For NUMUG99, Syracuse, New York, 1999

Ethics and Common Sense in Measurement

Thomas J. Lockhart, CCM, CMet
Meteorological Standards Institute
Box 26, Fox Island, WA 98333

If there is one absolute in the ethics of measurement, it is that the data (including METADATA) must speak for themselves. But what can numbers say? If one applies standard statistical processes to a series of numbers the answer is reproducible. Of course, there is some control in selecting a sub-set of a series of numbers (see "All That Is Labeled Data Is Not Gold" below). Even if a time series is continuous, the beginning point and the ending point can influence the outcome. There is nothing unethical in presenting a continuous sub-set of data and listing their statistical parameters, if the fact that it was a selected sub-set is disclosed. When the parameters are used to make a regulatory (or societal) point the ice is thinner.

The classic example of mixing science with "political" science comes from comments made by Stephen Schneider ["Our Fragile Earth," *Discover*, October 1987, p. 47], a proponent of the theory that CFCs are depleting the ozone layer. He said "[W]e have to offer up scary scenarios, make simplified, dramatic statements, and make little mention of any doubts we may have. Each of us has to decide what the right balance is between being effective and being honest." Those of us who subscribe to the ethics of measurement will have no problem with that decision. Science, including its sub-set measurement, requires honesty. There is no type-A or type-B honesty. If any consideration influences how numbers are gathered or generalized, that consideration must be defendable objectively.

The measurement community often expresses a genuine skepticism when considering the results of model simulation. The assumptions used for the model simplifications and for the "data" inputs deserve scrutiny. When assumptions go beyond the data in time, it is a forecast. There is real value in general circulation models and there has been real improvement in some aspects of weather forecasting as a result of model outputs. It will be some time, if ever, before models will contain the intelligence exhibited by local meteorologists (or local farmers) in local short-term forecasts. When assumptions go beyond the data in range, it is a guess. If anemometers are calibrated to 50 m/s and report speeds of 70 m/s it is impossible to know the uncertainty of the measurement. Extrapolation beyond experience is often the only course available but a footnote is required to warn the user of that fact. Extrapolation between measurement points may not agree with new measurements located to test the extrapolation. The measurement deserves the presumption of accuracy. There are instrument tests which can

confirm the instrument performance. The smoothed model value should not be presumed to be correct.

It is true that caveats ruin the literary quality of pronouncements. Perhaps there should be a sharp division between the statements of "scientists" on political issues and scientists reporting the results of their research. The public is not capable of sorting the "results" from a reputable scientist which are biased for advocacy from the results from a reputable scientist with all the dull but critical caveats which qualify the data supporting the results. It used to be that ethics took care of the problem. If ethics no longer apply, a method needs to be found to warn the public of the advocacy roll of "political" scientists.

All That Is Labeled Data Is Not Gold

Most of us are quite comfortable with the idea of evolution in the animal kingdom or in climate, by which small departures from the norm can spread and eventually change the entire picture. We are not as ready to accept a similar evolution in classical data sets, the gold standard of climatology.

Recently, I have been using a small subset of the data generated by the GEOSECS (Geochemical Ocean Sections) program (1973-1974), in particular, the ocean profiles of oxygen-18 and deuterium measured by Harmon Craig at Scripps Institution of Oceanography in La Jolla, Calif. Until recently, I was secure in the knowledge that the data I was using, gathered from the National Climate Data Center repository, were complete and unadulterated. Earlier this year, while presenting my results and using these data for comparison, a member of the audience wondered why I was not using the complete data set. Confused, I defended myself vigorously. Later, it appeared that although I had correctly presented the published data, there was an underground version of unpublished data that was known only to a small circle of initiates. My hosts were kind enough to include me among the cognoscenti and it was at this point that things started to become interesting.

Examining the new version, I started to find many inconsistencies in the two data sets: values different in one than the other, stations appearing in one or the other but not both, different measurement depths, and even different positions for the stations. Puzzled, I went further afield in search of information that could resolve the conflicts. Looking at the published tables, I worryingly found that the mistakes in station positions seemed to be due to typographical errors.

