Comparison of Quality and Completeness of National Climatic Data Center and Illinois Climate Assistance Service (CLASS) Data

WAYNE M. WENDLAND
Illinois State Water Survey, Champaign, Illinois

ALAN L. MCNAB AND VERNELL WOLDU
National Climatic Data Center, Asheville, North Carolina

15 September 1987 and 28 March 1988

ABSTRACT

We compare the completeness and quality of one year's daily temperature and precipitation observations from the National Climatic Data Center (NCDC) database with those received by the Illinois State Water Survey's (ISWS) Climate Assistance Service (CLASS). The NCDC data are keyed by hand from the E-15 forms received from the National Weather Service (NWS) Cooperative Observers, and are quality controlled. The CLASS data are transmitted by NWS Cooperative Observers via touch-tone phone to a Water Survey computer. These data are quality controlled only to the extent that each value is verbally repeated to the transmitting observer by voice synthesizer, and the ISWS computer checks to determine that the temperatures are consistent, i.e., maximum ≥ minimum and so on.

The NCDC database was more complete than that of CLASS. This situation was most obvious for precipitation entries where CLASS observers typically made no precipitation entry (rather than transmitting a zero) to save transmission time. National Climatic Data Center, on the other hand, automatically enters zeros when precipitation was unreported but was unlikely to have occurred. The NCDC database which is rigidly quality controlled was also more accurate (only 0.4% errors) that that of CLASS (3.6% errors).

It is possible that if telephone transmission became the official mechanism for data entry, missing data would be much reduced; and if each entry value had to be repeated by the observer, the accuracy of such transmissions would also improve.

1. Introduction

Weather and climate data from some 8000 National Weather Service (NWS) Cooperative Observers in the United States are continually received and archived by the National Climatic Data Center (NCDC). The data are received each month on paper forms (E-15s), keyed into a digital file, and checked for key entry and observer errors. The final data file contains original data, flags to identify suspected erroneous original data, and NCDC estimates (also identified by flags) for erroneous and missing data (NCDC 1987). This database becomes the primary climatological database for the United States to be used for climatic research, environmental assessment, and structure design characteristics. The keying, quality control and printing of the data are performed over a period of three to four months, thereby prohibiting distribution of printed copies of "Climatological Data" until three to four months after the fact.

From January 1984 until April 1987 38 Cooperative Observers in Illinois provided their observations to the Illinois Climate Assistance Service (CLASS), a computer-based weather and climate data system (Changnon et al. 1984, 1987) operated by the Illinois State Water Survey (ISWS). Climate Assistance Service accepts digitized temperature and precipitation observations transmitted daily by observers via touch-tone phones. The ISWS computer includes a voice synthesizer that repeats the received data to the observer for verbal verification. Software ensures that temperature data are consistent, i.e., maximum temperature is equal to or greater than minimum temperature and so on. If not consistent, voice synthesizer so states, and asks for retransmission.

It is important to understand that CLASS was designed for real-time data acquisition and dissemination, whereas the NCDC data are archived as a high quality database suitable for analyses in months and years after collection. In spite of the difference in timeliness of these two datasets, a comparison of the resulting data files was completed to contrast the accuracy and completeness of a digital file obtained via an observer centered data entry and verification method (CLASS), versus a centralized key entry and verification method (NCDC).

2. Method

For this study, daily data from the CLASS system, the NCDC database, and a sample of original manu-
script records (E-15s) for 38 Illinois stations from 1 February 1984 to 31 January 1985 were compared. The original measurements were the basis for accuracy assessment. This was the first year of CLASS operation. Data to be compared consisted of the daily maximum, minimum and observation-time temperatures and precipitation values.

The 38 stations are among the National Weather Service’s network of 172 Cooperative Observers in Illinois. The Cooperative Observers entered the daily data on E-15 forms and the forms were sent to the NCDC for keying into the NCDC Summary of the Day database. The database is checked for key entry errors and observer/instrument errors. When possible, NCDC provides estimated values for data that are missing or clearly erroneous. National Climatic Data Center checking also includes comparison of each observation to those of the station’s nearest neighbors.

