8.2 REPROCESSING OF HISTORIC HYDROMETEOROLOGICAL AUTOMATED DATA SYSTEM (HADS) PRECIPITATION DATA

Dongsoo Kim * and Brian Nelson
NOAA National Climatic Data Center, Asheville, North Carolina

Lawrence Cedrone
NOAA NWS Office of Hydrologic Development

1. INTRODUCTION

The Hydrometeorological Automated Data System (HADS) is a real-time and near real-time data acquisition, processing and distribution system operated by the Office of Hydrologic Development (OHD) of the National Weather Service (NWS). The real-time data are used at the NWS River Forecast Centers (RFC) for hydrologic modeling (i.e., flood forecasts). These data are used directly in some models and they are used for bias correction of radar rainfall data which are then input to other models. The NWS multi-sensor precipitation estimation (MPE) algorithm uses the HADS data to produce the multi-sensor precipitation estimate.

A recent collaboration between the National Climatic Data Center (NCDC) and OHD investigates the utility of a reanalysis of MPE. A reanalysis data set has many possible uses that range from climate applications to water resources management. Our main motivation to reprocess historical HADS precipitation has grown from the need for a high density long-term in-situ rainfall data set as input to the multi-sensor precipitation reanalysis (Nelson et. al., 2006). At the same time, under the NOAA Integrated Surface Observing System (ISOS) working plan, NCDC is to examine the suitability of archiving HADS data. Hence we began to study climatic values of precipitation elements in the HADS archive which would complement the existing NOAA's Cooperative Observers network (COOP) precipitation data. Numerous cooperators and network owners who participate in HADS report diverse measurements of meteorological and hydrological variables. Complete information can be found in http://www.nws.noaa.gov/oh/hads/. The transfer of historical HADS archive from OHD to NCDC began during March 2005, and data transmission continues.

Initial investigations showed the value of the HADS data for several applications. However, quality control issues appeared to be evident in the data set. And since we intend to use the data set for development of a long term high resolution gridded data set we wanted to resolve as many of the quality issues as possible. Therefore, the reprocessing of Shared Hydrometeorological Exchange Format (SHEF) encoded HADS precipitation variables allows us to produce a more complete data set than what might otherwise have been possible with the real-time processing. The steps involved in reprocessing HADS precipitation archive are:

- Decoding of SHEF for two precipitation elements (PC: Precipitation Accumulator, PP: Incremental Precipitation),
- Conversion of sub-hourly PC to hourly PP,
- Revision of missing data values into "no rain" values when conditions are met,
- Checking for gross range errors,
- Generation of neighborhoods that are within the +/- 0.25 degrees from the target gauge station.

After the reprocessing, we launch quality assurance modules such as: 1) Quality flagging against NEXRAD Level III Digital Processing Array (DPA); and 2) Quality flagging against its neighboring data values (nearest

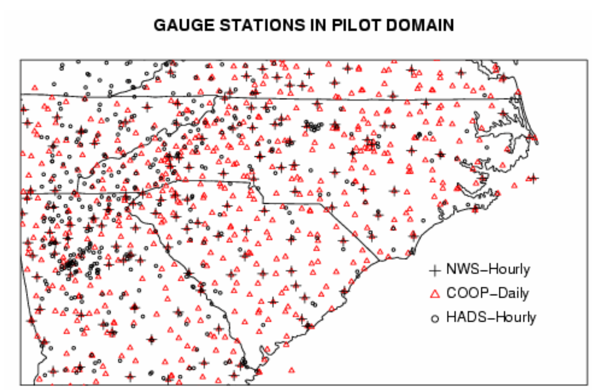

Figure 1. The pilot domain of historical HADS precipitation data during April, 2005. Three networks are displayed for the comparison of spatial density. Black circles are gauge stations in HADS.

* *Corresponding author address:* Dongsoo Kim, NOAA/NESDIS/NCDC, 151 Patton Ave, Asheville, NC 28801; e-mail: dongsoo.kim@noaa.gov.

neighbor checks). In the following sections, we describe detailed steps with examples in our pilot domain, the Carolinas (Figure 1). Quality modules presented here are still in development as quality threshold parameters will evolve as error characteristics are better understood.