Soon, though, I found someone else who had a subset of the unpublished data. This would clear things up. Unfortunately, it made things even worse. Where this data set purported to give the same data, it was different from both previous versions in seemingly random ways.

In despair, I tried to track down the source of the unpublished data. Finally, I found someone who had an old photocopy of an old-style computer printout whose provenance was claimed to be Harmon Craig's original measurements (including repeats). At last, truth!

It became clear very quickly that this printout was the original version of the unpublished data, and it was equally clear that the data had at some point been typed in by hand. The errors were of the sort caused by inadvertent line skipping, missing decimal points and minus signs, and incorrect averaging of repeat measurements. However, it was also evident that this was the source for the published tables. The published data though, had been subjected to some quality control; outliers had been thrown out, repeat samples with to much dispersion ignored, and possibly, further repeats had been performed. The measurement depths had clearly been refined using more accurate information.

The origin of the third data set remained mysterious until I was informed that it had been traced from a graph showing the data in an old article and then correlated to the original stations. That explained the small but random differences. Only with all this scientific detective work and

the discovery of the ancestral data was I now able to amalgamate the published and unpublished results into a consistent data set.

These data are only 25 or so years old and yet there are at least three (maybe more) different and mutually inconsistent versions floating around. The lesson we should take from this is that, if unchecked, small mutations will occur during transcriptions. Unless we are vigilant, "data sets" will evolve and data that have been so painstakingly collected and saved will be distorted and twisted beyond all usefulness. Our care in using data should be a force analogous to natural selection, keeping data sets fit for the purposes intended. If we do not take care, the gold standards of climatology may in time turn to lead.

Gavin Schmidt, NASA Goddard Institute for Space Studies, N.Y., USA; E-mail: gschmidt@giss.nasa.gov.

Schmidt, G., Eos, AGU Transactions, Vol. 79, No. 28, page 336, 1997.

During the process of writing a revision to the EPA Quality Assurance Handbook for Air Pollution Measurement Systems - Volume IV - Meteorological Measurements (EPA600/4-90-003) several subjects arose which could be resolved only by appeal to common sense. In the more than 30 workshops which have followed the publication of Volume IV one appeal that is always made is to use common sense. When all else fails (or even if it doesn't) common sense is a good standard.

For example, one always finds the requirement that calibrations be "traceable" to NBS (now NIST). What does this mean and how does one document compliance with the requirement? The common interpretation is that there needs to be some anemometer calibration at NIST which starts the transfer standard path. It might go to another wind tunnel where the NIST calibration is transferred to the new wind tunnel. A calibration in the new wind tunnel might be further transferred to a third wind tunnel. A series of documents could be in the file which records all these calibrations. The operator of the third wind tunnel might provide calibrations traceable to NIST.

A paper trail should not be enough when there is a possibility of error in the process. Two cases come to mind. The calibration at NIST has some uncertainty. Occasionally there is an outlier. The prudent summarization of a NIST calibration is a linear regression analysis, looking critically at differences at each point. Most mechanical anemometers are linear once the non-linear starting speeds are passed. Outliers or problems with the calibration can usually be seen with this analysis. One wind tunnel operator transferred each point of the NIST calibration to a new anemometer, even the one obvious outlier in the NIST data. This was defendable on paper but failed the common sense test.

Technology is always improving. Cup anemometers block some of the flow in a wind tunnel. How much is not well known. The amount varies with the design of the anemometer but also with the size of the wind tunnel test section. Common sense says that a cup anemometer calibrated in the large (3.25 square

meter) test section at NIST will have a small blockage error, probably 1% or less. When the calibrated anemometer is used in a smaller wind tunnel with a test section of 0.4 square meters, there will be more blockage. If the calibration method involves transferring the NIST wind speeds to the wind tunnel fan RPM and then transferring the wind tunnel fan RPM to another identical anemometer, the blockage cancels out. Common sense says you can ignore the blockage effect. If, on the other hand, the wind tunnel fan RPM is used to "calibrate" a different kind of anemometer, the relative blockage of each anemometer in the small test section must be known. Since this effect is difficult to quantify, it is common sense to use NIST transfer standards for each type of anemometer of interest.

