EASTERN SNOW CONFERENCE
Dearborn, Michigan 1994

A Commentary on the Pitfalls of Cleaning Snow Data

V. K. JONES[1]

ABSTRACT

Those who have worked with climatological data extensively are well aware of its problems and pitfalls, despite error-checking procedures applied by the National Climate Data Center. Snowfall data, by its inherent nature, has more problems than most other climate data.

Official snowfall data for Michigan for the snow seasons of 1988-89 through 1991-92 were further cleaned and corrected by knowledge-based techniques. This paper deals with problems and some solutions involved in the task, rather than the data. It is hoped that this paper will provide some awareness and assistance to others involved in such tasks, and to users of both official and cleaned data.

Key words: Snowfall, climate data, data management, data cleaning

INTRODUCTION

The National Climate Data Center (NCDC) processes, publishes, and archives a massive amount of climate data. These data come from sources ranging from professionally-operated first-order National Weather Service Offices (WSO's) to volunteer amateur observers. Despite effective computerized error-checking programs, these published official data still contain some missing, partial, and erroneous data present in the original observers' reports.

In this era of electronic number-crunching, it is often wrongly assumed that data obtained in published form, on magnetic disc or CD's, or interactively (often sanctified by responsible agencies), are highly accurate. Such an assumption can lead to unrealized and sometimes serious errors in your analyses and results.

Neff (1977) noted that records, once collected and published, often gain an aura of respectability and precision that is beyond tolerances that can legitimately be assigned to them. He urged that no more confidence be placed in the records than is justified by the measurement techniques.

These climatic records often have economic as well as meteorological significance. For many applications, such as hydrological studies or the allocation of state funds for snowfall removal, more accurate data are needed. Considering the often erratic areal distribution of snowfall and the nature of the missing and erroneous data, mathematical computer techniques and algorithms for correction and enhancement of raw datasets are generally inadequate. Techniques which allow for the input and application of human knowledge are required.

Such input includes knowledge of local patterns, topography, storm system behavior, snowfall-generating mechanisms, local and large-scale atmospheric circulation, time of year, temperatures, wind directions, other perturbing circumstances, and the application of objective and even subjective human judgement. In short, the cleaning and estimation of data becomes an art as well as a science.

Working under a contract with Michigan Office of the State Climatologist for the Michigan Department of Agriculture, I provided quality control and estimation of missing data for available snowfall records within the state of Michigan for the 1988-89 through 1991-92 snow seasons.

This paper represents experience gained during the above work and also work with other climate data over the years. It is not intended as a reproducible presentation of data, a "how-to" cookbook, justification of results, or even as a litany of intellectual misery. Its purpose is to alert the data processor and data user to real-world problems in most snow data sets, and to provide some general understanding of data-cleaning procedures.

FACTORS RELATED TO SNOWFALL MEASUREMENT

Sykes (1993) noted several factors related to snowfall measurements:

1) Site exposure along with site location relative to local geography; 2) Actual measurement means such as snow boards, catching container with saline/oil solution, weighing gauge, etc.; 3) Experience of observer(s); 4) Number of daily measurements; 5) Prevailing weather conditions, especially winds and variations in cloudy and sunny periods; 6) Actual times-of-day for measurements; Visibility and snowflake diameters in relation to surface snowfall accumulations;

[1] Adjunct Professor of Geography; Western Michigan University; Earth Systems Services; 3648 Packard Highway; Charlotte, Michigan 48813 USA

and, plain estimation when outside conditions are particularly difficult.

SNOWFALL MEASURING METHODS AND ACCURACIES

Neff (1977), in a review of literature regarding the accuracy of various types of rain gauges, reported agreement among investigators of a 5-10% average error in rainfall catch, and that the error is inconsistent since it varies from zero to 70%, depending on wind velocities. They also agreed that errors in measuring snowfall are greater than errors in measuring rain, because snow flakes are more susceptible to wind effects.

Goodison (1978) reported on a field study to assess the accuracy and comparability of precipitation gauge measurements of snowfall in Canada. At speeds up to 12 mi/hr (5.5 m/sec), the Nipher shielded gage, used as the Canadian standard, was within 10% of "ground true" as measured on snow boards at a sheltered site. At 11 mi/hr (5 m/sec), Alter-shielded Fischer/Porter and Universal (Belfort) gauges caught 40 and 51% of ground true water equivalent precipitation, respectively, while the figures for the same types unshielded were 21 and 32%. A Nipher shield is a solid trumpet-shaped device facing upward. An Alter shield is a ring of free-swinging leaves a few inches away from the gauge orifice.

