Defining coverage of an operational domain using a modified nearest-neighbor metric

Sez Atamturktur a,*, Matthew C. Egeberg a, François M. Hemez b, Garrison N. Stevens a

a Clemson University, Clemson, SC 29634, USA
b Los Alamos National Laboratory, Los Alamos, NM 87545, USA

1. Introduction

Numerical models are executed to predict within a range of settings known as the operational domain. The inability of the model to match observations within this domain can be represented by an empirically trained error model, known as discrepancy bias [1,2]. The discrepancy can be used to evaluate the predictive maturity of a model [3] and to bias correct the model predictions [1,2,4–6]. As the discrepancy bias is empirically trained from the available validation experiments, limiting validation experiments only to a region of the domain can result in a poor training of the discrepancy bias (Fig. 1), which in turn, can result in overconfidence in model predictions [7]. This potential oversight is illustrated in Fig. 1 by the
The concept of coverage has recently been included in Predictive Maturity Index (PMI), a metric developed to quantify the predictive capability of a numerical model [3]. Predictive capability, which is concerned with the ability of predictions to reproduce experimental measurements, inherently requires consideration of the coverage of validation experiments. Thus, coverage is treated as one of four components in the PMI and has a major role in quantifying the predictive maturity of a model. With such importance placed upon coverage, it is critical for coverage to be determined using the most refined definition available. In Section 2 of this paper, we identify four essential criteria for a satisfactory coverage metric.

Several coverage metrics are discussed in the literature (see for instance, Atamturktur et al. [8], Hemez et al. [3], and Stull et al. [9]) all of which have drawbacks. The metric developed in Atamturktur et al. [8] supplies a counterintuitive value and does not discern between interpolative and extrapolative regions of the domain. The metric from Hemez et al. [3] does not account for validation experiments added within the bounds of existing validation experiments, and the coverage metric in Stull et al. [9] is subjective, possibly leading to different conclusions between experts. In Section 3, we overview these three abovementioned coverage metrics from the literature and investigate the ability of each metric to meet the identified criteria from Section 2.

In Section 4 of this paper, a new coverage metric is proposed that alleviates the drawbacks of the existing metrics and satisfies all identified criteria. Section 5 demonstrates the applications of the proposed metric on a non-trivial problem of polycrystal plasticity and compares it to existing coverage metrics. In Section 6, the effect of dimensionality on the proposed coverage metrics is investigated focusing on a high-dimensional Rosenbrock function. Section 7 concludes the paper suggesting alternative uses of the proposed coverage metric.

2. Characteristics of exemplar coverage definition

Four criteria can be identified as essential characteristics for any coverage metric:

1. Coverage should improve if a new validation experiment is conducted at new, untested settings within the domain;
2. Poorer coverage should result from a clustered arrangement of validation experiments that limits exploration to certain regions of the domain, than an equal number of validation experiments spread more evenly throughout the domain;
3. Coverage should distinguish between interpolation and extrapolation, due to the lack of finite bounds for extrapolation;
4. Coverage should be objective, not subjective.

The first criterion is based on the postulation that conducting new validation experiments at untested settings provides additional information for model validation, leading to a greater predictive maturity. If a validation experiment has already been conducted at that setting, then a repeated validation experiment should provide no additional coverage.¹

The second criterion is focused on even distribution of validation experiments as suggested by distance-based experimental designs [10]. Design strategies that spread points throughout the domain, particularly in input dimensions that have significant influence on the output of the model, result in lower average prediction errors [11]. This is due to the fact that space-filling designs focus on global approximation of the model [12]. The second criterion therefore attempts to incorporate the benefits of space-filling designs into the coverage metric, which are favorable in the presence of systematic error [13].