2. PREPROCESSING

2.1 *Decoding*

The SHEF encoded HADS data have more than 20 meteorological and hydrological variables. We extract only two precipitation related variables, PC and PP.

2.2 *Assignment of Observation Time*

The observation frequency varies from station to station. Stations report at 5, 10, 15, 30, 60, 120, and 180 minute intervals. Most stations report at the hourly or 15-minute interval. Assignment of observation times involves two steps. First, we assign all sub-hourly and hourly reports to 15-minute bins according to its observation time. Any hourly PC report will fall into one of 4 bins according to observation time. Second, the hourly PP values are derived from PC values at top of hour. Any PC values whose observation time is not at the top hour are assigned to the nearest hour, and an associated quality flag file indicates minutes off top hour (discussed in section 3.1). The decrease of a PC value which results in a negative PP value is a strong indication of an error unless it happened during resetting of gauge.

2.3 *Revision of Missing Values*

Often times, missing values are reported intermittently and if the PC value before and after the missing report period is constant then it is reasonable to assume that the PC value should remain the same because the PC is a cumulative amount. We do not attempt to revise missing values if the period of consecutive missing values is more than 24 hours to avoid confusion with a possible “stuck gauge”. Any unrecoverable missing values of either -9999 or -99 are replaced as -88 (unavailable rather than missing).

2.4 *Gross Range Error Check*

We impose a constraint that the hourly PP value is less than 5 inches for the gross range error check. The failed hourly PP value is replaced as -88. Also, any recoverable missing values of either -9999 or -99 are replaced as -88 (unavailable rather than missing values). We then convert each of the preprocessing steps to produce the baseline hourly precipitation product.

2.5 *Neighborhood Database*

Another preprocessing step is to generate a list of neighboring stations. We generate a database of neighboring stations which are within 0.25 degrees latitude and longitude. We generate this database for future data processing, anticipating the need to quickly extract neighboring station information. The database includes information to retrieve the exact file along with distance from the target station.

3. QUALITY FLAGS

3.1 *Baseline Quality Flag*

The first quality flag that is generated is the baseline quality flag associated with the baseline hourly precipitation product. This file is intended to provide overall characteristics of the precipitation product following the approach of Tollerud et al. (2005). An example of the baseline quality flag file is shown below:

```
AVLN7 200409 715 716 716 716   1   0   3   0  98 620
```

Field 1: Station name (NWS Location ID)
Field 2: Year
Field 3: Month
Field 4: Number of reports in 15-minute interval
Field 8: The number of the columns used for HPD
Field 9: Number of negative rain but not -88.0
Field 10: Number of -88.0
Field 11: Number of heavy rains (>= 1.0 in/hr)
Field 12: Number of rains (> 0.0 but < 1.0 in/hr)
Field 13: Number of no rains (= 0.0 in/hr)

At the station AVLN7 (Asheville, North Carolina), the summary statistics show valuable information on the health of the hourly precipitation data set. These summary statistics are derived from the original SHEF reported 15-minute values (Field 4-7 are all non-zero). The baseline hourly precipitation is assigned to the top of the hour (i.e. Field 8 shows 1), and there are three unusable precipitation data values (i.e. Field 10 is 3). The number of hours for this month should be 24 (hours) X 30 (days) + 1 hour = 721, which is equal to the sum of Field 9 through 13. The difference of numbers from each bin (i.e., Fields 4 through 7) indicates the number of missing values. Many negative values in Field 9 are also helpful to trace a potential error as seen below:

```
HUGS1 200504 672 672 672 672   1 226  49   0  19 426
```

The station HUGS1 reports 49 missing values (Field 10) and 226 negative values (Field 9) during April, 2005. Investigations revealed that the station in question had been reporting stream gauge height (variable name is HG) in place of the PC value. We decided to eliminate this station from our database. The baseline quality flag file (extension *.QFL) will help identify such encoding errors of observer.

