

## 4. DATA COMPILATION AND PROCESSING PROCEDURES

A major task of this study was the acquisition, compilation and digitization of historical data relating to the quality of water and sediment within the Galveston Bay system. The data acquired in this project can be broadly categorized as digital format and hard-copy format. The former refers to any magnetic medium capable of manipulation on the digital computer, e.g. magnetic tape, floppy disks, CD's, etc. The latter refers to field sheets, tabulations, and (sadly) computer printout from digital files that no longer exist. In the case of the hard-copy data, all of the significant data sets, and most of the insignificant (that we had access to), were keyboarded as a part of this project effort. This proved to be an extensive process, undertaken by the employ of a welter of data-entry gnomes who hammered away at the data sets over a period of months. It is probably not inaccurate to observe that the probability of marshalling this kind of data-entry effort in the future is unlikely, so certainly one of the major products of this project is the digital data base itself, which is described below. The further analysis of these data requires their conversion, combination and transformation in various ways, all of which can circumscribe the interpretation of the data. The general procedures used in this project are outlined here.

### 4.1 Data Set Construction

Because the data in this compilation was to be analyzed in later tasks of the project, part of the effort was invested in integrating the data into a computer-manipulable data base. In designing the formats, emphasis was placed on data structure that is transferrable and manipulable via microcomputers (especially PC's), i.e. compact ASCII files. It may be noted that Tetra Tech (1987) recommends a specific hierarchical format for NEP data sets. While certain aspects of this format are satisfactory—or at least workable—the recommendation suffers from two great deficiencies: (1) the data structure contains numerous redundant fields, which will greatly expand the storage requirements for a data set, (2) the structure is specific to SAS, which will necessitate either that software for its use or specific codes for conversion. Therefore, data-base formats were devised specific to this project, to facilitate transfer and use of the data by other workers. (SAS has a fairly robust repertoire of input formats, so this is not so serious a limitation as for some other softwares, especially the data base managers, but the SAS requirement that missing data be represented by a decimal in the otherwise blank field will require pre-processing any ASCII files not observing that convention. It would be straightforward to create a program to read the data sets from this project and generate data files in the format of Tetra Tech, 1987.) Details on the data sets themselves, the formatting of the data base, and related processing information are given in a companion report, Ward and Armstrong (1992), which is intended to serve also as a User's Guide to the data.

One of the principles observed in the construction of the Galveston Bay data base was the maintenance of integrity of the original data from individual surveys. That is, in the compilation of data for a given parameter, say nitrogen series, the coded information included identification of the data source, say TWC Statewide Monitoring Program versus Corps of Engineers versus TWDB Coastal Data program, and was input without any modification, including retention of the original units of measurements. While the various data sources were later combined in various ways as a part of different analyses, it is mandatory that the data compilation be capable of separating and identifying, say, nitrogen data from divers agencies, as they may differ in accuracy, methodology and procedure, differences which could become crucial in interpreting apparent trends or in more specialized analyses.

This is one aspect of differentiating the *source* data base from *derivative* data bases. The source data base codifies (in machine format) the original measurements as reported by the originating agency. This data base therefore contains exactly the information in the original: nothing is lost or added. Even an apparently innocuous conversion of measurement units can introduce a distortion. For example, many units carry an implicit level of precision that is modified when converted to another system, such as converting depths in feet to depths in meters. Of course, in adapting the data file to the needs of the project, the source data file may be re-formatted. This might entail re-ordering of the variables, removing unneeded or redundant fields, or re-writing in a more compact format. An excellent example is use of data from the Texas Water Commission Statewide Monitoring Network (SMN) data file. This data was provided as a magnetic tape copy of a printed-page report, therefore containing headers, pagination and blank line fills (a file which contains about 40 million characters!). In our processing, data were extracted from this massive file and used to build up ASCII data bases of selected suites of variables that then functioned as the *source* data files for further analysis. These *source* files contain exactly the measurements in the original master tape, in the original units: only their format of ordering/storage is altered. One specific re-formatting to which all files were subjected was to be ordered in time; the resultant file we refer to as a *primary* file, but it is nothing more than a chronological source file.

For various analytical purposes, however, these data must be modified, for instance converted to common units, averaged in the vertical, aggregated, or screened out according to some criterion. The data set so processed is a *derivative* data base. Any number of derivative data bases can be created according to the needs of a scientific investigation; it is our opinion, however, that the source data base, once established, should remain inviolate and sacrosanct. Thus the basic approach in this project was to first create the source data base for a given parameter through the data compilation effort. Then various derivative bases were formed to selectively include certain subsets subjected to specific processing.

The goal of this data compilation, simply put, is to create a derivative record of time/space/concentration for each water/sediment-quality variable of concern. That is, each data entry must identify a point in space-time at which the measurement was performed and the associated parameter magnitude. (This in

turn introduces a sense of scale, or resolution, dictated by the resolution in the data as well as our conception of space-time variation in water/sediment quality, and underlying the analyses to be performed. This is discussed with respect to resolution and errors below.) Each data record also includes coded information identifying the data source, e.g. TWC SMN, Corps of Engineers, or TWDB Bays and Estuaries Program.

