J1.8 ASSIMILATING TIPPING BUCKET RAIN GAUGE DATA INTO THE PACRAIN DATABASE

Michael D. Klatt*, Mark L. Morrissey, and J. Scott Greene
University of Oklahoma, Norman, Oklahoma

1. INTRODUCTION

The Pacific Rainfall Project (PACRAIN) is part of the Environmental Verification and Analysis Center at the University of Oklahoma. One of the core components of PACRAIN is a comprehensive rainfall database (Greene et al. 2008). The database currently spans a period of 1874 to the present and contains more than 2 million observations from more than 800 sites. The data are compiled from various national agencies as well as sources unique to PACRAIN. Data are publicly available online at <http://pacrain.evac.ou.edu>.

The goal of the PACRAIN database is to be a starting point for any research requiring tropical Pacific rainfall data. To that end, it presents data in a format that is intended to be flexible enough for a wide range of applications while being robust enough to preserve the information content of a variety of rainfall measurement techniques. Data interpretation is minimal because users are in the best position to choose the most appropriate techniques for their needs.

PACRAIN is a partner in the Pacific Islands Global Climate Observing System (PI-GCOS) initiative to expand and enhance climate observation networks in the region. To date, a total of 50 tipping bucket rain gauges have been sent to Cook Islands, Guam, Kiribati, Niue, Samoa, Tonga, Tuvalu, and Vanuatu (see Fig. 1). Data collected from these gauges is being sent to PACRAIN (see Table 1) for inclusion in the PACRAIN database. Until now, all PACAIN data have been daily or monthly accumulations from manual-read gauges. Tipping bucket gauge (TBG) data represent a paradigm shift from traditional rainfall data, and thus it is necessary to examine how these data can be best integrated into the PACRAIN database.

Site	Begins	Ends	Tips
Nikao	2007-03-13	2007-06-11	1825
Hanan	2006-07-05	2006-08-07	814
Makefu	2005-08-19	2005-11-22	5216
Afiamalu	2006-01-20	2006-02-10	8150
Fua'amotu	2005-05-04	2007-08-08	16262
Funafuti	2007-03-29	2007-12-10	3779
Port Vila	2005-02-12	2005-12-31	6425

TABLE 1. TBG data received to date.

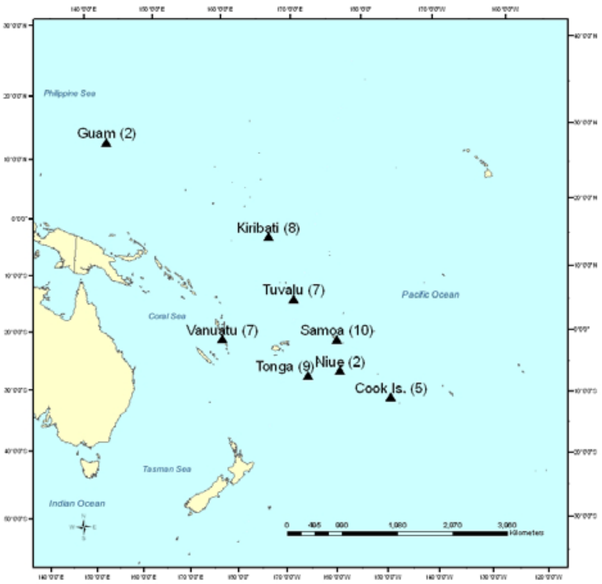

FIG. 1. The distribution of PI-GCOS tipping bucket gauges to date. The number in parentheses is the number of gauges sent to a particular location.

2. TIPPING BUCKET GAUGES

A tipping bucket gauge works by recording the amount of time it takes to accumulate a fixed amount of rainfall, e.g. 0.01 inches or 0.254 mm in the case of PI-GCOS gauges. A typical gauge contains a mechanism consisting of two small bins on a pivot (the "tipping buckets"). The mechanism is positioned underneath a collection funnel so that only one bin at a time can receive rainfall. The mechanism is calibrated so that weight of a known amount of rainfall will cause it to pivot (a "tip"), simultaneously placing the empty bin in position to collect rainfall, emptying the full bin, and actuating

*Corresponding author address: Michael Klatt, EVAC, University of Oklahoma, 100 E. Boyd, SEC 410, Norman, OK, USA 73019; e-mail: mdklatt@ou.edu a switch. A data logger is connected to the switch to record the tips in some fashion; the loggers used by PI-GCOS record the time of each tip.

There are several design limitations of tipping bucket gauges that can lead to errors in observed rainfall amounts. The finite response time of both the mechanism and the logger can cause underestimation during high-intensity events, while underestimation at low intensities is possible due to evaporation from a partially full bin. Underestimation can also occur if the gauge's collection funnel or mechanism is obstructed by ice or debris, or if the data logger reaches its capacity before the gauge is serviced. Spurious tips will cause overestimation of rainfall amounts. At very high rainfall rates it is possible for the mechanism to bounce off of its mechanical stop during a tip, causing a tip in the other direction—a phenomenon known as "double tipping." It is also possible for the gauge to record a spurious tip if it is jostled.

