A Review of Data Analysis Techniques for Application in Automated Quantitative Accuracy Assessments

Robert Kunz and John Mahaffy
Applied Research Laboratory
The Pennsylvania State University
P.O. Box 30
University Park, PA, 16804
rfk102@psu.edu, jhm@psu.edu

Keywords: code assessment, uncertainty analysis, validation, verification, best estimate

ABSTRACT

A survey of data conditioning and analysis techniques has been undertaken which focuses on their relevance to nuclear reactor systems (NRS) code accuracy and uncertainty assessment. A number of methods are considered for their applicability to the automated assessment of the accuracy of NRS code simulations, through direct comparisons with experimental measurements or other simulations. A variety of data types and computational modeling methods are considered from a spectrum of mathematical and engineering disciplines. The goal of the survey is to identify issues and techniques to be considered in the development of an automated code assessment procedure, to be used in United States Nuclear Regulatory Commission (USNRC) advanced T/H code consolidation efforts.

The motivation for the present work is first provided by background discussion that summarizes the relevance of this subject matter to the nuclear reactor industry. Next, the spectrum of NRS data types are classified into categories, in order to provide a basis for assessing individual comparison methods. The main body of the paper then summarizes the survey undertaken. There, each of the relevant issues and techniques considered are addressed. Several of the methods have been coded and/or applied to relevant NRS code-data comparisons. Results of these demonstration calculations are included.

1. INTRODUCTION

In recent years, the commercial nuclear reactor industry has focused significant attention on nuclear reactor systems (NRS) code accuracy and uncertainty issues. To date, a large amount of work has been carried out worldwide in this area, with significant involvement by the USNRC. Recently, the USNRC has sponsored the present authors to:

1. Survey available data conditioning and analysis techniques, focusing on their appropriateness in NRS code accuracy and uncertainty assessment
2. Develop software to deploy recommended techniques
3. Develop coding and interfaces for the software to enable automated assessment on a large number of data sets so as to facilitate code update efforts and modeling revalidation

The present paper documents the first of these efforts. As an outcome of effort 2, the
authors, and our colleagues, have developed a software platform, designated the Automated Code Assessment Program (ACAP), which has been delivered to the USNRC. An overview of the design and operation of ACAP is provided in Kunz et. al. (1998a, 1998b), along with several NRS code application examples which illustrate the software’s capabilities. More recently, effort 3 has resulted in the development of a spreadsheet-based batch capability for executing ACAP on a large number of data sets. This enables its use in rapidly providing an automated quantitative assessment of the change in the quality of a thermal-hydraulic analysis code from version to version. This aspect of our work is documented in Kunz et. al. (2000).

2. SYSTEMS CODE ACCURACY ASSESSMENT ISSUES

The computational fluid dynamics (CFD) field has matured to where production computer codes have now become well established at a large number of industrial design agencies. This status has been engendered by decades of vigorous research on numerical methods and physical modeling, and by rapid advances in computer hardware. The increased usefulness of CFD methods in design environments, has been characterized by more reliance on computer codes to produce accurate quantitative results, an evolution away from the traditional parametric assessment role that CFD held until recently.

As the need for, and ability to perform, higher accuracy simulations have increased, the CFD community has in turn increased its emphasis on assessing the quality of results. Analysis codes are no longer so much judged by their baseline algorithmic and physical modeling characteristics as by their reliability, accuracy and usability characteristics. CFD code developers and users now focus a great deal of their efforts on code verification and validation (V&V). Professional society journals which publish CFD analyses have all put into place standardized policies which ensure that a minimum level of V&V has been ascribed by authors to the published simulations (AIAA [1994], Gresho and Taylor [1994], Freitas [1993]). CFD reliability and usability topics themselves began to appear in the literature in the mid 1980s (Kmetyk et al. [1985], Bessette and Odar [1986], Roache, et al. [1986], for example), but far more widely so in the past seven years (Celik [1993], Coleman [1996, 1997], Huang [1995], Marvin [1995], Mehta [1995], Oberkampf [1995], Roache [1997], for example). Indeed, for several years, entire sessions have been devoted to this area in the AIAA’s annual Fluid Dynamics conference. NASA has recently sponsored an industrial consortium to investigate and formalize procedures for CFD code certification in the aerospace industry (Melnik et al. [1994, 1995]). Additionally, several national and international engineering organizations have established formal guidelines and/or standards for the V&V of CFD (and other analysis) codes (ANS [1987], IAHR [1994]).

The commercial nuclear reactor industry came to focus on code reliability issues significantly (perhaps a decade) earlier than other CFD design industries (e.g., aerospace, automotive, process). There are several reasons for this including the inherent safety (and concomitant code reliability and licensing) concerns associated with nuclear reactors, and fundamental differences between the CFD methods used in reactor systems codes and “standard” CFD methods. (In particular, nuclear reactor systems codes are primarily utilized for transient one-dimensional and quasi-one-dimensional analysis of complex reactor plant systems, whereas standard CFD analyses are characterized by reasonably well grid-resolved, usually steady-state solution of individual thermal hydraulic components).

Over a decade ago, the USNRC initiated an international effort to improve and standardize
the assessment of thermal hydraulic (TH) systems codes (Kmetyk et al. [1985], Bessett and Odar [1986], for example). Prior to that, the assessment of the performance of TH codes had been largely qualitative and subjective, and thereby difficult to use in plant safety certification. In 1984, the USNRC organized the International Thermal Hydraulic Code Assessment and Applications Program (ICAP), a major goal of which was the assessment of TH codes using relevant data from a wide range of international experimental facilities. Since that time, a large amount of research has been carried out worldwide in this area. As the USNRC has moved towards plant certification based on best estimate methodology, the establishment of V&V guidelines that incorporate quantitative accuracy and uncertainty measures has become even more important.

The issues associated with NRS code accuracy and uncertainty assessment are numerous and complex. They include:

1. **Scalability issues:**
   - In the design environments for virtually all fluid-thermal systems, there is a paucity of experimental data taken in full scale hardware, so most code accuracy assessments are made against scaled test facility data. This is true for NR plants as well.
   - Most NRS data is acquired in scaled facilities. Although the common theories of scaling are applied when designing the facilities, the facilities may not provide full dynamic similarity with a full scale reactor. In general, trends and the timing of key events in a NRS transient do scale well but in an effort to determine scaling effects, multiple scales are often utilized to generate the data used in assessment. These scaling complications have been addressed by D’Auria and his coworkers’ UMEA methodology (D’Auria et al. [1995b]) and other related methods (Bovalini and D’Auria [1992], D’Auria et al. [1995a]).

2. **Discretization and Model Setup Issues:**
   - NRS codes require nodalization of individual TH components. These discretizations are rarely grid converged in the conventional CFD sense due to the lumped parameter, quasi-1D modeling invoked in these codes. Indeed, it has been widely observed that significant differences in predictions can arise when different nodalizations are applied (Aksan et al. [1992], for example), and this has given rise to the growing practice of “qualifying a nodalization” against steady state data (Bonucelli et al. [1993]). (Such qualification establishes confidence in the appropriateness of a given nodalization as it is applied to full scale simulations as well.)
   - The selection of an appropriate computational time step and duration of a simulation affect the accuracy of a NRS code simulation.
   - The specification of boundary conditions in an NRS transient can introduce uncertainty since often a “best” value is difficult to define, and an experienced user may plausibly select from a range of values. Different choices from within this range can yield significantly different predictions of key parameters (D’Auria and Galassi [1997]). Boundary condition uncertainty must be assessed through parametric code runs.

3. **User Issues:**
   - The NRS code user can introduce uncertainty into a simulation, as evidenced by the widely shared observation that different users can easily produce different results using the same code applied to the same transient. Contributing to these differences are: the large number of available physical model specification input options, varying nodalization, time step and boundary condition selection (discussed above), and input errors (Aksan et al [1992], D’Auria et al. [1990b]).

4. **Software Reliability**
- The possibility of the presence of software/code errors including typographical or logical errors in a NRS code (especially a recently upgraded code being requalified) can introduce uncertainty into analyses, as can:
  - The possibility of compiler errors
  - Roundoff errors
  - Machine dependency
    Mueller et al. (1982) provide assessment of the role of these “operational” uncertainties in NRS codes.

5. Best estimate vs. conservative criteria:
   - In the last decade, the USNRC has begun to move towards basing licensing decisions on “best estimate” NRS code analysis. This contrasts with the historical approach of using models which conform to conservative requirements (spelled out in Appendix K of 10CFR50 [1997]). This move engendered “quantification of uncertainty” requirements on best estimate calculations being used for licensing purposes, as embodied within the Code Scaling Applicability and Uncertainty (CSAU) methodology and related approaches.

6. Key parameter selection:
   - The assessment of a simulation which models the complete physics of a NR transient can only be assessed if a prioritization is given to some parameters over others. Guidelines have been established (Kmetyk et al. [1985]) to identify the “key parameters” for a particular transient and particular reactor design. As a result, the code has to be assessed against each of the different sets of these key parameters for each of the identified transients for each of the reactor designs.
   - Often NRS transients are characterized by multiple time ranges, each associated with quite different dominant physical mechanisms. Accuracy assessment must accommodate these since certain key parameters are only relevant in certain of these “time windows”. Unambiguous and generally applicable specification of these time windows is also difficult.

7. Richness of Data:
   - As detailed in the next section, a wide variety of NRS data types are encountered including: single value key parameters, timing of events tables, scatter plots, 1-D (in space) steady state data, and time record data.
   - The latter of these are themselves characterized by a rich array of features.
     This variety of relevant data types complicates accuracy evaluation and broadens the scope of automated code assessment procedures.

8. Inconsistency of comparison quantities
   - There is, in general, not a one-to-one correspondence between available experimental and computed data. In particular, the same key parameters are not all measured in any given test program.
   - There is, in general, not a one-to-one correspondence between measured and computed time and space coordinates. This can be due to stability limitations of the NRS code and/or nodalization choices. Interpolation may then be required for direct comparison of data and analysis which itself introduces uncertainty into the comparison.