A career NWS meteorologist and former Climate Program Manager for Michigan informs me that the traditional method for measuring snowfall in Michigan is to select an apparently representative site, then take the average of three or more measurements with a calibrated measuring stick (Baldwin, 1994). Unshielded Universal (Belfort) weighing type gauges are fairly widely used in Michigan to determine time and rate of precipitation. Most cooperative observer stations use the standard 8-inch (20 cm) metal can rain gauge with dip stick. Station snowfall information does not include the type of measuring method used.

CONSIDERATIONS IN DATA EVALUATION

Station Quality

A few stations will always be less than reliable, due to long gaps in the data or a tendency toward obvious errors. It is difficult and tenuous to estimate large amounts of missing data. Some stations may have to be eliminated from your dataset.

Harder to detect are stations with questionable or inconsistent measurements. Erroneous data are worse than useless, since they may lead to wrong information and wrong answers without your being aware of the errors.

Temporal Resolution

The Monthly Climate Data publications (MCD's) by state, produced by NOAA's National Climate Data Center, report both monthly and daily data, although daily snowfall and snow depths are not published for all stations. It is difficult and time-consuming to deal with daily data on a state-wide and annual basis. However, daily data may provide clues to amounts and relationships between stations.

It is important to note the time of observation at the stations under consideration. Comparing daily totals may make no sense unless you realize that one station takes observations at 8 AM while another observes at 5 PM. What time of day did it snow?

Another form of temporal problem deals with time within the snowfall season. The behavior of the atmosphere may be rather different in December than in February. This is especially true with shoreline stations, whether in the case of temperatures near the freezing rain-snow line, or in partial or total ice cover on one or more of the Great Lakes.

Spatial Resolution

In the lee of the Great Lakes, the "lake effect" can have great impacts. For example, a mid-day lake effect snowfall left a measured 13 inches (33 cm) on my car in a Western Michigan University parking lot in Kalamazoo. On the east edge of town there was little more than a trace. Two miles (3 km) further east the pavement was dry. Those who are experienced in regional weather can easily top this with stories of 16 to 24 inches (40 to 60 cm) of partly cloudy just down the road.

Another problem of spatial resolution comes from the fact that we may have only one to three stations per county. Even when all are functional, accurate, and complete they are too sparse to provide a true picture of snowfall distribution and amounts due to unmeasured local variations.

Peck and Brown (1962) developed precipitation and elevation relationships for mountainous areas in Utah, based on good correlations between precipitation and station elevations. Departures of individual stations from the graphic curves were related to physiographic features. Similar relationships for relatively small differences in elevation and local relief appear in annual snowfall isopleth plots for Michigan, in combination with lake sources of precipitable moisture. This is noticeable for the higher elevations in the northern Lower Peninsula, but is especially striking on the Keweenaw Peninsula of Upper Michigan (Figure 1), where a high ridge lies normal to air flow from Lake Superior.

TASKS IN CLEANING OF DATA

There are two primary tasks in the cleaning of snowfall data, as in other data: a) detecting and correcting erroneous data; and b) providing estimates for missing data.

DETECTING ERRONEOUS DATA

Grossly obvious errors may be spotted by perusal of the data in tabular form. In large datasets a computer program may be used to flag figures which are out of allowable tolerances, but manual checking of suspect data is still required. The primary tool for detecting erroneous data is the map. Once the data are plotted on a base map, they are inspected for data values which appear illogical, such as significant differences in nearby stations. Isopleth plotting will make excessive amounts obvious, as well as showing unusually low values at single stations.

However, caution is required. Some data may appear out of line when they do in fact reflect the actual situation. Stations a few miles apart may actually receive significantly different amounts. The Alpena city station, near the shoreline of Lake Huron, often receives much less snowfall than the Alpena airport station about 10 miles (16 km) away - particularly when temperatures are near freezing. The Marquette city station on the shore of Lake Superior may have a seasonal total of only about 60% as much as the WSO station at the Marquette airport, which is some 12 miles (19 km) inland and 800 feet (244 m) higher in elevation.