3.2 *Buddy Quality Flag*

The preprocessed neighborhood database (*.BUD file) is used for the buddy quality flagging (buddy-check). Any station is compared with its neighbors located within +/- 0.25 degrees latitude and longitude. The station is an outlier if any hourly precipitation value is greater than 2.2 times the standard deviation of all the buddy values at that hour (flagged as -66.11). If there are no neighboring stations, then flagging statistics use default values. The station of AVLN7 for September, 2004 is shown below:

```
AVLN7 200409 715 716 716 716   1   0   3   0  98 620
      BUD   6 0.21 2.20   1   0 409 311
```

The column after BUD:

- 1st Number of stations including the target station (= 6)
- 2nd Average distance to neighbors (= 0.21 deg)
- 3rd The multiplier to standard deviation (=2.20)
- 4th Number of flagged stations in the neighbor (=1)
- 5th Cases when number of neighbors became less than or equal to 2 during the month (=0)
- 6th Cases when number of neighbors became greater than or equal to 6 during the month (=409)
- 7th Cases when number of neighbors are between 3 and 5 (=311).

If the number of stations used for the buddy check (1st column) is less than or equal to 2, no check is performed. The standard deviation test will be more reliable with increasing data values, i.e. more hourly reporting, neighboring stations. Further, the standard deviation is evaluated each hour with the intention of capturing the variability of hourly precipitation. The idea is to not reject isolated events.

3.3 *Radar Quality Flag*

Each station that went through quality flagging against the NEXRAD DPA values appends summary statistics after the code "RAD". This module of quality flagging can be done sequentially, but the example shown below is the result applied independently of the buddy-checking. This module follows a similar algorithm of Marzen (2004) who developed quality control procedures of gauge measurements for the Florida Water Management District network and NWS hourly precipitation products. Let us define R = Radar derived hourly precipitation, G = Gauge hourly precipitation, R_{max} = Maximum radar estimate from the surrounding 8 cells of DPA, D is absolute value of difference of G and R, namely, D = abs(G – R).

There are 6 cases of quality flags;

Case 1. If R_{max} = 0 and G > 0.4, then flag it as -77.11

Case 2. If R_{max} > 0 and G > 0.4 and G/R_{max} > 6.67, then flag it -77.22

Case 3. If R > 0 and G > 0 and D >= 1.0 and G/R_{max} > 5.0, then flag it -77.33

Case 4. If R > 0 and G > 0 and 0.5 < D < 1 and G/R_{max} > 6.67, then flag it -77.44

Case 5. If R > 0 and G > 0 and 0.25 =< D =< 0.5 and G/R_{max} > 10, then flag it -77.55

Case 6. If R > 0 and G < 0.1 and D > 0.5 and G/R< 0.11, then flag it -77.66

The flag file at the station of AVLN7 for September 2004 is shown below:

```
AVLN7 200409 715 716 716 716   1   0   3   0  98 620
      RAD   3   1   0   0   1   1   0
```

The first column after RAD is total number of flags and the following 6 columns represent the number of corresponding cases. In this example, three data values are flagged by NEXRAD DPA data. The threshold values are initially set to the same values as Marzen (2004) who derived thresholds from Florida Water Management District networks. Future work will refine these values.

4. COMPARISON WITH NCEP/UCAR ARCHIVE

The OHD's HADS program also derives hourly precipitation value (e.g., HADS PP) for the National Center for Environmental Prediction (NCEP), and they are being used for precipitation analysis and assimilation for operational numerical weather prediction (NWP) (Lin and Mitchell, 2005). These data are saved at NCEP and then archived at UCAR a month later (http://www.joss.ucar.edu/codiac/, search: Precipitation NCEP/EMC Gage Only Hourly Dataset). The NCEP/UCAR archived hourly precipitation data are originally derived from the HADS data set. So we decided to compare our baseline hourly precipitation data with NCEP/UCAR archived hourly PP product using two measures: 1) The number of missing hours, and 2) The number of stations reporting at the top of the UTC hour. A subset of NCEP/UCAR archive is selected to ascertain the same HADS origin. In the pilot domain, there were more than 500 stations each month during April through September, 2004.

4.1 *Number of Missing Values*

The NCEP/UCAR archive does not include periods when data are missing, so we have to insert a missing value when no report exists at that hour. Figure 2 shows the mean of the percentage of missing data values for all stations in the domain out of all available hours. This mean varies over the month because the number of stations reporting to the HADS varies. For the warm season of 2004, our quality control measures have revised missing data values such that the mean of the percentage of missing data values has decreased from a high of 11% to about 2%, (Figure 2).

