Reprinted from the Preprint Volume of the Eighth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, January 5-10, 1992, Atlanta, Ga. Published by the American Meteorological Society, Boston, Mass.

2.7

A DATA QUALITY DESCRIPTION MODEL FOR HYDROMETEOROLOGICAL OPERATIONS

Vernon C. Bissell
NOAA, National Weather Service
Northwest River Forecast Center
Portland, Oregon 97209

Jeffrey R. Zimmerman[1]
Research and Data Systems Corporation
Greenbelt, Maryland 20770

1. INTRODUCTION

Quality control of vast quantities of observed data is a major obstacle to efficient and rapid processing in modern forecasting and warning systems. The National Weather Service plans for the Advanced Weather Interactive Processing System (AWIPS) to have most data quality checking done at the system entry point so that the same steps need not be repeated at every point in the data distribution system. This implies a need to transmit quality description information along with data values. Furthermore, there is a need for consistent rules to identify and use quality aspects in processing databases. It is clear that a "quality description model" is essential in order to integrate many diverse and complex data networks and processing functions into a single efficient system. We will give the current status of a "data description model" now being considered for hydrologic data operations.

The key to data quality description is a pair of one-byte numeric descriptors, with an additional single character descriptor derived from the numeric pair. These are attached to each data value. The numeric pair permits modular quality checking, updating the quality measure attached to specific reported values even if the prior and subsequent checking processes are unknown to each other. The character descriptor gives efficient selective retrieval of data for quality-sensitive applications. The "data description model" also has guidelines for use of descriptive elements in a local database (e.g., the number of estimates and other special values permitted for a given observational element at a given time).

The definition of the numerical quality descriptors is outlined in the PROTEUS TECHNICAL NOTES. These are not final design documents, but rather are semi-formal working documents used by the National Weather Service in the PROTEUS project, a risk-reduction activity for planning hydrologic operations for AWIPS.

[1]Also affiliated with NOAA/NWS/Office of Hydrology

2. DATA QUALITY DESCRIPTORS

Some data quality description is already available in the "data qualifier" character of the Standard Hydrometeorological Exchange Format (SHEF; Pasteris, 1983; Bissell et al, 1984). This descriptor is very useful for selective retrievals and general description, but is deficient for most uses because it doesn't provide information needed to revise the "opinion" of a data value with later modular checking processes. This section discusses both numerical quality descriptors and the single character descriptor. The primary purpose for the numerical descriptors is to provide an objective way to revise the opinion data quality.

2.1 Numeric Descriptors (QDS,QPI) for Operational Data Quality

In data systems with individual reported values inspected by more than one checking scheme, there is a need to describe two things about data quality. These are:

(1) "what is the level of sophistication of the procedure which has been used to check the value?"; and

(2) "what was the opinion of that procedure, expressed in some numeric measure?"

The need for descriptors of these two elements is recognized in both the PROTEUS database design and the Columbia River Operational Hydromet Management System (CROHMS) (Speers, 1980). Integers in the range 0-255 will be used to measure each of these quality elements. Thus, full numeric quality description of a data value can be stored or transmitted in a packed (binary) form in a total of two bytes per data value. The two quality element measures are "quality process identifier" (QPI) and "quality departure score" (QDS) (Bissell, 1989a).

2.1.1 Quality Process Identifier (QPI)

The quality process identifier (QPI) measures sophistication of the inspection process used. QPI is zero when no checking of any kind has been applied, has low value for minimal checks, and increases for more discriminating procedures. The QPI should be a "generic" measure rather than a specific process number associated with only one procedure. Table 1 gives proposed guidelines for selecting QPI values. These use three levels of significance to derive QPI. The most significant is 'LEVEL' (having value between 0 and 3), broadly characterizing sophistication of the procedure being used. Within each 'LEVEL' there is a 'SCORE'(with value between 0 and 5), having individual components reflecting the degree to which adjacent station data and forecasts are used and the time range of data used in the test. Finally, an additive element (having value between 0 and 15) adds measure if the value is a revision or if there is a seasonal range check or if there is a fixed range check.