9. Subjectivity of analysis – experimental comparison
   - Recently, the USNRC has used qualitative code-experimental comparison measures such as “excellent, reasonable, minimal and insufficient”. These are well defined (Damerell and Simons [1993], Schultz [1993]). These measures allow a group of
experts to study a set of results and produce some meaningful statement on the value of a code. The process is useful for major releases on a code, but is time consuming, especially when it must be applied to large test matrices.

- Eliminating the inherent subjectivity of this current process is important in the USNRC code consolidation effort. In particular, doing so would allow code upgrades to be rapidly reassessed and would allow for quantitative tracking of improvements in the code’s capability.

10. Uncertainty in experimental measurements
- Several investigators (Bessette and Odar [1986], Coleman and Stern [1997]) have argued that experimental uncertainty must be considered in code-data comparisons, since simulation performance measures can be misleading when comparisons are made directly to reported measured values. Significant experimental uncertainty should be incorporated in code-data assessments to lessen the magnitudes of such difference measures.

11. Larger test matrices
- In the past, NRS code accuracy problems have been corrected in ways which have adversely impacted the comparisons of other untested transients. This has led the USNRC to introduce much larger test matrices.
- This, of course, translates to a significant increase of code reassessment work in a development environment, and therefore itself motivates an automated code assessment process.

12. Lack of suite of assessment tools.
- A standardized suite of automated code assessment tools is not currently available for NRS code-data or code-code comparisons.

These issues collectively motivate the need for automated code assessment in the USNRC's code consolidation effort, as well as in future V&V and licensing application environments. Ideally, in the future, when NRS code users are involved in licensing calculations of “real” plant transients, a single post-processor would be deployed. Based on all uncertainties involved, this post-processor would return, at a given confidence level, the maximum expected deviation between code and reactor of several key parameters (Wilson et al. [1985]). The methodology embodied in this “ideal” post-processor must address each of the uncertainty components summarized above. The need for such a methodology, has motivated a vast amount of research in the past decade (see D’Auria et al. [1995c] for a review of much of this work). Some progress has been made in all of these areas, however, reliable and general tools to quantify NRS code accuracy are not available today. An important contribution to meeting this ideal would be a universally available assessment tool for the users of NRS codes to post-process results in a way that would return quantitative accuracy measures of code-data comparisons. Such a tool would only address some of the uncertainties in real plant analysis. However, it would be part of a process which validates a code with scaled facility data, contributing an important component to total uncertainty in full scale plant simulations. As the USNRC pursues consolidation and advancement of a single NRS code, the need for such tools has never been greater, since such a tool would also greatly streamline revalidation against test matrix data.

It has been the overall goal of this research to initiate a software framework to automatically assess several of the NRS code uncertainty issues summarized above. In particular, a software package has been developed to objectively and quantitatively compare NRS simulations with data. This package, designated the Automated Code Assessment Program (ACAP) is described in detail in Kunz et. al. [1998a, 1998b, 2000]. Consistent with the observations made above, the code has been designed to:
• Tie into data bases of USNRC test data and code results
• Draw upon a mathematical toolkit to quantitatively compare user specified data and analysis suites
• Return unambiguous quantitative figures-of-merit associated with individual and suite comparisons
• Incorporate experimental uncertainty in the assessment
• Accommodate the multiple data types encountered in NRS environments
• Reduce subjectivity of comparisons arising from the “event windowing” process
• Provide a framework for automated, tunable weighting of key parameters in the construction of figures-of-merit for a given test and in the construction of overall figures-of-merit from component code-data comparison measures
• Accommodate inconsistencies between measured and computed independent variables (i.e. different time steps)

So the ACAP development program addresses issues 6-12 summarized above. The scope of this project therefore does not attempt to quantify the uncertainties introduced by user training issues, discretization issues or code operational issues. Nor does the present work address quantification of uncertainty associated with physical models being used on a best estimate basis, nor on scaling uncertainties. However, the present investigators feel that with modest modifications ACAP could be applied parametrically to complement uncertainty assessment in each of these other assessment areas.

In summary, our fundamental goal has been to develop a numerical toolkit to analyze discrete computational and experimental NR systems data, and, in particular, to use these data analysis procedures to develop code-data and code-code comparison measures. Discrete data analysis is, of course, an important element in a wide array of technical disciplines. Indeed, data analysis methods are important anywhere experimental data are used. Techniques to analyze data samples or records lie within the scope of the three overlapping fields: probability and statistics, approximation theory, and time-series analysis. Accordingly, much of the information on this subject is embodied in the mathematics literature. Also, the needs of several engineering and scientific communities have motivated the development of data analysis techniques, which although falling within the three general categories mentioned, are characterized by unique or extended features of relevance to the present research. In particular, methods developed in atmospheric/geologic sciences, economic forecasting, aerodynamic stability, demographics, digital signal processing, pattern (i.e., speech, optical, character) recognition and other fields have relevance to the analysis of NR systems data. Many of these methods, which are also surveyed here, are directly applicable or could be adapted to construct systems code-data or code-code comparison measures.

3. CATEGORIZATION OF NUCLEAR REACTOR SYSTEMS DATA

NRS data types are classified here into five categories, in order to provide a basis for assessing individual comparison methods. Specifically, scaled NR facilities are instrumented to provide a fairly wide array of key parameter and other data. These include:

I. Key parameters tables (Figure 1a).

II. Timing of events tables (Figure 1b)*.

* These data can be considered a subset of NRS data class I.
III. Scatter plots of nominally 0-D data (Figure 1c)†.

IV. 1-D (in space) steady state data (Figure 1d).

V. Time record data (Figure 1e).

Each of these data types is potentially important in any particular NRS code analysis, and thereby must be considered in automated code assessment procedures. Experimental uncertainty bounds are often available for NRS data (see Figures 1c – 1e). The emphasis of this work has been on the latter three. In particular, general comparison measures for single valued key parameters and timing of events tables can be straightforwardly introduced into an automated code assessment system. For this reason, simple techniques to do this are not considered in this review. Somewhat more sophisticated mathematical techniques are required for analysis of data Types III and IV, and data Type V in particular provides a significant challenge for several reasons:

1. The ubiquitous appearance and relevance of these transient data in NR systems
2. The typically long record (often $O(10^5)$ time steps) nature of these data, complicated significantly by their non-stationarity and diversity in characteristic features (long time scale damping, local quasi-periodicity, sudden changes due to active or passive phenomena, chatter (often of high amplitude), dependent variable limits (for volume fraction) between 0 and 1)
3. The fairly significant differences that often appear between computed and measured time trace data (see Figure 1e for example)

The focus of this survey is on methods applicable to Type V data, which include, as a subset, statistical and approximation methods that can be brought to bear on data Types III and IV as well.

In order to facilitate the discussion of the data analysis methods below, some nomenclature definition is appropriate. Random data can be defined as data which, in the absence of measurement error, will be unique for each observation. Nearly all experimental data satisfy this definition of randomness. Experimental NRS transient data is random data since any time a given facility is run, the response of the system will not be exactly the same (non-deterministic). Experimental NRS transient data is also non-stationary since generally, the measured parameter cannot be described as having a constant mean or autocorrelation function, that is, adjacent sections of the time trace will have different statistical measures.

It is not practical to repeat experimental transients enough times to generate a statistically significant ensemble. For this reason, there are not many practical techniques available to analyze non-stationary Type V data (Bendat and Piersol [1986]), though some which do exist are reviewed below. This paucity of analysis techniques contrasts with the wide range of powerful tools available to analyze stationary random data. Fortunately, many of these techniques may be applied to non-stationary data with some loss of rigor, or through some “pre-processing” of the non-stationary records (to render the data globally or locally closer to stationary), or both.

NRS code data, interestingly, cannot be viewed as random data at all. In particular, multiple runs of a NRS simulation will return identical results each time. However, one can conceptualize performing multiple runs of a NRS code using varying boundary conditions well within the uncertainty bounds that these boundary conditions are known. These runs would produce an ensemble of time records. One can view an available record as a representative of this ensemble in the same fashion that the experimental data is assumed (by necessity) representative of an ensemble, were it available. So hereafter, we consider both experimental and computed NRS data, and the difference between them (hereafter the absolute error) as non-stationary random data.

†. Often these data are rendered “0-D” by collapsing data obtained at multiple space-time coordinates to a single scatter plot.
Distinction is also drawn between dependent (measured physical) variables and independent (space-time coordinate) variables in NRS data. In NRS Types IV and V data, the uncertainty associated with the independent variables is much smaller than that associated with the dependent variables. This limits what kind of data modeling approximations that are appropriate (Press et. al., [1994]), and simplifies the consideration of experimental uncertainty (Coleman and Stern [1997]).

4. SUMMARY OF SURVEY

Data Analysis Methods

The data analysis methods surveyed herein, are classified into three broad categories:
1. Approximation theory based methods
2. Time series data analysis methods
3. Basic statistical analysis methods

A number of other methods are treated in the “Other Methods” section below.

The primary distinction between these categories of methods is the nature of the data to which they are applicable. These classes of methods are discussed here. For each, a brief overview of member techniques is provided. Several of these techniques have been adapted to NRS code-data comparison by other workers, and that literature is summarized. Discussion of the applicability of all reviewed techniques to NRS code assessment is provided. Several of the techniques are demonstrated through application to sample NRS code-data sets.

Approximation Theory Based Methods

Approximation theory encompasses mathematical techniques which provide useful (i.e. simple in some sense) functional approximations to discrete or continuous data. The three principal applications of approximation theory are data analysis (discrete approximation), functional representation (continuous approximation) and asymptotic theory. Discrete data analysis is of relevance here. Approximation theory techniques for discrete data could be useful as quantitative comparison measures for NRS data since they approximate discrete random data using deterministic functions. The parameters (i.e. coefficients) defining the functions that approximate the data and the code results can be compared directly. Alternatively, figures-of-merit could be constructed using the parameters defining an approximation to the absolute error (i.e. its proximity to zero quantified in some way). These approaches are illustrated below.

