JP2.23 A COMPREHENSIVE SINGLE-STATION QUALITY CONTROL PROCESS FOR HISTORICAL WEATHER DATA

Karen Andsager*, Michael C. Kruk, and Michael L. Spinar
Midwestern Regional Climate Center, Champaign, Illinois

1. INTRODUCTION

The National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center's (NCDC) Climate Database Modernization Program (CDMP) – Forts Database Build 1820s-1890s Project (Forts) has resulted in the scanning and indexing of hundreds of thousands of pages of meteorological records and journals from the Nineteenth Century, from microfilm held by the National Archives and Records Administration. These daily observations of temperature, precipitation, humidity, cloud cover, wind speed, and other variables, were made both as part of the United States military record and by volunteers under the supervision of various agencies such as the Smithsonian Institution, the Signal Service, and the Department of Agriculture. The digitization of these records for the 163 highest priority stations is well underway with an integrated process involving the keying and quality control of metadata and daily data. The metadata are digitized by CDMP staff located at NCDC and include specific instructions for the keying of each image for each station. A private contractor, SourceCorp, digitizes the daily data. The quality control process has been developed and applied by experienced climatologists at the Midwestern Regional Climate Center. The goal of the quality control of the daily data is to ensure that the digital record accurately preserves the daily weather record as recorded by the observer on the form. The spatially disparate nature of these historical records requires the use of a single-station data verification strategy. This process differs from the methods used by Kunkel et al. (1998) to quality control DSI-3205, and Kunkel et al. (2004) to quality control DSI-3206, which employed gridded estimates based on surrounding stations. The quality control process for the Forts digitization project is described herein.

2. QUALITY CONTROL PROCESS

In order to ensure the quality of the Forts dataset, rigorous quality control procedures have been implemented. To objectively flag monthly and daily data values for manual assessment, the quality control process employs a series of tests spanning eight different quality assurance categories (Table 1). The first four categories primarily address issues that may arise from the keying process itself, including, for example, a comparison between the specific instructions provided to the digitization contractor and the data elements received from them. In addition, a few tests within these categories address known systematic issues in the dataset, such as poor wet bulb observations in winter. The values failing these tests are automatically flagged in the final data. The second four categories address internal calculated and climatological consistencies, first on a monthly, then on a daily basis.

Each quality control test is run in succession, beginning with the gross error and formatting test, upon the translation of the daily data output-keying format into the NCDC DSI-3200 standard format. The outliers from each test are verified and any corrections are applied at the conclusion of the verification for each test. The resulting corrected dataset then becomes the source of input for the next quality control test. This step-wise quality control process continues until the tests in all eight categories are complete.

Manual verification with the scanned daily data forms is performed by experienced climatologists on all the flagged values that are not systematic errors. To expedite this process, a web tool has been developed that displays the flagged value, along with a table of all the available daily data keyed on the flagged image, and a list

* *Corresponding author address*: Karen Andsager, Illinois State Water Survey, 2204 S. Griffith, Champaign, IL 61820; e-mail: andsager@uiuc.edu.

of the corresponding elements from the metadata (Figure 1). The assessor compares the flagged value with the value on the scanned image, and may consider other information available on the image to help interpret poorly legible values due to the combination of handwriting and the process of microfilming the original documents. For example, to verify a flagged temperature from the exposed thermometer, the assessor may consider an available dry bulb temperature from the same observation time. Once the verification of the flagged value is complete, the assessor selects the option that best describes the verification. The data value may: 1) verify outright, 2) verify, noting that an error was apparently made by the observer, 3) require a correction due to a keying or other data entry error, 4) be set to missing, or 5) be deleted from the database. Values are declared observer errors when it is evident that the observer made a mathematical summation or division error on the form, improperly used the instrumentation (e.g. had difficulty measuring wet bulb temperatures during sub-freezing temperatures), or when it is clear that the observer mis-wrote the value on the form (usually by a factor of ten). When the assessor submits a verification and its associated justification for an outlier, this information is recorded and appended to a text file that contains all the previous verifications within each station. This method allows for the final data to be reconstructed, if necessary, by applying the set of corrections for the station in sequential order to the original raw output from the data entry process.

