Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 197 (2008) 2408–2430
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Model validation and predictive capability for the thermal challenge problem

Scott Ferson [a,*], William L. Oberkampf [b], Lev Ginzburg [c]

[a] *Applied Biomathematics, 100 North Country Road, Setauket, NY 11733, USA*
[b] *Validation and Uncertainty Estimation Department, Mail Stop 0828, Department 1544, Post Office Box 5800, Sandia National Laboratories, Albuquerque, NM 87185-0828, USA*
[c] *Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA*

Received 8 June 2007; received in revised form 5 July 2007; accepted 5 July 2007
Available online 23 December 2007

Abstract

We address the thermal problem posed at the Sandia Validation Challenge Workshop. Unlike traditional approaches that confound calibration with validation and prediction, our approach strictly distinguishes these activities, and produces a quantitative measure of model-form uncertainty in the face of available data. We introduce a general validation metric that can be used to characterize the disagreement between the quantitative predictions from a model and relevant empirical data when either or both predictions and data are expressed as probability distributions. By considering entire distributions, this approach generalizes traditional approaches to validation that focus only on the mean behaviors of predictions and observations. The proposed metric has several desirable properties that should make it practically useful in engineering, including objectiveness and robustness, retaining the units of the data themselves, and generalizing the deterministic difference. The metric can be used to assess the overall performance of a model against all the experimental observations in the validation domain and it can be extrapolated to express predictive capability of the model under conditions for which direct experimental observations are not available. We apply the metric and the scheme for characterizing predictive capability to the thermal problem.

Keywords: Validation; Predictive capability; Thermal challenge problem; Area metric

1. Introduction

Although there is increasing consistency in the formal definition of the term 'validation' [1,2], there is still wide disagreement about the precise steps involved in the validation process. In this paper, we use the terminology and the following three steps identified by [3]:

(i) *Validation assessment*: assessment of model accuracy by comparison of predictions against experimental data,
(ii) *Model extrapolation*: extrapolation (or possibly interpolation) of the model to the intended use, and
(iii) *Adequacy decision*: determination of whether the model is adequate for the intended use.

Validation is often contrasted with calibration or updating, which is the fitting of a model to empirical observations to maximize the match between predictions and observations, usually by changing model parameters but sometimes by introducing or omitting model components. Because we want to distinguish assiduously between these two activities, we will assume that a model is fixed while a validation is undertaken. As soon as the model is changed in structure or parameter values in any way to account for data, the activity is no longer validation in our strict sense.

The problem of validation might initially seem to be a relatively straightforward one: all one needs to do is take the prediction the modeler gives us and compare it to the

* Corresponding author. Tel.: +1 631 751 4350; fax: +1 631 751 3435.
E-mail address: scott@ramas.com (S. Ferson).

0045-7825/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cma.2007.07.030

new observation the empiricist gives us and see whether they match. Clearly, the answer could be that they match perfectly, that the model is a little off, or that the model is way off. But this simplicity is immediately corrupted by the complexity of the real world. In fact, analysts must have strategies to address several questions:

1. What if the prediction is a probability distribution rather than a point value?
2. What if there is experiment-to-experiment variability in the measured data?
3. What if there are statistical trends in the experimental data?
4. What if there are multiple predictions about different outputs to be assessed?
5. What if the predictions from the model are extremely expensive to compute?
6. What if the data were collected under conditions other than the intended application?
7. What if there is non-negligible experimental measurement uncertainty in the data?

This paper addresses the first six of these issues through the example of the Thermal Challenge Problem [4,5], which is the first of three problems considered at the Sandia Validation Challenge Workshop [6] in Albuquerque, New Mexico, over 21–23 May 2006. We interpreted the challenge problem to involve two goals. The first goal is to measure in some objective way the conformance of predictions from the model with the experimental measurements. The second goal is to use this measure to characterize the reliability, i.e., estimate the uncertainty, of other predictions made by the model.

Validation is not about checking whether a model is right or wrong per se. Indeed, we believe, as George Box famously asserted [7], that all models of any physical reality are wrong, at least in the narrow sense that they can never be perfect. Validation is about assessing the accuracy of the model and assessing whether a model is good enough for some intended purpose. For a deterministic model, validation can be a fairly straightforward affair. The model makes a point estimate for its prediction about some quantity. This prediction would be compared against one or more measurements about that quantity and the difference(s) would be understood as a measure of how accurate the model was. A model could be consistently inaccurate and yet close enough for its purpose. Likewise, even a highly accurate model might not be good enough if it is needed for some delicate, high-consequence decisions.

Two pervasive issues complicate these comparisons in general validation problems. The first is that, today, most serious simulation models generate entire *distributions* rather than point estimates as their predictions of system response quantities. These distributions often characterize the stochastic variability of these quantities and perhaps the epistemic uncertainty about these quantities. In many cases, experimental observations are small collections of numbers, although they can be abundant enough to be conveniently characterized as distributions.

The second issue is the data available for validation may not be directly relevant to the prediction of interest. In particular, we might be able to collect data under conditions that are similar, but not identical, to those for which a prediction is desired. This necessitates some sort of extrapolation that will allow us to characterize the validity of a model for prediction even when no immediately relevant data are available to compare against this prediction. In a strong sense, any forecast about future or general predictive capability is necessarily an extrapolation of some kind, even if directly relevant data are available, because such a forecast is always about hypothetical data that might be observed, rather than merely a summary of the consistency with past observations.

Section 2 gives a synopsis of the thermal challenge problem. Section 3 considers a risk-analytic approach to the problem that translates the variability observed in the material characterization data into exceedance probabilities of the system response quantity temperature. The analysis is refined and extended in Section 4 to account for a subtle trend in the material characterization data that causes values of one of the material properties to depend partially on the material's temperature. Section 5 suggests a general validation metric and procedures for its use that can be applied when a model's predictions take the form of probability distributions. Section 6 applies this metric to the challenge problem to assess the overall performances of the models developed in Sections 3 and 4 relative to all of the experimental observations over the validation domain. Section 6 also considers the extrapolation from the measured performances of the model in the face of the individual data points to express the predictive capability of the model under conditions of regulatory interest. Section 7 answers several specific questions posed as part of the validation challenge and Section 8 offers conclusions and outlines further research needs.

2. Sandia validation challenge problem

The formulation and numerical details of the thermal challenge problem are given in [4] and will not be reiterated here except in briefest outline. The problem consists of a mathematical model, three sets of experimental data which differ in size ('low', 'medium' and 'high'), and a regulatory requirement. The mathematical model is of the temperature under heating of a device constructed of some material and has the form

$$T(x,t) = \begin{cases} T_i + \frac{qL}{k}\left[\frac{(k/\rho C_p)t}{L^2} + \frac{1}{3} - \frac{x}{L} + \frac{1}{2}\left(\frac{x}{L}\right)^2 \right. \\ \quad \left. -\frac{2}{\pi^2}\sum_{n=1}^{6}\frac{1}{n^2}\exp\left(-n^2\pi^2\frac{(k/\rho C_p)t}{L^2}\right)\cos\left(n\pi\frac{x}{L}\right)\right], \quad t>0, \\ T_i, \quad t=0, \end{cases} \tag{1}$$

where T is temperature, x is location within the material, t is time since the onset of heating, T_i is the initial ambient temperature, q is the heat flux, L is the thickness of the material, and k and ρC_p are properties of the material. The regulatory requirement is

$$\text{Prob}(900\ ^\circ\text{C} < T_{x=0\ \text{cm},t=1000\ s,T_i=25\ ^\circ\text{C},q=3500\ \text{W/m}^2,L=1.90\ \text{cm}}) < 0.01, \tag{2}$$

given material properties of the device associated with a particular manufacturing process. The challenge is to use available empirical data to estimate whether the regulatory requirement in Expression (2) is satisfied. The empirical data includes material characterization data consisting of several measurements of the material properties k and ρC_p, and "ensemble" and "accreditation" data consisting of experimental observations of temperature for various values of x, t, q and L defining a validation domain, none of which were collected at the conditions of regulatory interest.

In this paper we use probability distributions to characterize both the observed variability in data and the forecasted variability of predictions. The predicted temperature would be a distribution because, although the values of x, t, T_i, q, and L are prescribed for us in the statement of the problem, the values of the material properties k and ρC_p are only known by sample data, which ought to be characterized by probability distributions. The regulatory requirement prescribes a critical temperature of 900 °C and a critical probability of 1%. If the predicted distribution for temperature ventures anywhere into the region of the temperature-probability plane where values larger than 900 °C are more probable than 0.01, then we know that we are out of compliance with the regulatory condition.

In principle, it would have been possible for those who designed the challenge problem to simply present us with the predictions from the model (expressed as distributions) and the observed data (expressed as collections of numbers), rather than giving us the model and asking us to generate the predictions. In a sense, the activity of producing predictions is not part of validation per se, but rather part of *modeling*. However, meeting the challenge in this paper requires delving into the modeling process to some extent. We recognize, moreover, that modeling and validation are usually intricately intertwined in practice, just as validation and calibration are often intertwined. Although we intend to distinguish between the generation of predictions and the validation of those predictions against data—and the focus of this paper is on the latter—for the purposes of the challenge, we don the modeler's hat where necessary and implement (and even slightly modify) the model to create predictions from it to be compared to observations.

The validation challenge problem was purposefully designed with some model weaknesses and inconsistencies in order that it realistically reflect the common situations analysts encounter. Thus, there are assumptions that seem arguable or incorrect, including assertions that certain parameters are constants, that interacting variables are mutually independent, and that there is no uncertainty in measured data. We believe that a comprehensive validation study demands that such assertions be critically examined, dealt with in a realistic manner, and perhaps rejected in favor of more realistic assumptions.

3. Risk-analytic approach to stochastic variation

The materials characterization data described in the challenge problem suggest there is variability in both the thermal conductivity k and heat capacity ρC_p. The step functions in Fig. 1 show the cumulative empirical distribution functions for the values for ρC_p and k from the 'medium' materials characterization data. These observed patterns likely understate the true variabilities in these parameters because they represent only 20 observations for each of them. If we had observed different sets of values, it is likely we would have seen slightly different patterns, and we may well have seen some values above or below the observed ranges from the 20 values. To model this possibility of more extreme values than were seen among the limited samples, risk analysts commonly fit a distribution to data to model the variability of the underlying population. We used normal distributions for this purpose, configured so that they had the same mean and

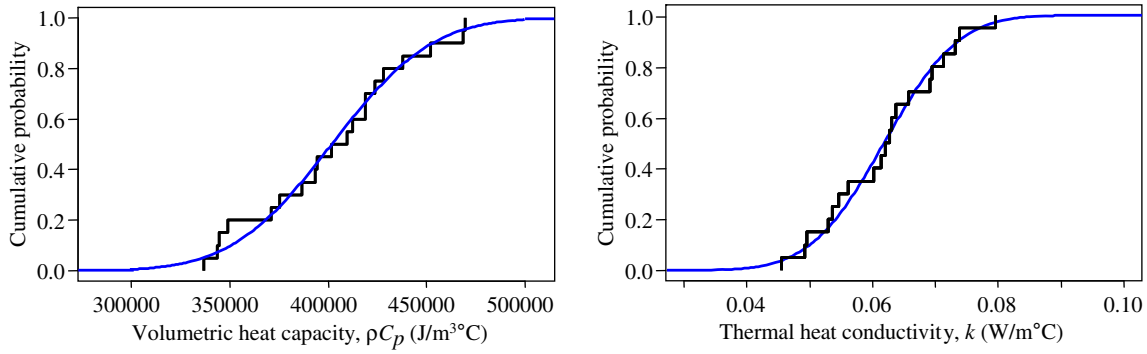

Fig. 1. The empirical distributions (step functions) for ρC_p and k seen during the 'medium' materials characterization and the normal distributions (smooth curves) fitted by the method of matching moments.

standard deviation as the data themselves, according to the method of matching moments [8]. The fitted normal distributions are shown in Fig. 1 as the smooth cumulative distributions. We do not consider this fitting of distributions to be model calibration because the distributions are not selected with reference to the system response quantity of temperature. Instead, the distributions merely summarize the material characterization data which are not otherwise used in the validation process.