The third criterion is motivated by the assertion from experts that empirical models should not be used outside the range of calibration experiments [14–17]. As Montgomery [17] warns, it is possible for a model to provide poor predictions outside the region of the available data even though the model may fit the observations well. Such objections to extrapolation are primarily due to the lack of clear bounds for extrapolation, which are well defined for an interpolative problem.²

¹ While conducting experiments at previously sampled settings may provide information about the experimental variability, coverage metrics neglect the benefits of replication.
² In the context of this work, “interpolation” refers to all predictions made within the region of validation experiments defined by a convex hull, and “extrapolation” refers only to predictions made outside the corresponding range of validation experiments. Under this definition, it is assumed that the
Experimental design strategies that concentrate runs near the boundaries of the domain, similar to an entropy-based experimental design as described in [13], tend to reduce the maximum prediction error [11]. Thus, the third criterion encourages experiments to be located near the boundaries of the domain, which is particularly favorable in the presence of random error [13].

The fourth criterion is straightforward; a coverage based upon hard evidence should be more credible and reliable than one based on an individual’s opinion. A metric that heavily relies on expert opinion may lead to different conclusions between different experts, whereas an objective metric is consistent and repeatable. Implementing methods that probabilistically quantify an expert’s opinion, such as those discussed in [18] however may alleviate the inconsistencies one might face due to subjectivity.

3. Earlier definitions of coverage

This section reviews and compares coverage metrics defined earlier in the published literature. Herein, the suitability of a coverage metric is measured by the ability to satisfy the four aforementioned criteria.

3.1. Atamturktur et al. [8]

Coverage is determined in Atamturktur et al. [8] using a sensitivity adjusted nearest-neighbor metric. By definition of this metric, control parameter ranges that define the operational domain are first normalized between 0 and 1. Next, each control parameter dimension is scaled according to the sensitivity of the model output to that particular control parameter, where a greater sensitivity causes the control parameter dimension to dilate placing focus on more sensitive model inputs. To approximate the sensitivity of each control parameter, Atamturktur et al. [8] exploits the correlation length of the Gaussian Process Model (GPM) emulators trained to replace the computationally demanding physics models.

The scaled hyper-dimensional domain is covered by a sufficient number of uniformly distributed grid points and each grid point is appointed to the nearest validation experiment. Fig. 2a shows the partitioning of the domain into nearest-neighbor regions. The distance between each grid point and associated nearest validation experiment is summed for all grid points and normalized by the total number of grid points, as shown in Eq. (1):

$$nnm = \frac{1}{g} \sum_{i=1}^{g} \min \left( d_{E_i} \right)$$

(1)

where parameter $nnm$ represents the nearest-neighbor metric value, $g$ represents the total number of grid points, and $\min \left( d_{E_i} \right)$ is the minimum distance of the $i$th grid point to the nearest validation experiment calculated within a sensitivity scaled multidimensional domain. The result is a value that represents the average normalized and sensitivity-scaled distance between each point in the domain to the corresponding nearest validation experiment. Decreasing this value improves the coverage.

The nearest-neighbor metric of Atamturktur et al. [8] is sensitive to the addition of a new validation experiment as well as the clustering of validation experiments. The metric is also objective. However, the metric is incapable of showing preference to interpolation over extrapolation. As shown in Fig. 2b, both grid point $A$ and $B$ are an equal distance from the nearest neighboring validation experiment and thus, are treated similarly by the nearest neighbor metric even though point $A$ involves an interpolative prediction, while point $B$ involves extrapolative prediction. Furthermore, the nearest-neighbor metric supplies a counter-intuitive value where improvement in coverage is represented by a decreasing value, whereas the coverage defined using the methods presented in Hemez et al. [3] and Stull et al. [9], as discussed in the following sections, supply intuitive indicators of coverage.

3.2. Hemez et al. [3]

In Hemez et al. [3], coverage is quantified in two steps. First, the convex hull, or multidimensional domain with the smallest convex volume, of the validation experiments is defined. Next, the ratio between the volume of the convex hull and the volume of the operational domain is calculated. The metric can be calculated according to Eq. (2):

$$\eta_c = \frac{V(\Omega_{CH})}{V(\Omega_t)}$$

(2)

where $\eta_c$ represents the coverage and $V(\cdot)$ is a function that calculates the volume of a multidimensional domain. $\Omega_{CH}$ is the convex hull that surrounds the validation experiments while $\Omega_t$ denotes the operational domain. The metric proposed by Hemez et al. [3] has a profound ability to show the distinction between interpolation and extrapolation. Moreover, the

Footnote continued

mechanics or physics principles do not change within the region of validation experiments relative to those captured by the model. If the phenomenology changes “between” these validation experiments, then one can no longer distinguish an interpolative prediction of the model from an extrapolative prediction.

