Validation of imprecise probability models

Scott Ferson, William L. Oberkampf and Lev Ginzburg
Applied Biomathematics (scott@ramas.com), Sandia National Laboratories, Stony Brook University

Abstract: Validation is the assessment of the match between a model's predictions and any empirical observations relevant to those predictions. This comparison is straightforward when the data and predictions are deterministic, but is complicated when either or both are expressed in terms of uncertain numbers (i.e., intervals, probability distributions, p-boxes, or more general imprecise probability structures). There are two obvious ways such comparisons might be conceptualized. Validation could measure the discrepancy between the *shapes* of the uncertain numbers representing prediction and data, or it could characterize the differences between *realizations* drawn from the respective uncertain numbers. When both prediction and data are represented with probability distributions, comparing shapes would seem to be the most intuitive choice because it sidesteps the issue of stochastic dependence between the prediction and the data values which would accompany a comparison between realizations. However, when prediction and observation are represented as intervals, comparing their shapes seems overly strict as a measure for validation. Intuition demands that the measure of mismatch between two intervals be zero whenever the intervals overlap at all. Thus, intervals are in perfect agreement even though they may have very different shapes. The unification between these two concepts relies on defining the validation measure between prediction and data as the shortest possible distance given the imprecision about the distributions and their dependencies.

Keywords: validation, observation, prediction, distribution, interval, p-box

1. Introduction

Validation is the comparison of the predictions of a theory or model against empirical data (AIAA 1998; ASME 2006; Oberkampf and Truncano 2002; Oberkampf et al. 2004; Oberkampf and Barone 2006; Hills 2006; Trucano et al. 2006; Romero 2007; Ferson et al. 2008). It is often contrasted with verification, which is the checking of a model's implementation against the intended specification (Oberkampf et al. 2004; Oberkampf and Trucano 2007). We also contrast validation with calibration, which is the adjustment of the model's parameters or its structure for the purpose of improving the match between its predictions and empirical reality (Kennedy and O'Hagan 2001; Trucano et al. 2006). Measures of validation might be useful in a calibration, but the processes are entirely different in their goals. Calibration seeks to correct a model, and validation seeks only to measure how correct the model is.

Several approaches to validation have recently been suggested based on simple comparisons of trends in means (e.g., Oberkampf and Barone 2006), more elaborate hypothesis testing (e.g., Hills and Trucano 2002; Hills and Leslie 2003; Rutherford and Dowding 2003; Chen et al. 2004;

Dowding et al. 2004), or still more comprehensive Bayesian schemes (e.g., Hanson 1999; Kennedy and O'Hagan 2001; Hazelrigg 2003; Zhang and Mahadevan 2003; O'Hagan 2006; Chen et al. 2006; 2007). This paper concerns only the basic question of how we should summarize and measure the discrepancies between a model's predictions and relevant empirical data. Oberkampf and Truncano (2007) called this problem the 'validation assessment'. Other important issues such as how such the measure could be used to inform or quantify the predictive capability of a model or deciding whether the model is adequate for some intended use are out of our present scope.

We consider validation assessment in a context where non-negligible uncertainty is present in the prediction or the data, or both. This uncertainty can come in different forms. It may arise from natural stochasticity or randomness in the world, perhaps owing to fluctuations in processes across space or through time, heterogeneity of individuals, or variability among engineered components. This uncertainty is objective in the sense that it exists irrespective of observation by humans and it is irreducible in the sense that empirical study does not necessarily reduce it. We call it aleatory uncertainty and recognize traditional probability theory as the primary calculus for addressing it. Aleatory uncertainty is often contrasted with epistemic uncertainty which is the partial ignorance, incertitude or imprecision that arises from incomplete or imperfect scientific study and comes from small sample sizes, missing data or data censoring or other measurement uncertainties, and perhaps doubt about the proper form of a model. Epistemic uncertainty is sometimes called subjective or reducible uncertainty because it's a function of the observer rather than physical reality and because it can in principle be reduced by empirical effort. Although probability theory has often been used to address epistemic uncertainty, other approaches are also employed, notably including interval analysis.

Recently, several researchers have suggested that methods beyond traditional probability theory might be necessary for models that must distinguish aleatory and epistemic uncertainty (Shafer 1976; Walley 1991; Klir and Wierman 1999; Oberkampf et al. 2001; Nikolaidis and Haftka 2001; Ferson et al. 2003; Helton and Oberkampf 2004; inter alia). We use the phrase 'uncertain number' (Ferson et al. 2003) to denote a varying or imperfectly known quantity that is mathematically characterized by an interval, probability distribution, p-box (Ferson et al. 2003), Dempster-Shafer structure (Shafer 1976; Oberkampf et al. 2001; Oberkampf and Helton 2005), random set (Matheron 1975; Molchanov 2006), set of probability measures or 'credal set' (Levi 1980), or similar structure from the theory of imprecise probabilities (Walley 1991). In general, an uncertain number can express both aleatory uncertainty and epistemic uncertainty. One might hold that a probability distribution, as a special case of an uncertain number, expresses purely aleatory uncertainty and an interval, also a special case, expresses purely epistemic uncertainty.

The engineering value of a model's quantitative prediction is a function of both its accuracy and its precision. The precision of a prediction expressed as an uncertain number is inversely related to the epistemic uncertainty encoded in the uncertain number. This uncertainty is sometimes called 'non-specificity' (Klir and Wierman 1999) and might be quantified as the width of an interval or the breadth between the left and right bounds of a p-box. A validation assessment

lets us quantify the second essential component determining the worth of the prediction: its accuracy in the face of empirical evidence.

Section 2 considers validation for the case where both prediction and data are represented by probability distributions. Section 3 considers the more elementary problem of validation when they are both intervals. Section 4 tries to harmonize the measures developed for these two special cases. Section 5 considers some alternative solutions, and section 6 offers some conclusions.

2. Validation Metric for Comparing Probability Distributions

The difference between two probability distributions can be characterized in many ways. The comparison could be conceived in terms of differences of their realizations (i.e., real numbers) or in terms of the discrepancies between their distribution shapes. In other words, if X and Y are random numbers distributed according to their respective cumulative distribution functions F and G, then we could talk about the distribution or average of $X - Y$, or we could focus on the difference between the shapes of F and G. The characterization that seems to be most useful in the context of validation of engineering models is based on comparing the shapes of the distributions of the random variables representing the prediction and relevant observations. Random variables whose distribution functions are identical are said to be 'equal in distribution'. If the distributions are not quite identical in shape, the discrepancy can be measured with any of many possible measures that have been proposed for various purposes in fields including statistical goodness of fit (e.g., Stephens 1974; Feller 1948; Kolmogorov 1941; Smirnov 1939), probability scoring rules (Winkler 1996; Lindley et al. 1979; de Finetti 1962; Brier 1950), information theory (Song 2002; Kullback 1959; Kullback and Leibler 1951), and texture analysis (e.g., Mathiassen et al. 2002).

