Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Computational Physics 217 (2006) 5–36
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Measures of agreement between computation and experiment: Validation metrics

William L. Oberkampf [a,*], Matthew F. Barone [b,1]

[a] Validation and Uncertainty Quantification Department, Mailstop 0828, P.O. Box 5800, Sandia National Laboratories, Albuquerque, NM 87185-0828, United States
[b] Aerosciences and Compressible Fluid Mechanics Department, Mailstop 0825, P.O. Box 5800, Sandia National Laboratories, Albuquerque, NM 87185, United States

Received 26 August 2005; received in revised form 17 March 2006; accepted 30 March 2006
Available online 12 June 2006

Abstract

With the increasing role of computational modeling in engineering design, performance estimation, and safety assessment, improved methods are needed for comparing computational results and experimental measurements. Traditional methods of graphically comparing computational and experimental results, though valuable, are essentially qualitative. Computable measures are needed that can quantitatively compare computational and experimental results over a range of input, or control, variables to sharpen assessment of computational accuracy. This type of measure has been recently referred to as a validation metric. We discuss various features that we believe should be incorporated in a validation metric, as well as features that we believe should be excluded. We develop a new validation metric that is based on the statistical concept of confidence intervals. Using this fundamental concept, we construct two specific metrics: one that requires interpolation of experimental data and one that requires regression (curve fitting) of experimental data. We apply the metrics to three example problems: thermal decomposition of a polyurethane foam, a turbulent buoyant plume of helium, and compressibility effects on the growth rate of a turbulent free-shear layer. We discuss how the present metrics are easily interpretable for assessing computational model accuracy, as well as the impact of experimental measurement uncertainty on the accuracy assessment.
Published by Elsevier Inc.

Keywords: Validation of computational models; Assessment of model accuracy; Model credibility; Uncertainty quantification; Experimental uncertainty; Fluid dynamics; Solid dynamics

1. Introduction

It is common practice in all fields of engineering and science for comparisons between computational results and experimental data to be made graphically. The graphical comparisons are usually made by plotting some

* Corresponding author. Tel.: +1 505 844 3799; fax: +1 505 844 4523.
E-mail addresses: wloberk@sandia.gov (W.L. Oberkampf), mbarone@sandia.gov (M.F. Barone).
[1] Tel.: +1 505 284 8686; fax: +1 505 844 4523.

0021-9991/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.jcp.2006.03.037

Nomenclature

C	confidence level chosen, $C = 100(1-\alpha)\%$
$\left\|\frac{\mathrm{CI}}{\bar{y}_e}\right\|_{\mathrm{avg}}$	average confidence indicator associated with the average of the absolute value of the relative estimated error over the range of the experimental data, see either Eq. (19) or (26)
$\left\|\frac{\mathrm{CI}}{\bar{y}_e}\right\|_{\max}$	confidence interval associated with the maximum absolute value of the relative estimated error over the range of the experimental data, see either Eq. (21) or (27)
E	true error of the computational model as compared to the true mean of the experimental measurements, $y_m - \mu$
$\widetilde{E}$	estimated error of the computational model as compared to the estimated mean of the experimental measurements, $y_m - \bar{y}_e$
$\left\|\frac{\widetilde{E}}{\bar{y}_e}\right\|_{\mathrm{avg}}$	average of the absolute value of the relative estimated error over the range of the experimental data, see Eq. (18)
$\left\|\frac{\widetilde{E}}{\bar{y}_e}\right\|_{\max}$	maximum of the absolute value of the relative estimated error over the range of the experimental data, see Eq. (20)
$F(\nu_1, \nu_2, 1-\alpha)$	F probability distribution, where ν_1 is the first parameter specifying the number of degrees of freedom, ν_2 is the second parameter specifying the number of degrees of freedom, and $1-\alpha$ is the quantile for the confidence interval chosen
n	number of sample (experimental) measurements
s	sample (estimated) standard deviation based on n experimental measurements
SRQ	system response quantity
t_ν	t distribution with ν degrees of freedom, $\nu = n-1$
$t_{\alpha/2\nu}$	$1-\alpha/2$ quantile of the t distribution with n degrees of freedom, $\nu = n-1$
$\bar{y}_e$	sample (estimated) mean based on n experimental measurements
y_{m}	mean of the SRQ from the computational model
α	arbitrarily chosen total area from both tails of the specified distribution
μ	population (true) mean from experimental measurements
$\vec{\theta}$	vector of coefficients of the chosen regression function, Eq. (22)
$\hat{\vec{\theta}}$	vector of regression coefficients that minimize the error sum of squares, Eq. (24)

computational system response quantity (SRQ) with the experimentally measured response over a range of some input parameter. If the computational results generally agree with the experimental data, the computational model is commonly declared, ''validated''. Comparing computational results and experimental data on a graph, however, is only incrementally better than making a qualitative comparison. With a graphical comparison, one rarely sees quantification of numerical solution error or quantification of computational uncertainties, e.g., due to variability in modeling parameters, missing initial conditions, or poorly known boundary conditions. In addition, an estimate of experimental uncertainty is not typically quoted, nor its statistical character quantified. A graphical comparison also gives little quantitative indication of how the agreement between computational results and experimental data varies over the range of the independent variable, e.g., a spatial coordinate, time, or Mach number. Further, a simple graphical comparison is ill suited for the purpose of quantitative validation because statistical methods are needed to quantify experimental uncertainty. It should be noted that some journals, such as those published by the American Institute of Aeronautics and Astronautics (AIAA) and the American Society of Mechanical Engineers (ASME), now require improved statements of numerical accuracy and experimental uncertainty.

The increasing impact of computational modeling on engineering system design has recently resulted in an expanding research effort directed toward developing quantitative methods for comparing computational and experimental results. In engineering and physics, the form of the computational models is predominantly given by partial differential equations (PDEs) with the associated initial conditions and boundary conditions. Although statisticians have developed methods for comparing models (or ''treatments'') of many sorts, their emphasis has been distinctly different from the modeling accuracy assessment perspective in engineering. Much

of the recent work has been conducted as part of the Department of Energy's Advanced Simulation and Computing (ASC) Program. Refs. [1,2] argue that quantification of the comparison between computational and experimental results should be considered as the evaluation of a computable measure or a variety of appropriate measures. They refer to these types of measures as a validation metric and recommend that both uncertainties and errors should be quantified in the comparison of computational and experimental results. The input data to the metric are the computational results and the experimental measurements of the same SRQ of interest. Uncertainties refer to quantities that are either a random variable, e.g., random measurement uncertainty in experiments, or unknown quantities due to lack of knowledge, e.g., a boundary condition not measured in an experiment but needed for input to the computational model. Errors are usually due to numerical solution inaccuracies, such as lack of spatial grid convergence and lack of time-step resolution in unsteady phenomena.

This paper develops a validation metric based on the concept of statistical confidence intervals. In Section 2, we review the terminology of verification and validation by distinguishing between code verification, solution verification, validation metrics, model calibration, and adequacy of a model for its intended use. We briefly review the perspectives of hypothesis testing in statistics, Bayesian statistical inference, and the recent engineering perspective in validation metrics. In Section 3, we recommend features that should be incorporated, or addressed, in validation metrics. We discuss our perspective for constructing our confidence interval-based validation metrics and situations where we believe our metrics may or may not be useful. In Section 4, we review some of the basic ideas of statistical confidence intervals and construct a simple validation metric for the case of the SRQ at one operating condition. We apply this metric to an example of thermal decomposition of a polyurethane foam. Section 5 extends the fundamental idea of the validation metric to the case where the SRQ is measured in fine increments over a range of the input parameter. These increments allow us to construct an interpolation function of the experimental measurements over the range of the input parameter. We apply this metric to the example of a turbulent buoyant plume of helium. In Section 6, we develop the metric for the situation where the experimental data are sparse over the range of the input parameter. This very common engineering situation requires regression (curve fitting) of the data. We apply this metric to the example of compressibility effects on the growth rate of a planar turbulent shear layer. Section 7 provides some observations on the present contribution and makes recommendations for future work.

2. Review of the literature

2.1. Review of the terminology and processes

The terms "verification" and "validation" have a wide variety of meanings in the various technical disciplines. The AIAA, through the computational fluid dynamics (CFD) committee on standards [3], the work of Roache [4–6], and Refs. [2,7], has played a major role in attempting to standardize the terminology in the engineering community. This paper will use the AIAA definitions [3].

2.1.1. Verification

The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model.

2.1.2. Validation

The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.

The definition of verification makes it clear that verification addresses the accuracy of the numerical solution produced by the computer code as compared to the exact solution of the conceptual model. In verification, how the conceptual model relates to the "real world" is *not* an issue. As Roache [5] stated, "Verification deals with mathematics". Validation addresses the accuracy of the conceptual model as compared to the "real world", i.e., experimental measurements. As Roache [5] further stated, "Validation deals with physics".

Verification is composed of two types of activities: code verification and calculation verification. Code verification deals with assessing: (a) the adequacy of the numerical algorithms to provide accurate numerical solutions to the PDEs assumed in the conceptual model; and (b) the fidelity of the computer programming to

implement the numerical algorithms to solve the discrete equations. (see Refs. [2,5,7–9] for further discussion of code verification.)

Calculation verification deals with the quantitative estimation of the numerical accuracy of solutions to the PDEs computed by the code. The primary emphasis in calculation verification is significantly different from that in code verification because there is no known exact solution to the PDEs of interest. As a result, we believe calculation verification is more correctly referred to as *numerical error estimation*; that is, the primary goal is estimating the numerical accuracy of a given solution, typically for a nonlinear PDE with singularities, discontinuities, and complex geometries. For this type of PDE, numerical error estimation is fundamentally empirical (a posteriori), i.e., the conclusions are based on evaluations and analysis of solution results from the code. (see Refs. [5,10–14] for further discussion of numerical error estimation.)

As logical principles, code verification and numerical error estimation should be completed before model validation activities are conducted, or at least before actual comparisons of computational results are made with experimental results. The reason is clear. We should have convincing evidence that the computational results obtained from the code reflect the physics assumed in the models implemented in the code and that these results are not distorted or polluted due to coding errors or large numerical solution errors. Although the logic is clear concerning the proper order of activities, there are examples in the literature where coding or solution errors discovered after-the-fact invalidated the conclusions related to the accuracy or inaccuracy of the physics in the models being evaluated. Stated differently, if a researcher/analyst does not provide adequate evidence about code verification and numerical error estimation in a validation activity, the conclusions presented are of dubious merit. If conclusions from a defective simulation are used in high consequence system decision-making, disastrous results may occur.

Ongoing work by the ASME Standards Committee on Verification and Validation in Computational Solid Mechanics is attempting to clarify that model validation should be viewed as two steps [15]: (1) quantitatively comparing the computational and experimental results for the SRQ of interest, and (2) determining whether there is acceptable agreement between the model and the experiment for the intended use of the model. The first step in validation deals with accuracy assessment of the model, which we will refer to as evaluation of a validation metric.

Fig. 1 depicts several important aspects of validation, as well as issues of prediction and calibration of models. The left-center portion of Fig. 1 shows the first step in validation. The figure illustrates that the same SRQ must be obtained from both the computational model and the physical experiment. The SRQ can be any type

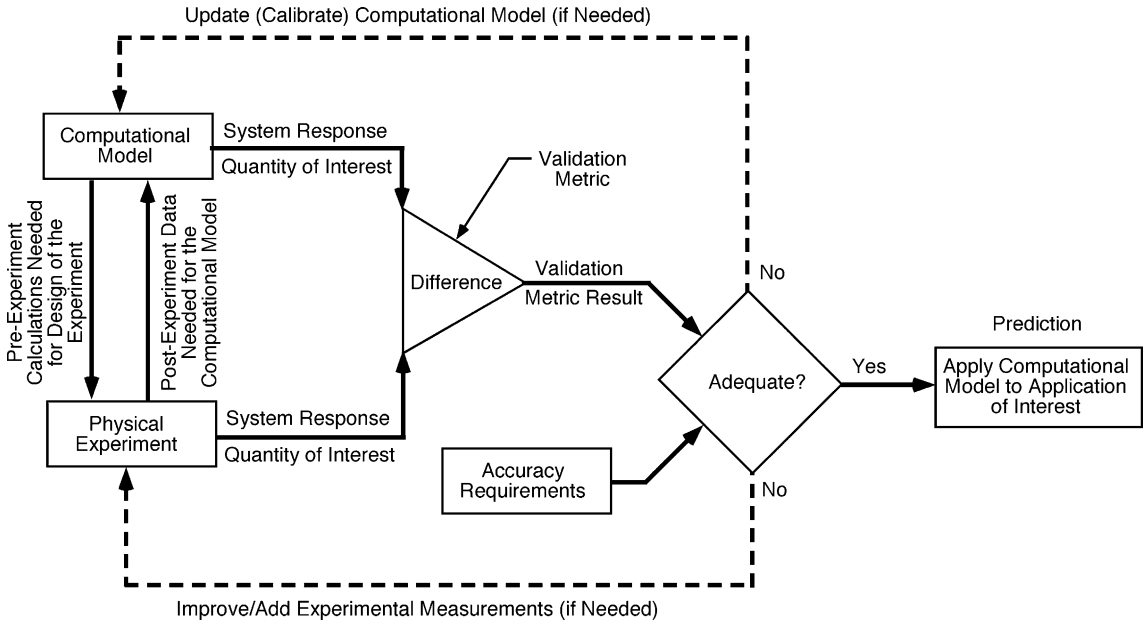

Fig. 1. Validation, calibration, and prediction [49]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

of physically measurable quantity, or it can be a quantity that is based on, or inferred from, measurements. For example, the SRQ can involve derivatives, integrals, or more complex data processing of computed or measured quantities such as the maximum or minimum of functionals over a domain. When significant data processing is required to obtain an SRQ, it is important to process both the computational results and the experimentally measured quantities in the same manner. The computational and experimental SRQs are input to the mathematical procedure, which can be considered as a difference operator, to compute a validation metric result. In this paper, when we refer to the "validation metric", we usually mean the mathematical procedure that operates on the computational and experimental SRQs. The SRQs are commonly one of two mathematical forms: (1) a deterministic quantity, i.e., a single value, such as a mean value or a maximum value over a domain; or (2) a probability measure, such as a probability density function or a cumulative distribution function. Each of these two forms can be functions of a parameter or multiple parameters in the computational model, such as a temperature or a Mach number; a function of spatial coordinates, such as Cartesian coordinates (x, y, z); or a function of both space and time. If both the computational and experimental SRQs are deterministic quantities, the validation metric will also be a deterministic quantity. If either of the SRQs is a probability measure, the result of the validation metric would also be a probability measure.