One must place some confidence in the field observers who are actually at the site and experience the local conditions. Such confidence is justified in most, though not all, cases.

METHODS OF ESTIMATING DATA

Missing Data - or No Occurrence?

Are the data really missing? Observers report only measured snowfall occurrence; there is no negative or zero report. If the station also reports temperatures, the temperature report may be missing also - or it may not! A snowfall-only or precipitation-only station provides no clue. Otherwise, missing data is usually flagged on the data print-out or in the MCD's.

Approaches Developed by Experience - and Inherent Problems

Following are a few methods which have been of help in estimating snowfall amounts - and also of some cautions in using such methods. Users can tailor them to their own needs and approaches, and could probably suggest many more methods.

1. Use of data from adjacent stations. This is not as simple and straightforward as it appears. There are natural differences between adjacent stations due to many factors such as micro-climatic variations and temporal differences in circulation, even on a monthly or annual basis. Comparison by regression techniques should provide reasonable estimates. However, in actual operation this often does not provide estimates which are credible. Another trap is that a closer look at an apparently reliable station may reveal questionable data. This can result in an ever-widening and ever-deepening quagmire of attempted corrections.

2. Use of daily data to aggregate to monthly data. If the monthly data in the MCD are incorrect, it is because the daily figures are inaccurate or incomplete. Use of daily station data from adjacent stations contains all the problems mentioned in Nr.1 above. In addition, time of daily observation may differ between stations. Yet such a method may prove useful.

3. Rules of thumb. The traditional rule of thumb for equating melted equivalent amounts uses a 10:1 snow/liquid ratio. This may occasionally prove useful in interpreting observer records. However, actual ratios may differ widely! A heavy, wet snowfall may have a ratio of less than 5:1, while a dry, fluffy fall, especially at low temperatures, may have a ratio of greater than 20:1. In fact, comparing data within an observer's report may provide apparent ratios which are beyond credibility.

4. Changes in observed snow depths. When both are reported, measured snowfall for the day often varies greatly from change in snow depth from the previous day. A 10-inch (25 cm) snowfall may be paired with a reported 3-inch (7.6 cm) increase in depth, or a 3-inch fall with a reported 4-inch (10