Ferson et al. (2008) proposed to quantify the mismatch between prediction and observation with the area between the prediction's probability distribution and the empirical distribution of observations. This area is the Minkowski L_1 metric

$$d(F, S_n) = \int_{-\infty}^{\infty} \left| F(x) - S_n(x) \right| \mathrm{d}x ,$$

where F is the cumulative distribution representing the model's prediction for the random variable and S_n is the empirical distribution function for relevant observations X_i, $i = 1,\ldots,n$, of that random variable. The empirical distribution function is

$$S_n(x) = \frac{\#\left\{X_i \le x\right\}}{n}$$

where # denotes the cardinality of the set, so $S_n(x)$ is the fraction of values in the data set that are at or below each magnitude x. The validation metric is thus computed solely from the prediction

F provided by the modeler and observations X_i provided by the empiricist. A small area means there is a good match, and a large area means that prediction and data disagree.

Figure 1 illustrates an example prediction distribution for rainfall as the smooth curve drawn in gray, together with the empirical distribution functions S_n for a hypothetical data set consisting of the values 770, 790, 820, 865 in millimeters of rain. The prediction distribution is approximately normal, with mean about 810 mm and variance of about 1700. The area of the shaded region between the two functions which measures their disagreement is almost 40 mm. Note that the empirical distribution function is zero for all values smaller than the minimum of the data and one for all values larger than the maximum of the data. Likewise, beyond the range of the prediction distribution, the value of $F(x)$ is either zero or one extending to infinity in both directions. For graphical clarity, however, these flat portions at probability zero or one are not depicted when the distributions are plotted.

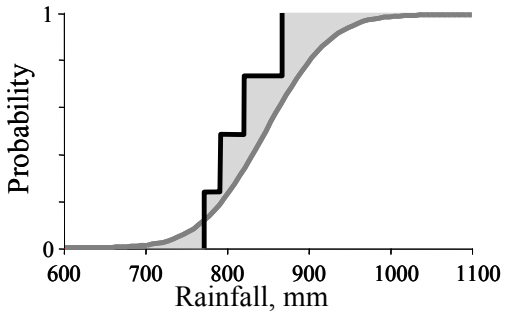

Figure 1. Area (shaded) between a prediction distribution (gray) and an empirical distribution function (black).

This metric can be computed for small data sets or even a single data value, in which case the S_n function would be the unit step function at that value. The approach can also be used even when the model is so complex and computationally expensive that it can only generate a small number of realizations for its prediction distribution. In such situations, the prediction distribution is modeled with an 'empirical' distribution formed from the sample realizations.

The area between the prediction distribution and the empirical distribution summarizing observations has several desirable properties as a formal validation measure of the mismatch between a model and evidence (Ferson et al. 2008). Most importantly, the area metric is an objective measure. Given a collection of observations and a prediction distribution, the area will be the same no matter who computes it because it does not depend on any judgments or parameters chosen by the analyst. Another important property is that the area metric generalizes deterministic comparisons between scalar values that have no uncertainty; if the prediction and the observation are both scalar point values, the area is equal to their difference. The area will tend not to be overly sensitive to minor discrepancies in the distribution tails (assuming the area is finite), but it obviously reflects the full distributions in assessing performance. In particular, it is

clearly not merely a measure of the difference in the means or even the means and variances, but takes account of any differences between the prediction and observation distributions. Because probability is dimensionless, the units of the area are the same as those of the system response quantity in which the prediction and data are expressed. This property is very important in making the measure intuitively meaningful to engineers. Its units are the same as one would expect for the result of a subtraction. If it were some dimensionless index or, worse, had some complex or esoteric statistical units, its physical interpretation would be difficult. The area measure is also unbounded in the sense that, if the prediction is completely off the mark of the observations, the area characterizing this discrepancy can in principle grow to be an arbitrarily large value, which is also an intuitive feature of distances. Finally, the area measure is mathematically well behaved and well understood. So long as the area converges to a finite value, it is a true metric in the mathematical sense, which means it has the essential features of a distance function. By definition, a mathematical metric d has four properties (Fréchet 1906):

non-negativity,	$d(x, y) \geq 0$,
symmetry,	$d(x, y) = d(y, x)$,
triangle inequality,	$d(x, y) + d(y, z) \geq d(x, z)$, and
identity of indiscernibles,	$d(x, y) = 0$ if and only if $x = y$.

All of these properties suggest that the area metric will be more comprehensive and easier to interpret than any of several alternative statistical measures or some distance measure based on merely matching prediction and observation distributions in the mean or in both mean and variance.

Ferson et al. (2008) also showed how the area metric could be extended to synthesize evidence of the conformance between model and data into a single measure when observations are to be compared to *different* prediction distributions. The trick is to transform each observation X_i to $u_i = F_i(X_i)$ where F_i is the prediction distribution against which X_i is to be compared. The u_i express all the available evidence on a universal scale of probability. By the probability integral transform theorem (Angus 1994), the u_i will be uniformly distributed over the unit interval [0,1] so long as the original X_i are distributed according to their respective prediction distributions F_i, which is to say, so long as the model is predicting the observations well. Statistical tests and diagnostics are straightforward to define for this synthesis. The model's performance can be assessed directly in terms of the u_i, or the values may first be back-transformed to a common axis that re-expresses the evidence in physical units. The back-transformation can be chosen so as to maximize the relevance of the assessment for a particular regulatory or performance question. This strategey can even be used to combine evidence about model-data conformance collected in entirely different dimensions (such as, for instance, rainfall and temperature). This synthesis abandons the interpretation of the area in original units of course, but it does allow analysts to compare the relative performance of the model for different system response quantities in a meaningful way.

2.1. WHY NOT BASE THE METRIC ON DIFFERENCES OF VALUES FROM THE TWO DISTRIBUTIONS?

One could imagine developing an alternative validation measure based on the absolute difference between a random value realized from the prediction distribution and a random value drawn from the data distribution. There would of course be a distribution of such differences. It might seem preferable to use this distribution of differences to characterize the disagreement between probability distributions (Menger 1942). A distribution could be more informative than the area metric which is a crude scalar summary that could not capture the information embodied in an entire distribution. The distribution of differences could be used itself as a characterization of the disagreement between the two distributions, or it might be summarized in various ways that might highlight aspects of the disagreement of special interest. But such a notion would need to consider the *stochastic dependence* between random values from the two distributions. Specifying an assumption about the dependence is necessary to define the distribution of differences $X - Y$ from specified distributions for X and Y. Are the values statistically independent? Do they have some correlation or a nonlinear dependence? Different assumptions can lead to starkly different distributions for the random difference.

Consider, for example, a weather model that predicts daily temperatures and, by aggregating these values, also predicts a distribution of daily temperatures over the course of a year. Suppose that relevant daily temperature observations are available. It may be the case that the predicted distribution of temperatures over the year matches the observed distribution of temperatures very well and yet the correlation between predicted and observed daily temperatures is markedly poor. For instance, if the model is out of phase with respect to seasons, it may be predicting summer temperatures during the winter and vice versa, which would lead to a correlation close to –1, even though it gets the distributions exactly right. The performance of such a model would have to be considered very poor in any sensible validation assessment. But note that this poor performance is really associated with the deterministic results from the model rather than the probabilistic ones per se. If the model had not made the deterministic predictions and confined itself to purely probabilistic forecasts, this problem would not have arisen.