3. VERIFICATION STATISTICS

Lunenburg, Vermont, and Oglethorpe Barracks, Georgia, are the first two stations to have been successfully taken through the quality control process for this project. Dr. Hiram Adolphus Cutting took all the observations for Lunenburg over the period 1859 until his death in 1892. Dr. Cutting struggled for many years to take the wet bulb observations accurately in the winter, when the temperature was below freezing. These observations can be difficult to take as too much water on the bulb causes the temperature to read too high: the water warms the bulb, then as itfreezes, it forms a crust, insulating the bulb, while the latent heat of fusion released artificially heats it. He eventually stopped recording any wet bulb temperatures in the cold season. Over one third (3300 values) of Dr. Cutting's winter wet bulb temperatures (those taken from October through April), have been systematically flagged as observer errors in the digital database under Category 1 (Table 1). These measurements are flagged because the recorded wet bulb temperature is greater than the sub-freezing dry bulb temperature. In contrast to Lunenburg, observations at Oglethorpe Barracks were taken by a number of Signal Service observers over the period 1866 through 1879. Wet bulb temperatures were recorded for 1869 through 1874. Due to the warmer climate in Georgia (minimum temperatures shown in Figure 2), only a handful of wet bulb temperatures were systematically flagged under Category 1.

The tests in Category 5 compare the monthly sum and mean as keyed from the observational form with the keyed daily values, for all element types with numeric values. For Lunenburg, this test was relatively efficient at flagging errors, with 39 out of 50 monthly flags (78%) containing an error of some sort (keying error, form not clear, or observer error) (Oglethorpe Barracks: 163 out of 177 flags, 92%). Tests in Category 6, which check the climatological consistency and extremes of the monthly totals and means, were relatively inefficient compared to Category 5, with 8 out of 46 monthly flags (17%) containing an error of some sort for Lunenburg (Oglethorpe Barracks: 2 out of 3, 66%). Category 7 examines the climatological consistency and extremes of daily values of temperature and precipitation. The extremes and temporal consistency tests were relatively inefficient at flagging erroneous values, with just 27 errors found in 500 daily flags (5%) for Lunenburg (Oglethorpe Barracks: 4 out of 61, 7%). When the appropriate data were available, daily consistency comparisons between element types, such as daily maximum temperature greater than or equal to the highest at-hour temperature reading for the day, were somewhat more efficient for Lunenburg, with 265 errors found in over 1080 daily flags (25%) for Lunenburg.

These consistency comparisons were not more efficient than the extremes and temporal consistency tests for Oglethorpe Barracks (21 out of 815, 3%).

Of the individual values found to contain errors, the majority are noted as observer errors, which are retained with an observer error flag in the database. For Lunenburg, 191 observer errors in individual values were noted from the tests of Category 7, while 101 keying errors were corrected in the database. Of the keying errors noted, about 90% were attributed more to the difficulty of reading the images than on actual key-stroke errors. For example, consider the entry for the 7 AM "thermometer in the open air" (first data column) on the 25th of February 1859, at Peoria, Illinois (Figure 3). The keyers entered "02". The wet bulb temperature for the same observation time on the same day (fourth data column) is "30". Also, several of the other "3"s resemble "0"s, particularly the 7 AM wet bulb temperature on the 14th. Thus, it would be more appropriate to interpret the 7 AM temperature as "32".

4. SUMMARY

A single-station quality control process has been developed for the Nineteenth Century daily data from Forts and other voluntary observers, with the primary purpose of ensuring accurate digitization of the data. At this time, two stations have passed completely through the quality control process. Each station may have its own set of quality issues, depending on the element types recorded and the general climate of the station. Some of the tests within the process are more efficient than others; given the volume of data and time considerations, limits for some of the tests may be adjusted to maximize the efficiency of the process at correcting issues with the largest impact on the climatology recorded in the data.

5. ACKNOWLEDGEMENTS

This project is funded through the Climate Database Modernization Program of the NOAA National Climatic Data Center. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Climatic Data Center or the Illinois State Water Survey.

6. REFERENCES

Kunkel, K.E., K. Andsager, G. Conner, W.L. Decker, H.J. Hillaker Jr., P. Knox, F.V. Nurnberger, J.C. Rogers, K. Scheeringa, W.M. Wendland, J. Zandlo, and J.R. Angel, 1998: An expanded digital daily database for climatic resources applications in the Midwestern United States. *Bull. Amer. Meteor. Soc.,* **79**, 1357-1366.