We also fitted normal distributions to the 'low' and 'high' data sets as well. The computed moments for the three data sets for each parameter are given in the following table:

	Low ($n = 6$)	Medium ($n = 20$)	High ($n = 30$)
Thermal conductivity k, W/m °C			
Arithmetic mean	0.06002	0.06187	0.06284
Standard deviation	0.01077	0.00923	0.00991
Volumetric heat capacity ρC_p, J/m^3 °C			
Arithmetic mean	405,500	402,250	393,900
Standard deviation	42,065	39,511	36,251

To express the uncertainty about the temperature prediction that arises from the stochastic variability of k and ρC_p observed in the characterization of the material properties, we need to project these normal distributions through the temperature response model in Expression (1). The projection can be effected with a straightforward Monte Carlo simulation [8,9]. Thermal conductivity and volumetric heat capacity are the only distributional variables in the simulation; the rest are constants:

Variable	Symbol	Value (s)	Units
Thermal conductivity	k	Normal (0.06187, 0.00923)	W/m °C
Volumetric heat capacity	ρC_p	Normal (402250, 39511)	J/m^3 °C
Heat flux	q	3500	W/m^2
Thickness	L	0.019	m
Initial temperature	T_i	25	°C
Location	x	0	m
Time	t	1000	s

There could also be variability in some of the other inputs too, especially in variables such as thickness L and heat flux q. We neglect these variabilities here only because the challenge problem instructs us to.

Whenever there is stochasticity in more than a single variable, there is a possibility that correlation or dependence between the variables may influence any arithmetic functions of those variables [10,11]. We looked for evidence of such dependence in the paired thermal conductivity and heat capacity data collected during materials characterization. Fig. 2 shows the scattergram of these two variables for the 'medium' data set, which reveals no apparent trends or evidence of statistical dependence. The Pearson correlation coefficient between these twenty points is 0.0595, which is not remotely statistically significant ($p \gg 0.5, \mathrm{df} = 18$). Because there are no physical reasons to expect correlations or other dependencies between these variable, at least over the variability ranges considered here, it would seem reasonable to assume that these quantities are statistically independent of one another. Plotting and correlation analysis for the 'high' and 'low' data sets gave qualitatively similar results.

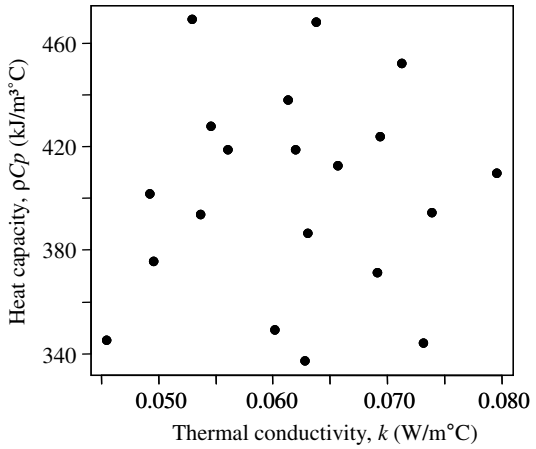

Fig. 2. Scattergram of ρC_p and k seen during the 'medium' materials characterization.

We used 10,000 replications in the Monte Carlo simulation, although many fewer replications could have sufficed if the model had been computationally intensive. We implemented the simulation in the R programming language [12]. When the stochasticity of k and ρC_p is projected through the heating model, it produces a distribution for the surface temperature after 1000 s. The output distribution produced is displayed as the complementary cumulative distribution in Fig. 3 so that the ordinate gives the

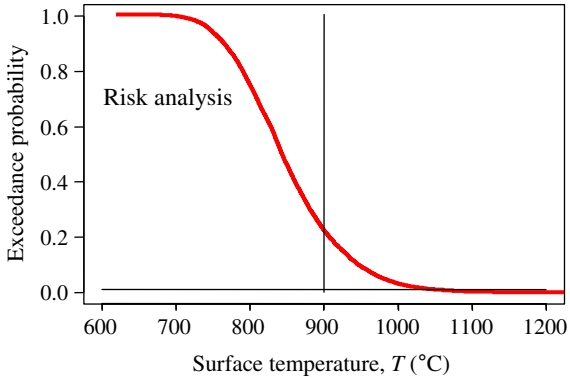

Fig. 3. Distribution (curve) resulting from risk analysis of stochasticity in material properties against the regulatory requirement (straight lines) for the exceedance probability when $x = 0$ cm, $t = 1000$ s, $T_i = 25$ °C, $q = 3500$ W/m^2, and $L = 1.90$ cm.

probability that the random variable temperature exceeds the value given on the abscissa. We use the complementary display for temperature because it makes it easier to visualize the probabilities of large values, which are the focus of concern in the regulatory statement. This result suggests that the probability of the temperature being larger than 900 °C is 0.22, much larger than the target of 0.01. This result argues strongly that the system is out of compliance with the regulatory requirement in Expression (2). This is obvious in the depiction in Fig. 3 because the distribution makes a deep incursion into the upper, right quadrant of the temperature-probability plane. The mean of the predicted temperature distribution is about 850 °C, which is similar to the value 838 °C from a deterministic calculation based on mean values for k and ρC_p. Although the mean is well below the temperature threshold of 900 °C, the stochasticity observed in the material properties implies that the regulatory threshold is regularly exceeded much more often than 1% of the time.

3.1. Robustness of the prediction

The robustness of this distributional result can be estimated by exploring its sensitivity to the choice of the distributional parameters for the k and ρC_p inputs, the normal shapes used to model their variation, and the assumptions about intervariable dependence. (An analyst might also want to explore the effects of variability or measurement error in the other inputs of the heating model, and the effect of uncertainty about the structure of the heating equation itself, but the challenge problem instructs us to ignore these uncertainties.) Because there were few observations collected in the material characterization study (6, 20 and 30 per variable in the 'low', 'medium' and 'high' data sets respectively), the estimates of the means and variances used to parameterize the normal distributions are associated with an appreciable degree of sampling uncertainty. Of course the effect of such sampling uncertainty on the estimate of the final temperature distribution is for the most part symmetric. That is, it creates bands of uncertainty on either side of the central distributional estimate displayed in Fig. 3. Therefore, if we were to enlarge our assessment by accounting for the sampling uncertainties about the k and ρC_p inputs, the result would be that the exceedance probability estimate of 0.22 would expand to an interval around that value. From a decision maker's point of view, this could only make the outcome seem *worse* for the hypothesis that the system is in compliance with the regulatory requirement in Expression (2), which specifies the probability not be larger than 0.01, because the uncertainty assessment reveals it might be even larger than 0.22.

An assessment accounting for uncertainty about the input distribution shapes has similar import. The model for k and ρC_p that would produce the smallest dispersion in the final temperature distribution, while still being consistent with the observed variability for k and ρC_p, would use the empirical distribution functions for these two inputs rather than parametric distributions such as normals. The empirical distribution functions simply summarize the observed data actually seen in the material characterization data. They are nonparametric estimates of the distributions because they do not require the analyst to select any parameters to specify the distribution. The model based on these distributions would be at least arguably reasonable, although it is likely to understate the chances of extreme values of the inputs. The result of the simulation based on this resampling strategy is very similar to the result shown in Fig. 3; in fact the final temperature distributions are largely indistinguishable from one another given the line thickness used in the display. The probability of exceeding 900 °C *increases* slightly to about 0.25, but the distribution tails contract so that, for instance, 1050 °C is the largest possible temperature (corresponding to the smallest observed values for k and ρC_p).

The only way that we might be in compliance with the regulatory requirement is if our risk analysis has overestimated the variation in the resultant temperature distribution. One assumption possibly worth reconsideration is whether k and ρC_p really have stationary distributions that are independent of temperature changes in the material. If there is some dependence of these material properties on temperature, it might be the case that our model predictions are in error. We consider this possibility in the next section.

4. Temperature dependence

In the description of the mathematical model for heat conduction in Expression (1), the volumetric heat capacity ρC_p and thermal conductivity k are assumed to be independent of temperature T. It is reasonable to ask whether this assumption is tenable given the available materials characterization data. Fig. 4 is the scattergram for the 'medium' data set for heat capacity as a function of temperature. Linear and quadratic regression analysis reveal no statistically significant trend among these points. The pictures are qualitatively the same for the 'low' and 'high' data sets in that no trend or other stochastic dependence is evident. Thus, the experimental data for heat capacity support the assumption in the mathematical model.

The materials characterization data for thermal conductivity, on the other hand, seem to be strongly related to temperature. Fig. 5 shows the scattergram of thermal conductivity as a function of temperature for the 'medium' data together with a regression line fitted by the least squares criterion. This regression is statistically significant ($p < 0.001, \mathrm{df} = 18$). The resulting model for this trend and residual scatter is

$$k \sim \alpha + \beta T + \mathrm{normal}(0, \sigma)$$
$$= 0.0505 + 2.25 \times 10^{-5} T + \mathrm{normal}(0, 0.0047),$$

where $\alpha = 0.0505$ and $\beta = 2.25 \times 10^{-5}$ are the fitted regression coefficients for the intercept and slope, the normal

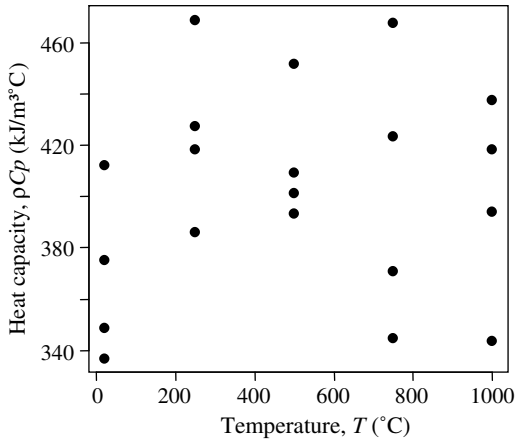

Fig. 4. Scattergram of heat capacities associated with different temperatures in the 'medium' materials characterization data suggesting these variables are independent.

function denotes a normal distribution with the given mean and standard deviation, and $\sigma = 0.0047$ is the residual standard error from the regression analysis. This σ is the standard deviation of the Gaussian distributions that, under the linear regression model, represent the vertical scatter of k at a given value of the temperature variable. There is no evidence that this trend is other than linear; quadratic regression does not provide a significant improvement in the regression fit. The visual impression that the data might be heteroscedastic—specifically that the variance among conductivities at the highest temperature is larger than for other temperatures—was not statistically significant in a post hoc test. The homoscedasticity and the strong linear trend of thermal conductivity on temperature are also evident in the 'low' and 'high' data sets, although the numerical details are of course slightly different.

This statistical dependence has implications for the analysis of the heating model. As given in the challenge problem, the mathematical model counterfactually assumes independence of thermal conductivity and temperature. If

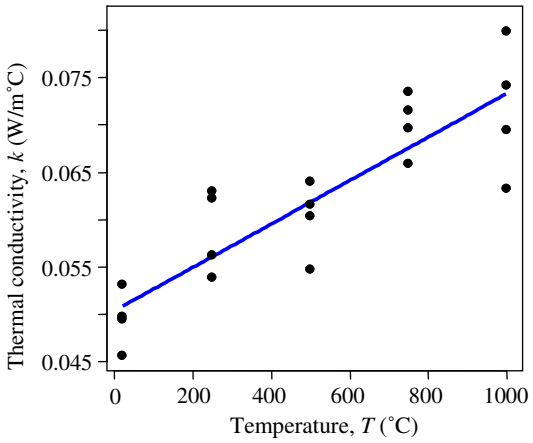

Fig. 5. Strong dependence in the scattergram of thermal conductivities and temperatures, together with the linear regression line, in the 'medium' materials characterization data.

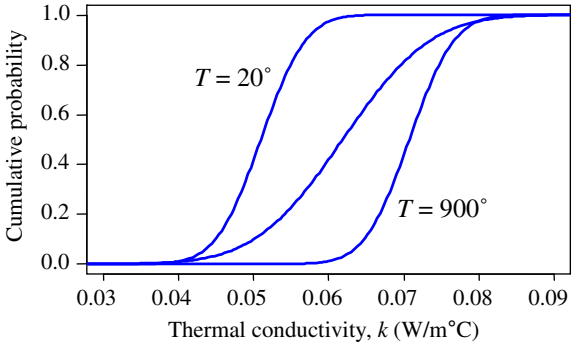

Fig. 6. Distributions of thermal conductivity, conditional on temperature at two values (outer curves), and unconditional (inner curve).

we could account for the observed dependency of k on T, we might be able to reduce the overall uncertainty in the model's predictions. The middle cumulative distribution function in Fig. 6 is the normal distribution fitted to the thermal conductivities k in the materials characterization data. (It is the same distribution previously shown in the right graph of Fig. 1.) We might be able to make the risk analysis described in the previous section more sophisticated by replacing this broad distribution with a *family* of much tighter normal distributions representing the variability of k conditional on temperature. Each of these tighter distributions is simply $k \sim 0.0505 + 2.25 \times 10^{-5} T +$ normal$(0, 0.0047)$ for a given scalar value of T. The breadth of each such distribution of k is the standard deviation of the residual term $\sigma = 0.0047$. Two distributions from this family are shown in the graph. When the temperature is 20 °C, the distribution of k's has a mean of 0.05 watts per meter degree. When the temperature is 900, the distribution has a mean of 0.07. For intermediate temperatures, the distribution of k has an intermediate central tendency, but always the same dispersion. Thus there is an entire continuous family of parallel normal distributions defined by the regression analysis of thermal conductivity on temperature. As the surface is heated, the distribution of thermal conductivities shifts higher. This means that, given a temperature of the material, the stochastic variability in thermal conductivity is constrained to a tighter distribution than would be suggested by the shallow middle distribution which ignores the temperature dependence. The conditioning of thermal conductivity on temperature reduces the variability compared to the traditional risk analysis of the previous section, although there remains an appreciable amount of variability in k.