Contrast the weather model with another model that does not predict individual daily temperatures, but only the summary *distribution* of daily temperatures. Essentially, this retreat changes the weather model into a climate model that does not make predictions about the temperature for any particular day, except to assert that, considered as a group over the course of many days, these temperatures will converge in distribution to the prediction. And it is certainly not making any predictions about the dependence between values that might be drawn from the prediction distribution and observed temperature values. The model is not even saying that such temperature pairs are independent. In fact, actual temperatures have strong autocorrelation from day to day, so supposing that temperatures should be drawn independently from the predicted distribution would obviously be empirically incorrect too. It is possible, of course, to construct a probabilistic weather model of daily temperatures. Such a model might predict a probability distribution for each and every day's temperature. But these predictions would not be saying

anything about dependence or even about randomness; they are asserting only that $F_i(X_i)$ are uniformly distributed, where F_i is the probability prediction for day i and X_i is the observed temperature for that day. In any case, if the model refrains from making deterministic forecasts and makes only purely probabilistic predictions about distributions without characterizing dependence, then the model would have excellent performance in a validation assessment.

If the model asserts nothing about the possible dependence between predicted and observed values of a system response quantity, then the distribution of differences between predictions and data cannot be uniquely defined. Thus, it would be seem to be impossible to base a validation metric on the distribution differences. It is possible, however, to *bound* the distribution of absolute differences even without specifying anything about the dependence between the subtrahend and the minuend. Elementary probability bounds analysis (Frank et al. 1987; Williamson and Downs 1990; Ferson 2002; Ferson et al. 2003) can be used to compute these bounds, which may be informative. Figure 2 depicts four examples of validation as characterized by the area metric and bounds on the distribution of differences. In the upper panel of graphs, prediction distributions F are depicted as gray curves, and data distributions S_n are depicted as black step functions. Under each of these four graphs, the corresponding area metric is plotted as a dotted spike. On the same graph, bounds on the distribution of absolute differences between random values from the prediction and data distributions are shown as thin lines. In each of the four comparisons, the prediction is a normal distribution with mean 2 and standard deviation 0.2, truncated at the 0.5^{th} and 99.5^{th} percentiles. In the first comparison, the data consists of a single observed value at 4, so the empirical distribution is degenerate. The validation metric in this case is 2 units, which is the area between the truncated normal and this degenerate step function. The distribution of absolute differences between random values drawn from the prediction distribution and the observed value 4 ranges between 1.5 and 2.5. The data forming the empirical distribution in the second comparison comes from 8 measurements scattered between roughly 2.2 and 3.2. The area validation metric in the second comparison is almost 0.6 units. Without specifying the stochastic dependence between the prediction and data distributions, it is impossible to define the distribution of their differences, but probability bounds analysis can bound the distribution (Ferson 2002). The thin lines in the second graph of the lower panel of Figure 2 represent the best-possible bounds on the distribution of absolute differences between predicted values and observed values. The breadth of the bounds comes from not making any assumption about the dependence between the two distributions. In the third comparison, the empirical data have a larger dispersion and the resulting area metric is somewhat larger. In the fourth comparison, the data values come much closer to the prediction distribution, so the area metric is much closer to zero. Note, however, that the distribution of differences could nevertheless include values close to one.

These few examples convey an idea for how the area metric and the bounds on the distribution of differences compare to each other. The bounds tell us how wrong we might be if dependence matters, but they do not contain the information needed to compute the area metric, so, insofar as the area metric is important or informative, the bounds on differences are

incomplete as a summarization of the disagreement. Likewise, the bounds contain information not encapsulated in the area metric as well, although engineering judgment does not seem to recognize the information in the bounds as particularly relevant to the question of whether the distributions match well.

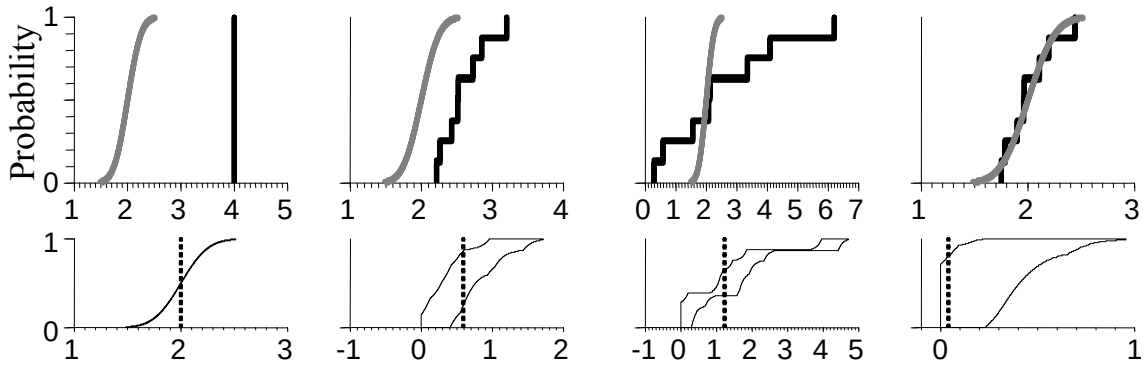

Figure 2. Predictions (gray) and data (black) yielding area metrics (dotted) and difference distributions (thin p-boxes).

3. Validation Measure for Comparing Intervals

Predictions should include epistemic uncertainty if it exists in our knowledge about the modeled physical process. Indeed, except in rare situations, precise predictions are not reasonable in real-world problems, or they only result from assumptions that modelers themselves do not unequivocally believe. Although a model may give point predictions, there is almost always an implied precision associated with each quantity. Modern notions of best practice argue that these implicit considerations be made explicit, and more and more modelers are accepting this and incorporating uncertainty analyses into their models. The simplest quantitative expression of epistemic uncertainty is an interval. Giving an interval as the representation of an estimated quantity is asserting that the value (or values) of the quantity lie somewhere within the interval. Intervals can arise in both predictions and observations. When a prediction is an interval, its width relates the modeler's inability to nail down the prediction precisely. The modeler is saying the quantity in question is within a particular range, but not saying any more than this. In particular, the modeler is not making any assertion about which possible values might be more likely than which other possible values. If there is such extra information available about a prediction, but too little to justify the selection of a particular probability distribution, the information can be expressed in a more general uncertain number such as a p-box, Dempster-Shafer structure or credal set.

Empirical observations can also contain epistemic uncertainty. Again, the simplest form of this is an interval. Uncertainty about measurements that is appropriately characterized by

intervals is called incertitude, and it arises naturally in a variety of circumstances, including plus-or-minus reports, significant digits, intermittent measurement, non-detects, censoring, data binning, rounding or bit compression in data transmission, missing data and gross structural ignorance (Ferson et al. 2007; 2004). When a collection of such intervals comprise a data set, one can think of the breadths of the intervals as representing epistemic uncertainty while the scatter among the intervals represents variability or aleatory uncertainty. Recent reviews (Manski 2003; Gioia and Lauro 2005; Ferson et al. 2007) have described how interval uncertainty in data sets produces uncertain numbers containing epistemic uncertainty. When empirical observations have uncertainty of this form that is too large to simply ignore, these elementary techniques can be used to characterize it is a straightforward way.

