A Deterministic Approach to the Validation of Historical Daily Temperature and Precipitation Data from the Cooperative Network

Thomas Reek,
Stephen R. Doty, and
Timothy W. Owen
National Oceanic and
Atmospheric Administration,
National Climatic Data Center,
Asheville, North Carolina

Abstract

It is widely known that the TD3200 (Summary of the Day Cooperative Network) database held by the National Climatic Data Center contains tens of thousands of erroneous daily values resulting from data-entry, data-recording, and data-reformatting errors. TD3200 serves as a major baseline dataset for detecting global climate change. It is of paramount importance to the climate community that these data be as error-free as possible. Many of these errors are systematic in nature. If a deterministic approach is taken, using empirically developed criteria, many if not most of these errors can be corrected or removed. A computer program utilizing Backus Normal Form structure design and a series of chain-linked tests in the form of encoded rules has been developed as a means of modeling the human subjective process of inductive data review. This objective automated correction process has proven extremely effective. A manual review and validation of 138 stations of a 1300-station subset of TD3200 data closely matched the automated correction process. Applications of this technique are expected to be utilized in the production of a nearly error-free TD3200 dataset.

1. Introduction

For over 100 years, volunteer weather observers, comprising what is known today as the Cooperative Observer Network (COOP), have faithfully recorded daily weather data on forms supplied by the National Weather Service (NWS) and its predecessors. Today, the COOP system consists of approximately 8300 stations located throughout the contiguous United States, Alaska, Hawaii, Pacific Islands, Puerto Rico, and the Virgin Islands. About two-thirds of the stations record daily measurements of maximum (TMAX), minimum (TMIN), and observing time temperature (TOBS), precipitation (rain, snowfall, hail water) (PRCP), snowfall (SNOW), and snow depth (SNWD). About one-third record only PRCP, SNOW, and SNWD. A limited number also record soil temperature, evaporation, and wind measurements. Over the years, some 25 000 stations have, at one time or another, participated in the COOP system. A detailed description of the COOP system has been given by the NWS (1987). For the purposes of this discussion, it is important to note that observing stations are divided into three distinct and often dissimilar "reader groups." One group records 24-h maximum and minimum temperatures (and precipitation) in the morning, typically around 8 A.M., and are known as A.M. readers. Another group records in the afternoon, presumably after the daily maximum has been reached, typically about 4 P.M. This is the P.M. reader group. The third group reads at midnight and are the midnight (MID) readers. Still another much smaller group reads at varying times, usually at sunrise or sunset, and are grouped with the A.M. and P.M. groups accordingly.

Next, the recorded observation forms are sent directly to the National Climatic Data Center (NCDC) for purposes of quality control (QC), dissemination, and archiving. Prior to the centralization of this function at NCDC in 1962, COOP observation forms found their way to regional processing centers, where the data were extracted for use in publications such as the current *Climatological Data* (CD). As punch cards became popular in the late 1940s, the production of publications was automated. Still, according to Davis (1991, personal communication), punched cards were viewed as an interim medium for hard-copy production and dissemination; errors corrected on hard copy were not always corrected in the card files. The collection of punch cards continued to grow until magnetic tape became the automation/storage medium of choice in the late 1960s. As computers became more commonplace, the value of such a large wealth of data became apparent. Agreements were made with various state universities to digitize the pre-1948 daily observations in order to expand the database. To date, this treasure trove of climatic data, known as TD3200 (NCDC 1991), contains many decades of data from some 25 000 stations of varying lengths of record. As its usage for climate-related studies grew, the need for quality data became more apparent.

2. Quality control background

Over the years, QC ranged from none to complex spatial comparison techniques. Duchon et al. (1982) describe in detail one such method introduced at NCDC in January 1982. Reek and Crowe (1991) introduced the use of Geographic Information Systems (GIS) and Expert Systems (ES) techniques as additional components to the QC process. Operational procedures introduced at NCDC in January 1982 provide for the retention of both the original and corrected values in the digital database. Prior to the advent of these and predecessor procedures, identified errors were often corrected on the manuscript and/or CD publications only. The punched cards (pre-1970s database) were not always likewise updated. Other problems of common occurrence include the loss of punched cards; electronic data processing (EDP) machines mangled cards or misread/mispunched them; the mass transfer from cards to tape resulted in more errors; card storage procedures/transfers and environmental hazards such as humidity and temperature variations adversely affected the integrity of the cards; transfer from cards to FOSDIC (Film Optical Sensing Device for Input to Computers) film medium and then to magnetic tape did not always reproduce the original manuscript data with certainty.

Processing the tremendous volume of historical data requires that GEA be as fully automated as possible. GEA requires a removal of gross errors, presently achieved at NCDC by the double keying (key verification) of current input data, which has been the method of digitization since 1989.

