Agricultural and Forest Meteorology 237–238 (2017) 362–370

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Inferring random component distributions from environmental measurements for quality assurance

E. John Sadler[a,*], Kenneth A. Sudduth[a], Scott T. Drummond[a], Allen L. Thompson[b], Jiaxun Chen[c], Patrick R. Nash[a]

[a] USDA-ARS Cropping Systems & Water Quality Research Unit, Rm 269 Ag. Eng. Bldg, University of Missouri, Columbia, MO 65211, USA
[b] Bioengineering Department, Rm 251 Ag. Eng. Bldg, University of Missouri, Columbia, MO 65211, USA
[c] Statistics Department, Rm 9 Middlebush Hall, University of Missouri, Columbia, MO 65211, USA

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 14 February 2017
Accepted 18 February 2017
Available online 6 March 2017

Keywords:
Quality assurance
Quality control
Environmental measurement
Air temperature
Bias
Random variation

ABSTRACT

Environmental measurement programs can add value by providing not just accurate data, but also a measure of that accuracy. While quality assurance (QA) has been recognized as necessary since almost the beginning of automated weather measurement, it has received less attention than the data proper. Most QA systems examine data limits and rate of change for gross errors and examine data for unchanging values. Others compare data from other locations using spatial tools or examine temporal consistency. There exists a need for analytical tools that can increase the likelihood of detecting small errors, such as a calibration drift, or increased variation in a sensor reading. Two such empirical tools are described herein that can inform a first level QA process. One operates on data from a single sensor, using comparisons between a current and a prior datum; the other leverages additional information from a duplicate sensor and operates on only the current datum. The objectives of this paper are to describe the computational methods, illustrate results with multiple-month datasets representing both nominal and failing sensors, provide some indications of validity of assumptions made in the derivation, and suggest where in a quality assurance program these methods could be applied. With little additional datalogger programming to obtain both the period average and the ending value, these tools could be added to QA toolkits in many automated weather stations.

1. Introduction

Information obtained in an ongoing environmental measurement program is inherently more valuable if it is accompanied by indicators of data quality. Generally speaking, the indicators of the quality constitute part of the metadata about the measurement, although they have often not been recognized as such. As datasets become increasingly available online and to unknown users, there is an increasing need to categorize data quality as, for example, good, suspect, missing, or gap-filled (see proposed guidelines, WMO, 2004). However, even the judgement of 'good' data is governed by the intended use. Such a judgement may not be appropriate for a different use. Accommodating other uses would appear to be better served if the quality of the data were quantified. Attempts to approach quantification have usually been limited to citing manufacturer's specifications of instrument accuracy. However, accuracy specifications may be variable across the range of the parameter measured, may be dependent on other parameters, may be affected adversely by calibration drift or sensor failure, or may be heavily affected by the exposure of the instrument to the conditions being measured.

Thus, it is increasingly desirable for information on accuracy to be tied to each measured value. Current trends in database development and publication suggest that sensor readings will be expected to be accompanied by real-time estimates of the confidence in the measurement, as a movement toward self-describing data (Gray, 2009), perhaps including record-level estimates of sensor uncer-

Abbreviations: *A*, actual or true value corresponding to *O* (overbar represents the mean); *B*, bias corresponding to *O* (overbar represents the mean); CMRB, Central Mississippi River Basin; *i*, *j*, index for current and prior time; LTAR, Long-Term Agro-ecosystem Research network; *O*, observation at point in time (overbar represents the mean); *p*, *s*, denote primary and secondary sensor values for *O*, *A*, *B*, and *R*; *R*, random component corresponding to *O* (overbar represents the mean); QA, quality assurance; QC, quality control; *T*, temperature of air; *Tmax*, maximum air temperature during period; *Tmin*, minimum air temperature during period.

* Corresponding author.
E-mail addresses: John.Sadler@ars.usda.gov, sadlerj@missouri.edu (E.J. Sadler), Ken.Sudduth@ars.usda.gov (K.A. Sudduth), Scott.Drummond@ars.usda.gov (S.T. Drummond), ThompsonA@missouri.edu (A.L. Thompson), jczz7@mail.missouri.edu (J. Chen), Pat.Nash@ars.usda.gov (P.R. Nash).

http://dx.doi.org/10.1016/j.agrformet.2017.02.021
0168-1923/Published by Elsevier B.V.

tainty (Sadler, 1983). Such metadata would be an incremental evolution of historical quality assurance and control (both generally termed QA) of meteorological data collected at automated weather stations. Gandin (1988) noted that automated QA followed the first automated weather stations in the middle 1950s, and that complex QA techniques have existed since the late 1960s. However, Gandin (1988) also noted that adoption was slow because QA was considered a purely technical activity and of secondary importance by both meteorologists and administrators. Gandin (1988) also provides a useful structure for the current paper, by categorizing errors as random, systematic, and rough (gross) errors. The latter include such events as abject sensor failure, as by broken cables or shorts. Systematic errors could be caused by calibration errors, for example. Random errors could perhaps be better termed as other variation of unknown causes, including environmental variation. Gandin's review also provides a classification of QA tests, listing plausibility checks (either range limits or distributional outliers), checks for contradictions among sensors (e.g., heavy rain with no clouds), spatial consistency checks (separate similar sensors), and a comparison to models. Later literature also includes rate-of-change limits and persistence checks (Meek and Hatfield, 1994; National Oceanic and Atmospheric Administration, 1998; Shafer et al., 2000; WMO, 2004; Hubbard et al., 2005; Hubbard et al., 2012; NOAA-NWS, 2015). In distributed weather networks, spatial interpolation has been widely deployed (Shafer et al., 2000; Hubbard et al., 2005; Williams et al., 2011; Hubbard et al., 2012). In retrospective QA assessments, temporal persistence (i.e., stuck sensor) or repetition (i.e., repeated months or years) can be evaluated (Durre et al., 2010).