The comparison between two fixed real numbers reduces to the scalar difference between the two. Suppose that, instead of both numbers being reals, at least one of them is an interval range representing acknowledged uncertainty. If the prediction and the observation overlap, then we should say that the prediction is *correct*, in an important sense, relative to the observation. If the prediction is an interval, this means that the model, or perhaps one would say the modeler, is being modest about what is being claimed. For example, the assertion that a regional maximum temperature will be between 20 and 40 °C is a weaker claim than saying it will be exactly 30. And it is a stronger claim than saying the temperature will be between 10 and 60. In the extreme case, a vacuous prediction, while not very useful, is certainly true, if just because it isn't claiming anything that might be false. For example, predicting that some probability will be between zero and one doesn't require any bravery, but at least it is free from contradiction. It is proper that a prediction's express uncertainty be counted toward reducing any measure of mismatch between theory and data in this way because the model(er) is admitting doubt. If it were not so, an uncertainty analysis could otherwise have no epistemological value. From the perspective of validation, when the uncertainty of prediction encompasses the actual observation, the prediction ought to be regarded as true, because *validity is distinct from precision*. Both are important in determining the usefulness of a model, but it is reasonable to distinguish them and give credit where it is due.

A reciprocal consideration applies, by the same token, if the datum is an interval to be compared against a prediction that's a real number. Validation has to give to the model whatever benefit of the doubt that arises because of the uncertainty about the datum. For instance, if the prediction is, say, 30% and the observation tells us that it was somewhere between 20% and 50%, then we would have to admit that the prediction might be perfectly correct. If on the other hand the evidence was that it was between 35% and 75%, then we would have to say that the disagreement between the prediction and the observation might be as low as 5%. We could also be interested in how bad the comparison might be, but a validation metric shouldn't penalize the model for the empiricist's imprecision. In most conceptions of the word, the 'distance' between two things is the length of the shortest path between them. The distance between England and France is the breadth of the English Channel between Dover and Calais; it doesn't matter that Newcastle and Marseilles are much further apart. Similarly, the validation measure between a

point prediction and an interval datum, or vice versa, should be the shortest difference between the characterizations of the quantities. Likewise, the validation measure between an interval prediction and an interval datum is the shortest distance between the two intervals, which will be zero if they overlap. Symbolically, the validation measure for comparing intervals A with B is

$$\inf_{\substack{X \in A \\ Y \in B}} |X - Y|.$$

where inf denotes the infimum (which just generalizes minimum for intervals that might be open or partially open). Although this choice for the validation measure shares a similar graphical intuition with the area metric discussed in section 2, this measure is quite different from it. Note, for instance, that this measure is not a mathematical metric. It violates the property of identity of indiscernibles, because a value of zero for the measure does not imply that the intervals are identical. Mathematicians call a non-negative, symmetric function that satisfies the triangle inequality but not identity of indiscernibles a 'pseudometric'. More fundamentally, this measure is not based on the shapes of the intervals like the area metric was based on the shapes of the probability distributions. Indeed, the shape of the intervals could be wildly different yet still yield a value of zero for the validation measure if they overlap at all. In fact, the formula above suggests that the measure is based instead on considering possible *realizations* of values X and Y from the respective intervals.

4. Unification of the Two Conceptualizations for General Uncertain Numbers

The key to harmonizing the shape-based comparison described in section 2 with the realization-based comparison described in section 3 is to recognize that both are essentially special cases of the Wasserstein distance (Vallender 1973; Dobrushin 1970)

$$\inf_{\substack{X \sim F \\ Y \sim Sn}} \mathrm{E}|X - Y|,$$

where the E denotes the expectation operator, and the infimum is taken over all possible random variables X and Y that are distributed according to F and S_n respectively. When the prediction F and the data distribution S_n are probability distributions, the infimum searches over all possible stochastic dependencies between the random variables X and Y (constrained by the fact that they must respect their marginal distributions F and S_n). The Wasserstein distance is a metric for any distributions for which the infimum is finite (Dobrushin 1970). When the random variables are univariate, then it equals the area metric (Vallender 1973). The infimum occurs when the X and Y are comonotonic, that is, when the dependence between X and Y is perfect, and the correlation between them is a large as is possible given their marginal distributions. It is this fact that creates the graphical interpretation as the area between the distributions.

When the prediction and data are intervals, we interpret the tilda to mean 'is an element of' and ignore the E operator (because intervals do not have probability measures defined over them) so that the Wasserstein distance is the same as our intuitive formula for the validation measure for intervals described in section 3.

The generalization of the Wasserstein distance for uncertain numbers is now clear: it should be the infimum expectation of the absolute value of the difference between the variates, where the infimum is taken over all possible distribution with respective uncertain numbers *and* under all possible dependencies between those distributions. The computational task of identifying this infimum may be challenging for some uncertain numbers such as credal sets, but it turns out to be rather simple for p-boxes. The area measuring mismatch for general p-boxes is the integral

$$\int_{-\infty}^{\infty} \Delta([F_{\mathrm{R}}(x), F_{\mathrm{L}}(x)], [S_{n\mathrm{R}}(x), S_{n\mathrm{L}}(x)])\, \mathrm{d}x$$

where F and S_n denote the prediction and the data distributions, respectively, and the subscripts L and R denote the left and right bounds for those distributions, and

$$\Delta(A, B) = \min_{\substack{a \in A \\ b \in B}} | a - b |$$

is the shortest distance between two intervals, or zero if the intervals touch or overlap. This measure integrates the regions of non-overlap between the two sets of bounds, for every value along the probability axis.

The thin p-boxes in the lower panel of graphs in Figure 2 are bounds on all possible distributions of the difference between the two random values. Instead of all possible distributions, we want the mean of the precise distribution of differences assuming perfect dependence between the prediction F and data distribution S_n. We might therefore characterize this measure as the *mean perfect absolute difference of deviates*, but perhaps it will suffice to continue to call it the 'area measure'. It is important to keep in mind that we're *not* selecting perfect dependence as our model of how the prediction and observation distributions are expected to be related to each other. Perfect dependence would mean that locally large observations would always be associated with locally large predicted values, and small with small, in a very strict fashion. We certainly do not believe that they would be related in this way in reality. Perfect dependence just falls out of the formula because it is the dependence that leads to the smallest possible value of the mean of the absolute differences. The smallest area is the one of interest because the distance between two things is the length of the shortest connection between them. At least for p-boxes, this also has the happy graphical interpretation as the area between the prediction and the observation.