A QPI value of 255 represents an absolute check which can not be superseded except by a 'revision'.

Table 1. Proposed guidelines for QPI values

```
QPI = 16 * process score (0-15)
     + 8 * seasonal range check flag(0-1)
     + 4 * fixed range check flag (0-1)
     + 2 * extra discriminator flag #1
           (0-1)
     + 1 * extra discriminator flag #2
           (0-1) (local use flag)
     0 <= QPI <= 255

Set QPI=254      if this comes out at 255
Set QPI=8        if calculated is less than
                 8 when revision flag is on
Set QPI=255      for absolute acceptance or
                 rejection

process score = LEVEL *[ group data score
                 (0-2) + forecast score
                 (0-2) + time range of
                 vals used (0-1)   ]
where

LEVEL = 0 fixed criteria (minimal
          acceptance)
        1  screen  with  some  dynamic
          element
        2 basic verification applied
        3 advanced verification applied
```

2.1.2 Quality Departure Score (QDS)

Table 2 describes the proposed "quality departure score" (QDS). The QDS is a departure from the conditional expected value expressed in tenths of standard deviations, multiplied by four. For data checking procedures which have inherent means of stating the number of standard deviations from expected value, assignment of QDS will be fairly easy. For procedures which do not have this capability inherent, the assignment of QDS for specific results will be somewhat subjective, but nevertheless can be done.

The proposed QDS has two elements. First, departure from expected can be expressed out to 6.3 standard deviations (a real occurrence about once in every one or two billion in a normal distribution). In 'smart' procedures, this measure will be based on the conditional distribution of values (based on a forecast, persistence, correlation with adjacent stations, or some other method) which reflects information available.

Second, by 'spreading' the scores by a factor of four, the score can also tell if the value is higher or lower than the value expected. Each quadruplet of QDS scores represents the same level of departure from expected. For example, a QDS value of 44 means that the subject value is between 1.0 and 1.1 standard deviations from the conditionally expected value, with no indication if it is above or below. A QDS of 45/46 indicates the same standard deviation range, but says the subject value is less/greater than conditionally expected. A QDS of 47 would be reserved for future use. A QDS value of 48 then begins the range between 1.1 and 1.2 standard deviations.

Table 2. Proposed QDS description

```
     QDS = 4 * IFFT + IS,  4 <= QDS <=255

where

     IFFT = integer frequency factor
            (tenths)
          = tenths of standard deviations
            from expected value
            (1 <= IFFT <= 63)

     IS   = 0  departure score is
               unsigned;
            1   value  tested  less  than
               expected;
            2  value tested greater than
               expected;
            3  reserved for future use;
```

When QDDS is less than 3 (IFFT=0), the data value is considered NULL (no departure score assigned).

---

3. REPLACEMENT/RETENTION CRITERIA FOR NUMERIC QUALITY DESCRIPTORS

3.1 The Need

Numeric quality descriptors (QDS,QPI) are associated with each and every data value to be used operationally. Each value may be checked by several different validation procedures, so we must know what to do with each new "opinion" (QDS, QPI) as it becomes available (Bissell, 1990). One approach could store many (QDS,QPI) "opinion" pairs with each data value, one for each quality checking routine applied, and leave it to every user program to sift through the opinions in deciding what to do. This would be inefficient from a storage/retrieval perspective, and would add substantial communication overhead in distributed systems where multiple checks are performed at different sites for example.

It seems clear that criteria must be established by which one opinion can be selected as preferred over an alternate opinion at each step of the data inspection process. The single opinion would then in some sense represent the composite effort of all checking done up to that point. Storage and transmission requirements for quality description can then be reduced to a workable level. So the question boils down at each step to asking "Do I RETAIN my prior opinion of the data value, or do I REPLACE the prior opinion with the one just generated, or do I replace the prior opinion with a COMPOSITE of the prior and most recent opinions?"

In terms of the discussion which will follow, our problem is:

i. Given a prior opinion (QDSp,QPIp) of a data value, and
ii. Given the opinion (QDSt,QPIt) of the most recent test,
iii. What will be the single resultant opinion (QDSn,QPIn) reflecting all quality checks conducted?