When, in the past, wet-bulb and dry-bulb temperatures were measured with a sling psychrometer it was a new technology. Other methods for measuring dew-point temperature and relative humidity allowed for difference comparisons to be made. Then it became clear that siting bias was a big problem for the sling psychrometers. Even if the thermometer is moving rapidly, solar radiation will heat the thermometer. If shade is found there is still the problem with reflected radiation. Body heat and humidity can bias the reading by modifying the air if the thermometer is down wind of the operator. Understanding the measurement process and the application of common sense will minimize these errors.

A very accurate air pressure transducer can be calibrated in the laboratory but when it is installed in the atmosphere the wind effect must be considered. Static ports are now available to minimize wind pressure effects but two or more decades ago the exposure of the pressure transducer was not considered. The transducer would be mounted in a weather proof box but the inside box pressure was assumed to be the same as atmospheric pressure. When the wind was blowing there was a bias, conditional on the speed and direction of the wind, which could be several times the calibration uncertainty. The common sense in this example must be uncommon until the knowledge of such effects reaches the operational meteorologist who must make decisions about the design of the instrument systems.

There is an "official" data archive for the United States at the National Climatic Data Center (NCDC) in Asheville, North Carolina. If one needs a copy of the "official" data for some purpose, often related to some law suit, one can get it. It comes with a gold seal and trailing ribbons. Judges and juries are impressed by the trappings of truth. It is true that the copy is certified to be correct by the head of NCDC, but this does not say anything about the accuracy of the data. The National Weather Service is responsible for the accuracy of the measurements it makes, NCDC is only responsible for faithfully recording the numbers and copying them when required.

EPA and NRC list the performance specifications required for measurement systems used on projects under their authority. Performance audits are usually

required on a periodic schedule to verify that the systems meet those specifications. The auditor may use auditing methods designed to document conformance. For wind speed there are two tests. One is starting threshold measuring the starting torque of the shaft and bearings with the cup wheel or propeller removed. Bearings will degrade over time. The time is a function of the exposure environment. There is a starting torque which has been shown to be equivalent of 0.5 m/s. The audit will document the starting torque expressed as a speed. If the result is 0.6 m/s and the requirement is 0.5 m/s, are all of the speed data rejected? If the last audit was six months ago, perhaps on the last 30 days of data will be rejected since the bearings will degrade with time. The application of common sense by the auditor, operator, and regulator will result in keeping all the data. The bearings would be changed but the performance of the anemometer was probably acceptable. The wind in the atmosphere does not start at steady slow speeds as the wind tunnel test requires. The simulation of starting speed by starting torque has some uncertainty. Average winds under 1 m/s are usually increased to 1 m/s for EPA models.

If the audit showed a starting speed of 2 m/s, the answer becomes more difficult. The operator should examine the speed data for the six months or year since the last audit where the starting speed was shown to be 0.4 m/s. Perhaps the record will show a period where something happened to the anemometer. If the site is windy there may not be many periods with winds less than 2 m/s, in which case a higher starting speed would not degrade the data. Common sense and critical examination will suggest an answer which the regulator and operator will agree to. The last answer to accept is the auditor rejecting all the speed data because of the starting torque test results.

The other auditing method for anemometers is imposing a series of known rates of rotation to the anemometer shaft. This challenges the ability of the measurement system to sense the rate of rotation of the shaft and express this rate in terms of wind speed at the output of the system. When the sensor output is a frequency and the data logger is digital, the test will always pass. The only thing being challenged is the ability of the system to count pulses and apply an algorithm to express frequency as speed. There is nothing in this test that confirms the algorithm of wind speed to frequency. This takes a wind tunnel test or the acceptance of the manufacturer's claim that the generic transfer function for the product is correct.

When dealing with regulators, operators, and consultants, a common sense discussion will usually result in an acceptable solution. What is even more important, it will result in the exchange of information which leaves all parties more experienced and better prepared for the next question.