Another feature that should be stressed in Fig. 1 is the appropriate interaction between computation and experimentation that should occur in a validation experiment. To achieve the most value from the validation experiment, there should be in-depth, forthright, and frequent communication between computationalists and experimentalists during the planning and design of the experiment. Also, after the experiment has been completed, the experimentalists should measure and provide to the computationalists all the important input quantities needed to conduct the computational simulation. Examples of these quantities are actual freestream conditions attained in a wind-tunnel experiment (versus requested conditions), as-fabricated model geometry (versus as-designed), and actual deformed model geometry due to aerodynamics loads and heating. What should *not* be provided to the computationalists in a rigorous validation activity is the measured SRQ. Stated differently, it is our view that a blind computational prediction be compared with experimental results so that a true measure of predictive capability can be assessed in the validation metric. For an extensive discussion of the philosophical viewpoint, planning, design, execution, and analysis of validation experiments, see Refs. [2,11,16–20].

The second step in validation deals with comparing the validation metric result with the accuracy requirements for the intended use of the model. That is, validation, from a practical or engineering perspective, is *not* a philosophical statement of truth. The second step in validation, depicted in the right-center portion of Fig. 1, is an engineering decision that is dependent on the accuracy requirements for the intended use of the model. Accuracy requirements are, of course, dependent on many different kinds of factors. Some examples of these factors are: (a) the complexity of the model, the physics, and the engineering system of interest; (b) the difference in hardware and environmental conditions between the engineering system of interest and the validation experiment; (c) the increase in uncertainty due to extrapolation of the model from the validation conditions to the conditions of the intended use; (d) the risk tolerance of the decision makers involved; and (e) the consequence of failure or underperformance of the system of interest. Although the uncertainty estimation methodology and risk assessment issues involved in the second step are critically important in the application of a computational model for its intended use, these issues are beyond the scope of this paper. Here, we deal only with the first step in validation: validation metrics.

2.2. Review of approaches

Traditional approaches for quantitatively comparing computational and experimental results can be divided into three categories (here, we exclude graphical comparisons). First, in the 1960s the structural dynamics community began developing sophisticated techniques for assessing agreement between computational and experimental results, as well as techniques for improving agreement. These latter techniques are commonly referred to as parameter estimation, model parameter updating, or system identification. Two recent texts that provide an excellent discussion of this topic are Refs. [21,22]. In the approach followed by the structural dynamics community, certain model input parameters are considered as deterministic (but poorly known) quantities that are estimated by a numerical optimization procedure so that the best agreement between computational and experimental results can be obtained for a single SRQ or a group of SRQs. Multiple solutions of the computational

model are required to evaluate the effect of different values of the model parameters on the SRQ. Although these techniques are used to compare computational and experimental results, their primary goal is to improve agreement based on newly obtained experimental data.

The second approach is hypothesis testing or significance testing [23,24]. Hypothesis testing is a well-developed statistical method of deciding which of two contradictory claims about a model, or a parameter, is correct. In hypothesis testing the validation assessment is formulated as a ''decision problem'' to determine whether or not the computational model is consistent with the experimental data. The level of consistency between the model and the experiment is stated as a probability, based on what has been observed in comparing SRQs from the model and the experiment. This technique is regularly used in the operations research community for comparing mutually exclusive models. Hypothesis testing has recently been used in a model validation setting by Refs. [25–30]. Two features of this recent work are noteworthy. First, a validation metric is not specifically computed as a stand-alone measure that indicates the level of agreement or disagreement between computational and experimental results. The result of a hypothesis test is focused, instead, on obtaining a yes–no statement of computational-experimental consistency for a pre-specified level of significance. Second, this work deals with an SRQ from the computational model that is represented as a probability distribution. That is, multiple realizations of the SRQ are computed from the model using sampling techniques, such as Monte Carlo sampling, and then the ensemble of these realizations is compared with the ensemble of experimental measurements.

The third approach is the use of Bayesian analysis or Bayesian statistical inference [31–33]. Bayesian analysis has received a great deal of attention during the last two decades from statisticians, risk analysts, and some physicists and structural dynamicists. Although the process is rather involved, Bayesian analysis can be summarized in three steps. Step 1 is to construct, or assume, a probability distribution for each input quantity in the computational model that is chosen to be a random variable. Step 2 involves conditioning, or updating, the previously chosen probability models for the input quantities based on comparison of the computational and experimental results. To update the probability models, one must first propagate input probability distributions through the computational model to obtain probability distributions for the SRQs commensurate with those measured in the experiment. The updating of the input probability distributions, using Bayes equation to obtain posterior distributions, commonly assumes that the computation model is correct, i.e., the updating is conditional on the correctness of the computational model. Step 3 involves comparing new computational results with the existing experimental data or any new experimental data that might have been obtained. The new computational results are obtained by propagating the updated probability distributions through the computational model. Much of the theoretical development in Bayesian estimation has been directed toward optimum methods for updating statistical models of uncertain parameters in the computational model. In validation metrics, however, the emphasis is on methods for assessing the fidelity of the physics of the *existing* computational model. Although many journal articles have been published on the topic of Bayesian inference, the recent work of Refs. [34–40] is noteworthy.

From this very brief description of parameter estimation and Bayesian inference, it should be clear that the primary goal of both approaches is ''model updating'' or ''model calibration''. Although this goal is appropriate and necessary in many situations, it is a clearly different goal from that used to evaluate a validation metric. Our emphasis in validation metrics is in blind assessment of the predictive capability of a computational model (how good is the model?), as opposed to optimizing the agreement between a given model and experimental measurements. Fig. 1 depicts the goal of model calibration as the dashed-line upper feedback loop. In the figure, the loop is taken if the model does not adequately meet the specified accuracy requirements. It should also be noted that the upper feedback loop can also be taken even if the model is adequate. In such a case one wants to incorporate the latest experimental information into the model and not waste valuable information obtained from the experiment. The lower feedback loop in Fig. 1 could be taken if improvements or changes are needed in the experimental measurements or if additional experiments are needed to reduce experimental uncertainty.

Several researchers have taken approaches that differ from the three just mentioned; however, such approaches exhibit a common characteristic. Refs. [2,41–49] focus only on comparing a deterministic value of the SRQ from the computational model with the experimental data. That is, they do not propagate uncertain input parameters through the computational model to obtain multiple realizations or an ensemble of

SRQs. Oberkampf and Trucano [2] compute a validation metric for the case of multiple experimental measurements over a range of the input parameter. They assume the experimental measurement error is given by a normal (Gaussian) distribution, and they scale their validation metric result over the range from zero to unity. A value near zero occurs when there is a very large difference between computational and experimental results, and a value near unity occurs when nearly perfect agreement occurs. The precise implication of values between zero and unity is, of course, open to interpretation.

3. Construction of validation metrics

3.1. Recommended features of validation metrics

We believe that validation metrics should include several intuitive properties that would make them useful in an engineering and decision-making context. Extending the ideas of Refs. [2,7,49], the following is a list of conceptual properties that we believe a validation metric should satisfy:

(1) A metric should either: (a) explicitly include an estimate of the numerical error in the SRQ of interest resulting from the computational simulation or (b) exclude the numerical error in the SRQ of interest only if the numerical error was previously estimated, by some reasonable means, to be small. The primary numerical error of concern here is the error due to lack of spatial and/or temporal resolution in the discrete solution. Numerical error could be explicitly included in the validation metric, such as inclusion of an upper and a lower estimated bound on the error in the SRQ of interest. Although explicit inclusion of the numerical error in the metric seems appealing, it would add significant complexity to the theoretical derivation, calculation, and interpretation of the metric. By estimating beforehand that the numerical error is small, one can eliminate the issue from the calculation and interpretation of the metric. Taking this latter approach, the numerical error should be judged small in comparison to the estimated magnitude of the experimental uncertainty.

(2) A metric should be a quantitative evaluation of predictive accuracy of the SRQ of interest, including all of the combined modeling assumptions, physics approximations, and previously obtained physical parameters embodied in the computational model. Stated differently, the metric evaluates the aggregate accuracy of the computational model for a specific SRQ. Consequently, there could be offsetting errors or widely ranging sensitivities in the model that could show very accurate results for one SRQ, but poor accuracy for a different SRQ. If there is interest in evaluating the accuracy of submodels or the effect of the accuracy of individual input parameters within the computational model, one should conduct a sensitivity analysis of the SRQ. However, sensitivity analysis is a separate issue from constructing a validation metric.

(3) A metric should include, either implicitly or explicitly, an estimate of the error resulting from postprocessing of the experimental data to obtain the same SRQ that results from the computational model. Examples of the types of postprocessing of experimental data are as follows: (a) the construction of a regression function, e.g., least-squares fit, of the data to obtain a continuous function over a range of an input (or control) quantity; (b) the processing of experimental data that are obtained on a very different spatial or temporal scale than what is modeled in the computational model; and (c) the use of complex mathematical models of the physically measured quantities to process the experimental data. A case where the postprocessing described in Example (b) might be necessary is when there are localized underground measurements of a pollutant concentration and the computational model contains a large-scale, spatially averaged permeability model. One might require the type of postprocessing defined in Example (c) when very similar models of the physics in the computational model are also needed to process and interpret the experimental data. Note that in the recommended Property (2) mentioned above, any error associated with the postprocessing of the numerical solution of PDEs should be considered as part of the error in the computational model.

(4) A metric should incorporate, or include in some explicit way, an estimate of the measurement errors in the experimental data for the SRQ that are the basis of comparison with the computational model. The possible sources for measurement errors depend on a very wide range of issues, but a discussion of these

is clearly beyond the scope of this paper [50,51]. However, measurement errors are commonly segregated into two types: bias (systematic) errors and precision (random) errors. At a minimum a validation metric should include an estimate of precision errors, and, to the extent possible, the metric should also include an estimate of bias errors. The most practical method of estimating a wide range of bias errors is to use design-of-experiment techniques to transform them into precision errors so that statistical procedures can be used [18,24,51].

(5) A metric should depend on the number of experimental measurements that are made of a given SRQ of interest. The number of measurements can refer to a number of situations: (a) multiple measurements made by the same investigator using the same experimental diagnostic technique and the same experimental facility, (b) multiple investigators using different facilities and possibly different techniques, and (c) multiple measurements of a given SRQ over a range of input quantities (or levels) for the SRQ. The reason for including this issue in our recommendations is to stress the importance of multiple measurements in estimating the accuracy of the experimental result. We contrast our recommendation with the situation where one experimental measurement is made of an SRQ and then the experimental uncertainty is estimated based on many assumptions, such as previous experience concerning the error sources and the interaction and propagation of contributing error sources through the data reduction process. One measurement with an estimated uncertainty band has much less credence than multiple measurements, particularly when the multiple measurements vigorously seek to identify possible sources of error in the measurements or they are from independent sources. (See the classic paper by Youden [52] and the discussion by Morgan and Henrion [53] concerning the consistent tendency to underestimate experimental uncertainty.)

(6) A metric should *exclude* any indications, either explicit or implicit, of the level of adequacy in agreement between computational and experimental results. Examples of the level of adequacy that have been improperly used, in our view, in validation metrics are: (a) comparisons of computational and experimental results that yield value judgments, such as "good" or "excellent"; and (b) computational results that are judged to be adequate if they lie within the uncertainty band of the experimental measurements. We have stressed this issue in Section 2.1, particularly with regard to Fig. 1, and in Section 2.2. Validation metrics should be measures of agreement, or disagreement, between computational models and experimental measurements; issues of adequacy or satisfaction of accuracy requirements should remain separate from the metric.

Although these six conceptual properties in a validation metric seem intuitive, the published literature demonstrates a wide variety of views regarding what a validation metric should embody and how that metric should be interpreted. Refs. [1,2] propose a metric that satisfies all of these properties. Their metric took the approach of combining Properties 2, 3, and 4 above into one mathematical quantity: the metric itself. Specifically, they combined the measure of agreement between computational and experimental results, the estimate of experimental uncertainty, and the number of experimental replications into a single expression for the metric. Although this is a reasonable approach, the present authors have concluded that combining all three properties into the same quantity is not the best approach. The present approach constructs a validation metric that separates the accuracy estimation of the computational model from the level of confidence in estimation of the accuracy. Note that hypothesis testing combines these two issues, accuracy and confidence, into one measure: a probability measure.

3.2. Perspectives of the present approach

The present approach assesses the accuracy of the model based on comparing deterministic computational results with the estimated mean of the experimental measurements. The primary differences in the present perspective, and most of the work cited above, are that: (a) a stand-alone validation metric is constructed to provide a compact, statistical measure of quantitative disagreement between computational and experimental results; and (b) a statistical confidence interval is computed that reflects the confidence in the accuracy of the experimental data. We concur with Ref. [7] that such a validation metric would be most effective in moving beyond the "viewgraph norm" mode of comparing computational and experimental results so that quantita-

tive statements of accuracy can be made. This type of metric would be useful for situations in which a computational analyst, a model developer, or competing model developers are interested in quantifying which model among alternate models is most accurate for a given set of experimental data. In addition, this type of metric would be useful to a design engineer or a project engineer for specifying model accuracy requirements in a particular application domain of the model. It should be noted that if the application domain is outside the experimental measurement domain, one must account for the additional uncertainty of extrapolation of the model. Although we recognize that the extrapolation procedure should be dependent on both the error structure and the uncertainty structure in the validation domain, how this extrapolation should be accomplished is a complex, and unresolved, issue.

The primary reason for our interest in deterministic computational results, as opposed to the approach of propagating computational input uncertainties to determine output uncertainties in the SRQ, is the much-lower computational costs involved in deterministic simulations. Many computational analysts argue that computational resources are not available to provide both spatially and temporally resolved solutions, as well as nondeterministic solutions, for complex simulations. Risk assessment of high-consequence systems, for example, safety of nuclear power reactors and underground storage of nuclear waste, has shown that with an adequate, but not excessive, level of physical modeling detail, one *can* afford the computational costs of nondeterministic simulations. However, we recognize that there is substantial resistance in many fields to attain both grid-resolved and nondeterministic simulations. Consequently, we believe there is a need to construct validation metrics that require only deterministic computational results. As the presently resistant fields mature further, we believe validation metrics will be constructed that compare probability distributions of the SRQ from the computational model with probability distributions from the experimental results.

The validation metrics developed here are applicable to SRQs that do not have a periodic character and that do not have a complex mixture of many frequencies. For example, the present metrics would not be appropriate for analysis of standing or traveling waves in acoustics or structural dynamics. Another example of an inappropriate use would be the time-dependent fluid velocity at a point in turbulent flow. These types of SRQs require sophisticated time-series analysis and/or mapping to the frequency domain. Validation metrics constructed by Geers [41], Russell [42,43], and Sprague and Geers [45] are better suited to periodic systems or responses with a combination of many frequencies.

4. Validation metric for one condition

4.1. Development of the equations

In this section, the fundamental ideas of the present validation metric are developed for the case where the SRQ of interest is defined for a single value of an input or operating-condition variable. This will allow some discussion of how the present approach implements the recommended conceptual properties mentioned previously, as well as give an opportunity to review the classical development of statistical confidence intervals. Since it may be confusing why we begin the development of validation metrics with a discussion of confidence intervals, we make the following point. We are interested in obtaining an error measure between a deterministic computational result and the mean of a population of experimental measurements for which only a finite sequence of measurements has been obtained. When this is grasped, it is realized that the key issue is the statistical nature of the sample mean of the measured system response, *not* the accuracy of the agreement between the computational result and the individual measurements. With this perspective, it becomes clear that the point of departure should be a fundamental understanding of the statistical procedure for estimating a confidence interval for the true mean. In traditional statistical testing procedures, specifically hypothesis testing, the point of departure is the derivation for the confidence interval of the difference between two hypotheses: the computational mean and the experimental mean.