3 Sufficiency of the number of grid points will be discussed later in Section 4.
metric is objective. This metric however, is controlled by the positioning of the experiments at the boundaries of the domain, where the addition of experiments within the convex hull fails to reflect improvement in the coverage, as shown with experiment A in Fig. 3. Hemez et al. [19] suggests that better definitions of coverage could be developed and applied to the PMI to account for the number and overall spread of validation experiments.

3.3. Stull et al. [9]

The metric defined in Stull et al. [9] creates a convex hull around each individual validation experiment rather than a single convex hull containing every experiment. The coverage is then defined as the ratio of the summation of the convex hulls surrounding the validation experiments to the convex hull defining the domain, which can be expressed as:

$$\eta_c = \sum_{k=1}^{N} (-1)^{k+1} \left( \sum_{1 \leq i_1 < \ldots < i_k \leq N} |V(\Omega_{E_{i_1}}) \cap \ldots \cap V(\Omega_{E_{i_k}})| \right) / V(\Omega_V)$$

where $\Omega_{E_i}$ is the convex hull surrounding the $k$th validation experiment and $N$ is the total number of validation experiments. Note that in Eq. (3), the intersections of convex hulls that double count the coverage are subtracted from the total volume according to the well-known principle of inclusion and exclusion [20]. Therefore, if the convex hulls from more than one validation experiment overlap, the area is only counted once.

The metric proposed by Stull et al. [9] is subjective as the size of the convex hull surrounding each validation experiment is based on expert opinion. Furthermore, with this metric, a validation experiment could be added without improving the

---

4 A more objective criterion could also be used, where the size of each convex hull surrounding a validation experiment is based on a gradient-based sensitivity analysis. The size of the convex hull can be inversely proportional to the gradient of the model predictions around that particular experiment.
coverage, provided that the existing convex hulls completely engulf the convex hull of an additional validation experiment, as shown in Fig. 4 with experiment A.

Stull et al. [9]'s metric neither recognizes large unexplored regions in the domain nor differentiates between interpolation and extrapolation as the validation experiments could be clustered in one region of the domain and achieve the same coverage as a more distributed arrangement provided that there is no overlap of the convex hulls, as shown in Figs. 5 and 6.

The metric proposed by Stull et al. [9] should be given more credit than Figs. 4–6 suggest, which present carefully-chosen, problematic situations for this metric. Using expert opinion to vary the size of the convex hull associated with each individual validation experiment may alleviate some limitations and provide an improved quantification of coverage. However, doing so forces the metric to rely on expert opinion and increases subjectivity.

The discussion presented in this section is summarized in Table 1. Note that each metric fails at least one criterion but each criterion is passed by at least one metric.

4. Proposed coverage definition

Due to the ability to already pass three of the four criteria, the coverage metric presented in Atamturktur et al. [8] is modified to account for the difference between interpolative and extrapolative predictions by adding an extrapolation penalty based upon the convex hull utilized in Hemez et al. [3]. Additionally, the metric is transformed to provide a more intuitive coverage value, in which a greater value indicates improved coverage of the domain.
4.1. Penalizing extrapolative predictions in the coverage metric

The nearest-neighbor metric is first modified to differentiate between interpolative and extrapolative predictions. A convex hull encompasses the validation experiments as in Hemez et al. [3], dividing zones of interpolation and extrapolation, as shown in Fig. 7. Grid points that lie outside the zone of interpolation are subject to an extrapolation penalty equal to the minimum distance between the grid point and the zone of interpolation. This penalty is added to the distance between the grid point and the nearest validation experiment, as shown in Eq. (4):

\[
nnm = \frac{1}{g} \sum_{i=1}^{g} \min(d_{E,i}) + d_{ZI,i}
\]

where \(d_{ZI,i}\) is the minimum distance between the ith grid point and the zone of interpolation. Distances \(d_E\) and \(d_{ZI}\) are shown in Fig. 7. Applying this extrapolation penalty increases \(nnm\) and thus, reduces coverage. Through this penalty, validation experiments are encouraged to be positioned nearer the boundaries of the domain, reducing the zone of extrapolation.