The fundamental approximation problem for discrete data can be stated: Given a set of m data points \((f_i(x_i), i=1, m)\), find an analytical functional representation whose exact form (i.e., component magnitudes) is determined by minimizing in some sense the differences between this functional representation and the basis data. Here, we limit the scope of approximation theory discussion to single valued discrete functions of a single independent variable, as characterize Types IV and V data. Type III data dismiss spatial-temporal dependence by collapsing the independent variable to a single scatter plot. Accordingly, these data cannot be interpreted as single valued (Methods related to the approximation theory techniques discussed here, but applicable to Type III data, are treated in the Basic Statistical Analysis section below). Two subcategories of discrete approximation methods are best approximation methods‡ and interpolation methods. The discussion here is limited to linear methods, that is methods based on linear combinations of basis func-

‡.Best approximation methods are also termed regression methods in the literature.
The best approximation problem is characterized by an overdetermined system. Specifically, a functional approximation basis will have fewer degrees of freedom (say coefficients of an nth order polynomial) than the number of data points defining the discrete function to be approximated. The problem is then closed by minimizing an appropriate norm of the difference between the discrete data and approximating function. So the best approximation process involves: 1) Specification of a basis family of functions (e.g. polynomial, exponential), 2) Selection of appropriate norm(s) for assessing the accuracy of the representation and 3) Determination of functional coefficients which minimize the selected norms. It is important that both the basis functions and the norm selected in steps 1 and 2 be chosen with careful consideration for what the approximation is to be used for. In particular, basis functions should be selected that retain the important features of the data while ignoring the “noise” or unimportant features of the data.

By far the most employed norm in best approximation methods is the $L_2$ norm. Best approximation methods which employ minimization of an $L_2$ norm are termed least-square methods and are characterized by a minimum “energy” of total error, and overall efficiency of the method when orthonormal basis functions are used. If the chosen basis functions are linearly independent, and an $L_2$ norm is selected for minimization, the approximation problem involves the solution of the normal equations, an $n \times n$ linear system, where $n$ is the number of points defining the discrete function. If in addition to linear independence the basis functions are orthonormal, then this linear system can be far more easily solved. Weighted least squares or “chi-square” minimization is a generalization to $L_2$ minimization, where each data point is ascribed a weight in the least squares procedure. If the standard deviation of the datum about its mean is known a priori, this weight can be selected to be inversely proportional to it. This approach can "anchor" a fit towards regions where the standard deviation is known to be low.

Other than $L_2$, norms used for best approximation are the $L_{\infty}$ and $L_1$ norms. All three are subsets of the generalized $L_p$ norm. $L_p$ norms are seldom used when $p$ is other than 1 or 2, except to approximate the $L_{\infty}$ norm (Fletcher et al. [1971]), and it is therefore not considered beyond that role here. The $L_{\infty}$ norm has been widely used for discrete data approximation, with minmax or Chebychev basis polynomials. These polynomials have the desirable feature of the smallest (or nearly so in the case of Chebychev) maximum deviation (for a given polynomial order) from the approximated discrete function. In general, the $L_{\infty}$ minimization problem in approximation theory is not restricted to polynomial basis functions, and this gives rise to significant mathematical complexities related to uniqueness, convergence and computational intensity, beyond the scope of the present discussion.

The $L_1$ norm minimizes the average absolute value of a functional approximation to discrete data and therefore can be a desirable minimization norm when a small percentage of the data can be deemed erroneous, as characterized by obvious deviation from trends set by the remainder of the data. This is because the effective weight given these points is smaller in the $L_1$ norm than any other $L_p$ norm. Such differences in the sensitivity of approximations to “outliers” is the topic of robust estimators” in approximation theory (e.g. M-, L-, R-estimates, see Press et al. [1994] for example). Details of these methods are beyond the scope of this review, but linear $L_1$ minimization can be considered a subset of M-estimation.

The most efficient algorithms to use to obtain the best approximation fit, for a given basis, varies with the norm chosen for minimization. However, for all norms, the best approximation problem can be viewed as a multidimensional minimization (or optimization) problem, and stan-
standard optimization methods can be deployed.

In order to demonstrate the relative merits of these various norms, an NRS data example is provided here. In particular, a time segment of an OSU SBLOCA test from Lee and Rhee (1997) is considered. Figure 2a shows a plot of measured and RELAP5 predicted vessel pressure vs. time for the NRC12 case. In Figure 2b, the absolute error is plotted vs. a normalized time coordinate. A quadratic fit was selected to represent the absolute error, and $L_1$, $L_2$ and $L_\infty$ norms were used for minimization. Here a “three-dimensional” (corresponding to three polynomial coefficients) downhill simplex optimization technique was used to determine the fit. The norm features described above are observable. In particular, the $L_\infty$ norm fit responds to the very large spikes in error and thereby gives rise to an obviously poor fit. $L_1$ and $L_2$ norm fits are similar with $L_1$ responding less to the large spikes in absolute error early in the time segment, as expected.

In summary, best approximation methods define a subspace (basis) of possible approximations and the best approximation from this space is determined by minimization of an appropriate norm. Another approach to approximating a discrete function is to exactly fit a basis with $n$ degrees of freedom to the $n$ data points. This defines the interpolation methods subset of approximation theory. The most common of these is polynomial interpolation, where an $(n-1)$th order polynomial is fit to $n$ data points. It is well known that such a fit exists and is unique (Weierstrass Approximation Theorem) for single valued functions, $f_i(x_i)$.

Approximating polynomials can be constructed which interpolate between function values at discrete points (Lagrangian) and also, if specified, derivatives of the function values (Hermite). (Hermite can be generalized to higher order derivatives as well (the classical Taylor polynomial fit at a point is an example). Each of these interpolation methods are special cases of the general Hermite-Birkhoff interpolation problem. Hermite interpolation is more relevant to continuous function approximations, where analytical derivatives of the basis data are available. Accordingly, the relevance of this method to discrete NRS data is likely small. Interpolation is obviously not appropriate for Type III data since both variables in these sets are independent (and functional relationships between them are therefore not single valued). Simple Lagrangian interpolation can be relevant to automated code assessment for Type IV data.

Polynomial interpolation can yield unrealistic variation between discrete data points (Runge phenomenon), especially when a large number of data are being fit (large $n$) and the interpolated variable spacing is uniform**. This is often the case for Types IV and V NRS data (Figures 1d, 1e) where $\Delta x$ and $\Delta t$ are typically constant or near constant, and records can be long (often $O(10^5)$). In general, polynomial interpolation is not a good choice for data characterized by sharp rises surrounded by weakly stretched curves, as can describe some Types IV and V data. Also, for large $n$, the polynomial interpolation problem can be computationally intensive.

Though many discrete functions cannot be adequately approximated using a single polynomial applied across its range, locally applied polynomial fits can effectively represent discrete data. Cubic splines are by far the most common of these methods. There, a 3rd order polynomial is fit between each data pair. The four degrees of freedom (coefficients) of this cubic are closed using the two $(x, y)$ values at the endpoints and enforcing gradient continuity with the two adjacent splines. The compact support offered by cubic splines, and other related splines (some classes of B-splines, exponential splines) ameliorate the Runge phenomenon, and thus often return far more realistic function distributions between data pairs.

**. The Chebychev class of polynomials minimizes such over/undershoots, when the locations of independent variables can be specified, but this is not relevant to NRS data.
Figures 3a and 3b illustrate some of the above interpolation techniques for sample Type IV NRS data. In particular, MIT-Siddique test data, digitized from Shumway [1995], is approximated. In Figure 3a, the failings of a seventh order polynomial interpolated to the eight data pairs are observed. Unrealistic variations between pairs are observed for experimental, RELAP5 and absolute error. A standard cubic spline is applied in Figure 3b and this interpolation procedure is seen to provide a far more realistic distribution of the measured and computed quantities and the absolute error.

There appears to have been no direct application of approximation theory methods to Types IV and V NRS data in the literature (though best approximation analysis has been used for Type III data as discussed in the Basic Statistical Analysis section below). However, the various norms considered above (and hybrids of them) have been widely used in the NR and Atmospheric Science communities to directly measure code-data differences for Type III data. These are discussed below.

As discussed and illustrated above, there is a significant opportunity to usefully bring elements of approximation theory into NRS code-data and code-code comparisons. For example, low order polynomial best approximation with \( L_1 \) and/or \( L_2 \) minimization can be used to smooth and integrate NRS data Type V absolute error. Also spline fits can be used to approximate Type IV data. If applied to absolute error, such fits could also be integrated (perhaps more meaningfully than discrete numerical integration) yielding figures-of-merit.

**Time-Series Data Analysis Methods**

Time-series data analysis techniques are designed to estimate properties of a measured or computed process from a time series of repeated successive observations which are not necessarily independent. Time series data analysis techniques are considered here for NRS Type V data.

In general, data which are amenable to time-series analysis are those which can be modeled as stochastic processes, that is, processes which can be described using probabilistic laws. Time series methods are themselves broadly sub-classified between probabilistic methods and spectral methods. Both are considered here.

Probabilistic methods model processes based on assumptions concerning the nature of the process being studied, and using basic statistical measures. Most of these techniques are formulated for stationary processes, though a number of methods are available to transform data sets so as to render stationary techniques applicable (at least locally). These transformation approaches are discussed below. Assuming stationary data for now, the first step in the application of a probabilistic time series data analysis technique is the determination of an appropriate model of the process under consideration. Such models include purely random processes, moving average (MA) processes, autoregressive (AR) processes, random walk processes and more general combinations or extensions of these (e.g., ARMA, ARIMA). Particular classes of data are well described by particular process models. For example, economic data is often well suited to moving average process modeling.

Once a particular class of process model is selected, the model is “fit” to the data. Standard statistical measures (mean, variance, autocovariance) and other model coefficients are determined which define the fit. “Goodness of fit” measures are deployed (residual analysis) which can provide quantitative measure of how good the model has performed, and how reliable forecasting based on the model is.

The potential usefulness of probabilistic time series data analysis techniques to NRS data is demonstrated in Figure 4, where a “nearly stationary” segment of the OSU SBLOCA test introduced above is analyzed. Figure 4a shows the measured and RELAP5 predicted results between 10000 and 14000s for this case. In Figure 4b, the autocorrelation function of measured and com-
Puted pressure traces are plotted vs. time lag for the experimental data and RELAP5 simulation. Also appearing there is an approximate MA process fit to the RELAP-5 simulation. This fit models the data at a given time step as a weighted linear combination of the data values at some number of previous time steps. Two pieces of information are clearly accessible from the autocorrelation plots. First, variations in the experimental measurements are far more random in nature than the RELAP5 results in this region. The computed results show significant autocorrelation out to a lag of more than ten time steps. Second, it is observed that a MA process can do a reasonable job of modeling this feature of the predicted transient.