4.1.1. Construction of a statistical confidence interval

A short review and discussion will be given for the construction of a statistical confidence interval. The development of confidence intervals is discussed in most texts on probability and statistics. The following development is based on the derivation by Devore [23], Chapter 7.

Let X be a random variable characterizing a population having a mean μ and a standard deviation σ. Let $x_1, x_2, \ldots, x_n$ be actual sample observations from the population, and these are assumed to be the result of a random sample $X_1, X_2, \ldots, X_n$ from the population. Let $\overline{X}$ be the sample mean, which is a random variable, based on the random sample $X_1, X_2, \ldots, X_n$. Provided that n is large, the central limit theorem implies that $\overline{X}$ has approximately a normal distribution, *regardless* of the nature of the population distribution. Then it can be shown that the standardized random variable

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{1}$$

has an approximate normal distribution with zero mean and a standard deviation of unity. S is the sample standard deviation, which is a random variable, based on random samples $X_1, X_2, \ldots, X_n$. It can also be shown, provided n is large, that the probability interval for Z can be written as

$$P(z_{-\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \tag{2}$$

where $z_{\alpha/2}$ is the value of the random variable Z at which the integral of Z from $z_{\alpha/2}$ to $+\infty$ is $\alpha/2$. Since Z is symmetrical and has its mean at zero, the integral of Z from $-\infty$ to $z_{-\alpha/2}$ is also equal to $\alpha/2$. The total area from both tail intervals of the distribution is α.

Eq. (2) can be rearranged to show that the probability interval for μ, the mean of the population that is the unknown quantity of interest, is given by

$$P\left(\overline{X} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha \tag{3}$$

Eq. (3) can be rewritten as a confidence interval, i.e., a probability interval, for the population mean using sampled quantities for the mean and standard deviation

$$\mu \sim \left(\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) \tag{4}$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively, based on n observations. Note that $\bar{x}$ and s are computed from the realizations $X_1 = x_1$, $X_2 = x_2$, …, $X_n = x_n$. The term $s/\sqrt{n}$ is the standard error of the sample mean that measures how far the sample mean is likely to be from the population mean. The level of confidence that μ is in the interval given by Eq. (4) can be shown to be $100(1 - \alpha)\%$. The value of α is arbitrarily assigned and is typically chosen to be 0.1 or 0.05, corresponding to confidence levels of 90% or 95%, respectively.

The confidence interval for the population mean can be interpreted in a strict frequentist viewpoint or in a subjectivist, or Bayesian, viewpoint. Let C be the confidence level chosen, i.e., $C = 100(1 - \alpha)\%$, for stating that the true mean μ is in the interval given by Eq. (4). The frequentist would state, "μ is in the interval given by Eq. (4) with probability C," which means that if the experiment on which μ is estimated is performed repeatedly, in the long run μ will fall in the interval given by Eq. (4) C % of the time. The subjectivist would state [54], "Based on the observed data, it is my belief that μ is in the interval given by Eq. (4) with probability C". The reason that it *cannot* be strictly stated that C is the probability that μ is in the interval given by Eq. (4) is that the true probability is either zero or one. That is, the true mean μ is either in the interval or it is not; we simply *cannot know with certainty* for a finite number of samples from the population. Not withstanding these fine points of interpretation, we will essentially use the subjectivist interpretation in a slightly different form than is presented above: μ is in the interval given by Eq. (4) with confidence C.

Now consider the case of estimating a confidence interval for an arbitrary number of experimental observations n, with n as small as two. It can be shown [23] that the equation analogous to Eq. (4) is

$$\mu \sim \left(\bar{x} - t_{\alpha/2,v} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,v} \cdot \frac{s}{\sqrt{n}}\right) \tag{5}$$

where the level of confidence is given by $100(1 - \alpha)\%$ and $t_{-\alpha/2,v}$ is the $1 - \alpha/2$ quantile of the t distribution for $v = n - 1$ degrees of freedom. For n greater than 16, the cumulative t distribution and the cumulative standard normal distribution differ by less than 0.01 for all quantiles. In the limit as $n \to \infty$, the t distribution approaches the standard normal distribution.

Eq. (5) is regularly used for hypothesis testing in classical statistical analysis. However, our perspective in the construction of validation metrics is notably different. We wish to quantify the difference between the computational results and the true mean of the experimental results. Stated differently, we wish to measure shades of gray between a computational model and an experiment – not make a "yes" or "no" statement about the congruence of two hypotheses.

4.1.2. Construction of a validation metric based on confidence intervals

As discussed with regard to Fig. 1, the input quantities that should be used in the computational simulation of the SRQ of interest are those that are actually realized in the validation experiment. Some of these input quantities from the experiment may not be known precisely for various reasons, for example: (a) a quantity may not have been specifically measured but was estimated by the experimentalist, taken from an engineering handbook of physical properties, or simply taken from a fabrication drawing of hardware to be used in the experiment; (b) a quantity may not have been specifically measured but is known to be a sample from a well-characterized population; and (c) a quantity in the experiment may not be controllable from one experiment to the next, but the individual realizations of the quantity are measured so that the population for the entire experiment could be fairly well characterized. If these input quantities are considered as random input variables to the computational model, the proper procedure is to propagate these uncertain quantities through the model to characterize the SRQ as a random variable. To avoid this computational cost, as discussed previously, it is commonly assumed that the mean value of the SRQ can be approximated by propagating *only* the mean, i.e., the expected value, of all uncertain input parameters through the computational model. This approach is accurate only for linear models, or non-linear models where very limited scatter is associated with the random variables. We will briefly discuss this assumption, however it is addressed in many texts on propagation of uncertain inputs through a model. (See, for example, Ref. [55].)

A Taylor series can be written that shows the approximation: Let Y_m be the random variable SRQ from the computational model; let $g(\cdot)$ represent the PDE with the associated initial conditions and boundary conditions that map uncertain inputs to the uncertain SRQ; and let χ_i, where $i = 1, 2, \ldots, n$, be the uncertain input random variables. Then, the Taylor series for uncorrelated input random variables can be expanded about the mean of each of the input variables, μ_{χ_i}, and written as [55]

$$\mathscr{E}(Y_m) = g(\mu_{\chi_1}, \mu_{\chi_2}, \ldots, \mu_{\chi_n}) + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial^2 g}{\partial \chi_i^2}\right)_{\mu_{\chi_i}} \mathrm{Var}(\chi_i) + \cdots \tag{6}$$

where $\mathscr{E}(Y_m)$ is the expected value, i.e., the mean, of the SRQ and $\mathrm{Var}(\chi_i)$ is the variance of the input variables. It is seen from Eq. (6) that the first term of the expansion is simply g evaluated at the mean of the input variables. The second term is the second derivative of g with respect to the input variables. This term, in general, will be small with respect to the first term if either: (a) g is nearly linear in the input variables or (b) the variance of all of the input variables is small. Linearity in the input variables essentially never occurs when the mapping of inputs to outputs is given by a differential equation, even a *linear* differential equation. Note that when using this approximation one could obtain poor agreement between computational and experimental results, and the cause is *not* the model per se, but the inaccuracy of the computational mean caused by the assumption of the propagation of the mean of the inputs. With this approximation clarified, we now move on to the construction of a validation metric.

For the validation metric we wish to construct, we are interested in two quantities. First, we want to estimate an error in the SRQ of the computational model based on the difference between the computational model and the estimated mean of the population based on the experimentally measured samples of the SRQ. Let y_m be the SRQ from the computational model, i.e., the first term of the series expansion given in Eq. (6). Changing the notation used previously for the experimental measurements from $\bar{x}$ to $\bar{y}_e$, we define the estimated error in the computational model as

$$\widetilde{E} = y_m - \bar{y}_e \tag{7}$$

where $\bar{y}_e$ is the estimated, or sample, mean based on n experiments conducted. $\bar{y}_e$ is given by

$$\bar{y}_e = \frac{1}{n}\sum_{i=1}^{n} y_e^i \tag{8}$$

where $y_e^1, y_e^2, \ldots, y_e^n$ are the individually measured results of the SRQ from each experiment.

Second, we wish to compute an interval that contains the true error, which we do not know, at a specified level of confidence. Let the true error E be defined as

$$E = y_m - \mu \tag{9}$$

where μ is the true mean of the population. Writing the confidence interval expression, Eq. (5), for μ as an inequality relation and changing the notation as just mentioned, we have

$$\bar{y}_e - t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}} < \mu < \bar{y}_e + t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}} \tag{10}$$

where s is the sample standard deviation given by

$$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_e^i - \bar{y}_e\right)^2\right]^{1/2} \tag{11}$$

Multiplying Eq. (10) by –1 and adding y_m to each term, we have

$$y_m - \bar{y}_e + t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}} > y_m - \mu > y_m - \bar{y}_e - t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}} \tag{12}$$

Substituting the expression for the true error, Eq. (9), into Eq. (12) and rearranging, one obtains

$$y_m - \bar{y}_e - t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}} < E < y_m - \bar{y}_e + t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}} \tag{13}$$

Substituting the expression for the estimated error, Eq. (7), into Eq. (13), we can write the inequality expression as an interval containing the true error where the level of confidence is given by $100(1-\alpha)\%$:

$$\left(\widetilde{E} - t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}}, \widetilde{E} + t_{\alpha/2,v}\cdot\frac{s}{\sqrt{n}}\right) \tag{14}$$

Using the traditional level of confidence of 90%, one can state the validation metric in the following way: the estimated error in the model is $\widetilde{E} = y_m - \bar{y}_e$ with a confidence level of 90% that the true error is in the interval

$$\left(\widetilde{E} - t_{0.05,v}\cdot\frac{s}{\sqrt{n}}, \widetilde{E} + t_{0.05,v}\cdot\frac{s}{\sqrt{n}}\right) \tag{15}$$

Three characteristics of this validation metric should be mentioned. First, the statement of confidence is made concerning an interval in which the true error is believed to occur. The statement of confidence is *not* made directly concerning the magnitude of the estimated error, nor concerning an interval around the computational prediction. The reason such statements cannot be made is that the fundamental quantity that is uncertain is the *true* experimental mean. Stated differently, although we are asking how much error there is in the computational result, the actual uncertain quantity is the *referent*, i.e., the true experimental value, *not* the computational result.

Second, the interval believed to contain the true error is symmetric around the estimated error. We can also state that the rate of decrease of the magnitude of the interval is a factor of 2.6 when going from two experiments to three experiments. For a large number of experiments, the rate of decrease of the magnitude of the interval is $1/\sqrt{n}$. Additionally, the size of the interval decreases linearly as the sample standard deviation decreases.

Third, for small numbers of experimental measurements the assumption must be made that the measurement uncertainty is normally distributed. Although this is a very common assumption in experimental uncertainty estimation, and probably well justified, it is rarely *demonstrated* to be true [50,51]. However, for a large number of experimental measurements, as discussed above, the confidence interval on the mean is valid regardless of the type of probability distribution representing measurement uncertainty.

As a final point in the development of our approach to validation metrics, we stress the primacy we give to the experimental data. As can be clearly seen from Eq. (9), the referent for the error measure is the experimental data, not the computational model or some type of weighted average between the computational model and the experimental data. However, our trust in the accuracy of experimental measurements is not without some risk, specifically, if an undetected bias error exists in the experimental data. (See, for example, Refs. [7,52,53] for further discussion.)

4.2. Example: thermal decomposition of foam

As an example of the application of the validation metric just derived, consider the assessment of a computational model for the rate of decomposition of a polyurethane foam due to thermal heating. The computational model solves the energy equation and is composed of three major components: (a) thermal diffusion through the materials involved, (b) chemistry models for the thermal response and decomposition of polymeric materials due to high temperature, and (c) radiation transport within the domain and between the boundaries of the physical system. The foam decomposition model predicts the mass and species evolution of the decomposing foam and was developed by Hobbs et al. [56]. Dowding et al. [28] computed the results for this example using the computer code Coyote which solves the mathematical model using a finite element technique [57]. Three-dimensional, unsteady solutions were computed until the foam decomposes, vaporizes, and escapes from the container. Solution verification for the computational results relied on the grid-refinement studies previously conducted by Hobbs et al. [56]. These earlier grid-refinement studies estimated that the mesh discretization error was less than 1% for the velocity of the foam decomposition front for mesh sizes less than 0.1 mm.

The experiment to evaluate the computational model was composed of a polyurethane foam enclosed in a stainless steel cylinder that was heated using high-intensity lamps (Fig. 2). The experiment was conducted by Bentz and Pantuso and is reported in Hobbs et al. [56]. The position of the foam-gas interface was measured as a function of time by X-rays passing through the cylinder. The steel cylinder was vented to the atmosphere to allow gas to escape, and it was heated in three directions: top, bottom, and side. For some of the experiments, a solid stainless steel cylinder or hollow aluminum component was embedded in the foam.

The SRQ of interest is the average velocity of the foam decomposition front when the front has moved between 1 and 2 cm. The SRQ was measured as a function of imposed boundary-condition temperature. Since we are only considering one operating condition for the present validation metric example, we pick the temperature condition of $T = 750$ °C because it had the largest number of experimental replications. Some of the

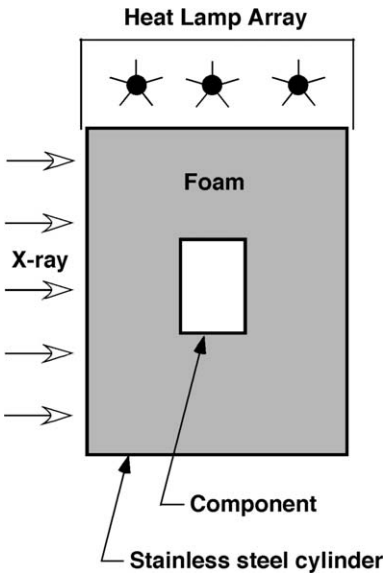

Fig. 2. Schematic of foam decomposition experiment [49]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

Table 1
Experimental data for foam decomposition, Ref. [56]

Experiment number	Temperature (°C)	Heat orientation	V (experiment) (cm/min)
2	750	Bottom	0.2323
5	750	Bottom	0.1958
10	750	Top	0.2110
11	750	Side	0.2582
13	750	Side	0.2154
15	750	Bottom	0.2755

replications, shown in Table 1, were the result of different orientations on the heat lamps. Computational simulations by Dowding et al. [28] showed that cylinder orientation had little effect on the velocity of the decomposition front. Since we are only interested in a single deterministic result from the code, we picked one of the Dowding et al. results for the computational SRQ. The result chosen for the computational SRQ was 0.2457 cm/min. With this approximation, we assigned the variability resulting from the heating orientation of the cylinder to uncertainty in the experimental measurements.

