0%)

Table 2: Fitted efficiency averaged over all station and pressure level combinations. Numbers in parentheses indicate the percent of station and pressure level combinations where the given model obtained the highest fitted efficiency.

Intercept Only	6240630
Linear Terms	4631181
2-Way Interactions	4101383
3-Way Interactions	4058965
4-Way Interactions	4049724
5-Way Interactions	4047592

Table 3: Deviance table for the TPR logistic regression models.

Outlier Contamination	Number of Years	Other	Ran-Sys
0%	20	48.4% (0%)	**56.9% (100%)**
1%	20	48.6% (0%)	**55.2% (100%)**
2%	20	48.8% (0%)	**53.4% (100%)**
5%	20	**49.4% (100%)**	48.0% (0%)
10%	20	**50.4% (100%)**	39.2% (0%)
0%	40	49.2% (0%)	**53.8% (100%)**
1%	40	49.3% (0%)	**51.9% (100%)**
2%	40	49.4% (17%)	**50.0% (83%)**
5%	40	**49.6% (100%)**	44.3% (0%)
10%	40	**50.0% (100%)**	35.2% (0%)
0%	80	**50.8% (100%)**	47.5% (0%)
1%	80	**50.7% (100%)**	45.4% (0%)
2%	80	**50.6% (100%)**	43.3% (0%)
5%	80	**50.1% (100%)**	37.1% (0%)
10%	80	**49.4% (100%)**	27.8% (0%)

Table 4: Fitted TPR 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 TPR.

Intercept Only	5835529
Linear Terms	1244423
2-Way Interactions	279232
3-Way Interactions	258216
4-Way Interactions	255145
5-Way Interactions	254845

Table 5: Deviance table for the FPR logistic regression models.

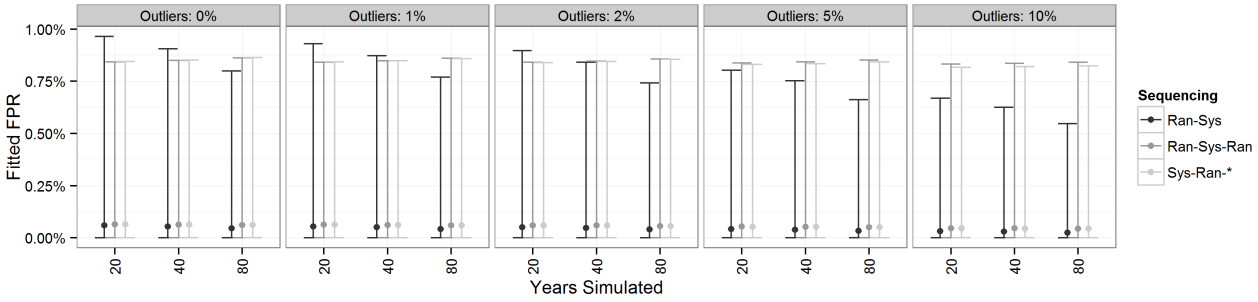

Figure 6: This graph depicts the estimated FPR from the logistic regression model. The middle of each errorbar is the estimated FPR averaged over all station and pressure level combinations. The maximum (minimum) of the errorbar is the highest (lowest) FPR obtained across all station and pressure level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

Outlier Contamination	Number of Years	Ran-Sys	Ran-Sys-Ran	Sys-Ran-*
0%	20	**0.059% (97%)**	0.065% (3%)	0.065% (0%)
1%	20	**0.055% (97%)**	0.063% (3%)	0.063% (0%)
2%	20	**0.052% (97%)**	0.060% (0%)	0.060% (3%)
5%	20	**0.043% (100%)**	0.054% (0%)	0.054% (0%)
10%	20	**0.032% (100%)**	0.046% (0%)	0.046% (0%)
0%	40	**0.054% (97%)**	0.064% (3%)	0.064% (0%)
1%	40	**0.051% (97%)**	0.061% (3%)	0.061% (0%)
2%	40	**0.048% (100%)**	0.059% (0%)	0.059% (0%)
5%	40	**0.039% (100%)**	0.053% (0%)	0.053% (0%)
10%	40	**0.030% (100%)**	0.046% (0%)	0.045% (0%)
0%	80	**0.046% (100%)**	0.061% (0%)	0.061% (0%)
1%	80	**0.043% (100%)**	0.059% (0%)	0.059% (0%)
2%	80	**0.040% (100%)**	0.057% (0%)	0.057% (0%)
5%	80	**0.034% (100%)**	0.051% (0%)	0.051% (0%)
10%	80	**0.025% (100%)**	0.044% (0%)	0.043% (0%)

Table 6: 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.

