Time of Observation Error (TOBs) in temperature maxima can be reliably measured from real data (rather than estimated from models).

Ron House
University of Southern Queensland

Abstract

TOBs is a phenomenon concerning the time of day at which measurements are taken, whereby some maximum or minimum temperatures are not recorded; instead, a faulty, but always high (for maxima) or low (for minima), value is recorded from the 'detritus' of a more extreme value the previous day. This paper explains why such a phenomenon should leave a detectable signature in the statistics of maximum and minimum temperature changes from day-to-day. The entire US unadjusted temperature data, over 200 million data points, is divided into yearly baskets and examined for average occurrences of certain day-to-day temperature change patterns whose probability and/or magnitude would be expected to change, if the TOBs hypothesis is true, under changes in measurement time of day at recording stations. Whereas official estimates of TOBs are made by inference from models or from pairwise homogenisation (a process with many severe critics, but beyond the scope of this paper), this paper obtains direct estimates of TOBs error in daily maxima (Tmax) from the real data, along with statistical reliability estimates. This method detects the systematic error that actually exists, rather than one inferred from modelling. We find that the official estimates of the errors due to TOBs are significantly over-estimated. We also assess the use of the same method to find the TOBs error in daily minima (Tmin).

1 Introduction

This paper will address the TOBs problem using a new analysis of actual historic data, but first we shall set the scene by reference to two main sources discussing the problem: a good general explanation by Zeke Hausfather (Hausfather (2015): Understanding Time of Observation Bias), who is a scientist on the Berkeley Earth climate science project, which is responsible for its own analysis of historic temperature data; the other is a paper referred to by Hausfather: Menne et al. *(2009).*

Hausfather's explanation captures the dilemma nicely:

> "Until the late 1950s the majority of stations in the U.S. record recorded temperatures in the late afternoon, generally between 5 and 7 PM. However, volunteer temperature observers were also asked to take precipitation measurements from rain gauges, and starting around 1960 the U.S. Weather Service requested that observers start taking their measurements in the

> morning (between 7 and 9 AM), as that would minimize the amount of evaporation from rain gauges and result in more accurate precipitation measurements. Between 1960 and today, the majority of stations switched from a late afternoon to an early morning observation time, resulting [in] a systemic change (and resulting bias) in temperature observations.
>
> "...[Weather stations] use what are called minimum-maximum thermometers that record both maximum and minimum temperatures between resets of the instrument. The time at which the instrument is reset and the measurements are written down in the observers logbook is referred to as the time of observation.
>
> "...At first glance, it would seem that the time of observation wouldn't matter at all. After all, the instrument is recording the minimum and maximum temperatures for a 24-hour period no matter what time of day you reset it. The reason that it matters, however, is that depending on the time of observation you will end up occasionally double counting either high or low days more than you should. For example, say that today is unusually warm, and that the temperature drops, say, 10 degrees F tomorrow. If you observe the temperature at 5 PM and reset the instrument, the temperature at 5:01 PM might be higher than any readings during the next day, but would still end up being counted as the high of the next day. Similarly, if you observe the temperature in the early morning, you end up occasionally double counting low temperatures. If you keep the time of observation constant over time, this won't make any different to the long-term station trends. If you change the observations times from afternoons to mornings, as occurred in the U.S., you change from occasionally double counting highs to occasionally double counting lows, resulting in a measurable bias."

Hausfather's Figure 2 shows the USHCN official historic adjustments to the record, reproduced here as Figure 1.

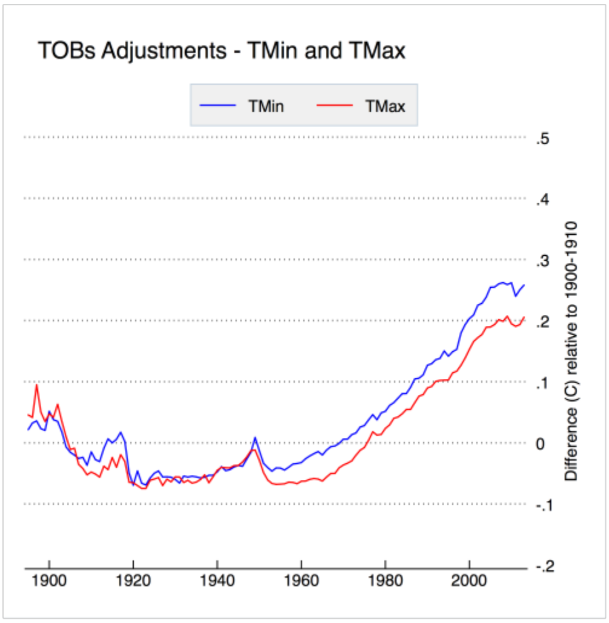

Figure 1: Net impact of TOBs adjustments on U.S. minimum and maximum temperatures via USHCN.

It is important to note that the official adjustments are inferred from models and theoretical assumptions. Even when performed by pairwise comparison of nearby stations, assumptions are still being made about how nearby stations should behave, in order to decide how to adjust data. In contrast, the method used here actually measures, in a statistical sense, the effect on almost all of the data that is susceptible to this distortion. The new methodology will be described in the following section.

One final general point needs to be made before proceeding:

TOBs error is *not* a "bias"!

A bias is an inaccurate measurement where the inaccuracy falls preferentially to one side. The TOBs effect is an error, not a bias.[1] True, it is an error that always has the same sign, but it is an error nonetheless. Considering afternoon measurements of maxima, sometimes the value we record today as today's maximum is actually a spurious value "left over" from yesterday, as Hausfather explains above. If today's real maximum is less than the leftover value from yesterday, yesterday's leftover will be recorded as today's maximum, regardless of how much today's true maximum falls below that number. Today's true value is not recorded with a biassed offset; it is lost entirely.

So it is one thing for Hausfather to present us with Figure 1, the "adjustments" to the historic temperature record, if all that is intended is to gain a better assessment of overall trend, since even an error might have a reliable average value. But it is quite another when the temperature recon-

[1] For the sake of simplicity I shall discuss the effect of evening measurements upon maxima, with the understanding that most of the discussion applies also to morning measurements upon minima, with the sign of the error reversed.

struction thus obtained is portrayed as if it is a more accurate representation of the past than the actual measured values. In particular, the highest of a series of maximum readings will *always* be unaffected by TOBs error. *All of those recorded high temperature records from the U.S. in past decades actually happened.*

We note that Christy et al. (2016) argue that Tmax is the most reliable estimator of global oceanic heat content. For this reason, and because Tmax can be analysed most simply, we shall now proceed to examine the situation with regard to maximum records.

2 Overview of Methodology

[Note: Throughout this paper, for simplicity, I shall call any measurement that is not affected by TOBs, "valid", without any implication that the measurement is valid on any other measure than freedom from TOBs. All locally-highest Tmax are valid. All locally lowest Tmin are valid. In the maxima, TOBs can only affect a measurement lower than that of the day before; in the minima, it only affects those that are higher.]

In the following we only consider the **temperature record for the maxima**.

Our goal is to detect and measure the signal from the TOBs error in actual historic maxima. In brief, the plan is:

1. Firstly we characterise the properties of those data that are susceptible to the error.
2. Then we identify a subset of the susceptible data that can be identified in the record regardless of whether the error actually was possible for a particular datum (i.e., it is possible if the reading was taken in the afternoon, but not if taken in the morning; keep in mind that we do not know the time of day any specific datum was recorded).
3. Lastly, on a year-by-year basis, we perform a linear regression analysis, for the identified data, of the mean size of the fall in temperature from a previous valid datum (dependent variable) against the percentage of readings in that year that were recorded in the afternoon (TOBs – independent variable). Luckily we have yearly data concerning how many stations in the U.S. operated in the afternoon, morning, or other times, as explained below. The slope of this regression (degrees C per percent afternoon measurements) allows us to calculate the actual historic TOBs effect for each year from 1895 to 2005, relative to our base year (2005). These may be directly compared with the official adjustments shown in Figure 1.

We note the following important facts:

1. All readings higher than the previous day's are valid; the only way to get a higher reading than the true one on any given day is for the previous day's higher temperature to push it up; therefore such a reading is today's reading, not detritus from yesterday.
2. Readings lower than yesterday's might or might not be affected.

3. Readings equal to yesterday's also might or might not be affected; but in this case, they can only be affected if the readings are taken at precisely the time of the maximum temperature. Since the U.S. record is largely the work of committed amateurs, such an occurrence, leading to an obvious error, seems unlikely—perhaps extremely unlikely. (Nevertheless, below we measure the frequency of this sad occurrence, and find it to be very low.)