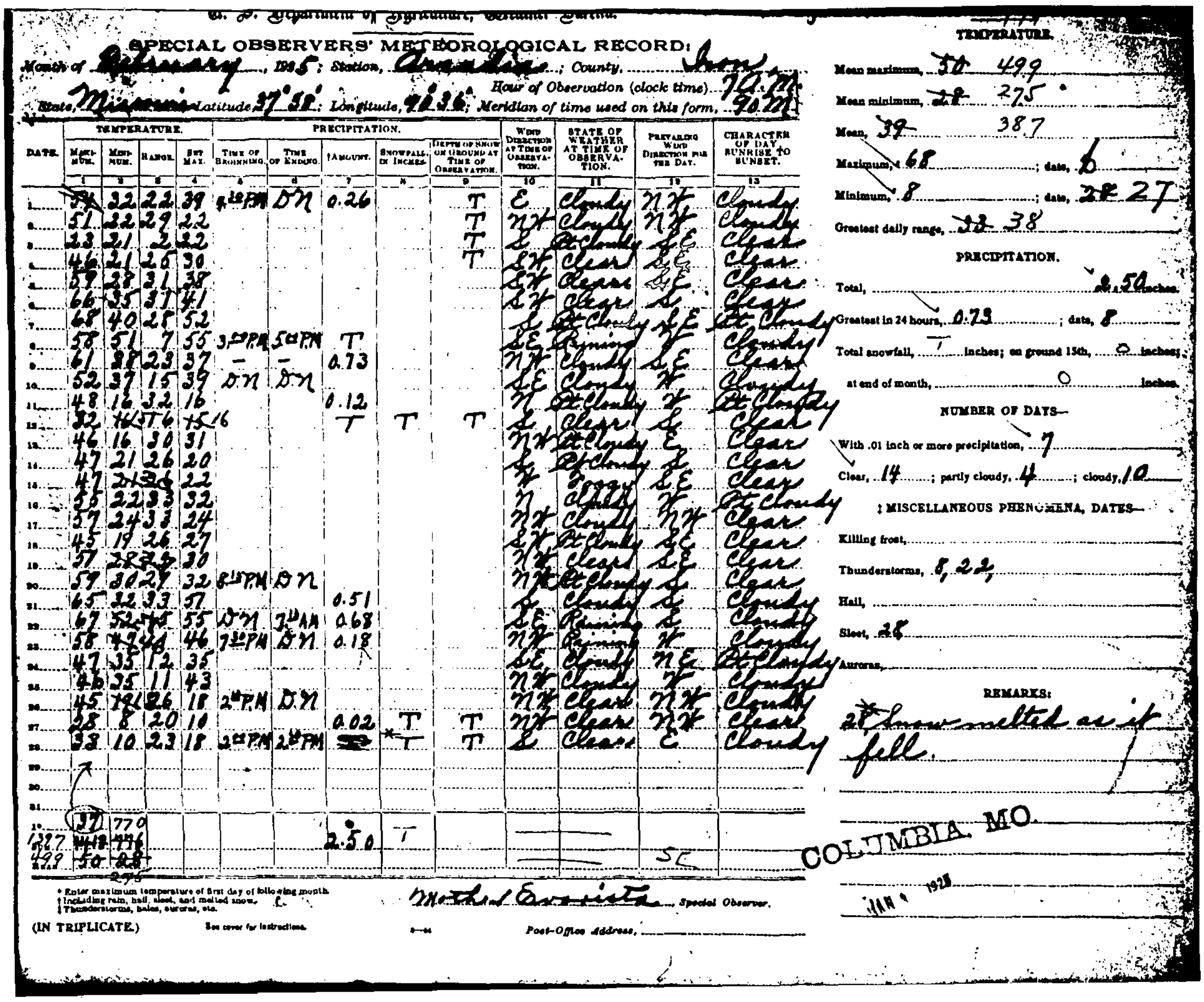
Prior to 1962, it is unknown how many of these data were key verified. Probably, data-entry procedures varied among the many key-entry sites. The single-key entry procedures established at NCDC in 1962 continued until 1989, when it was determined to be both cost and quality effective to begin key verification (double keying). Prior to this, key-entry errors were addressed only when uncovered by the QC procedures in effect at that time. Beginning in January 1982, when the TD3200 database format was established, provision was made for retaining both the originally observed (entered) value and an NCDC-edited (suggested) value for erroneous (suspect) data. Prior to the establishment of TD3200, erroneous observed values were discarded. Certain regionalized data-entry (QC) procedures also resulted not only in the loss of "original" data, but also introduced errors as well.

An examination of Fig. 1, the official observation form for Arcadia, Missouri, for February 1925, provides a typical example of inconsistency errors introduced by regionalized key-entry procedures. Notice that, as recorded, there are no internal inconsistency errors; that is, each maximum temperature is greater than or equal to the minimum temperature of that and the previous day. Given that this station is an A.M. reader, regionalized key-entry procedures dictated a shift of all maximum temperatures to the previous day. Notice the "\" on the first day of the month and the arrow pointing to day 28 at the bottom of the form. Data from this form were entered with the value recorded on day 2 as the value for day 1. Day 3 was entered on day 2, day 4 on day 3, and so on. The maximum value of 23 on day 3, when moved to day 2, produced an inconsistency with the minimum (32°F) of day 1. Similar errors were introduced on days 11, 23, and 26. This method of data entry attempted to alter a 7 A.M. observation time record to a midnight observation time equivalent. In addition, monthly averages now contained in a separate dataset (TD3220) were computed from the original unshifted entries. Restoration of the value to the date as recorded ensures agreement with this monthly dataset (assuming both datasets have otherwise been keyed correctly).

3. Recent quality control developments

In early 1991, NCDC began a pilot project entitled Validation of Historical Daily Data (ValHiDD). The plan called for an examination of the efficacy of using the current operational synoptically based quality control system, known as GEA (Geographical Edit and Analysis) (Reek and Owen 1991), to validate historical daily data. GEA employs an expert system using specially designed interactive graphical presentations. GEA also employs an automated neighbor-selection module that incorporates manually derived selections as well. Adequate neighbor selection is germane to the spatial comparison techniques used to validate anomalous data, particularly in the western United States, where previous techniques were sometimes inadequate.

