8-2014

Optimal Design of Validation Experiments for Calibration and Validation of Complex Numerical Models

Matthew Egeberg
Clemson University, megeber@clemson.edu

OPTIMAL DESIGN OF VALIDATION EXPERIMENTS FOR CALIBRATION AND VALIDATION OF COMPLEX NUMERICAL MODELS

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Civil Engineering

---

by
Matthew C. Egeberg
August 2014

Accepted by:
Dr. Sez Atamturktur, Committee Chair
Dr. Hsein Juang
Dr. Abdul Khan

ABSTRACT

As prediction of the performance and behavior of complex engineering systems shifts from a primarily empirical-based approach to the use of complex physics-based numerical models, the role of experimentation is evolving to calibrate, validate, and quantify uncertainty of the numerical models. Oftentimes, these experiments are expensive, placing importance on selecting experimental settings to efficiently calibrate the numerical model with a limited number of experiments. The aim of this thesis is to reduce the experimental resources required to reach predictive maturity in complex numerical models by (i) aiding experimenters in determining the optimal settings for experiments, and (ii) aiding the model developers in assessing the predictive maturity of numerical models through a new, more refined coverage metric.

Numerical model predictions entail uncertainties, primarily caused by imprecisely known input parameter values and biases, primarily caused by simplifications and idealizations in the model. Hence, calibration of numerical models involves not only updating of parameter values but also inferring the discrepancy bias, or empirically trained error model. Training of this error model throughout the domain of applicability becomes possible when experiments conducted at varying settings are available. Of course, for the trained discrepancy bias to be meaningful and a numerical model to be predictively mature, the validation experiments must sufficiently cover the operational domain. Otherwise, poor training of the discrepancy bias and overconfidence in model predictions may result. Thus, *coverage* metrics are used to quantify the ability of a set of validation experiments to represent an entire operation domain.

This thesis is composed of two peer-reviewed journal articles. The first article focuses on the optimal design of validation experiments. The ability to improve the predictive maturity of a plasticity material model is assessed for several index-based and distance-based batch sequential design selection criteria through a detailed analysis of discrepancy bias and coverage. Furthermore, the effect of experimental uncertainty, complexity of discrepancy bias, and initial experimental settings on the performance of each criterion is evaluated. Lastly, a technique that integrates index-based and distance-based selection criteria to both exploit the available knowledge regarding the discrepancy bias and explore the operational domain is evaluated. This article is published in Structural and Multidisciplinary Optimization in 2013.

The second article is focused on developing a coverage metric. Four characteristics of an exemplar coverage metric are identified and the ability of coverage metrics from the literature to satisfy the four criteria is evaluated. No existing coverage metric is determined to satisfy all four criteria. As a solution, a new coverage metric is proposed which exhibits satisfactory performance in all four criteria. The performance of the proposed coverage metric is compared to the existing coverage metrics using an application to the plasticity material model as well as a high-dimensional Rosenbrock function. This article is published in Mechanical Systems and Signal Processing in 2014.

DEDICATION

I would like to dedicate this thesis to my parents Gary and JoAnne Egeberg.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Sez Atamturktur for her inspiration, encouragement, and constant dedication throughout my degree. I would also like to thank my committee members Dr. Khan and Dr. Juang for their continuous support and guidance.

In addition, I would also like to thank my collaborators, Brian Williams, Cetin Unal, and François Hemez of Los Alamos National Laboratory and Garrison Stevens of Clemson University. Thank you to Ricardo Lebensohn and Carlos Tome of Los Alamos National Laboratory for sharing the VPSC code as well as Eddie Duffy of Clemson University for his technical support in use of the Palmetto Cluster. Also, thanks to Murat Hamutcuoglu, a former post-doctoral fellow and RJ Cadotte, an undergraduate student of Clemson University for their assistance in the preparation of the manuscript as well as Godfrey Kimball for his editorial review.

This work is funded in part by the Verification and Uncertainty Quantification (VU) program element of the Nuclear Energy Advanced Modeling and Simulation (NEAMS) program at Los Alamos National Laboratory (LANL): subcontract Number 84093-RFP-10. This research is being performed in part using funding received from the Department of Energy Office of Nuclear Energy's Nuclear Energy University Programs (Contract Number: 00101999).