Table 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pass</td>
<td>Fail</td>
<td>Improved but imperfect</td>
</tr>
<tr>
<td>2</td>
<td>Pass</td>
<td>Fail</td>
<td>Improved but imperfect</td>
</tr>
<tr>
<td>3</td>
<td>Fail</td>
<td>Pass</td>
<td>Fail</td>
</tr>
<tr>
<td>4</td>
<td>Pass</td>
<td>Pass</td>
<td>Fail</td>
</tr>
</tbody>
</table>

Fig. 6. Possible effect of interpolation/extrapolation ratio on coverage metric proposed by Stull et al. [9].

Fig. 7. Example zone of interpolation and extrapolation for a two dimensional domain.

4.2. Transforming the proposed coverage metric into an intuitive indicator

The modified nearest-neighbor value yields a counterintuitive description of coverage that decreases as the number of experiments increases. To provide a more intuitive interpretation of coverage that can be straightforwardly integrated in the PMI, the metric value is transformed, utilizing the upper and lower bounds of the modified nearest-neighbor value.

The lower bound of the nearest-neighbor value ($\text{nnm}_{\text{min}}$) occurs if a validation experiment is located at each grid point, producing a metric value equal to 0. However, for the grid points to sufficiently represent the entire operational domain, there must be more grid points than validation experiments; therefore, as the number of validation experiments increases, the metric value asymptotically approaches 0.

The upper limit of the nearest-neighbor metric value ($\text{nnm}_{\text{max}}$) is achieved using only one validation experiment. For a rectangular domain, defined by the minimum and maximum values (i.e. ranges) of each input parameter, the location for an experiment that yields the worst coverage occurs at a corner of the domain. With a single experiment, the convex hull encompasses zero volume; hence, the extrapolation penalty is equal to the nearest distance between each grid point and validation experiment. The average distance between validation experiment and each grid point is equal to the integration of the distance from the validation experiment over the entire domain, divided by the multidimensional volume of the domain:

$$\text{nnm}_{\text{max}} = \frac{\int_{\Omega} \sum_{i=1}^{n} d_{i} d\Omega}{\int d\Omega}$$

where $\Omega$ represents the multidimensional volume of the domain. For a rectangular domain as the grid is refined, the numerically obtained maximum value converges to the theoretical value from Eq. (5). This is demonstrated in Fig. 8 for a two-dimensional domain with the total number of grid points increasing from four to 40,000. In Fig. 8, the numerical value converges to the theoretical value of 1.5304 as the grid is refined. As expected, the maximum metric value increases with increased dimensionality of the domain (see Fig. 9).

The metric is transformed to range between zero and infinity with zero representing the poorest possible coverage and infinity representing perfect coverage. This transformation is accomplished using the following functional form:

$$\eta_c = \frac{1}{\text{nnm}} - \frac{1}{\text{nnm}_{\text{max}}}$$

Under this definition, when one experiment is located in the worst possible location, the coverage will equal zero. As more experiments are added at new, untested settings, the coverage will increase up to infinity.

Fig. 8. Convergence of maximum metric value to the theoretical value as the number of grid points increases.

Fig. 9. Maximum metric value as a function of dimensionality (for unit sensitivity in each direction).
Table 2
PMI Term Definitions [9].