Almost all of the data sets include the time of sampling, at least to some resolution. The point in space is more problematic. Most sampling programs express position by an alphanumeric station name. In order to be able to process the data spatially, this point must be expressed quantitatively. In this project, latitude/longitude coordinates were used to locate the horizontal position of the sample, and depth (i.e., distance below the water surface) to locate the vertical position. The former required precisely plotting the sampling stations from descriptions or from project maps and determining by manual measurement the coordinate positions, which were then keyboarded into a digital data base. Much of this effort had already been carried out for the Data Inventory Project (Ward and Armstrong, 1991), but the new data sets located during this work required station positioning and digitization. In a minority of instances, the data-collecting agency includes latitude/longitude coordinates for the sampling stations (although, as described elsewhere, we have encountered numerous errors, and were forced to plot and re-measure many of these). This station location data is entered into a separate file, and the horizontal coordinates merged with the measurements at a later stage of the processing into the derivative files.

## **4.2 Quality Assurance, Reliability and Uncertainty**

The limits of resolution of measurements and the associated imprecision, and the extent of infection of a data set with errors contribute a degree of uncertainty to each entry in the data record. The obverse concept is the reliability of that data set for scientific analysis. The need for determining the reliability of historical data and discounting measurements that are judged to be "unreliable" is clearly important. This is recognized by EPA and general methods for accomplishing this are outlined by Tetra Tech (1987). Further, this need was identified specifically in the NEP draft Scope of Work for this project. It is the PI's conviction that such judgements must be formulated carefully, and the rejection of data be given close consideration.

In data compilation and processing in this study, a major concern was the detection of errors capable of elimination and the quantification of the residual uncertainty in the data. This includes, but is not restricted to, the procedures commonly referred to as Quality Assurance/Quality Control.

### *4.2.1 Data screening and data-transfer quality assurance*

The NEP primary data bases were compiled from various original data sources, some digitally and some manually, and because a transfer of information is

involved, there is the possibility of error. Therefore, specific measures were introduced to minimize the occurrence of error, and maximize its detection, as follows:

(1) All data available in machine readable form from an originating agency were obtained, manipulated and entered in that form. Further, intermedia transfers were minimized, i.e., copies were sought as ASCII or LOTUS files on floppy discs. (In a few instances, magnetic tape was necessitated due to the size of the data files.)

(2) Data entry by hand employed standardized formats that mimicked the hard-copy sources, and the data entry methods employed standard, simple software, viz. EXCEL, MS WORD, LOTUS or the IBM Personal Editor. The data entries were verified by line-by-line comparison of a hardcopy printout of the entry form with the original data source. (*NB*, screen-versus-original comparison is prone to misinterpretation and fatigue, hence the emphasis on comparison of the two hardcopies.) Following the entry and verification steps, the data were scanned and spot checked personally by one of the PI's.

(3) Each new data set that involved a large file of information (and hence especially prone to errors of fatigue or oversight) was subjected to machine screening to verify that the variables lie within expected ranges and exhibit "natural" variability. When aberrancies were detected, the entries were verified against the original source. In many instances this screening detected apparent blunders in the source file itself. These are discussed separately below and in Ward and Armstrong (1992). Further, additional steps in the data processing process included various error traps and cross checks, which serve as further error-checking.

Early in the project work, an attempt was made to use optical scanners for data transfer. It was quickly learned, by experimentation with several scanners of varying easel dimensions and manufacturer, that the error rate (i.e., character-recognition problems or noise responses) was unacceptably high, even when the source document was clean and high-contrast. Therefore, manual entry had to be pursued.

Particular note should be given the term "mimicked" in (2) above. This is a significant departure of the procedure of this project from that recommended by Tetra Tech (1987), who require that re-formatting into a uniform format, as well as conversion and/or mathematical transformation, be carried out as part of the data entry process. We believe this strategy is seriously flawed. The entry of thousands of numbers by keyboarding personnel demands maximizing efficiency and accuracy. Any differences between the keyboard format and the hard copy are an invitation to misinterpretation and transcription mistakes. Further, since keyboarding personnel are rarely equipped to interpret the numbers they are entering, they should not be expected to carry out calculations of any kind, but to simply input what they see. The Tetra Tech procedure, we believe, reduces efficiency and requires an additional level of oversight that could be totally replaced by machine screening. Moreover, we take exception to the philosophy of

altering the source data, even by units conversion or rounding, as discussed above, and this is precisely what Tetra Tech recommends.