Using the data in Table 1 and Eqs. (5), (7), (8), (11), and (15), we obtain

$$\text{number of samples} = n = 6$$
$$\text{sample mean} = \bar{y}_e = 0.2314 \text{ cm/min}$$
$$\text{estimated error} = \widetilde{E} = 0.2457 - 0.2314 = 0.0143 \text{ cm/min}$$
$$\text{sample standard deviation} = s = 0.0303 \text{ cm/min}$$
$$\text{degrees of freedom} = n - 1 = \nu = 5$$
$$t \text{ distribution for 90\% confidence } (\nu = 5) = t_{0.05,\nu} = 2.015$$
$$\pm t_{0.05,\nu} \cdot \frac{s}{\sqrt{n}} = \pm 0.0249 \text{ cm/min}$$
$$\text{true mean with 90\% confidence} = \mu \sim (0.2065,\ 0.2563) \text{ cm/min}$$
$$\text{true error with 90\% confidence} \sim (-0.0106,\ 0.0392) \text{ cm/min}$$

Fig. 3 depicts the sample mean, the model mean, the estimated interval of the true mean, and the estimated error, with 90% confidence. In summary form, the result of the validation metric is $\widetilde{E} = 0.0143 \pm 0.0249$ cm/min with 90% confidence. Since the magnitude of the uncertainty in the experimental data is roughly twice the estimated error, one cannot make any more precise conclusions than ± 0.0249 cm/min (with 90% confidence) concerning the accuracy of the model. Whether the estimated accuracy, with its uncertainty, is adequate for the

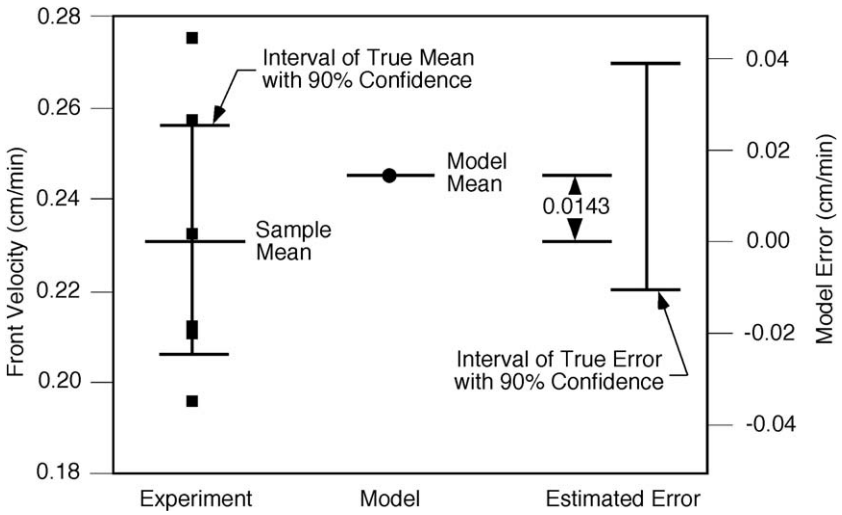

Fig. 3. Statistical and validation-metric results of foam decomposition.

intended use of the model is the second step in validation, as was discussed with regard to Fig. 1. If the estimated accuracy, with its uncertainty, is not adequate for a model-use decision, then one has two options. The first option, which is the more reasonable option for this case, is to reduce the experimental uncertainty in future measurements by obtaining additional experimental measurements or by changing the experimental procedure or diagnostic method to reduce the experimental uncertainty. The second option would be to improve, or update, the model so that it gives more accurate results. However, in the present case, the error in the model is small with respect to the experimental uncertainty. As a result, this option would make little sense.

5. Validation metric using interpolation

5.1. Development of the equations

We are now interested in the case where the SRQ is measured over a range of the input variable or the operating-condition variable. For example, in the foam decomposition experiment just discussed, we would be interested in the velocity of the foam decomposition front as a function of the heating temperature of the cylinder. Another example would be the thrust of a rocket motor as a function of burn time. Here we consider the case of one input variable while all others are held constant. This type of comparison is probably the most common between computational and experimental results. The present ideas could be extended fairly easily to the case of multiple input variables as long as the input variables were independent.

The following assumptions are made with regard to the computational results:

(1) The mean value of the SRQ is obtained by using the mean value of all uncertain input parameters in the computational model, i.e., the first term of the series expansion given in Eq. (6). Input parameters include, for example, initial conditions, boundary conditions, thermodynamic and transport properties, geometric quantities, and body forces such as electromagnetic forces on the domain of the PDEs.
(2) The SRQ is computed at a sufficient number of values over the range of the input variable, thus allowing an accurate construction of an interpolation function to represent the SRQ.

The following assumptions are made with regard to the experimental measurements:

(1) The input variable from the experiment is measured much more accurately than the SRQ. Quantitatively, this means that the standard deviation of the input variable is much smaller than the standard deviation of the SRQ. Note that this assumption allows for the case where the input variable is uncontrolled in the experiment but assumed to be accurately measured.
(2) Two or more experimental replications have been obtained, and each replication has multiple measurements of the SRQ over the range of the input variable. Using the terminology of Coleman and Steele [50], it is desirable that Nth-order replications have been obtained, and possibly even replications made by different experimentalists using different facilities and different diagnostic techniques.
(3) The measurement uncertainty in the SRQ from one experimental replication to the next, and from setup to setup, is given by a normal distribution.
(4) Each experimental replication is independent from other replications; that is, there is zero correlation or dependence between one replication and another.
(5) For each experimental replication, the SRQ is measured at a sufficient number of values over the range of the input variable so that a smooth and accurate interpolation function can be constructed to represent the SRQ.

With these assumptions, the equations developed in Section 4.1 are easily extended to the case in which both the computational result and the experimental mean for the SRQ are functions of the input variable x. Rewriting Eq. (15), the true error as a function of x is in the interval

$$\left(\widetilde{E}(x) - t_{0.05,\nu}\cdot\frac{s(x)}{\sqrt{n}}, \widetilde{E}(x) + t_{0.05,\nu}\cdot\frac{s(x)}{\sqrt{n}}\right) \tag{16}$$

with a confidence level of 90%, and the standard deviation as a function of x is given by

$$s(x) \sim \left[\frac{1}{n-1} \sum_{i=1}^{n} \left(y_e^i(x) - \bar{y}_e(x) \right)^2 \right]^{1/2} \tag{17}$$

Note that $y_e^i(x)$ is interpolated using the experimental data from the ith experimental replication, i.e., the ensemble of measurements over the range of x from the ith experiment. Each experimental replication need not make measurements at the same values of x because a separate interpolation function is constructed for each ensemble of measurements, i.e., each ith experimental replication.

5.2. Global metrics

Although these equations provide the results of the validation metric as a function of x, there are some situations where it is desirable to construct a more compact, or global, statement of the validation metric result. For example, in a high-level project management review, it may be useful to quickly summarize measures of disagreement for a large number of computational models and experimental data. A convenient method to compute a global metric would be to use a vector norm of the estimated error over the range of the input variable. The L_1 norm is useful to interpret the estimated average absolute error of the computational model over the range of the data. Using the L_1 norm, one could form an average absolute error or a relative absolute error over the range of the data. We choose to use the relative absolute error by normalizing the absolute error by the estimated experimental mean and then integrating over the range of the data. We define the *average relative error metric* to be

$$\left| \frac{\widetilde{E}}{\bar{y}_e} \right|_{\text{avg}} = \frac{1}{(x_u - x_l)} \int_{x_l}^{x_u} \left| \frac{y_m(x) - \bar{y}_e(x)}{\bar{y}_e(x)} \right| \mathrm{d}x \tag{18}$$

where x_u is the largest value and x_l is the smallest value, respectively, of the input variable. As long as $|\bar{y}_e(x)|$ is not near zero, the average relative error metric is a useful quantity.

The confidence interval that should be associated with this average relative error metric is the average confidence interval normalized by the absolute value of the estimated experimental mean over the range of the data. We define the *average relative confidence indicator* as the half-width of the confidence interval averaged over the range of the data:

$$\left| \frac{\mathrm{CI}}{\bar{y}_e} \right|_{\text{avg}} = \frac{t_{0.05,\nu}}{(x_u - x_l)\sqrt{n}} \int_{x_l}^{x_u} \left| \frac{s(x)}{\bar{y}_e(x)} \right| \mathrm{d}x \tag{19}$$

We refer to $\left| \frac{\mathrm{CI}}{\bar{y}_e} \right|_{\text{avg}}$ as an indicator, as opposed to an average relative confidence interval, because the uncertainty structure of $s(x)$ is not maintained through the integration operator. $\left| \frac{\mathrm{CI}}{\bar{y}_e} \right|_{\text{avg}}$ would provide a quantity with which to interpret the significance of $\left| \frac{\widetilde{E}}{\bar{y}_e} \right|_{\text{avg}}$. Stated differently, the magnitude of $\left| \frac{\widetilde{E}}{\bar{y}_e} \right|_{\text{avg}}$ should be interpreted relative to the magnitude of the normalized uncertainty in the experimental data, $\left| \frac{\mathrm{CI}}{\bar{y}_e} \right|_{\text{avg}}$.

There may be situations where the average relative error metric may not adequately represent the model accuracy because of the strong smoothing nature of the integration operator. For example, there may be a large error at some particular point over the range of the data that should be noted. It is useful to define a maximum value of the absolute relative error over the range of the data. Using the L_∞ norm to accomplish this, we define the *maximum relative error metric* as

$$\left| \frac{\widetilde{E}}{\bar{y}_e} \right|_{\max} = \max_{x_l \leqslant x \leqslant x_u} \left| \frac{y_m(x) - \bar{y}_e(x)}{\bar{y}_e(x)} \right| \tag{20}$$

A significant difference between $\left| \frac{\widetilde{E}}{\bar{y}_e} \right|_{\text{avg}}$ and $\left| \frac{\widetilde{E}}{\bar{y}_e} \right|_{\max}$ would indicate the need to more carefully examine the trend of the model with respect to the trend of the experimental data.

The confidence interval that should be associated with the maximum relative error metric is the confidence interval normalized by the estimated experimental mean. Both the confidence interval and the estimated exper-

imental mean are evaluated at the point where the maximum relative error metric occurs. Let the x value where $\left|\frac{\widetilde{E}}{\bar{y}_e}\right|_{\max}$ occurs be defined as $\hat{x}$. Then the confidence interval associated with the maximum relative error metric is

$$\left|\frac{\mathrm{CI}}{\bar{y}_e}\right|_{\max} = \frac{t_{0.05,\nu}}{\sqrt{n}}\left|\frac{s(\hat{x})}{\bar{y}_e(\hat{x})}\right| \tag{21}$$

Note that in this section, Section 5, all of the functions of x, e.g., $y_m(x)$ and $s(x)$, are considered as continuous functions constructed by interpolation. In the next section, Section 6, we consider the case where these functions are constructed using regression.

5.3. Example: turbulent buoyant plume

As an example of the validation metric just derived, consider the assessment of a computational model for a turbulent buoyant plume of helium that is exiting vertically from a large nozzle. Turbulent buoyant plumes, typically due to the combustion of fuel–air mixtures, have proven to be especially difficult to model in CFD. This is primarily because of the strong interaction between the density field and the momentum field dominated by large turbulent eddies. The slowest turbulent scales are on the order of seconds in large fires, and this large-scale unsteadiness is beyond the modeling capability of a Reynolds-Average Navier–Stokes (RANS) formulation. The computational model to be evaluated here solves the continuity equation and the temporally filtered Navier–Stokes (TFNS) equations. The TFNS equations are similar to RANS equations, but a narrower filter width is used so that large-scale unsteadiness can be captured [58]. DesJardin et al. [59] have also computed turbulent buoyant plumes using large-eddy simulation (LES), but these simulations are even more computer intensive than TFNS simulations. Tieszen et al. [60] conducted an unsteady, three-dimensional simulation of a large-scale helium plume using the TFNS model and the standard k–ε turbulence model. These models, among others, are implemented in the SIERRA/Fuego computer code [61] being developed at Sandia as part of the ASC Program.

The experimental data for the validation metric were obtained in the fire laboratory for accreditation of models and experiments (FLAME) facility at Sandia. The FLAME facility is a building designed for indoor fire experiments so that atmospheric winds do not influence the buoyant plume, and all other boundary conditions affecting the plume can be measured and controlled. For the present experiment, instead of the fuel-

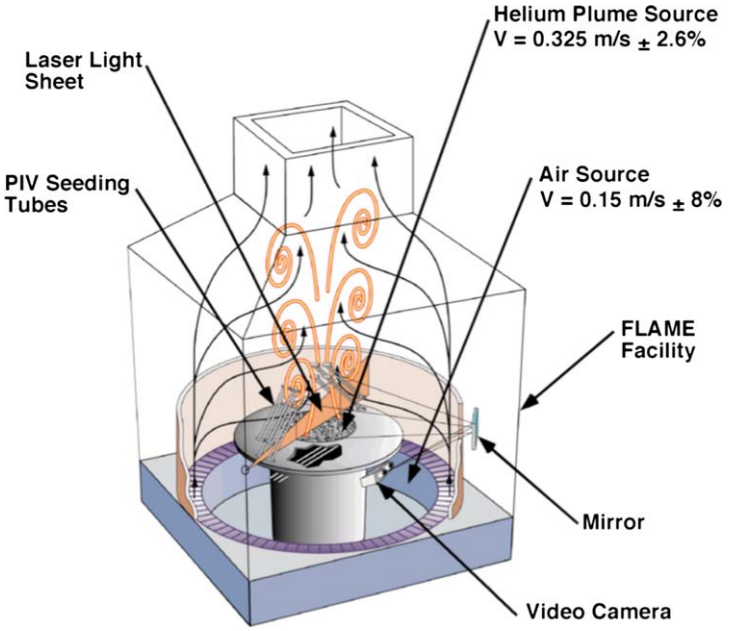

Fig. 4. Experimental setup for measurements of the helium plume [49,59,62]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

pool fire producing a buoyant plume, an inflow jet of helium was used (Fig. 4) [59,62]. The helium source is 1 m in diameter and is surrounded by a 0.51 m wide surface to simulate the ground plane that is typical in a fuel-pool fire. Inlet air is injected from outside the building at the bottom of the facility and is drawn by the accelerating helium plume over the ground plane surrounding the plume source.

The experimental data consist of velocity field measurements using particle image velocimetry (PIV) and scalar concentration measurements using planar-induced fluorescence (PLIF). Here we are interested in only the PIV measurements, but details of all of the diagnostic procedures and uncertainty estimates can be found in O'Hern et al. [62]. The PIV data are obtained from photographing the flowfield, which has been seeded with microspheres of glass beads, at 200 images/s. Flowfield velocities are obtained in a plane that is up to 1 m from the exit of the jet and illuminated by a laser light sheet. The flow velocity of interest here, i.e., the SRQ that is input to the validation metric, is the vertical velocity component along the centerline of the helium jet. For unsteady flows such as this, there are a number of different oscillatory modes that exist within the plume. The SRQ of interest is time-averaged for roughly 10 s in the experiment, which is roughly seven cycles of the lowest mode in the jet. Shown in Fig. 5 are four experimental measurements of time-averaged vertical velocity along the centerline as a function of axial distance from the exit of the helium jet. The experimental replications were obtained on different days, with different equipment setups, and with multiple recalibrations of the instrumentation. A large number of velocity measurements were obtained over the range of the input variable, the axial distance, so that an accurate interpolation function could be constructed.