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_1$</td>
<td>${ \tanh(\gamma_1 \times \eta) } \quad \eta \leq 1$</td>
</tr>
<tr>
<td>$\psi_2$</td>
<td>$\tanh(\frac{\eta}{\gamma} \tau)$</td>
</tr>
<tr>
<td>$\psi_3$</td>
<td>$1 - \tanh(\frac{\eta}{\gamma} \tau)$</td>
</tr>
<tr>
<td>$\psi_4$</td>
<td>${ 1 - e^{\frac{-\epsilon}{\delta}} } \times e^{-e^{-w \times \delta_5 \times e^{-\alpha_5}}}$</td>
</tr>
</tbody>
</table>

4.3. Incorporation of the proposed coverage metric in the predictive maturity index (PMI)

The PMI has been established as a quantitative and objective metric to evaluate predictive capabilities of numerical models and has been applied to for instance the Preston–Tonks–Wallace model of plastic deformation [3], the Viscoplastic Self-Consistent (VPSC) material model [21], and the nuclear fuel performance code, LIFEIV [22]. Recently modified by Stull et al. [9], the PMI includes four attributes: coverage of the domain, $\eta_c$, robustness to model parameter uncertainty, $\alpha_5$, scaled discrepancy bias, $\delta_5$, and model complexity, $N_K$ as described in Eq. (7):

$$PMI(\eta_c; N_K; \delta_5; \alpha_5) = \prod_{i=1}^{5} \psi_i$$  \hspace{1cm} (7)

where $\psi_i$ are shown in Table 2, with positive, user-defined coefficients $\gamma_1$, $\gamma_2$, $\gamma_3$, and $\gamma_4$. The purpose of the gamma values is to weigh the effect of each attribute on the PMI. Note $N_K$ in Table 2 represents a reference number of knobs, or uncertain model parameters. As each attribute is bounded between 0 and 1, the PMI is naturally bounded between 0 and 1. The exponential or hyperbolic tangent functions in Table 2 are used to provide asymptotic limits between these bounds.

The functional terms shown in Table 2 are designed around the coverage definition in Stull et al. [9] in which coverage is allowed to vary between 0 and infinity. This range is equal to the range for the proposed coverage metric and allows incorporation of the proposed coverage metric into the PMI in a straightforward manner.\(^5\)

5. Demonstrating the use of coverage metric on a practical application

The proposed coverage metric is applied to quantify the coverage of the domain achieved by synthetic experiments selected through Batch Sequential Design. These synthetic experiments are used to calibrate the Viscoplastic Self-Consistent (VPSC) code for modeling stress–strain response and textural evolution of 5182 aluminum alloy.

5.1. VPSC material model

The VPSC code developed in [23] predicts plastic deformations considering both climb and glide dislocation at the single-crystal level. The governing equation is written as [23]:

$$\frac{d\varepsilon}{dt} = \gamma_0 \sum_{i=1}^{N_c} \left\{ m^i \left( \frac{m^i \cdot \sigma}{\tau_0^i} \right)^{\eta_0} \times \text{sgn}(m^i \cdot \sigma) + c^i \left( \frac{c^i \cdot \sigma}{\sigma_0^i} \right)^{\eta_0} \times \text{sgn}(c^i \cdot \sigma) \right\}$$  \hspace{1cm} (8)

where $d\varepsilon/dt$ denotes the strain rate, and $\sigma$ represents the stress applied to the crystal. The terms $c^i$ and $\sigma_0^i$ are the climb tensor and critical stress associated with climb, respectively. Similarly, $m^i$ and $\tau_0^i$ are the Schmid tensor and critical resolved shear stress associated with glide. In Eq. (8), $\eta_0$ is the glide stress exponent and $\eta_0$ is the climb stress exponent. The single crystal equation is summed over all active slip systems, $N_c$. Finally, $\gamma_0$ is a normalization factor [23]. A large number of parameters are required to completely describe the crystallographic textures using weights associated with a partition of 3-D orientation space [21]. However, for calibration and validation purposes, the final textures can be characterized by two components: (i) intensity associated with a retained (001) cube texture and (ii) intensity associated with a (101) compression texture. The 001 and 101 poles represent corners of the inverse pole figure [21].