Processing the tremendous volume of historical data requires that GEA be as fully automated as possible. GEA requires a removal of gross errors, presently achieved at NCDC by the double keying (key verification) of current input data, which has been the method of digitization since 1989. The gross errors present in the historical database statistically corrupt the GEA assessment of good data. Though the num-



U. S. Department of Agriculture, Weather Bureau.

SPECIAL OBSERVERS' METEOROLOGICAL RECORD:

Month of February, 1925; Station, Arcadia; County, Iron

Hour of Observation (clock time) 7 A.M.

State, Missouri Latitude 37° 38'; Longitude, 90 36; Meridian of time used on this form, 90 M.

	TEMPERATURE.				PRECIPITATION.								
DATE.	Maximum.	Minimum.	Range.	Set Max.	Time of Beginning.	Time of Ending.	†Amount.	Snowfall, in Inches.	Depth of Snow on Ground at Time of Observation.	Wind Direction at Time of Observation.	State of Weather at Time of Observation.	Prevailing Wind Direction for the Day.	Character of Day Sunrise to Sunset.
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	~~74~~	32	22	39	4^10 PM	DN	0.26		T	E	Cloudy	NW	Cloudy
2	51	22	29	22					T	NW	Cloudy	NW	Cloudy
3	23	21	2	22					T	S	Pt Cloudy	SE	Clear
4	46	21	25	30					T	SW	Clear	SE	Clear
5	59	28	31	38						SW	Clear	SE	Clear
6	66	35	31	41						SW	Clear	S	Clear
7	68	40	28	52						S	Pt Cloudy	SE	Pt Cloudy
8	58	51	7	55	3^20 PM	5^00 PM	T			SE	Raining	W	Cloudy
9	61	38	23	37	–	–	0.73			NW	Cloudy	SE	Clear
10	52	37	15	39	DN	DN				SE	Cloudy	W	Cloudy
11	48	16	32	16			0.12			N	Pt Cloudy	W	Pt Cloudy
12	32	~~16~~	~~T~~6	~~15~~ 16			T	T	T	S	Clear	S	Clear
13	46	16	30	31						NW	Pt Cloudy	E	Clear
14	47	21	26	20						SW	Pt Cloudy	S	Clear
15	47	21	26	22						W	Foggy	SE	Clear
16	55	22	33	32						N	Cloudy	W	Pt Cloudy
17	57	24	33	24						NW	Cloudy	NW	Clear
18	45	19	26	27						SW	Pt Cloudy	SE	Clear
19	57	28	29	30						NW	Clear	SE	Clear
20	59	30	29	32	8^15 PM	DN				NW	Pt Cloudy	S	Clear
21	65	32	33	57			0.51			S	Cloudy	S	Cloudy
22	67	52	15	55	DN	7^00 AM	0.68			SE	Raining	S	Cloudy
23	58	47	[illegible]	46	7^30 PM	DN	0.18			NW	Raining	W	Cloudy
24	47	35	12	35						SE	Cloudy	NE	Pt Cloudy
25	46	35	11	43						NW	Cloudy	W	Cloudy
26	45	19	26	18	2^00 PM	DN				NW	Clear	NW	Cloudy
27	28	8	20	10			0.02	T	T	NW	Clear	NW	Clear
28	33	10	23	18	2^00 PM	2^45 PM	[illegible]	*T	T	S	Clear	E	Cloudy
29													
30													
31													
1*	37	770											
1327	1418	776					2.50	T					
499	50	28										SE	

275

* Enter maximum temperature of first day of following month.
† Including rain, hail, sleet, and melted snow.
‡ Thunderstorms, halos, auroras, etc.

Mother Evarista, Special Observer.

(IN TRIPLICATE.) See cover for instructions. Post-Office Address, ...

TEMPERATURE.

Mean maximum, ~~50~~ 499

Mean minimum, ~~28~~ 275

Mean, ~~39~~ 387

Maximum, 68; date, 6

Minimum, 8; date, ~~28~~ 27

Greatest daily range, ~~33~~ 38

PRECIPITATION.

Total, 2.50 inches.

Greatest in 24 hours, 0.73; date, 8

Total snowfall, T inches; on ground 15th, 0 inches;

at end of month, 0 inches.

NUMBER OF DAYS—

With .01 inch or more precipitation, 7

Clear, 14; partly cloudy, 4; cloudy, 10

‡ MISCELLANEOUS PHENOMENA, DATES—

Killing frost, ...

Thunderstorms, 8, 22,

Hail, ...

Sleet, 28

Auroras, ...

REMARKS:

28th Snow melted as it fell.

COLUMBIA, MO.
JAN 1925

FIG. 1. Copy of original manuscript from Arcadia, Missouri, February 1925, showing introduction of maximum temperature versus minimum temperature of previous day internal inconsistency errors.

ber of errors in the historical data is small (estimated at less than 0.05% of all values), the sheer volume of data and systematic nature of some of the erroneous values exceeds the present capabilities of GEA. In addition, resorting the data into a synoptic format is a nontrivial task, since the historical database (TD3200) is organized in station element sort.

A potential resolution to the problem of efficient access to synoptically sorted data lies in a parallel development effort at NCDC for the spatial display of historical climate data, known as CLIBASE. CLIBASE is a packed binary database subset organized in synoptic sort (Fig. 2). It is via the CLIBASE that NCDC plans to perform the GEA spatial analysis and also improve upon the serial completeness of the database. Before this can be attempted, a removal of the gross errors, especially those that defy the CLIBASE packing algorithms, must be accomplished.