The nature of TBG data makes it impossible to determine the exact temporal distribution of the recorded rainfall. For high rainfall rates, when the interval between tips is small, TBG data closely approximates the instantaneous rainfall rate. As the interval between tips increases, however, TBG data becomes like traditional coarse accumulation data; it is known how much rainfall there was during a specific time interval, but not exactly when within that interval the rainfall occurred. Furthermore, with TBG data it is impossible to determine periods of no rain. For tip size V and tip times t_i and t_{i+1}, the observed rainfall rate is $V\cdot(t_{i+1} - t_i)^{-1}$; this is asymptotic to zero because V cannot be zero. TBG data are most useful at high rainfall rates where temporal ambiguities are minimized.

3. RAINFALL DATA MODELS

Unlike other meteorological parameters, rainfall is typically expressed as an aggregate value over a finite time interval rather than an instantaneous value. This interval is usually not explicit in the data, which can lead to ambiguity (Klatt et al. 2006). For example, daily rainfall records from the National Climatic Data Center (NCDC) give the observation date, which is the end of the interval (NCDC 2006). However, in data from New Zealand's National Institute for Water & Atmospheric Research (NIWA), the given date is the beginning of the interval (NIWA, personal communication). Even rainfall intensity, which itself is an instantaneous value, is often expressed in the context of an interval. For example, the Global Precipitation Climatology Project monthly data set provides rainfall rates, but these are monthly rainfall rates and are thus aggregate values (Huffman and Bolvin 2002). Although tipping bucket gauges are sensitive to rainfall rate, they are actually accumulation gauges where sampling is fixed with respect to amount (the tip size) rather than time; therefore, TBG data are also aggregate values.

The current PACRAIN data model is based on traditional fixed interval observations. Each record contains an accumulation and the start time and duration of the interval for that accumulation. By making the duration explicit, this model is capable of handling data of any resolution. It also allows for the duration to vary from record to record for a given site; this is particularly useful for storing data from the Schools of the Pacific Climate Experiment (Postawko et al. 1994), where observations are not always made at the same time every day. Although this model was not conceived with TBG data in mind, it is capable of storing such data since they are in fact accumulations of variable duration.

An important factor to consider, though, is database capacity. With 50 gauges distributed to date, an average rainfall of ~2500 mm·year^{-1}, and a tip size of 0.254 mm, the PI-GCOS network has the potential to generate almost 500,000 tips per year. With one record per tip, this would increase the amount of data being added to the database by a factor of six. This would greatly increase the data storage requirements and have an adverse effect on database performance. There may be an alternate data model appropriate for TBG data that would minimize their footprint while still being compatible with the existing fixed-interval data.

One way to compress the TBG data would be to aggregate them using a fixed time interval. Attempting to preserve the character of high-intensity events by choosing a small interval would result in data redundancy for low-intensity events, and might increase the total number of records. Choosing an interval is thus a tradeoff between losing information content or increasing data size. Even if a suitable interval could be chosen, aggregating TBG data using any fixed interval is problematic. Given tip size V, tip time t, and fixed interval boundary T such that $T < t$, how should V be apportioned between the two intervals bounded by T? The fraction of V that occurred before or after T cannot be determined, so any method to apportion it will be arbitrary and will introduce random errors into the data. It is PACRAIN policy to present data "as-is" as much as possible, so anything that will decrease information content while introducing random errors can be ruled out *a priori*.

Sansom (1992) describes a rainfall data model which he calls “breakpoint representation”. This was motivated by efforts to digitize pluviographs from automatic siphon gauges, and has the advantages of preserving information content while reducing redundancy. As Sansom observed, rainfall events are a series of transitions from one steady rainfall rate to another. Thus, a rainfall event can be fully characterized by a sequence of data pairs that denote a rainfall rate and its duration. For TBG data, then, it is only necessary to store changes in rainfall rate. Given tip times t_i, t_{i+1}, and t_{i+2} such that $t_{i+1} - t_i = t_{i+2} - t_{i+1}$, it is not necessary to store both interval (t_i, t_{i+1}) and (t_{i+1}, t_{i+2}) because the rainfall rate is the same. This results in a 2:1 compression ratio without losing any information about the rainfall event. This is a powerful concept that can be applied to traditional fixed-interval data as well as TBG data. If the entire PACRAIN database used a breakpoint representation it might result in a significant reduction in storage requirements.

Fixed-interval data have many applications, so fixed-interval accumulation data presented as breakpoint data should ideally be convertible back to their original representation without loss of information. While a breakpoint representation preserves the information of data with respect to rainfall intensity, in eliminating redundancies it eliminates some of the fixed-interval information of the data. Consider consecutive daily records, all having the same amount A. These records could be collapsed into a single breakpoint with an intensity of $A \cdot \text{day}^{-1}$ and a duration of T days. When trying to convert the breakpoint data to daily rainfall, there is no way to determine which days contain which fraction of the amount $T \cdot A$. Even if it was known that the breakpoint was derived from daily data, it could not be assumed that the breakpoint was the result of T equal daily values. A breakpoint might describe T equal daily amounts, but it might also describe, for example, a T-day accumulation due to missed observations. Therefore, if breakpoint representation is used there needs to be additional information stored to preserve the character of the original data.