For stationary or weakly non-stationary data, code-data comparisons of autocorrelation function can be made. In particular, the magnitude of autocorrelation function at a given time lag, or the integral of the autocorrelation or autocovariance to a given time lag can be compared. Alternatively, MA and other probabilistic time series data analysis models can be used to directly compare the computed and measured time histories through direct comparison of the coefficients of the process fitting procedure.

The other class of time series data analysis techniques are spectral techniques. In these methods, the time series is assumed to be composed of sin and cosine waves at different frequencies, that is, a process is modeled through assumed spectral characteristics as opposed to probabilistic characteristics. The most common spectral time series data analysis methods are Fourier series or discrete Fourier transform techniques. These can be viewed as best approximation procedures using trigonometric basis functions (which form an orthonormal set) and employing $L_2$ minimization. Such techniques are ubiquitously applied in experimental methods, functional analysis and numerous other fields.

The sample spectrum, power spectrum and spectral density function are also important tools in time series data analysis. Like the discrete Fourier transform, these can be used to isolate fundamental frequencies in a noisy signal, and are sometimes better suited to this task than the discrete Fourier transform. They are closely related to the autocovariance function, indeed, they are transforms of one another. This mathematical equivalence does not render consideration of both spectrum and autocorrelation function redundant, since their "representational value" may differ, that is, important information may be more accessible in one over another (Box and Jenkins [1976]).

The discrete Fourier transform has been used in the NR community for automated code assessment by D’Auria and his coworkers (Ambrosini et al. [1990], for example). In this approach, the discrete Fourier transform of the measured and computed time trace is obtained. From the amplitudes of the component frequencies, two characteristic quantities are computed, the average amplitude, AA, and the weighted frequency, WF. The AA sums the difference between experimental and code discrete Fourier transform amplitudes at each frequency. The WF weights each frequency difference in the summation appearing in the AA with the frequency itself. Each measure is non-dimensionalized. The AA clearly provides a measure of the absolute amplitude error for a simulation, and WF provides an indication of where the frequency errors are largest.

To illustrate this method, “artificial” data sets used by D’Auria and his colleagues have been reproduced in Figure 5a. Here an “experimental” transient and six “code” results, digitized from Ambrosini et al. [1990] are reproduced. The code results were originally selected to characterize a variety of code-data discrepancy features. In Figure 5b, the present authors have computed the AA and WF quantities for the six cases and these results closely correspond to those previously published, as expected.

In automated code assessment, the D’Auria FFT approach can be used to quantify code accuracy in a number of ways. For example, threshold “contours of acceptability” can be defined in
the AA-WF plane, each simulation then returning a single figure-of-merit.

Rigorous application of both probabilistic and spectral time series data analysis methods to automated code assessment is limited to stationary periodic data. In addition, spectral approaches which employ global transforms (such as the discrete Fourier transform) are well known to give poor representations of signals characterized by local phenomena. Indeed, square waves, reminiscent of the artificial experimental data in Figure 5a are often used to illustrate this (i.e., Gibb’s phenomena).

Despite these limitations of D’Auria’s discrete Fourier transform method, useful information on code-data accuracy has been obtained using this method by several researchers in the literature. Accordingly, the present investigators have incorporated this method in ACAP.

Basic Statistical Analysis Methods

The two classes of methods considered so far encompass data analysis procedures that are inherently applicable to successive data. As such, the approximation and time series analysis methods model data in a fashion which describes discrete functional behavior with respect to time or space, making them more appropriate for Types IV and V NRS data. Basic statistical analysis methods can also be brought to bear in analyzing NRS data. The field of statistics can be broadly defined to incorporate approximation theory and time series data analysis methods. Basic statistical methods are here distinguished, as methods that describe random data in a fashion that is unconcerned with the spatial or temporal ordering of the data. Data are treated as a sample of k observations of one or more variables. Index k designates a running index over individual realizations in this data set. An example of data ideally suited to basic statistical description and analysis would be the test scores and IQs (x_k, y_k) for a sample of k students.

Single random variables are of fundamental concern in statistics. Here, a single variable, x_k, say test scores, are sampled, and then standard descriptive measures of the sample are computed. Such measures include the mean, variance, median, skewness and other more arcane measures. For automated code assessment, these descriptive measures can be applied to the absolute error, and as such have been termed statistical difference measures, and been fairly widely used in the atmospheric sciences community (Fox [1981, 1984], Wilmott [1982], Rao [1987], for example).

Also, multiple random variables can be identified with individual realizations (e.g., x_k = test score, y_k = IQ) and the relationships between these can be studied using cross-correlation and regression procedures. Again in concert with designations adopted by the atmospheric sciences community (ibid.) these methods are here termed statistical correlation measures when applied to code-data comparisons. Predicted value and measured value are treated as paired random variables in these automated code assessment applications. Both statistical difference measures and statistical correlation measures are discussed here.

Straightforward application of basic statistical analysis methods, as just defined, dismiss spatial and temporal localization information. Data are considered from a basic statistical viewpoint as samples comprising one or two random variables (experimental value and/or computed value) with a priori notion of an independent variable ignored. Accordingly, if there are significant spatial or temporal trends in the data (as is the norm in NRS data), quantities like mean, standard deviation and correlation coefficient can be misleading and/or useless. However, if time trends can be removed, or if statistics measures are applied locally (in time), these techniques can provide, if not rigorous, at least useful information. Measures that preprocess the data so as to improve the stationarity assumption are discussed below. If time (or space) localization information is eliminated a priori (as is the case with NRS Type III data), basic statistical measures can also be usefully applied.

A number of statistical difference measures have been applied in the NR community
(Kmetyk et al. [1985], Wilson et al. [1985], Ambrosini et al. [1990], D’Auria [1995a], for example) and in the atmospheric sciences community (Fox [1981, 1984], Wilmott [1982], Ku et al. [1987], for example). These include:

1. Mean error (or average absolute error), ME
2. Variance of error (square of standard deviation), VE
3. Mean square error, MSE
4. Mean error magnitude, MEM
5. Mean relative error, MRE

Measures 2 and 4 are closely associated with the $L_2$ and $L_1$ norms discussed above, respectively. Relative error measures normalize the absolute error by the local magnitude of the data (measured and/or computed). In addition to these basic difference measures, NR and atmospheric sciences workers have deployed other derived difference measures including:

6. Index of agreement (Wilmott [1982]), IA
7. Systematic and unsystematic mean square error (Wilmott [1982]), MSE$_S$, MSE$_U$
8. Mean fractional error (Ku et al. [1987]), MFE

These latter three non-standard statistical difference measures have some potentially appealing features for automated code assessment. In particular, the index of agreement distinguishes between the predicted and measured quantity in its definition, and has been defined as the “measure of the degree to which the observed [quantity] is accurately measured by the simulated [quantity]” (Ku et al. [1987]). The index of agreement is non-dimensional. Systematic and unsystematic mean square errors measure, for the observed and predicted data respectively, difference from a linear least squares fit of their correlation (predicted = linear function of observed). By introducing these two measures, and comparing their magnitudes to the mean square error, one can determine how close the predictions are to “as good as possible”. This is illustrated below. The mean fractional error was defined in an attempt to reduce the bias afforded larger magnitude data by statistical measures based on absolute error, as well as the bias afforded smaller magnitude data by relative error based measures.

To illustrate the utility of these measures, they are each applied to a sample Type III NRS data set. Figures 6a and 6b show sample data adapted from Shumway [1995]. These plots show comparisons of RELAP5 simulations of UCB wall condensation tests (separate effects tests that simulate PCCS conditions). For this demonstration calculation, these data were digitized directly from the printed reference and then analyzed. The descriptive measures introduced above were computed and are given in Table 1 for two RELAP simulations (which represented code runs that implemented default and “improved” diffusion models respectively).

<table>
<thead>
<tr>
<th>Descriptive Statistical Measure</th>
<th>RELAP5 – default diffusion</th>
<th>RELAP5 – new diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean error, ME (average absolute error)</td>
<td>$-0.143 \times 10^{-1}$</td>
<td>$0.755 \times 10^{-2}$</td>
</tr>
<tr>
<td>Variance of error, VE</td>
<td>$0.131 \times 10^{-2}$</td>
<td>$0.483 \times 10^{-3}$</td>
</tr>
<tr>
<td>Mean square error, MSE</td>
<td>$0.130 \times 10^{-2}$</td>
<td>$0.480 \times 10^{-3}$</td>
</tr>
<tr>
<td>Mean error magnitude, MEM</td>
<td>$0.271 \times 10^{-1}$</td>
<td>$0.159 \times 10^{-1}$</td>
</tr>
<tr>
<td>Mean relative error, MRE</td>
<td>$-0.982 \times 10^{-1}$</td>
<td>$0.953 \times 10^{-1}$</td>
</tr>
</tbody>
</table>
These statistics consistently confirm the superiority of the new model. Several observations apply:
1. The ME, VE and MEM are significantly smaller for the new model.
2. VE and MSE are nearly identical owing to the small values of ME.
3. The ME and MRE indicate the degree of bias in the predictions. The tabulated values of ME suggest a significant average underprediction of the data for the original model and a small average overprediction for the newer model. This is a manifestation of the favoritism afforded the cluster of lower magnitude data for the original model. This is observable in Figure 6c which plots the UCB data relative error. As discussed above, the MFE is a more consistent measure of bias. The ratio of MFE between the two models (1.45) lies between the ratio of ME (1.89) and MRE (1.03).
4. The IA is significantly better (i.e. closer to the perfect agreement value of 1.0) for the new model.
5. The new model predictive improvements quantified by the above measures are accompanied by an increase in the systematic component of the variance (increased MSE_s/MSE). This suggests that further improvements to the new diffusion model would likely be possible.