The following study compares the CLASS database as received from the observer, to the corrected database of NCDC.

3. Number of observations

The first step of the comparison consisted in counting the data missing from CLASS and NCDC files. The total number of values for each element was counted and subtracted from the number of potential values for each observation. All 38 stations recorded precipitation, snowfall and snow depth; therefore, the number of possible precipitation observations was 366 days × 38 stations = 13 908. Only 36 stations recorded maximum, minimum, and time of observation temperature, for both CLASS and NCDC networks; therefore, for these elements, the number of possible observations was 366 days × 36 stations = 13 176.

Nine stations did not participate in the CLASS program for the entire 12 month study. Consequently, 35 station-months (1084 observations) for such cases were not included when calculating the percentage of missing CLASS data.

The National Climatic Data Center provides estimates for maximum and minimum temperatures in cases when a station has fewer than nine missing temperatures in a month and data from nearby stations are available. This procedure added 54 values to the NCDC files. Table 1 presents the total number of values in each file; percentages are based on the total possible observations.

4. File accuracy

The accuracy of the NCDC and CLASS files is described here in terms of the number of correct values determined by several methods. Each successive method exhibits a decreasing confidence in the ability to determine if and where (CLASS, NCDC, or E-15 data) an error occurred. The first determination (Step 1) is the clearest definition of “correct,” and subsequent steps introduce more complex situations. Fortunately, the Step 1 condition accounts for about 96% of data common to both files.

The Step 1 definition of correct data includes only those observations for which values from all three sources (NCDC, CLASS and E-15 manuscript) are available, equal and the NCDC value is not flagged as suspect. The equality of values implies that no errors were introduced during the NCDC key entry of the manuscript data, and no errors were introduced as the observer entered the data via touch tone system to CLASS.

The fact that the NCDC data are unflagged means the temperatures have successfully passed (i) range checks, (ii) interelement consistency checks (e.g., max > min) and (iii) area edit checks. Precipitation, snowfall and snow depth data have also passed several checks but these are primarily designed to identify major (obvious) entry errors and to clarify incomplete, improper or unclear entries in the manuscript. For instance, observers sometimes fail to record zero precipitation amounts (the entry is left blank, which could mean the precipitation is missing). However, if the temperatures are recorded on the day of the blank precipitation entry, NCDC usually interprets the blank as zero and appends a flag to the zero. As illustrated in this example, such flags are not used to indicate questionable accuracy, but rather that an interpretation of the E-15 entry was necessary. For the purposes of this study, the NCDC values of precipitation, snowfall and snow depth have been counted as “not flagged.”

It was not necessary to compare the manuscript data with the keyed data to identify those observations that are correct per the Step 1 definition. A computer program simply compared all unflagged NCDC values to

<table>
<thead>
<tr>
<th>Table 1. Number of observations, 1 February 1984–31 January 1985. (Percentages based on number of possible observations)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Element</strong></td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Max temp.</td>
</tr>
<tr>
<td>Min temp.</td>
</tr>
<tr>
<td>At obs temp.</td>
</tr>
<tr>
<td>precip.</td>
</tr>
</tbody>
</table>
the corresponding CLASS values. If the two values were equal it is extremely unlikely that they would differ equally from the manuscript value. Table 2 shows the frequency of the observations that are correct per the Step 1 definition (96.3%). The remaining 3.7% show disagreement between files and/or have been flagged by NCDC. The entries in column 2 suggest likely CLASS errors, and those in column 3 likely NCDC errors.

The Step 2 definition of erroneous data concerns the observations listed in column 2 of Table 2 as “Number of unflagged values for which CLASS ≠ NCDC.” The Step 2 definition counts an observation correct if it agrees with the manuscript entry and the NCDC value is not flagged. This definition is designed to identify key entry (NCDC) or touch tone (CLASS) errors in uncomplicated situations; i.e., when no flag indicated there could be other reasons for the difference between the NCDC and CLASS data.