#### 4.2.2 Data screening and data-base quality

The errors introduced by the data transfer procedures of this project were the simplest to deal with, because their existence (i.e., that they were in fact errors of entry) could be confirmed by comparison with the original source, and corrections could be expediently implemented. The same screening process, i.e. testing for values within "reasonable" bounds (discussed below), spatial continuity (as reflected by simultaneous data from different depths or nearby stations) and temporal continuity (comparison with measurements at the same station before and after the sampling time), occasionally detected aberrant values in the source data files themselves. When possible, we contacted the agency source to verify the reported information. For most of the data files, however, there is no longer an authoritative source with which to compare the reported data: the original field sheets are discarded, or the principal investigator or originating agency is not accessible (or even extant). This forced us to make probability judgements. Consonant with our philosophy of leaving the source data files sacrosanct, "corrections" were introduced into these data files only when the typographical error was "patently obvious." For example, in the computer printout of the Bureau of Commercial Fisheries hydrographic data from the 1960's (Pullen and Trent, 1968), in the record for Station 83, we find the following entries:

MO	DA	YR	TIME	SD	TEM	SAL	P	N	O
12	18	63	1540	0	083	249			
12	18	63	1540	2	084	248			
12	18	63	1540	5	083	253			
12	18	63	1540	10	083	260			
12	18	63	1540	15	084	026	5		
12	18	63	1540	20	084	267			
12	18	63	1540	25	080	274			

An unlikely value of salinity, given the other values in this vertical profile, an isolated determination of phosphorus at an unlikely depth, and the likelihood of a data entry clerk accidentally entering an extra space in the salinity column, motivated us to correct the 1540 salinity to 265 and remove the "5" in the P column. (The decimal points are omitted in this format.) This correction is "patently obvious." Errors such as this, obviously misplaced (or omitted) decimal points, ppm entered instead of ppt (or vice versa), dropped or inverted digits in a date where there are other data from the same sampling run to confirm the date, are regarded as "patently obvious," and represent the limit to which we entered corrections into the source data files. If there is any reasonable possibility that the source data could be entered correctly, or if it is probably wrong but we have no logical, near-certain means of supplying the correct value, then the entry was allowed to stand. Clearly, most apparently aberrant values fell into this category. In the process of creating the derivative data bases later, and certainly in the later analyses, there is the opportunity to reject apparently aberrant data, so leaving

such values in the source files causes no harm to the analyses and preserves the integrity of the source data base.

Latitude and longitude coordinates were also subjected to screening. This employed a "range of limits" screen to verify that the positions fell within the latitude-longitude range of Galveston Bay of 29° 00' to 29° 50', 94°30 to 95° 15 (which helped in identifying wildly incorrect points) and a comparison of station descriptions to where the station plotted. In a few instances, enough information was given on the boat tracks during sampling to allow some judgement as to the likelihood of error. Generally, finer corrections were reserved for the derivative data-base screening unless some independent information was available. Errors in the positions determined in this project or the earlier Data Inventory project proved to be rare, due to the procedures of cross-checking and proofing used during these projects. However, the latitude/longitude coordinates provided by some of the agencies exhibited problems. For the TWC SMN data, when SMN station 1014.2700 was discovered to plot on top of St. Thomas Junior High School, it was easy to access the description of station location, plot the station on an accurate map and determine the correct coordinates. For the Parks and Wildlife Department hydrographic data base, only latitude/longitude coordinates were provided by the agency; station location information is kept on file at the field office. In this data set, the "range of limits" screen disclosed 52 erroneous coordinates out of about 12,000 records (one such point plotted due south of the Azores), a very small incidence of error, due to the good quality-controlled and formalized data-entry system of this agency. Positions this much in error were obviously due to incorrect digits in the degrees position. While the TPWD Seabrook Lab staff kindly tracked these down for us and provided the corrected locations, errors in minutes or seconds would not be so easily detected, if detectable at all. This does give an indication of the probable fraction of location errors in this data set due to simple data entry or map-reading errors. (Also in this same data set, we located only one "patently obvious" data entry error, a date given as 1930 instead of 1990. On the other hand there were numerous errors in parameter magnitudes, especially dissolved oxygen, whose correction was not "patently obvious.")

#### *4.2.3 Uncertainty measures and data quality*

The screening procedures outlined in the two preceding sections address data errors of the typographical or "blunder" variety. There remains, of course, a residual error in any set of measurements, deriving from the omnipresent sources of imprecision, inaccuracy and mistakes (including data-entry errors). In this project, data bases for specific variables were created by the combination of data sets from different sources, with differing analytical methodologies, different agency objectives, and differences in field procedures. In order to be able to attach a degree of uncertainty (or its complement, a level of confidence) to such a data set, it is necessary to assess the uncertainty in each of the component data sets, and devise a means of transferring this information to the composite data set. A data user then has the basic information to further determine how the

uncertainty is affected by whatever processing of aggregation, units and proxy transformations, and averaging the data may be subject to.

Clearly, the first step is to define carefully and precisely the formulation of uncertainty, i.e. if we denote the measured value of some parameter as

$$v \pm e,$$

$e$  must be defined unambiguously and its dependence upon the value of  $v$  and other factors carefully specified. Unfortunately, there is terminological chaos in the practice of reporting uncertainty in science and engineering. In the above expression,  $e$  may mean the standard deviation of measurements about their mean, the standard error about the "true" value, the tolerance of measurement, the absolute bounds on the range of  $v$ , or the magnitude of some fixed multiple of standard deviation or standard error. With careful definition, the measures relative to the mean are usually interconvertible (the exception being the absolute bounds on range). The problem is that in the literature " $\pm e$ " frequently appears without any associated clarification. Here we will define how the expression is used in this report, and what meaning we will assume when the data source or reference uses but does not define the expression.