Statistical correlation measures can also provide quantitative descriptions of the correspondence between the data. In particular, the magnitude of cross-correlation coefficient and a “goodness” measure for a polynomial fit could be used to quantitatively provide a figure-of-merit for code-data comparisons. For example, four linear statistical correlation measures were computed for the same UCB data set appearing in Figure 6: 1) the cross-correlation coefficient, \( \rho_{xy} \), 2) the L_2 norm of a linear least squares fit to the data, L_2-standard, 3) the L_2 norm of a linear least squares fit to the data constrained to pass through the origin, L_2-constrained and 4) the L_2 norm of the difference between the data and the “perfect agreement line” defined by \( q''_{\text{EXPT}} = q''_{\text{RELAP}} L_2\text{-deviation} \). This is a measure of absolute error. The calculated values of these measures appear in Table 2.

<table>
<thead>
<tr>
<th>Descriptive Statistical Measure</th>
<th>RELAP5 – default diffusion</th>
<th>RELAP5 – new diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index of agreement, IA</td>
<td>0.847 x 10^0</td>
<td>0.916 x 10^0</td>
</tr>
<tr>
<td>Systematic mean square error, MSE_s</td>
<td>0.236 x 10^-3</td>
<td>0.128 x 10^-3</td>
</tr>
<tr>
<td>Unsystematic mean square error, MSE_u</td>
<td>0.106 x 10^-2</td>
<td>-0.352 x 10^-2</td>
</tr>
<tr>
<td>MSE_s/MSE</td>
<td>0.181 x 10^0</td>
<td>0.267 x 10^0</td>
</tr>
<tr>
<td>MSE_u/MSE</td>
<td>0.819 x 10^0</td>
<td>0.733 x 10^0</td>
</tr>
<tr>
<td>Mean fractional error, MFE</td>
<td>-0.378 x 10^-1</td>
<td>0.260 x 10^-1</td>
</tr>
</tbody>
</table>

Table 1: Descriptive Statistical Measures for Type III Data.

<table>
<thead>
<tr>
<th>Statistical Correlation Measure</th>
<th>RELAP5 – default diffusion</th>
<th>RELAP5 – new diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_{xy} )</td>
<td>0.782</td>
<td>0.847</td>
</tr>
<tr>
<td>L_2-standard</td>
<td>0.421</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Table 2: Statistical Correlation Measures for Type III Data.
The correlation coefficient shows modest improvement for the new model (i.e., it is somewhat closer to the perfect correlation value of 1.0), whereas all $L_2$ norms are significantly lower (indicating much better agreement). (The lines corresponding to the two least squares analyses performed and the “perfect fit” $q''_{\text{EXPT}} = q''_{\text{RELAP}}$ line appear in Figures 6a and 6b).

The usefulness of basic statistical analysis measures for assessing Type III data is suspect when, as in the above example, the mean value of the absolute error is not statistically relevant. This is because data at various spatial locations are included, and significant spatial trends exist. In this circumstance, the mean and the variance cannot be deemed “good statistical estimators” (Bendat and Piersol [1986]) since the mean does not necessarily represent an expected value of absolute error. Similar serious difficulties arise for Types IV and V data. For these, application of basic statistical analysis techniques again dismiss the temporal or spatial nature of the data. They are treated as random samples. This is appropriate only if the data is stationary. Otherwise, as above, even the basic statistical measures of mean and variance are of questionable merit.

Even if a particular NRS data set were stationary, and reasonable mean and variance values could be determined, it is not appropriate to assume any distribution of the absolute error about its mean. Accordingly, the most powerful aspect of basic statistical analysis techniques, statistical inference, cannot be deployed. Specifically, because we have no knowledge whatsoever regarding the probability density function of the absolute error about its mean, we cannot make assumptions on its form (say Gaussian, Student t, Chi-square). Therefore, we cannot establish uncertainty or confidence interval bounds on the absolute error. This makes it difficult to determine whether differences in basic statistical analysis measures are statistically significant enough to draw meaningful conclusions. The foregoing arguments are illustrated in Figures 6e and 6f. There, the probability density function (PDF) for the absolute error of the two UCB “samples” are seen to be neither consistent with one another nor with any well defined form. Nevertheless, assuming a normal distribution of the absolute error about its mean, a 95.4 (or 2$\sigma$) confidence interval can be easily determined: $-0.087 \leq ME_{\text{default diffusion}} \leq 0.058$, $0.036 \leq ME_{\text{new diffusion}} \leq 0.052$. Confidence intervals can also be constructed for other descriptive measures and correlation measures. The Fischer-Z variable (Nadler and Smith [1993], for example) can be used to construct a significance measure for $\rho_{xy}$.

Wilson et al. [1985] also assumed a normal distribution of absolute error about the mean of a locally near-stationary NRS data set, and then proceeded to construct a 95% confidence limit on the mean absolute error. They, of course, recognized the limitations discussed above and presented their results as “reasonable confidence limits [that] would be at the 95% level if [the absolute error was normally distributed stationary data].” Fox [1980] also constructed confidence intervals for average absolute error in atmospheric analysis-data comparisons, but similarly noted, “If the assumptions concerning the use of the distribution upon which the interval construction is based are seriously violated, the interval statement itself will be inaccurate”. The present authors believe that the definitive unavailability of a known distribution function for absolute error in NRS code-data comparisons renders such statistic inference approaches inappropriate. Constructing statistical esti-

Table 2: Statistical Correlation Measures for Type III Data.

<table>
<thead>
<tr>
<th>Statistical Correlation Measure</th>
<th>RELAP5 – default diffusion</th>
<th>RELAP5 – new diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_2$-constrained</td>
<td>0.421</td>
<td>0.235</td>
</tr>
<tr>
<td>$L_2$-deviation</td>
<td>0.465</td>
<td>0.265</td>
</tr>
</tbody>
</table>
mators for NRS data, as above, can provide a useful indication of code accuracy, but in the authors’ opinion, rigorous statistical inference measures should not be computed and used to assess code-data uncertainty. This position seems consistent with Wilmott’s [1982] position: “Confidence bands and tests of statistical significance are not nearly as illuminating as an informed scientific evaluation of the summary and difference measures.”

In summary, the authors incorporated each of the statistical difference and correlation measures summarized above in ACAP. As discussed below, a code accuracy figure-of-merit is constructed based on some subset of these measures, but statistical inference, including the construction of significance measures, is not implemented.

**Trend Removal and Time Windowing**

As discussed above, a common characteristic in the application of time series data analysis and basic statistical analysis class methods to Type V NRS data is that non-stationarity of NRS data renders many of the powerful methods within these classes less useful or inapplicable. Nevertheless, in the discussion and demonstration computations above, the application of several of these methods to “raw” NRS data was provided, and was seen to be of some use in automated code assessment. Two alternatives to this “apply-it-anyway” approach are:

1. Pre-processing of the NRS data, rendering it amenable to more rigorous application of time series data analysis and basic statistical analysis methods.
2. Application of methods expressly designed for non-stationary data.

Techniques in the first category include trend removal and time-windowing, and these are discussed here. Several non-stationary analysis methods are considered below.

There are a number of methods available for transforming data to more closely satisfy the stationary process assumption. Such trend removal techniques can therefore, in principle, increase the usefulness of time series data analysis and basic statistical analysis methods. These techniques are best applied when the non-stationarity is not of principal interest, as in the removal of “drift” from an experimental data set. Examples of techniques used in trend removal include simple curve fitting (i.e. best approximation methods from approximation theory discussed above), smoothing, high pass filtering, running averages, evolutionary spectra, complex demodulation, intervention analysis and Bayesian forecasting (several of these are closely related).

In trend removal, the modeled trend (or its deviation from its mean) is subtracted or filtered from the raw data. If the removed non-stationarity is not of principal interest, then this information is discarded. However, if the trend itself is of importance (as is usually the case in USNRC data), the “removed information” should be retained for concomitant analysis. In this light, the trend removal can be considered as a linear decomposition. For the present automated code assessment application, the authors believe that it is more general, and often appropriate, to assume that both stationary and non-stationary components of NRS data sets are important from an accuracy standpoint. Accordingly, separate data analysis techniques can be brought to bear on the stationary and non-stationary components of the time series. If the underlying assumptions of these separate analyses are not violated, they can be considered together in constructing a code-data comparison measure. This should yield a more rigorous, more robust (less susceptible to pathological exceptions) and more accurate comparison measure.

Two approaches to automated code assessment which accommodate this view were considered:

1. Trend removal is first performed on both experimental and systems code data. Time series data analysis and/or basic statistical analysis methods are then applied to the two residuals
(raw data - trend) which should be “closer” to stationary than the absolute error, and at least have a mean much closer to zero than the absolute error. Approximation theory and/or basic statistical analysis comparison measures are then applied to experimental and computed trends.

2. Trend removal is first performed on experimental data only. The experimental trend is then subtracted from both experimental and systems code raw data yielding a more nearly stationary experimental residual trace and presumably a significantly non-stationary trace associated with the simulation.

Option 2 has several advantages over option 1. First, since the experimental data will be fixed in the automated code assessment data base, only one trend removal modeling process need be performed per time trace. Once modeled, the continuous or discrete representation of the trend can be fixed in the experimental data base as well. Second, since no trend removal is applied to the code data, arbitrariness of non-stationary trend modeling is not introduced there, an item which could be important since a large number of simulations will be compared with the same experimental data set over the course of time. Unfortunately, the authors have found that in practice option 2 comes at the cost of significant non-stationarity being present in the “raw code - experimental trend” trace, since the simulation trends will not necessarily coincide with the experimental data. This renders the analysis of the absolute error of the residuals nearly as inappropriate for time series data analysis and best approximation methods as the absolute error of the original traces. So option 1 has been chosen for further consideration and is demonstrated here.

Figure 7a shows a comparison of the predicted and measured rod temperature in the core heatup and reflood stages of a FLECHT SBLOCA test vs. a TRAC-B simulation (Paige [1998]). A particular axial location in the core has been selected. For reference, the absolute error is plotted in Figure 7b. This data is clearly nonstationary, and though the basic trend of the data is well captured by the simulation, there is some underprediction of the peak rod temperature and some oscillatory features in the simulation, both of which should be captured quantitatively using an automated code assessment procedure.