We can combine this regression model relating k and T with the challenge problem's heating function Expression (1), albeit in an ad hoc fashion because one of the assumptions underlying Expression (1) is that k is independent of T. This combination creates a system of two equations that can be solved iteratively. In this iterative approach, we start from the (unconditional) distribution of k observed in the materials characterization data, and compute from it the resulting distribution of T (just as we did in Section 3).

We then project this distribution of T through the regression function to compute another distribution of k. That is, we compute the new distribution $k \sim 0.0505 + 2.25 \times 10^{-5}T + \text{normal}(0, 0.0047)$, where T is the just computed distribution of temperatures and the normal function generates normally distributed random deviates centered at zero with standard deviation 0.0047 (which are independent of the temperatures). The resulting distribution of k's conditional on temperature is then used to reseed the process, which is repeated until the distribution of T converges. We found that only two or three iterations were sufficient for convergence.

When we undertake this iterative solution, we are clearly trespassing into the domain of the physics modeler. We said we did not want to do this because it confounds the activities of validation with modeling, but the challenge seems designed to invite us to do it, so we are taking the bait because the model seems clearly deficient as originally stated. We offer the result, not as our belief that it is the best approach, but merely as an alternative model which we will subject to a validation process. Note that we are certainly not asserting, nor do we necessarily believe, that this regression model is the best or even an appropriate way to account for the dependence of k and T. Accounting for the dependence is delicate; one might prefer to send the issue back to the modeler who could devise a new model with a solution to an altered differential equation. A physics modeler who knows about their interactions should really be making pronouncements about such things. We are just exploring this as an exercise to see whether it can reduce the variability of the resulting surface temperatures and improve the fit to data (which we will consider in Section 6).

We revisited the Monte Carlo simulation described in the previous section with the ad hoc model for the dependence of k and T. The resulting predicted distribution of surface temperatures after 1000 s is shown as the solid distribution in Fig. 7. This distribution of temperatures has a smaller range and reduced variance compared to that seen in the traditional risk analysis of the previous section, but it is still in violation of the regulatory requirement in Expression (2) because it also encroaches into the danger zone of the upper, right quadrant. However, it does so much less than the result from the risk analysis conducted in Section 3 with the unmodified model. The estimated probability that temperature exceeds the critical value of 900 °C is about 0.05, only five times larger than that specified in the regulatory requirement.

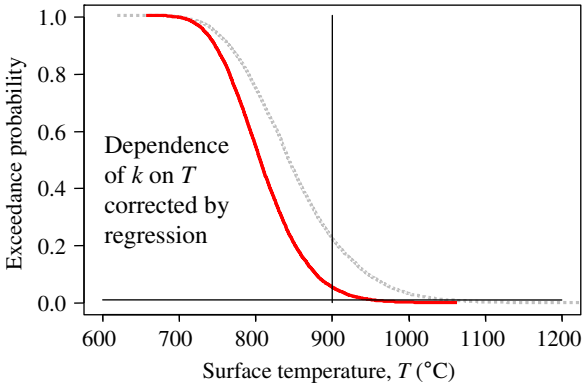

Fig. 7. Result of risk analysis with correction for dependence between k and T using regression (solid line), together with previous result assuming k and T are independent (dotted).

5. Validation assessment

The topic of model validation has received considerable attention recently [13–16]. Three main approaches are commonly used for validation in engineering settings, including hypothesis testing [17–21,cf. 10], Bayesian methods [22–28,cf. 29,cf. 30], and mean-based comparisons [31,32,3]. All three approaches have drawbacks. For instance, the purpose of hypothesis testing is to identify statements for which there is compelling evidence of truth. This is a rather different goal than that of validation, which is focused on assessing the quantitative accuracy of a model. The Bayesian approach to validation is primarily interested in evaluating the probability (i.e., the belief) that the model is correct. Yet, to our minds, this is not the proper focus of validation. We are not concerned about anyone's belief that the model is right; we are interested in *objectively measuring the conformance of predictions against data* that have not previously been used to develop or calibrate the model. The main limitation of approaches based on comparing means or other summary statistics is that it considers only the central tendencies or other specific behaviors of data and predictions and not their entire distributions. When predictions are distributions, they can contain a considerable amount of detail and it is not always easy to know what is important, nor to be sure that a comparison of means will be sufficiently informative for a particular application.

In this section, we introduce the notion of comparing probabilistic quantities, describe the desirable properties that a validation metric for making such comparisons should have, and suggest a particular one that has these properties. Section 6 applies this metric to the thermal challenge problem.

5.1. Comparing values that vary randomly

There are a variety of standard ways to compare random variables in probability theory (http://www.wikipedia.org/Random_variables#Equivalence_of_random_variables). If random numbers X and Y always have the same value, the random variables are said to be "equal", or sometimes "surely equal". A much weaker notion of equality is often useful. If we can only say that the expectation (i.e., the average) of the absolute values of the differences between X and Y is zero, the random variables are said

to be "equal in mean". If they are not quite equal in mean, we can measure their discrepancy in this sense by the mean metric

$$d_E(X, Y) = E(|X - Y|) \neq |E(X) - E(Y)|,$$

where E denotes the expectation operator. (Note that this difference is not the same as the absolute value of the difference between the means.) The idea can be generalized to higher-order moments, and equality in a higher-order moment implies equality in all lower-order moments.

The notion of equality for randomly varying quantities can be relaxed further still by comparing only the *shapes* of the probability distributions of the random variables. Random variables whose distributions are identical are said to be "equal in distribution". This is often denoted as $X \sim Y$, or sometimes by $X =^d Y$. This is really a rather loose kind of equality, because it does not require the individual values of X and Y to ever be equal, or even to ever be close. For instance, suppose X is normally distributed with mean zero and unit variance. Then X and $Y = -X$ are obviously equal in distribution, but are about as far from equality as can be imagined. Nevertheless, equality in distribution is an important concept because a distribution often represents all that is known about the values of a random variable. 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. For instance, a very common such measure is the maximal vertical difference between the two cumulative distribution functions

$$d_S(X, Y) = \sup_z |\Pr(X \leqslant z) - \Pr(Y \leqslant z)|,$$

which is the Smirnov distance used, for example, in defining the Kolmogorov–Smirnov statistical test for comparing distributions [33–35].

One of the properties of the Smirnov distance is that it is symmetric, which is to say that $d_S(X, Y)$ always equals $d_S(Y, X)$. The symmetry might be considered unnecessary or even counterintuitive as a feature for validation. We do not view predictions and observations as exchangeable with each other; it matters which is which. Suppose, for instance, that we inadvertently switched the theoretical distribution with the data distribution. One might expect to obtain a different result from having made such a mistake, but the Smirnov distance does not change whether the data and prediction are exchanged or not.

The Kullback–Leibler divergence [36,37] is another very widely used measure of the discrepancy between distributions that is not symmetric. It is defined, in its discrete formulation, between a probability mass function p for X and a probability mass function q for Y as

$$\sum_z p(z) \log_2 \frac{p(z)}{q(z)},$$

where z takes on all values in the common range of X and Y. The p distribution summarizes the observations and the q distribution represents the theoretical prediction. The continuous formulation is similar except that the summation is replaced by an integration. The word 'divergence' may be misleading because this quantity has nothing to do with notions of divergence familiar from calculus as the inner product of partial derivatives or the flux per unit volume. Instead, the term is used here in its other meaning as deviation from a standard. The Kullback–Leibler divergence is commonly used in information theory where it is interpreted as the expected extra message length per datum that must be transmitted to identify a particular value of z drawn from among all those in the common range of X and Y that would be needed to communicate a value when an encoding is optimal for a given incorrect distribution q, rather than using an encoding based on the true distribution p. It is also well known in physics as the relative entropy between p and q, i.e., the entropy of the distribution p with respect to the distribution q.

As we have mentioned, there are, in fact, many other measures that could be used to compare data and prediction distributions. But, given the broad acceptance and ubiquity of the Smirnov and Kullback–Leibler measures in probability and physics, it is perhaps necessary to explain why we simply do not use one of these as our validation metric. The next section considers the features of such a metric that would make it most useful for the application to the challenge problem and other similar problems in validation. In the following section we suggest still another measure to assess the discrepancy between the prediction and observations that we think is best suited for validation of probabilistic models in engineering.

5.2. Desirable properties of a validation metric

A validation metric is a formal measure of the mismatch between predictions and data that have not previously been used to develop the model. A low value of the metric means there is a good match, and a higher value means that prediction and data disagree more. We are interested in validation metrics that can be applied when predictions are probabilistic and are still reasonably intuitive to engineers and project managers. There are many desirable properties of a validation metric that would be useful in assessing the accuracy of models used in engineering simulations. Perhaps naturally, the first such property is that the validation metric be an *objective* measure of the distance, in some sense, between prediction and new data. Objectiveness means that, given a collection of observations and predictions, a validation metric will produce the same assessment no matter what analyst conducts it. This is a basic tenet of scientific and engineering practice, that the conclusion be reproducible and that it not depend on the attitudes or predilections of the analysts. If some inescapable subjectivity must enter the assessment, it would be good to keep this intrusion as small and as limited as possible so as to emphasize the objectiveness of the method and minimize

the elements subject to dispute. Objectiveness helps to minimize chance that the distortions from particular agendas enter into the analysis.

It is also important that the validation metric in some reasonable way *generalizes deterministic comparisons* between scalar values that have no uncertainty. That is, if the prediction and the corresponding measurement are both point values, then the natural metric is simply their difference. We would like the metric for the case when there is uncertainty to generalize this idea, and to reduce back to the simple distance when the prediction and observations are very tight distributions. There are, of course, many ways to make a generalization of this difference, but maintaining the backward compatibility, as it were, with the intuitive methods of validation for point values seems to us to be essential.

The validation metric should *reflect differences in the full distribution* of the predictions and the data. This is to say that a distributional comparison should be sensitive not just to differences in the means, or to differences in the means and variances, but to differences in the entire statistical distributions. Having said this, we probably would not like the result to be totally swamped by peculiarities in the extreme tails of the distributions, especially if those tails represent highly unlikely behaviors or very rare occurrences. Some degree of robustness that makes the validation measure *not too sensitive to long tails* will tend to make it more practical in real-world applications of validation.

We argue that the validation metric should express its result in *physical units* rather than some esoteric statistical units. For instance, if the prediction and the measurement are in degrees Celsius, then the metric that measures their discrepancy should also be a value in degrees Celsius rather than some anonymous normalized scale that has little intuition for engineers and project managers. The desirability of this feature is related to the aforementioned need that it generalize deterministic comparisons.

For closely related reasons, the measure should be *unbounded* in the sense that, if the prediction is completely off the mark of the measurement, the metric characterizing this discrepancy should be able to grow to be an arbitrarily large value. The alternative is a bounded of scaled measure which produces some upper value once the prediction and the data become sufficiently dissimilar but then can no longer distinguish even larger dissimilarities.

Finally, the measure used in validation should be also be mathematically well behaved and well understood. It would probably be useful if the measure were a *true metric* in the mathematical sense, or a similar function, which has the essential features of a true distance function. By definition, a mathematical metric d has four properties [38]:

Non-negativity	$d(x,y) \geqslant 0$,
Symmetry	$d(x,y) = d(y,x)$,
Triangle inequality	$d(x,y) + d(y,z) \geqslant d(x,z)$, and
Identity of indiscernibles,	$d(x,y) = 0$ if and only if $x = y$.

If not a complete metric, the measure would surely have most of these properties. We have already mentioned the possibility that the measure not be symmetric. A non-symmetric measure that satisfies the other metric properties is called a 'quasimetric'. A non-negative, symmetric measure that has the identity of indiscernibles property is called a 'semimetric'. A non-negative, symmetric measure that satisfies the triangle inequality is called a 'pseudometric'.