Since we can identify certain data that are known to be valid (local highs), we can examine the sizes of the falls from there to the next day's lower reading. If TOBs is in effect, we would expect the average magnitude of the fall to be reduced, since some of those following day's readings will be invalid. This is shown in Figure 2.

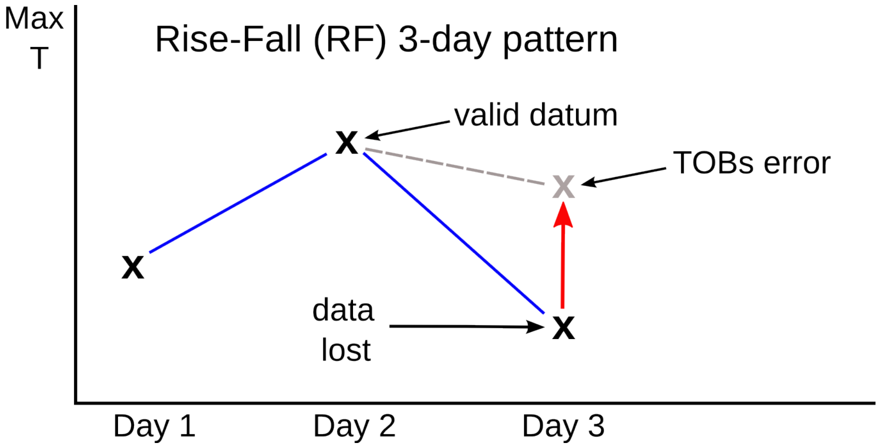

Figure 2: When TOBs is in effect, some data after local highs will be lost and a higher value recorded instead.

But knowing that a reading might be erroneous doesn't help us estimate the error. But as mentioned above we have yearly statistics for U.S. stations operating in the afternoon, morning, or other times. Therefore we can do a regression analysis of the proportion of sites observing in the afternoon (i.e. TOBs error is possible) *versus* the size of the drop in temperature from the valid datum to the following, possibly affected, datum. The relevant site data is provided by both Menne (his Fig. 3) and Hausfather (his Fig. 1), shown here as Figure 3.

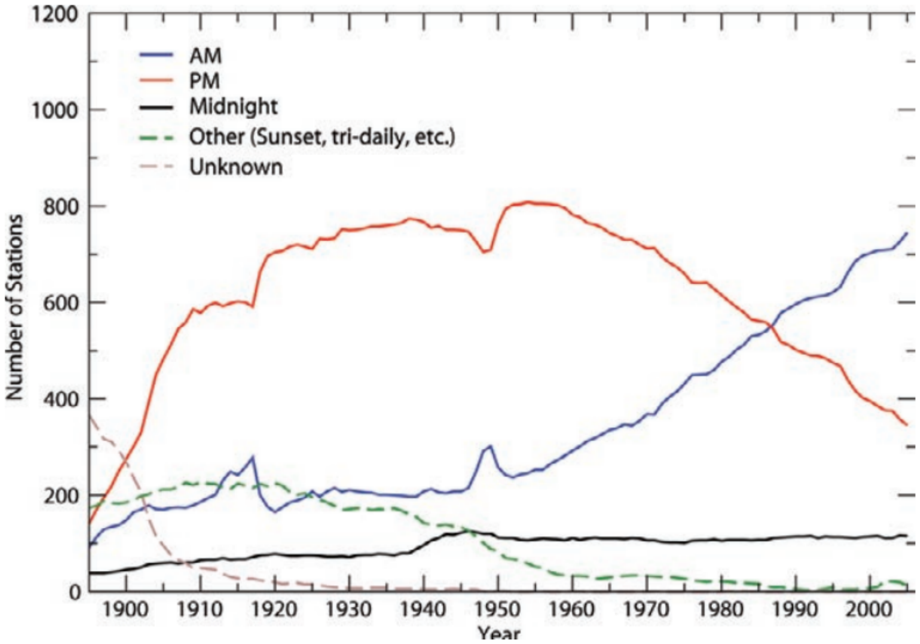

Figure 3: Number of stations using each observation time window, per year.

Unfortunately this diagram is not directly useful, since the size of any systematic error due to TOBs depends upon the proportion of stations measuring by each method, not the absolute number. Accordingly the data in this graph has been digitised and converted to percentages, as shown in Figure 4.

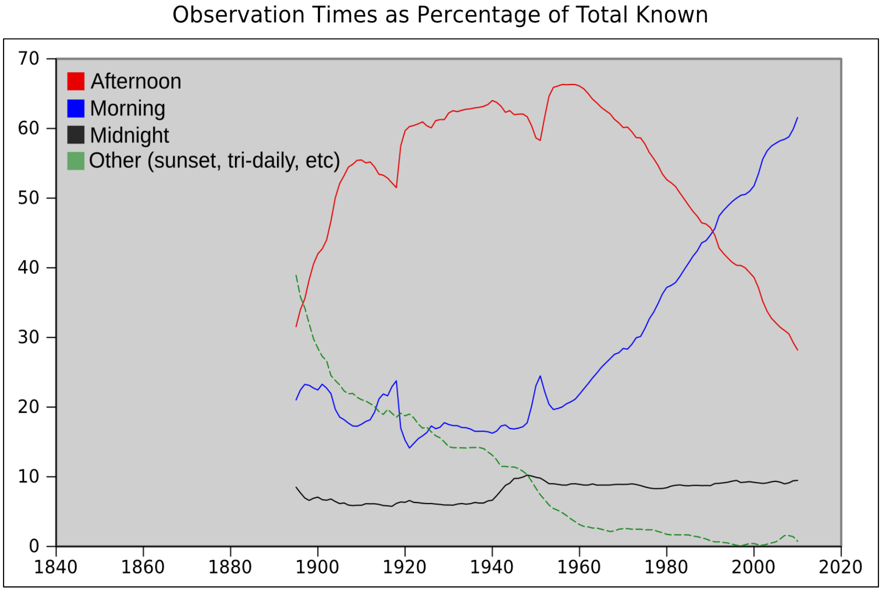

Figure 4: Observation times as percentages can be used in regression.

In this conversion, sites with unknown TOBs are ignored, which is equivalent to assuming that they are distributed in the same proportions as the known sites. (From about 1910 onwards, this assumption cannot introduce a large error because the number of unknown sites falls off dramatically.)

In outline, the method used here is to perform a linear regression of the mean size of the temperature drops in the days after a valid reading against the percentage of afternoon readings. We know that the true relationship, if any, must be linear because the error is present, yes or no, depending on whether a station does, or does not, make its recordings in the afternoon. This is not to say that other effects do not influence the average size of the error; for example, a station recording early in the afternoon would be expected to have greater or more errors than one recording late. But such effects should be random in relation to our independent variable (TOBs). The size of such effects should display on a residual plot as scatter around a straight line, whereas non-random effects should distort the shape of the line. In fact the results as highly satisfactory, as we shall see.

3 Analysis

3.1 Obtaining a reliable subset of the data

We explained the basic method in the previous section, but there are some pitfalls that we need to address in detail. Before we can perform regression, we have to ensure that we are using a reliable subset of the data; that is, a subset that will be present (except in circumstances too uncommon to significantly change the statistics) in both affected and unaffected stations. All local highs are valid, as we have seen, but in affected stations we might not recognise them all as the valid highs that they are, if the previous datum followed an even higher value, and was lifted, due to TOBs, above the following high. To understand this, consider the four consecutive daily readings shown in Figure 5.

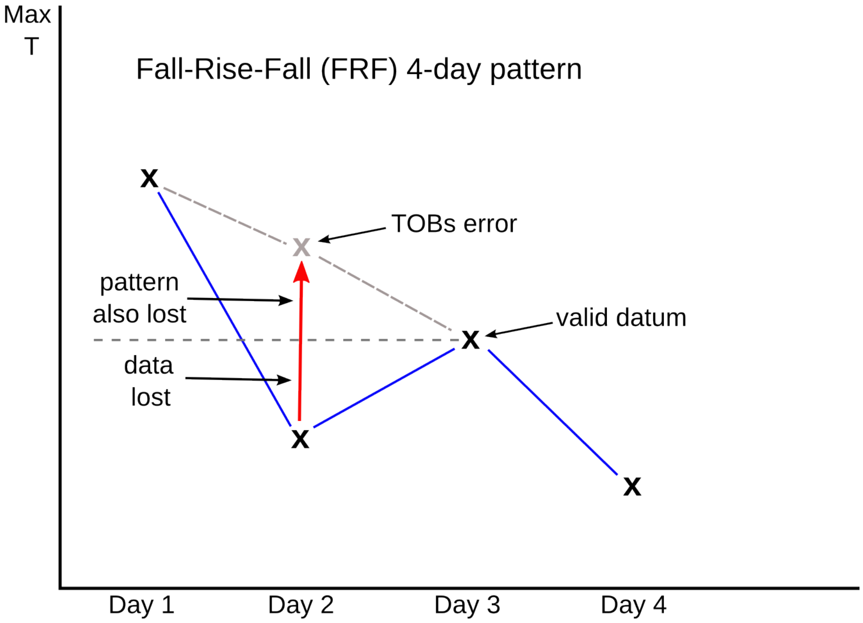

Figure 5: The FRF pattern identifies one valid datum amongst the daily maxima.