The option 1 procedure was applied. A running average (of 80 time steps) was performed on the raw experimental and TRAC-B data sets. Figure 7c shows the computed trends arising from this process. The residuals of these two data sets (raw data - trend) are shown in Figure 7d. Inspection of Figures 7c and 7d clearly suggest the utility of trend removal in isolating two classes of discrepancy. The absolute error of the trend defines the overall peak rod temperature error level, and can be quantified using several of the approximation theory and basic statistical analysis methods described above. The residuals and their absolute error are plotted in Figure 7d are clearly more amenable to time series data analysis and basic statistical analysis than the raw absolute error plotted in Figure 7b. These methods provide figures-of-merit quantifying the significant oscillation appearing in the absolute error.

In this example, the trend removal process selected was not ideal for capturing the global trend associated with reflood. In particular, the running average smoothes the steep temperature drop features in both the measured and computed traces, as seen in Figure 7c, and this manifests itself in a transfer of error content from trend to residual in this region (t > 230 s). It is likely that an alternative global trend removal process could do a better job at capturing this local feature (e.g. smaller range running average or high pass filter). However, the time scale of this sharp descent feature (which has been chosen to be part of the global trend) is commensurate with the time scale of the oscillation feature in the absolute error (which has been designated part of the residual.) Global trend removal processes cannot therefore completely distinguish the two. This difficulty motivates the next topic, time windowing.

Time windowing, that is separating regions of the time trace prior to data analysis, can ame-
liorate some of the ambiguities associated with global approximation theory, time series data analysis or basic statistical analysis methods. Indeed, when the techniques defined so far, including trend removal, are successively applied to a few suitable, predefined time windows, more meaningful and robust comparison measures can be constructed. This is illustrated here.

Figure 7e shows the FLECHT data with two defined time windows associated with transitional and reflood segments of the SBLOCA. For the transition window, the same running average trend removal deployed above was used, but a smaller running average range (6 time steps) was used for the reflood window. The desired original trend in now well captured, as seen in Figure 7f. Also, Figure 7g illustrates that undesirable transfer of error content to the residuals has been mitigated. For this case, a reasonable choice for trend figure-of-merit is MRE. Reasonable choices for residual figures-of-merit are ME, VE, AA and WF. These five figures-of-merit are given in Table 3.

Unfortunately, predefining time windows introduces, by definition, some subjectivity into the automated code assessment process. This issue has been treated extensively in the NR automated code assessment literature, and several investigators have resolved that time windowing is required (Kmetyk et al. [1985], D’Auria et al. [1995]). In the view of the present authors, time windowing can be incorporated definitively in an automated code assessment process. Specifically, each experimental set in the automated code assessment data base can have associated with it pre-defined time windows. These ranges will be agreed upon for each test matrix trace prior to incorporation within the automated code assessment data base. The process for defining and achieving consensus on these is not treated here, but would presumably become part of a formal process in augmenting the automated code assessment data base. Once in the data base, these time windows become fixed. This approach eliminates subjectivity in the process. ACAP incorporates both trend removal and time windowing options, as summarized in Kunz et. al. [1998a, 1998b, 2000].

Table 3: Figure-of-Merit for FLECHT Data.

<table>
<thead>
<tr>
<th>Figure-of-merit</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRE$_{\text{trend}}$</td>
<td>0.054</td>
</tr>
<tr>
<td>ME$_{\text{residual}}$</td>
<td>-0.2 °K</td>
</tr>
<tr>
<td>$\sigma = (\text{VE}_{\text{residual}})^{1/2}$</td>
<td>8.2 °K</td>
</tr>
<tr>
<td>AA$_{\text{residual}}$</td>
<td>1.1</td>
</tr>
<tr>
<td>WF$_{\text{residual}}$</td>
<td>29.5</td>
</tr>
</tbody>
</table>

The MRE$_{\text{trend}}$ captures the 5% average underprediction in trend in the transition region. The proximity of ME$_{\text{residual}}$ to 0 provides that the trend removal process was effective. The residual standard deviation of 8 °K is primarily due to the low frequency oscillation in absolute error. The AA value of greater than 1 indicates a discrepancy in average amplitude between the data and prediction larger than the average amplitude of the data residual itself. This discrepancy is contained in lower wave numbers and identified by the nondimensional WF of 29.5 (which indicates centering of AA near mode 30. There are 260 modes in the discrete Fourier transform).

Unfortunately, predefining time windows introduces, by definition, some subjectivity into the automated code assessment process. This issue has been treated extensively in the NR automated code assessment literature, and several investigators have resolved that time windowing is required (Kmetyk et al. [1985], D’Auria et al. [1995]). In the view of the present authors, time windowing can be incorporated definitively in an automated code assessment process. Specifically, each experimental set in the automated code assessment data base can have associated with it pre-defined time windows. These ranges will be agreed upon for each test matrix trace prior to incorporation within the automated code assessment data base. The process for defining and achieving consensus on these is not treated here, but would presumably become part of a formal process in augmenting the automated code assessment data base. Once in the data base, these time windows become fixed. This approach eliminates subjectivity in the process. ACAP incorporates both trend removal and time windowing options, as summarized in Kunz et. al. [1998a, 1998b, 2000].
Other Methods

Time-Frequency Methods

NRS data is characterized by non-stationarity, and this limits the applicability and power of many of the techniques introduced above. As discussed above, trend removal and time windowing can be effective in rendering the data closer to stationary. Alternatively, time-frequency methods are directly applicable to non-stationary data, and are therefore of interest here. Two principal techniques in this class are the short time Fourier transform and the wavelet transform.

A Fourier transform can be considered an alternative way of looking at a time varying data record. As discussed above, this alternative view may carry some useful qualitative and quantitative information, especially for stationary data. However, the FT does not provide easily interpretable time localization information which is relevant to non-stationary data records.

The short time Fourier transform defines a time window which slides along the time trace. At each timestep, a discrete Fourier transform is applied to this local time window. From this, local frequency and phase content of the signal is obtained. If the sample size within the window is large and the trace within the window is near stationary, the short time Fourier transform will capture accurate, time localized spectral information of the signal. Such a time-frequency method ameliorates the problems associated with global transforms applied to data with local features (discussed in time series data analysis section above). This, of course, comes at the expense of increased dimensionality in the problem.

The short time Fourier transform is characterized as having a “fixed resolution” over the entire time-frequency domain. This manifests itself by limiting the short time Fourier transform analysis to having good temporal resolution or good frequency resolution (depending on choice of window size) but not both. This limitation is overcome using wavelet transforms. Wavelet transforms differ from short time Fourier transform methods in that their choice of basis functions are not necessarily sinusoidal, and their resolution effectively varies in the time-frequency plane. This allows for more accurate representation of features in a time trace than short time Fourier transform methods, especially when important features appear at widely varying times and/or frequencies. These basic ideas are illustrated in Figure 8. In Figure 8a a segment of the same OSU SBLOCA data plotted in Figure 2a appears (experimental data only here). A short time Fourier transform and wavelet transform of this data appear in Figures 8b and 8c. These plots are spectrograms of the transforms, that is, contour plots of the square modulus of the transform coefficients in the time-frequency plane. Both transforms capture the higher energy associated with the oscillatory feature in the time trace, in a time localized fashion. The wavelet transform is seen to provide a better resolved representation of the feature.

There are a wide array of discrete and continuous wavelet transforms available, and the proper choice depends primarily on the nature of the features being extracted (Morlet continuous wavelet transforms were deployed for the results presented here). Some of these target features include local periodicity, local minima and maxima and, importantly, their variation in time. Accordingly, it is likely that a suite of wavelet transform tools could be used effectively in automated code assessment for NR applications (currently, the Morlet transform is available in ACAP). Some examples of the fields where time-frequency methods have been used for time series data analysis are the atmospheric sciences (Torrence and Compo [1998], for example), fluid turbulence research (Wang et al. [1997] for example), aircraft engine stability (Hendricks et al. [1997]),

††.It is this feature that motivates the principal use of wavelets, data compression.

(20)
mechanical system monitoring and speech pattern recognition.

The question arises as to what to do with the large amount of data that is generated by a wavelet analysis. One approach, devised here, and installed in ACAP, hybrids D’Auria’s method within a time-frequency approach. A wavelet transform of the artificial D’Auria data shown in Figure 5a was taken. Parameters analogous to D’Auria’s average amplitude and weighted frequency were constructed at each time step, and the locus of these points are plotted for each simulation in Figure 8d. This plot illustrates that local accuracy can vary widely in the AA-WF plane. A scalar figure-of-merit can be defined within a prespecified acceptability threshold contour. Such a threshold is defined in Figure 8d. Table 4 lists these “percent acceptable” figures-of-merit.

Table 4: Wavelet Based Figure-of-Merit for D’Auria Data.

<table>
<thead>
<tr>
<th>Case</th>
<th>Percent Acceptable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>98</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

So time-frequency and in particular wavelet analysis offers the possibility of mitigating some of the limitations of discrete Fourier transforms. To summarize, these include applicability to non-stationary data and improved capturing of time local features. These two benefits warranted the incorporation of wavelet transform techniques into ACAP.

Pattern Recognition Methods

Pattern recognition is concerned with the extraction of useful information from large quantities of noisy data. Three common applications of this technology are digital image processing, optical character recognition and speech recognition. The generic pattern recognition problem includes signal conditioning, segmentation, feature extraction and decision making. The first three of these have relevance to trend removal, time windowing and key event identification, respectively, in the analysis of NR systems data, and are therefore treated here.

Signal conditioning involves removal of unimportant information, and approaches for doing so were briefly summarized above.

Segmentation of a time series in pattern recognition theory can refer to either the identification of the starting point of each cycle in a quasi-periodic signal or detection and separation of major regions with common “basic character” (Nadler and Smith [1993]). The latter is relevant here. Statistical or structural segmentation methods can be deployed. In statistical methods some measures of “short but sufficiently long” sequences of the time series are compared, and if sufficiently different, a boundary is defined there. These are most appropriately applied to slowly varying chaotic waveforms, and are therefore relevant to time windowing of NRS data. Structural methods seek known waveform shapes and segment along these boundaries. These methods are therefore most appropriate for quasi-periodic signals and less relevant here.