The TD3200 database requires a filtering process using a set of rules that recognize the way observers record(ed) data, the way card-punch operators key(ed) the data, and the systematic nature by which EDP procedures corrupt(ed) the data. By the fall of 1991, NCDC, in cooperation with the Soil Conservation Service (SCS) Climate Data Access Facility (CDAF), developed, tested, verified, and validated a rules-based method for detecting and correcting discrepancies in the TD3200 database. The gross errors can now be identified, categorized, and eliminated from the TD3200 database.

4. Quality control techniques

Certainly, many approaches exist and have been tried in an effort to improve the quality of the digital

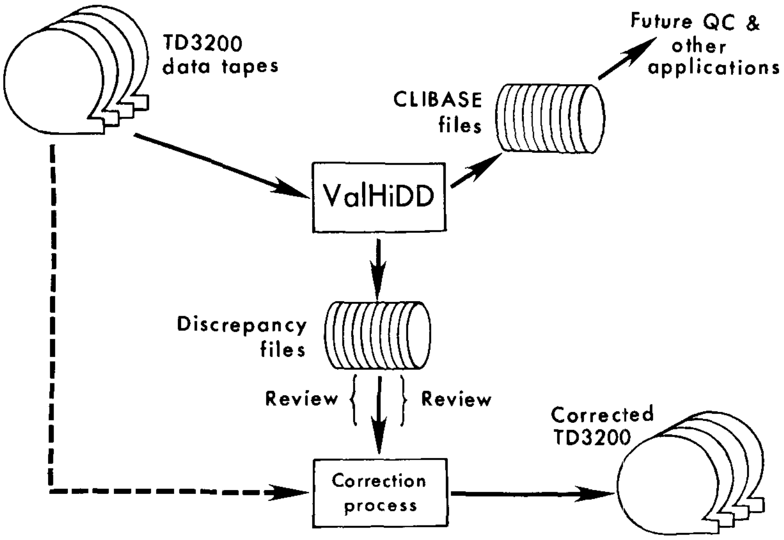


FIG. 2. ValHiDD data-flow diagram.

daily data for the 25 000 stations spanning the past 100 years. Wallis et al. (1990), for example, conducted a filtering process in which TD3200 data failing to pass *gross-limits criteria* were flagged as suspect but were not adjusted or deleted. NCDC often utilizes statistical methods (i.e., standard deviation outlier elimination) when performing applications. The ValHiDD system differs in that it is an inference process, based on a predetermined, systematic, empirically derived set of rules that provide for both the recognition and correction of erroneous data.

According to Moninger and Cote (1987), an inference engine (IE) is the part of an expert system (ES) that attempts to apply the rules from a knowledge base to the data in working memory to converge on a solution. In the ValHiDD case, knowledge of how volunteer observers recorded and made observations, as well as knowledge of card-punching techniques and storage problems, nourished the formulation of a rather large set of inference rules. The resultant IE is a well-structured Fortran program operating on a BNF (Backus Normal Form) structure that converges on a verify/fix/flag-as-suspect solution. BNF is a formal language structure first used for syntax parsing in the design of the ALGOL-60 programming language. It is an effective tool for orchestrating the examination of daily weather observations.

Facilitating the identification of errors and suspected errors while at the same time trying to prescribe correct values is no easy task. The method chosen was to build a discrepancy file.containing the station identification; date of the offending value; the data element in error (#); TMAX, TMIN, PRCP, SNOW, SNWD, and an error code (EC) that identifies the rule violated; the severity of the infraction; and the suggested prescription for remedy. Error codes are two-digit numerical values organized into rule groups. The first digit (1–9) signifies the rule group responsible for its production. The second digit (1–9) indicates a specific corrective action, if odd, or a level of suspicion (but no action), if even. As shown in Fig. 2, the discrepancy file can then be used as control input in the subsequent update of TD3200. The software was developed by first designing a BNF grammar that differentiates good-data and good-data-relationships from bad ones. Data are examined 3 days at a time, with the exception of some longer time-series continuity tests. In the ValHiDD BNF grammar, a successful terminal stack or "Good_Day goal symbol" is reached according to the BNF example given in Fig. 3.

For those unfamiliar with BNF, the symbols Good_max, Good_min, etc., represent tokenized symbols (called tokens) placed onto a program stack. Individual lines are called promotions. In the cases where data are also manipulated, they are called productions. The promotion/production flow is read bottom up. A grammar is considered complete when the uppermost promotion, known as the terminal stack token or goal symbol, is reached, indicating a successful transversal of the grammar structure. The ";" indicates "followed on the stack by." The ";" indicates "promote stack to the value on the left when stack contains tokens on the right." The "," means "or" and the "." signifies the end of a promotion/production group. So, a Fixed_max is defined as a maximum temperature that failed some condition (Bad_max = violates etc. lines 22–24), which was then followed by an Ok_max resulting from a correction procedure at line 21 (procedure not shown), the result of which successfully passed all maximum temperature checks leading to the production of the Fixed_max token (line 15) as the next stack entry. Had the corrected maximum failed any test, a second Bad_max token would have been added to the stack (agaih at lines 22–24) and the resulting sequence "Bad_max; Bad_max" (line 17) would have dictated the production of a "Delete_max" operation (line 17). The "Delete_max" token would become a "Good_max" (line 11) just as an initially "Ok_max" (line 09) or "Fixed_max" (line 10) would. This demonstrates the grammar facility that disallows an erroneous value from passing successfully through ValHiDD without either being corrected or deleted.