TABLE OF CONTENTS

Table of Contents (Continued)

Page

LIST OF TABLES

LIST OF FIGURES

List of Figures (Continued)

Figure Page

List of Figures (Continued)

Figure Page

List of Figures (Continued)

Figure Page

CHAPTER ONE

INTRODUCTION

1.1 Motivation

Design of complex engineering systems is changing from a primarily empirical enterprise to the use of complex physics-based numerical models (Jacobson et al. 2009). As such, physical experiments once relied upon to reveal the relationship between input and output parameters, are now used to calibrate, validate, and quantify uncertainty of complex numerical models. Due to the high consequences associated with the use of these numerical models, predictive maturity of these models becomes of interest (Unal et al. 2011). Several attempts have been made to assess the predictive maturity of numerical models through qualitative, expert judgment ranking systems (Balci et al. 2002, Oberkampf et al. 2007, Green et al. 2008). However, these ranking systems rely on expert judgment and therefore are naturally subjective. Herein, the quantitative and objective Predictive Maturity Index (PMI) metric proposed by Hemez et al. 2010 is used to define predictive maturity.

In this thesis, the interest is on improving the predictive maturity of numerical models through experimental campaigns. Validation experiments are used to calibrate numerical models through comparisons with model predictions (Trucano et al. 2006). Calibration entails inference of the uncertain model parameters as well as the discrepancy bias, or empirically trained error model (Draper 1995, Kennedy & O'Hagan 2001), throughout the operational domain (Hemez et al. 2010). Herein, focus is on design of validation experiments to improve the inference of the discrepancy bias throughout the

operational domain due to the need to bias-correct the numerical models (Atamturktur et al. 2011). Improved inference of the discrepancy bias directly correlates to more accurate model predictions. Meanwhile, experimental campaigns are limited by the cost and time demands of conducting physical experiments (Rosner 2008), placing an importance on efficient experimental designs. Therefore, numerical models must be calibrated with a limited number of validation experiments at finite experimental settings within the operational domain and then used to make predictions at untested settings throughout the domain (Unal et al. 2011, Atamturktur et al. 2011).

As the discrepancy bias is empirically trained, limiting validation experiments only to a region of the operational domain can result in a poorly trained discrepancy bias, and as a result, overconfidence in model predictions (Atamturktur et al. 2011). To mitigate this problem, experiments must sufficiently explore the operational domain. To capture this phenomenon, the PMI metric incorporates the concept of *coverage* (Hemez et al. 2010). Coverage is the ability of a set of validation experiments to represent the entire operational domain. Since coverage is a major component in determining the predictive maturity of a numerical model through the PMI, it is important to properly identify coverage.

1.2 Scope

This thesis, consisting of two peer-reviewed journal articles, aims to reduce the experimental resources required to achieve predictive maturity of complex numerical models through two tasks (i) improving the efficiency of experimental campaigns, and (ii) refining the tools used to determine the predictive maturity of numerical models.

The first article, presented in chapter two and published in Structural and Multidisciplinary Optimization[1], contributes to the first task. Several index-based and distance-based batch sequential design (BSD) selection criteria are applied to the Visco Plastic Self-Consistent (VPSC) material plasticity model in order to assess the performance of various selection criteria from the literature on a nontrivial application. The predictive maturity of the VPSC model is evaluated using PMI (Hemez et al. 2010) to compare the performance of each selection criterion. Furthermore, a detailed analysis of discrepancy bias and coverage reveal the driving factors behind the differences in performance. The importance of discrepancy and coverage are varied to simulate possible real-world situations in which an analyst may place more significance on discrepancy rather than coverage or vice versa. In addition, the study investigates the effect of experimental uncertainty, complexity of the discrepancy bias, and settings of initial experiments on the performance of each selection criterion. This study provides guidance to analysts when determining the best selection criterion to use. Under this guidance, analysts are more likely to use a selection criterion that will achieve desired predictive maturity using fewer experiments when compared to an alternative selection criterion.