5.3. Proposed metric for validation of probabilistic predictions

This section defines a metric between a probabilistic prediction and a set of one or more empirical observations. Any probabilistic prediction of the form we are considering in this paper can always be characterized as a cumulative distribution function $F(x)$ or, equivalently, as a complementary cumulative distribution function $1 - F(x)$. In this notation, x denotes whatever variable the prediction is about. The prediction is specified by the model, and we presume that it has been given in this format. The observation(s), on the other hand, are usually provided as a collection of point values in a data set. The distribution function for a data set, which is sometimes called its empirical distribution function or EDF, summarizes the data set as a function suitable for graphical depiction. It is a function from the x-axis to a probability scale on the interval $[0,1]$. It is constructed as a non-decreasing step function with a constant vertical step size of $1/n$, where n is the sample size of the data set. The locations of the steps correspond to the values of the data points. Such a distribution for data x_i, $i = 1,\ldots,n$, is

$$S_n(x) = \frac{\sum_{i=1}^{n} I(x_i, x)}{n},$$

where

$$I(x_i, x) = \begin{cases} 1 & x_i \leqslant x, \\ 0 & x_i > x, \end{cases}$$

so that $S_n(x)$ is simply the fraction of data values in the data set that are at or below each magnitude x. A distribution preserves the statistical information in the data set about its central tendency or location, its dispersion or scatter, and, in fact, all other statistical features of the distribution. The only information in the original data set that is not in the distribution is the order in which the values were originally given, which is meaningless whenever the data were sampled at random. When the data set consists of a single value, then the S_n function is a simple spike at the location along the x-axis given by that value, that is, it is zero for all x less than that value and one for all x larger than that value. For graphical clarity, however, it becomes convenient not to depict these flat portions at zero and one when the functions are plotted.

We propose to use the *area* between the prediction distribution F and the data distribution S_n as the measure of the mismatch between them. Mathematically, the area

between the curves is the integral of the absolute value of the difference between the functions

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

It is clear geometrically that this quantity is also equal to the average horizontal difference between the two functions, $\int |F^{-1}(p) - S_n(p)^{-1}| \mathrm{d}p$, but this is not the same as the average of the absolute differences between *random values* from these two distributions. (Such an average would not be zero if the distributions were coincident.) This area is thus a function of the shapes of the distributions, but is not readily interpretable as a function of the underlying random variables. The area measures the disagreement or 'badness of the fit' between theory and empirical evidence. It is a metric so long as the integral exists. The rest of this section explains the features of this metric and justifies its usefulness as a measure for validation assessment.

Fig. 8 illustrates this area measure of mismatch for two data sets against a prediction distribution of temperature. The prediction distribution, shown as the smooth gray curve, is the same in both graphs. This prediction distribution might be obtained by solving the mathematical model analytically or perhaps by propagating a large number of replicate samples through it in a Monte Carlo simulation. Superimposed on these graphs are distribution functions S_n for two hypothetical data sets. On the left graph, the data set consists of the single value 252 °C, and on the right, the data set consists of the values {226, 238, 244, 261}. The areas measuring the mismatches between the prediction and the two data sets are shaded. In the left graph, the area consists of a region to the left of the datum at 252, and a region to right of it. In the right graph, there are four shaded regions composing the total area between the prediction distribution and the data distribution.

Fig. 9 shows how the area metric differs from a validation measure based on merely matching in the mean or matching in both the mean and variance. In each of three cases, the prediction distribution is shown as a smooth gray curve. It is the same in all three graphs, although the scale in the third graph is a bit different from the other two. The black step functions represent three different data sets as empirical distribution functions S_n. In the leftmost graph, the prediction distribution and the observed data have the same mean. But, otherwise, the data look pretty different from the prediction; the data appear to be mostly in two clusters on either side of the mean. Indeed, so long as the average of the data balances at that mean, those data clusters could be arbitrarily far away from each other, and any validation measure based only on the mean would not detect any discrepancy between the theory and the data, even though the data might bear utterly no resemblance to the prediction, apart from their matching in the mean. In the middle graph, both the mean and the variance match between the observed data and the theoretical prediction, but still one should not be very proud of the fit of the data here because of how deviant the prediction is with respect to the left tail of the distribution. Smaller values are more prevalent in the real data than were predicted. In the third graph, the conformance between the prediction and the data is good overall. This is reflected in the smallness of the area between the prediction distribution and the data distribution. The only way for the area to be small is for the two distributions to match closely in all respects. In each of these cases, the overall fit can be measured by the area between the two curves. It measures disagreements that the lower-order moments like the mean and variance cannot address.

In complex engineering systems it is not uncommon to have only one experimental test of the complete system.

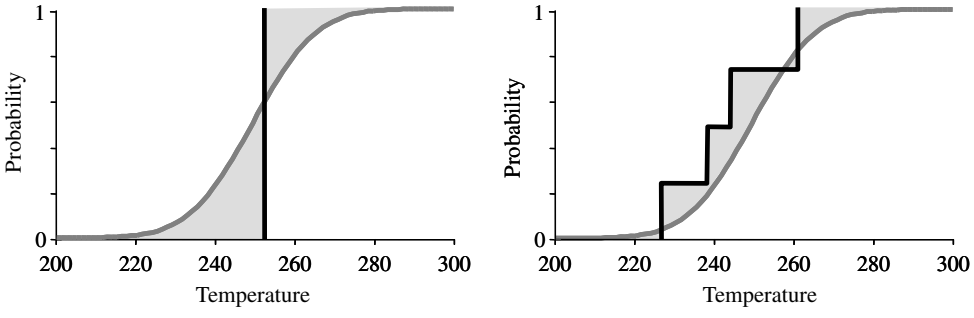

Fig. 8. Example data sets, with $n = 1$ on the left and $n = 4$ on the right, shown as S_n distributions (black) against a prediction distribution (gray), and areas (shaded) between the prediction distribution and the two data sets.

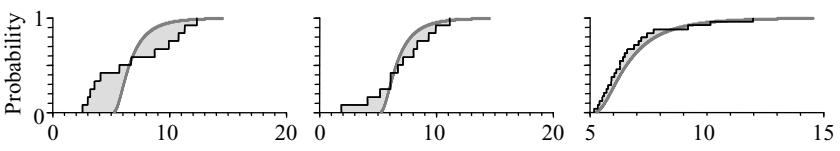

Fig. 9. Examples of matches between a prediction distribution (gray) and different empirical data sets (black).

This situation is reflected in the thermal challenge problem, particularly in the 'low' data set where there is a single observation to compare against each prediction distribution. In such cases, the empirical distributions are not complex step functions, but instead single-step spikes representing point values (i.e., degenerate distributions). Fig. 10 shows how the area metric varies with different values for a single datum matched against a prediction distribution. In these three examples, the prediction distribution is centered at 2 and ranges between 0 and 4. The most important thing to notice is that a single value can never perfectly match the entire distribution, unless that distribution is itself a degenerate point value. The first and second graphs of the figure compare the prediction distribution to a single observation at 2.25 and 1.18, respectively, yielding corresponding values for the area metric of 0.44 and 0.86. About the best possible match that a single datum could have occurs when the datum is located at the distribution's median, but, even there, the area metric will often be pretty large. In the case of the distribution depicted in Fig. 10, the area metric will be smallest when the observation is 2, which yields a value of 0.4 for the metric. That value depends on the shape of the prediction distribution, especially its kurtosis. If, for example, the prediction distribution were uniform over the range $[a, b]$, a single observation cannot be any 'closer' to it than $(b - a)/4$, which is the value of the area metric if the point is at the median. That's the best match possible with a single data point. How *bad* could the match be? The match could be very bad; indeed, it can be bad to an arbitrarily large degree. The rightmost graph in Fig. 10 shows another example of a single datum compared to the same prediction distribution. In this case the data point is at 26, which means that it is about 24 units away from the distribution. The area metric can be arbitrarily large, and it reduces to the simple difference between the datum and the prediction when both are point values. Because the ordinate probability is dimensionless, the units of the area are always the same as the units of the abscissa.

The area metric depends on the scale in which the prediction distribution and data are expressed. The two graphs in Fig. 11 depict a pair of comparisons in which the corresponding shapes are identical but the scales are different, as though the left graph were expressed in meters and the right graph in centimeters. Although the shapes are the same, the area metric is different by 100 fold. It would, of course, be possible to normalize the area measure, perhaps by dividing it by the standard deviation of the prediction distribution, but we do not believe this would be a good idea because the result would no longer be expressed in the physical units of the abscissa. Such a normalization would destroy the *meaning* of the metric.

Fig. 12 illustrates why retaining the scale and physical units (degrees, seconds, meters, Newtons, etc.) of the data is important for the intuitive appeal of a validation metric. The two graphs are drawn with the same x-axis, and they

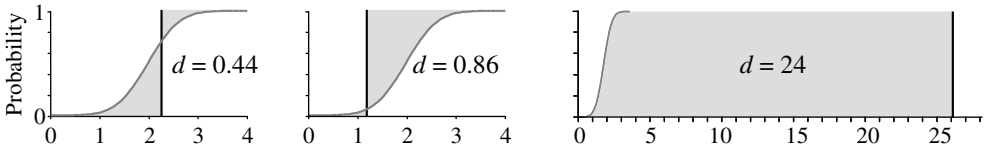

Fig. 10. Comparisons of a prediction distribution (gray curve) with three different data points (black spikes).

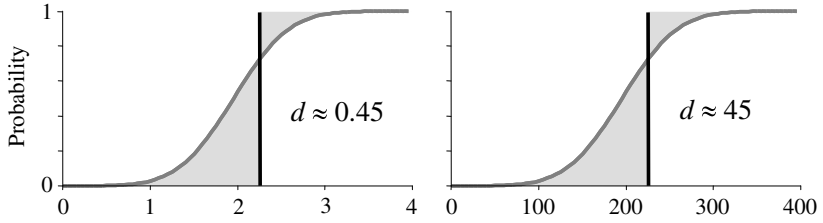

Fig. 11. Further examples showing the metric's dependence on the scale.

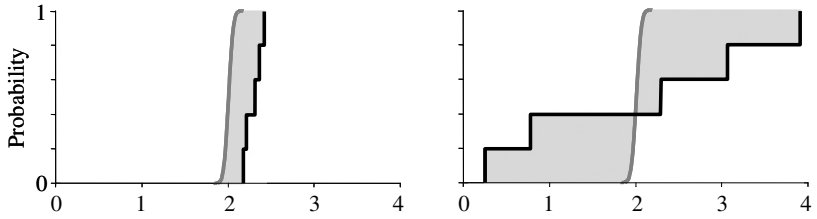

Fig. 12. Why physical units are important for a validation metric.

depict the same prediction distribution as a gray curve concentrated around 2. Two data sets are summarized as the S_n distributions shown as black step functions on the graphs. A statistician might argue that the comparison in the right graph reveals a better match between the theory and the data than the comparison in the left graph. In the left graph, the two distributions do not even overlap, whereas in the right graph the distributions at least overlap and are similar in their means. Using a traditional Kolmogorov–Smirnov test for differences between the two distributions, a statistician would find statistically significant evidence the distributions in the left graph are different ($d_S = 1.0$, $n = 5$, $p < 0.05$), but would *fail* to find such evidence for the distributions in the right graph ($d_S = 0.6$, $n = 5$, $p > 0.05$). But this is not at all how engineers would understand these two comparisons. For engineers, the main focus is on the discrepancy, in units along the x-axis, between the two distributions. In this sense, the comparison on the left is a much better match between theory and data than is the comparison on the right. Engineers have a strong intuition that the data–theory comparison on the left has a better match than that on the right, even though the distributions on the left do not even overlap with each other. The discrepancy on the left is never larger than half a unit along the x-axis, whereas the discrepancy on the right could be larger than two units. It is this physical distance measuring the disagreement that really matters to engineers, not some arcane probabilistic distance. This is the reason why the validation metric should be expressed in the original units, as is the case for the area metric.

Finally, consider the behavior of the area metric as theory and evidence diverge further and further. Fig. 13 shows two graphs, each with a prediction distribution drawn in gray and data distribution drawn in black. Neither data distribution overlaps its respective prediction distribution. The traditional and commonly used Smirnov's distance (which is the maximum vertical distance between the two distributions) cannot distinguish between these two comparisons. The maximal vertical distance in both cases is just unity, so the distributions are both as far apart as they can be according to the Smirnov metric. Under this measure, each data distribution is simply 'far' from its prediction distribution. The area metric, on the other hand, is about 2 for the left graph and about 40 for the right graph. The area metric therefore identifies the left graph as having considerably more concordance between data and prediction than the right graph. If the criterion for an acceptably accurate prediction is that it is within 10 units of the actual data, then it might be that the prediction in the left graph is acceptable for the intended purpose, even though the prediction in this case does not overlap with the data. Likewise, given the same accuracy requirement, the prediction in the right graph is not acceptable for the intended use.