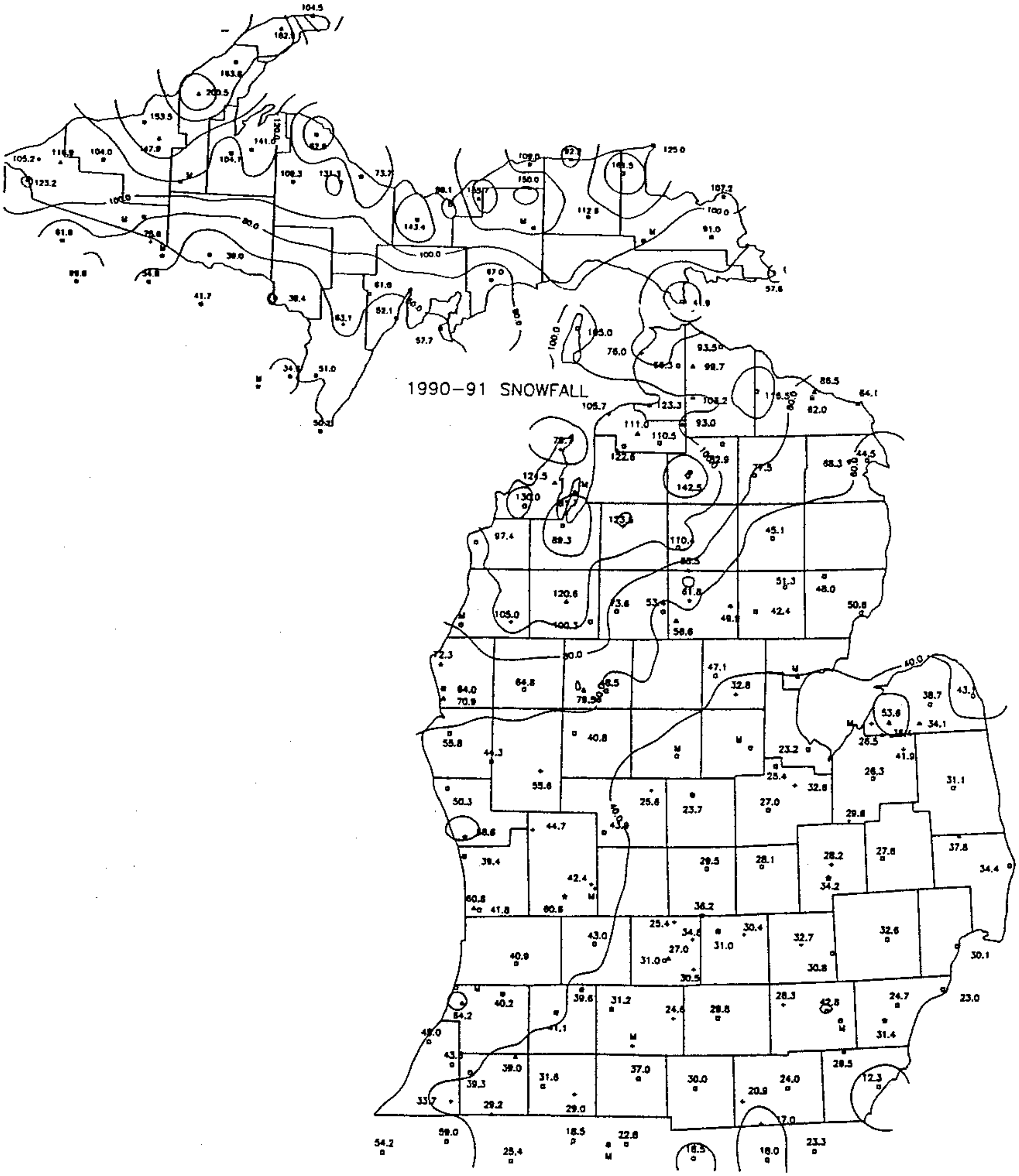

Figure 1. Michigan Annual Snowfall, 1990-91 Season
(after cleaning)

cm) increase. Ambient temperatures, long settling times, or the presence of liquid precipitation may help to explain the variations in some cases. Unfortunately, the reliability of this method as a means of estimating snowfall turns out to be rather low.

5. Plotting storm totals. In some situations, it may be necessary to plot the total snowfall readings from the path of a single storm passing through an area. Isopleth patterns, combined with knowledge of the area, may help to make sense out of apparently erratic readings. This can be time-consuming, but can be enlightening. GIS-type computer techniques would be helpful in making such an approach time-efficient.

6. Use of observers' notes and comments. It is often useful to go back to the original observers' forms such as E-15's, if available. Notes and comments, if any, are often useful. For example, notations of "rain changing to snow" or "windy" may alert you to a problem.

EXAMPLES OF PROBLEMS

The sequence of numbers is sometimes inadvertently switched. Decimals are misplaced or missing. Clarity of penmanship is not necessarily a mandatory skill for observers. Observers may not always approach consistent perfection in their measurement techniques. Some stations, such as those at certain government facilities, may not be manned on weekends.

It is very difficult to measure snowfall under windy conditions. At the Michigan Agricultural Experiment Station site on the MSU campus I placed my snowfall site at the center of an old one-acre (0.4 ha) orchard which had grown up into a thicket. Despite this, I have seen drift accumulation on my snow-boards. When the snow is blown horizontally across the orifice on a precipitation gauge, the percent of catch is anybody's guess!

Sitings and microclimates of local stations may vary greatly, despite efforts to select good locations. Visits to station sites by the data validator, if possible, may provide clues to variations. Some stations tend to be consistently high or low in comparison to nearby stations. There may or may not be a valid or discernible micro-climatological reason for this.

When the snowfall data are to be used to allocate snow removal funds, it is humanly possible for a systematic subjective bias to creep into the measurements when the ad hoc seasonal observation station is located at the road commission garage.

There has been some evidence that a few observers may have melted the daily accumulation of snow, then used a 10:1 ratio to estimate depth of snowfall. Of course this is how all-purpose precipitation gauges work, on a basis of weight or liquid volume. Others may have used the 10:1 ratio from measured snow to get liquid equivalent. Such data must be used with caution.

When you are attempting to estimate missing data for Station A from Stations B and C, you may discover problems from Station C's data which you have previously overlooked. This then throws a previous estimate for Station D into question. A series of attempted corrections, if not brought quickly under control, can result in paralysis of your analysis.

Finally, despite the application of any and all possible valid - and even creative - techniques, you will still end up with a number of data points for which you cannot make a remotely confident estimate.