Tieszen et al. [60] investigated the sensitivity of their solutions to both modeling parameters and numerical discretization on an unstructured mesh. The key modeling parameter affecting the TFNS solutions is the size of the temporal filter relative to the period of the largest turbulent mode in the simulation. Four spatial discretizations were investigated: 0.25M, 0.50M, 1M, and 2M elements ($1M = 1 \times 10^6$). Each of these solutions was time-averaged over roughly seven puffing cycles, as were the experimental data. In comparing their 1M- and 2M-element solutions, we found little reason to be convinced that the 2M-element solution was spatially converged. A finer mesh, say, 4M elements, would greatly help in determining whether the computational results are actually converged. However, computational resources were not available to compute the 4M-element solution. As a result, we will use their 2M-element solution as only representative data with which to demonstrate the present validation metric.

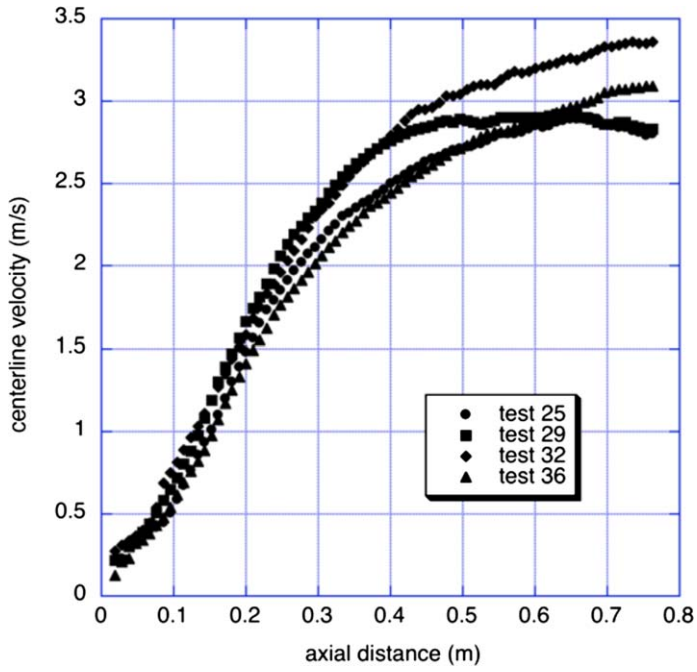

Fig. 5. Experimental measurements of time-averaged vertical velocity along the centerline for the helium plume [49,59,62]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

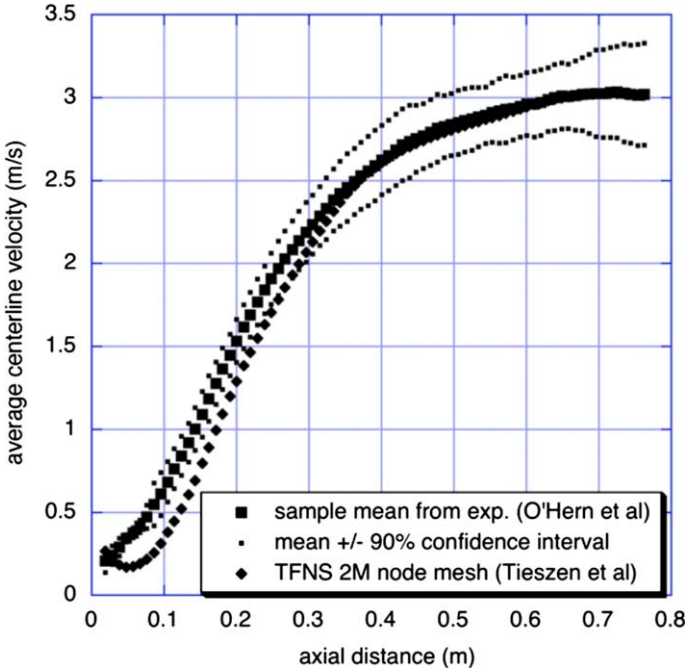

Fig. 6. Experimental sample mean with 90% confidence interval and computational result for vertical velocity in the helium plume [49]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

Using the experimental data shown in Fig. 5, noting that $n = 4$, one obtains the sample mean of the measurements, $\bar{y}_e(x)$, shown in Fig. 6. Also, using the interpolated function for the experimental sample mean and the confidence interval for the true mean, one obtains the interval around the estimated mean in which the true mean will occur with 90% confidence (Fig. 6). The computational solution obtained from the 2M-element mesh is also shown in Fig. 6. As is commonly done in the literature, an author would conclude that there is "good" agreement between computational and experimental results or, more boldly, claim that the code has been "validated". However, as discussed previously, such statements ignore critical issues: (a) "good" has not been quantified; and (b) accuracy requirements for the intended use of the model have been ignored, rendering any claim of "good" agreement questionable.

The level of (dis)agreement between computational and experimental results can be more critically seen by plotting the estimated error, $\widetilde{E}(x) = y_m(x) - \bar{y}_e(x)$, along with the 90% confidence interval from the experiment (Fig. 7). The type of plot shown in Fig. 7 is the result of the validation metric derived in Section 5.1. Examining these quantities provides a magnifying glass, as it were, to both the error in the computational model and the uncertainty in the experimental data. Only courageous modelers, experimentalists, or decision makers using the model will be eager to examine matters this closely. Two points should be made from Fig. 7. First, the largest modeling error, although not large, occurs very near the beginning of the plume. Second, near this region the magnitude of the modeling error is outside the 90% confidence interval of the experimental data, giving credence to the estimated modeling error. We remind the reader that these conclusions can only be defended if it is *assumed* that the TFNS solution is mesh converged.

The validation metric result shown in Fig. 7 can be quantitatively summarized, or condensed, using the global metrics given in Eqs. (21)–(24). Over the range of the data, these results are as follows:

Average relative error $= 11\% \pm 9\%$ with 90% confidence
Maximum relative error $= 54\% \pm 9\%$ with 90% confidence

Thus, the average relative error could be as large as 20% and as small as 2% (on average) over the range of the data, with 90% confidence due to uncertainty in the experimental data. The average relative error shows that the model accuracy, on average, is comparable to the average confidence indicator in the experimental

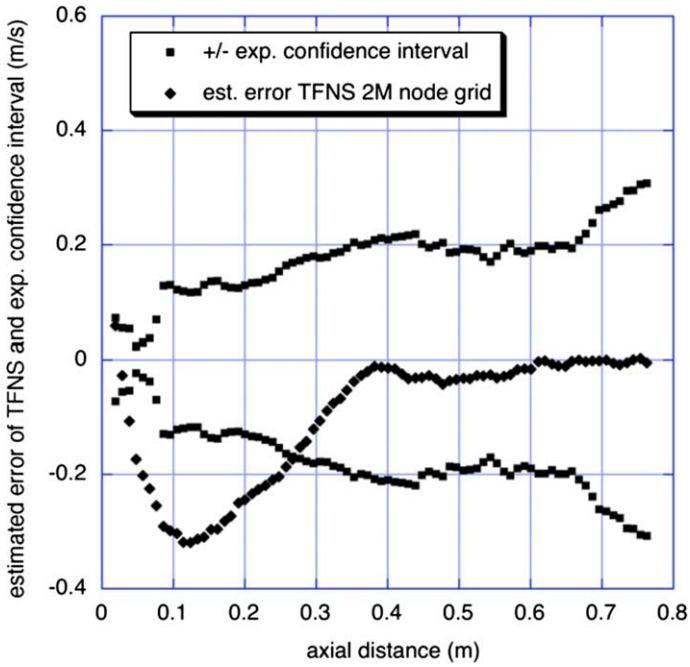

Fig. 7. Validation metric result and 90% confidence interval for centerline velocity [49]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

data. Similarly, the maximum relative error could be as small as 45% and as large as 63%, with 90% confidence due to uncertainty in the experimental data. The maximum relative error, 54%, which occurs at $x = 0.067$ m, is five times the average relative error, indicating a significant difference in the local character of the computational model and the experimental data. Note that for these experimental data, the average relative confidence

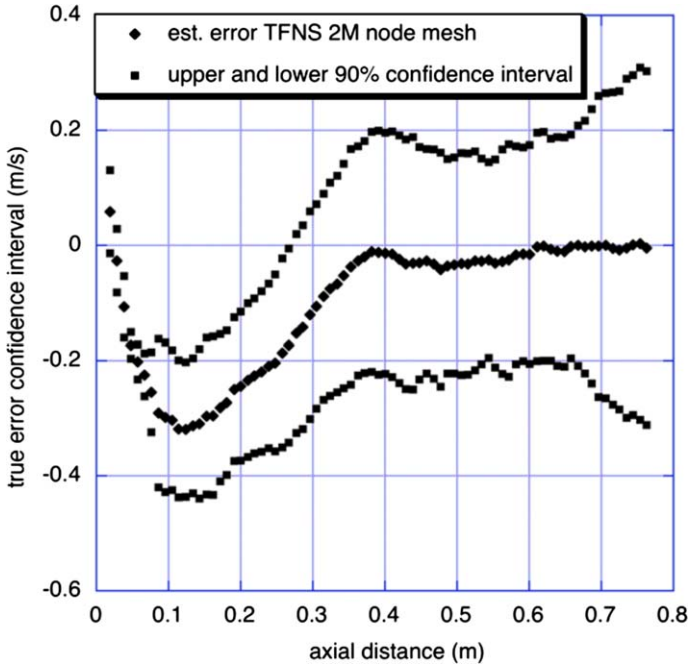

Fig. 8. Estimated error and true error in the model with 90% confidence interval uses. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

indicator, 9%, happens to be essentially equal to the relative confidence interval at the maximum relative error. If one uses both the average relative error and the maximum relative error for a "first look" evaluation of the model, a large difference between these values should prompt a more careful examination of the data, for example, examination of plots such as Figs. 6 and 7.

The final method of displaying the results of the validation metric is to plot the 90% confidence interval of the true error in velocity predicted by the computational model as a function of the axial distance from the exit of the jet. Using Eq. (16), one obtains the result shown in Fig. 8. Our best approximation of the true error in the model is the estimated error. However, with 90% confidence we can state that the true error is in the interval shown in Fig. 8.

Although Fig. 8 displays essentially the same data as shown in Fig. 7, Fig. 8 allows us to consider slightly different perspectives for assessing the model. For example, we could view Fig. 8 from the perspectives of those who might use the validation metric results to evaluate the predictive capability of the computational model. A model builder, for example, would likely investigate the cause of the largest error, i.e., near $x = 0.01$ m, and explore ways to improve the model. For an analyst, i.e., a person who is going to use the model for predictions of flowfields that are related to the present flowfield, the perspective is somewhat different from that of the model builder. The analyst might conclude that the accuracy of the model is satisfactory for its intended use and simply apply the model as it is. Alternatively, the analyst might decide to use Fig. 8 to incorporate a bias-error correction directly on the SRQ, i.e., the vertical velocity on the centerline of the plume. For example, the analyst might take any new result for the SRQ computed from the model and correct it according to the curve for the estimated error in Fig. 8. If the analyst was conducting a nondeterministic analysis, the analyst might assign a bias correction using a normal probability distribution to the interval shown in Fig. 8 with the expected value set to the estimated error and the upper and lower intervals set to the 90% quantile of the distribution. This procedure for model correction would clearly involve risk because it completely ignores the physical cause of the error. However, if the schedule or budget for completing the analysis does not allow further investigation, this procedure could prove useful for the analyst and decision maker.

6. Validation metric requiring regression

6.1. Development of the equations

We are now interested in a case similar to that described in Section 5, where there is still one SRQ that is measured over a range of one input or operating condition variable but the quantity of experimental data for this new case is not sufficient to construct an interpolation function. Consequently, a regression function (curve fit) must be constructed to represent the estimated mean over the range of the data. Some examples are lift (or drag) of a flight vehicle as a function of the Mach number, turbopump mass flow rate as a function of backpressure, and depth of penetration into a material during high-speed impact. Construction of a regression function is probably the most common situation that arises in comparing computational and experimental results when the input variable is *not* time. When time-dependent SRQs are recorded, the temporal resolution is typically high so that the construction of a validation metric would be analogous to the situation discussed in Section 5.

Regression analysis procedures are well developed in classical statistics for addressing how two or more variables are related to each other when one or both contain random uncertainty. We are interested here in the restricted case of univariate regression, i.e., how one variable (the SRQ) relates to another variable (the input variable). The two assumptions pertaining to the computational results discussed in Section 5.1 are also made for the present case. The first four assumptions pertaining to the experimental measurements discussed in Section 5.1 are also made for the present case. In addition to these, the following assumption is made with regard to the experimental uncertainty: the standard deviation of the normal distribution that describes the measurement uncertainty is constant over the entire range of measurements of the input parameter. It should also be noted that this assumption is probably the most demanding of the experimental measurement assumptions listed.

In the present development, it was initially thought that traditional confidence intervals in regression analysis could be applied to the construction of the validation metric. (See, for example, Ref. [23] for

a description of the traditional development of confidence intervals in regression analysis.) We realized, however, that the traditional development only applies to the case of a specific, but arbitrary, value of the input parameter. That is, the traditional confidence interval is a statement of the accuracy of the estimated mean as expressed by the regression for *point values* of the input parameter x. The traditional confidence interval is written for μ conditional on a point value of x, say, x^*, i.e., $\mu[\bar{y}_e(x)|x^*]$ As a result, the traditional confidence interval analysis cannot be applied to the case of a validation metric over a range of the input variable.

A more general statistical analysis procedure was found that develops a confidence interval for the entire range of the input parameter [63–65]. That is, we wish to determine the confidence interval that results from uncertainty in the regression coefficients over the complete range of the regression function. The regression coefficients are all correlated with one another because they appear in the same regression function that is fitting the experimental data. This type of confidence interval is typically referred to as a simultaneous confidence interval, a simultaneous inference, or a confidence region, so that it can be distinguished from traditional (or single-comparison) confidence intervals.

Since the quantity of experimental data is not sufficient to construct an interpolating function, we can represent the estimated mean of the data, $\bar{y}_e(x)$, as a general nonlinear regression function

$$\bar{y}_e(x) = f(x;\vec{\theta}) + \varepsilon \tag{22}$$

where $f(x;\cdot)$ is the chosen form of the regression function over the range of the input parameter x; $\vec{\theta} = \theta_1,\ \theta_2, \ldots, \theta_p$ are the unknown coefficients of the regression function; and ε is the random measurement error. Let the set of n experimental measurements of the SRQ of interest be given by

$$(y_e^i, x_i) \quad \text{for } i = 1,\ 2, \ldots, n \tag{23}$$

Using a least-squares fit of the experimental data, it can be shown [64,65] that the error sum of squares $S(\vec{\theta})$ in p-dimensional space is

$$S(\vec{\theta}) = \sum_{i=1}^{n} \left[y_e^i(x) - f(x_i;\vec{\theta}) \right]^2 \tag{24}$$

The vector that minimizes $S(\vec{\theta})$ is the solution vector, and it is written as $\vec{\hat{\theta}}$. This system of simultaneous, nonlinear equations can be solved by various software packages that compute solutions to the nonlinear least-squares problem.