The second article, presented in chapter three and published in Mechanical Systems and Signal Processing[2], contributes to the second task. Based on the premise that coverage is the ability of a set of validation experiments to represent the entire

[1] Atamturktur S, Williams B, Egeberg M, and Unal C (2013) Batch Sequential Design of Optimal Experiments for Improved Predictive Maturity in Physics-Based Modeling. Structural and Multidisciplinary Optimization (Springer) 48(3): 549-569

[2] Atamturktur S, Egeberg M, Stevens G, and Hemez F (2014) Defining Coverage of an Operational Domain Using a Modified Nearest-Neighbor Metric. Mechanical Systems and Signal Processing (Elsevier), DOI 10.1016/j.ymssp.2014.05.040

operational domain, four characteristics of an exemplary coverage metric are identified. Coverage should (i) improve if a new experiment is added at untested settings, (ii) favor a more uniform distribution of experiments over a clustered arrangement, (iii) distinguish between interpolation and extrapolation, and (iv) be objective. The inability of any coverage metric from the literature to satisfy all four criteria prompts the proposal of a new coverage metric which satisfies all four criteria. The effectiveness of the proposed metric is demonstrated alongside the existing coverage metrics on the VPSC model as well as the high-dimensional Rosenbrock function. This study helps provide a more precise quantification of predictive maturity by proposing a refined metric for coverage, a crucial component in predictive maturity. A more precise measure of predictive maturity reduces uncertainty of whether or not a model has reached predictive maturity; therefore, resources may be saved on unnecessary experimentation when a numerical model has in fact already reached predictive maturity.

References

Atamturktur S, Hemez F, Williams B, Tome C, Unal C (2011) A forecasting metric for predictive modeling. Computers & Structures 89:2377-2387

Balci O, Adams RJ, Myers DS, Nance RE (2002) Credibility assessment: a collaborative evaluation for credibility assessment of modeling and simulation applications, In Proceedings of the 34th winter simulation conference: exploring new frontiers, San Diego, California, USA, 214-20

Draper D (1995) Assessment and Propagation of Model Uncertainty. Journal of the Royal Statistical Society 57:45–97

Green LL, Blattnig SR, Hemsch MJ, Luckring JM, Tripathi RK (2008) An uncertainty structure matrix for models and simulations. American Institute of Aeronautics and Astronautics AIAA-2008-2154

Hemez F, Atamturktur S, Unal C (2010) Defining predictive maturity for validated numerical simulations. Computers and Structures Journal 88:497-505

Jacobson JJ, Matthern GE, Piet SJ, Shropshire DE (Aril 2009) Vision: Verifiable Fuel Cycle Simulation Model. Advances in Nuclear Fuel Management IV (ANFM). Hilton Head, South Carolina, USA

Kennedy MC, O'Hagan A (2001) Bayesian Calibration of Computer Models. Journal of the Royal Statistical Society 63: 425-464

Oberkampf WL, Pilch M, Trucano TG (2007) Predictive capability maturity model for computational modeling and simulation. Sandia National Laboratories Report; SAND2007-5948

Rosner R (2008) Making nuclear energy work How shifting research goals and improving collaboration with industry will help U.S. national labs spur new nuclear energy development. Bulltin of Atomic Scientists 64(1):28-33

Trucano TG, Swiler LP, Igusa T, Oberkampf WL, Pilch M (2006) Calibration, validation, and sensitivity analysis: What's what. Reliability Engineering & System Safety Journal 91: 1331-1357

Unal C, Williams B, Hemez F, Atamturktur SH, McClure P (2011) Improved best estimate plus uncertainty methodology, including advanced validation concepts, to license evolving nuclear reactors. Nuclear Engineering and Design Journal 241:1813-1833

CHAPTER TWO

BATCH SEQUENTIAL DESIGN OF OPTIMAL EXPERIMENTS FOR IMPROVED PREDICTIVE MATURITY IN PHYSICS-BASED MODELING

2.1 Introduction

Advanced modeling and simulation are increasingly relied upon to predict the performance of new generations of nuclear fuels (Jacobson et al. 2009). When fuel performance predictions are used in support of high consequence decisions, questions naturally arise about the predictive maturity of these models (Unal et al. 2011). An intrinsic component for achieving predictive maturity is model calibration, a methodology used to infer both the uncertain input model parameters and the discrepancy bias of a model (Hemez et al. 2010). Invariably, calibration is achieved through systematic comparisons of model predictions against validation experiments (Trucano et al. 2006).