The Step 2 definition required an inspection of the E-15 form in each case when the unflagged NCDC value was not equal to the CLASS value. Because of time constraints, the comparison was completed in only a 16% random sample of the data. The number and percentage of total in each category according to Step 2 are shown in Table 3.

The Step 3 definition of correct concerns temperature data that have been flagged by NCDC. (All of the precipitation, snowfall and snow depth values were accounted for in Steps 1 and 2 because these elements have been treated as unflagged data.) National Climatic Data Center values are flagged under the following conditions: i) Flag A: Missing data, outliers or inconsistent values were estimated; or ii) Flag B: Value was apparently inconsistent, but could not be estimated.

The following five conditions define “correct” according to Step 3:

1) If the CLASS value is missing, count NCDC value as correct (except for condition flag B, above). The occurrence of a flag demonstrates that the value keyed from the manuscript is temporally or spatially inconsistent, is an outlier, or is missing from the E-15 form. CLASS data are not available to clarify the problem.

2) If the CLASS value equals the NCDC value, count both as correct (except for flag B cases, above). Both files indicate that the manuscript value is in error or that NCDC correctly estimated a value that was missing in the manuscript.

3) If the NCDC value is not equal to the CLASS value and the CLASS value is not equal to the manuscript value (i.e., the NCDC flag does not indicate CLASS is wrong), count CLASS as correct. The NCDC value may be an improvement over the inconsistent, outlier or missing manuscript value, but is not equal to the CLASS value.

### Table 3. NCDC and CLASS errors according to Step 2 definition of correct (Unflagged NCDC ≠ CLASS; correct value determined by manuscript based on ~16% random sample of dataset).

<table>
<thead>
<tr>
<th>(1) Number of unflagged values for which CLASS ≠ NCDC (CLASS 999)</th>
<th>(2) Sample number of observations checked against manuscript</th>
<th>(3) Number of samples for which NCDC correct (CLASS error)</th>
<th>(4) Number of samples for which CLASS correct (NCDC error)</th>
<th>(5) Neither correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Temp</td>
<td>439</td>
<td>80</td>
<td>78</td>
<td>97</td>
</tr>
<tr>
<td>Min Temp</td>
<td>253</td>
<td>70</td>
<td>69</td>
<td>98</td>
</tr>
<tr>
<td>At Obs Temp</td>
<td>223</td>
<td>35</td>
<td>35</td>
<td>100</td>
</tr>
<tr>
<td>Precip</td>
<td>652</td>
<td>82</td>
<td>77</td>
<td>94</td>
</tr>
<tr>
<td>Snowfall</td>
<td>132</td>
<td>13</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Snow depth</td>
<td>198</td>
<td>21</td>
<td>20</td>
<td>95</td>
</tr>
<tr>
<td>Mean</td>
<td>96.6%</td>
<td>1.7%</td>
<td>1.7%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

Unauthenticated | Downloaded 05/30/22 03:05 PM UTC
4) If the NCDC value is not equal to the CLASS value and the CLASS value equals the manuscript value, count NCDC as correct (except for flag B cases, Table 4). The NCDC flag indicates that both the CLASS and manuscript values are outliers or inconsistent with another element; i.e., both manuscript and CLASS are incorrect.

5) If the NCDC value should not have been flagged, count CLASS as correct. This occurs in cases when flags identify entry errors (such flags are supposed to be removed before the data are entered into the final NCDC database file). The number of occurrences in categories (1) through (5), Step 3 are summarized in Table 4.

The absolute number of cases is small in each instance. As expected, quality-controlled NCDC data exhibit fewer errors than the raw CLASS data, by about a 1:1.5 ratio (99:155).

The accuracy and completeness measures presented in Tables 1–4 are combined and summarized in Table 5. The percentages of Table 5 are based on 53,436 total possible observations of maximum, minimum and observation time temperatures, and precipitation values (column 1, Table 1). Snowfall and snow depth are not included in Table 5 because of the large number of missing values due to the failure to record “zero” amounts.