Datum 3 is valid. It cannot be affected by TOBs. The only question is, can we find it? If there is no TOBs error on datum 2, or if the error is small enough that datum 2 is still lower than datum 3 (the valid one), then the FRF pattern will remain in the data, and we can find datum 3 regardless of TOBs. But if the error lifts datum 2 above datum 3, the FRF pattern is transformed into FFF, and we now do not know that datum 3 is valid (even though it still is).

We can measure how critically TOBs robs us of the ability to spot valid highs by measuring the frequency of the FRF pattern shown above in the historic data. We can 'lose' an FRF pattern as shown above, but we can also gain one: an FRRF pattern can be converted into FFRF, which contains an FRF pattern. However, since FRF is four data long, and FFRF is five, FFRF will occur on average only half as often, meaning that when TOBs is active, there are twice as many losses from the set of FRF patterns as there are gains. If, therefore, TOBs raises data points in any significant number so as to alter the relationships of rises and falls, we will observe this by the simple measure of counting FRF patterns in the historic data. The result is shown in Figure 6.

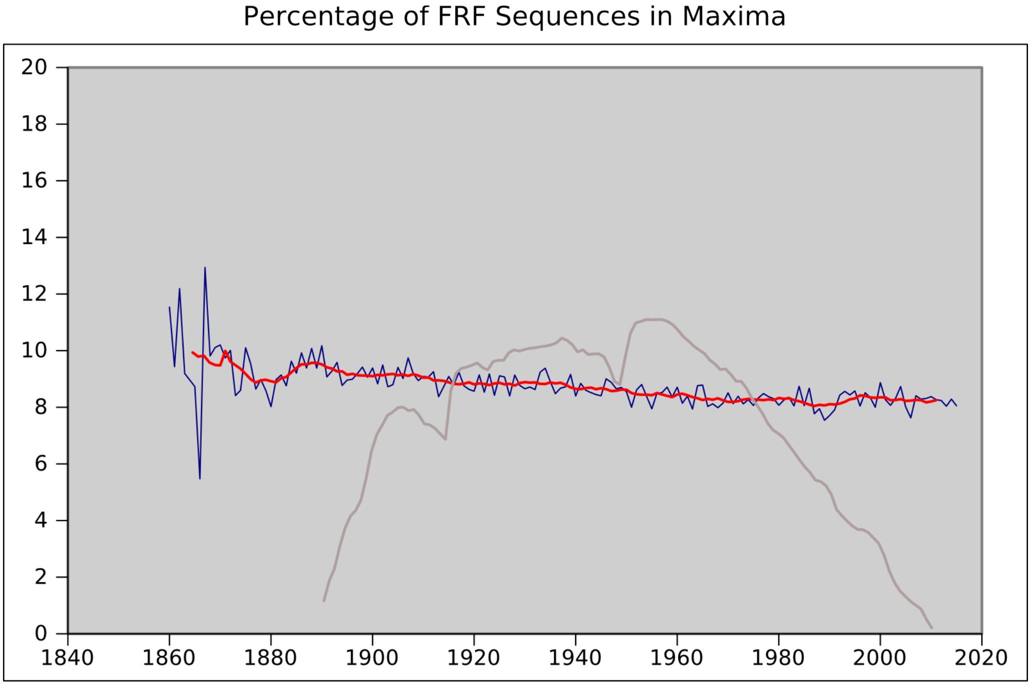

Figure 6: Percentage of FRF patterns in historic maximum data, with 9-year moving average; The shape of the graph of percentage of afternoon measurements is shown in grey for comparison.

Inspection of this Figure shows some fluctuation in the early years when there is little data (which is to be expected), then the line settles down to a very gentle decline over the next century or so. The slow decline is unexplained; in truly random data on such a massive dataset, we would expect small wiggles on a flat line. Instead, we get small wiggles on a slowly falling line. But what is crystal clear is that there is no visible effect that bears any relation to the proportion of afternoon measurements.

This may be verified with a regression check. The TOBs effect should preferentially reduce the frequency of this pattern, the higher the proportion of afternoon measurements. The regression slope of the above FRF percentage against afternoon measurement percentage is 0.00474781: it is very small, and in the wrong direction! Therefore the TOBs effect is insufficient to promote or evict any significant number of sequences into or out of the set of FRF patterns.

A related problem is as follows. Consider all patterns consisting simply of a fall (F); that is, a measurement followed by a lower measurement. This can be seen in days 2 and 3 of Figure 2, or days 1 and 2, or days 3 and 4, of Figure 5. If rises in the second datum are extreme enough to break the patterns of rises and falls in any significant number, then sometimes the second value will be the same as the first. This would happen if the thermometer was reset at the time of maximum itself. It would result in our not considering some sequences: we would be failing to include in our analysis measurements that were wrecked in the worst possible way. If we want to know if declines are reduced in magnitude due to TOBs, we cannot omit declines that decline all the way to zero!

However, if any significant number of declines are being counted as unchanged (I shall use S, “same”, for such patterns), then changes in time of observation should introduce a disturbance in the percentage of S patterns in the total data set. We have to remember with S patterns that there are other confounding factors. Observations are not real or rational numbers, they are quantised by the fineness of the recording apparatus, so issues such as Fahrenheit *vs* Celcius, digital *vs* analog, and thermometer markings, all confound a simple analysis of the S pattern. However, the results are shown in Figure 7.

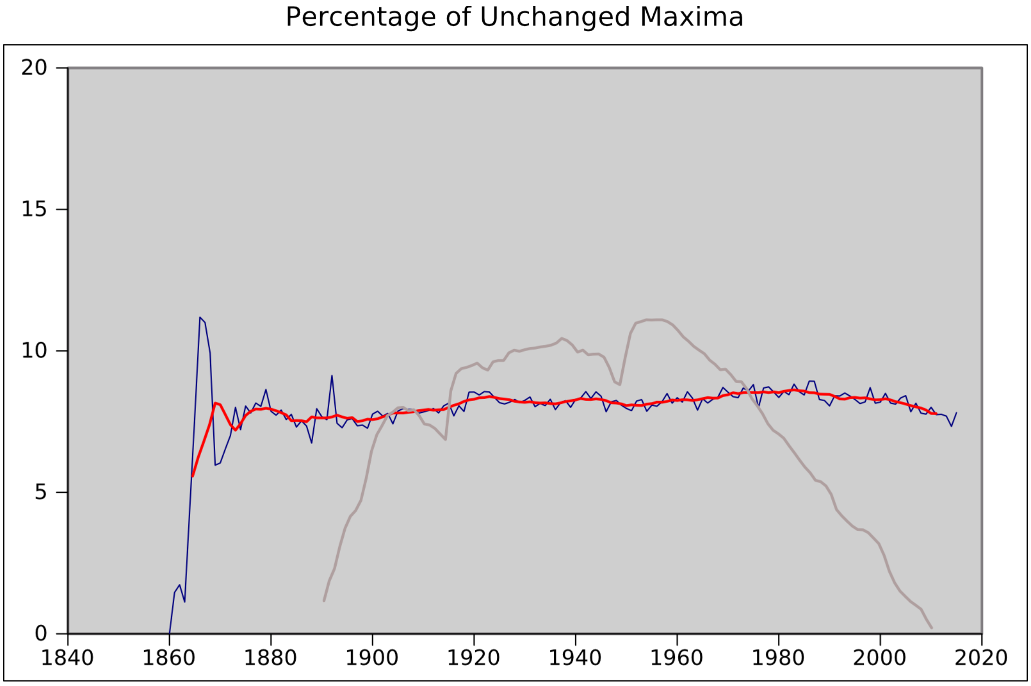

Figure 7: Percentage of observations with the same recorded maximum as the day before.