Figure 3 depicts four more examples. As before, predictions are depicted in the upper panel in gray, and data are depicted in black, but now they are p-boxes rather than precise distributions. Under each of these four graphs, the corresponding area measure is shown as a dotted spike and

bounds on the all distribution of differences between random values from the prediction and observation p-boxes are shown as thin lines. In each of the four comparisons, the prediction is a p-box of normal distributions whose means are in the interval [1.75, 2.25] with standard deviation 0.2, truncated at the 0.5^{th} and 99.5^{th} percentiles. In the first comparison, the data consists of a single interval [4,5], so the resulting area measure is 1.75. It is the area between the rightmost normal distribution inside the gray p-box and the leftmost scalar inside the black interval. It seems reasonable that the discrepancy between the prediction and data in this case is only 1.75 units even though the difference between a predicted value and an observed data value could be larger than 3.5 units. The wide breadth of the bounds on the differences comes from the epistemic uncertainty about the prediction distribution and the data distribution within their respective p-boxes and also from not making any assumption about the dependence between them. The data in the second comparison comes from 8 measurements for which measurement incertitude was ±0.25. The 8 intervals implied by this incertitude were cumulated into a p-box describing epistemic uncertainty about the empirical distribution function (Ferson et al. 2007). The area in the second comparison is about 0.34, which is the between the right edge of the graph prediction and left edge of the black data p-box. In the third comparison, the empirical data had the same sample size and the same incertitude as in the second comparison, but the values happened to have a larger dispersion. In this case, the area is the sum of the two areas where the gray prediction p-box and the black data p-box do not overlap. In the fourth comparison, the data values had the same measurement uncertainty but a smaller dispersion and central tendency so the area measure is zero because there exist distributions that lie within both the prediction and data p-boxes.

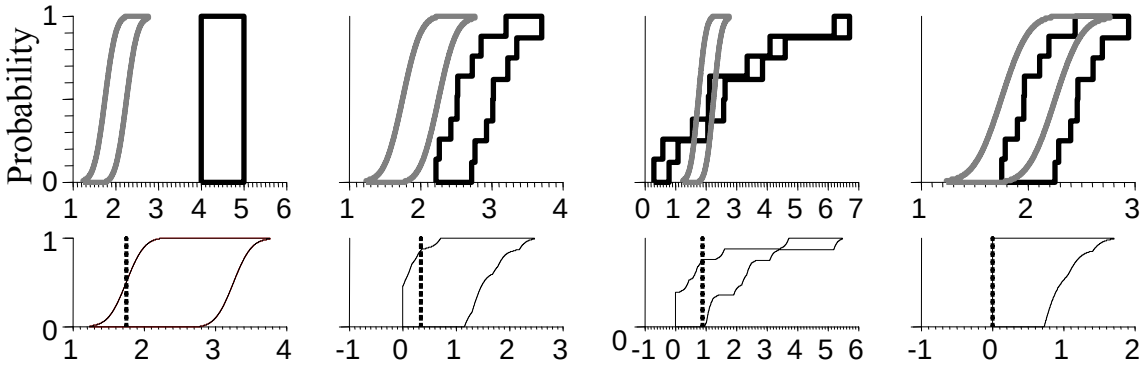

Figure 3. Predictions (gray) and data (black) yielding area measures (dotted) and difference p-boxes (thin).

5. 'Same Shape' versus 'Possibly Equal'

Although we think that using the area distance when the prediction and observations are uncertain numbers as described in the previous section is appropriate both mathematically and in practical engineering terms, we acknowledge that there are several other ways this generalization could be conceived. This section introduces three alternative generalizations of the area metric for use when uncertain numbers are used to characterize predictions or observations.

The area metric proposed in section 2 is based on the distribution functions of the predictions and the data, as distinguished from the random variables those distributions summarize. Although we chose to compare the shapes of the probability distributions when the quantities had only aleatory uncertainty, this choice does not seem satisfactory when there is epistemic uncertainty present as well. The area measure between the prediction and data in the general case as described in section 4, is no longer a mathematical metric when at least one is an interval or a more general uncertain number because the area can fall to zero without the prediction and data becoming identical (as in the rightmost graph of Figure 3). In section 4, the application of the area measure when prediction and data are characterized as uncertain numbers was based on the conventional idea that distance between two things is the length of the *shortest* line between them. There are, however, different ways to look at the question. A standard mathematical way to construct a metric between two potentially overlapping sets is to define

$$\max\left(\sup_{F \in x} \inf_{G \in y} d(F,G), \sup_{G \in y} \inf_{F \in x} d(G,F) \right)$$

where F is an element of the first set x and G is an element of the second set y and d is a metric on the space containing the sets (Pompieu 1905), which is our case would just be the area metric. The elements F and G are possible distribution functions taken from the respective prediction and data uncertain numbers x and y. This function is zero if and only if the set of distributions representing the prediction is the same as the set of distributions representing the data, that is, if their respective uncertain numbers had identical shapes. This function constitutes a much stricter view about agreement between prediction and data. It holds that perfect agreement involves not only overlapping but having exactly the same imprecision. Generalizing the area distance using this function would mean that our measure would remain a true mathematical metric, but it seems overly strict about what constitutes perfect agreement. For instance, suppose that the theoretical prediction is a simple interval and is to be compared with an observation that is also an interval and that the prediction interval is a *subset* of the observation interval. In other words, the prediction and observation agree in that they overlap, but the imprecision about the observation is wider than that of the prediction. It doesn't seem reasonable to insist that the theory and data are somehow not in perfect agreement in this situation, nor to require that the theory somehow inflate the uncertainty of its prediction simply to match the poorer precision associated with the observation.

Another way to generalize the area metric for uncertain numbers considers comparisons between distributions realized from the uncertain numbers, rather than the shapes of the uncertain numbers. For example, it might be natural to find upper and lower bounds on the areas between distributions that are consistent with the two uncertain numbers. Rather than differences between pairs of bounds, this would be bounds on differences between pairs of distributions. In this case, the measure would be the smallest and largest possible values of the underlying metric

$$\left[\inf_{F \in x, G \in y} d(F,G), \sup_{F \in x, G \in y} d(F,G) \right],$$

where F and G are distribution functions within (consistent with) the respective uncertain numbers x and y. The range would be degenerate, i.e., the infimum and supremum would be the same if the two uncertain numbers are actually particular probability distributions, neither having any epistemic uncertainty. The range being double-zero would mean that the prediction and the data distribution are identical, and that neither has any epistemic uncertainty. This generalization is not a metric because it does not have the property of identity of indiscernibles; x and y could be identical and not yield a double-zero.

Note that this scheme, like the Pompieu scheme, can be very difficult computationally because there are infinitely many distributions within the uncertain numbers to be compared. It obviously does not suffice to compare extreme distributions corresponding to the edges of the uncertain numbers. For example, consider the leftmost graph of Figure 4. It is intuitively clear that that the smallest possible value of the area between a distribution inside the prediction bounds and a distribution inside the observation bounds corresponds to the shaded area. This area corresponds to a prediction distribution that follows the left edge of the prediction bounds (smooth gray bounds) for small probability levels and follows the right edge of the prediction bounds for large probabilities. The corresponding distribution consistent with the observation bounds (black step bounds) conversely follows the right edge of those bounds for small probabilities and the left edge for large probabilities. For intermediate probabilities, the prediction distribution and the empirical distribution are coincident monotone curves in the region where the bounds overlap. The *largest* possible area, however, is not so easy to discern from the graph. The two distributions that lead to the largest possible area are depicted on the rightmost graph of Figure 4. The distribution from within the prediction bounds is shown as a dashed line; the distribution from within the observation bounds is shown as a dotted line. The area between these two distributions is shaded in the middle graph of the figure. The non-intuitive shape of the shading gives a hint at the computational complexity of bounding the area metric. This scheme of bounding the area is not itself a mathematical metric. Firstly, it produces two numbers rather than a single scalar. Secondly, it does not satisfy the property of identity of indiscernibles. Even if the prediction uncertain number is identical to the data uncertain number, the upper bound will not be zero (unless there is no epistemic uncertainty).