_____, D.R. Easterling, K. Hubbard, K. Redmond, K. Andsager, M.C. Kruk, and M.L.Spinar, 2004: Quality control of pre-1948 cooperative observer network data. *J. Tech*, (accepted for publication).

TABLE 1. QUALITY ASSURANCE CATEGORIES

1	Gross Translation and Error	This test is applied on translation of the data from the output-keying format used by SourceCorp to the TD-3200 format used by NCDC. This test flags individual daily values with cutoffs set to the extreme range of each element type as well as addresses formatting issues associated with the output keying and DSI-3200 data formats. Systematic data quality issues are also automatically flagged or corrected.
2	Metadata Element Agreement	This test checks for a match between the element types keyed by the daily data keyers and the element list keyed in the metadata for each image.
3	Element Duplication and Consistency	This test checks for duplication of element types within each month for each station. The test also checks for consistency in observation time within each element type and assignment of element types within broader classes of related types.
4	Data Completeness	This test employs a series of manual and automated checks to ensure data completeness. Specific tests examine months with a small amount of daily data and gaps in or abrupt terminations of the station record.
5	Consistency of the Keyed Daily and Monthly Data	This test checks for consistency between the keyed monthly means and sums and the respective calculated totals from the daily values for all element types within each station.
6	Internal Temperature and Precipitation Consistency of the Monthly Means/Totals	This test checks for the climatological internal consistency of temperature and precipitation daily data elements on a monthly basis. It compares independently calculated monthly means and totals using the keyed daily values.
7	Internal Extremes and Climatological Consistency of Daily Values	This test checks for extremes and internal consistency of the keyed daily values within each station data element. The daily extremes limits are set using the climatology of the station in the CDMP-Forts data set.
8	Consistency of Daily Values Calculated From Other Data	This test checks for consistency between keyed daily values derived from other keyed daily values within each station record, e.g., relative humidity calculated from wet/dry bulb temperatures and daily mean temperature from maximum/minimum temperatures.

CDMP-Forts Outlier Verification for AL_Mobile

Station ID	Start Year	Start Month	Day	Hour	End Year	End Month	AE Number	Flag Type	Flag Sub-Type	Element	Value
015478	1884	01	33	23	1884	01	AE03150D\00000128	05	01	DPTP	66.8

Errors on the page? Questions? E-mail Mike K. for help.

- ◉ Verifies (no change)
- ○ Daily Data Needs to be Keyed
- ○ Delete Value
- ○ Set Value to Missing
- ○ Change Value
- ○ Change Element Type
- ○ Change Date
- ○ Change Obs Time
- ○ Change Station ID Number

- ◉ Save Entry and Set up Next
- ○ Save Entry and Listall
- ○ Save Entry and Go to the Main Page
- ○ Save Entry and Quit

Outliers Remaining: [Calculate]

[Submit Entry] [List All To Be Keyed]

[Go Back to the Main Page]

[Add an Outlier]