5.4. Validation assessment when predictions are sparse

Engineers naturally strive for high fidelity between their computer models and the underlying physical reality they try to represent. In many disciplines, this leads to extremely complex and detailed models. In fluid dynamics modeling of turbulence, for example, it is possible today for computer models to use 10^{15} grid points! But there are practical tradeoffs to such fidelity. Because the computational burden is so large, it is sometimes difficult or, in some extreme cases, even practically impossible to extract any type of uncertainty information from these highly detailed models.

The area metric proposed here is applicable even when the predictions are sparse. Suppose, for instance, it is practical to conduct only a small number of simulations of a complex model to produce a handful of quantitative predictions. Although these models cannot produce the smooth prediction distributions that other models generate with many simulation runs, it may be reasonable to consider the values computed to be *samples* from that smooth distribution. If they are random samples (as they would be if inputs are selected randomly), then the 'empirical' distribution function formed from these values is an unbiased nonparametric estimator of the true distribution that would emerge with asymptotically many runs. The prediction distributions for the challenge problem that are displayed in Figs. 3 and 7 were each based on 10,000 Monte Carlo replications, but it is clear that the area validation metric could be applied even if many fewer replications had been employed. In fact, the area metric can in principle be applied to the case when the prediction distribution is characterized with only few simulation runs of the model. Fig. 14 illustrates the idea of constructing S_n functions for both data and predictions, the latter being the values from the sample runs of the model. In this case, there were only three simulations of the model conducted, whose values are random samples from an underlying distribution. They are to be compared against a data set consisting of five values.

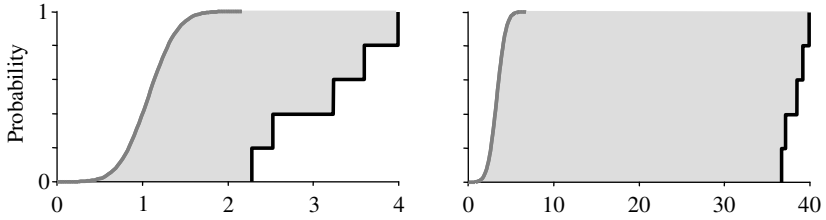

Fig. 13. Distinguishing non-overlapping data and prediction distributions.

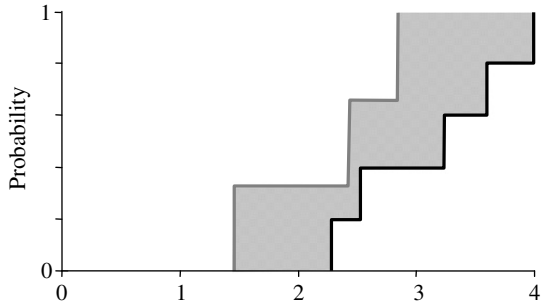

Fig. 14. The area metric when the prediction distribution (gray) is characterized by only three simulation runs.

Having very few simulation runs really means that analysts can construct only an impoverished picture of what the model is actually predicting. This implies the prediction will be accompanied by substantial epistemic uncertainty arising from sampling error. Future work will address the issue of expressing this and other forms of epistemic uncertainty in the prediction distribution and accounting for it with the area validation metric.

5.5. Pooling incomparable comparisons

This section describes an approach based on the area metric that can be used to integrate the evidence from all relevant data over the entire validation domain into a single measure of overall mismatch.

When several observations are collected for a single prediction distribution, the empirical distribution S_n is used to pool these experimental data into a single object for comparison against the prediction distribution. This pooling is not possible, however, when data are to be compared against *different* distributions. If, for example, the model predicts temperature at several instances as a function of time, it would not be reasonable to pool observed temperatures collected along the time axis (unless their predicted distributions happened to be identical). In this case, the data must be compared with their respective prediction distributions. Does that imply that we will have multiple values of the validation metric to compute? One could certainly compute all the areas separately for each pair of prediction distribution and its observation(s). But, if we did this, how could we combine the resulting areas together in some sensible way into an aggregate measure of the overall discrepancy between the model's predictions and the experimental data?

The most workable strategy around this problem is to express the conformance of theory and data on some universal scale. The natural scale for this purpose is probability. By transforming every datum x_i according to its corresponding prediction distribution F_i, we obtain a value $u_i = F_i(x_i)$, which ranges on the unit interval $[0,1]$. Fig. 15 shows three such transformations for hypothetical observations depicted as spikes and their corresponding prediction distributions shown as gray curves. For example, in the thermal challenge problem these three graphs could be the temperature responses at three different values of time t or three different values of location x. The intersections of the spikes and their distribution functions identify values on the probability scale for each u-value. The prediction distributions F_i can be any shape at all, and they need not be the same for different observations. The u's are always defined because $F(x) = 1$ for any value of x larger than the largest value in the distribution, and $F(x) = 0$ for any value smaller than the smallest value in distribution.

The various resulting u-values can then be pooled to obtain an *overall* summary metric assessing the accuracy of the model's predictions. Under the assumption that the x_i really are distributed according to their respective distributions F_i, these $u_i = F_i(x_i)$ will have a uniform distribution on $[0,1]$. This fact is called the probability integral transform theorem in statistics [39]. This is what it means for a random variable to be "distributed according" to a distribution. The converse of this fact is perhaps more familiar to engineers because it is often used to generate random deviates from any specified probability distribution: given a distribution F and a uniform random value u between zero and one, the value $F^{-1}(u)$ will be a random variable distributed according to F. Conversely, if x is distributed according to F, then $u = F(x)$ is distributed according to a uniform distribution over $[0,1]$. None of this changes if there happen to be multiple x- and u-values, and, in fact, none of it changes if there are multiple distribution functions, so long as the x-values are properly matched with their respective distributions. The x-values are made into compatible u-values by this transformation. Because all the u-values are randomly and uniformly distributed over the same range, then pooling them together yields a set of values that are collectively randomly and uniformly distributed over that range. If, however, we find that the u_i are not distributed according to the uniform distribution over $[0,1]$, then we can infer that the x observations must *not* have been distributed according to their prediction distribution functions.

The distribution of pooled u-values can be studied to infer characteristics of the overall match between the x-values and their respective prediction distributions. For instance, the area metric can be applied directly to the u-values compared against the standard uniform distribution. Also, the model's (mis)match for different predictions, generated from their particular observations, can also be compared to each other. This would allow one to conclude, for example, that a model predicts well for, say, high temperatures but not for low temperatures. The reason this is possible is that we have transformed all the observations into the same universal probability scale for the comparisons.

Transforming the observations into a universal probability scale is useful for aggregating incomparable data, but, by itself, it has the disadvantage of abandoning the original physical units of the comparisons. This deficiency can be repaired by back-transforming the u-values through a suitable distribution function G that restores the scale

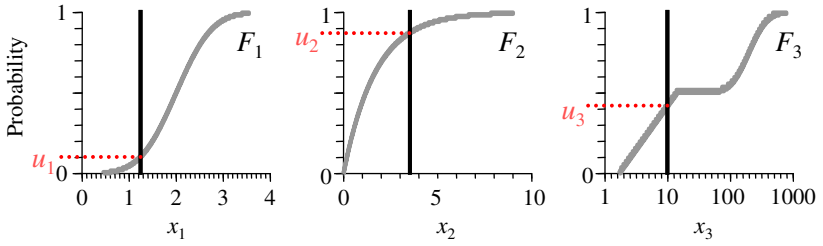

Fig. 15. Translation of observations (spikes) through prediction distributions (gray) to a universal probability scale.

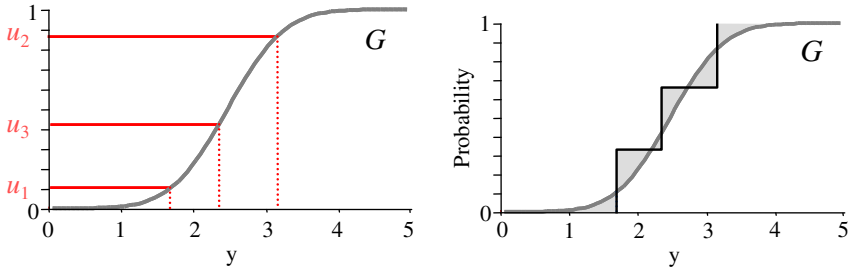

Fig. 16. Back-transformation from the u-scale to an archetypal scale determined by a distribution G (left) and the area metric for the pooled back-transformed values against the G distribution (right).

and its interpretation. Fig. 16 shows how this would work for the three u's that were computed in the previous figure. The right graph shows the empirical distribution function S_n of those three back-transformed data values

$$y_i = G^{-1}(u_i) = G^{-1}(F_i(x_i)).$$

All of these y_i have the same scale, which are inherited from G. This back-transformation depends on the specification of the G distribution. What distribution should be used? In the case of the challenge problem we are considering in this paper, it makes sense to use the prediction distribution about the regulatory requirement as the back-transformation distribution. This is the distribution, after all, that spells out *where* we are specifically interested in the model's predictive capability. Using the regulatory prediction distribution as the G distribution allows all the available observations that are germane to *any* predictions made by the model to be used to characterize the uncertainty about this most important prediction. The right graph in Fig. 16 also shows with shading the area metric between the back-transformed y_i and the prediction distribution G. This metric is in the physically meaningful units as a result of the back-transformation.

We can use the name 'u-pooling' for this procedure of transforming data to the probability scale and back-transforming to some specified scale. Just as forming the empirical distribution S_n from a collection of data allows us to pool those observations, u-pooling allows us to pool observations that correspond to different prediction distributions.

5.6. Testing the validation metric

Oberkampf and Barone [32] considered the important question of how an analyst might justify a conclusion that there is a significant disagreement between a theory's mean predictions and the means of its validation data based on quantitative analysis. Methods are needed to allow the analyst to answer the 'so what?' question about particular values of the validation metric. Consider, for instance, two situations. In the first, experimental observations have been exhaustively collected so that there is essentially no sampling uncertainty about the data distribution, and likewise the function evaluations are cheap so the prediction can also be specified without any sampling uncertainty. Suppose that we compute the validation metric in this situation to be $d = 1$. In the second situation, we compute the validation metric to be $d = 10$, but it is based on a very small sample size of empirical observations, or a small number of function evaluations, or both. In the first situation, the disparity between the predictions and the data is statistically significant in the sense that it cannot be explained by randomness arising from the sampling uncertainty (because there is none), but must rather be due to inaccuracies in the model. In the second situation, however, it is not clear that the disagreement between the predictions and the data is significant, even though it is ten times larger. The computed discrepancy might be entirely due to the vagaries of random chance that were at play when the observations and function evaluations were made. Some kind of statistical analysis is required to give a *context* for these two d values to understand when a value is big and when it is not really so big.

Oberkampf and Barone [32] offered a way to formally determine whether there is a substantial mismatch between validation data and a model's predictions. They compared the scatter of sampling variability seen among multiple experimental observations (quantified as traditional Neyman–Pearson confidence intervals) to the difference

between the mean output from the predictive model and the sample mean of the observational data. They show how the approach can be applied when the experimental data is autocorrelated, as it often is when observations are collected as functions of time or some other control variable. Their approach depends on replicated experimental observations, and cannot be used when there is only a single observation per prediction distribution. Moreover, their approach lacks a formal statistical underpinning because it is based on an intuitive argument about the plausible magnitude of the error between model and data in the face of sampling uncertainty about the latter.

We suggest that statistical methods to detect evidence of significant mismatch between a model and its validation data can be constructed by applying standard statistical tests to the u-values or the y-values derived in the previous section. These methods can be used by an analyst to formally justify an impression or conclusion that the experimental observations disagree with the model's predictions. Transforming the x-values to u-values and pooling all the u-values together can substantially increase the power of the statistical test because the sample size is larger in a single, synthetic analysis. For example, statistical tests for departures from uniformity, such as the traditional Kolmogorov–Smirnov test [35], applied to the u-values can identify significant overall failure of the model's predictive capability. This test assumes that the experimental data values are *independent* of one another, which is not always true in practice, especially when observations have been collected along a time course as in the challenge problem. There are other statistical tests that are also commonly applied in this situation, including the traditional chi-squared test and Neyman's smooth tests [40,41] and references therein. The test can also be applied to compare the y_i values against the predicted distribution G in the physically meaningful scale. One could also define statistical tests of whether the discrepancy between data and theory is larger than some threshold size.