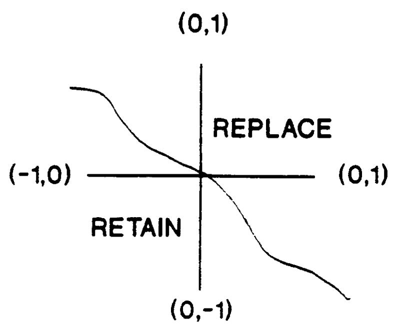

Figure 1. Nature of the replacement criterion line.

3.2 Philosophy of Replacement/Retention

Figure 1 demonstrates the general nature of the criterion replacement question, showing a hypothetical "line of equal opinion" on a plot of the (QPI,QDS) "opinion pairs". By definition the line of equal opinion must pass through the prior opinion (QPI1,QDS1) and is a locus of possible most recent opinions which would be considered of equal merit to assess reliability of the observed value.

The line of equal opinion delineates two regions. One retains the prior opinion and the other replaces the prior opinion with the most recent opinion (we're not yet considering compositing of the two).

Opinion replacement occurs if the most recent quality test result falls above and to the right of the "line of equal opinion". This criterion is conservative because descriptors which indicate more suspicion of the data value's integrity tend to be retained. One way of looking at the replacement criterion is to say that in order to replace numeric quality descriptors, the gain in process sophistication going from prior opinion to most recent opinion must be greater than the loss of reasonableness going from the prior to the most recent opinion.

The rationale for this approach is explained by considering increments (1,0), (-1,0), (0,1), and (0,-1) from prior to most recent (QPI,QDS). If the increment is (1,0), we would REPLACE the old opinion with the new opinion, and if the increment is (-1,0) we will RETAIN the prior description. This is because we wish to take a conservative approach to operational data which may be fed into models and other guidance tools. We therefore choose to take the more pessimistic view of the observed value's integrity if the two test procedures have equal skill.

Going the other direction, if the increment is (0,1), we would REPLACE the prior descriptors (opinion) with the more recent opinion. If the increment is (0,-1) we will RETAIN the prior description. The reason here is that if two quality control checks give the same measure of reasonableness of the value, we should retain the descriptor of the more sophisticated of the two checking procedures.

So reference to Figure 1 shows the general nature of the decision line (line of equal opinion). For continuity, the decision line must pass through the prior opinion (QDS1,QPI1), and because of the foregoing discussion the decision line must pass in some fashion through the second and fourth quadrants of the plot. Thus the general nature of the replacement criterion is obtained.

For now, we select the decision line to be a straight line simply because we have no basis for anything fancier. The slope of the line is (-1*KQ), where KQ is the parameter of the relationship. A nominal value of KQ=1 will be used at the outset unless/ until a different value can be justified. The general form and specific parameters of the replacement line ultimately depend explicitly on guidelines (like Table 1) which assign quality process identifier (QPI) values. It's expected that some adjustment to the replacement line and the QPI guidelines will be made as experience is gained with the system.

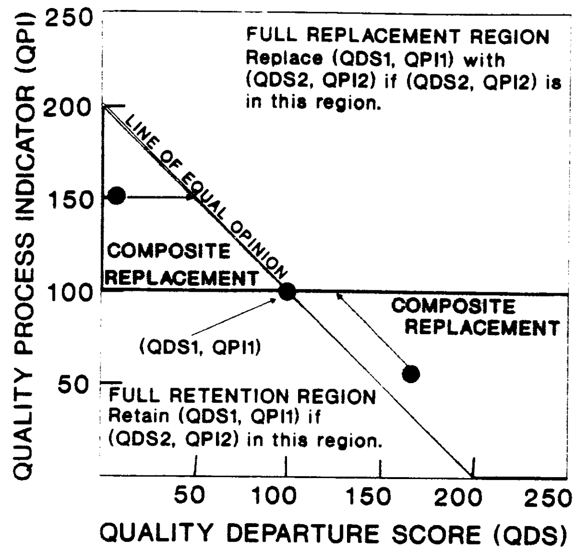

Figure 2. Composite/continuous replacement.