This hierarchical grammar insists that all data elements be validated, fixed, or deleted before proceeding to the next element and that all elements must be likewise processed before proceeding to the next day. The BNF grammar provides for an effective coding technique, which insists that all attempted corrections pass the same tests as the original data. The stacked token values assigned to each production provide for the efficient and succinct orchestration of the integrated logic; that is, we can conveniently keep track of where we are, how we got there, what has been done, and what is to be done.

5. ValHiDD rules

The errors detected by ValHiDD fall into two distinct categories: digitizing errors and observer errors. Digitizing errors are often the most obvious and are thus reasonably easy to locate. A maximum temperature of 548° is an undisputable error. Generally, observer errors are much more subtle. Repeating a maximum temperature for 3 days is possible, but repetition of the maximum or minimum for 5 days, 10 days, or longer becomes increasingly unlikely. Specific sets of rules, divided into logical groups, have been defined for specific types of errors, with nine rule groups in all.

a. Rule group 1: Limits check

The first rule group checks against an extremes table (i.e., it is a limits check). An extremes table was constructed by state and month for maximum and minimum temperature and precipitation. It should be noted that TD3200 temperature, rainfall, snowfall, and snow depth are represented in whole degrees Fahrenheit, hundredths of inches, tenths of inches, and whole inches, respectively. For snowfall, a state annual extremes table was developed, and for snow depth a single extreme value was used (451 inches). Failure of the limits check requires the value to exceed (not exceed, for minimum temperature) the table value for the current month, the previous month, and the subsequent month. Failure at this stage may result in deletion from the file.

Certain conditions, such as a temperature of 198°F, may qualify for a subtraction of 100, if it can be determined that the resultant value (in this case, 98) is quite reasonable. It was determined by empirical review that a common systematic artifact, characteristic of EDP (electronic data processing) methods, is the presence of a "1" in the first position of the three-position maximum-temperature field. A sign-reversal algorithm may also successfully adjust a −89 maximum temperature. The erroneous negative sign in maximum and/or minimum temperature fields is another EDP artifact. Provision is also made for the possibility that temperature was entered to tenths of degrees, as in the earlier example of 548. Consider the examples shown in Table 1.

Values that are within 4°F (inclusive) of the all-time extreme are coded as suspect unless a condition of persistence is determined to be present. In this way, isolated near-extreme values are also noted but not deleted. For values where a fix is indicated and the predetermined fix does not result in a reasonable value as determined by a second pass through all subsequent and previous checks, the datum is coded for deletion.

b. Rule group 2: Internal inconsistency check

The second rule group tests for internal inconsis-

```
00  Good_day     = Good_value; END.
01  Good_value   = Good_max,
02               = Good_value;Good_min,
03               = Good_value;Good_prcp,
04               = Good_value;Missing_prcp,
05               = Good_value;Good_snow,
06               = Good_value;Missing_snow,
07               = Good_value;Good_depth,
08               = Good_value;Missing_depth.
09    Good_max   = Ok_max,
10               = Fixed_max,
11               = Delete_min.
12    Good_min   = Ok_min,
13               = Fixed_min,
14               = Delete_min.
15   Fixed_max   = Bad_max;Ok_max.
16   Fixed_min   = Bad_min;Ok_min.
17  Delete_max   = Bad_max;Bad_max,
18               = Missing_max.
19  Delete_min   = Bad_min,Bad_min,
20               = Missing_min.
21      Ok_max   = etc....;etc,
..               = etc....;etc.
22     Bad_max   = Violates_limits,
23               = violates_etc,
24               = violates_etc,
..               = violates_etc.
```

FIG. 3. Example of BNF grammar used to define ValHiDD logic flow.

TABLE 1. Sample systematic temperature discrepancies.

(Sign reversal)			(Add 100)			(Subtract 100)			(Entered to tenths)		
Day	TMAX	TMIN	Day	TMAX	TMIN	Day	TMAX	TMIN	Day	TMAX	TMIN
N − 1	87	54	N − 1	103	76	N − 1	87	52	N − 1	56	27
N	−85	52	N	5	77	N	185	52	N	548	31
N + 1	78	56	N + 1	108	75	N + 1	78	56	N + 1	67	37
Result = 85			Result = 105			Result = 85			Result = 55		

tencies. Internal inconsistencies come in different forms, from the maximum temperature recorded being less than the minimum temperature, to the snow depth increasing by at least three inches without snowfall on the current or previous day. Other internal inconsistencies involve a maximum temperature that is less than the minimum of the previous day.