6. Validation and prediction for the challenge problem

Sections 3 and 4 described two distinct implementations of the heating model and extracted from each a prediction about the distribution of surface temperatures after 1000 s. In Section 5 we discussed the development of the area validation metric and a scheme to pool comparisons made on different scales. In this section, we apply the validation metric to the thermal challenge problem. Using this metric we ask two questions: How well do the predictions match the actual observations that are available? And what does the match, or rather the mismatch, of the empirical data with the model's predictions tell us about what we should infer about other predictions? We first assess the performance of the model over the validation domain by u-pooling to compute a *summary* measurement of the overall mismatch between the model and the data. We also *extrapolate* the validation measures to characterize the predictive capability of each model to address the regulatory requirement in the challenge problem. These two uses of the validation metric are independent; the predictive capability is *not* based directly on the result of the validation assessment, but rather derived separately from the values of the validation metric from comparisons of individual experimental observations to the respective model predictions.

The designers of the thermal challenge problem introduced several subtleties that must be addressed. For instance, the experimental design that produced the validation data used several configurations of the environmental heat flux q and thickness L of the device material, none of which corresponded exactly to the configuration of interest in the statement of the regulatory requirement in Expression (2). There are also some temperature data at different positions within the device material, at the surface, in the middle and on the other side. Apparently, the purpose of this bracketing was to explore and account for any trends that might be present in the performance of the model. The validation data were divided in the challenge problem into ensemble and accreditation data, distinguished primarily by the closeness of the conditions to the configuration of interest in the regulatory requirement. Although we did not treat the ensemble and accreditation observations differently, doing so would not be incompatible with the use of the validation metric proposed in this paper.

In the ensemble data, temperatures were measured at each of 11 points in time between 0 and 1000 s, and, in the accreditation data, temperature was measured at 21 points in time during the heating. However, we presume that the initial temperatures of 25 °C that appear in the top rows (at $t = 0$ s) of Tables 6 and 8 in [4] are *not* to be compared against the model predictions because they are really initial conditions rather than observations of the course of heating. Thus, there are only 10 observations per experiment in the ensemble data set and only 20 time observations per experiment in the accreditation data set.

6.1. Model performance over the validation domain

To evaluate the overall performances of the two models we used the u-pooling approach described in Section 5.5. Every observation in the validation domain (which consists of the ensemble data plus the accreditation data) is associated with values of the control parameters q, L, x and t as determined by the experimenter at which the observation was collected. These four parameters, along with the distributions characterizing the variability of k and ρC_p, were used in Monte Carlo simulations (with 10,000 replications each) to compute the prediction distribution of temperatures from the heating model in Expression (1) defined by these control parameters. Every temperature observation in the validation domain is thereby paired with a prediction distribution of temperature. There are 140 of these pairs in the 'medium' data set. The pairs define the u-values $u_j = F_j(T_j)$, where T denotes an observed temperature and F denotes the associated prediction distribution and

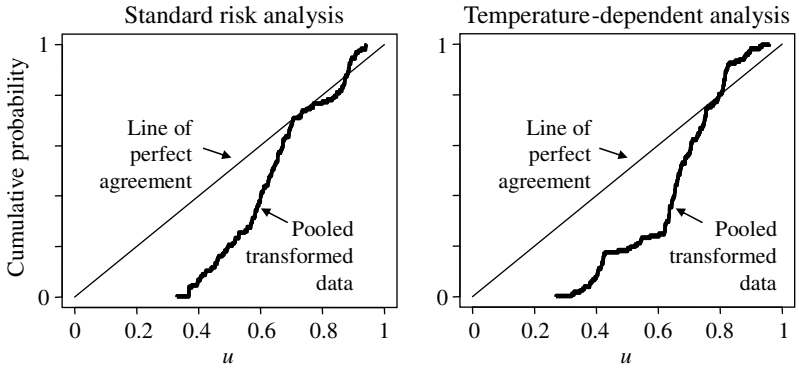

Fig. 17. Model performance for two analyses as summarized by observed distribution functions of u-values (step functions) compared to theoretical uniform distributions (straight lines).

$j = 1, \ldots, 140$. Fig. 17 shows empirical distributions of these u-values. These empirical distributions constitute summaries of the performances of the two analyses described in Sections 3 and 4. The step function in each graph represents all 140 observations of temperature in the 'medium' data set compared to their respective prediction distributions generated under that analysis. The step functions are the empirical distributions of the u-values produced by these comparisons. These distributions are to be compared against the uniform distribution over $[0, 1]$, whose graph appears in each graph as the 45° line. These are the lines of perfect agreement. If the observed temperatures were drawn from distributions matching those that were predicted by the model under an analysis, then the step function would match the uniform distribution, to within fluctuations due to random chance.

Fig. 18 shows the back-transformations of these distributions to the physically meaningful scales of temperature. The smooth curves in the graphs are the prediction distributions under the two analyses for the conditions of the regulatory requirement. (As explained in Section 2, these conditions were $x = 0$ cm, $t = 1000$ s, $Ti = 25$ °C, $q = 3500$ W/m^2, and $L = 1.90$ cm, which correspond to the distributions previously displayed in Figs. 3 and 7.) Also displayed in Fig. 18 as step functions are the corresponding pooled distributions of u-values *back-transformed onto the same temperature scale* via the inverse probability integral transforms specified by the respective prediction distributions. That is, they are the distributions of the quantity $G^{-1}(u_j)$ where G is the prediction distribution of regulatory interest under each of the two parallel risk analyses.

The graphs in Fig. 18 are just nonlinear rescalings of the graphs in Fig. 17. Under these rescalings the straight lines become the smooth curves, and the tails of the step functions are stretched relative to central values of the distributions. For each graph, the transformation of the abscissa is exactly the one that changes the standard uniform distribution into the prediction distribution. The ordinate is also flipped from cumulative to exceedance probability. The result translates the evidence embodied in all 140 observations in the ensemble and accreditation data sets onto the scale defined by the prediction distribution for the regulatory requirement in Expression (2). These distributions should not be interpreted as though they were themselves data collected on this very temperature scale. They were, after all, collected on a variety of temperature scales and are pooled for the sake of this comparison. Thus, they do not represent direct evidence about what the temperatures

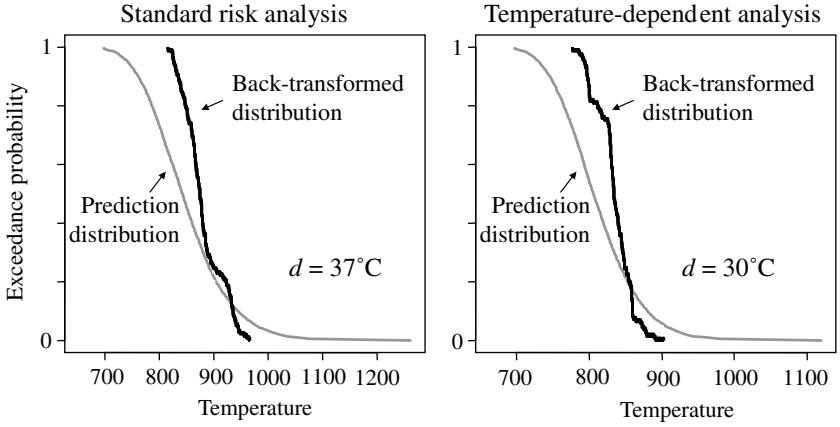

Fig. 18. Distributions of back-transformed u-values (step functions) compared to predicted temperature distributions (smooth lines) for the two analyses.

will be under the conditions of the regulatory requirement, but only how well the prediction distributions produced by the model have matched temperatures so extreme under other conditions in the validation domain.

The temperature-dependent risk analysis, illustrated in the right graph, has a somewhat better match with the available empirical data than the standard risk analysis, illustrated in the left graph. The distribution of back-transformed u-values is closer in this analysis to its prediction distribution shown as the smooth curve. The superiority of this match is reflected by the area metric d in the following table.

Analysis	d	95% confidence interval
Standard risk analysis	37.2 °C	[34.0, 42.7] °C
Temperature-dependent analysis	30.4 °C	[27.4, 33.7] °C

Recall that the area metric d measures the area between the empirical distribution of back-transformed pooled u-values and the prediction distribution which is our expectation about them. The 95% confidence intervals were computed by a nonparametric bootstrap method [42] based on resampling from the 140 u-values. That is, we took a random sample of size 140 from the distribution of u-values (with replacement obviously) and recomputed the value of d by comparing the distribution of these randomly selected u's to the prediction distribution. We repeated this process 10,000 times and sorted the resulting array of d-values. The 95% confidence interval was estimated as $[d_{(2.5N/100)}, d_{(N-(2.5/100)N)}] = [d_{(250)}, d_{(9750)}]$, i.e., the interval between the 250th and 9750th values from the sorted list of 10,000d's. Each confidence interval estimates the sampling uncertainty associated with the actual value of d arising from having computed it from only 140 observations.

Both of the graphs in Fig. 18 suggest that Expression (1) is somewhat better at predicting temperatures close to 900 °C than it is at predicting much lower temperatures. This may be the result of the model in [4, (Eq. (2))] having been calibrated to perform well around this temperature. In any case, good model performance for such temperatures is likely to be of considerable interest in an application like the challenge problem. The temperature-dependent analysis used regression and an iterative simulation scheme to represent the dependence evident in the materials characterization data between temperature and thermal conductivity. The reward for the substantial extra calculation is a modest improvement in the model's performance vis-à-vis the data. The match for this analysis was quantitatively better than that of the standard risk analysis.

Calculations were also done with the 'low' and 'high' data sets, which yielded qualitatively similar results; the temperature-dependent analysis was somewhat better than the standard risk analysis for both data sets in the area metric. The performances of the model under the two analyses for all three data sets in terms of the area metric d and a 95% bootstrap confidence interval around it are given in units of °C in this table:

Analysis	Low ($n = 100$)	Medium ($n = 140$)	High ($n = 280$)
Standard risk analysis	52.6, [49.4, 55.9]	37.2, [34.0, 42.7]	18.5, [15.3, 23.9]
Temperature-dependent analysis	34.1, [30.0, 38.1]	30.4, [27.4, 33.7]	11.6, [9.5, 15.0]

Fig. 19 summarizes the performance results graphically as distributions of the back-transformed u-values compared to the respective prediction distributions for the two analyses under the 'low' and 'high' data sets.

The graphs in Figs. 18 and 19 reveal that the area metric d as used in the validation assessment is strongly sensitive to the sample size of the observations on which it is based. This is not because the model is getting more accurate with more data. Rather, it is the result of there being *more evidence* of a good match between the model and the data. For the challenge problem, as the number of observations increases through the three data sets, the value of the area metric declines considerably. Of course, this might not have been true if the model were making inaccurate predictions. Insofar as the model is accurate, however, increasing the sample size decreases d to its lower limit of zero. The dependence of d on the number of observations means that we should only compare performances that are based on the same sample size, i.e., within columns of the table above. In this case, the temperature-dependent analysis has a consistently better (smaller) overall d than the standard risk analysis at all three sample sizes described in the challenge problem.

6.2. Extrapolation and predictive capability

What does overall model uncertainty *as assessed by comparing predictions from each analysis to observations* tell us about the reliability of the model's prediction about the regulatory requirement? To answer this question we employed linear regression to extrapolate the evidence about the (mis)match between predictions and data seen under several experimental conditions to the set of conditions relevant to the regulatory requirement in Expression (2).