Change the value here: **66.8**

Change the element type here: **DPTP**

Change the start year here: **1884**

Change the start month here: **01**

Change the day here: **33**

Change the hour here: **23**

Change the station ID number here: **015478**

Station ID: AL_Mobile YEAR: 1884 MONTH: 01 AE03150D\00000128

Day	MNTO HF 99	TMAX HF 99	TMIN HF 99	TRNG HF 99	WDVL MH 15	TAHR HF 07	TAHR HF 15	TAHR HF 23	TMPW HF 07	TMPW HF 15	TMPW HF 23	DPTP HF 07	DPTP HF 15	DPTP HF 23	BPTI IT 07	BPTI IT 15	BPTI IT 23
01	04880	06180	03600	02580	00120	06100	04930	03600	06020	04700	03150	05940	01440	02250	29969	29997	30114
02	03050	03800	02400	01400	00080	02460	02680	03000	02110	03050	02600	01140	01730	01700	30248	30220	30372
03	03220	03900	02430	01470	00030	02460	03900	03300	02310	03200	02900	01960	01600	02100	30381	30347	30367
04	04700	05600	03190	02410	00040	03400	05600	05100	03140	05000	04900	02680	04300	04700	30311	30225	30176
05	02780	05100	01900	03200	00120	03500	02950	01900	03380	02450	01600	03140	01200	00500	30239	30291	30404
06	02690	03450	01390	02060	00030	01510	03100	03450	01310	02600	02990	00610	01400	02090	30377	30315	30270
07	04000	04850	03280	01570	00200	03400	04500	04100	02900	04390	04100	01900	04270	04100	30204	29885	29890
08	02940	04100	02500	01600	00140	03000	03210	02600	02890	02910	02400	02670	02310	02000	29955	30034	30150
09	03230	03800	02180	01620	00060	02200	03700	03800	02100	03200	03700	01900	02100	03500	30150	30108	30109
10	05030	05800	03280	02520	00030	03700	05800	05600	03500	05600	05500	03200	05400	05400	30077	30024	30071
11	05050	06700	03900	02800	00160	05840	05400	03900	05740	05300	03700	05640	05200	03400	30032	30103	30256
12	04040	05200	02890	02310	00060	03810	05000	04100	02850	04000	03500	02530	02200	02300	30323	30315	30333
13	04490	05400	03300	02100	00070	03320	05350	04800	03120	04600	04500	02720	03600	04100	30272	30167	30112
14	05500	06500	04350	02150	00120	04490	06200	05810	04350	06000	05680	04170	05800	05550	30031	29891	29918
15	05390	06100	04900	01200	00080	05220	06050	04900	05080	05400	04700	04940	04750	04500	30047	30041	30264
16	04680	05200	04200	01000	00090	05310	05120	04600	04000	04400	04150	03500	03360	06450	30322	30278	30351
17	04590	04900	04040	00860	00090	04130	04900	04750	04030	04650	04650	03930	04350	04550	30377	30269	30212
18	05750	06380	04650	01730	00110	05370	05990	05900	05300	05900	05800	05230	05850	05100	29996	29811	29856
19	04780	06100	04450	01650	00120	04900	05000	04450	04540	04500	04100	04080	03800	03600	29947	30032	30208
20	05730	04450	03300	01150	00120	03700	04200	03300	03290	03500	02900	02460	02100	02100	30357	30383	30536

Figure 1 – A sample from the verification web tool highlighting the monthly mean dewpoint temperature (DPTP) of 66.8˚F (19.3˚C) as an outlier inconsistent with the mean of the daily values in Mobile, Alabama, at 11 PM local time during January of 1884.