Draper and Smith [64] and Seber and Wild [65] discuss a number of methods for the computation of the confidence regions around the point $\vec{\hat{\theta}}$ in p-dimensional space. For any specified confidence level $100(1-\alpha)\%$, a unique region envelops the point $\vec{\hat{\theta}}$. For two regression parameters, (θ_1, θ_2), we have a two-dimensional space, and these regions are contours that are similar to ellipses with a curved major axis. For three parameters, $(\theta_1, \theta_2, \theta_3)$, we have a three-dimensional space, and these regions are contours that are similar to bent ellipsoids and shaped like a banana. A procedure that appears to be the most robust to nonlinear features in the equations [65] and that is practical when p is not too large, is to solve an inequality for the set of $\vec{\theta}$:

$$\vec{\theta} \text{ such that } S(\vec{\theta}) \leqslant S(\vec{\hat{\theta}}) \left[1 + \frac{p}{n-p} F(p, n-p, 1-\alpha) \right] \tag{25}$$

In Eq. (25), $F(v_1, v_2, 1-\alpha)$ is the F probability distribution, v_1 is the first parameter specifying the number of degrees of freedom, v_2 is the second parameter specifying the number of degrees of freedom, $1-\alpha$ is the quantile for the confidence interval of interest, and n is the number of experimental measurements.

If we would like to make a quantitative assessment of the global modeling error, then we can extend the global measures expressed in Eqs. (21)–(24). The average relative confidence indicator, Eq. (19), and the confidence interval associated with the maximum relative error, Eq. (21), are based on symmetric confidence intervals derived in Section 5.1. Since we no longer have symmetric confidence intervals, we approximate these by taking the average half-width of the confidence interval over the range of the data and the half-width of the confidence interval at the maximum relative error, respectively. As a result, we now have

$$\left|\frac{\mathrm{CI}}{\bar{y}_e}\right|_{\mathrm{avg}} = \frac{1}{(x_u - x_l)} \int_{x_l}^{x_u} \left|\frac{y^+_{\mathrm{CI}}(x) - y^-_{\mathrm{CI}}(x)}{2\bar{y}_e(x)}\right| \mathrm{d}x \tag{26}$$

for the average relative confidence indicator. $y^+_{\mathrm{CI}}(x)$ and $y^-_{\mathrm{CI}}(x)$ are the upper and lower confidence intervals, respectively, as a function of x. As discussed in the next section, $y^+_{\mathrm{CI}}(x)$ and $y^-_{\mathrm{CI}}(x)$ are found by substituting into the regression function $f(x;\vec{\theta})$, all $\vec{\theta}$ that satisfy Eq. (25). As stated earlier, $\left|\frac{\mathrm{CI}}{\bar{y}_e}\right|_{\mathrm{avg}}$ provides a quantity with which to interpret the significance of $\left|\frac{\widetilde{E}}{\bar{y}_e}\right|_{\mathrm{avg}}$.

Also, we have

$$\left|\frac{\mathrm{CI}}{\bar{y}_e}\right|_{\mathrm{max}} = \left|\frac{y^+_{\mathrm{CI}}(\hat{x}) - y^-_{\mathrm{CI}}(\hat{x})}{2\bar{y}_e(\hat{x})}\right| \tag{27a}$$

for the half-width of the confidence interval at the maximum relative error point, $\hat{x}$. $\left|\frac{\mathrm{CI}}{\bar{y}_e}\right|_{\mathrm{max}}$ provides a quantity with which to interpret the significance of $\left|\frac{\widetilde{E}}{\bar{y}_e}\right|_{\mathrm{max}}$. The maximum relative error point is defined as the x value where $\left|\frac{\widetilde{E}}{\bar{y}_e}\right|$ achieves its maximum, that is,

$$\hat{x} = x \text{ such that } \left|\frac{y^+_{\mathrm{CI}}(x) - y^-_{\mathrm{CI}}(x)}{\bar{y}_e(x)}\right| \text{ is a maxium for } x_l \leqslant x \leqslant x_u \tag{27b}$$

6.2. Solution of the equations

We consider a geometric interpretation of Eq. (25) to facilitate the numerical evaluation of the inequality. We seek the complete set of $\vec{\theta}$ values that satisfy the inequality. For a given confidence level α, the inequality describes the interior of a p-dimensional hypersurface in $\vec{\theta}$ space. Thus, for $p = 2$, it describes a *confidence region*, bounded by a closed contour, in the parameter space (θ_1, θ_2). An example of a set of such contours is depicted in Fig. 9. As the confidence level increases, the corresponding contours describe larger and larger regions about the least-squares parameter vector $\vec{\hat{\theta}}$.

The numerical algorithm employed in the present work discretizes the interior of the confidence region using several contour levels that lie within the highest confidence contour. For example, suppose we wish to calculate the 90% confidence interval given the confidence regions depicted in Fig. 9. We would evaluate the regression equation at a number of points, say, 20%, along the entire 90% contour. Then, we would do

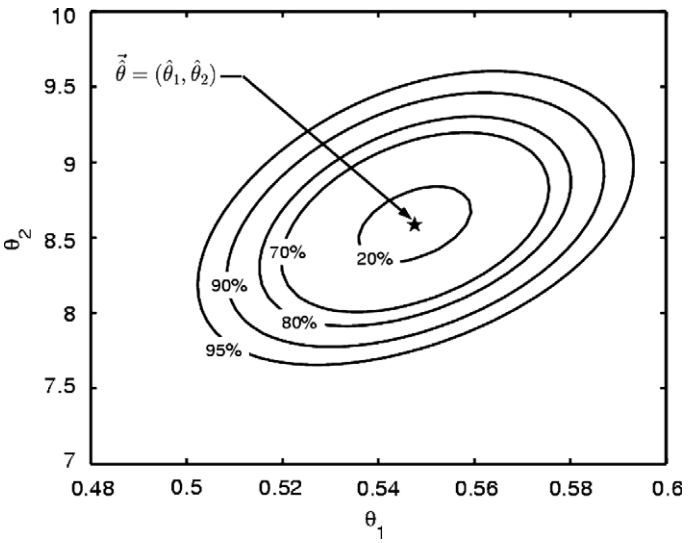

Fig. 9. Example of various confidence regions for the case of two regression parameters [49]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

the same along the 80% contour, the 70% contour, and so on down to the 10% contour. With all of these regression function evaluations, we would then compute the maximum and minimum of the function over the range of the input parameter x. This would provide reasonably good coverage of the 90% confidence interval of the regression function. If more precision was needed, one could choose more function evaluations along each contour and compute each contour in 1% increments of the confidence level.

For a three-dimensional regression parameter space, slices can be taken along one dimension of the resulting three-dimensional surface, and each slice can be discretized in the manner described for the two-dimensional case. Generalizing to N dimensions, one may generate a recursive sequence of hypersurfaces of lower dimension until a series of two-dimensional regions are obtained, and evaluation over all of the two-dimensional regions gives the desired envelope of regression curves.

To determine the upper and lower confidence *intervals* associated with the regression equation, Eq. (22), we use the solution to Eq. (25), i.e., all $\vec{\theta}$ lying within (and on) the desired contour. The confidence intervals are determined by computing the envelope of regression curves resulting from *all* $\vec{\theta}$ lying within the confidence region. If we think of the solution to Eq. (25) as given by a set of discrete vectors of $\vec{\theta}$, then we can substitute this set of parameter vectors into the regression equation, Eq. (22). For each element in this set of $\vec{\theta}$s, we obtain a specific regression function. If we evaluate the ensemble of all regression functions by using all of the $\vec{\theta}$s, we can compute the maximum value of the regression function, $y^{+}_{\mathrm{CI}}(x)$, and the minimum value of the regression function, $y^{-}_{\mathrm{CI}}(x)$, over the range of x. As a result, $y^{+}_{\mathrm{CI}}(x)$ and $y^{-}_{\mathrm{CI}}(x)$ define the upper and lower bounds on the confidence intervals, respectively, over the range of x. One may ask why the regression function must be evaluated over the entire confidence region. This must be done because the nonlinear regression function can have maxima and minima anywhere within the confidence region.

6.3. Example: compressible free-shear layer

The example chosen for the application of the validation metric derived in Section 6.1 is prediction of compressibility effects on the growth rate of a turbulent free-shear layer. An introduction to the problem is given, followed by a discussion of the available experimental data. Details of the computational model and verification of the numerical solutions are then described along with the validation metric results. A more detailed discussion of the experimental and computational analysis can be found in a paper by Barone et al. [66].

6.3.1. Problem description

The planar free-shear layer is a canonical turbulent flow and a good candidate for use in a unit-level validation study. Fig. 10 shows the general flow configuration in which a thin splitter plate separates two uniform streams (numbered 1 and 2) with different flow velocities and temperatures. The two streams mix downstream of the splitter-plate trailing edge, forming the free-shear layer within which momentum and energy are diffused. For a high Reynolds-number flow, the boundary layers on both sides of the plate and the free-shear layer are inevitably turbulent. In the absence of any applied pressure gradients or other external influences,

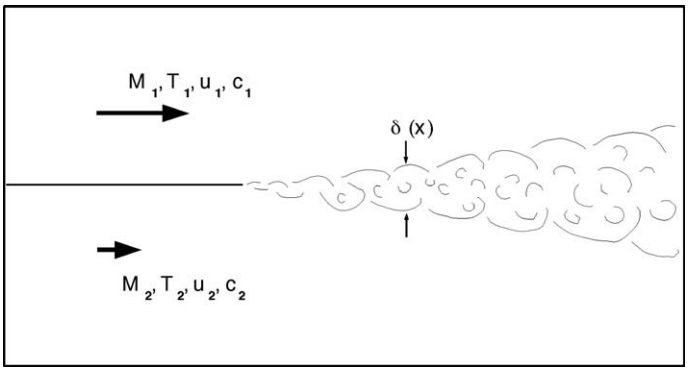

Fig. 10. Flow configuration for the turbulent free-shear layer [49]. The figure is reprinted by permission of the American Institute of the Aeronautics and Astronautics, Inc.

the flowfield downstream of the trailing edge consists of a shear layer development region near the edge, followed by a similarity region. Within the development region, the shear layer adjusts from its initial velocity and temperature profiles inherited from the plate boundary layers. Further downstream in the similarity region, the shear layer thickness, $\delta(x)$, grows linearly with streamwise distance x, resulting in a constant value of $\mathrm{d}\delta/\mathrm{d}x$.

Of particular interest in high-speed vehicle applications is the behavior of the shear layer as the Mach number of one or both streams is increased. A widely accepted parameter correlating the shear layer growth rate with compressibility effects is the convective Mach number, that was defined by Bogdanoff [67] for mixing two streams of the same gas:

$$M_c = \frac{u_1 - u_2}{c_1 + c_2} \tag{28}$$

where u is the fluid velocity and c is the speed of sound. It has been found experimentally that an increase in the convective Mach number leads to a decrease in the shear layer growth rate for the fixed velocity and temperature ratios of the streams. This is usually characterized by the compressibility factor Φ, which is defined as the ratio of the compressible growth rate to the incompressible growth rate at the same velocity and temperature ratios:

$$\Phi = \frac{(\mathrm{d}\delta/\mathrm{d}x)_c}{(\mathrm{d}\delta/\mathrm{d}x)_i} \tag{29}$$

6.3.2. Experimental data

Experimental data on high-speed shear layers are available from a number of independent sources. The total collection of experimental investigations employs a wide range of diagnostic techniques within many different facilities. Comparisons of data obtained over a range of convective Mach numbers from various experiments indicate a significant scatter in the data. (See, e.g., Lele [68].) Recently, Ref. [66] carefully reexamined the available data and produced a recommended data set that exhibits smaller scatter in the measurements. The guidelines for filtering and reanalyzing the data were as follows:

1. Shear layer thickness data based on pitot measurements or optical photographs, such as Schlieren photographs, were not considered reliable and were rejected.
2. A data point was eliminated if there was clear evidence, based on the description of the experiment, the facility, or the data, that the required experimental conditions were not met, e.g., $\delta(x)$ showed that shear layers were not fully developed in the test section of the wind tunnel. Note that no data was removed simply because it was an "outlier" or "looked bad", as is commonly done in reporting experimental results.

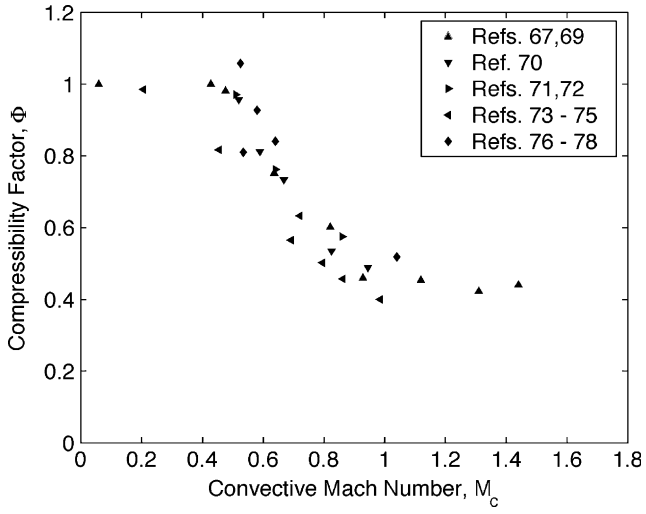

Fig. 11. Experimental data for compressibility factor versus convective Mach number [66].

3. A consistent method was used to estimate the incompressible growth rate $(d\delta/dx)_i$ given the experimental flow conditions for each experiment considered.

The resulting ensemble of data from Refs. [67,69–78] is presented in Fig. 11. The data are organized into groups of sources, some of which are themselves compilations of results from several experiments.

6.3.3. Computational model

For the present study, we use the simulation results computed by Barone et al. [66]. They used the Favre-averaged compressible Navier–Stokes equations with the standard k–ε turbulence model [79]. The low Reynolds-number modification to the k–ε model of Nagano and Hishida [80] was applied near the splitter plate. Most turbulence models in their original form do not correctly predict the significant decrease in shear layer growth rate with increasing convective Mach number, necessitating inclusion of a compressibility correction. Several compressibility corrections, derived from a variety of physical arguments, are widely used in contemporary CFD codes. In this study, the dilatation-dissipation compressibility correction of Zeman [81] is used.

The solutions were computed using the Sandia advanced code for compressible aerothermodynamics research and analysis (SACCARA) [82,83], that employs a block-structured, finite volume discretization method. The numerical fluxes are constructed with the Symmetric TVD scheme of Yee [84], which gives a second-order convergence rate in smooth flow regions. The equations are advanced to a steady-state using the LU-SGS scheme of Yoon and Jameson [85]. Solutions were considered iteratively converged when the L2 norm of the momentum equation residuals decreased eight orders of magnitude. Numerical solutions were obtained over the convective Mach number range of the experimental data, from 0.1 to 1.5, in increments of 0.14.