For the sake of simplicity, we describe this extrapolation only for the temperature-dependent risk analysis under the 'medium' data set. Entirely analogous calculations are possible for the other analysis and the other data sets. Assessing the predictive capability is essentially putting uncertainty bands around forecasts. In the case of the thermal challenge problem under the temperature-dependent risk analysis, Fig. 7 showed that the *expected* risk of being larger than 900 °C after 1000 s is already five times larger than 0.01, which is the upper limit on this probability allowed by the

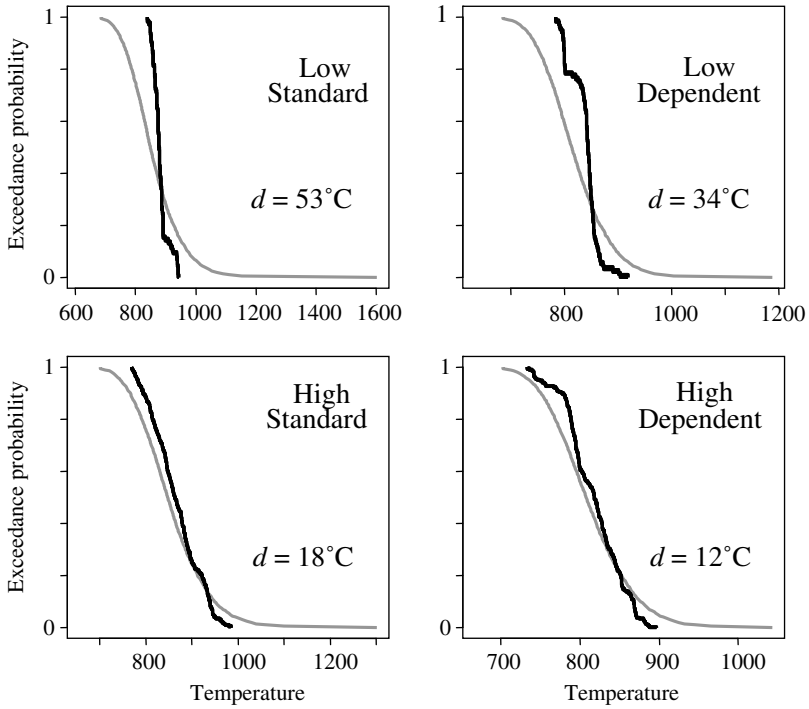

Fig. 19. Model performances as distributions of back-transformed u-values (step functions) compared to prediction distributions (smooth curves) for 'low' and 'high' data sets and the standard and temperature-dependent risk analyses.

regulatory requirement. Adding uncertainty around this risk will only make the upper limit on the probability get larger. Nothing that is computed in this section could ever suddenly make the system seem to come into compliance with the regulatory requirement. We note at the start that the question about predictive capability is already a moot one for whether the device can sufficiently insulate under the specified heat flux to meet the regulatory requirement. We make the calculations described below not in answer to the challenge problem—we already know that this answer is 'no'—but as an example of how we suggest one could estimate uncertainty due to the empirical error and extrapolation of the model to the conditions of intended use.

There were 140 temperature observations made at various heat fluxes q, thicknesses L, positions x, and times t in the 'medium' data set. (Interestingly, only one initial temperature T_i was considered in the challenge problem.) As described in Section 6.1, the prediction distribution from the heating model in Expression (1) under each of these 140 specifications was computed with a Monte Carlo simulation. Each observed temperature was then compared to its respective prediction distribution and a u-value was computed. Instead of first pooling these 140 u-values as we did in the previous section, we back-transformed each u-value directly to the temperature scale using the prediction distribution associated with the conditions of the regulatory requirement in Expression (2). That is, we computed the back-transformed temperature $T_j^* = G_{-1}(F_j(T_j))$ where T denotes observed temperature, F denotes the associated prediction distribution, G is the distribution depicted in Fig. 7, and $j = 1, \ldots, 140$ indexes the observations. We then computed the area metric $d(T_j^*, G)$ between the back-transformed temperature and the prediction distribution of regulatory interest. These comparisons of scalar values and the prediction distribution for temperature at the conditions for the regulatory requirement thus yielded 140 values of the area metric. The mean of these areas was 69, with values ranging between 56 and 129.

The 140 areas were regressed against the input variables q, L, x, t with a linear model to look for any trends that might be present, and to develop a statistical model of the sampling uncertainty associated with a prediction of the value of the area metric expected at the conditions for the regulatory requirement. The regression yielded strongly significant regression coefficients for q, L and t (with $p < 0.001$) but found no significance for the x variable. The best fitting linear model for the expected value of the area metric for a single (new) observation as a function of heat flux, thickness, position and time is

$$126 - 0.0160q - 914L + 201x - 0.0124t + N(0, 10.8),$$

where the last term is the normally distributed residual error arising from unexplained scatter in the area values. At the configuration for the regulatory requirement, this becomes

$$126 - 0.0160(3500) - 914(0.019) + 201(0) - 0.0124(1000) + N(0, 10.8),$$

which equals $40.2 + N(0, 10.8)$. Thus, the regression model under those conditions is saying we expect a value for the area metric to be about 40, with some uncertainty so that

it typically ranges from about 30 to 50. Other regression models might have been used for this extrapolation, possibly including nonlinear relationships or other variables, perhaps even the predicted temperature itself.

Because we are extrapolating to conditions specified in the regulatory requirement that were never directly studied, it is especially important to account for sampling uncertainty in predicting the magnitude of the area metric under those conditions. A standard way to do this for linear regressions is to compute the prediction intervals. The 95% prediction interval for the value of the area metric predicted under the conditions $x = 0$ cm, $t = 1000$ s, $T_i = 25$ °C, $q = 3500$ W/m^2, and $L = 1.90$ cm, is $[17, 63]$ °C. Because we are trying to capture the epistemic uncertainty about the model's predictions vis-à-vis the data, we would always use the *upper* bound from the prediction interval as our estimate of the possible error of the model.

How should this extrapolation of the area metric as a function of q, L, x, t be interpreted as predictive capability? This is an open question. Fig. 20 shows one possible way. The graph depicts a prediction distribution as the inner curve, with a parallel distribution on either side of it displaced in the positive or negative direction just enough so that the area between each curve and the prediction in the middle is the specified amount extrapolated from the validation assessment. Given the performance of the model, one could not reasonably expect the prediction to have accuracy any better than that characterized by these outer distributions. In our view, they represent the minimal uncertainty associated with the predictive capability of the model as evidenced by the validation assessment on available data. Thus, the predictive capability of the model for future data sets is at least as bad as depicted in this graph.

This assessment applies, in particular, to the prediction distributions from the traditional risk analysis (Section 3) and the temperature-dependent risk analysis (Section 4) from which we made inferences about the probability of surface temperatures being larger than 900 °C. By displacing the prediction distributions (in both directions) by the upper bound of the prediction interval for the value of the area metric from the regression extrapolated to the conditions of regulatory interest, lower and upper bounds on the probability of exceeding 900 °C can be read from the graphs as the vertical interval on probability between the bounds at that temperature. These bounds characterize the predictive capability with respect to the regulatory requirement.

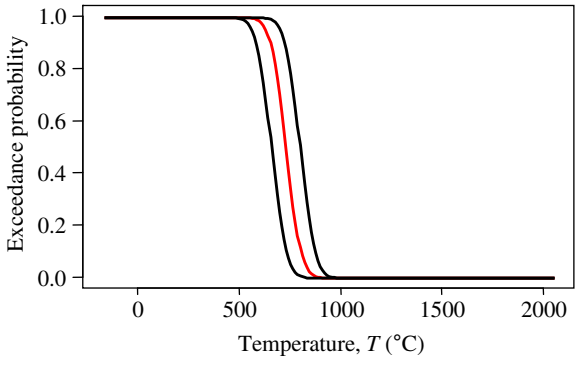

Fig. 20. One way to interpret an estimated area metric value as predictive capability.

The expectation that future data would form a distribution within these bounds amounts to assuming that the shape of the distribution is correct and that only its location along the abscissa is in real doubt. There are other more conservative ways one might extrapolate the area metric to forecast predictive capability. For instance, the outer curves might be constructed as the envelope of all distribution functions that are as close to the prediction distribution as the area metric observed in the validation.

Would it still be reasonable to use regression if there had been validation data at the particular configuration of interest? In the case of the thermal challenge problem, if observations of temperature had been collected under the precise conditions mentioned in the regulatory requirement, would there be any point to developing a regression? Even though the regression is not needed for an extrapolation, analysts might still prefer to make use of the ancillary information collected under other configurations to improve the assessment of the predictive capability. In addition to a better estimate of the expectation, regression analyses also permit the estimate of confidence intervals that would be useful in expressing the effect of sampling uncertainty on the reliability of the estimate.

As mentioned in the introduction to Section 6, the estimate of the model's predictive capability is not based on the summary validation metric computed by first pooling u-values. Instead, the empirical evidence is expressed in terms of the validation metric and then aggregated via a regression analysis. The reason it has to be done this way is that pooling uses up all the data to compute a single value of the validation metric and does not allow us to later decompose the evidence about the data–model mismatch in a way that is necessary for the extrapolation to characterize the uncertainty about the model's prediction at the conditions of regulatory interest.

7. Specific taskings

The tasking document [5] for the challenge problems poses several specific questions to be asked about any proposed solution to the problem. We do not reproduce those questions here, but only give our responses to them. These answers serve as a summary of the approach we have used to address the thermal challenge problem. In the preambles to the questions, Hills et al. [5] assert several times that replicates within the ensemble and accreditation databases are independent. Although they were not clear in this matter, they must have been referring to the replication of individual experiments and not referring to measurements at different times during the same experiment. That is, *experiments* consisting of ten or twenty observations as a function of time are independent of one another, but the

individual temperature *observations* at various times are certainly not mutually independent.

Both Hills et al. [5] and Dowding et al. [4] draw a distinction between "ensemble" and "accreditation" data, but we have not observed such a distinction in the analyses described here. Both are experimental observations of the system response quantity, collected under conditions variously close to or far from the conditions specified by the regulatory requirement. For us, there is no qualitative difference between the two kinds of observations, and we treat them in a unified way, both in the analyses described in this paper and in our responses to the questions of the tasking document.

7.1. Material characterization responses

1. The variability observed in the material characterization data is recognized as stochastic variation that accompanies manufacture of the device out of the shielding material. Different devices will have shielding with different material properties. We used normal probability distributions to characterize this variation, and estimated the distributions' parameters using the method of matching moments. Sensitivity analysis was used to characterize the effect of uncertainty about the normality assumption.
2. The benefit of additional data (from 'low' to 'medium', and from 'medium' to 'high') on material characterization could be quantified with any of many standard statistical goodness-of-fit measures used in fitting probability distributions. We did not bother to compute these measures for the material characterization here (but see the response to question #5 in Section 7.2 below).

7.2. Validation with ensemble and accreditation data responses

1. The area between the model's prediction distribution and the empirical distribution function of measurements of the system response quantity is used as the validation metric.
2. The area validation metrics d and their 95% confidence intervals for the standard risk analysis and the temperature-dependent analysis are summarized in this table:

Analysis	Low ($n = 100$)	Medium ($n = 140$)	High ($n = 280$)
Standard risk analysis	52.6, [49.4, 55.9]	37.2, [34.0, 42.7]	18.5, [15.3, 23.9]
Temperature-dependent	34.1, [30.0, 38.1]	30.4, [27.4, 33.7]	11.6, [9.5, 15.0]

Increasing sample size decreases the value of the validation metric, and tends to narrow its confidence interval, at least in the case of the temperature-dependent analysis which exhibits the better match to the available data.

3. Analysts should not specify the requirements on accuracy; such requirements are the province of decision makers. The reason is that different decision makers and different uses have widely different demands on a model's accuracy. Analysts should only *determine* the accuracy of the model and estimate the uncertainty about a prediction, which may be good or poor given the needs of the decision maker. In response to this quantification of a model's accuracy, a decision maker might reject the model's use for a particular application, or alternatively the decision maker might infer from the model's estimated inaccuracy that an engineered system must be designed more conservatively with larger margins of tolerance to account for the poor model performance.
4. A calibration step is not recommended as a part of validation. We argue that calibration and validation should be carefully distinguished. (Fitting distributions to the thermal conductivities and heat capacities in the material characterization data, although part of the modeling process, is not considered to be calibration because the distributions were not selected with reference to the output system response quantity of temperature.)
5. Having more data (from 'low' to 'medium', and from 'medium' to 'high') increases the evidence for model's apparent accuracy, as reflected in the summary validation metric. Having more data generally allows us to be more confident about the assessment of that accuracy as reflected in the confidence intervals for the validation metric. The impact of increasing the number of tests was not specifically quantified under this exercise. However, the validation metric is sensitive to sample size and this is reflected in the results of the analyses of the challenge problem. Because the area validation metric quantifies *evidence* of model accuracy, it reflects sample size as well as the closeness of the values to the predictions. In comparisons with continuous prediction distributions, the metric can approach zero only when sample size is large. Thus, the benefit of additional data could in principle be quantified by the decrease in the summary area metric assessing overall mismatch. In the case of the standard risk analysis, when data increases from 100 to 140 and then to 280 observations, the summary area metric decreased from 53 to 37 and then to 18 °C. For the temperature-dependent analysis, the decrease was from 34 to 30 and then to 12 °C.
6. A minimum of one function evaluation per prediction is required to use the proposed validation approach. Increasing the number of function evaluations per prediction improves the characterization of the prediction distribution, typically smoothing the distribution. For the challenge problem, ten thousand function evalua-

tions per prediction were used because the heating equation was so simple to compute. In cases where the model is computationally expensive, the proposed validation metric can be applied with many fewer evaluations.