One again, random data would exhibit a flat line. The actual data, unlike the previous case, is somewhat more ‘active’, even in the years where vast numbers of observations are available to smooth the statistics. We might speculate about the reasons for the variation; for example, the slow fall in recent years could be due to the introduction of digital thermometers, which might be more precise (even if not more accurate). But such speculations are beside the point: we can test the dependence by performing a linear regression to find the dependency of the percentage of S sequences upon the percentage of afternoon readings. The regression line for S percentage, *s*, upon afternoon reading percentage, *a*, is:

$$s = 0.00822485a + 7.75539 \qquad (1)$$

The R^2 statistic for this regression is 0.0707484. Thus there is almost no dependency upon TOBs, and even this is swamped by other factors. To put this in context, the difference between *no* afternoon measurements and *all* afternoon measurements is just 0.8% more unchanged measurements. That is, almost no unchanged measurements are due to a TOBs error caused by resetting the thermometer at exactly the hottest time of day.

3.2 Analysis methodology

These results allow us to conclude that we can find a reliable subset of the falling data for analysis: almost all of it. To correct for the loss of some falls (as explained above), we can simply pad our figures with an appropriate number of zero-valued extra "falls". A more complex adjustment method would be unlikely to make much difference, due to the small fraction of lost falls under consideration.

We can collect statistics from various patterns within the data and perform linear regressions as follows:

- The independent variable is the percentage of afternoon observations in a calendar year;
- the dependent variable is the size of the fall of a (possibly compromised) maximum reading from an earlier day's valid maximum reading..

We collect "scenarios" from the data by looking for patterns in the data for daily maxima. To denote these patterns, we use the notation introduced earlier: "R" stands for a rise (a measurement following a lower measurement the day before); "F" for a fall; "S" for the same recorded value as the day before; and we also use lower case "n" to stand for "not": thus "nF" is "not a fall", and "nR" is "not a rise".

The simplest pattern suiting our purpose is the shortest: RFR, which is four data long. This collects those falls that have a valid datum immediately preceding, but are not followed by another fall. The reason we do not collect, in a single step, "just any" fall following a valid datum is that this is not a random data set; it is a physical system limited in an approximate sense by the maximum change that can occur between very high and very low values. If two falls occur in a row, we expect that they are likely to be individually smaller, on average, than a single fall, since the entire sequence must fit between physically realistic possibilities for the highest and lowest values. It turns out that this expectation is confirmed by examining the data itself. This is illustrated in Figure 8; note that this Figure is provided to illustrate the point only, the numbers plotted are not used in the analysis.

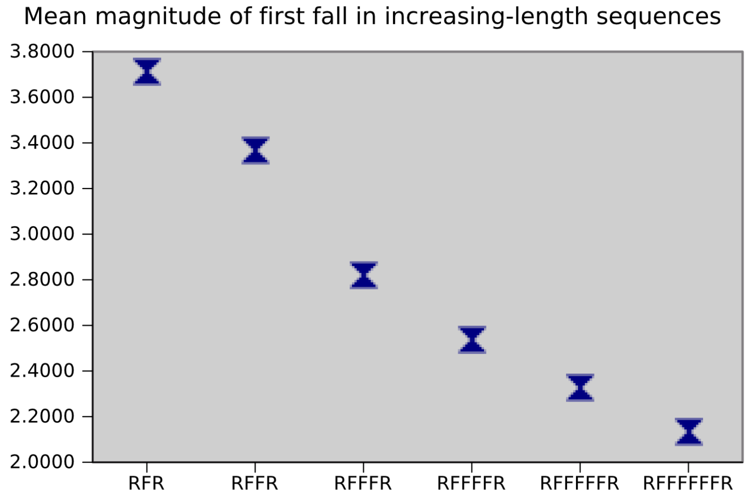

Figure 8: Mean magnitude in °C of fall in maximum temperature on the day after a valid reading, sorted by length of the falling sequence.

Similar results follow for the second and subsequent falls. The practical consequence of this is that we must perform a separate linear regression on each kind of fall: the first fall in a given sequence has different statistical properties from the second in the same sequence, etc., and the *n*th fall in one sequence is different from the *n*th fall in any other.

Choice of sequence. Another issue is which sequences we should use for the analysis. Our goals are, (a) to collect and analyse as many falls as we can dependably relate to a preceding valid value for statistical analysis; and (b) extract overall results of the highest reliability. These goals can conflict if collecting extra data of slightly less 'soundness' greatly increases the overall coverage.

Our complete dataset has 174 million data points. But the total of all the Fs, Rs, and Ses in the data does not equal 100% because the first datum can never be counted; every missing data point also introduces a datum (the next day) that cannot be identified as an R, F, or S. Our goal is to find as many useful Fs in the complete data set as possible. As it happens, we have two possible ways to perform our analysis.

(The first way) In the above we have used sequences RF...R, up to six falls in length. Very few sequences have length greater than this, so no useful purpose is served by including them. However, we can capture some of the (very few) even longer sequences by, on the longest pattern, omitting the final R; thus the six Fs of the longest sequence are in fact the first six Fs of sequences of six or more Fs. Specifically, our six patterns are RFR, RFFR, RFFFR, RFFFFR, RFFFFFR, and RFFFFFF. It turns out from complete analysis of the entire data set, that overall 43.57% of them are Fs. These six RF...R sequences pick up Fs amounting to 32.09%, or roughly 75% of the possible F data.

(The second way) Once again we use sequences up to six falls in length, but this time we use the nFF...nF sequences; that is, instead of demanding an R at the start and end, we allow an R or an S. These data sets include Fs summing to 41.99% of the total data, or over 96% of the total Fs.

(Most of the 3-odd percent not found are because of sequences that straddle missing data or the start or end of a data sequence; very few are from sequences longer than six Fs.) An S is, of course, a datum that is the same as the value the day before. Since these can be compromised by TOBs error, they are not all valid. Our reason for considering this method of analysis is the fact, discussed previously, that very few Ses are in practice introduced by TOBs. But perhaps the additional data collected makes it worthwhile. In similar manner to the first way, our final sequence omits the ending nF.

It turns out that the two methods give similar results; but we leave it to the reader to choose which they prefer. Each analysis is shown below.

3.3 Results – RF...R sequences

Our analysis covers the years 1895 to 2005 inclusive, as these are the years for which we have the TOBs breakdown. We show the effect of TOBs error as a ΔT difference from 2005.

The RF...R sequences account for around 75% of the possible F data, as described above. We consider six sequences, RFR ... RFFFFFR, RFFFFFF. Each F has different statistical properties, so we perform a total of 21 separate computations.

Each computation is a linear regression of the amount of fall from a previous higher valid datum. The first F after the initial R in each of the six sequences is straightforward: it is the fall from that previous valid value. For the final five sequences, the datum after the second F is subtracted, not from the previous datum, since that might not be valid, but from the same datum after the initial R as were the initial Fs, that is, two days previously. Likewise, the third F (occurring in the final four sequences) is subtracted once again from the previous valid high, but now this was three days previously; and so on for the fourth, fifth, and sixth Fs where these are present.

This is shown in the example Figure 9.

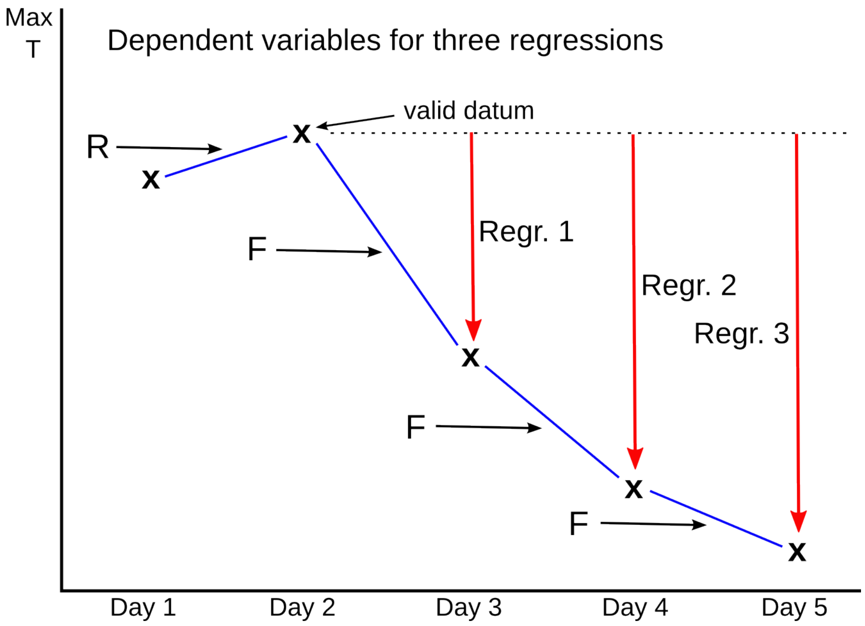

Figure 9: Dependent regression variables are all measured from the previous valid datum.