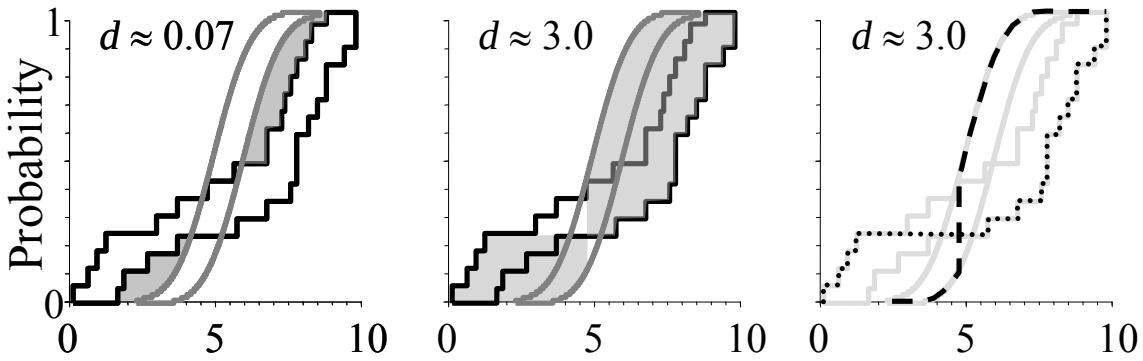

Figure 4. Smallest (left) and largest (middle and right) possible areas between a distribution inside the uncertain predictions (gray bounds) and a distribution inside the uncertain empirical observations (black step functions). The extremal distributions yielding the largest area are depicted in the right graph.

As yet another alternative, we could generalize our validation metric as the *two-dimensional* vector $Д(x,y) = (d(x_L, y_L), d(x_R, y_R))$ where the subscript L denotes the left side of a p-box and the subscript R denotes the right side, and d is our regular area metric for distributions. The left value of the pair reflects the difference between the left side of the prediction and the left side of the observations. Likewise, the right side of the distance pair reflects the difference between the right side of the prediction and the right side of the observations. This pair would constitute what we might call a double metric, $Д: \mathbf{B} \times \mathbf{B} \rightarrow \Re^+ \times \Re^+$, where $\mathbf{B}$ is the set of all p-boxes (which includes intervals, probability distributions and scalars as special cases), and $\Re^+$ is the set of all positive real numbers, satisfying the following generalizations of the four metric properties:

$Д(x, y) = (a, b)$ implies both $a \geq 0$ and $b \geq 0$ (non-negativity),

$Д(x, y) = Д(y, x)$ (symmetry),

$Д(x, y) = (0,0)$ if and only if $x = y$ (identity of indiscernibles), and

$$\left\{ \begin{array}{l} Д(x, y) = (a_1, b_1) \\ Д(y, z) = (a_2, b_2) \\ Д(x, z) = (a_3, b_3) \end{array} \right\} \text{ imply } a_1 + a_2 \geq a_3 \text{ and } b_1 + b_2 \geq b_3 \quad \text{(triangle inequality).}$$

Figure 5 shows three examples of this double metric. In the leftmost graph, a scalar prediction at $x = 7$, depicted as a gray spike, is compared to an interval observation $y = [14, 19]$ shown in black. The value 7 is compared against both sides of the interval to yield $Д(x, y) = (|14–7|, |19–7|) = (7, 12)$. In the middle graph, the comparison is between two intervals, and the two-dimensional difference is $Д([4, 9], [13, 18]) = (|13–4|, |18–9|) = (9, 9)$. In the rightmost graph of Figure 5 the black observation interval overlaps with the gray prediction interval. The double metric is $Д([3, 11], [8, 17]) = (|8–3|, |17–11|) = (5, 6)$. The value of the double metric would be (0,0) when the corresponding edges coincide exactly. Being double-zero would not mean that the uncertainty in

either the evidence or prediction has gone to zero, but only that they match in both location and imprecision.

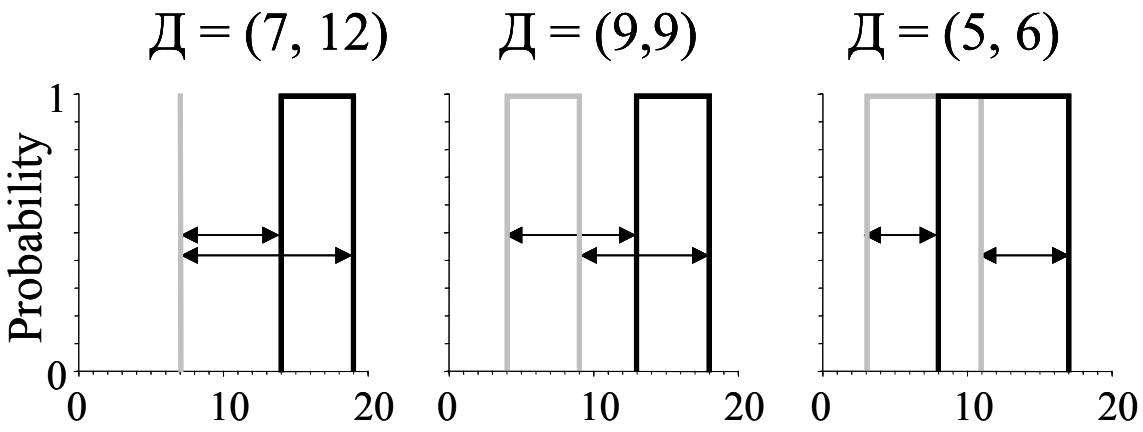

Figure 5. A generalized, two-dimensional metric between uncertain numbers (intervals).

Four possible generalizations of the area metric for epistemic uncertainty in predictions or observations have been discussed in this and the previous section. None has all the properties one might desire. Neither the shortest distance nor the range of possible areas is a true mathematical metric because they do not have the property of identity of indiscernibles. In the case of the shortest distance, the distance being zero does not guarantee that the representation of the prediction is identical to the representation of the observations. In the case of the range of possible areas, if the prediction and observation representations are identical, the value will not generally be zero. The double metric and Pompieu's max-sup-inf both have formal metric properties (or at least generalizations of them), but they seem to be overly strict in that predictions must match observations in their uncertainties even though there's no physical or engineering reason to demand this. The double metric is the easiest to compute, followed by the shortest distance. Pompieu's max-sup-inf and the range of possible areas are hardest to compute. The shortest distance measure and the double metric are both based on the comparing the shapes of the representations of the prediction and observations, whereas the other two measures are based on comparing individual elements (i.e., distribution functions consistent with those representations). The table below summarizes these observations.