A SUGGESTED PROCEDURE FOR CLEANING SNOWFALL DATA

Following is a suggested procedure for cleaning sets of snowfall data. The reader is strongly reminded that these are subject to the caveats contained in the "Methods" and "Other Problems" sections of this paper, and to other conditions and situations in which the analyst may find him or herself. Each analyst will develop his/her own appropriate techniques. It is important to resist rationalization and maximize data integrity as much as possible throughout the process.

As confident estimates are made at any point in the process, these former problems are dropped from the "to do" part of the analysis and replaced by data flagged as estimated. Some stations with large gaps or highly unreliable data may have to be rejected.

1. Print out tables of existing data by station and month, by climatic divisions. Missing and incomplete data are flagged.

2. Inspect each station for data which are absent, contain missing data for the month, or are obviously suspect.

3. Plot a map of snowfall by station for each month and for the total season, with missing and incomplete data flagged. Plotting isopleths will often highlight problems with data. Mark these problem locations for further study.

4. Determine seasonal thresholds of snow/no snow for the various regions. For late Fall and early Spring months, determine when earliest and latest snowfalls occurred and thus when data are blank due to non-occurrence rather than due to missing data or non-reporting.

5. Determine regional patterns from isopleth plots, for guidance in evaluating possible problem data and checking later estimates for reasonableness. Especially note regional relief and proximity to large bodies of water, relative to wind directions under various micro- and meso-scale conditions.

6. Check table data against MCD data. Even if your data came from MCD/NCDC sources, you may find some discrepancies. Where published, check daily MCD data. A couple of missing days, when no snowfall occurred elsewhere in the area, may allow confidence in the longer-term figure given despite the lack of record for those days.

7. With a thorough knowledge of the area, it may be possible to make a reasonable estimate from surrounding stations. Beware of stations with quirks of location, topography, microclimate, and other factors which make them atypical of the surrounding area.

8. Obtain copies of the original station observers' reports(E-15's) for the month in question. These are generally archived by the State Climatologist.

9. In some cases, it may be of help to plot storm totals in a region over several days.

10. After the easy problems are solved you may create a working table of problem data points, with columns for comparing estimates made by different methods. You may wish to note such clues as day-to-day changes in depth of snow on the ground, or liquid equivalents, if given. If available, station information such as NWS Form B44 may provide information on the type of precipitation-measuring equipment in use.

11. From the above techniques and others, your flagged estimates are inserted into a copy of the original dataset. Reprint and plot the corrected dataset.

12. You should now have a much reduced list of problem data points. A few iterations of the process may be required, adding further ingenious and creative techniques as needed.

13. Some data point problems may not be solvable. In early or late season months, these single points may not be significant in seasonal totals. Highly tenuous estimates may have to be made for a few station-months, with adequate warning to the data user.

CONCLUSIONS

Official climate records are not immutable, inviolate, eternal truth. Working with them requires the application of a healthy, though not cynical, skepticism and often some analysis and cleaning. In working with weather data, and especially with snowfall records, it is essential that you proceed with caution, awareness, and a large amount of common sense.

ACKNOWLEDGEMENTS

Special appreciation is expressed to Dr. Fred V. Nurnberger, State Climatologist, Michigan Department of Agriculture, for guidance and assistance with the data cleaning process and for helpful reviews of this paper. Peter Kurtz and Cathy Sernick provided valuable assistance in the printing format. The suggestions of anonymous reviewers were of significant help in the improvement of this paper.

REFERENCES

Baldwin, Stanley, 1994. Personal conversation.

Goodison, Barry E. 1978. Accuracy of Canadian Snow Gage Measurement. J. of Applied Meteorology, V. 17, October 1978 pp. 1542-1548

Neff, Earl L., 1977. How Much Rain Does a Rain Gage Gage? J. of Hydrology, Vol. 35. pp. 213-220

Peck, Eugene L. and Merle J. Brown, 1962. An Approach to the Development of Isohyetal Maps for Mountainous Areas. J. of Geophysical Research Vol.67 No. 2, February 1962 pp. 681-694

Sykes, R. B., Jr. 1993. Seasonal Snowfall Totals to 1992-1993 from 1884-1885 for Oswego, New York. 50th Eastern Snow Conference/ 61st Western Snow Conference pp. 315-324