Table 1 shows the occurrence frequency, regression slope, and R-squared and p-values of each regression included in the assessment.

Table 1: Statistical properties of each fall (F) included in the calculation

Which F	**Freq. (%)**	**Regression slope**	**R-squared**	**p-value**
RFR(F1)	9.09	-0.017449	0.592	5.602E-23
RFFR(F1)	5.56	-0.010775	0.414	2.655E-14
RFFR(F2)	5.56	-0.008003	0.0643	7.259E-03
RFFFR(F1)	2.27	-0.008099	0.336	2.578E-11
RFFFR(F2)	2.27	-0.005970	0.0453	0.02497
RFFFR(F3)	2.27	-0.005366	0.0196	0.14270
RFFFFR(F1)	0.79	-0.004901	0.120	3.783E-04
RFFFFR(F2)	0.79	-0.003585	0.0147	0.20539
RFFFFR(F3)	0.79	-0.001525	0.00135	0.70235
RFFFFR(F4)	0.79	-0.001817	0.00127	0.71052
RFFFFFR(F1)	0.25	-0.000335	0.000405	0.83395
RFFFFFR(F2)	0.25	-0.000399	0.000132	0.90481
RFFFFFR(F3)	0.25	0.000257	2.86E-05	0.95554
RFFFFFR(F4)	0.25	0.000169	8.07E-06	0.97639
RFFFFFR(F5)	0.25	-0.001434	0.000419	0.83113
RFFFFFFR(F1)	0.11	0.000532	0.000779	0.77115
RFFFFFFR(F2)	0.11	0.003585	0.00781	0.35625
RFFFFFFR(F3)	0.11	0.007759	0.0198	0.14098
RFFFFFFR(F4)	0.11	0.011097	0.0272	0.08338

Which F	Freq. (%)	Regression slope	R-squared	p-value
RFFFFFFR(F5)	0.11	0.010531	0.0174	0.16761
RFFFFFFR(F6)	0.11	0.008710	0.00915	0.31791

The main features of the regressions shown here are:

- in general, as one progresses to longer sequences and to Fs further on in the same sequence, the regression slope decreases until it is no longer significantly different from zero (indicated by a large p-value—where 0.05 represents the familiar 95% confidence level);
- the R-value, which is interpreted as the fraction of the variation that is explained by the regression's independent parameter, also decreases—as it must, if TOBs is having a decreasing effect on the longer sequences.

3.3.1 Statistical validity

There are good reasons to believe that these regressions have reasonably well detected and measured the actual average effect upon the data of TOBs error. For the larger initial falls with larger amounts of data, the high R-squared and low p-values are as expected. These become problematic when the dependent variable really is indistinguishable from zero, and is lost in the noise. The TOBs error is becoming small at the same time as the noise is getting larger due to the smaller sample sizes. We can check, however, that what remains really is indistinguishable from random noise about zero, by plotting the residuals.

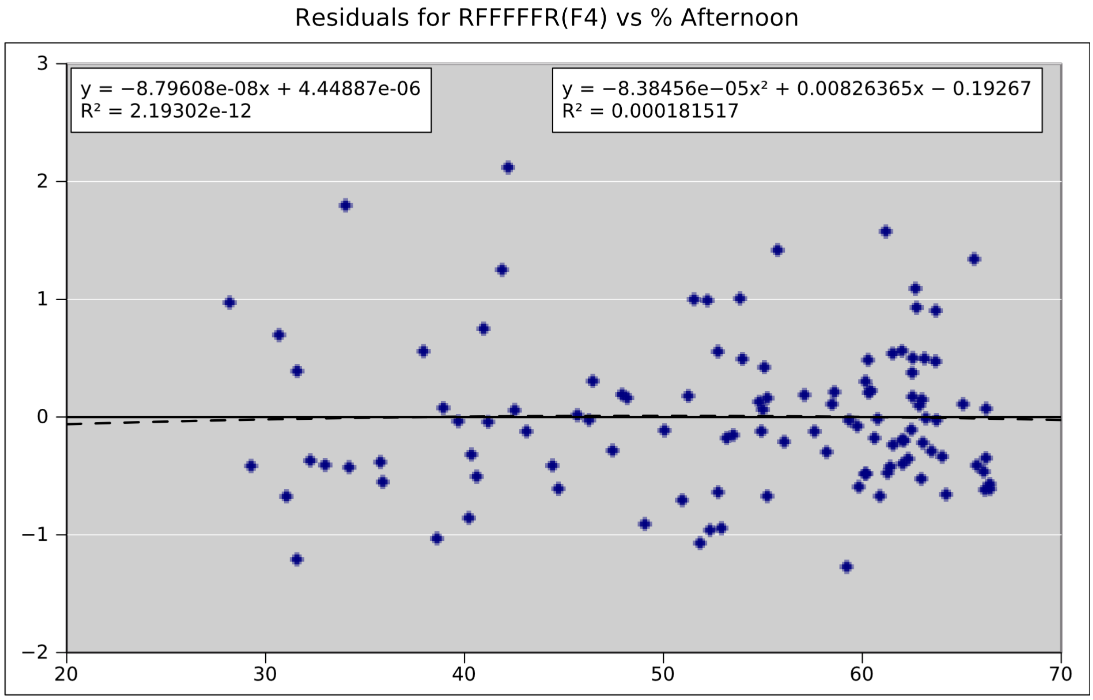

Figure 10: Residuals (°C) for the regression with the poorest statistical properties, RFFFFFR(F4).

Let us start with the very worst regression, as indicated by high p-value and low R-squared, namely RFFFFFR(F4). See Figure 10. The residuals, along with linear and quadratic fits, are shown. That these are both virtually flat indicates that the TOBs regression slope really does capture the entirety of the TOBs effect upon these data.

More residual plots are shown below, representing the individual regressions with the greatest impact upon our final answer. See Figures 11, 12, and 13.

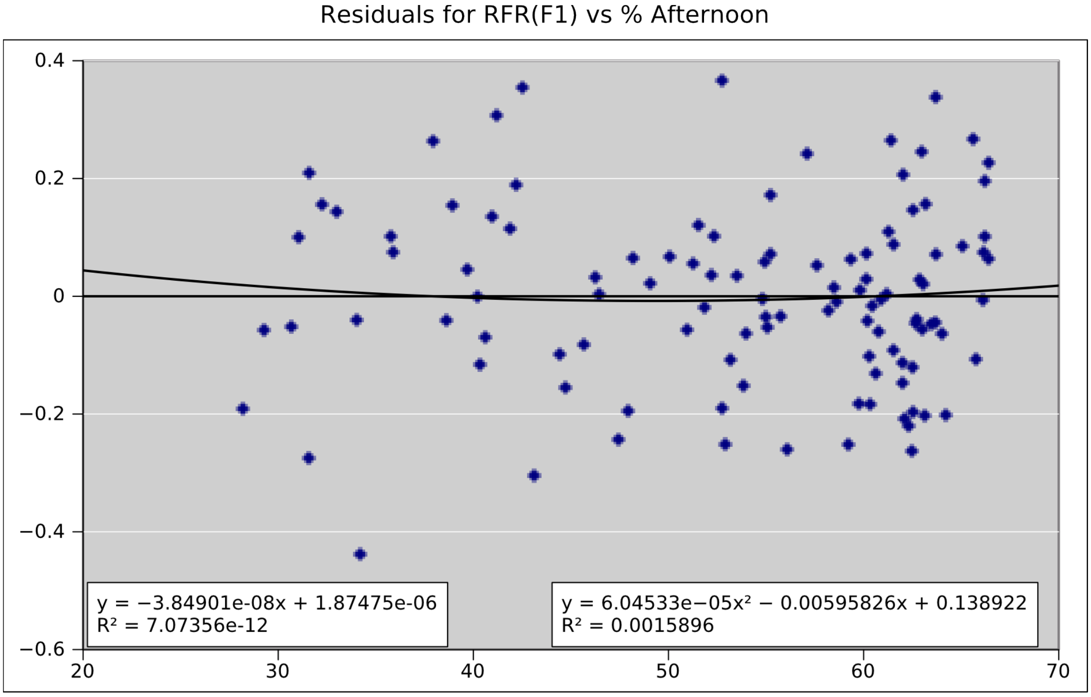

Figure 11: Residuals for the most common F, RFR(F1) (20.86% of all Fs, 9.09% of total data)