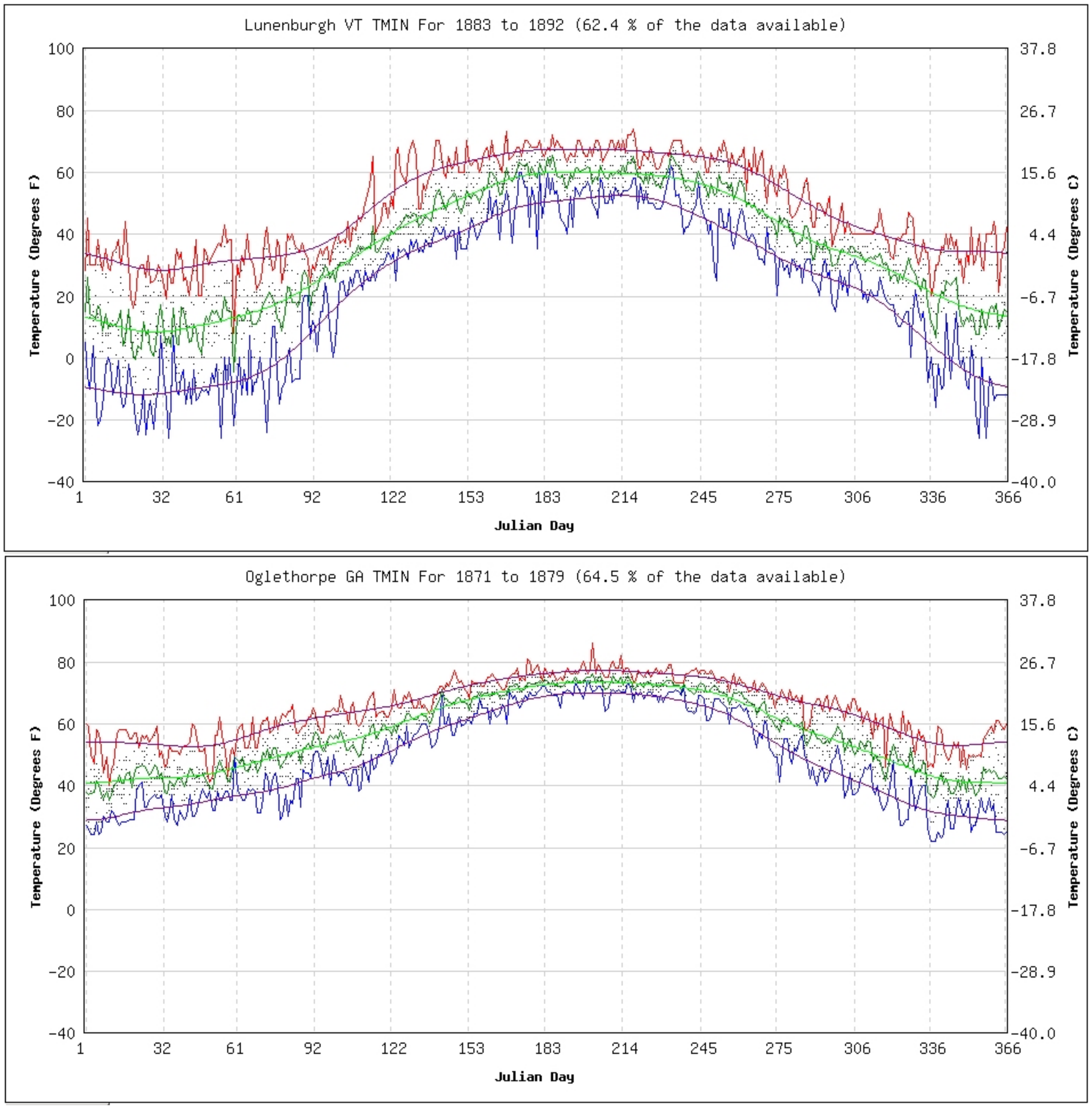
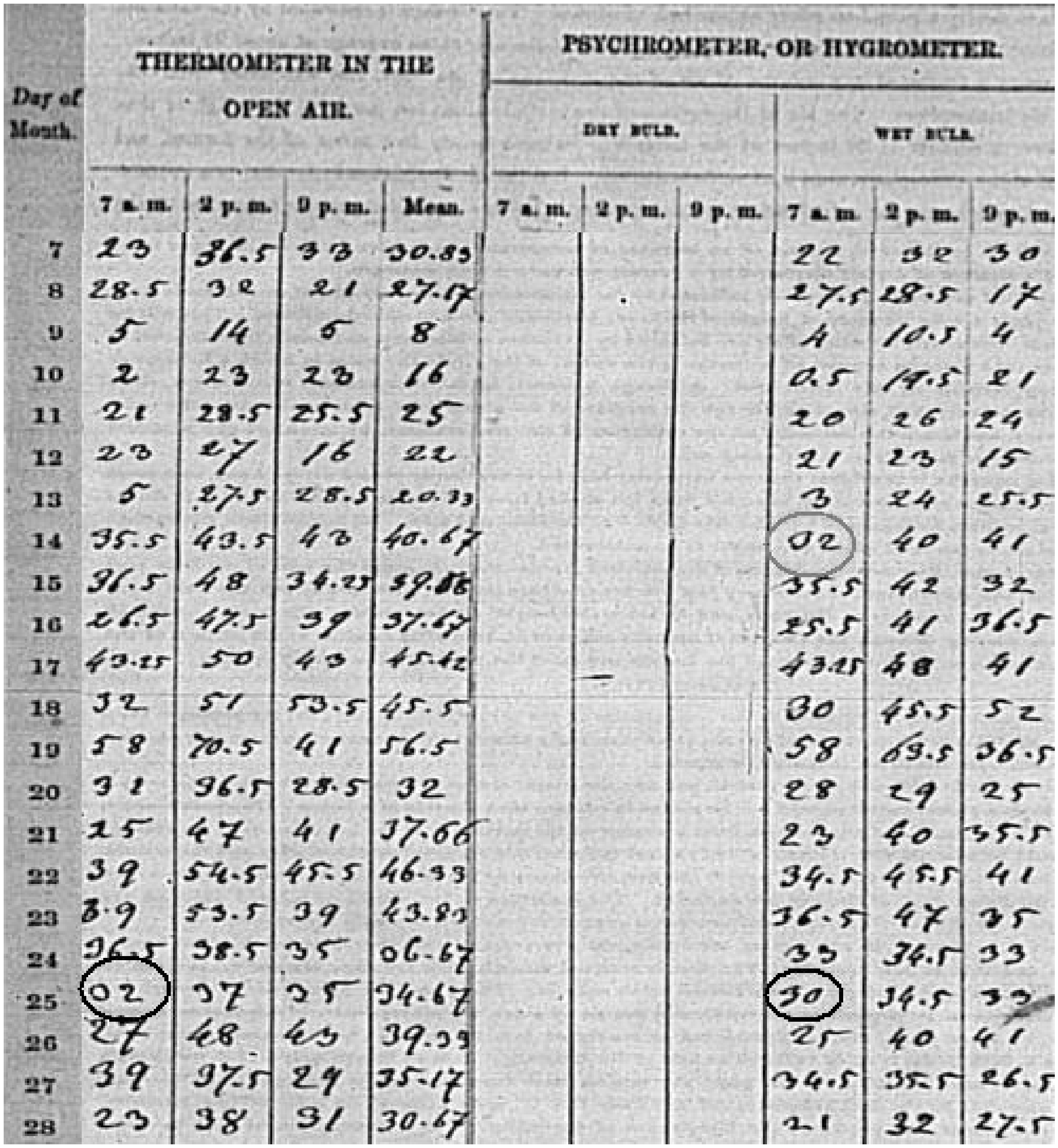