Standard fixes in the internal inconsistency group include the swapping of the maximum and minimum temperatures, assigning the maximum of today equal to the minimum of the previous day, or deleting the datum most likely in error. In some cases, due to specific observation and/or key-entry procedures (Fig. 1), it is necessary to move all daily maximum values one day forward. Reek et al. (1991) offers a detailed description of the ValHiDD rules and automated repair procedures for these cases.

c. Rule group 3: Flatliner temperature check

The third rule group involves the identification of flatliner temperatures. A flatliner is defined to be a run of at least five consecutive days of the same maximum or minimum temperature. Five identical consecutive values are coded as suspect. A run of more than five days but less than ten is coded as highly suspect. Any run greater than or equal to ten days' duration is coded for deletion, with the exception of the first day, which is coded as suspect. An exception to this rule is made for places such as the Caribbean, where such events are commonplace.

d. Rule group 4: Excessive diurnal range check

The fourth set of rules checks for excessive diurnal range, an idea adopted from Wallis et al. (1990). These checks are very similar to those mentioned in group 2. When the diurnal range exceeds 75°, both maximum and minimum temperatures are examined for serial continuity as described in the spike tests of rule group 6, to be discussed later. If evidence of discontinuity is found for one variable or both, the value(s) is marked for deletion.

e. Rule group 5: Invalid PREC/SNWF/SNWD relationship check

The fifth group of rules involves the relationship among precipitation, snowfall, and snow depth. These rules seek to identify those values that are inconsistent with the other two elements. For example, when snowfall is indicated and the snow depth is increasing proportionately, a zero amount for precipitation is invalid. Additionally, a trace value for snowfall with legitimate snowfall temperatures infers that a zero precipitation must be changed to a trace amount. This rule alone accounted for 25% of all the fixes to precipitation. Group 5 expects the ratio of precipitation to snowfall to be within reason, based on the maximum and minimum temperature according to Table 2. In some cases, the rules allow for the snowfall or precipitation to be divided by or multiplied by 10, thus allowing for a fix, or a standard precipitation-to-snowfall ratio may be used to estimate (replace zero nontrace)

TABLE 2. Standard snowfall-to-precipitation ratios.

	Temperature ranges (in degrees Fahrenheit)						
	>28	20 to 27	15 to 19	10 to 14	0 to 9	−1 to −20	< −20
Ratios	20:1 or less	10:1 to 30:1	20:1 to 40:1	30:1 to 50:1	40:1 to 60:1	50:1 to 70:1	60:1 to 100:1

precipitation amounts. We have found that a significant number of these errors result from the existence of a zero where a blank is intended. While zero values are changed to a missing value, ValHiDD makes no attempt to estimate replacements for missing data in any rule group.

f. Rule group 6: Temperature spike check

The sixth group of rules identifies spikes in the temperature series. A spike is defined to be the smallest absolute result from the comparison of the singular differences of three consecutive days, centered on the day in question. A nonmonotonic change must also be present. Spikes greater than 50° result in the value being marked for deletion. Spikes of greater than 40° but less than 50° are marked as suspect. Smaller spikes are used as determinants by rule groups 2, 3, and 4 above.

g. Rule groups 7 and 8: Multiple rule-group failures check

The seventh and eighth rule groups recognize the coexistence of two or more simultaneous rule failures and assign combination code flags accordingly. For example, a near-record maximum temperature assigned a moderate suspect code might later be detected as a discontinuity by the spike test or as a corespondent in the diurnal range test.

h. Rule group 9: Failed fix check

A final rule group, the ninth level of control, recognizes those elements that were subject to a correction but failed the second iteration (i.e., attempted fixes that failed the rule tests). It is important to keep track of these events for system-monitoring and feedback purposes. Human review of these events has played an important role in the empirical development of the correction methods used and the erudition (determination) of certain systematic properties attributed to many of the errors.

It should be noted that particular attention is given to zero values, as it has been determined that missing data [e.g., maximum temperature not recorded (blank)] were often keyed as zero. ValHiDD contains the necessary mechanisms to replace these zeros with "missing" values. The ValHiDD procedures, as well as all the ValHiDD rules, have been incorporated into the present operational QC system at NCDC.

6. Test results

A 1300 station subset of TD3200 data (approximately 15% of the database) was used to test the software during the iterative process of empirical development of the inference rules base. These stations, recording both temperature and precipitation data, represent the long-term stations in the database and are representative of all geographical areas of the conterminous United States. Over 1000 errors and their automated corrective actions were manually verified against the original manuscripts on file at the NCDC to measure and improve ValHiDD's ability to correctly reproduce the original manuscript data. The efficacy of the sign reversal and +/– 100 algorithms were found to be extremely high—in fact, in excess of 99%. The Max/Min swap algorithm proved to be the least effective, verifying at less than 60%. This algorithm is undergoing reexamination. In all instances, however, the correction produced more reasonable data. The remaining algorithms had a verification rate of 95% or better.

The results of the processing, with respect to the data itself, were also very encouraging, as only a 0.04% data error rate was encountered for the 1300-station test dataset. However, over 75 000 (the number is much higher if individual days of each flatliner scenario are considered) potential discrepancies were flagged, deleted, and/or corrected, a number far too large (projected at nearly 1 million for the entire database) not to warrant attention. In addition, a second test dataset comprising 138 stations, known as HCN/D (Historical Climatology Network Daily) data (Hughes et al. 1991), was borrowed from another NCDC project involving intensive manual subjective review and validation. The data for these 138 stations, validated and unvalidated, were processed, and the results compared. ValHiDD performed exceedingly well, as no presumed correct data from the previously validated control dataset were flagged or manipulated. In addition, ValHiDD uncovered a few missed or improperly corrected items. ValHiDD data corrections applied to the 138 stations from the unvalidated source were, with only minor exception, equal to those of the human (validated) manually intensive correction process.