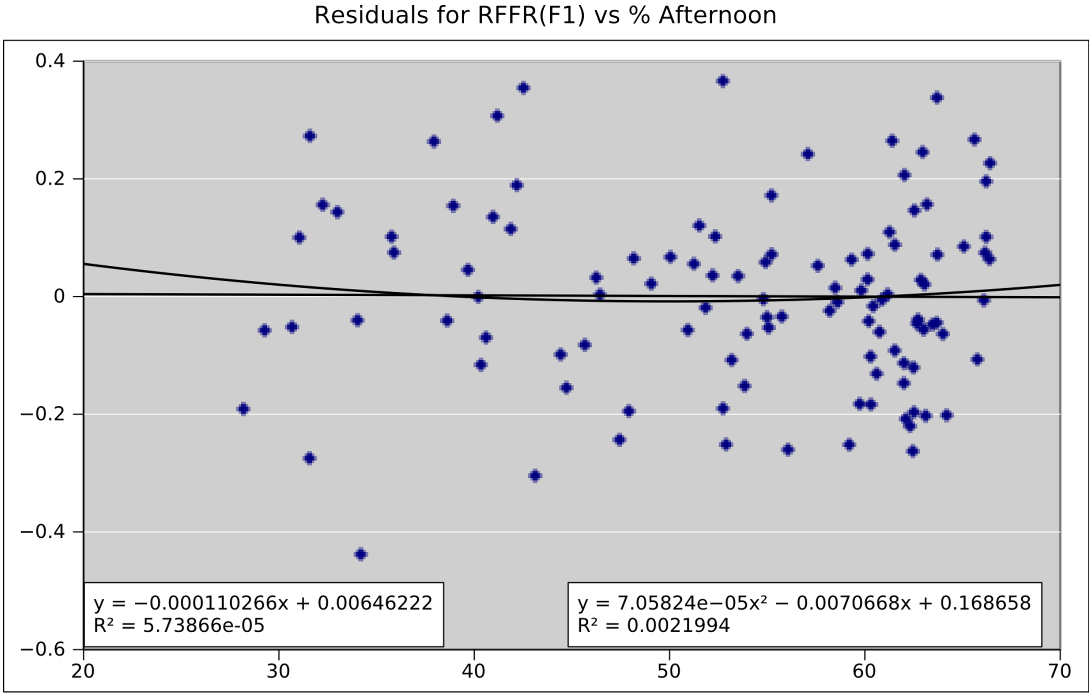

Figure 12: Residuals for the first F in RFFR (12.76% of all Fs, 5.56% of total data)

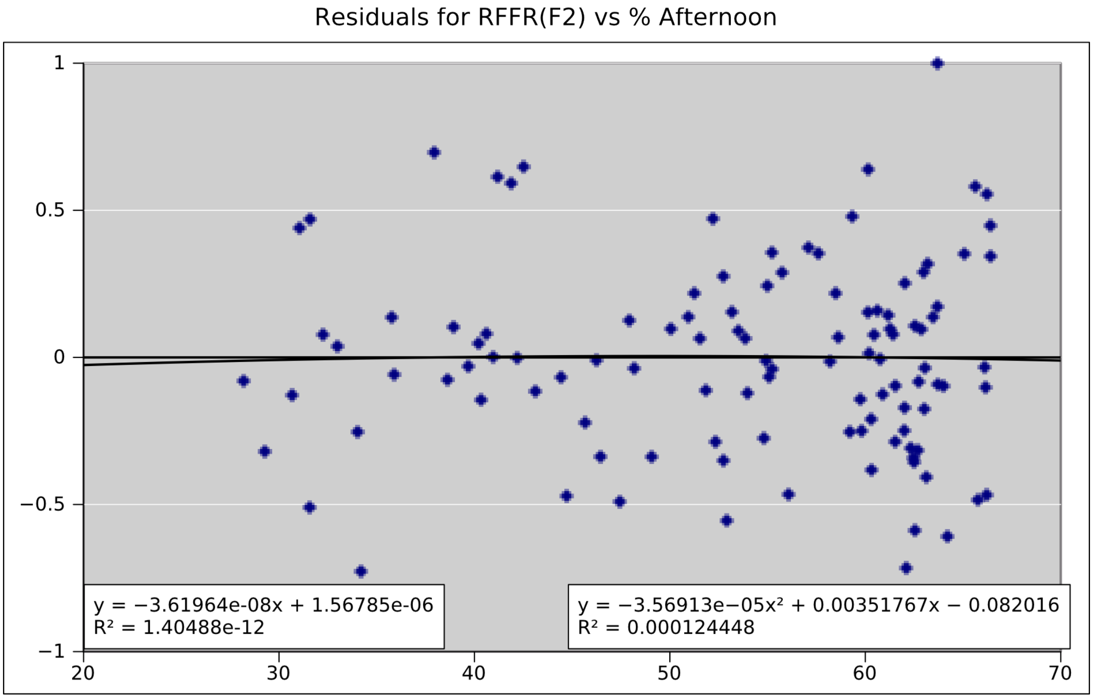

Figure 13: Residuals for 2nd F in RFFR (12.76% of all Fs, 5.56% of total data)

These regressions display all the qualities of a successful match of a dependent variable that is linearly related to its independent variable. We can have high confidence that no other dependence (such as a non-linear one) of size-of-fall upon percentage of afternoon readings exists. As noted previously, this is almost a necessity given that TOBs error is either possible (if readings are in the afternoon) or not (if they are not), barring unlikely scenarios such as year-dependent, system-wide systematic alterations of the precise timing of afternoon readings.

The final test of the validity of these regressions is the summation of all these separate regression slopes, weighted by their frequency, into a single overall result. This is shown in Figure 14.

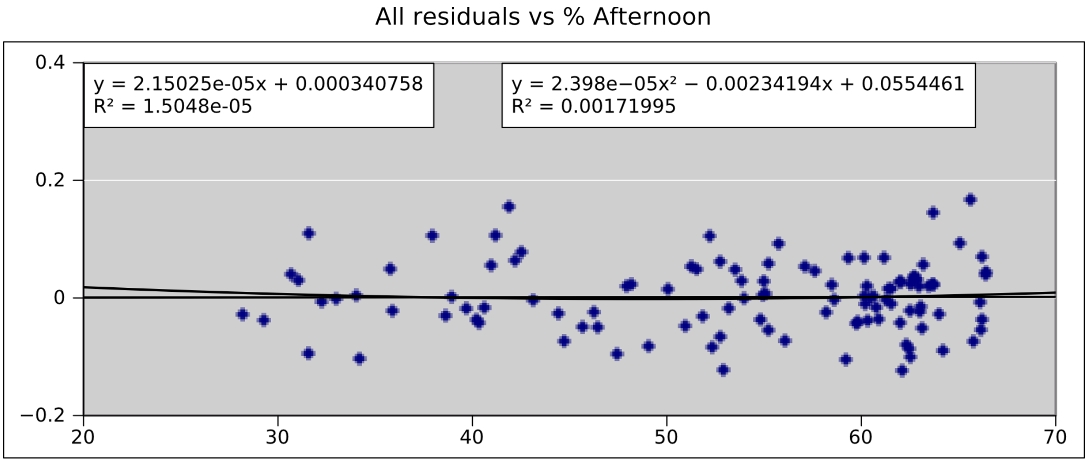

Figure 14: Residuals from the summed calculated TOBs effect due to afternoon readings.

The result is highly satisfactory. Once again, the outcome is as flat and random as could possibly be expected, and clusters closer to zero, as would be expected from the greater sample size.

3.3.2 RF...R Results

We have seen that the regressions available in the RF...R patterns may reliably be used to measure all, or virtually all, of the TOBs error in the falls that they make use of. We calculate as follows:

1. For each year 1895 – 2005:
    1. for each individual regression:
        1. multiply the regression slope by the percentage of afternoon measurements;
        2. multiply this by the fraction of the data that these Fs comprise;
    2. sum these values to give the effect of all Fs used in all regressions;
    3. assuming Fs not considered behave, on average, like those that were, multiply by the total percentage of Fs in this year’s data, and divide by the total percentage used in the regressions.
    4. subtract the value thus calculated for 2005, to obtain a zero-based offset.

A similar process is followed to obtain observed values for comparison, except in step 1.1.1 above, in which, for each individual regression, instead of using the regression slope times the afternoon percentage, we use the observed fall minus the intercept computed in the regression. This subtraction is necessary because each category of observed data has a very different magnitude for its average fall. (Recall that we are subtracting from the previous valid datum, not from the previous day’s datum.) Thus the magnitude of the first fall from a valid datum will be very different from, say, the fifth fall.