3.3 Composite/Continuous Replacement

One problem with the simple retain-or-replace algorithm is that the resulting final opinion can be erratic, taking big jumps given only a small change in the most recent test result. We now add the requirement that the resultant opinion must be a continuous function of the most recent opinion. The algorithm used to accomplish this can produce a resultant opinion which is a composite of the prior and most recent opinions (instead of simply choosing one or the other), as shown by the areas labeled "composite replacement" in Fig. 2. These are produced by requiring that in compositing a prior opinion P=(QDSp,QPIp) and a new test result T=(QDSt,QPIt), the new result N=(QDSn,QPIn) will have QPIn which is the higher of QPIp and QPIt.

Figure 2 shows how test results falling in the composite replacement region to the left and above from the prior opinion will move to the new composite opinion by "sliding" to the right to the line of equal opinion. Test results falling in the composite region to the right and below from the prior composite opinion will move to the new composite opinion by "sliding" up a line parallel to the line of equal opinion to the point where QPIn=QPIp. Thus, given

```
Prior opinion   = (QDSp,QPIp)
Recent test     = (QDSt,QPIt)
```

we have the new composite opinion given by

```
QPIn = max [QPIp, QPIt]
QDSn = ( QPIEXTmax - QPIn ) / KQ
```

where

```
QPIEXTmax = max [ QPIEXTp, QPIEXTt ]
QPIEXTp   = QPIp + KQ*QDSp
QPIEXTt   = QPIt + KQ*QDSt
```

We may wish to think of QPIEXTmax as the lowest QPI process which will fully extinguish memory of all prior tests in the composite.

It is very fundamental that this formulation replaces a single former opinion using a single recent opinion from a single test. Thus only the replacement opinion needs to be carried forward to the next step if sequential tests are being performed. The final result is independent of the order in which sequential tests are performed. This can be seen in considering the alternate approach of saving all test results (QDSj,QPIj) for j=1,..m. The final composite "next" opinion is given by

```
QPIn =  max [ QPI1, QPI2, ...,QPIm ]
QDSn = (QPIEXTmax - QPIn ) / KQ
```

where

```
QPIEXTmax = max [ QPIEXT1,...,
                  QPIEXTm ]
```

and

```
QPIEXTj   = QPIj + KQ*QDSj
            for tests j=1,..m
```

This is a very fast simple algorithm. Selection of the composite opinion from several stored opinions is not computationally intensive, so system designers may choose either to keep only the composite as testing progresses, or keep all test results and calculate the composite opinion when needed. Probably the best choice would be to store (QDS, QPI) for the test having the highest QPI and for the test having the highest QPIEXT.

3.4 Application example

Consider the case where three tests are performed on a report of six-hour precipitation which would be exceeded only once in twenty years in the given month. Tests will be climatology, persistence and forecast. Details in selection of QPI and QDS are omitted. Figure 3 shows results.

TEST 1: Climatology

The probability of exceedance is 1/2400, giving QDS=132. Assign QPI=12 to the seasonal climatology check.

```
(QDS1,QPI1) = (132, 12)
QPIEXT1 = 12 + 1*132 = 144
```

TEST 2: Persistence

Suppose the observation is 3.5 standard deviations from the conditional mean based on prior observation(s), then QDS = 3.5*40 = 140. Assign QPI=20 to the persistence test.

(QDS2,QPI2) = (140, 20)
QPIEXT2 = 20 + 1*140 = 160

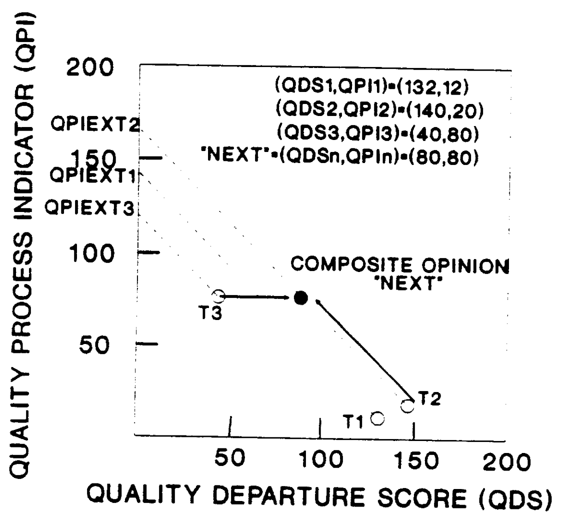

Figure 3. Example of "opinion replacement."