The NCDC errors in Table 5 are obtained by adding columns 4 and 5 in Table 3 and all of the categories designated as “NCDC-Error” in Table 4. The CLASS errors are obtained by adding columns 3 and 5 in Table 3 and all of the categories designated “CLASS-Error” in Table 4. The number of correct values is obtained by subtracting the number of missing and error values from 53,436. This procedure places the “unflagged” and “uncompared” values into the category “correct.” A value is “uncompared” if either the CLASS or NCDC value is missing, i.e., not available for comparison. The 1084 observations not recorded because nine CLASS stations did not contribute data for a complete 12 months (section 3) are also considered “uncompared.” About 10% of the NCDC values are classified as correct via this procedure.

5. Comparison of NCDC and CLASS data

Several differences between the NCDC file and the CLASS file have been identified. One is completeness. The CLASS file lacks 5.2% of the temperature and precipitation data, compared to 2.8% missing from the NCDC file after quality-control (Table 5). There is also a difference in the distribution of the missing temperature and precipitation data. About 90% of the missing NCDC data occur because the monthly E-15 forms (which are the basis of the NCDC file) are missing. The forms that are received are essentially complete; only about 0.5% of the data on forms received by the NCDC are missing. In contrast to NCDC, less than 10% of the missing CLASS data were missing because an entire month was wanting. Of those months with some data having been received, 5% of the total possible data on CLASS were missing. Although some missing data are due to the observers not making any entry to save time (for example, no entry as opposed to entering a zero when no precipitation occurred), many data apparently were missing because observers on vacation did not transmit late observations upon return.

The CLASS snowfall/snow depth data appeared to be much more incomplete than those of NCDC (47% complete for the former; 96% for the latter). The CLASS observers typically passed over these two queries without entering a zero, to save transmission time. This was particularly common in fall, before the snow “season” was perceived to have begun, and again in

---

**Table 4. Number of cases in each category of Step 3 definition of “correct” temperature.**

<table>
<thead>
<tr>
<th>Flag A indicates value is an estimate, or correction was made for a manuscript value that was missing, inconsistent, or an outlier.</th>
<th>CLASS missing</th>
<th>NCDC = CLASS</th>
<th>NCDC ≠ CLASS</th>
<th>NCDC ≠ CLASS</th>
<th>NCDC ≠ CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCDC-correct</td>
<td>NCDC-correct</td>
<td>NCDC-error</td>
<td>NCDC-correct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(85)</td>
<td>(48)</td>
<td>(61)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flag B indicates value is inconsistent but no correction was made.</th>
<th>CLASS missing</th>
<th>NCDC = CLASS</th>
<th>NCDC ≠ CLASS</th>
<th>Can not occur</th>
<th>NCDC ≠ CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCDC-error</td>
<td>NCDC-error</td>
<td>NCDC-error</td>
<td>Cannot occur</td>
<td>NCDC-error</td>
<td></td>
</tr>
<tr>
<td>(16)</td>
<td>(47)</td>
<td>(23)</td>
<td></td>
<td>(6)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5. Accuracy and completeness of maximum and minimum temperatures and precipitation.**

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Error</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCDC</td>
<td>96.8%</td>
<td>0.4%</td>
<td>2.8%</td>
</tr>
<tr>
<td>CLASS</td>
<td>91.2%</td>
<td>3.6%</td>
<td>5.2%</td>
</tr>
</tbody>
</table>
late winter and spring. Such missing entries occur in
the NCDC database as well; however, NCDC adds
zeros to a record for days on which a zero did not
actually appear on the E-15, but snow was unlikely to
have occurred. This problem can be somewhat alle-
viated by noting precipitation at nearby stations.

Substitute observers apparently are more reluctant
occasionally to transmit data via phone to a computer
than to complete the E-15 form. If data transmission
to the computer were to become the “official” pro-
cedure for sending observations to a repository, greater
compliance may result in fewer missing entries and
greater accuracy.

For temperature and precipitation data, the NCDC
file contains about 0.4% errors and the CLASS file con-
tains about 3.6% errors (Table 5). In 97% of the more
complex Step 2 cases, NCDC data were correct and
CLASS data erroneous, whereas the converse occurred
in only 1.7% of these cases. For the study as a whole,
the Step 2 definition identified about 80% of all errors.