7.3. Regulatory assessment responses

1. The variabilities observed in the materials characterization were used to define probability distributions representing parameters of the heating model. This probabilistic model was used to make predictions about surface temperatures after heating which were compared to observed temperatures. The error between the observed and predicted temperatures was extrapolated by a statistical regression to characterize the uncertainty of prediction of the probabilistic model under the conditions of regulatory interest. The resulting assessment was adequate to determine whether or not the regulatory requirement about the surface temperature after heating will be met.
2. We constructed two analyses based on the heating model described in the thermal challenge problem. Both analyses predict the probability that the system response exceeds the regulatory temperature criterion of 900 °C, and both suggest that this probability is far larger than the regulatory requirement that it be less than 0.01. The following table gives the estimated probabilities for the two analyses and the three data sets:

Analysis	Low	Medium	High
Standard risk analysis	0.28, [0.13, 0.52]	0.22, [0.17, 0.29]	0.24, [0.17, 0.32]
Temperature-dependent	0.09, [0.04, 0.18]	0.05, [0.03, 0.09]	0.05, [0.03, 0.09]

   Also given in square brackets are bounds on those probabilities implied by the model's predictive capability which was assessed in the validation with the area metric. The bounds were obtained from displaced distributions (such as shown in Fig. 20). Note that these bounds on the probabilities are wider than had been characterized in the robustness analysis of Section 3.1, because they now include the uncertainty arising from the observed disagreement of the model compared to the available data. None of the analyses suggest that the system will conform with the regulatory criterion.
3. The confidence of these predictions about this probability was assessed by extrapolating the observed error of predicted temperature distributions compared to relevant data. The resulting uncertainty bounds envelope the threshold probability of the regulatory requirement and therefore prevent a conclusion about whether the system is in compliance. Nevertheless, because the central estimates are already out of compliance, we can conclude that there is no evidence that the system is in compliance. This conclusion represents the third step, the adequacy decision, in our conceptual view of validation described in the introduction.

8. Conclusions

Although a deterministic analysis of the thermal challenge problem would suggest that the surface temperature after 1000 s of heating would be less than 900 °C, a standard risk analysis clearly reveals that it will not satisfy the stated regulatory requirement that surface temperature exceeds 900 °C with probability less than 0.01. Indeed, simulations that account for the stochasticity observed among the materials characterization data suggest that the probability of exceeding this temperature is many times more likely than this threshold probability. The variation in this result arises from uncertainty about how the materials characterization data should be modeled. Using the assumption suggested by the original challenge problem that the material's thermal conductivity is independent of its temperature, we obtain the result that temperatures will exceed the probability specified by the regulatory requirement by a factor of 22. Accounting for the observed correlation between temperature and thermal conductivity, we predict that the probability will exceed the regulatory requirement by only a factor of five. Under both analyses, the system appears to be clearly out of compliance with the regulatory requirement.

This paper has proposed to use the area between the prediction distribution and the data distribution (i.e., the empirical distribution function) as a validation measure that has several desirable properties, including objectiveness and robustness, being a true unbounded metric, retaining the units of the data themselves, generalizing the deterministic difference, and applicable even very few experimental or computational realizations are available. The proposed area metric can be used to measure the overall error of the model in the face of observational data. It represents an empirical assessment of model-form uncertainty that should inform the interpretation of any predictions the model makes. This uncertainty is apart from any uncertainty that arises internally within the model as a consequence of variability or uncertainty about model parameters, and it can be assessed from validation data whether or not any uncertainties about parameters or about the form of the model itself have been propagated through the model. This simple area metric can be generalized by various weighting schemes to account for situations when, for example, overestimates are preferred to underestimates or for situations in which matching in the distribution tails is more important than an overall good match. This flexibility may turn out to be an especially important feature of the proposed metric.

We take validation to involve two closely related questions. The first question asks how good the model is, and

the second asks what this match or mismatch might imply about the performance of the model for future predictions in some intended application. The first can be answered by assessing the disagreement between the model's predictions and whatever new data may be available using the proposed area validation metric. The second question can be answered by assessing the observed error between the model's predictions and relevant empirical observations that are available under given conditions, and extrapolating this error to the condition under which the prediction is to be made. In the case of the challenge problem, the mismatch observed between the heating model and the empirical observations is large enough to obscure the performance of the system so that, in fact, we cannot conclude that the system could not possibly be in compliance with the regulatory requirement.

The approach to validation as described in this paper based on the proposed area metric is fundamentally unlike the more common validation approaches based on Bayesian methods or hypothesis testing. In its focus on updating, the Bayesian approach integrates validation with calibration, which we believe ought to be rigorously separated. Bayesians want to use whatever data is available to improve the model as much as possible. Our approach reserves the first use of any validation data for assessing the performance of the model so as to reveal to decision makers and other would-be users of the model an unvarnished—and unrevisited—characterization of its predictive abilities.

Further research is needed on several issues, including

1. How measurement uncertainty associated with experimental observations should be addressed in validation and how it can be incorporated in the proposed area metric.
2. Whether and how the area metric should be applied when observational data is entirely outside the range of the model's prediction so as to be theoretically 'impossible' under the model; and
3. How the *u*-pooling and back-transformation strategies might be extended to aggregate data for simultaneous predictions of different system response quantities (such as temperature and electric resistance) which may have entirely different units.

A forthcoming paper will explore these issues in some depth.

Acknowledgements

We thank Marty Pilch who has been the force behind the formulation of the validation challenge problems and the Sandia validation workshop. We thank the other organizers Kevin Dowding, John Red-Horse, Richard Hills (New Mexico State University), Ivo Babuska (University of Texas), Raul Tempone (Florida State University) and Tom Paez, for inviting us to participate in the validation challenge. We also thank Marvin Adams (Texas A&M), Gianluca Iaccarino (Stanford University), Chris Roy (Auburn University), Vladik Kreinovich (University of Texas at El Paso), Chris Paredis (Georgia Tech), Arnold Neumaier (Universität Wien), Tony Cox (Cox Associates) and the attendees at the Validation Challenge Workshop in Albuquerque, 21–23 May 2006, for stimulating discussions about validation. Our effort was funded by the Sandia Epistemic Uncertainty Project, directed by William Oberkampf.

References

[1] AIAA, Guide for the verification and validation of computational fluid dynamics simulations, AIAA G-077-1998, American Institute of Aeronautics and Astronautics, Reston, Virginia, 1998.
[2] ASME, Guide for verification and validation in computational solid mechanics, ASME V&V 10-2006, American Society of Mechanical Engineers, 2006, Available from: <http://catalog.asme.org/Codes/PrintBook/VV_10_2006_Guide_Verification.cfm>.
[3] W.L. Oberkampf, T.G. Trucano, Verification and validation benchmarks, SAND2007-0853, Sandia National Laboratories, Albuquerque, NM, Nucl. Engrg. Des., in press.
[4] K.J. Dowding, M. Pilch, R.G. Hills, Formulation of the thermal problem, Comput. Methods Appl. Mech. Engrg. 197 (29–32) (2008) 2385–2389.
[5] R.G. Hills, M. Pilch, K.J. Dowding, I. Babuska, R. Tempone, Model validation challenge problems: tasking document, Comput. Methods Appl. Mech. Engrg., this workshop.
[6] Reference to the editor's introduction.
[7] G.E. Box, N.R. Draper, Empirical Model-building and Response Surfaces, Wiley, 1987 [quotation appears on page 424].
[8] M.G. Morgan, M. Henrion, Uncertainty A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press, 1990.
[9] A.C. Cullen, H.C. Frey, Probabilistic Techniques in Exposure Assessment, Plenum Press, 1999.
[10] J.L. Devore, Probability and Statistics for Engineers and Scientists, Duxbury Pacific Grove, California, 2000.
[11] S. Ferson, R.B. Nelsen, J. Hajagos, D.J. Berleant, J. Zhang, W.T. Tucker, L.R. Ginzburg, W.L. Oberkampf, Dependence in probabilistic modeling, Dempster–Shafer theory, and probability bounds analysis, SAND2004-3072, Sandia National Laboratories, Albuquerque, NM, 2004, <http://www.ramas.com/depend.zip>.
[12] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2004, See <http://www.R-project.org>.
[13] R.G. Hills, I. Leslie, K.J. Dowding, Statistical validation of engineering and scientific models: application to the abnormal environment, SAND2004-1029, Sandia National Laboratories, Albuquerque, NM 2004.
[14] W.L. Oberkampf, T.G. Trucano, C. Hirsch, Verification validation and predictive capability in computational engineering and physics, Appl. Mech. Rev. 57 (5) (2004) 345–384.
[15] R.G. Hills, Model validation: model parameter and measurement uncertainty, J. Heat Transfer 128 (2006) 339–351.
[16] V.J. Romero, Validated model? Not so fast, The need for model "conditioning" as an essential addendum to model validation, in: AIAA-2007-1953 in Proceedings of the 2007 AIAA Non-Deterministic Approaches Conference, Honolulu, American Institute of Aeronautics and Astronautics, 2007.
[17] R.G. Hills, T.G. Truncano, Statistical validation of engineering and scientific models: a maximum likelihood based metric, SAND2002-1783, Sandia National Laboratories, Albuquerque, NM 2002.

[18] R.G. Hills, I. Leslie, Statistical validation of engineering and scientific models: validation experiments to application, SAND2003-0706, Sandia National Laboratories, Albuquerque, NM 2003.
[19] B.M. Rutherford, K.J. Dowding, An approach to model validation and model-based prediction – polyurethane foam case study, SAND2003-2336, Sandia National Laboratories, Albuquerque, NM 2003.
[20] W. Chen, L. Baghdasaryan, T. Buranathiti, J. Cao, Model validation via uncertainty propagation, AIAA J. 42 (2004) 1406–1415.
[21] 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, SAND2004-3632, Sandia National Laboratories, Albuquerque, NM 2004.
[22] K.M. Hansen, A framework for assessing uncertainties in simulation predictions, Physica D 133 (1999) 179–188.
[23] M. Kennedy, A. O'Hagan, Bayesian calibration of computer models (with discussion), J. Royal Statist. Soc. Series B 63 (2001) 425–464.
[24] G. Hazelrigg, Thoughts on model validation for engineering design. in: Proceedings of ASME 2003 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Chicago, 2003.
[25] R. Zhang, S. Mahadevan, Bayesian methodology for reliability model acceptance, Reliab. Engrg. System Safety 80 (2003) 95–103.
[26] A. O'Hagan, Bayesian analysis of computer code outputs: a tutorial, Reliab. Engrg. System Safety 91(2006) 1290–1300, Original technical report available at http://www.tonyohagan.co.uk/academic/pdf/BACCO-tutorial.pdf.
[27] W. Chen, Y. Xiong, K.-L. Tsui, S. Wang, Some metrics and a Bayesian procedure for validating predictive models in engineering design, in: Proceedings of ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Philadelphia, American Society of Mechanical Engineers, 2006, <http://ideal.mech.northwetern.edu/pdf/DAC_validation06.pdf>.
[28] W. Chen, Y. Xiong, K.-L. Tsui, S. Wang, A design-driven validation approach using Bayesian prediction models, J. Mech. Des. 130 (2) (2008).
[29] D. Draper, Assessment and propagation of model uncertainty, J. Royal Statist. Soc. Series B 57 (1995) 45–97.
[30] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York, 1985.
[31] W.L. Oberkampf, T.G. Trucano, Verification and validation in computational fluid dynamics, Progr. Aerospace Sci. 38 (2002) 209–272.
[32] W.L. Oberkampf, M.F. Barone, Measures of agreement between computation and experiment: validation metrics, J. Comput. Phys. 217 (2006) 5–36.
[33] A. Kolmogoroff, Confidence limits for an unknown distribution function, Ann. Math. Statist. 12 (1941) 461–463.
[34] N. Smirnoff, 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, 1939, (fasc. 2).
[35] W. Feller, On the Kolmogorov–Smirnov limit theorems for empirical distributions, Ann. Math. Statist. 19 (1948) 177–189.
[36] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.
[37] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
[38] M. Fréchet, Sur quelques points du calcul fonctionnel (Thèse), Rendiconti Circolo Mathematico di Palermo 22 (1906) 1–74.
[39] J.E. Angus, The probability integral transform and related results, SIAM Rev. 36 (4) (1994) 652–654.
[40] M.A. Stephens, EDF statistics for goodness of fit and some comparisons, J. Am. Statist. Associat. 69 (1974) 730–737.
[41] G.D. Rayner, J.C.W. Rayner, Power of the Neyman smooth tests for the uniform distribution, J. Appl. Math. Decision Sci. 5 (3) (2001) 81–191.
[42] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, 1993.