Figure 2 – A comparison of the minimum temperature climatology for Lunenburgh, Vermont (top) and Oglethorpe Barracks, Georgia (bottom). Vertical axis is in units of ^{o}F on the left and ^{o}C on the right. The red line denotes the warmest minimum temperature in the dataset within the period of record for a given day (where the purple line is the smoothed average of these values). The green line denotes the average daily minimum temperature (where the green line is the smoothed average of these data). The blue line represents the lowest daily minimum for a given day (and the purple line again represents the smoothed average of these values).



Day of Month.	THERMOMETER IN THE OPEN AIR.				PSYCHROMETER, OR HYGROMETER.					
					DRY BULB.			WET BULB.		
	7 a. m.	2 p. m.	9 p. m.	Mean.	7 a. m.	2 p. m.	9 p. m.	7 a. m.	2 p. m.	9 p. m.
7	23	36.5	33	30.83				22	32	30
8	28.5	32	21	27.17				27.5	28.5	17
9	5	14	6	8				4	10.5	4
10	2	23	23	16				0.5	14.5	21
11	21	28.5	25.5	25				20	26	24
12	23	27	16	22				21	23	15
13	5	27.5	28.5	20.33				3	24	25.5
14	35.5	43.5	43	40.67				32	40	41
15	36.5	48	34.25	39.58				35.5	42	32
16	26.5	47.5	39	37.67				25.5	41	36.5
17	43.25	50	43	45.42				43.25	48	41
18	32	51	53.5	45.5				30	45.5	52
19	58	70.5	41	56.5				58	63.5	36.5
20	31	36.5	28.5	32				28	29	25
21	25	47	41	37.66				23	40	35.5
22	39	54.5	45.5	46.33				34.5	45.5	41
23	39	53.5	39	43.83				36.5	47	35
24	36.5	38.5	35	36.67				33	34.5	33
25	32	37	35	34.67				30	34.5	33
26	27	48	43	39.33				25	40	41
27	39	37.5	29	35.17				34.5	35.5	26.5
28	23	38	31	30.67				21	32	27.5

Figure 3 – An example of a keying error for Peoria, Illinois in February of 1859. The 7am at-hour temperature on the 25^{th} day (circled in black) was keyed incorrectly as 2˚ F (-16.7˚ C). This is compared with the wet bulb temperature at the same observation time (circled in black), as well as with the handwriting of the 7am wet-bulb temperature on the 14th (circled in grey). Based on this comparison, it is evident that the 7am at-hour temperature should have been keyed as 32˚ F (0˚ C).