In this report, we will employ the error bound  $e$  to be the magnitude of the population standard deviation ( $\sigma$ ) about a fixed value of the variate. Specifically, for a given measurement procedure under a static set of controls (same concentration, same lab, same personnel, same equipment, same coffee),  $e$  is the standard deviation about the known value of the variate for a theoretically limitless set of replications. For practical purposes, we usually have to estimate  $e$  by the standard deviation about the mean of the measurements under the same idealized limitless conditions. (The distinction is one of accuracy versus precision.) This standard deviation is estimated in practice by a finite set of measurements: if the set is large the estimate is good; if the set is small, the sample standard deviation may have to be corrected to estimate the population standard deviation. For many of the trace organics of current concern in water quality, e.g. the priority pollutants, available precision data may be limited to only 3 or 4 replicates for a given set of controls, so the correction may be substantial. Technically, the correction is the factor  $\sqrt{[N/(N-1)]}$ ; further, other statistical inferences must be altered if the sample standard deviation departs significantly from the population standard deviation. This and related matters are treated in any standard textbook on statistical methods, e.g. Mood (1950), Hamilton (1964).

Standard Methods (APHA, 1971, 1985, 1989) and ASTM Standards (ASTM, 1976, 1980) recommend the use of " $\pm e$ " as a standard deviation. Unfortunately, these proposals for a uniform reporting of precision compete with practice and intuition in the literature. Many authors use " $\pm e$ " to specify, in effect, tolerance limits, i.e., the range in which "most" of the measurements fall. "Most" seems to mean substantially more than 95%. Tolerance specification has traditionally assigned a level to  $e$  of about  $3\sigma$  (e.g., Kennedy and Neville, 1976); exactly  $3\sigma$  implies a 2.7% probability of a measurement with normally-distributed error falling outside the

indicated range, while  $3.09\sigma$  implies exactly a 2% probability of violation. This usage seems to lie much closer to the intuitive connotation of precision expressed as  $\pm e$  than the use of standard deviation, especially among water resources scientists. Another competing concept is the precision latent in the expression of significance. A measurement reported as 5.36, for example, with no additional qualifiers, implies a tolerance (in the above terminology) of  $\pm 0.01$  or no worse than  $\pm 0.02$ : by writing the third digit, the author is indicating relatively strong certainty of its significance (EPA, 1979). (Mathematicians are generally more fastidious, demanding a tolerance  $< 0.5$  times the last significant unit, e.g.  $\pm .005$  in the above example, Scarborough, 1966: then the statement "correct to  $n$  significant figures" means *correct* to  $n$  significant figures. Skougstad et al., 1979, state that the last significant digit is the "first doubtful digit," but the meaning of "doubtful" is not elaborated.) Thus, for a worker known to be scrupulous in the expression of significance, some measure of that worker's judgment of precision can be inferred.

The uncertainty may vary with the magnitude of the measurement, and the dependency may be generalized as

$$e = a + mv \quad \text{for } v_0 < v < v_1 \text{ and } v > v_t$$

Actually,  $e$  may vary nonlinearly with mean value  $v$  of the measured parameter, but for present purposes an at-most-linear variation is sufficient (because the limited data usually available on precision will not support the assignment of a nonlinear variation). This formulation calls explicit attention to a range of applicability of the measurement from  $v_0$  to  $v_1$ . Any analytical method has limits on its range of validity, though for some procedures these limits are so broad relative to the natural range of the variate that they are non-limiting in practice. For a specific parameter, often the constant term  $a$  or the linear variation  $m v$  will dominate the dependency of error  $e$  on variate value  $v$ , and the other can be neglected. In the case of the former, the precision is constant over the range of applicability, and may be expressed simply as a constant value with the units of  $v$ . In the case of the latter,  $e$  may be conveniently stated as a fraction (a percentage) of  $v$ . (The suggestion appearing in recent editions of *Standard Methods* to report standard deviation as a percentage of the mean is unfortunate, in that it suggests that  $e$  varies directly with  $v$ , when in fact it may not.) Frequently there is inadequate data to determine which, if either, dominates. Sometimes, both may be important. For example, in Skougstad et al. (1979), the analysis for sediment boron is stated to have a precision (as a standard deviation) of about 7 mg/kg at the lower end of the range of applicability at 10 mg/kg, and about 50 mg/kg at the upper end of the range at 250 mg/kg. Substitution of these values in the above equation yields:  $a \approx 5$  mg/kg and  $m \approx 20\%$ , so that the total precision is

$$e = 5 + 0.20v \quad \text{mg/kg}$$

It is of course even better if there are multiple values of  $e$  for a range of values of  $v$ , whereupon a regression line can be estimated, and the best-fit values of  $a$  and  $m$  determined statistically. This is the format used in the most recent USGS

manual (Fishman and Friedman, 1989) for dissolved analyses (see also Friedman and Erdmann, 1982), and when data warrant in the ASTM Annual.