Table 3 summarizes the results of the 1300 station test by element and outcome. Several entries stand out, such as the absence of fixes for snow depth and the seemingly large number of fixes for the maximum temperature. Since no rules were defined for fixing snow depth (snow depth can be present even with very warm temperatures, since hail amounts are included) other than the deletion of extreme values, a zero fix total is to be expected. The large number of fixes for the maximum temperature is explained by the large number of A.M. reader stations for which the maximum temperature was shifted, during the key-entry process, to the previous calendar day, thus creating an occasional inconsistency with the mini-

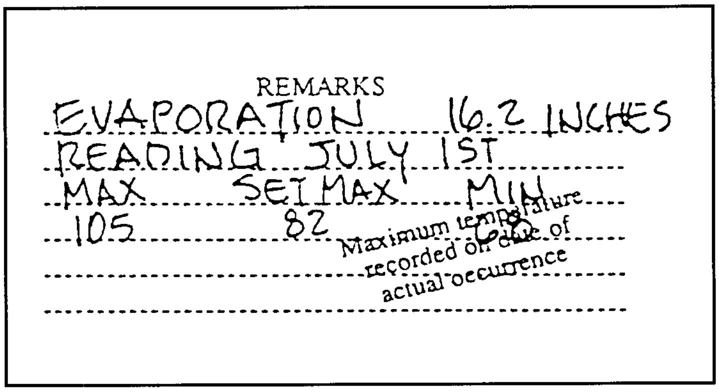
TABLE 3. The number and percentage of discrepancies by element taken in the 1300-station test.

	TMAX	%	TMIN	%	PRCP	%	SNOW	%	SNWD	%	TOTAL	%
Deletes	1481	.0052	1295	.0045	465	.0016	3263	.0121	42	.0002	6546	.0047
Fixes	16886	.0592	707	.0025	6325	.0220	193	.0007	0	0	24111	.0175
Suspect	2335	.0082	653	.0023	12157	.0423	14786	.0547	5518	.0218	35449	.0257
Flatliners	5411	.0190	7759	.0272	—	—	—	—	—	—	13170	.0095
Total	26113	.0916	10414	.0365	18947	.0659	18242	.0675	5560	.0220	79276	.0574

mum of the previous day (Fig. 1). The largest category of precipitation errors involves zero precipitation with legitimate snowfall. The largest number of suspected temperature discrepancies is the 5–9-day flatliner scenario.

The largest category of temperature internal inconsistencies was found to be the maximum temperature being less than the minimum of the previous day, which is a classic error of what is commonly referred to as a "TMAX shifter." The term shifter has come about due to the visual presentation given when an offending station's data for a given month are displayed as a time series along with data for neighboring stations that have followed the prescribed reporting procedures. The resultant temporal shift is often readily apparent. It should be noted that temporal internal inconsistencies fall into two classes. Class 1 is an intentional effort on the part of the A.M. observer to report the maximum temperature on the perceived day of occurrence. This is often indicated by a qualification stamp on the observation form, as shown in Fig. 4. In this case, the observer exercised judgment as to the actual day of occurrence. Class 2 is the result of data-entry procedures for pre-1948 data that systematically moved (shifted) each TMAX for A.M. readers back one day, as shown earlier by Fig. 1. Since observer judgment was not involved, data for this type of station contain far greater numbers of internal inconsistencies than the class 1 type. It is this feature that distinguishes the two shifter signatures.

The largest category of temperature internal inconsistencies was found to be the maximum temperature being less than the minimum of the previous day, which is a classic error of what is commonly referred to as a "TMAX shifter."


REMARKS

EVAPORATION 16.2 INCHES

READING JULY 1ST

MAX SET MAX MIN

105 82 68

Maximum temperature recorded on date of actual occurrence

FIG. 4. Qualification stamp.

The large number of suspect values in the precipitation and snowfall categories (Table 3) is the result of internal inconsistency (ratio) problems, presumably caused by recording SNOW to whole inches rather than to tenths, or undermeasurement of the catch in the raingage. ValHiDD rules cannot always determine which parameter was in error and, therefore, which to adjust, so both are flagged as suspect.

Class 2 temporal inconsistency errors, included in rule group 2 (Table 4), ended in 1948. This is just as one would expect, since the data-entry procedures changed at that time. Class 1 errors, although contrary to the present rules of observation, still exist today. The steady increase in the number of cases of flatliners (group 3), 90% of which are of exactly 5 days' duration, is a mystery. Review of original manuscripts has revealed these to be present on the manuscript and not EDP artifacts. Further investigation is planned. The increased number of suspect flags in the 1970s

Table 4. Decadal analysis of discrepancies versus rules groups.

Rule Group	Decades										TOTAL
	<1900	1900	1910	1920	1930	1940	1950	1960	1970	1980–87	
1	23	68	60	157	282	198	229	43	7	12	1079
2	1179	2101	3255	4250	4501	3870	879	291	74	638	21038
3	67	410	491	526	1204	1314	1946	2137	2602	2423	13120
4	2	7	5	8	9	11	5	0	3	4	54
5	199	522	893	1012	1511	1195	758	706	642	891	8329
6	3	7	11	8	16	18	61	9	5	5	143
7	1	1	5	4	1	3	3	0	0	1	19
8	0	4	26	3	7	2	1	1	0	1	45
Suspect	421	1688	2967	3009	4534	4118	3703	4810	6933	3266	35449
Category	Number of elements processed by decade (thousands)										
TMAX	344	1018	1421	1802	2782	3212	4758	4748	4687	3753	28524
TMIN	344	1018	1420	1804	2786	3211	4751	4748	4702	3755	28536
PRCP	376	1037	1445	1821	2793	3221	4799	4775	4738	3765	28769
SNOW	241	892	1279	1628	2530	2998	4656	4608	4551	3634	27017
SNWD	128	727	1023	1337	2267	2775	4575	4504	4456	3523	25316
TOTAL	1432	4692	6588	8391	13157	15417	23539	23382	23134	18430	138161

are due to an unusually large number of snowfall entries of zero when precipitation and cold temperatures (–21°F) are present, and snow depth is not available. As explained earlier, ValHiDD is not always able to establish which element, if any, is in error. Therefore, for the zero snowfall case in question, suspect flags are assigned to both precipitation and snowfall. A sample list of actual discrepancies is shown in Table 5.