TEST 3: Forecast

Suppose a test using forecasts has QPI=80, and the observed value is one standard deviation greater than the mean forecast. Then QDS=1.0*40.

(QDS3,QPI3) = (40,80)
QPIEXT3 = 80 + 1*40 = 120

The final composited opinion resulting from the three tests is

QPIn = max [QPIj] = 80
QDSn = (QPIEXTmax - QPIn)/KQ = (160-80)/1 = 80

Figure 3 shows that TEST 2 completely erases any memory of TEST 1 (full opinion replacement), while RESULT 3 is a composite of the TEST 2 result and the TEST 3 result.

4. TRANSLATION BETWEEN NUMERIC DATA QUALITY DESCRIPTORS AND SHEF "DATA QUALIFIER"

4.1 The need

The numerical quality descriptors are intended primarily for use during sequential quality checking processes, although they may also be used by specialized user programs. On the other hand, the SHEF "data qualifier" character gives a simplified external transmission of data quality description, allows simplified selective retrieval of data by generalized quality level, and gives a simple visual means to visually identify data quality on a display or printout. SHEF data qualifiers which show quality are V="verified", S="screened", Z="null", Q="questioned", and 'R'=rejected. This report uses two proposed changes to the SHEF data qualifier. First, X is used for rejected instead of R. Second, C="coarse check" is used to indicate "minimal acceptance".

Since checking procedures would typically produce numerical descriptors, we need to translate numeric results into the user-friendly SHEF data qualifier for most user programs. There is also a need to define defaults by which data qualifiers can be translated into numeric descriptors. The following description of these translations follows Bissell (1989b).

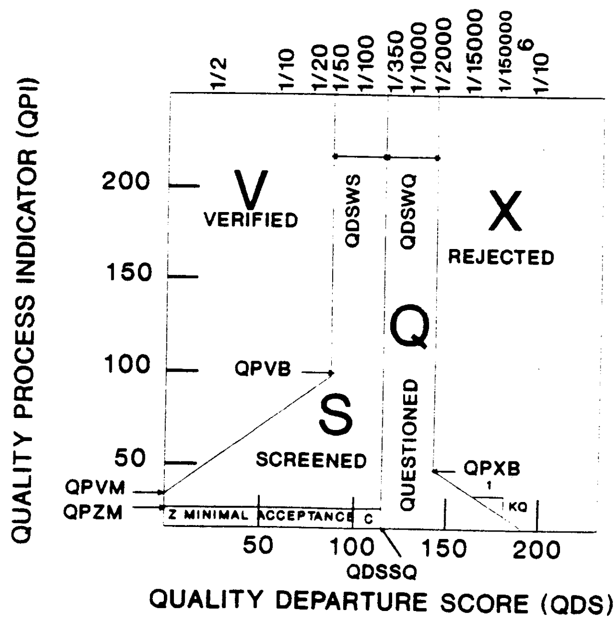

Figure 4. Translate (QDS,QPI) to data qualifier.

4.2 Translation from numeric description (QDS,QPI) to SHEF Data Qualifier

Figure 4 shows assignment areas of SHEF data qualifier based on (QDS,QPI). These are defined by parameters (see Table 3) which would be system installation standards, except for one (QDSSQ). QDSSQ is unique to each individual data time series based on the sensor quality and reliability. This permits a highly reliable time series (high QDSSQ) to accept extremely uncommon values as screened or verified, while restricting poor quality time series (low QDSSQ) to accepting only the most reasonable values as screened or verified. The assignment regions are very specifically related to the definitions of QDS and QPI as given in Table 1 and Table 2. Nominal values of the parameters are given in Table 3.