The final element in the above formulation is the threshold value  $v_t$ , which  $v$  must exceed for the analysis to be meaningful. Again, such a threshold value always exists, due perhaps to mechanical friction in a gauge or the limits of resolution of a probe, but it may be much smaller than the lowest value of  $v$  encountered, or be much smaller than  $e$  for  $v \approx 0$ , and thus be practically negligible. For trace concentration determinations based upon gas chromatography, however, the threshold is a singularly important element of the procedure, establishing the detection limits of the analysis.

In order to completely characterize a measurement, we must include an estimate of the uncertainty, including any limiting values, such as the detection limit. This was approached in this study in several ways depending upon the extent of documentation for the data set, in decreasing order of preference:

- (a) review of QA/QC procedures observed by the collecting agency, as reflected in practices memos, manuals and directives,
- (b) identification of the specific methodologies used and their established accuracy,
- (c) statistical variation of the measurements themselves, relative to some external standard, e.g. a more accurate proxy relation or data from a contemporary, independent source.
- (d) judgement, based upon experience with the method or equipment, and upon the practice of workers in the field using that methodology, as inferred from their explicit or implicit uncertainty statements.

This first task was to document the different agency procedures and their implications for precision and accuracy. For recent data with well-established procedures and QA/QC protocols, this was generally straightforward, though many agencies have no written descriptions and our information had to be obtained from personal communications. For older data, the methodologies and probable care of the observers must be judged (following the above procedures). Where possible, measurements of related parameters from the same program or measurements of the same (or related) parameter by more than one agency were cross-correlated to detect systematic differences. Unfortunately, the general sparsity of data in space and time frequently prohibited this kind of test, but for some variables such as salinity, the data were sufficiently dense to allow it (see Section 3.2 above). In some instances, we were forced to judge fairly low levels of accuracy (i.e., broad confidence limits).

The best published sources of precision data for specific analytical methods are *Standard Methods*, the ASTM annuals, and the USGS Techniques of Water-Resources Investigations. Generally, there is more information—and more quantitative scope—on precision in the later editions than the earlier, which

raises a dilemma: when precision information changes, should we utilize the data contemporaneous with the measurements, i.e. assumed to be reflective of the technology and procedures of the time, or should we presume that the more recent data derives from a larger base of measurements, and represents an improved estimate of precision applicable to the older techniques as well? Considering that the reported precision for many trace metals and organics is *lower* (i.e., greater standard deviations) in more recent publications (e.g., Fishman and Friedman, 1989) than in the older (e.g., Skougstad et al., 1979), this is not a merely pedantic concern. No doubt there are elements of truth in either alternative, but we have elected the former. This is not an irreversible decision, as any later user of the data base has the option of employing a different measure of precision, and consequently a different data rejection procedure. (Of course, if one does not use the standard deviation as a basis for data rejection, then the issue of the source of precision information becomes irrelevant to the analysis.)

Also, we note that the precision data available is generally much more complete and accurate for the water-phase analyses than the sediment. Indeed, in the USGS manuals (Wershaw et al, 1987, Fishman and Friedman, 1989), for each of the bottom-material analyses there is simply the statement: "It is estimated that the percent relative standard deviation for [parameter name] in bottom material will be greater than that reported for dissolved [parameter name]." When precision data are presented for water-suspended sediment mixtures, we have used that preferentially over the dissolved data to estimate uncertainty for the sediment analysis.

A separate concern in data processing is the handling of anomalous values lying well beyond the expected range of the variate. Most of these are the result of human error at some point in the process from laboratory or field measurement to entry into the data base. A frequent manifestation is a decimal point mislocation, resulting in multiplying the true value by one or several orders-of-magnitude. (See the aberrant point identified in the lower panel of Fig. 3-9.) A screening rule can be formulated to reject such points. The problem is how to assign a rejection trigger so as to exclude points certainly in error, but not to exclude points that happen to deviate widely from "normal" values, since such deviations may in fact be real and therefore significant.

It has become traditional in data processing to differentiate between values that are so extreme as to be rejected as "unlikely" (including "impossible") and those that are "unusual" but within the realm of possibility, see, e.g., Bewers et al. (1975). This is the approach recommended by Tetra Tech (1987) who provide "A" and "B" values for an extensive list of estuarine variables, corresponding respectively to "unusual" and "unlikely." It must be noted that the normal strategy is to use these limits to identify anomalous points *during the data analysis and entry process*, to provide feedback to the originators of the data for verification and correction. In our present study, there is no prospect of tracing back to the originator of the data (except for verifying data entry performed during this project), so we need to determine a criterion for data rejection. We also note that any such rejection trigger would be applied at the earliest to the compilation

of the Derivative Data Files, not to the source data (except, of course, for the "patently obvious" category described earlier).