Edge based detection methods are statistical segmentation approaches which key on features
at the boundaries of time segments. Discrete differential operators (i.e. gradient, Laplacian) or spectral methods (including time-frequency methods) can be used for edge detection. The possibility of using such methods to define time windows in a NRS transient is reasonable. The authors have not to date pursued such an automated segmentation procedure in ACAP, but, as indicated above, time windowing can be prespecified in the automated code assessment data base.

Feature extraction is concerned with obtaining the key information required for decision analysis from the segmented data. Methods used for feature extraction include moment analysis, Fourier transforms (including time-frequency), Karhunen-Loeve transforms and Hadamard transforms. Features such as slope, inflection points and local extrema are important in NRS Type V data and therefore deployment of feature extraction techniques from the pattern recognition literature could be incorporated in ACAP in the future.

Multi-Variate Analysis Methods

NRS data Type IV and V definitions are based on a single independent variable, x and t respectively. Often, multiple NRS Type V time traces are taken simultaneously at different spatial locations in the reactor. This is illustrated in Figure 9, which shows the FLECHT data introduced above at all computed axial locations in a TRAC-B model (Paige [1998]). One or more of these traces may be taken as representative for code-data comparison and analyzed independently, but another possibility is to consider the relationship between them. Such multi-variate approaches characterize the dependent variable, say temperature, with respect to multiple independent variables, say x, t. These methods can capture important features in the data, such as phase lag, which are inaccessible by univariate approaches.

Most multi-variate methods are extensions of the approximation theory, time series data analysis and basic statistical techniques discussed above in the context of a single independent variable. More sophisticated multi-variate approaches, such as the multi-channel singular spectrum analysis method have also been deployed in the atmospheric sciences community (Plaut and Vautard [1994]), to model weather patterns over large areas (space dimensions) and long time periods (multi-year).

The authors believe that multi-variate analysis can eventually be used effectively for automated code assessment, but we have so far implemented only univariate methods in ACAP.

Experimental Uncertainty

The USNRC has emphasized for over a decade (Kmetyk et al. [1985], Bessette and Odar [1986]) the importance of including the contribution of experimental data uncertainty when assessing the accuracy associated with CFD simulations. Indeed, the current USNRC hierarchal quality definitions for code-data comparisons are specified with reference to the experimental uncertainty: “[for excellent agreement], the code will, with few exceptions, lie within the uncertainty bands of the data. Whereas for reasonable agreement, quantitative differences between code and data are generally observed to be greater that the experimental uncertainty.” (Damerell and Simons [1993], Schultz [1993]). This issue has also received attention recently from the industrial CFD community at large (Coleman and Stern [1997]).

Experimental uncertainty is usually reported with a 95% confidence interval, that is, the true value of a quantity is expected to lie within ± the reported uncertainty of the reported data value 95% of the time. If the experimental uncertainty band is large, consideration of experimental uncertainty in the construction of code-data comparison measures lessens the significance and/or magnitudes of these measures. To illustrate this, consider an example adapted from Coleman and Stern
applied to the OSU SBLOCA data used in the examples above. Figure 10a shows predicted and measured integrated mass flow through an Automatic Depressurization System (ADS) for the NRC12 case. As can be seen in the figure, an experimental uncertainty bound is known for this quantity (± 250 kg). If the absolute error is plotted vs. time and considered with the experimental uncertainty for this case, the absolute error is a less meaningful quantity. This is illustrated in Figures 10a and 10b. In particular, an artificial code solution was constructed which, as seen in Figure 10a, clearly exhibits significant differences from both RELAP and measured values (a straight line was taken for the artificial data). However, Figure 10b illustrates that the artificial solution cannot be deemed much less accurate than the RELAP simulation if taken in light of the experimental uncertainty.

This simple example serves to motivate the incorporation of experimental uncertainty in automated code assessment metrics. Building such measures into basic statistical analysis techniques is straightforward. Two possibilities are: 1) reporting mean error magnitude with the experimental uncertainty (MEMEU) or 2) constructing a “percent validated” (PV) metric defined as the percentage of the computed data that lies within the experimental uncertainty (0 ≤ PV ≤ 100). These definitions are consistent with the criteria used for code validation given by Coleman and Stern [1997]. For the data shown in Figure 10, these metrics are given in Table 5.

Table 5: Automated Code Assessment Measures Incorporating Experimental Uncertainty

<table>
<thead>
<tr>
<th>Statistic</th>
<th>RELAP5</th>
<th>Artificial Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEMEU</td>
<td>187 ± 250 kg</td>
<td>313 ± 250 kg</td>
</tr>
<tr>
<td>PV</td>
<td>73 %</td>
<td>66 %</td>
</tr>
</tbody>
</table>

These measures declare the RELAP simulation slightly superior to the artificial data, though both are seen to remain mostly within the experimental uncertainty. The MEMEU and PV metrics are less distinguishing than MEM alone, as desired.

For more sophisticated automated code assessment tools, incorporation of experimental uncertainty is not as straightforward. This is because for more refined error measures, the component contributions to experimental uncertainty must be individually ascertained in order to properly incorporate it within an automated code assessment measure. Consider an example where a zero drift experimental error gives rise to an experimental uncertainty that is of the same order of magnitude as well defined periodic features within the signal. This is illustrated in Figure 11a. Here, an experimental uncertainty of ± .3 is associated with the drift in the measured data. The two analyses shown return similar MEMEU and PV metrics (MEMEU₁ = .36 ± .3, MEMEU₂ = .37 ± .3, PV₁ = 46%, PV₂ = 50%). However, simulation 2 is clearly the superior one if taken in light of its capturing the dominant oscillatory feature of the data.

One approach to resolving this issue is to ascribe the experimental uncertainty to the trend in the data. Now, as indicated above, the automated code assessment data base must include, in addition to the raw experimental data, the experimental uncertainty. Also, trend removal and/or time windowing information may be included with each set. Therefore, including more detailed information on the experimental uncertainty (such as which component in the trend decomposition process it is associated with) is reasonable.

This proposal is illustrated in Figures 11b and 11c. There, trend removal has been performed on each of the signals appearing in Figure 11a. If the experimental uncertainty is ascribed entirely to the trend, MEMEU and PV measures for the two code-data trends are, as before, inconclusive.

Table 5: Automated Code Assessment Measures Incorporating Experimental Uncertainty

<table>
<thead>
<tr>
<th>Statistic</th>
<th>RELAP5</th>
<th>Artificial Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEMEU</td>
<td>187 ± 250 kg</td>
<td>313 ± 250 kg</td>
</tr>
<tr>
<td>PV</td>
<td>73 %</td>
<td>66 %</td>
</tr>
</tbody>
</table>

These measures declare the RELAP simulation slightly superior to the artificial data, though both are seen to remain mostly within the experimental uncertainty. The MEMEU and PV metrics are less distinguishing than MEM alone, as desired.
This says that the figures-of-merit associated with the trends do not establish significant superiority of either simulation. However, examination of the residuals clearly establishes the superiority of simulation 2 over simulation 1. This is obvious upon inspection of Figure 11c. In an attempt to capture this quantitatively, several automated code assessment measures were brought to bear on this residual data. These are presented in Table 6.

Table 6: Automated Code Assessment Measures Applied to Residuals in Figure 11c

<table>
<thead>
<tr>
<th>Figure-of-merit</th>
<th>Simulation 1</th>
<th>Simulation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>-0.0005</td>
<td>0.0039</td>
</tr>
<tr>
<td>MEM</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>(\sigma = (VE)^{1/2})</td>
<td>0.21</td>
<td>0.30</td>
</tr>
<tr>
<td>AA</td>
<td>1.00</td>
<td>1.30</td>
</tr>
<tr>
<td>WF</td>
<td>15.3</td>
<td>13.5</td>
</tr>
<tr>
<td>(\rho_{xy})</td>
<td>(4 \times 10^{-9})</td>
<td>(-0.6 \times 10^{-3})</td>
</tr>
</tbody>
</table>

Several interesting findings apply. First, the ME, MEM and VE measures are very poor indicators of the level of agreement. Indeed, simulation 2 exhibits apparently worse agreement with data if these measures were to be considered. This observed behavior is a manifestation of the slightly different frequency of the simulation 2 residual, which yields large absolute errors where the traces are locally out of phase. This again highlights the care which must be taken in deploying basic statistical analysis methods for automated code assessment. D’Auria’s approach fares as poorly here. The AA for simulation 2 is larger than simulation 1. The explanation for this can be gleaned from Figure 11d, where discrete Fourier transforms for the three traces are shown. Again, the slight difference between measured and predicted amplitudes, designated |absolute error| is included in the figure. The WF parameter captures the not-very-useful fact that the error for simulation 2 is centered at a slightly lower frequency than simulation 1. The final parameter appearing in Table 5 is the cross correlation coefficient. This parameter does a good job in illustrating the superiority of simulation 2.

In summary, it is important to incorporate experimental uncertainty in automated code assessment, and this motivated the authors to include the PV strategy (as well as time windowing and trend removal) in ACAP. Several general shortcomings with basic statistical analysis and spectral time series data analysis methods have been observed here, but selection of an appropriate metric (in this case cross-correlation coefficient) brought to light the essential features in the comparison.

5. IMPLEMENTATION ISSUES

Inconsistency of Comparison Quantities

In the background section, two inconsistencies between computed and measured data were
mentioned, which introduce some uncertainty into code-data comparisons. The first of these is that there is often not a one-to-one correspondence between available measured and computed dependent variables. This arises because all key parameters are not necessarily always measured in a test program. The second is that time and/or space coordinates in the NRS simulation may not be the same as those in the experiments.

From an automated code assessment standpoint, code and experimental data must both be available for a comparison. If not, a contributing figure-of-merit associated with the unavailable parameter cannot be constructed. This issue complicates the determination of the relative performance of a single code version applied to two similar facilities/tests if the same data is not available for the two. However, the main role of automated code assessment is to compare the relative performance of two similar versions of a code against a single facility/test or against each other. So the same figure-of-merit can always be consistently built for a given test.