As models are executed to predict fuel performance at vastly different operational regimes, for calibration of these advanced models to be meaningful, the entire operational domain of the nuclear fuel must be explored through a sufficiently large quantity (and of course, quality) of validation experiments. However, proper exploration of the operational domain is challenged by the cost and time demands of physical experiments which prohibit extensive experimental campaigns (Rosner 2008). The problem is further compounded due to the infeasibility of reproducing extreme operational regimes in a laboratory environment to obtain physical experiments, as in the case of fusion reactors (Yoshiie 2005). As a result, the current trend is shifting towards calibrating numerical

models with a limited number of validation experiments for making predictions with the calibrated models at untested settings (Unal et al. 2011). Therefore, the next natural step for advancement in fuel performance predictions entails reducing the extent of the experimental campaign required to reach the desired predictive maturity in these numerical models. With these new trends and goals, the design and execution of validation experiments must be closely associated with modeling and simulation efforts (Jiang and Mahadevan 2006).

Validation experiments can help improve the predictive ability of a numerical model by (i) mitigating the uncertainty in the model parameters, and (ii) inferring the discrepancy bias throughout the domain (Box and Draper 1959). Regarding mitigating uncertainty, there is extensive literature on experimental designs for emulator training, specifically on various alphabet-optimal designs, such as A-optimality, D-optimality, G-optimality and V-optimality, all of which focus on improving the calibrated values of the input parameters (Evans and Manson 1978, Shao 2007). Of particular interest to the current work however, is the second benefit of validation experiments. In particular, we implement design approach formulated by Williams et al. (2011) for optimal design of experiments that focuses on improving the inference of model discrepancy bias throughout the domain. This focus is justified by the need to bias-correct the numerical models for interpolative or extrapolative purposes (Atamturktur et al. 2011). Herein, the design of validation experiments aims to achieve stability in the inferred discrepancy bias as new validation experiments become available. The desired stability in the inferred

discrepancy can be quantified with various metrics, which will henceforth be referred to as *selection criteria*.

Selection criteria define the targeted benefits of future experiments (analogous to the utility function in Lindley 1972). In this manuscript, we are concerned with evaluating the performance of various selection criteria, including index-based criteria, such as the expected improvement for predictive stability, the expected improvement for global fit, maximum entropy, and distance-based criteria, such as weighted Euclidean distance and Mahalonobis distance. Herein, the performance of these selection criteria is judged strictly from the perspective of *predictive maturity* of the numerical model.

To provide a quantitative and objective evaluation of predictive maturity, we implement the Predictive Maturity Index (PMI) proposed by Hemez et al. (2010). PMI integrates three distinct attributes of model development, experimentation and calibration efforts: discrepancy, coverage and complexity, where design of optimal experiments has a direct influence on two of the three attributes of PMI: coverage and discrepancy. Investigation of this influence for various selection criteria is the focus of this paper.

The problem of optimal design of experiments has been widely studied for the development of fast running emulators that are used in lieu of computationally demanding physics-based models (Dersjö and Olsson 2012; Li, Aute and Azarm 2010). However, until recently methods for designing optimal validation experiments have been lacking. This manuscript aims to contribute to the recent advancements in optimal design of validation experiments, focusing on a practical, non-trivial problem of predicting poly-crystal plasticity (Lebensohn et al. 2010). Visco Plastic Self-Consistent (VPSC) is a

meso-scale code for modeling the creep of core reactor clad and duct components subjected to in-service conditions of irradiation, stress, and thermal cycling. The performance of alternative selection criteria is compared for both exact and inexact versions of the VPSC plasticity model, with a parametric study undertaken not only for the complexity and variance of the model discrepancy but also for the experimental uncertainty.

2.2 Discrepancy Bias

Draper (1995) emphasizes the two aspects of developing a numerical model, η that links known quantities of x to unknown quantities of y, the first involving the physics or engineering principles invoked to establish a link between these two quantities, x and y; and the second involving unknown parameters, t associated with the chosen physics or engineering principles, such that $y(x)= \eta(x,t)$. Here, x represents the control parameter settings defining the domain of applicability[3], within which the model will be executed in predictive capacity. *Model form error* arises due to the inevitable incompleteness of physics or engineering principles in η, which often also leads to missing parameters (Farajpour and Atamturktur 2012). This section demonstrates the role of the model form error and missing parameters in predictive modeling through a proof of concept example.

Herein, we compare the predictions of a numerical model, $\eta(x,t)$ to its corresponding *truth*. The truth function is executed to generate five experiments at randomly selected control parameter settings (indicated by squares in Figure 2.1). The numerical model includes an imprecise parameter, t with a value falling between -1 and

[3] Note that validation experiments must be conducted to explore this domain.