Generally, as a matter of personal philosophy, we reject very little data in the formulation of the Derivative Files, and reserve further data screening for the specific analyses to which the Derivative Data Bases are subject. Data were rejected if the date, position, or depth were obviously impossible and there were no satisfactory means of judging the correct value. Generally, we did not reject data at this stage based on the parameter value, but reserved that for later steps in the analysis. Rejection triggers were assigned to many (not all) of the variables based upon the suggestions of Tetra Tech (1987) or on judgement of the PI's. These are given in Tables A-1 and A-2 of the Appendix. Therefore, the summary of data in Tables 3-7 *et seq.*, which are unscreened, reflect some probably-incorrect values. For example, the maximum sediment volatile solids value of 98% (Table 3-8), though lying within the range of *possibility*, lies outside the range of *probability*. (Surely there would have been reports of lab technicians lacking eyebrows.) The same remark can be addressed to the 19% zinc and the 9.4% iron. Below this, though, the demarcation between the two is less certain: the 3.2% value for oil & grease or the 0.6% DDT may be more unlikely than unusual, but we are hesitant to dismiss them on strictly an *a priori* basis.

Tables A-1 and A-2, in the appendix summarize the measures of uncertainty and rejection criteria assigned in this study. These uncertainty criteria were based upon available information on precision of various methodologies and procedures for different parameters, summarized in Table A-6 of the Appendix, suitably rounded and supplemented by estimates of accuracy from analysis of data from Galveston Bay when available, by precision data from similar compounds if primary data were not available (e.g., total DDT estimated from precision data for p,p'-DDT), and judgement of the Principal Investigators when no solid information was available. The rejection triggers were assigned as a combination of Tetra Tech (1987) and judgement calls by the PI's. (We note that many of the values in the Tetra Tech report are inapplicable to Galveston Bay because they are either patently obvious, e.g., no concentration greater than 100%, or inappropriate, e.g. a temperature limit of 30°C.) Both the uncertainty and the rejection triggers are provided more as guidance to the future users of these data sets than as absolute bounds on data inclusion, and reflect as much our judgement of the quality of the different data programs as statistical constructs.

Data rejection can be performed based upon either the level of uncertainty of the measurement or its magnitude relative to the rejection trigger (when one is provided). Each measurement in the Derivative Data Base is accompanied by the specified level of confidence, transformed into units of the variable and scaled (when appropriate) to the magnitude of the measurement. Thereafter, any data processing can be preceded by an assignment of acceptable accuracy of measurement; any measurements failing this level would be excluded from that analysis. But these measurements would still be retained in the data base. We believe this to be a superior approach to merely deleting data, especially older data, by a sharply defined criterion of "reliability". This is closely related to the notion of preservation of data integrity discussed above.

### 4.3 Data Set Processing

The principal steps in data-processing in this study were:

- (1) For each historical data program in the Galveston Bay system, compile a Primary Data File, consisting of the digital record of measurements ordered chronologically;
- (2) For each parameter of concern, sift through the Primary Data Files, applying whatever screening, proxy relationships, and units conversions are necessary, to create a Master Derivative File for that parameter;
- (3) Sort the Master Derivative Files into geographic segments for the Galveston Bay system;

At this point, for each parameter there would be a chronological record of measurements of that parameter for each segment of the bay, which can be subjected to various statistical analyses to expose time-space variations. The general processing procedures are shown schematically in the flow charts of Figs. 4-1 through 4-3.

The construction of the Primary Data File was described in Section 4.1 above. Because the ultimate product is to be a chronological Derivative File, and the process of chronologizing a file can be resource-intensive, it was decided to chronologize the data records as early in the process as feasible, then to design subsequent data handling in such a way that ordering is preserved. This is the principal difference between the Primary File and the Source File, the digital record in the format and units of the agency that obtained the data, i.e. the Primary File contains exactly the same data records except ordered chronologically. A secondary difference is that the first tier of the Q/A screening is applied in this process, see Fig. 4-1, so that the Primary File will have entry errors and "patently obvious" data errors corrected or deleted.

The creation of the Derivative Data Files is fundamentally a matter of merging information from various files and re-formatting the product. The various steps in this procedure are shown in Fig. 4-2. The sampling station latitude/longitude coordinates are collected in a separate file, and accessed according to the agency station designations to merge the coordinates with the data taken at that station. At this stage, all units conversions are applied, as well as any proxy relationships by which one parameter may be transformed into another. Because we anticipate analyzing data on a time scale of days to weeks, the information on clock time (i.e., time of day) of each sample is not carried through to the Derivative Data Files, but the full date is retained. In addition to the parameter value itself, the uncertainty is estimated and included in the data record.