In special cases, individual weather stations may deploy duplicate or triplicate sensors as described below. Some modern weighing rain gauges have triplicate vibrating wire sensors for complex comparisons within each reported measurement. These special cases often provide other tests important in a local setting. In general, however, range limits, rate-of-change limits, and static/persistence checks are amenable to use in a single weather station (Meek and Hatfield, 1994).

However, these tests are much more suited to detecting the rough or gross errors in Gandin's (1988) classification, which are extreme outliers, and are not as effective detecting statistically valid but still erroneous data (NOAA-NWS, 2015). For example, bounds on ingest systems usually set one pair each of range and rate-of-change limits for air temperature, and those limits are quite broad ($\sim\pm50\,^{\circ}$C for air temperature, and $\pm20\,^{\circ}$C/h for rate-of-change limits). Seasonal limits can improve detection abilities (e.g., Hubbard et al., 2005). However, using historical distributions or extremes is unlikely to detect small changes in calibration. There exists scope to identify or develop test statistics that could improve detection of subtle sensor errors. Such statistics would be most valuable were they to approach a known central tendency, vary over a well-defined distribution during nominal sensor performance, and vary over a much broader range during times of errors in the sensor measurement.

Range limits based on deviation from mean values enable a probabilistic QA statement based on those range limits. This approach would be an extension from fixed limits as by Hubbard et al. (2005), who used limits of 3 standard deviations from the mean to flag suspect outliers. If available, instantaneous expected variation values to compare with historical distributions could inform confidence in measurements, by simply stating the current value's position on the historical empirical distribution. Historical distributions of the random term could place constraints on expected values and therefore inform detection of a degraded sensor. Either of these criteria requires appropriate thresholds to discriminate between probable correct or erroneous data, with appropriate balance between Type I and Type II errors (Durre et al., 2008). Finally, if the manufacturer's accuracy or equivalent information from a lab test is known, empirically derived variation values could inform judgments on additional sources of variation, such as environmental exposure.

We describe two methods to infer near-current distributions of random variation while a sensor is deployed, to be used in a near-real-time QA assessment conducted at the time of downloading data from a datalogger. One method is for a single-sensor configuration; the other is to infer the difference of variation for two sensors in a duplicate sensor configuration. In both cases, data collected both as instantaneous samples at the end of the interval and as an average over the interval are leveraged to infer the random term.

The objectives of this paper are to describe the computational methods, illustrate results with multiple-month datasets representing both nominal and failing sensors, provide some indications of validity of assumptions made in the derivation, and suggest where in a quality assurance program these methods could be applied.

2. Materials and methods

2.1. Example data and context

We illustrate the procedure below using air temperature collected at a recently installed weather station in the Central Mississippi River Basin (CMRB) node of the Long-Term Agro-ecosystem Research (LTAR) network. This station was installed in August of 2015 for several purposes, one of which was to provide near-real-time weather data to the LTAR network's online portal (https://ltar.nal.usda.gov/ltar/met/query?station=5&_action_query=Display+meteorology+data). This weather station was designed with redundant sensors for most common measurements. One set was to the standards chosen by the LTAR working group on meteorology (under development), and the other was to more nearly match the legacy weather station in operation since 1992 (Sadler et al., 2015) to provide longitudinal overlap with which to test non-homogeneities in the record as a result of the changes.

In the CMRB station, the air temperature sensors are the same (Campbell Scientific Inc., Logan Utah, model HC2S3 Temperature and Relative Humidity Probe), but the shields are different. The LTAR standard is aspirated (Apogee Instruments, Logan Utah, model TS-100), where the legacy weather station used a passive radiation shield (Campbell Scientific Inc., Logan Utah, model 41003-5 10-Plate Solar Radiation Shield). The air temperature sensor has a stated accuracy of $\pm0.1\,^{\circ}$C at $23\,^{\circ}$C, diverging slightly at higher and lower temperatures, but always within $\pm0.25\,^{\circ}$C in the expected environmental range of $-20\,^{\circ}$C to $40\,^{\circ}$C. The aspirated shield has been demonstrated by the manufacturer to perform very nearly the same as industry-standard high-powered aspirated shields, with differences in the range of $\pm0.02\,^{\circ}$C (their stated resolution) (http://www.apogeeinstruments.com/comparison-of-three-fan-aspirated-solar-radiation-shields/). No information is available from the manufacturer on radiation-induced errors in the passive shield, but the manufacturer acknowledges that ventilation in shields improves measurement accuracy.

The sensors were read by an automated datalogger (Campbell Scientific Inc., Logan Utah, model CR3000) on a 3-s interval, with data reported as a sample, average, maximum, and minimum at 5-min intervals. These values enabled statistical aggregation to longer periods, including the LTAR network's 15-min interval, with exact averages and extrema. Software on a central server downloaded the data from the datalogger on hourly intervals, and the 5-min data records were combined to monthly files and stored.

The system operated from installation 3 September 2015 onward, except that the secondary sensor was put in place a