The VPSC model has two control parameters (temperature and strain rate) and three outputs (stress–strain response, pole 001 texture, and pole 101 texture) that define the operational domain. Two calibration parameters ($\tau_0^i$ and $\sigma_0^i$) are found to exhibit a dependency on both temperature and strain rate and therefore are each replaced by four parameters that describe the functional relationship as explained in [21]. As a result, the VPSC model possesses 10 total calibration parameters. In [21], the 10 calibration parameters are calibrated against physical validation experiments measuring stress at

\(^5\) Under the Stull et al. [9] definition, a coverage value greater than 1 indicates that the coverage exceeds the dimensions of the domain. Hence, the $\psi_1$ term is equal to 1 for all coverage values equal to or greater than 1. In the proposed coverage metric, a coverage value equal to 1 does not represent perfect coverage of the domain. Therefore, the condition that $\psi_1$ equals 1 if the coverage equals or exceeds 1 is removed.
a strain equal to 0.6, textural intensity of the 001 pole, and textural intensity of the 101 pole. In [24], these calibrated values are considered to be “true” values to allow for a simulated Batch Sequential Design (BSD) study as discussed in the following section.

5.2. Selection of experimental settings through batch sequential design (BSD) for VPSC model

In BSD, information from available experiments is used to select the optimum (according to a predefined criterion) settings of future experiments sequentially in batches of user-selected sizes [11]. Numerous different criteria are available to be used in the optimization process of the BSD approach [25]. In application to the VPSC model, BSD is deployed to determine the optimal locations of experiments through use of the Euclidean distance EDIST criterion [24]. EDIST is a distance-based criterion that selects design settings that minimize the maximum correlation between discrepancy values of the proposed design and existing design.

The initial experimental settings (batch 0) as well as the BSD selected settings (batches 1–10) are shown in Fig. 10. These experiments are simulated by running the VPSC code using the settings of control parameters (temperature and strain-rate) selected by BSD and the so-called exact values of the calibration parameters determined in [21]. With the addition of new experimental data, the model is recalibrated and the process is repeated until completion of the tenth batch. During model calibration, VPSC code is replaced with a fast-running Gaussian Process Model (GPM) emulator [2, 5] that is trained to approximate the input and output relationship of the VPSC code. When using an emulator, it is necessary to validate the accuracy of the trained emulator predictions for settings other than the training set (typically referred to as hold-out simulations) [26]. Fig. 11 demonstrates that accuracy of the GPM emulator used in our study through cross-validation for stress predictions at a strain equal to 0.6, where the holdout simulations and the GPM predictions show satisfactory agreement.

5.3. Coverage obtained through batch sequential design (BSD) selected experiments

In our application, the operational domain is defined by temperatures between 200 and 550 °C and strain-rates between 0.001 and 1 s⁻¹ [24]. The VPSC code predicts stress–strain response, texture 001 evolution and texture 101 evolution, and thus the metric value is determined for each output separately and then averaged [21]. In this application, the sensitivity of each control parameter to each of the three model outputs is determined by the spatial dependence parameter, β of the GPM emulator for each output separately. The β parameter describes the dependence of the output on each particular input;

![Fig. 10.](image1.png)

**Fig. 10.** Experimental settings selected through BSD (marker number denotes batch number).

![Fig. 11.](image2.png)

**Fig. 11.** Cross validation of the Gaussian process model used to substitute for simulations.
therefore, a control parameter with greater influence on the output yields a larger $\beta$ value than a control parameter with less influence. In this application, this sensitivity is determined after the tenth batch. Using the proposed metric, the coverage obtained considering all three outputs of the VPSC code is shown in Fig. 12 for each batch.

The coverage is also evaluated using the metrics presented in Hemez et al. [3] and Stull et al. [9]. The metric from Atamturktur et al. [8] is omitted, as the proposed metric is a close revision. The results using the Hemez et al. [3] metric are presented in Fig. 13. Between second and third batches, as well as between fifth and ninth batches, the coverage is not affected by the addition of new validation experiments. This is because the validation experiments added in those batches are located inside the existing convex hull (recall Fig. 10). In contrast, the coverage metric proposed herein yields improvement of coverage between every batch (Fig. 12), recognizing the experiments located inside the convex hull.