1. In Figure 2.1 (Left), an ensemble of model predictions obtained with sampled values of *t* is compared to the five available physical experiments. This comparison shows that $\eta(x,t)$ fails to reproduce the physical experiments to within observational uncertainty regardless of the parameter value used for *t*. Therefore, the numerical model is incomplete and possibly missing input parameters that are necessary to fully describe the *truth*. Figure 2.1 (Right) gives a quantitative representation of the degree of inaccuracy and incompleteness (herein referred to as model form error) of this hypothetical numerical model if the true values for the uncertain input parameters were known with certainty.

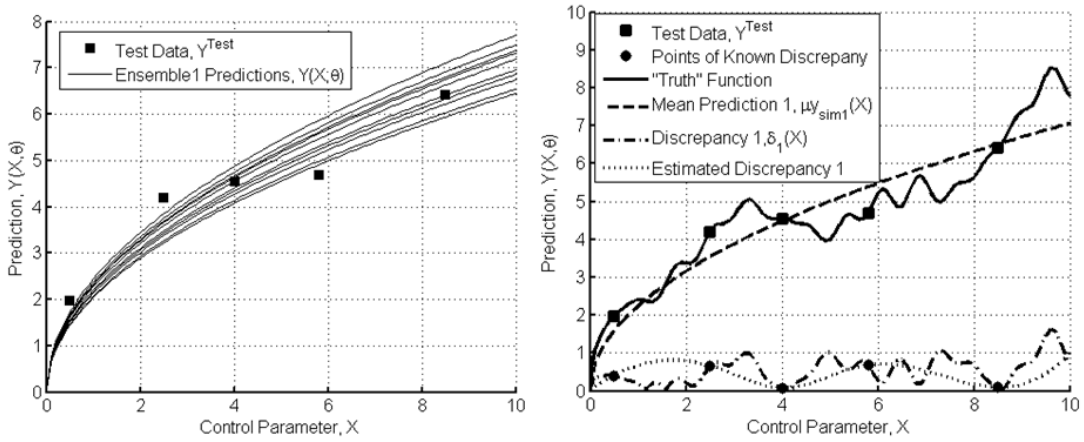

Fig. 2.1 (Left) Comparison of ensemble simulation model predictions against physical experiments, (Right) Model form error representing the degree of incompleteness of a simulation model

Kennedy and O'Hagan (2001) emphasizes that due to the inevitable inexactness of numerical models, calibration of model parameters and estimation of the inherent model form error must be completed simultaneously. Failing to do so can result in parameters being calibrated to mathematically viable but physically incorrect values to

compensate for model form error. Such compensating effects typically lead to over-confidence in the predictive ability of the model. Accordingly, we utilize a Bayesian implementation (Higdon et al. 2008) of the equality originally proposed by Kennedy and O'Hagan (2001), which simultaneously considers parameter uncertainty and model bias. In this formulation, physical observations, $y(x)$, are defined as the summation of truth, $\xi(x)$, and experimental error, $\varepsilon(x)$:

$$y(x) = \xi(x) + \varepsilon(x) \qquad (2.1)$$

If the model form error is known, the truth $\xi(x)$ is defined as the sum of model predictions, $\eta(x, \theta)$ obtained with best fitted input parameter values ($t=\theta$, where θ is the best fitted value for t) and the model form error.

Model form error is unknown however, and thus, must be estimated exploiting the experiments. The empirically trained model form error–henceforth referred to as *discrepancy bias* $\delta(x)$–can be obtained by quantifying the disagreement between experiments and the model predictions with the best fitted input values. The best estimate of truth can then be defined over the entire operational domain in terms of the model predictions $\eta(x, \theta)$, the discrepancy bias $\delta(x)$, and experimental error, $\varepsilon(x)$:

$$y(x) = \eta(x, \theta) + \delta(x) + \varepsilon(x) \qquad (2.2)$$

In a fully Bayesian interpretation of Eq. 2.2, the posteriors for θ and $\delta(x)$ can be explored via Markov Chain Monte Carlo (MCMC) (Metropolis 1953, Higdon et al. 2003). For computationally demanding numerical models, in which MCMC explorations are infeasible, the physics model can be replaced with a fast-running surrogate model (van Keulen and Vervenne 2004, Hemez and Atamturktur 2011). We replace the

numerical model, $\eta(x, \theta)$ with a constant mean Gaussian Process Model (GPM) and the independent error model for discrepancy bias, $\delta(x)$ with a zero mean GPM (see Williams et al. 2006 for further discussion on GPM).