Measure	*Scheme*	*Metric*	*Compute*	*Strictness*
Shortest distance	Shape	No	Medium	Reasonable
Pompieu's max-sup-inf	Element	Yes	Hard	Too strict
Range of possible areas	Element	No	Hard	Reasonable
Double metric	Shape	Yes	Easy	Too strict

We expect that the shortest distance will be most useful in many practical applications. In some situations, the range of possible areas will be most informative.

The comparison between random numbers characterized by probability distributions could be understood in terms of their difference as real numbers that are *realizations* from those distributions or in terms of the discrepancies between the *shapes* of those distributions. When there is only aleatory uncertainty associated with the prediction and observations, it seems reasonable to use the latter comparison based on distribution shapes for the purposes of validation. The analogous comparison between uncertain numbers, i.e., characterizations of numerical quantities that express both aleatory and epistemic uncertainty, can also be considered in these two senses. But comparing the shapes of distributions does not seem completely satisfactory when there is epistemic uncertainty present as well. There are several approaches possible for handling epistemic uncertainty based on the area metric. Two of these approaches seem most promising. The first is based on comparing shapes and considers the measure of the disagreement to the smallest possible value of the area metric that would be consistent with distributions from within the express uncertainty. The second approach, based on realizations, considers the range of possible values of the area metric consistent with distributions within the uncertainty.

6. Conclusions

The comparison between random numbers that are characterized by probability distributions can be understood in terms of their difference as real numbers that are realizations from those distributions, or in terms of the discrepancies between the shapes of their distributions. It seems reasonable to use the latter comparison based on distribution shapes for the purposes of validation for (precise) probabilistic models. The analogous comparison between uncertain numbers, i.e., characterizations of numerical quantities that simultaneously express both aleatory and epistemic uncertainty, can also be considered in these two senses. But, whereas we chose to compare the shapes of the probability distributions when the quantities had only aleatory uncertainty, this choice does not seem satisfactory when there is epistemic uncertainty present as well. In the case of comparing two simple intervals which contain only epistemic uncertainty, if the prediction interval overlaps with the datum interval, then the prediction is perfectly correct from the perspective of a validation assessment. The shapes of the two intervals could be quite different, and indeed, their overlap could be very small, yet the validation measure of their mismatch is zero if they overlap at all.

There are several ways to unify and extend these apparently disparate notions of validation for the case of general uncertain numbers that include both epistemic and aleatory uncertainty. Perhaps the most workable is the smallest area between the uncertain numbers. This is the smallest possible area between probability distributions contained in the respective uncertain numbers under any possible dependence. For many situations in which p-boxes are used to characterize the prediction and the data, the smallest area is easy to compute when the edges of

the p-boxes represent admissible distributions. In these cases, the smallest area is the mean of the distribution of differences of the extremal distributions computed under the assumption of perfect dependence.

Acknowledgements

We thank Marty Pilch, Kevin Dowding and Laura Swiler at Sandia National Laboratories, Bob Nau of Duke University and Wei Chen of Northwestern University. This paper is a product of work supported by the Sandia Epistemic Uncertainty Project under contract 19094 and supported by NASA Small Business Innovation Research grants NNL06AA40P and NNL07AA06C. All opinions are those of the authors and not necessarily of the supporting agencies.

References

AIAA. 1998. Guide for the Verification and Validation of Computational Flui Dynamics Simulations. AIAA G-077-1998. American Institute of Aeronautics and Astronautics, Reston, Virginia.

Angus, J.E. 1994. The probability integral transform and related results. *SIAM Review* 36(4): 652–654.

ASME. 2006. Guide for Verification and Validation in Computational Solid Mechanics. ASME V&V 10-2006. American Society of Mechanical Engineers. http://catalog.asme.org/Codes/PrintBook/VV_10_2006_Guide_Verification.cfm.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Brier, G.W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78: 1–3.

Chen, W., L. Baghdasaryan, T. Buranathiti and J. Cao. 2004. Model validation via uncertainty propagation. *AIAA Journal* 42: 1406-1415.

Chen, W., Y. Xiong, K.-L. Tsui, and S. Wang. 2006. Some metrics and a Bayesian procedure for validating predictive models in engineering design. *Proceedings of ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Philadelphia. American Society of Mechanical Engineers. http://ideal.mech.northwestern.edu/pdf/DAC_ validation06.pdf.

Chen, W., Y. Xiong, K.-L. Tsui, and S. Wang. 2007. A design-driven validation approach using Bayesian prediction models. *Journal of Mechanical Design* [in press].

Colyvan, M. 2004. The philosophical significance of Cox's theorem. *International Journal of Approximate Reasoning* 37(1): 71–85. http://homepage.mac.com/mcolyvan/papers/cox.pdf.

Cox, R.T. 1946. Probability, frequency and reasonable expectation. *American Journal of Physics* 14:1–13.

Devore, J.L. 2000. *Probability and Statistics for Engineers and Scientists*. Duxbury, Pacific Grove, California.

Dobrushin, R.L. 1970. Prescribing a system of random variables by conditional distributions. *Theory of Probability and its Applications* 15: 458–486.

Dowding, K.J., M. Pilch and R.G. Hills. 2008. Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering* [in press].

Dowding, K.J., R.G. Hills, I. Leslie, M. Pilch, B.M. Rutherford an M.L. Hobbs. 2004. *Case Study for Model Validation: Assessing a Model for Thermal Decomposition of Polyurethane Foam*. SAND2004-3632, Sandia National Laboratories, Albuquerque, NM.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B* 57: 45–97.

Feller, W. 1948. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics* 19: 177–189.

Ferson, S. 2002. *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Lewis Publishers, Boca Raton, Florida.

Ferson, S. and L.R. Ginzburg. 1996. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and Systems Safety* 54:133–144.

Ferson, S., V. Kreinovich, L. Ginzburg, D.S. Myers, and K. Sentz. 2003. *Constructing Probability Boxes and Dempster-Shafer Structures.* SAND2002-4015, Sandia National Laboratories, Albuquerque, New Mexico. http://www.ramas.com/unabridged.zip.

Ferson, S., C.A. Joslyn, J.C. Helton, W.L. Oberkampf and K. Sentz. 2004. Summary from the epistemic uncertainty workshop: consensus amid diversity. *Reliability Engineering and System Safety* 85: 355–370.

Ferson, S., V. Kreinovich, J. Hajagos, W.L. Oberkampf and L. Ginzburg 2007. *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. SAND2007-0939, Sandia National Laboratories, Albuquerque, NM. http://www.ramas.com/intstats.pdf.

Ferson, S., W.L. Oberkampf and L. Ginzburg. 2008. Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering* [in press]. Available at http://www.ramas.com/thermval.pdf.

de Finetti, B. 1962. Does it make sense to speak of "good probability appraisers"? Pages 357–363 in *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, I.J. Good (ed.). Wiley, New York.