For each convective Mach number, solutions were calculated on three grids: coarse, medium, and fine. The grids are uniform in the streamwise, or x, direction, and stretched in the cross-stream, or y, direction, so that grid cells are clustered within the shear layer. The cells are highly clustered in the y direction near the trailing edge and become less clustered with increasing x to account for the shear layer growth. Richardson extrapolation [5,10,11,86,87] was used to estimate the discretization error on $d\delta/dx$. The maximum error in the fine-grid solution was estimated to be about 1% at $M_c = 0.1$ and about 0.1% at $M_c = 1.5$.

We defined δ using the velocity layer thickness definition. (See Ref. [66] for details.) As mentioned previously, the thickness grows linearly with x only for large x due to the presence of the development region, which precedes the similarity region. Given that the growth rate actually approaches a constant value only asymptotically, the thickness as a function of x is fit with a curve that mimics this functional form. The function used for the fit is

$$\delta(x) = \beta_0 + \beta_1 x + \beta_2 x^{-1} \tag{30}$$

which leads to a growth rate that approaches β_1 as x becomes large. The coefficient β_1 is taken to be the actual fully developed shear layer growth rate.

Following extraction of the compressible growth rate, $(d\delta/dx)_c$, the incompressible growth rate, $(d\delta/dx)_i$, must be evaluated at the same velocity and temperature ratio. Incompressible or nearly incompressible results are difficult to obtain with a compressible CFD code. Therefore, the incompressible growth rate was obtained by computing a similarity solution for the given turbulence model and flow conditions. The similarity solution is derived by Wilcox [79] in his turbulence modeling text and implemented in the MIXER code, which is distributed with the text. The similarity solution is computed using the same turbulence model as the Navier–Stokes calculations, but under the assumptions that (a) the effects of laminar viscosity are negligible and (b) there exists zero pressure gradient.

6.3.4. Validation metric results

The quantities δ and $d\delta/dx$ are post-processed from the finite-volume computational solution and the MIXER code, but the SRQ of interest for the validation metric is the compressibility factor Φ. Before the validation metric result can be computed, we must prescribe a form for the nonlinear regression function to represent the experimental data in Fig. 11. It is important that the proper functional behavior of the data, established through theoretical derivation or experimental measurement, be reflected in the form of the regres-

sion function. For the compressible shear layer, we know that Φ must equal unity, by definition, in the incompressible limit $M_c \to 0$. Experimental observations and physical arguments also suggest that $\Phi \to$ constant as M_c becomes large. These considerations lead to the following choice of the regression function, taken from Paciorri and Sabetta [88]:

$$\Phi = 1 + \hat{\theta}_1 \left(\frac{1}{1 + \hat{\theta}_2 M_c^{\hat{\theta}_3}} - 1 \right) \tag{31}$$

Using Eq. (31) and the experimental data shown in Fig. 11, we used the MATLAB [89] function *nlinfit* from the Statistics Toolbox to calculate the following regression coefficients:

$$\hat{\theta}_1 = 0.5537, \quad \hat{\theta}_2 = 31.79, \quad \hat{\theta}_3 = 8.426 \tag{32}$$

We now compute the 90% confidence interval of the regression function Eq. (31) with the $\vec{\hat{\theta}}$ values given in Eq. (32) and the inequality constraint given by Eq. (25). We use the method outlined in Section 6.2 to compute the 90% confidence region in the three-dimensional space described by θ_1, θ_2, and θ_3. The resulting confidence region, pictured in Fig. 12, resembles a curved and flattened ellipsoid, especially for small values of θ_2. The elongated shape in the θ_2 direction indicates the low sensitivity of the curve fit to θ_2 relative to the other two regression parameters. Evaluation of the regression function Eq. (31) for all $\vec{\theta}$ lying within the 90% confidence region yields the desired simultaneous confidence intervals.

Fig. 13 shows the final result of the analysis in graphical form: a plot of the experimental data along with the regression fit, the 90% confidence interval, and the computational simulation result. Concerning the 90% confidence interval, it is seen that the largest uncertainty in the experimental data occurs for large M_c. This

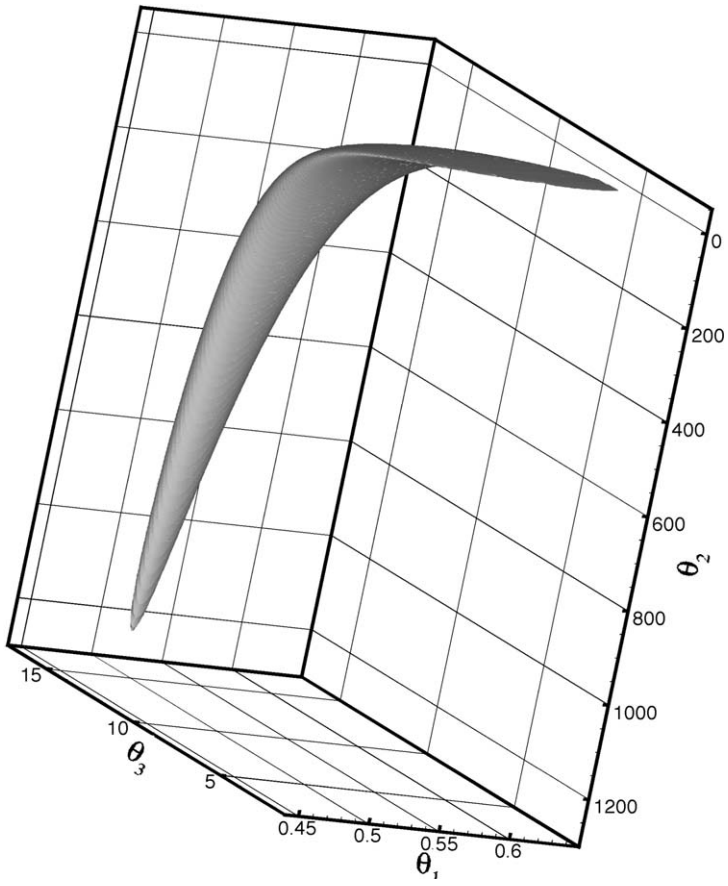

Fig. 12. Three-dimensional 90% confidence region for the regression fit to the shear layer experimental data.

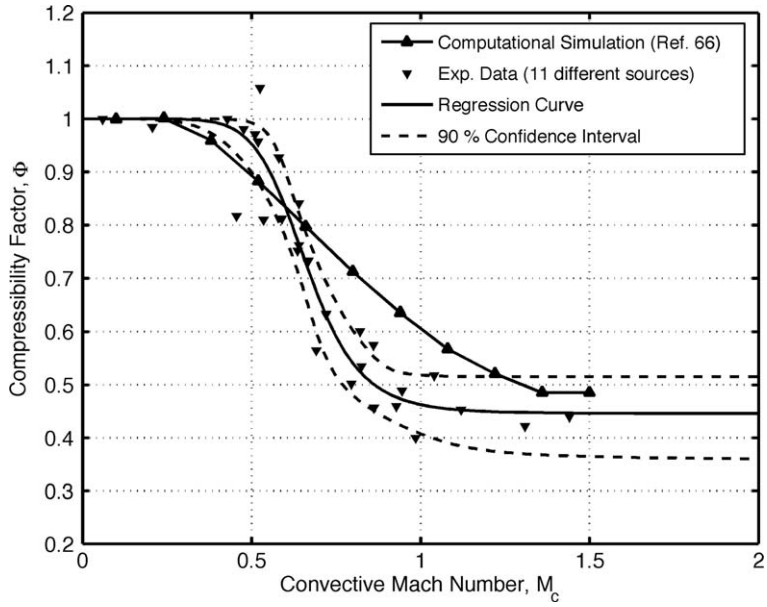

Fig. 13. Comparison of the simulation result with the experimental data, nonlinear regression curve, and 90% simultaneous confidence interval.

uncertainty is primarily a result of the uncertainty, i.e., the 90% confidence interval, in the θ_1 parameter of the regression function. From the viewpoint of the design of needed validation experiments, one can conclude that future experiments should be conducted at higher convective Mach numbers to better determine the asymptotic value of Φ. Concerning the error assessment of the k–ε model, it is seen that the Zeman compressibility correction predicts a nearly linear dependence of the compressibility factor on M_c over the range $0.2 \leqslant M_c \leqslant 1.35$. One could claim that the trend is correct, i.e., the Zeman model predicts a significant decrease in the turbulent mixing as the convective Mach number increases; however, the Zeman model does

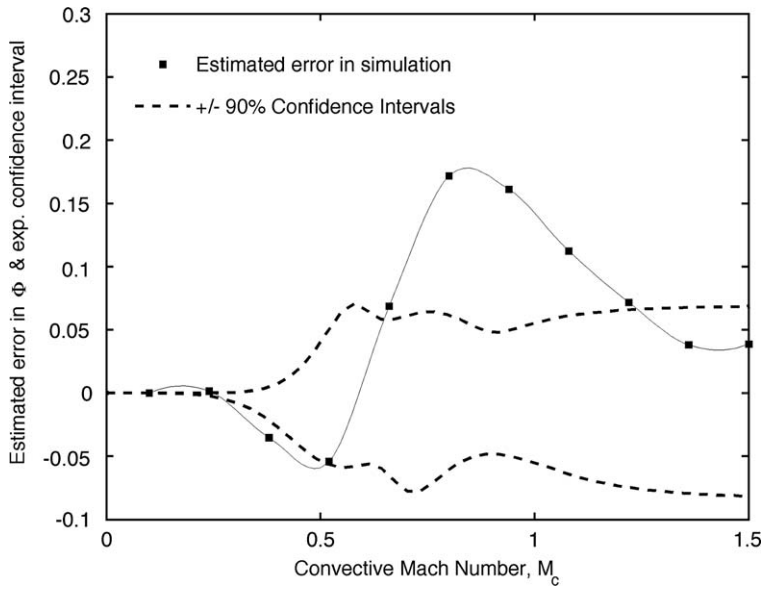

Fig. 14. Validation metric result and 90% confidence interval for Φ.

not predict the nonlinear dependency on M_c. We did not compute any simulation results for $M_c > 1.5$ and, as a result, did not determine the asymptotic value of Φ for the Zeman compressibility correction. However, the solutions for $M_c = 1.36$ and $M_c = 1.50$ suggest that the asymptotic value is near $\Phi = 0.49$.

The estimated error, $\widetilde{E}(x)$, of the model as a function of M_c is plotted in Fig. 14 along with the 90% confidence interval from the experimental data. This plot presents the validation metric result, i.e., the difference between computation and the regression fit of the experimental data, along with the 90% confidence interval representing the uncertainty in the experimental data. As pointed out previously in the helium plume example, the validation metric makes a critical examination of both a computational model and the experimental data. With this plot it is seen that there is a slight underprediction of turbulent mixing in the range $0.3 \leqslant M_c \leqslant 0.6$ and a significant overprediction of turbulent mixing in the range $0.7 \leqslant M_c \leqslant 1.3$. Examining an error plot such as this, one could conclude that the Zeman model does not capture the nonlinear trend of decreasing turbulent mixing with increasing convective Mach number. Whether the model accuracy is adequate for the requirements of the intended application is, of course, a completely separate conclusion.

Note that in Fig. 14 the confidence intervals are not symmetric with respect to zero. In the case of nonlinear regression, specifically Eq. (29) here, the nonlinear function need not possess any symmetric properties with respect to the regression parameters. Therefore, evaluation of the nonlinear function over the set of $\vec{\theta}$ satisfying Eq. (25) results in asymmetric confidence intervals over the range of the input parameter. For the shear layer example, Eq. (31) is evaluated over the volume of regression coefficients shown in Fig. 12.

Using Eqs. (18), (20), (31), and (32), the results for the k–ε model with the Zeman compressibility correction over the range $0 \leqslant M_c \leqslant 1.5$ are as follows:

Average relative error $= 13\% \pm 9\%$ with 90% confidence
Maximum relative error $= 35\% \pm 10\%$ with 90% confidence

The average error of 13%, though not alarmingly large, is clearly larger than the average experimental confidence indicator. As in the helium plume example, we encounter a maximum error that is noticeably larger than the average error, i.e., roughly a factor of three. From Fig. 14 it can be found that the maximum absolute error occurs at $M_c = 0.83$. The maximum relative error, however, occurs at $M_c = 0.88$. At this value of M_c one determines that the 90% confidence interval is $\pm 10\%$.

7. Conclusions and future work

The validation metrics derived here are relatively easy to compute and interpret in practical engineering applications. When nonlinear regression functions are required for the metric, the nonlinear regression function requires a software package, such as Mathematica or MATLAB, to perform the computations. The interpretation of the present metrics in engineering decision making should be clear and understandable to a wide variety of technical staff (analysts, model builders, and experimentalists) and management. The metric result has the following form: estimated error of the model $\pm$ an interval that represents experimental uncertainty with 90% confidence. The present metrics are only measures of error for the mean response of the system. More descriptive metrics, for example, those that would measure the accuracy of the variability of the response, should be developed in the future. The present metrics can be used to compare the modeling accuracy of different competing models, or they can help to assess the adequacy of the given model for an application of interest. We point out that how the result of a validation metric relates to an application of interest is a separate and more complex issue, especially if there is significant extrapolation of the model. Although this issue is not addressed here, it is critical to the estimation of computational modeling uncertainty for complex engineering systems.

The validation metrics presented here should apply to a wide variety of physical systems in fluid dynamics, heat transfer, and solid mechanics. If the SRQ is a complex time-varying quantity, such as velocity at a point in a turbulent flow, then the quantity should be time-averaged to obtain a steady-state. If it is inappropriate to time-average the SRQ of interest and it has a periodic character or a complex mixture of many periods, such as modes in structural dynamics, then the present metrics would not be appropriate. These types of SRQs require sophisticated time-series analysis and/or mapping to the frequency domain. In addition, the present metrics

directly apply to single SRQs that are a function of a single input, or control, quantity. Future work will extend the present approach to metrics that would apply to single SRQs that are a function of multiple input quantities.

Acknowledgements

The authors thank Jon Helton, consultant to Sandia, for the enlightening discussions and insights. We also thank Tim Trucano, Basil Hassan, and Marty Pilch of Sandia for reviewing an earlier version of the paper and making a number of suggestions for improvement, and Harold Iuzzolino of Gram, Inc. for computational assistance. The cooperation of Fred Blottner, Kevin Dowding, Sheldon Tieszen, and Tim O'Hern of Sandia in sharing recent results used in the example problems is greatly appreciated. In addition, we thank the reviewers for the Journal of Computational Physics for providing in depth comments and constructive criticisms for improving the manuscript. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-AL85000.