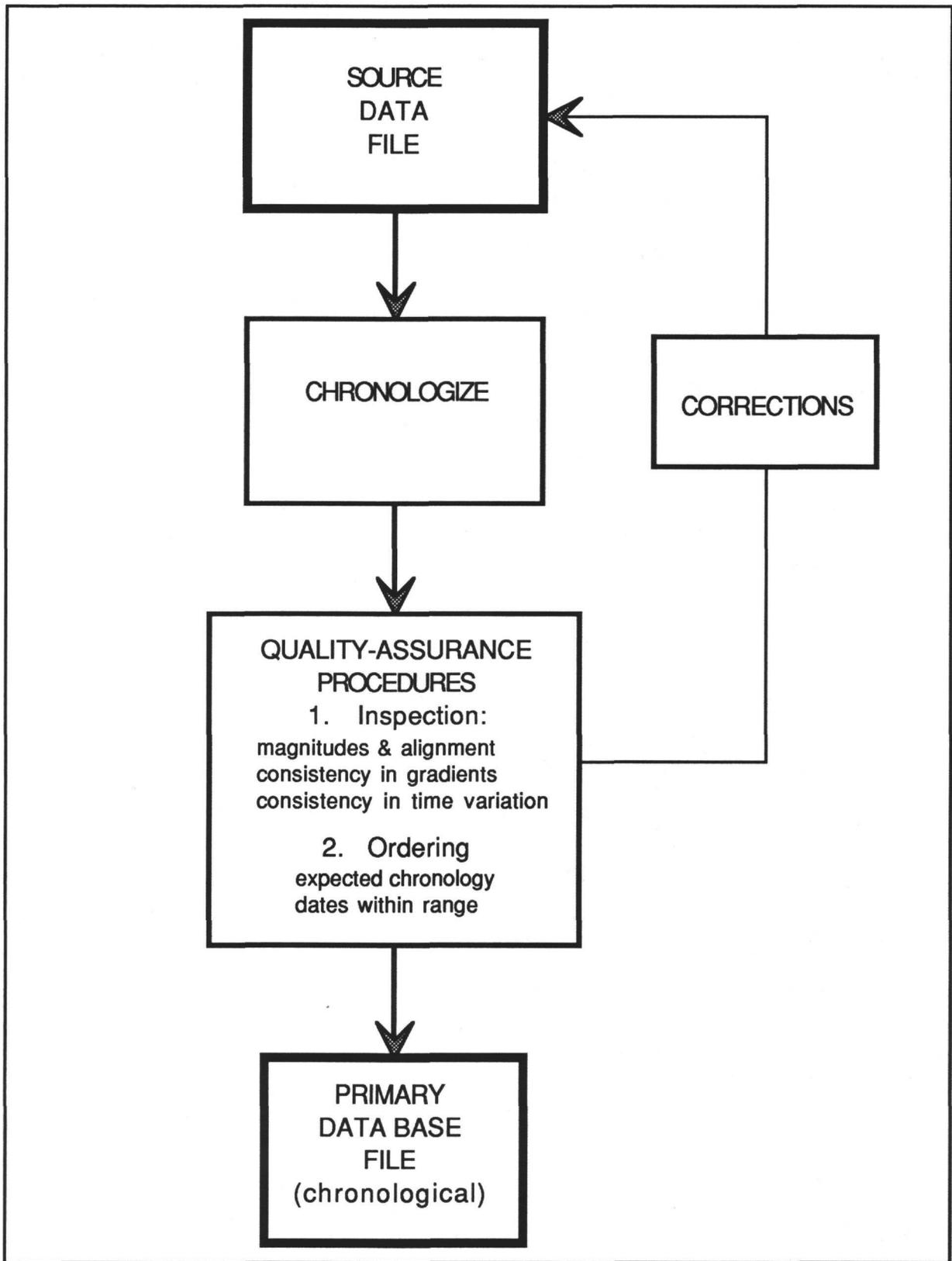


Fig. 4-1. Procedure for pre-processing and generation of primary data base

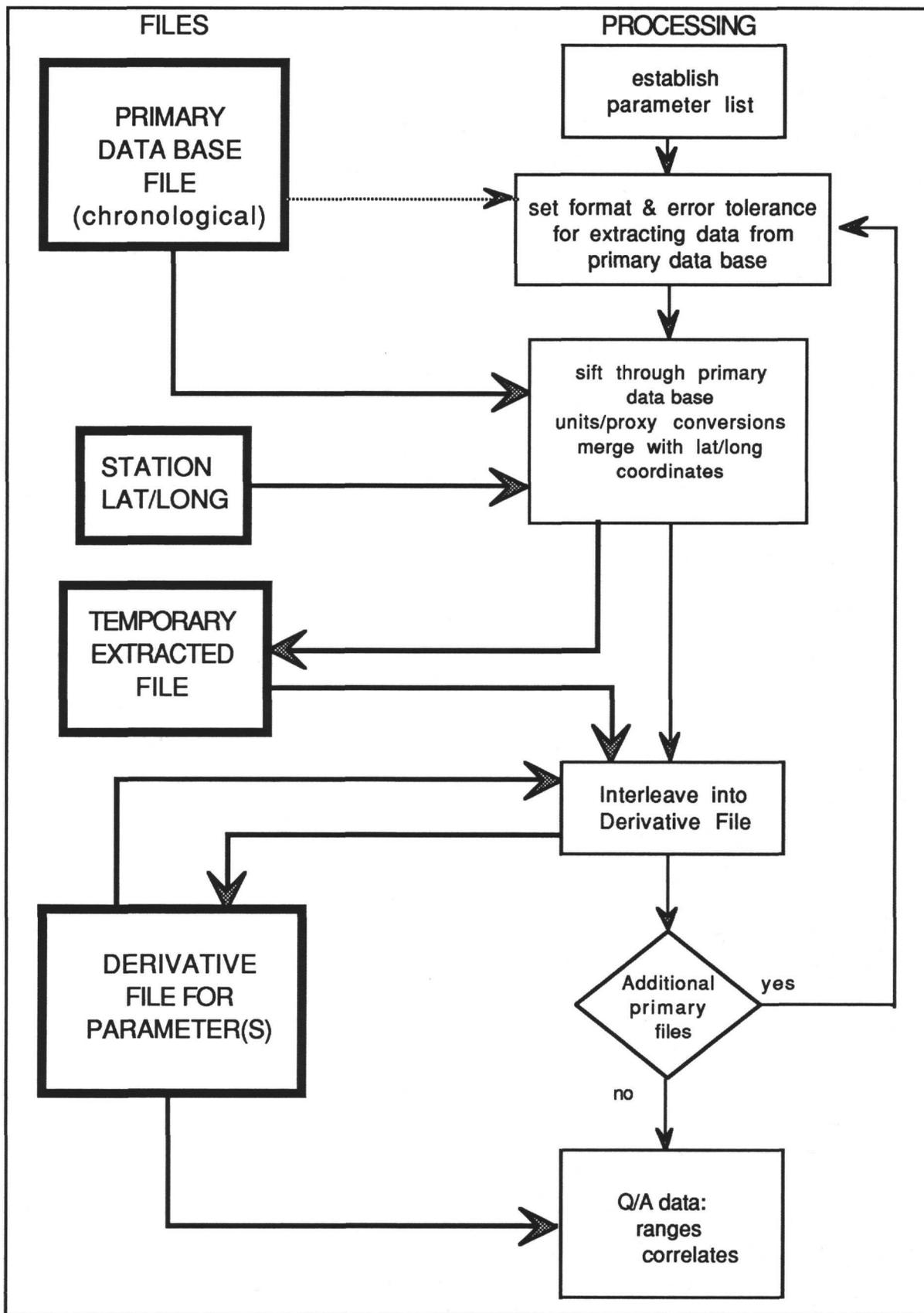


Fig. 4-2. Procedure for creation of derivative data bases



The format of each record in the Derivative Data Files is:

DATE LATITUDE LONGITUDE DEPTH MEASRMT UNCERNTY PRJ

where DATE, LATITUDE AND LONGITUDE are 6-digit fields (YRMODA and the latitude/longitude coordinates are degrees/minutes/seconds), the sample depth is in meters, MEASRMT is the measured value of the parameter (retaining four significant figures), UNCERNTY is the uncertainty as a standard deviation following the convention of Section 4.2.3 above (to three significant digits), and PRJ is a 3-digit integer flag that identifies the agency or project that was the source of the measurement (see Table 3.6). Thus, each record of the Derivative Data File represents a point in time (to resolution of a day) and space (horizontal and vertical position), together with the measurement and its uncertainty. Each such record requires 50 bytes of storage.

Throughout this process there are numerous error traps and cross-checks, not only to ensure that the data is not corrupted by a bug in the processing but also to detect entry errors or aberrancies in the data as reported by the agency. The temporary extracted file shown in Fig. 4-2 contains the data in the above format from a single source. This file is examined closely for errors or anomalies before it is interleaved into the Master Derivative File. Once the Derivative File is created, it can be subjected to various screenings and data rejection, according to the preferences of the researcher.