The sloped portion of the line separating the 'Q' region from the 'X' region has slope (-1*KQ), same as the slope of the replacement/retention line.

The "minimal acceptance" region (proposed SHEF data qualifier 'C') is provided because some minimal checking procedures will be insufficient to label a passing value as "screened". Values which have had no data checking of any kind will be assigned the 'Z' ("null") qualifier.

Across the top of Figure 4 is a fraction giving the probability that a valid data value would have a QDS of that value or higher (since QDS shows some number of standard deviations away from the conditionally expected value). With the nominal value of QDSSQ=115, the cutoff between acceptable values ('C' or 'S') and questioned values ('Q') would occur at a little less than three standard deviations from the conditional mean, or about one in every 300 valid occurrences. Data from a very good sensor could have QDSSQ=140 so that only one in every 2000 valid occurrences would be marked as questioned.

Table 3. Nominal parameters in numeric-to-SHEF

SYSTEM-FIXED PARAMETERS

QPZM = 16 (minimum QPI for "screened")
QPVM = 32 (minimum QPI for "verified")
QPVB = 96 (QPI at V-S breakpoint)
QPXB = 48 (QPI at Q-X breakpoint)
QDSWS = 25 ("screened" region width)
QDSWQ = 25 ("questioned" region width)
KQ = 1 (neg. of replacement line slope)

TIME-SERIES SPECIFIC PARAMETERS

QDSSQ = 115 (QDS at S-Q boundary)

4.3 Translation from SHEF data qualifier to numeric (QDS,QPI)

Mapping the (QDS,QPI) regions into data qualifiers is a mapping of many-to-one, and is unique. Going the other way we have a one-to-many mapping and must select defaults in translating a data qualifier character into numeric descriptors. The suggested defaults are shown in Figure 5 and tabulated in Table 4. There are two cases. First consider a value which has been received with only with a data qualifier. The default values for (QDS,QPI) in this case are shown as V*, Z*, C*, S*, Q*, X*, and F*. The second case occurs if a value of QPI is received along with the data qualifier. Defaults in this case are represented by the dashed lines. The nominal location of these default lines is suggested at two-thirds of the way from the lowest QDS to the highest QDS possible within that data qualifier region for the given QPI.

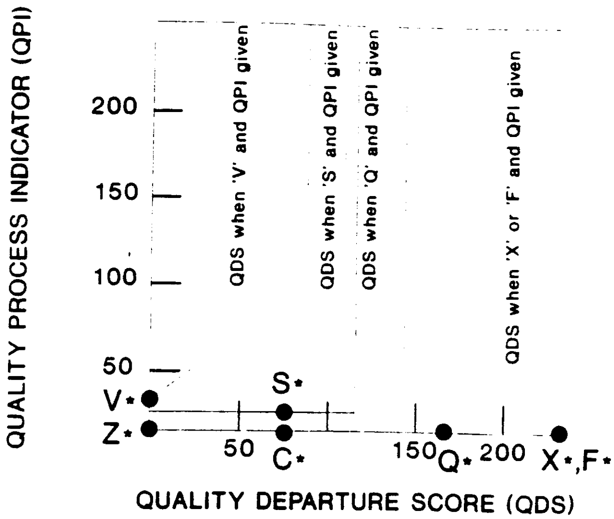

Figure 5. Defaults in SHEF-to-numeric translation.

Table 4. Nominal defaults for SHEF-to-numeric

Input DQ	Resultant Default (QDS,QPI)
'V'	V* = (0 ; QPVM)
'S'	S* = (2*QDSSQ/3 ; QPZM +1)
'C'	C* = (2*QDSSQ/3 ; 1)
'Z'	Z* = (0 ; 0)
'Q'	Q* = (QDSSQ+2*(QDSWQ+QPXB/KQ)/3 ; 1)
'X'	X* = (170+(QDSSQ+QDSWQ+QPXB/KQ)/3;1)
'F'	F* :treat "flagged" like reject X

5. USE CONVENTIONS OF QUALITY DESCRIPTORS

Fundamental to the use of this data quality description model is the capability of the data base to recognize three different categories of data values: actual; estimated; and wrong, replaced (Bissell, 1989c). Data falling in the actual category will have those character quality descriptors described in section 4.2. An estimated piece of data will have a character quality descriptor of "E", while a data falling in the wrong, replaced category will have a character quality descriptor of "W".