Intercept Only	159942
Linear Terms	136467
2-Way Interactions	132980
3-Way Interactions	132585
4-Way Interactions	132421
5-Way Interactions	132411

Table 7: Deviance table for the efficiency logistic regression models.

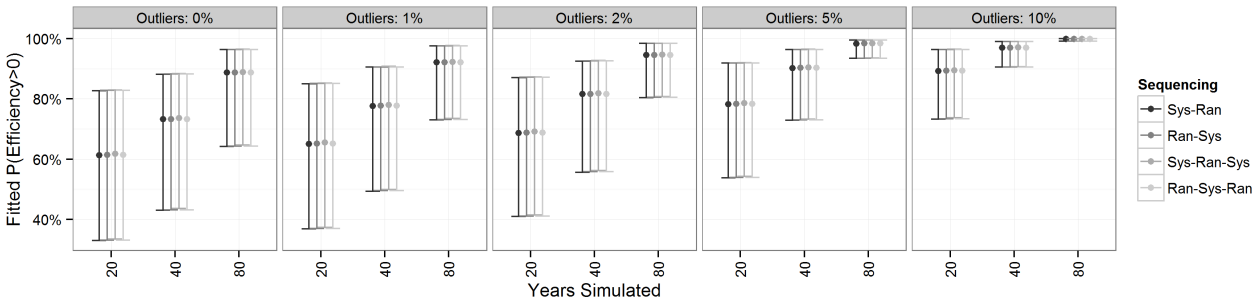

Figure 7: This graph depicts the estimated efficiency from the logistic regression model. The middle of each errorbar is the estimated efficiency averaged over all station and pressure level combinations. The maximum (minimum) of the errorbar is the highest (lowest) efficiency obtained across all station and pressure level combinations.

Outlier Contamination	Number of Years	Ran-Sys	Ran-Sys-Ran	Sys-Ran	Sys-Ran-Sys
0%	20	61.4% (0%)	61.4% (0%)	61.3% (0%)	**61.8% (100%)**
1%	20	65.2% (0%)	65.2% (0%)	65.1% (0%)	**65.6% (100%)**
2%	20	68.8% (0%)	68.8% (0%)	68.7% (0%)	**69.2% (100%)**
5%	20	78.4% (0%)	78.4% (0%)	78.3% (0%)	**78.6% (100%)**
10%	20	89.3% (0%)	89.3% (0%)	89.2% (0%)	**89.4% (100%)**
0%	40	73.3% (0%)	73.3% (0%)	73.2% (0%)	**73.6% (100%)**
1%	40	77.8% (0%)	77.8% (0%)	77.7% (0%)	**78.0% (100%)**
2%	40	81.7% (0%)	81.7% (0%)	81.6% (0%)	**81.9% (100%)**
5%	40	90.2% (0%)	90.2% (0%)	90.2% (0%)	**90.4% (100%)**
10%	40	97.0% (0%)	97.0% (0%)	97.0% (0%)	**97.0% (100%)**
0%	80	88.7% (0%)	88.7% (0%)	88.7% (0%)	**88.9% (100%)**
1%	80	92.1% (0%)	92.1% (0%)	92.1% (0%)	**92.2% (100%)**
2%	80	94.6% (0%)	94.6% (0%)	94.6% (0%)	**94.7% (100%)**
5%	80	98.3% (0%)	98.3% (0%)	98.3% (0%)	**98.4% (100%)**
10%	80	99.8% (0%)	99.8% (0%)	99.8% (0%)	**99.8% (100%)**

Table 8: Fitted 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 efficiency.