The ultimate application of an automated code assessment procedure will be in a revalidation or licensing environment where an overall figure-of-merit is constructed from the weighted sums of those associated with individual code-data comparisons (see next section). It must be decided how to weigh, relative to each other, figures-of-merit from individual tests, similar in most regards except data availability. Though important, this issue is beyond the scope of the present work. ACAP provides a software framework for automated, tunable weighting of component figures-of-merit, while recommended strategies for specifying them is evolving.

The second code-data consistency issue mentioned above is within the scope of the present work. Specifically, data analysis modules of the automated code assessment procedure must accommodate, where needed, the differences between code and data space-time coordinates. This issue is relevant, of course, only to Types IV and V data where discretization choices and/or numerical stability issues will generally return NRS predictions of dependent variables at different locations in space-time than where the data was taken.

Of the techniques analyzed, approximation theory and all of the basic statistical analysis methods considered except correlation measures, do not inherently require that experiment and computation have coincident independent variables. Such consistency is however required for basic statistical analysis correlation measures, probabilistic time series data analysis methods and the multivariate methods considered. Also, valid application of some trend removal processes including running averages require independent variable consistency. Discrete Fourier transform and time-frequency methods can be more accurately deployed if samples are taken at the same time steps, but this is not a requirement of their implementation.

ACAP incorporates a resampling/interpolation pre-processor for bringing experimental and computed data to the same independent variable basis.

Subjectivity Removal – Automated Construction of Figures-of-Merit

So far, a number of techniques for quantifying code-data or code-code comparisons have been summarized. Most of these techniques have been installed within a “toolkit” of assessment modules in ACAP. The input to each of these modules is the data to be compared; the output from each of them is one or more figures-of-merit.

ACAP constructs one (or at most a few) overall figures-of-merit defining the fidelity of a suite of NRS code runs applied to the automated code assessment data base. As mentioned in the previous sections, defining the best way to construct overall figures-of-merit is beyond the scope of the present work. Rather, we focused on providing a general software framework for doing so.

In application, and automated code assessment run will involve extraction of multiple code-
data or code-code “raw” data sets. For each a number of data will be available, most generally several from each of Types I-V. For each of these x vs. y sets, one or more comparison measures could be deployed, each returning a “local” figure-of-merit.

Willmott [1982] and Fox [1984] have both recommended that multiple difference based statistical accuracy indices should be presented when reporting model (i.e., simulation) performance. Within the scope of the present work, the authors have accommodated this philosophy by implementing a general performance figure-of-merit weighting construct in ACAP. Specifically, a single figure-of-merit for a give simulation can be constructed from an arbitrarily (i.e., user specified or “canned”) weighted sum of several statistical accuracy measures. Removal of subjectivity is achieved once this figure-of-merit construction is frozen.

The overall figures-of-merit constructed in this process are to be interpreted as relative performance measures. These can then form the basis of acceptance/rejection tests in code revalidation. As relative measures, they must accommodate the basic requirement that superior solutions yield superior figures-of-merit. It has been observed above that a given figure-of-merit may or may not satisfy this basic “sanity” check depending on the application. The authors anticipate that a good deal of the effort involved in developing a robust figure-of-merit assembly procedure will be focused on satisfying this requirement.

6. CONCLUSION AND SUMMARY RECOMMENDATIONS

A number of mathematical data analysis methods have been surveyed for their applicability in the construction of NRS code-data and code-code comparison measures. The goal of the survey was to identify issues and techniques to be considered in the development of an automated code assessment procedure, ACAP, to be brought to bear in USNRC advanced T/H code consolidation efforts. Techniques from the overlapping fields of approximation theory, time-series data analysis, basic statistical analysis, as well as several other methods have been considered. Several techniques were demonstrated using example NRS code-data sets.

A number of conclusions apply:
1. Most of the methods considered can be applied to provide useful quantitative measures of accuracy for at least a subset of NRS data Types III, IV and V.
2. Inappropriate use of some methods can yield incorrect results, that is, return figures-of-merit that are worse for more accurate simulations. This motivates:
   • Definition of a robust comparison measure or suite of measures as one that reliably returns better figures-of-merit for superior comparisons and worse figures-of-merit for inferior comparisons
   • That great care be taken in the selection of the suite of analysis tools chosen for each particular comparison
3. The inherent limitations to stationary data of most available methods render straightforward application to NRS Type V data less than rigorous. Trend removal techniques can be brought to bear to preprocess the data, thereby yielding more robust comparison measures, especially when deployed in concert with time-windowing.
4. Experimental uncertainty can be effectively incorporated in code-data accuracy assessment within the framework of the “toolkit” of analysis procedures considered. Experimental uncertainty should be included with the “raw” experimental data in the code reassessment test matrix.
5. Inconsistency between the computed and measured independent variable range and basis (i.e. different time steps) motivates the incorporation of resampling and range trimming condition-
ers within ACAP. Such “synchronization” is required for most comparisons.

6. For Type V data, techniques that are intrinsically appropriate for non-stationary data analysis can be utilized in the construction of comparison measures. These include best approximation fits and, most promising in the view of the present investigators, time-frequency techniques.

7. There is a fundamental lack of rigor in applying basic statistical analysis procedures to most NR systems data. This arises due to non-stationarity of the data and the unavailability of a known distribution of error about its mean. This renders the construction of statistical inference measures suspect at best. Basic statistical difference and correlation measures can be deployed to construct useful figures-of-merit, but uncertainty bounds should not be inappropriately constructed.

8. As indicated in conclusion 2 above, great care must be taken in deploying comparison measures. In particular, for each experimental data set, a demonstrably robust assessment strategy must be developed. The present investigators feel that this requirement defines a process whereby expert assessors “calibrate” and document a suite of robust data analyses for each experimental data set in the code reassessment matrix. This assessment configuration will in general include preconditioning strategies, data comparison measures, figure-of-merit weighting assembly factors, and should be included with the “raw” experimental data in the reassessment matrix. Such configured assessments will then be used to define ACAP sessions in future code re-assessments.

In concert with the method assessment findings in this paper, a set of baseline techniques for code-data (or code-code) comparisons, data preconditioning, figure-of-merit-assembly and incorporation of experimental uncertainty were selected and implemented in ACAP. Details of these and a more general overview of the software are available in Kunz et. al., [1998a, 1998b, 2000].

7. REFERENCES

No. 4.
Figure 1a. Sample NRS data Type I. Key parameters tables (from Bessette and Odar [1986]).

Figure 1b. Sample NRS data Type II. Timing of events table (from Jo and Connell [1985]).

Figure 1c. Sample NRS data Type III. Scatter plot of nominally 0-D data (from Shumway [1995]).

Figure 1d. Sample NRS data Type IV. 1-D (in space) steady state data (from Shumway [1995]).

Figure 1e. Sample NRS data Type V. Time record data (from Lee and Rhee [1997]).
Figure 2a. Comparison of measured and RELAP5 predicted vessel pressure vs. time for the NRC12 case (from Lee and Rhee [1997]).

Figure 2b. Comparison of several best approximation fits to the absolute error associated with data in Figure 2a.

Figure 3a. Polynomial fit to measured (closed circles) and two RELAP5 predicted (open symbols) heat flux distributions for MIT-Siddique test data (digitized from Shumway [1995]).

Figure 3b. Cubic spline fit to measured (closed circles) and two RELAP5 predicted (open symbols) heat flux distributions for MIT-Siddique test data (digitized from Shumway [1995]).
Figure 4. a) Segment of measured and RELAP5 predicted vessel pressure vs. time for the NRC12 case (from Lee and Rhee [1997]). b) Autocorrelation of experimental and computed time series, and approximate MA model of computed time trace.

Figure 5a. D’Auria artificial code assessment data, digitized from Ambrosini et al. [1990].

Figure 5b. D’Auria figure-of-merit computed from data appearing in Figure 5a.
Figure 6a. UCB wall condensation test data from Shumway [1995]. Computed vs. measured wall heat flux. Default RELAP5 diffusion model. Lines correspond to L_2-standard (solid), L_2-constrained (dotted) and Perfect agreement (dashed).

Figure 6b. UCB wall condensation test data from Shumway [1995]. Computed vs. measured wall heat flux. New RELAP5 diffusion model. Lines correspond to L_2-standard (solid), L_2-constrained (dotted) and Perfect agreement (dashed).

Figure 6c. UCB wall condensation test data from Shumway [1995]. Absolute error vs. measured wall heat flux. Default RELAP5 diffusion model.

Figure 6d. UCB wall condensation test data from Shumway [1995]. Absolute error vs. measured wall heat flux. New RELAP5 diffusion model.

Figure 6e. UCB wall condensation test data from Shumway [1995]. PDF of absolute error. Default RELAP5 diffusion model.

Figure 6f. UCB wall condensation test data from Shumway [1995]. PDF of absolute error. New RELAP5 diffusion model.
Figure 7. a) Segment of measured and TRAC-B predicted rod temperature vs. time for FLECHT SBLOCA test (Paige [1998]). b) Absolute error associated with data in a). c) Running average trends and the absolute error associated with the data in a). d) Residuals and the absolute error associated with data in a), c). e) Time windows defined for data in a). f) Running average trends and the absolute error associated with the data in a) with separate running averages applied to transition and reflood time windows. g) Residuals and the absolute error associated with data in a), f).
Figure 8.  
a) Measured vessel pressure vs. time for the NRC12 case (from Lee and Rhee [1997]).
b) Short time Fourier transform spectrogram of data in Figure 8a.
c) Morlet continuous wavelet transform spectrogram of data in Figure 8a.
Figure 8d. Continuous wavelet representation of solution accuracy applied to D’Auria’s artificial data appearing in Figure 5a.
Figure 9. TRAC-B predicted rod temperature vs. time for FLECHT SBLOCA test at multiple axial locations in core (Paige [1998]).
Figure 10. a) RELAP5 predicted and measured integrated mass flow through an ADS vs. time for NRC12 case (from Lee and Rhee [1997]). Linear “artificial” data and experimental uncertainty band also plotted.

b) absolute error associated with data in Figure 10a.
Figure 11a. Artificial data, simulations and experimental uncertainty band.
Figure 11. b) Trends for artificial data traces in Figure 11a. c) Residuals for artificial data traces in Figure 11a.
Figure 11d. Discrete Fourier transform of artificial data in Figure 11a and the absolute error of the amplitudes.