Frank, M.J., R.B. Nelsen and B. Schweizer 1987. Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probability Theory and Related Fields* 74:199–211.

Fréchet, M. 1906. Sur quelques points du calcul fonctionnel (Thèse). *Rendiconti Circolo Mathematico di Palermo* 22:1–74.

Gioia, F., and C.N. Lauro. 2005. Basic statistical methods for interval data. *Statistica Applicata* [Italian Journal of Applied Statistics] 17(1): 75–104.

Hansen, K.M. 1999. A framework for assessing uncertainties in simulation predictions. *Physica D* 133: 179-188.

Hazelrigg, G. 2003. Thoughts on model validation for engineering design. *Proceedings of ASME 2003 Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Chicago.

Helton, J.C., and W.L. Oberkampf (eds.) 2004. *Reliability and Engineering System Safety* 85 (issues 1–3).

Hills, R.G., M. Pilch, K.J. Dowding, I. Babuska, and R. Tempone. 2008. Model validation challenge problems: tasking document. *Computer Methods in Applied Mechanics and Engineering* [in press].

Hills, R.G. 2006. Model validation: model parameter and measurement uncertainty. *Journal of Heat Transfer* 128: 339–351.

Hills, R.G., and T.G. Truncano. 2002. *Statistical Validation of Engineering and Scientific Models: A Maximum Likelihood Based Metric*. SAND2002-1783, Sandia National Laboratories, Albuquerque, NM.

Hills, R.G., and I. Leslie. 2003. *Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application*. SAND2003-0706, Sandia National Laboratories, Albuquerque, NM.

ISO [International Organization for Standardization]. 1993. *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization, Geneva, Switzerland.

JCGM [Joint Committee for Guides in Metrology]. 2006. Evaluation of measurement data—Supplement 1 to the "Guide to the expression of uncertainty in measurement"— Propagation of distributions using a Monte Carlo method. http://www.intermet.jp/JCGM/0610news/Supplement_to_GUM.pdf.

Kennedy, M. and A. O'Hagan. 2001. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B* 63: 425-464.

Klir, G., and M.J. Wierman. 1999. *Uncertainy-based Information: Elements of Generalized Information Theory*. Physica-Verlag, Heidelberg.

Kolmogorov [Kolmogoroff], A. 1941. Confidence limits for an unknown distribution function. *Annals of Mathematical Statistics* 12: 461–463.

Kullback, S. 1959. *Information Theory and Statistics*. Wiley, New York.

Kullback, S., and R.A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22: 79–86.

Levi, I. 1980. *The Enterprise of Knowledge.* MIT Press, Cambridge, Massachusetts.

Lindley, D.V., A. Tversky and R.V. Brown. 1979. On the reconciliation of probability assessments. *Journal of the Royal Statistical Society* A 142 (Part 2): 146–180.

Manski, C.F. 2003. *Partial Identification of Probability Distributions*, Springer Series in Statistics, Springer, New York.

Matheron, G. 1975. Random Sets and Integral Geometry. J.Wiley, New York

Mathiassen, J.R., A. Skavhaug and K. Bø. 2002. Texture similarity measure using Kullback-Leibler divergence between gamma distributions. Pages 19–49 in *Computer Vision - ECCV*

2002: 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28–31, 2002. Proceedings, Part III. Lecture Notes in Computer Science, volume 2352. Springer, Berlin.

Menger, K. 1942. Statistical metrics. *Proceedings of the National Academy of Science U.S.A.* 28: 535-537.

Molchanov, I. 2005. *Theory of Random Sets*. Springer, London.

Neapolitan, R.E. 1992. A survey of uncertain and approximate inference. Pages 55–82 in *Fuzzy Logic for the Management of Uncertainty*, L. Zadeh and J. Kacprzyk (eds.), John Wiley & Sons, New York.

Nikolaidis, E., and R. Haftka. 2001. Theories of uncertainty for risk assessment when data is scarce. *International Journal of Advanced Manufacturing Systems* 4(1): 49–56.

Oberkampf, W.L., and M.F. Barone. 2006. Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics* 217: 5–36.

Oberkampf, W.L., and J.C. Helton. 2005. Evidence theory for engineering applications. Pages 10-1–10-30 in *Engineering Design Reliability Handbook*, E. Nikolaidis, D.M. Ghiocel, and S. Singhal (eds.), CRC Press, Boca Raton, Florida.

Oberkampf, W.L., and T.G. Trucano. 2007. *Verification and Validation Benchmarks*. SAND2007-0853, Sandia National Laboratories, Albuquerque, NM. To appear in *Nuclear Engineering and Design*.

Oberkampf, W.L., J.C. Helton and K. Sentz. 2001. Mathematical representation of uncertainty. American Institute of Aeronautics and Astronautics Non-Deterministic Approaches Forum, Seattle, WA, Paper No. 2001-1645, April, 2001.

Oberkampf, W.L., T.G. Trucano and C. Hirsch. 2004. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews* 57(5): 345–384.

Pompeiu, D. 1905. *Sur la continuité des fonctions de variables complexes* (Thèse). Gauthier-Villars, Paris. *Ann. Fac. Sci. de Toulouse* 7:264-315.

Rabinovich, S. 1993. *Measurement Errors: Theory and Practice*. American Institute of Physics, New York.

Romero, V.J. 2007. Validated model? Not so fast. The need for model "conditioning" as an essential addendum to model validation. AIAA-2007-1953 in *Proceedings of the 2007 AIAA Non-Deterministic Approaches Conference*, Honolulu. American Institute of Aeronautics and Astronautics.

Rutherford, B.M., and K.J. Dowding. 2003. *An Approach to Model Validation and Model-based Prediction—Polyurethane Foam Case Study*. SAND2003-2336, Sandia National Laboratories, Albuquerque, NM.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.

Song, K.-S.. 2002. Goodness-of-fit tests based on Kullback-Leibler discrimination information. *IEEE Transactions on Information Theory* 48:1103–1117

Stephens, M.A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69: 730–737.

Smirnov [Smirnoff], N. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin de l'Université de Moscou, Série internationale (Mathématiques)* 2: (fasc. 2).

Taylor, B.N. and C.E. Kuyatt. 1994. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results.* NIST Technical Note 1297, National Institute of Standards and Technology, Washington, DC. http://physics.nist.gov/Pubs/guidelines/contents.html. See also web guidance at http://physics.nist.gov/cuu/Uncertainty/index.html.

Trucano, T.G., L.P. Swiler, T. Igusa, W.L. Oberkampf and M. Pilch. 2006. Calibration, validation and sensitivity analysis: what's what. *Reliability Engineering and System Safety* 91: 1331–1357.

Williamson, R.C. and T. Downs 1990. Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4:89–158.

Winkler, R.L. 1996. Scoring rules and the evaluation of probabilities. *Test* 5: 1–60.

Yates, F. 1934. Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* (Supplement) 1: 217–235.

Vallender, S.S. 1973. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability and its Applications* 18: 784–786.

Walley, P. 1991. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.

Zhang, R., and S. Mahadevan. 2003. Bayesian methodology for reliability model acceptance. *Reliability Engineering and System Safety* 80:95-103.