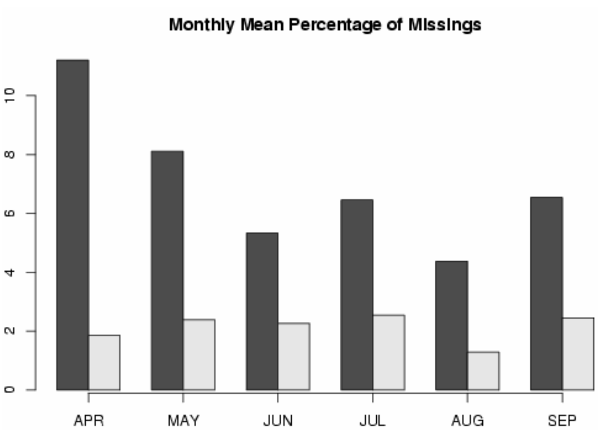

Figure 2. The mean percentage of missing data during April through September, 2004. Darker bars stand for percent of missing data from NCEP/UCAR archive, while juxtaposed bars are from reprocessed HADS data. Reprocessed HADS reduced the number of missing data values in the baseline hourly precipitation.

4.2 *Number of Observations at Top Hour*

The NCEP/UCAR archive includes observation time to the minutes in hourly PP data. While most of the reports are at the top of hour (i.e., 0^{th} minute), about 35% were *non-top-of-the-hour* observations. Figure 3 shows distribution of *non-top-of-the-hour* in 15-minutes bins in September 2004. Even though we do not know processors involved in NCEP/UCAR archive, most of the reprocessed HADS precipitation data are at the top hour. We have noticed that several stations of 30-minutes off top hour in NCEP/UCAR archive were actually reporting semi-hourly (i.e., 0^{th} and 30^{th} minute), in the historical HADS data.

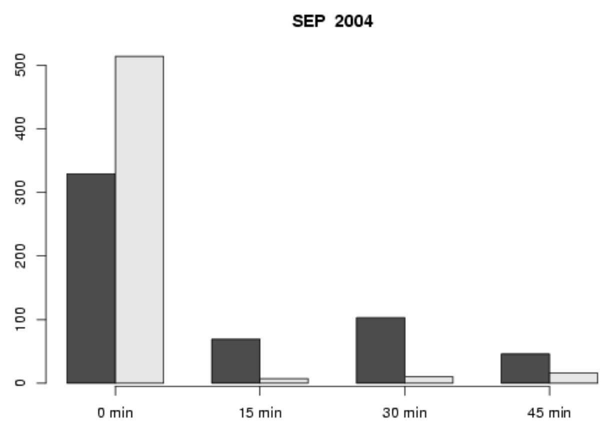

Figure 3. The average observations that are off the top hour in 15min bins during the month of September, 2004 in the Carolina domain. Darker bar stand for station numbers from NCEP/UCAR archive, while juxtaposed bars are from reprocessed HADS data. Reprocessed HADS reduced temporal representative errors.

5. EXAMPLES OF HOURLY PRODUCT

In September 2004, four heavy rainfall events were associated with hurricanes. Figure 4 shows the time-series of hourly precipitation during September 2004 at the station in Asheville, North Carolina (AVLN7). Three data points were missing as marked by the “red dot” at 0000UTC September 1, and 1400UTC and 1500UTC September 16. The bottom of Figure 4 shows the same time-series with red open circles which correspond to the radar quality flag and green open circles which correspond to the buddy quality flag. The two quality modules did not flag the same data. The three values that were flagged by the radar quality module (red circles) are at 0300UTC, 0500UTC and 0700UTC on September 17 when all are heavy rainfall cases. The valued that is flagged by the buddy check (green circle) is more effective to pick up small rainfall amounts which are inconsistent with 5 neighbors reporting 0 in/hr at 2200 UTC September 28 as opposed to the target value of 0.01 in/hr.