The NCDC Step 2 errors occurred because key entry
errors were not identified and corrected. The CLASS
data have a much larger proportion of data entry errors
than NCDC; the CLASS errors are attributed to ob-
server error. Only about 20% of the NCDC errors are
uncorrected key entry errors. The positive impact of
NCDC’s quality control system is obvious.

6. Conclusions

As expected, the NCDC climate database is of higher
quality than that of CLASS. It has been keyed by hand
from the monthly E-15s and is checked for spatial and
temporal discontinuities; and when missing, aug-
mented with estimated values when possible. However,
these data are available from NCDC only three to four
months after the fact.

The data from CLASS, on the other hand, are avail-
able within an hour after the time of observation.
CLASS quality-control is limited to 1) a voice synthe-
sizer repeating each parameter for the observer’s ver-
ification, and 2) requiring that temperatures are inter-
ally consistent; i.e., maximum temperature is equal
to or greater than the minimum and so on. Missing
data are not estimated. These data are available each
day, current to the day, and are primarily used for near
real-time weather/climate assessments.

Comparing one year of daily observations of max-
imum, minimum and time of observation temperature,
precipitation, snowfall, and snow depth for 38 Illinois
stations during the first year of CLASS’s operation with
those data from NCDC’s database showed the NCDC
file to be considerably more complete than the CLASS
file. This is due to the fact that 1) NCDC estimates
missing data, whereas CLASS does not, and 2) data
not transmitted to CLASS on the day of observation
are sometimes not transmitted at a later date, although
the opportunity is offered with each phone call. It is
relatively easy for a substitute observer to complete an
E-15 form during an observer’s absence, or for the ob-
server to complete the form from notes made by the
substitute. Apparently substitutes are not as likely to
learn the touch-tone method of transmitting data, nor
is late entry by the observer as compelling as completing
the E-15 form.

The error rate in the NCDC file is considerably less
than the CLASS error rate. About 90% of the total
(3.6%) CLASS errors occurred during touch-tone data
entry. The voice synthesizer, while undoubtedly a help
in reducing errors, apparently does not provide suffi-
cient impetus for the observer to listen for erroneous
entries. Requiring each parameter to be entered twice
might reduce such errors.

Experience with CLASS suggests that observers
should be required to key a decimal (with either the
“*” or “#” key of a phone pad) to reduce order of
magnitude errors. These were particularly noted in
snowfall and snow on the ground. Observers sometimes
reported the parameters to hundredths of an inch,
whereas the former is only recorded to tenths of an
inch, and the latter to whole inches.

Although there are differences in data quality be-
tween the NCDC quality controlled files and those of
CLASS, each serves a decidedly different audience; i.e.,
NCDC serving the need for quality climate data with
a data density in Illinois of about one station per 845
square kilometers (320 square miles), first available
three to four months after the fact or anytime thereaf-
other; whereas CLASS offers virtually immediate climate
assessment from a database density of about 1 station
per 3,700 square kilometers (1,400 square miles). The
record of the use of CLASS data (Changnon et al. 1987)
suggests that perhaps a similar outlet should be imple-
mented for the Remote Observation System Auto-
mation (ROSA) cooperative data being received daily
by the National Weather Service from several mid-
western states.

Acknowledgments. We thank Ms. Marsha Webb and
Debbie Mitchell for typing the manuscript; and S. A.
Changnon, Jr., R. G. Semonin, and J. L. Vogel for
reviewing the manuscript. Partial support from this
work was provided by Grant COMO5A5A-A-D-MC-
056 from the Climate Analysis Center of the National
Weather Service, and by Grant COOMN5A81AA-
D-00-112 from the National Climate Program Office,
both to the Illinois State Water Survey.

REFERENCES

Soc., 65, 704–705.
W. M. Wendland and J. L. Vogel, 1987: Usage of near real-
time climate information. J. Climate Appl. Meteor., 26, 1072–
1079.
NCDC, 1987: TD-3200, Summary of day co-operative. National Cli-
matic Data Center, 25 pp.