References

[1] W.L. Oberkampf, T.G. Trucano, Validation methodology in computational fluid dynamics, AIAA Paper 2000-2549.
[2] W.L. Oberkampf, T.G. Trucano, Verification and validation in computational fluid dynamics, Progress in Aerospace Sciences 38 (2002) 209–272.
[3] AIAA, Guide for the Verification and Validation of Computational Fluid Dynamics Simulations, American Institute of Aeronautics and Astronautics, 1998. AIAA-G-077-1998.
[4] P.J. Roache, Need for control of numerical accuracy, Journal of Spacecraft and Rockets 27 (1990) 98–102.
[5] P.J. Roache, Verification and Validation in Computational Science and Engineering, Hermosa Publishers, Albuquerque, NM, 1998.
[6] P.J. Roache, Verification of Codes and Calculations, AIAA Journal 36 (1998) 696–702.
[7] W.L. Oberkampf, T.G. Trucano, C. Hirsch, Verification, validation, and predictive capability in computational engineering and physics, Applied Mechanics Reviews 57 (2004) 345–384.
[8] P. Knupp, K. Salari, Verification of Computer Codes in Computational Science and Engineering, Chapman & Hall/CRC, Boca Raton, FL, 2002.
[9] P.J. Roache, Code verification by the method of manufactured solutions, Journal of Fluids Engineering 124 (2002) 4–10.
[10] D. Pelletier, E. Turgeon, D. Lacasse, J. Borggaard, Adaptivity, sensitivity, and uncertainty: towards standards of good practice in computational fluid dynamics, AIAA Journal 41 (2003) 1925–1933.
[11] C.J. Roy, M.A. McWherter-Payne, W.L. Oberkampf, Verification and validation for laminar hypersonic flowfields, Part 1: Verification, AIAA Journal 41 (2003) 1934–1943.
[12] J. Peraire, A.T. Patera, Bounds for linear-functional outputs of coercive partial differential equations: local indicators and adaptive refinement, in: P. Ladeveze, J.T. Oden (Eds.), Advances in Adaptive Computational Methods in Mechanics, Elsevier, 1998.
[13] M. Ainsworth, J.T. Oden, A Posteriori Error Estimation in Finite Element Analysis, John Wiley, New York, 2000.
[14] I. Babuska, T. Strouboulis, The Finite Element Method and its Reliability, Oxford University Press, Oxford, UK, 2001.
[15] ASME, Council on Codes and Standards, Board of Performance Test Codes: Committee on Verification and Validation in Computational Solid Mechanics, American Society of Mechanical Engineers, 2003.
[16] W.L. Oberkampf, F.G. Blottner, Issues in computational fluid dynamics code verification and validation, AIAA Journal 36 (1998) 687–695.
[17] W.L. Oberkampf, Design, execution, and analysis of validation experiments, in: F. Grasso, J. Periauz, H. Deconinck (Eds.), Verification and Validation of Computational Fluid Dynamics, von Karman Institute for Fluid Dynamics, Rhode Saint Genese, Belgium, 2000.
[18] D.P. Aeschliman, W.L. Oberkampf, Experimental methodology for computational fluid dynamics code validation, AIAA Journal 36 (1998) 733–741.
[19] T.G. Trucano, M. Pilch, W.L. Oberkampf, General Concepts for Experimental Validation of ASCI Code Applications, Sandia National Laboratories, 2002, SAND2002-0341.
[20] C.J. Roy, W.L. Oberkampf, M.A. McWherter-Payne, Verification and validation for laminar hypersonic flowfields, Part 2: Validation, AIAA Journal 41 (2003) 1944–1954.
[21] J.S. Bendat, A.G. Piersol, Random Data: Analysis & Measurement Procedures, Wiley, New York, 2000.
[22] P. Wirsching, T. Paez, K. Ortiz, Random Vibrations: Theory and Practice, Wiley, New York, 1995.
[23] J.L. Devore, Probability and Statistics for Engineers and Scientists, Duxbury, Pacific Grove, CA, 2000.
[24] D.C. Montgomery, Design and Analysis of Experiments, John Wiley, Hoboken, NJ, 2000.
[25] R.G. Hills, T.G. Trucano, Statistical Validation of Engineering and Scientific Models: A Maximum Likelihood Based Metric, Sandia National Laboratories, 2002, SAND2001-1783.

[26] T.L. Paez, A. Urbina, Validation of mathematical models of complex structural dynamic systems, in: Proceedings of the Ninth International Congress on Sound and Vibration, Orlando, FL, 2002.
[27] R.G. Hills, I. Leslie, Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application, Sandia National Laboratories, 2003, SAND2003-0706.
[28] K.J. Dowding, R.G. Hills, I. Leslie, M. Pilch, B.M. Rutherford, M.L. Hobbs, Case Study for Model Validation: Assessing a Model for Thermal Decomposition of Polyurethane Foam, Sandia National Laboratories, 2004, SAND2004-3632.
[29] B.M. Rutherford, K.J. Dowding, An Approach to Model Validation and Model-Based Prediction–Polyurethane Foam Case Study, Sandia National Laboratories, 2003, SAND2003-2336.
[30] W. Chen, L. Baghdasaryan, T. Buranathiti, J. Cao, Model validation via uncertainty propagation, AIAA Journal 42 (2004) 1406–1415.
[31] A.B. Gelman, J.S. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman & Hall, London, 1995.
[32] J.M. Bernardo, A.F.M. Smith, Bayesian Theory, John Wiley, New York, 1994.
[33] T. Leonard, J.S.J. Hsu, Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers, Cambridge University Press, Cambridge, UK, 1999.
[34] K.M. Hanson, A framework for assessing uncertainties in simulation predictions, Physica D 133 (1999) 179–188.
[35] M.C. Anderson, T.K. Hasselman, T.G. Carne, Model correlation and updating of a nonlinear finite element model using crush test data, Paper No. 376, in: 17th International Modal Analysis Conference (IMAC) on Modal Analysis, Kissimmee, FL, 1999.
[36] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society Series B-Statistical Methodology 63 (2001) 425–450.
[37] T.K. Hasselman, G.W. Wathugala, J. Crawford, A hierarchical approach for model validation and uncertainty quantification, in: Fifth World Congress on Computational Mechanics, Vienna, Austria. Available from: <http://wccm.tuwien.ac.at>, 2002.
[38] B. DeVolder, J. Glimm, J.W. Grove, Y. Kang, Y. Lee, K. Pao, D.H. Sharp, K. Ye, Uncertainty quantification for multiscale simulations, Journal of Fluids Engineering 124 (2002) 29–41.
[39] M.J. Bayarri et al., A framework for validation of computer models, in: Proceedings of the Workshop on Foundations for Verification and Validation in the 21st Century, John Hopkins University/Applied Physics Lab., 2002.
[40] R. Zhang, S. Mahadevan, Bayesian methodology for reliability model acceptance, Reliability Engineering and System Safety 80 (2003) 95–103.
[41] T.L. Geers, An objective error measure for the comparison of calculated and measured transient response histories, The Shock and Vibration Bulletin 54 (1984) 99–107.
[42] D.M. Russell, Error measures for comparing transient data: Part I, development of a comprehensive error measure, in: Proceedings of the 68th Shock and Vibration Symposium, Hunt Valley, MD, 1997.
[43] D.M. Russell, Error measures for comparing transient data: Part II, error measures case study, in: Proceedings of the 68th Shock and Vibration Symposium, Hunt Valley, MD, 1997.
[44] H.W. Coleman, F. Stern, Uncertainties and CFD code validation, Journal of Fluids Engineering 119 (1997) 795–803.
[45] M.A. Sprague, T.L. Geers, Response of empty and fluid-filled, submerged spherical shells to plane and spherical, step-exponential acoustic waves, Shock and Vibration 6 (1999) 147–157.
[46] F. Stern, R.V. Wilson, H.W. Coleman, E.G. Paterson, Comprehensive approach to verification and validation of CFD simulations-Part 1: methodology and procedures, Journal of Fluids Engineering 123 (2001) 793–802.
[47] R.G. Easterling, Measuring the Predictive Capability of Computational Models: Principles and Methods, Issues and Illustrations, Sandia National Laboratories, 2001, SAND2001-0243.
[48] R.G. Easterling, Statistical Foundations for Model Validation: Two Papers, SAND2003-0287, Sandia National Laboratories, 2003.
[49] W.L. Oberkampf, M.F. Barone, Measures of agreement between computation and experiment: validation metrics, AIAA Paper 2004-2626.
[50] H.W. Coleman, W.G. Steele Jr, Experimentation and Uncertainty Analysis for Engineers, John Wiley, New York, 1999.
[51] J.R. Taylor, An Introduction to Error Analysis: The Sudy of Uncertainties in Physical Measurements, University Science Books, Sausalito, CA, 1997.
[52] W.J. Youden, Enduring values, Technometrics 14 (1972) 1–11.
[53] M.G. Morgan, M. Henrion, Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press, Cambridge, UK, 1990.
[54] R.L. Winkler, An Introduction to Bayesian Inference and Decision, Holt, Rinehart, and Winston, New York, 1972.
[55] A. Haldar, S. Mahadevan, Probability, Reliability, and Statistical Methods in Engineering Design, John Wiley, New York, 2000.
[56] M.L. Hobbs, K.L. Erickson, T.Y. Chu, Modeling Decomposition of Unconfined Rigid Polyurethane Foam, SAND99-2758, Sandia National Laboratories, 1999.
[57] D.K. Gartling, R.E. Hogan, M.W. Glass, Coyote – A Finite Element Computer Program for Nonlinear Heat Conduction Problems, Part I – Theoretical Background, Sandia National Laboratories, 1994, SAND94-1173.
[58] C.D. Pruett, T.B. Gatski, C.E. Grosch, W.D. Thacker, The temporally filtered Navier–Stokes equations: properties of the residual stress, Physics of Fluids 15 (2003) 2127–2140.
[59] P.E. DesJardin, T.J. O'Hern, S.R. Tieszen, Large eddy simulation of experimental measurements of the near-field of a large turbulent helium plume, Physics of Fluids 16 (2004) 1866–1883.
[60] S.R. Tieszen, S.P. Domino, A.R. Black, Validation of a simple turbulence model suitable for closure of temporally-filtered Navier–Stokes equations using a helium plume, Sandia National Laboratories, 2005, SAND 2005–3210.

[61] C.D. Moen, G.H. Evans, S.P. Domino, S.P. Burns, A multi-mechanics approach to computational heat transfer, IMECE-2002-33098, in: 2002 ASME International Mechanical Engineering Congress and Exhibition, New Orleans, LA, 2002.
[62] T.J. O'Hern, E.J. Weckman, A.L. Gerhart, S.R. Tieszen, R.W. Schefer, Experimental Study of a Turbulent Buoyant Helium Plume, SAND2004-0549J, Sandia National Laboratories, 2004.
[63] R.G. Miller, Simultaneous Statistical Inference, Springer-Verlag, New York, 1981.
[64] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley, New York, 1998.
[65] G.A.F. Seber, C.J. Wild, Nonlinear Regression, John Wiley, New York, 2003.
[66] M.F. Barone, W.L. Oberkampf, F.G. Blottner, Validation of compressibility corrections for two-equation turbulence models, AIAA Journal (in press).
[67] D.W. Bogdanoff, Compressibility effects in turbulent shear layers, AIAA Journal 21 (1983) 926–927.
[68] S.K. Lele, Compressibility effects on turbulence, in: J.L. Lumley, M. Van Dyke (Eds.), Annual Review of Fluid Mechanics, Annual Reviews, Inc., Palo Alto, CA, 1994.
[69] D. Papamoschou, A. Roshko, The compressible turbulent shear layer: an experimental study, Journal of Fluid Mechanics 197 (1988) 453–477.
[70] N. Chinzei, G. Masuya, T. Komuro, A. Murakami, K. Kudou, Spreading of two-stream supersonic turbulent mixing layers, Physics of Fluids 29 (1986) 1345–1347.
[71] M. Samimy, G.S. Elliott, Effects of compressibility on the characteristics of free shear layers, AIAA Journal 28 (1990) 439–445.
[72] G.S. Elliot, M. Samimy, Compressibility effects in free shear layers, Physics of Fluids A 2 (1990) 1231–1240.
[73] S.G. Goebel, J.C. Dutton, Experimental study of compressible turbulent mixing layers, AIAA Journal 29 (1991) 538–546.
[74] J.C. Dutton, R.F. Burr, S.G. Goebel, N.L. Messersmith, Compressibility and mixing in turbulent free shear layers, in: 12th Symposium on Turbulence, Rolla, MO, 1990.
[75] M.R. Gruber, N.L. Messersmith, J.C. Dutton, Three-dimensional velocity field in a compressible mixing layer, AIAA Journal 31 (1993) 2061–2067.
[76] J.R. Debisschop, J.P. Bonnet, Mean and fluctuating velocity measurements in supersonic mixing layers, in: W. Rodi, F. Martelli (Eds.), Engineering Turbulence Modeling and Experiments 2: Proceedings of the Second International Symposium on Engineering Turbulence Modeling and Measurement, Elsevier, New York, 1993.
[77] J.R. Debisschop, O. Chambers, J.P. Bonnet, Velocity-field characteristics in supersonic mixing layers, Experimental Thermal and Fluid Science 9 (1994) 147–155.
[78] S. Barre, P. Braud, O. Chambres, J.P. Bonnet, Influence of inlet pressure conditions on supersonic turbulent mixing layers, Experimental Thermal and Fluid Science 14 (1997) 68–74.
[79] D.C. Wilcox, Turbulence Modeling for CFD, DCW Industries, 2002.
[80] Y. Nagano, M. Hishida, Improved form of the k-epsilon model for wall turbulent shear flows, Journal of Fluids Engineering 109 (1987) 156–160.
[81] O. Zeman, Dilatation dissipation: the concept and application in modeling compressible mixing layers, Physics of Fluids A 2 (1990) 178–188.
[82] C.C. Wong, F.G. Blottner, J.L. Payne, M. Soetrisno, Implementation of a parallel algorithm for thermo-chemical nonequilibrium flow solutions, AIAA Paper 95-0152.
[83] C.C. Wong, M. Soetrisno, F.G. Blottner, S.T. Imlay, J.L. Payne, PINCA: A Scalable Parallel Program for Compressible Gas Dynamics with Nonequilibrium Chemistry, SAND94-2436, Sandia National Laboratories, 1995.
[84] H.C. Yee, Implicit and Symmetric Shock Capturing Schemes, NASA-TM-89464, NASA, 1987.
[85] S. Yoon, A. Jameson, An LU-SSOR Scheme for the Euler and Navier–Stokes Equations, AIAA Paper 87-0600.
[86] P.J. Roache, Perspective: a method for uniform reporting of grid refinement studies, Journal of Fluids Engineering 116 (1994) 405–413.
[87] I. Celik, W.M. Zhang, Calculation of numerical uncertainty using Richardson extrapolation: application to some simple turbulent flow calculations, Journal of Fluids Engineering 117 (1995) 439–445.
[88] R. Paciorri, F. Sabetta, Compressibility correction for the Spalart–Allmaras model in free-shear flows, Journal of Spacecraft and Rockets 40 (2003) 326–331.
[89] MathWorks, Matlab, The MathWorks, Inc.