After processing the observed and computed values in these similar ways, the observed changes in fall magnitude are our observational “gold standard” against which the calculated TOBs error may be compared for goodness of fit. We can also include the USHCN modelled TOBs error shown in Figure 1. The three curves (observation, our calculated TOBs error, and USHCN modelled TOBs error) are shown in Figure 15.

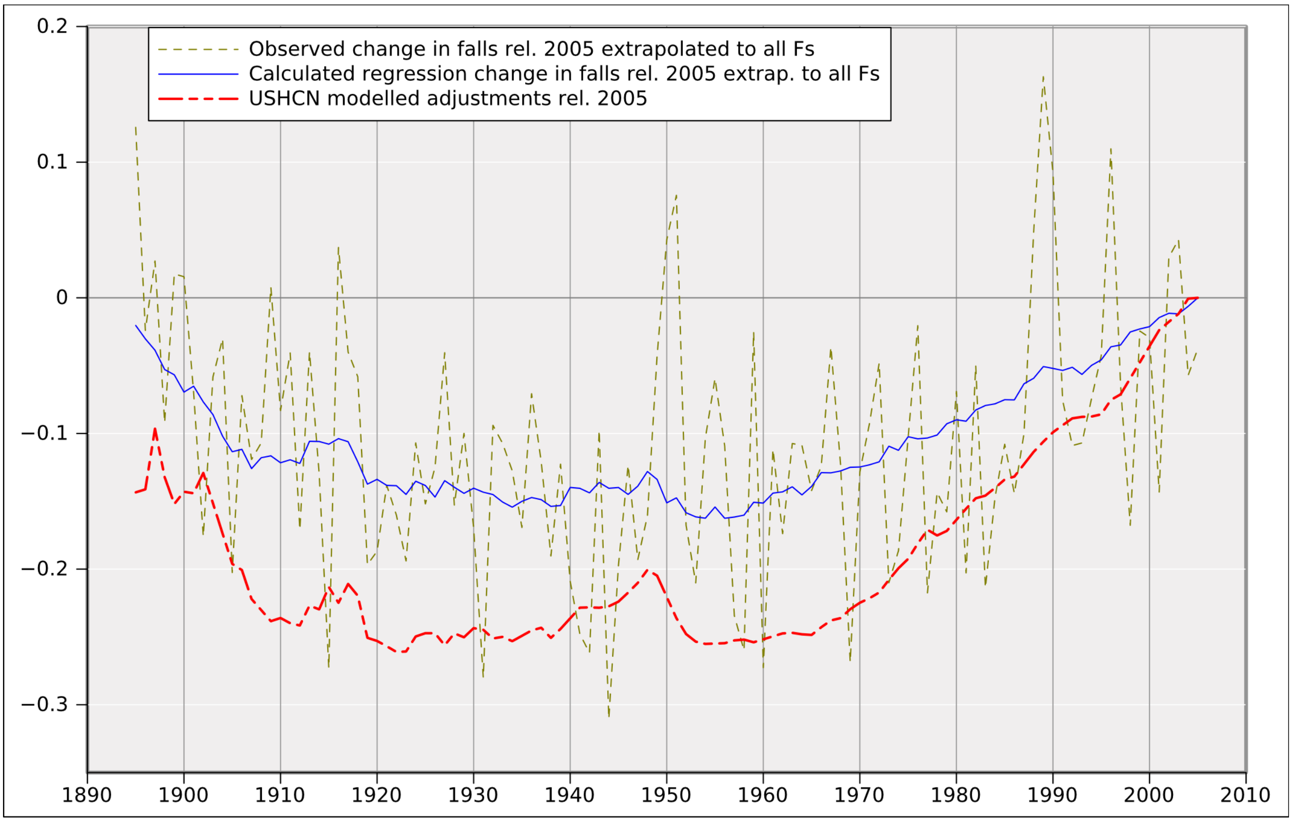

Figure 15: Detected TOBs error (°C) in RF...R sequences, USHCN adjustment, and observations

The overall USHCN modelled result is 177.9% of the effect measurable from the actual data; that is, corrections to temperature trends in official records should really be only 5/9ths of their current value.

3.3.3 Further check

It may seem surprising that the value recovered from actual data is so much less than that obtained from modelling. Hausfather explains that modelling proceeds by using the past several years of computerised records, where readings are available throughout the day, and testing various measurement times throughout the day. Without asserting any particular reason why this overestimates reality, it is worth noting again Figure 7, the percentage of observations with the same recorded maximum as the day before. This shows a distinct "phase change" starting around 1985, where an accelerating dropoff commences. As we contemplated earlier, this may be due to more precise measurements recently. It is hard to think of a reason, *in nature*, that might cause such a change in a huge amount of observational data.

An independent check that might help choose between the two contenders for the truth about the actual historic TOBs error is its linear nature. Barring random fluctuations, it is strictly linear in the percentage of affected stations. Therefore, the putative historic curve tracking its magnitude would be expected to closely follow the shape of the curve representing the percentage of afternoon observations. We can modify Figure 15 to remove the observational data, and include the percentage of afternoon observations (Figure 16).

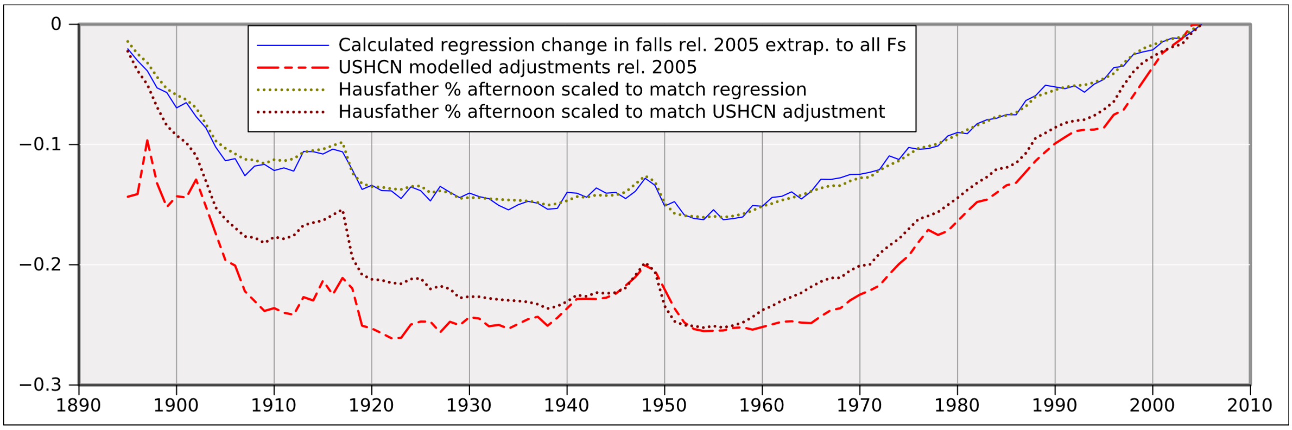

Figure 16: This paper's results and USHCN's modelling, each overlaid with scaled % afternoon

There is little to be said for Figure 16, except that our results do match the curve and USHCN's do not. It bears repeating: the effect of TOBs error is linear in the percentage of afternoon measurements. Our curve of measured TOBs error closely overlays the afternoon curve, with expected random errors, and bearing in mind we cannot collect measurements for all Fs. The USHCN curve is simply a different shape. Setting aside entirely its failure to match the data in magnitude, its shape shows that it cannot be a competent assessment of a linear effect. This conclusion follows from comparison of two USHCN profiles; it does not depend in any way on the calculations given in this paper.

3.4 Results – nFF...nF sequences

Whereas RF...R sequences capture almost 74% of all Fs, these sequences capture over 96% of the Fs, but as explained earlier, at the cost of a small possibility that the starting value is not actually valid. This means that a datum which is included in the nth fall after a valid datum might actually belong in the (n+1)th fall. As we have seen, the number of "faulty" S sequences is small, but some degree of error will be introduced by this process. Without repeating the explanations from the previous section, we first check the plot of overall residuals (Figure 17).