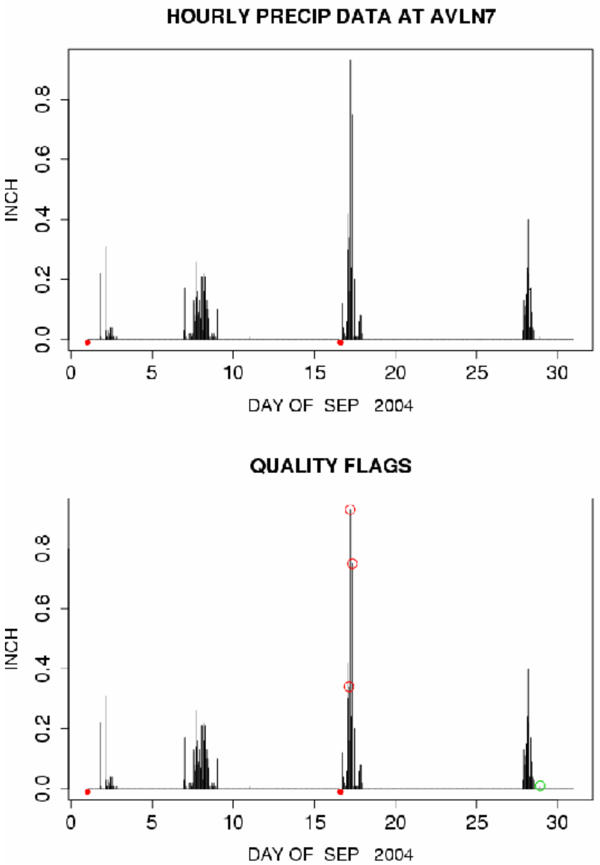

Figure 4. (Top) The time-series of hourly precipitation during September 2004 at the station in Asheville (AVLN7), North Carolina. In this data, two were recorded as missing as marked “red dot”. (Bottom) The same as above except additional open circles of “red” as flagged by NEXRAD data, and “green” as flagged by buddy check.

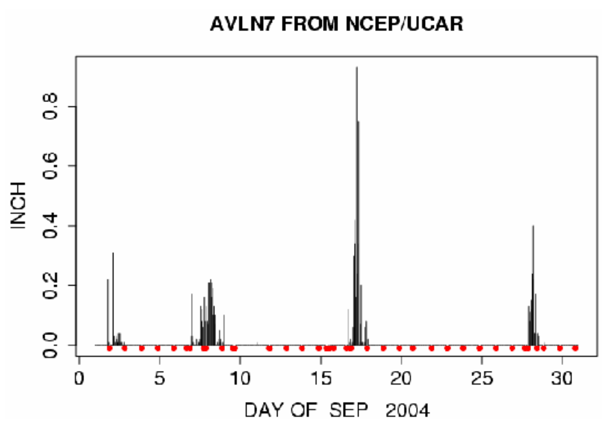

Figure 5. The time-series of hourly precipitation at AVLN7 from NCEP/UCAR archive. There are 48 missing data as marked by “red dots”. All the observed data show identical time-series to Figure 4. The original HADS report was 15 minute PC.

Figure 5 shows the hourly precipitation at AVLN7 from the NCEP/UCAR archive. There are 48 cases of missing data. More than half of them were at 2000UTC and 2100 UTC. Some missing data values occurred during rain events. We are speculating that data transmission missed the cut-off time.

Figure 6 shows data from the station at Leicester, North Carolina (MHLN7), which is located to the northwest of the station AVLN7. Both the radar and buddy-check flagging modules worked as expected. The buddy check was more effective as it picked up 7 data values which are all 0.01 in/hr in the baseline hourly precipitation and all neighborhood data were no rain. The possibility of over-flagging by the radar quality flagging module is under examination in this case as a possible result of under-estimation by the radar.