2.3 Batch Sequential Design: Selection Criteria

Myers et al. (1989) advocates the use of sequential approaches in optimal design of experiments. Design augmentation in an iterative, sequential manner uses the information learned from the available set of experiments to improve upon the existing design with future experiments. Moreover, the sequential approach allows the previously existing experiments to be incorporated into the optimal design process (Thompson 2010). Sequential designs can be performed in either a batch sequential or fully sequential (one at a time) manner (Müller and Pötscher 1989). While batch sequential design is sub-optimal compared to fully sequential approach, the practical aspects of conducting physical experiments may suggest the use of the batch-sequential approach with an experimenter-defined batch size (Williams et al. 2011). Herein, we will implement a batch-sequential design (BSD) approach for selecting future optimal experiments according to the stability of the discrepancy bias.

The BSD selects optimum settings for the future validation experiments based upon an existing set of validation experiments, where the optimality condition (i.e., objective function) is defined by a selection criterion. The experiments are selected in the batches of predefined size. BSD continues the selection of batches of experiments until the experimental budget is consumed or a threshold gain in stability of posterior density of discrepancy bias fails to be met.

For optimization of the design criteria, the implemented BSD algorithm uses an exchange algorithm, specifically modified Fedorov exchange method (Fedorov 1972). Exchange algorithms update an experiment in the initial design to improve the desired benefit, as quantified by the selection criteria. Design updates are continued until the relative improvement falls below a given threshold level (Bulutoglu and Ryan 2009, Ogungbenro et al. 2005), which herein is set to 10^{-4}. While the original Fedorov exchange algorithm only performs the 'best' exchange, the modified Fedorov algorithm, implemented herein, executes any beneficial exchange increasing the efficiency of the algorithm (Cook and Nachtsheim 1980).

This section reviews several selection criteria from the literature (see Loeppky et al. 2010 and Williams et al. 2011), while the next section discusses the implementation of these selection criteria for optimal design of experiments for the VPSC code.

2.3.1 Index- based criteria:

Index-based criteria are related to the information content of the design, which is proportional to the inverse of the covariance matrix. Crudely put, an optimal design minimizing the variance maximizes the information content of the experimental design.

Expected Improvement for Predictive Stability (EIPS):

The EIPS criterion evaluates stability of the discrepancy term based upon the expected Kullback-Leibler distance between the current and the proposed future predictive distributions of discrepancy. The maximum expected improvement represents the largest entropy loss between the initial predictive density and the predictive density obtained if proposed experiments at new settings are indeed conducted. Design settings

are chosen to minimize the maximum entropy loss, which approaches smaller values as additional experiments are conducted, resulting in a greater stability in the predictive distribution of discrepancy.

Expected Improvement for Global Fit (EIGF):

In Lam and Notz (2008), the goal of the EIGF algorithm is to obtain one-step sequential additions of simulation runs that efficiently train surrogate models so that predictions at unsampled control parameter settings adequately represent simulation model output. This concept has been further developed to obtain batches of settings for future experiments specifically for application to discrepancy prediction. The criterion chooses the batch of new design settings for future experiments to improve discrepancy prediction by balancing the potential for variance reduction and bias mitigation using information in currently available experimental data.

Maximum Entropy (ENT):

Originally developed to select data that minimizes entropy in predictions at un-sampled settings in a finite system (Shewry and Wynn 1987), the ENT criterion has been extended to accommodate batch updates of existing designs to maximize information in the predictive distribution of discrepancy at untested design settings. In the context of generalized regression modeling, ENT selects new design settings to maximize the determinant of the correlation matrix associated with the distribution for predicting discrepancy at the proposed new design settings conditional on currently available experimental data.

(Integrated- and Maximum) Mean Square Error (MSE) Criteria:

The mean square error (MSE) criteria are used to select design points by minimizing functions of posterior discrepancy variance. The integrated MSE (IMSE) criterion selects a batch of experiments that minimizes the closed form integration of discrepancy variance over the input domain whereas the maximum MSE criterion adds design settings to minimize the maximum discrepancy variance over the input domain (Sacks et al. 1988; Sacks et al. 1989).