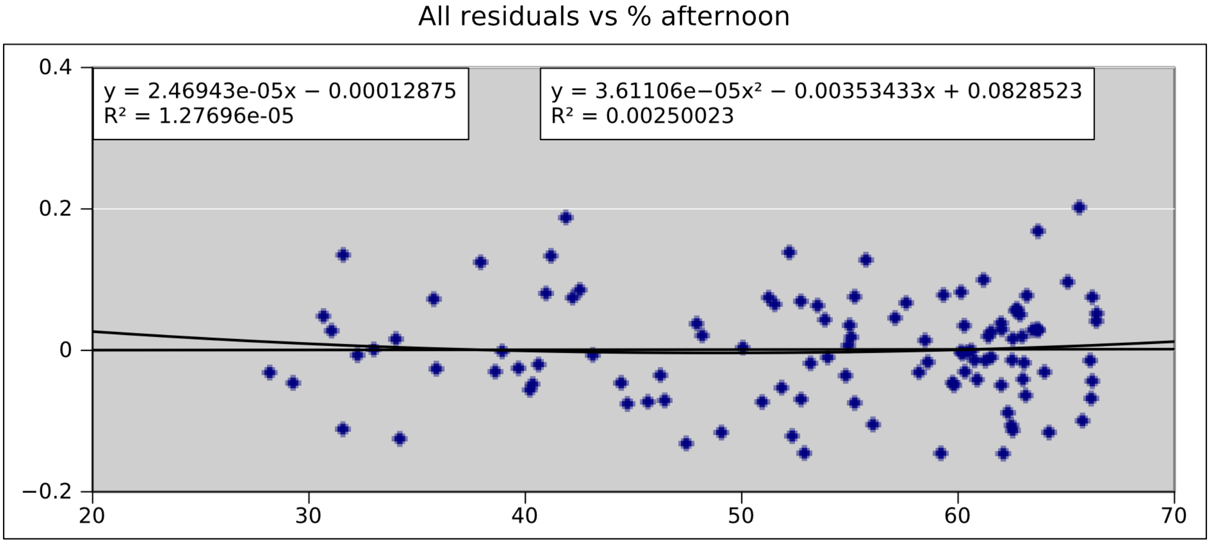

Figure 17: Summed residuals from the nFF...nF sequences

The residuals in this case resemble those from the RF...R sequences (Figure 14), but have a small amount of extra spread. The result for nFF...nF sequences, corresponding to Figure 15 in the RF...R sequences, is shown in

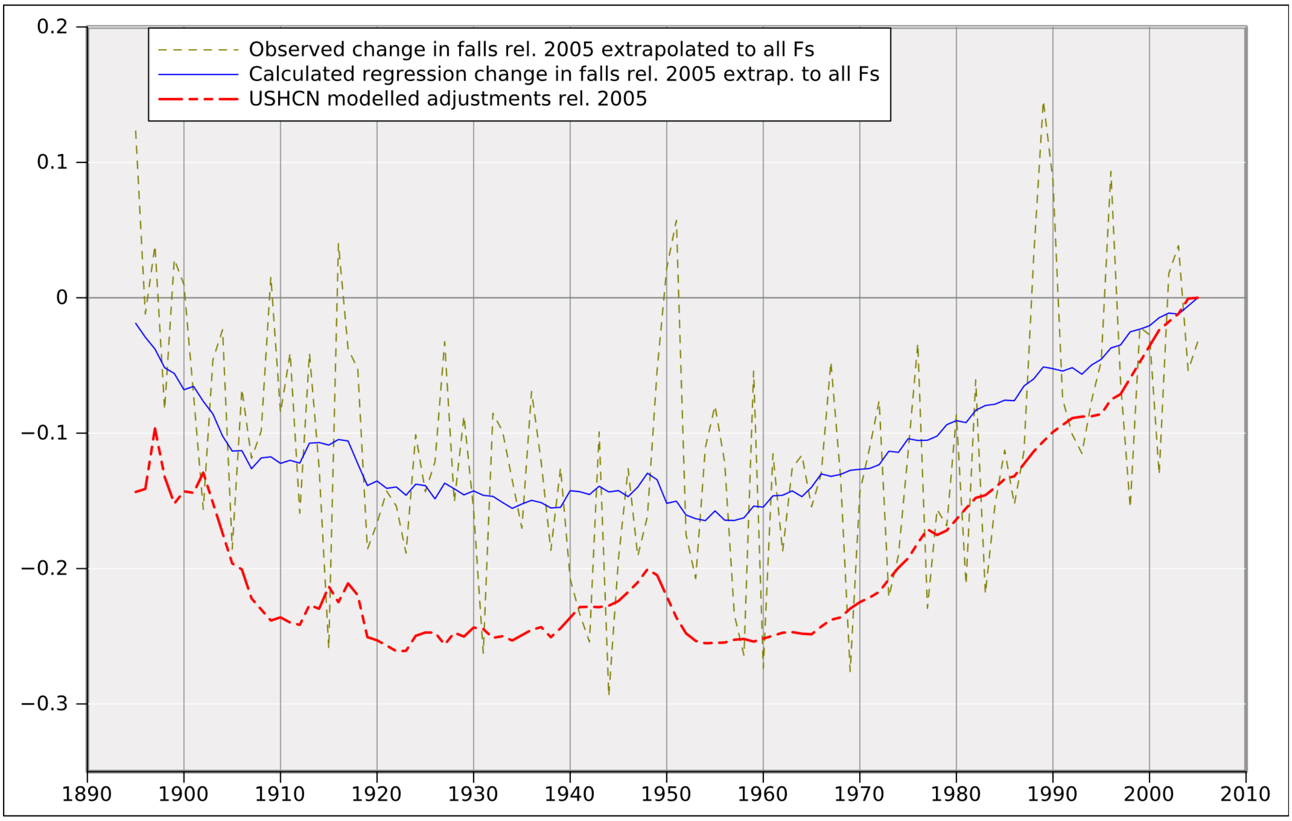

Figure 18: Detected TOBs error (°C) in nFF...nF sequences, USHCN adjustment, and observations

The result is visually almost indistinguishable from the RF...R cases, although these data include all but about 3% of Fs in the data set. If we take this as the more reliable reconstruction, then the overall USHCN modelled result is 175.9% of the effect measurable from the actual data. (Recall that in the RF...R case it was 177.9%; this is very slightly better for the official estimate, though it is still far greater than the actual effect.)

Finally, what about the 3.6% of the Fs that we cannot use? As explained earlier, these are mostly due to sequences that straddle the start, end, or a missing value, in a station observation record. Since longer sequences have a greater probability of interruption, and they also have lower sensitivity to TOBs, our results are almost certainly a slight over-estimate of the TOBs effect. Thus our “worse” result, the nFF...nF sequence, paints the most flattering picture that is at all credible for the official estimates: they are at least $\simeq$176% of the real amount, on average.

One last comparison: this 176% is a ratio of area under the curve. But in some years the official estimate is far worse. Huge overestimates are common. In 1895, it is an astonishing 760%; 214% in 1914 and 1916; 186% in 1966; and 236% in 1998.

4 Conclusion and Summary

Our method is based on the observation that, if TOBs error affects measurements then, over the course of a year, keeping in mind the large number of observations across all stations in the US, the sizes of measured daily falls in maximum temperature relative to a preceding unaffected datum (from the previous day or a day a few days earlier) should be statistically reduced by an increasing fraction of stations measuring in the afternoon. Furthermore, this effect must be linear unless there have been systematic changes in the exact time of measurement in the afternoon (for stations taking measurements then). Given the complete absence of information on that account, and its unlikelihood, such systematic confounders have been discounted.

We have found that linear regression has worked extremely well based on this assumption. Statistical measures R-squared and p-value have been very satisfactory, and residual plots look completely random, as they should. Finally we have learned that the official TOBs adjustments, based upon modelling using recent data and projected back upon the historic data, are, on average, at least 176% of the value found in the historic data itself, with some individual years overestimated even more.

This judgement is in regard to the official modelled values based on experiments with synthetic changes to recent computerised hourly measurements. We cannot speculate on why those experiments give a TOBs estimate greater than that in the real data, except to say that they clearly do. Hausfather also mentions that more recent analyses, such as Berkeley Earth, do not need to make this TOBs adjustment, as they use a pairwise homogenisation algorithm to detect and correct for step changes in station data. Again, any comment on those methods is beyond the scope of this paper, except to say that, if they do indeed closely resemble the official TOBs adjustments discussed here, they, to, are necessarily over-estimated.

5 References

Christy, John R., and McNider, 2016. Richard T. Time Series Construction of Summer Surface Temperatures for Alabama, 1883-2014, and Comparisons with Tropospheric Temperature and Climate Model Simulations. *Journal of Applied Meteorology and Climatology* 55(3), 811–826, 2016. DOI: 10.1175/JAMC-D-15-0287.1

Hausfather, Zeke. 2015. Understanding Time of Observation Bias. http://judithcurry.com/2015-/02/22/understanding-time-of-observation-bias

Menne, Matthew J., Williams, Claude N. Jr., Vose, Russell S., 2009. The U.S. Historical Climatology Network Monthly Temperature Data, Version 2. *BAMS*, 993–1007, July 2009. ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/papers/menne-etal2009.pdf