Several alternative coverage values can be obtained using the Stull et al. [9] metric depending on the bound chosen by the expert (Fig. 14). Accordingly, the convergent properties of the coverage may change. For example, assuming each
experiment covers 45% of the domain in each dimension causes the coverage to reach a value of 1 (perfect coverage) after the sixth batch. However, if 25% bounds are used, the gain in coverage is nearly linear from the first batch to the tenth and a final coverage equal to 0.7959 is achieved after the tenth batch. In contrast, the coverage metric proposed herein is objective and thus unsusceptible to the potential variability between the opinions of two experts.

6. Curse of dimensionality

In this section, the effect of dimensionality on the proposed metric is investigated and compared to existing metrics using purely mathematical functions.

6.1. Effect of dimensionality

As a constant number of experiments are used to cover a domain of increasing dimensionality, the coverage is expected to decrease as the density of experiments decreases. To investigate this phenomenon known as curse of dimensionality, domains ranging between two and ten dimensions are populated by 100 experiments selected using Latin Hypercube Sampling (LHS). The Stull et al. [9] metric is evaluated assuming 25% bounds around each experiment and the proposed coverage metric is evaluated using four grid points for each dimension to keep computational time reasonable at high dimensions. Sensitivity values of one are assumed for all dimensions. The analysis is repeated 50 times and coverage is computed for each using the proposed coverage metric as well as the Hemez et al. [3] and Stull et al. [9] metrics (Fig. 15).

In Fig. 15(a), using the Hemez et al. [3] metric, the coverage quickly decreases. When eight dimensions are analyzed, only 2.5% of the domain is covered, and when 10 dimensions are being analyzed, only 0.3% of the domain is covered. Therefore, for a domain with 100 dimensions, the coverage achieved using the Hemez et al. [3] metric is nearly negligible. Similarly, the Stull et al. [9] metric displays a steep decline in coverage as the dimensionality increases and quickly becomes computationally prohibitive (Fig. 15(b)). As with the Hemez et al. [3] and Stull et al. [9] metrics, the coverage metric proposed herein also suffers from the curse of dimensionality, displaying a decreasing value as the number of dimensions increases (Fig. 15(c)). The rate of decrease in coverage is largest when the number of dimensions is low and the calculated coverage exhibits a slower rate of decrease as the dimensionality increases.

In short, Fig. 15 demonstrates that a greater number of experiments are required in a higher dimensional domain to achieve the same coverage as a lower dimensional domain, as expected. It must be noted that the Hemez et al. [3] metric requires a greater number of experiments than the number of dimensions to evaluate the metric. Therefore, it would not be possible to evaluate the Hemez et al. [3] metric if there were 10 or fewer experiments. On the other hand, the proposed coverage metric may be evaluated using as few as one experiment. Furthermore, the Stull et al. [9] metric poses computational constraints at higher dimensions. Therefore, for a high-dimensional domain with an equal or fewer number of experiments, the proposed coverage metric remains to be the only choice to evaluate the coverage.

6.2. Demonstrating the coverage of high-dimensional domain on the Rosenbrock function

The performance of the proposed coverage metric for a higher dimensional domain (i.e. 10 dimensional domain) is studied using the Rosenbrock function:

\[ Y = \sum_{k=1}^{N-1} (1 + X_k)^2 + C_{k+1}(X_{k+1} - X_k^2)^2 \]  

(9)

Fig. 15. Average coverage (solid line) ± 3 standard deviations (dashed lines) obtained with 50 simulations of a 100 experiment LHS design.

Data can only be collected for a large sample size for up to four dimensions due to the high computational cost at higher dimensions. Using an Intel Core 2 Quad CPU (Q9400) at 2.66 GHz with 4.00 GB memory, results for a single LHS design are obtained in 0.36, 1.38, and 88.4 seconds for two, three, and four dimensions, respectively. Results for five dimensions cannot be obtained in under one hour. For eight or more dimensions, the proposed coverage metric is evaluated assuming that the entirety of the domain is an extrapolative regime.
In Eq. (9), \( N \) represents the number of dimensions while \( C_k \) are user defined coefficients to weigh the effect of each input. Predictions generated by a two-level, full-factorial \( (2^{10} = 1024 \text{ runs}) \) design of experiments are analyzed with an analysis-of-variance (ANOVA) to determine the statistical significance of each input. A larger \( R^2 \) value indicates a parameter that exhibits greater influence. As such, the main effect \( R^2 \) value is scaled as a percentage and used as the sensitivity scaling factor for each dimension. Values of \( C_k \) and results from the ANOVA are given in Table 3.