7. Conclusion and intentions

Although the number of discrepancies uncovered and resolved is small compared to the total number of data values examined, their removal/correction is significant for several reasons:

- They represent the most conspicuous errors and are often culprits in statistical discordancies.
- They serve to discredit the database.
- They contribute to confusion in establishing accurate historical climatological record means and extremes.
- They infect climate models and climate summaries.
- They are systematic in nature.

ValHiDD has proven an effective tool for removing those errors most often highlighted in critical reviews of the TD3200 database. The ValHiDD correction-to-deletion ratio of 4:1 verifies the systematic nature of those errors. By removing these derogating values, the confidence level of the database is greatly elevated.

It is the intent of NCDC to fine tune the ValHiDD rule

TABLE 5. A sample list of actual discrepancies from the state of Arizona.

EC	#	STN ID#	Year	JDY	TMAX	TMIN	PRCP	SNOW	S NWD	Station name	CORR
17	2	021026	1898	115	100	−60	0	0	0	Buckeye	60
17	1	024639	1935	340	−67	38	0	0	0	Kingman	67
17	1	025467	1924	152	−77	missing	trace	0	0	Mesa	77
17	1	027370	1947	312	−61	24	0	0	0	Sacaton	61
17	1	029652	1930	352	−73	42	0	0	0	Yuma Citrus Sta	73
17	1	029652	1951	057	−70	38	0	0	0	Yuma Citrus Sta	70
19	1	021614	1965	243	187	64	19	0	0	Childs	87
19	1	021614	1965	244	198	64	0	0	0	Childs	98
19	1	028815	1905	259	1	71	0	0	0	Tucson Univ of	101
61	1	022659	1982	071	8	missing	0	0	0	Douglas	Delete
42	1	021634	1936	364	51	−25	58	58	missing	Chinle	Suspect
12	2	027435	1931	317	25	−21	0	0	missing	St. Johns	Suspect

base and process the entire TD3200 database through the ValHiDD program. The resulting period of record discrepancy file will be initially reviewed and verified at the CDAF in early 1992. An updated period of record-merged TD3200 series will be produced by NCDC later in 1992. ValHiDD correction files and implementation software will be made available for use by past recipients of TD3200. NCDC will continue to pursue the historical spatial/areal validation and serial completeness of the TD3200 database.

Acknowledgments. The authors wish to thank Eric Gadberry, Pam Hughes, and Elaine Mason of NCDC for their invaluable assistance in the data-acquisition and evaluation phases of this project. We also wish to express our appreciation to Jerry Sawyer of NCDC for his untiring efforts in the manual review and comparison of the ValHiDD discrepancy file to the original manuscripts. We thank Scott Miller, also of NCDC, for preparing the figures and tables presented in this manuscript.

We also acknowledge the Soil Conservation Service Climate Data Access facility for their partial funding of this project, and, in particular, Phil Pasteris for his valuable input and suggestions, Kelly Redmond of the Western Regional Climate Center, Myron Molnau of the University of Idaho, and Nathaniel Guttman of NCDC for their review of an earlier version of the rules criteria.

References

Duchon C., M. Mignogno, and R. Knight, 1982: *Cooperative Data Processing Systems Guide.* Internal NCDC document, 200 pp. [Available from NCDC, chief, Data Operations Branch.]

Hughes, P. Y., E. H. Mason, T. R. Karl, and W. A. Brower, 1991: *United States Historical Climatology Network Daily (HCN/D) temperature and precipitation data,* ORNL/CDIAC-50, NDP-042, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN, 127 pp.

Moninger, W. R., and Cote, D. F., 1987: Summary of the First Conference on Artificial Intelligence Research in Environmental Sciences (AIRES). *Bull. Amer. Meteor. Soc.* **68,** 793–800.

National Climatic Data Center, 1991: *Surface Land Daily Cooperative, Summary of the Day TD-3200.* NCDC, 25 pp.

National Weather Service, 1987: *Cooperative Program Management, Weather Service Operations Manual B-17* (revised), NOAA-NWS, Silver Springs, MD, 50 pp.

Reek, T., and T. Owen, 1991: *GEA Documentation and Users Guide,* Internal NCDC Document, 30 pp. [Available from NCDC, chief, Data Operations Branch.]

——, and M. Crowe, 1991: Advances in quality control technology at the National Climatic Data Center. Preprints, *Seventh Intl. Conf. on Interactive Info. and Processing Sys. for Meteorol. Oceanogr. and Hydrol.,* New Orleans, Amer. Meteor. Soc. 397–403.

——, S. R. Doty, and T. Owen, 1991: *ValHiDD Documentation and Users Guide,* Internal NCDC document, 20 pp.

Wallis, J. R., D. P. Lettermaier, and E. F. Wood, 1990: A daily hydroclimatological data set for the continental U.S. IBM Research Report, RC 16607, 20 pp.