Run length encoding (RLE) is a lossless data compression scheme (Salomon 2006) that can be combined with breakpoint representation to preserve the original non-breakpoint information. The basic concept of RLE is that repetitive data values can be replaced with a single value and a count of how many times that value is repeated. The addition of a repetition count would allow breakpoint data to be transformed back into its original format, fixed-interval or otherwise. A breakpoint of duration T and repetition count n can be resolved into fixed-interval accumulations of duration $T \cdot n^{-1}$. Continuing with the example from above, if $T = n$ it is known that the breakpoint represents n daily observations of amount A.

RLE can reduce the size of a data set by eliminating redundant records, but its implementation requires additional data to be stored. If the number of duplicate records is small, RLE may actually increase the overall data storage requirement. In order to implement RLE with the PACRAIN database, a 4 byte integer (the repletion count) would need to be added to every 68 byte record. For RLE to be cost-effective, it must reduce the total number of records by at least 6%.

A preliminary analysis shows that RLE would reduce the total size of TBG data received to date by less than 2%. However, if RLE were applied to the rest of the PACRAIN data it would reduce the number of records by 32%. The discrepancy between the benefit of RLE for TBG data and for fixed-interval data is not surprising. For a TBG record to be a duplicate it has to have the same duration as a preceding record; given a precision of 0.5 s, the odds of this are small. For fixed interval data, on the other hand, it is common to have periods of no rain that can be collapsed into a single record. The size reduction for fixed-interval data makes RLE cost-effective given the current ratio of TBG data to fixed interval data. However, as the proportion of TBG data increases over time RLE will become less cost-effective.

4. OTHER ISSUES

Aside from choosing the best data model, there are some miscellaneous issues that need to be resolved before TBG data can be added to the PACRAIN database. The first is the temporal resolution of the TBG data. The database can currently handle time stamps down to a precision of 1 second, but the PI-GCOS data loggers have a precision of 0.5 seconds. The database management system (DBMS), PostgreSQL, can handle time data of that precision, but modifications need to be made to the supporting applications that handle data ingest and retrieval. A minor modification of the data format used for online access is also necessary.

The data logger records the start time and stop time of a service cycle as well as the time of each tip event. When the logger is serviced using a PC or a portable device its internal clock is set to the system time of that PC or device. Thus, it is possible for the logger times to be inconsistent from one cycle to the next. Inconsistencies will not always be obvious, but there is one instance so far where the start time for one cycle is earlier than the stop time for the previous cycle by a few minutes. Problems like this can be noted in the database by a footnote attached to the affected records.

The end of a service cycle needs to be handled differently than a normal tip event. It cannot be assumed that no rain fell during the interval from the last tip time to the logger stop time; it is possible that there was rainfall, just not enough to cause an additional tip. Thus, this period has an unknown rainfall amount in the range [0, 0.254) mm. These incomplete tips will need to be denoted with a special value in the database.

5. CONCLUSIONS

Work continues to determine the optimum data model for integrating TBG data with existing PACRAIN data. The current version of the database is capable of storing data from the PI-GCOS tipping bucket gauge network, although the potentially large volume of new data might be overwhelming. Sansom's breakpoint model, which is very similar to the existing PACRAIN model, is promising, but it needs to be combined with a repetition count. A breakpoint model with a repetition count would greatly reduce the required number of traditional fixed-interval records in the database. However, if the TBG data received to date are representative, the repetition count will actually increase the storage requirements for TBG data because the number of redundant records is very small. Once the data model is finalized, several issues specific to the PI-GCOS data can be addressed, and the data can finally be added to the database.

6. REFERENCES

Greene, J. S., M. Klatt, M. Morrissey, and S. Postawko, 2008: The Comprehensive Pacific Rainfall Database. *J. Atmos. Oceanic Technol.*, **25**, 71-82.

Huffman, G. J. and D. T. Bolvin, 2002: GPCP Version 2 combined precipitation data set documentation. NASA Goddard Space Flight Center, 36 pp.

Klatt, M., M. L. Morrissey, and J. S. Greene, 2006: Temporal comparison of the Comprehensive Pacific Rainfall Database (PACRAIN) with satellite rainfall estimates. Preprints, *22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*, Atlanta, GA, Amer. Meteor. Soc., 4.4.

National Climatic Data Center, 2006: Data documentation for data set 3200 (DSI-3200). National Climatic Data Center, 18 pp.

Postawko, S., M. L. Morrissey, and B. Gibson, 1994: The Schools of the Pacific Rainfall Climate Experiment: Combining research and education, *Bull. Amer. Meteorol. Soc.*, **75**, 1260-1266.

Salomon, David, 2006: *Data Compression: The Complete Reference*. Springer, 1092 pp.

Sansom, J., 1992: Breakpoint representation of rainfall. *J. Appl. Meteor.*, **31**, 1514-1519.