Experimental data are generated from a LHS design. The proposed coverage metric is evaluated under the assumption that all grid points are penalized for extrapolation. The coverage calculations, repeated 50 times for different LHS designs, are shown in Fig. 16. The coverage metric proposed herein displays diminishing returns as the number of experiments increases. The increase in coverage from 1 experiment to 3 experiments is greater than the gain realized from increasing the number of experiments from 75 to 200. In other words, two experiments when the coverage is poor are more valuable to improving the coverage than 125 experiments after 75 experiments have already been conducted. Analysts may use a plot similar to Fig. 16 to help determine when the gains in coverage do not justify the cost of further experiments, thus when the experimental campaign should be terminated. The proposed coverage metric may be most useful for high-dimensional applications where the Hemez et al. [3] and Stull et al. [9] metrics experience limitations either in the form of high computational cost or the inability to evaluate the metric with fewer experiments than dimensions.

### 7. Conclusion

A quantitative metric is defined to assess the coverage provided by a set of validation experiments within an operational domain. The proposed coverage metric is designed around four criteria: (i) coverage should improve if a new validation experiment is conducted at new, untested settings within the domain, (ii) poorer coverage should result from a clustered arrangement of validation experiments that limits exploration to certain regions of the domain, than an equal number of validation experiments spread more evenly throughout the domain, (iii) coverage should distinguish between interpolation and extrapolation, and (iv) coverage should be objective, not subjective. This paper modifies the sensitivity adjusted nearest neighbor metric developed in Atamturktur et al. [8] to encourage experimental designs with validation experiments nearer the boundaries of the domain, thus reducing extrapolation. The authors also propose a transformation of the proposed coverage metric which allows the metric to be implemented in the Predictive Maturity Index (PMI). The proposed coverage metric is demonstrated on the multivariate Viscoplastic Self-Consistent code as well as a high-dimensional variant of the Rosenbrock function.

### Table 3

Coefficients of the Rosenbrock function and statistics for main-effect analysis.

<table>
<thead>
<tr>
<th>Variable ((X_k))</th>
<th>Coefficient ((C_k))</th>
<th>(R^2) Statistic (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>22.5</td>
</tr>
<tr>
<td>2</td>
<td>5.0</td>
<td>3.6</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>11.2</td>
</tr>
<tr>
<td>4</td>
<td>3.0</td>
<td>10.8</td>
</tr>
<tr>
<td>5</td>
<td>4.0</td>
<td>10.4</td>
</tr>
<tr>
<td>6</td>
<td>5.0</td>
<td>10.0</td>
</tr>
<tr>
<td>7</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>8</td>
<td>5.0</td>
<td>4.9</td>
</tr>
<tr>
<td>9</td>
<td>2.0</td>
<td>15.4</td>
</tr>
<tr>
<td>10</td>
<td>6.0</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Fig. 16. Average coverage (solid line) ± 3 standard deviations (dashed lines).
The usefulness of the proposed coverage metric extends beyond implementation in the PMI. The metric can be used to directly compare multiple designs of experiments. Furthermore, the metric could be implemented as a Batch Sequential Design selection criterion to select the future settings of validation experiments. As a distance-based criterion, the metric could be combined with an index-based criterion to create a selection condition, similar to the Coverage Augmented Expected Improvement for Predictive Stability (C-EIPS) criterion developed in [24], which simultaneously explores the entire domain and exploits regions with high variance in the discrepancy bias.

Acknowledgments

This research is being performed using funding received from the Department of Energy Office of Nuclear Energy's Nuclear Energy University Programs (Contract no. 00101999). The authors would like to thank Ricardo Lebensohn and Carlos Tomé of Los Alamos National Laboratory for sharing the VPSC code. The authors would like to thank Godfrey Kimball for his editorial review of this paper.

References