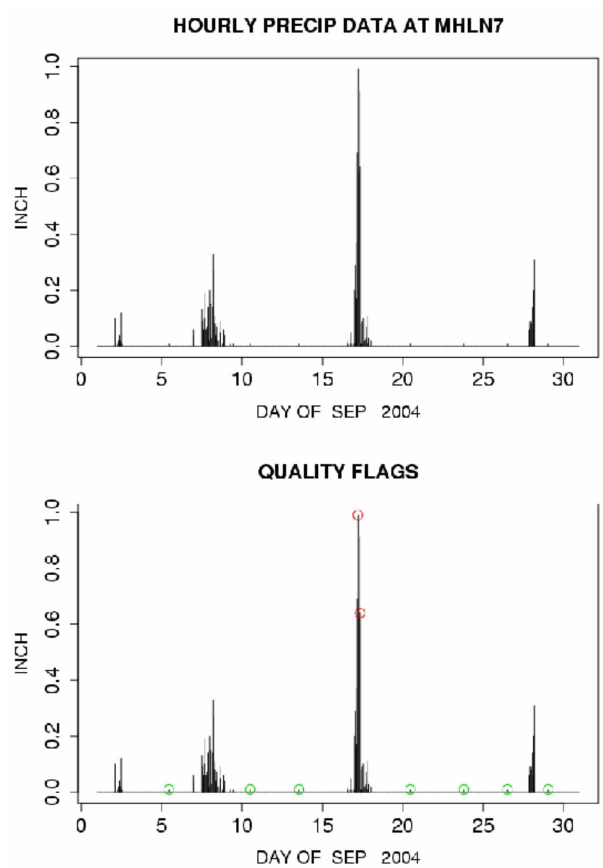

Figure 6. (Top) The time-series of hourly precipitation during September 2004 at the station in Leicester (MHLN7), North Carolina. None are missing In this data. (Bottom) Two are flagged by NEXRAD data, and 7 are flagged by buddy check.

6. SUMMARY AND PLAN

The reprocessing of historical HADS precipitation offers opportunities to revise missing data values and to reducing processing errors associated with *non-top-of-the-hour* observations. The preliminary results of early version already show the benefits of reprocessing of HADS data, but we plan refinements of quality flagging modules. Refinements will include rigorous comparisons with accumulated HADS hourly precipitation estimates and other daily precipitation data (NOAA and non-NOAA networks).

Our main motivation to reprocess historical HADS precipitation has grown from the need for a high density long-term in-situ rainfall data set as input to the multi-sensor precipitation reanalysis. The reanalysis of multi-sensor precipitation estimate is also planned.

The quality assurance of precipitation data is the most important and difficult problem that users have been facing (Shi et. al, 2003; Smith et. al, 2004). The development of QC with historical HADS precipitation can lead to QC automation of real-time data. Efforts in progress at the NWS include compiling manual precipitation QC practices at River Forecast Centers.

The early version baseline hourly precipitation data and associated quality flag files in pilot domain for April-September, 2002-2005, are available upon request.

7. ACKNOWLEDGMENT

The authors acknowledge reviews by Drs. Bates and LeDuc. Prof. Fuelberg provided information of his research activity at The Florida State University. The authors also extend thanks to Drs. Kondragunta of OHD and Tollerud of NOAA/FSL (now NOAA/ERSL) for continued discussions on precipitation QC issues.

8. REFERENCES

Lin, Y. and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and application. Preprints, AMS 19th Conf on Hydrology, San Diego.

Marzen, J., 2004: Development of a Florida high-resolution multisensor precipitation dataset for 1996-2001 – quality control and verification, M.S. Thesis, Department of Meteorology, Florida State University, 2004, pp86.

Nelson, B., D. Kim, J. Bates and D. J. Seo, 2006: Multi-sensor Precipitation Estimation Reanalysis, AMS 20th Conf on Hydrology, 29 Jan.- 2 Feb., 2006, Atlanta, GA

Shi, W., E. Yarosh, R.W. Higgins and R. Joyce, 2003: Processing daily rain-gauge precipitation data for the Americas at the NOAA Climate Prediction Center, 19th Conf on IIPS, 10-13, Feb. 2003, Long Beach, CA

Smith, M. and coauthors, 2004: NOAA NWS Distributed Hydrologic Modeling Research and Development, NOAA Tech Rep NWS 45, U.S. Dep. Of Commerce, pp62, 2004

Tollerud, E., R. Collander, Y. Lin, and A. Loughe, 2005: On the performance, impact, and liabilities of automated precipitation gauge screening algorithms, 21st Conf on Weather Analysis and Forecasting, 1-5, Aug. 2005, Washington DC.