5.1 Actual

When a specific piece of data enters the system, a number of different quality control procedures will be performed and an opinion of the data will be generated. The appropriate descriptors will be appended to the data, and the data value and the descriptors will be posted to the data base. Subsequent to posting, some additional information may be received which will initiate more sophisticated quality control checks. These checks may result in a revised opinion of the data (see section 3). The new descriptors will be posted; however, the reported data value will remain unchanged.

5.2 Estimated

The forecaster would have the option of estimating the value of a specific piece of data. This estimation may arise due to missing data, or the forecaster may have some information that will enable a reasonable estimate to be made. The specifications can be established to allow multiple estimates to be retained; however, if more than one estimate is retained, procedures will need to be established to determine a priority for using estimated values in applications.

5.3 Wrong, Replaced

In those instances where a revised data report is recorded, the initial report will not be discarded. The initial report will be assigned a character quality descriptor of "W". The revised report will undergo the quality checking procedures, and be assigned a character quality descriptor in the actual category.

6. CURRENT STATUS

The CROHMS data base, and the ALERT data base designed by the California Nevada RFC, were designed with this data description model in mind. This data description model is also being recommended by the Office of Hydrology for inclusion in the AWIPS-90 design currently being specified by the NWS.

7. SUMMARY

A data quality description model has been described. The model quantifies the sophistication of the checking procedure, and the opinion of the data quality by that checking procedure into a numeric pair of descriptors. The numeric descriptors serve a multitude of purposes. A single, character descriptor can be derived from the numeric pair. The character descriptor may then be used to facilitate retrieval of data by quality-sensitive application programs, and compilation of statistics. The numeric pair provides a methodology for updating the opinion of the quality of a piece of data. Separate quality control checks may be carried out, and the numeric results compared. A set of rules governs the determination of the resulting opinion, and the process is independent of the order in which the quality control checks were performed. The model allows for estimates, and provides guidelines for the use of the descriptors with a local database.

REFERENCES

Bissell, V.C., P.A. Pasteris, and D.G. Bennett,1984: Standard Hydrologic Exchange Format (SHEF). J. Wtr. Resources Pln. and Mgmt. Div.,110 392-401.

______, 1989a: DATA QUALITY DESCRIPTION: Numeric Descriptors (QDS,QPI) for Operational Data Quality. PROTEUS Technical Notes, NW0001/Rev. 1, NOAA, National Weather Service, Silver Spring, MD, 5 pp.

______, 1989b: DATA QUALITY DESCRIPTION: Translation between numeric data quality descriptors and the SHEF data qualifier. PROTEUS Technical Notes, NW0003/Rev. 1, NOAA, National Weather Service, Silver Spring, MD, 5 pp.

______, 1989c: DATA QUALITY DESCRIPTION: Suggested conventions to use data qualifiers and numeric quality descriptors. PROTEUS Technical Notes, NW0004/Rev.1, NOAA, National Weather Service, Silver Spring, MD, 3 pp.

______, 1990: DATA QUALITY DESCRIPTION: Replacement/retention criterion for selection of 'primary' numeric quality control descriptors (QDS, QPI). PROTEUS Technical Notes, NW0002/Rev. 1, NOAA, National Weather Service, Silver Spring, MD, 6 pp.

Pasteris, P.A., 1983: Standard Hydrologic Exchange Format (SHEF) Version I, U.S. Dept. Commerce, National Weather Service, Tech. Memo NWS-WR-180, Salt Lake City, UT.

Speers, Douglas D., 1980: Columbia River Operational Hydromet Management System. Case Studies of Applied Advanced Data Collection and Management, Aerospace Division, ASCE Task Committee Pub., New York, N.Y., pp 337-345.