We regard the Master Derivative Data files to be our principal data resource product from this study. These contain all of the data for each parameter that we were able to locate and digitize, and incorporate our judgment on which data should be retained or rejected. In order to address the concerns of characterizing the ambient quality of Galveston Bay and its historical trends, these data bases are subjected to additional processing, as indicated in Fig. 4-3. Principally, this involves further filtering of the data and sorting the data into the separate geographical segments. Because the source data files can contain duplicate measurements, e.g. the TWDB Coastal Data System may contain TWC measurements that are also in the Statewide Monitoring Network system, some research projects may share the same data files, keyboarding personnel may have inadvertently duplicated entries, etc., there is the possibility that duplicate measurements may be present in the file. Therefore, there is a preliminary screening step to detect such duplicates. (This is repeated after the data are sorted by segment to detect "near-duplicates.") "Vertical processing" in Fig. 4-3 refers to selection of data from only one depth (or range of depths) or to averaging data in the vertical.

Finally, the defining quadrilaterals for a system of segmentation are applied to the data to sort into the various segments. For this project, two segmentation systems were employed, as discussed in Chapter 2. The first is the Texas Water Commission water quality segments, which are defined and their associated geographical quadrilaterals given in Tables A-3 and A-4, resp., in the Appendix. The second is the system of hydrographic segments shown in Figs. 2-5 through 2-7, and defined by the quadrilaterals of Table A-5 in the Appendix. As some of

these segments are represented by the union of two or more quadrilaterals, once the initial sorting is completed, these quadrilaterals must be consolidated into a single data record for that segment, hence the process of "consolidation" in Fig. 4-3. At this point, the data files are in a form suitable for statistical analysis.

page 166  
Blank