2.3.2 Distance- based criteria:

Distance-based criteria view the experimental designs as candidate points spread in the n-dimensional domain of applicability defined by control parameters, x, where n is the number of control parameters of the numerical model, $\eta(x_i, \theta), i=1,\ldots,n$. The objective is then to explore the domain as uniformly and broadly as possible.

Weighted Distance (WDIST) Criteria:

Distance-based approaches have been proposed as batch sequential design criteria to select future experiments to improve discrepancy prediction when calibrating computer models (Johnson et al. 1990; Morris and Mitchell 1995). Two measures of weighted distance are considered: Euclidean (EDIST) and Mahalonobis (MDIST), with weights related to sensitivities of the control parameters. More sensitive control parameters have greater weight and thus, are allowed to be more densely sampled. The sensitivity-weighted distance criterion chooses new design settings that minimize the maximum correlation between predicted discrepancy values on the proposed design and between the proposed and existing designs. That is to say, new design settings are placed in locations

where the ability to borrow strength from available data for discrepancy inference is most limited.

Compared to the index-based criteria, the distance-based criteria are more computationally efficient in that they avoid the slower matrix manipulations required by the index-based criteria. Conversely, distance-based criteria are only indirectly related to the more explicit notions of variance and bias reduction embodied by the index-based criteria.

2.4 Predictive Maturity of Numerical Models

Over the last decade, there have been numerous efforts to assess the predictive maturity of numerical models developed in academic institutions, industry (Balci et al., 2002), National Laboratories (Oberkampf et al., 2007), and NASA (Green et al., 2008). These frameworks seek to assess the overall predictive capability of a numerical model for intended use through qualitative, expert-judgment based ranking systems. In an effort to supply a holistic and quantitative metric for assessing the predictive capability of a simulation model, the PMI metric proposed by Hemez et al. (2010), integrates three distinct aspects of the model development, experimentation and model calibration processes. These aspects are:

- The extent to which experiments cover the domain of applicability; referred to as 'coverage;'
- The fundamental inability of the model to represent the underlying physics; referred to as 'discrepancy bias;'
- The degree of physics sophistication of the model; referred to as 'complexity.'

An obvious advantage of PMI is its quantitative nature, which results in a repeatable and scientifically defendable metric removing the subjective nature of expert opinion from the assessment of simulation model predictability. The basic formulation of the PMI index is expressed as:

$$PMI(c, N_K, \delta_S) = c\left(\frac{N_R}{N_K}\right)^{\gamma_1} \left(1-\delta_s\right)^{\gamma_2} e^{(1-c^2)^{\gamma_3}-\delta_S^{\,2}} \qquad (3)$$

where the parameters (γ_1, γ_2, γ_3) are user-defined weighting coefficients that control the relative impact of coverage, c, scaled discrepancy, δ_S, and complexity, N_K on PMI. Here, γ_1, γ_2 and γ_3 values are taken as 0.5, 0.25 and 2, respectively, to provide uniform weight for coverage and discrepancy as suggested by Hemez et al. (2010). These weight coefficients are kept constant to maintain uniformity between the PMI of various selection criteria.

Coverage is related to the settings of physical experiments performed in the domain of applicability. The adopted strategy to quantitatively measure coverage of the domain is based on the convex hull—that is the smallest convex domain, within which all physical experiments fit. In this study, coverage is calculated as the ratio of the convex hull of the physical experiments to that of the operational domain. As coverage of the operational domain increases, the predictive maturity naturally increases. Discrepancy, δ(x) as introduced earlier, supplies an independent estimate of errors due to either missing or inaccurate numerical modeling. To maintain a standard definition of discrepancy, the estimated discrepancy values are normalized with respect to the mean value of the corresponding simulation predictions ($\eta(x, \theta)$ in Eq.2). A scaled discrepancy,

δ_S is obtained over the entire domain of applicability. Herein, the complexity attribute is constant for all investigated cases, and thus will not influence the PMI calculations.

We envision that BSD selection criteria will elicit various effects on the discrepancy and coverage attributes of PMI. We are particularly interested in classifying the selection criteria introduced in Section 2.2 for their tendency to improve normalized discrepancy versus coverage attributes.

2.5 Visco-Plastic Self Consistent (VPSC) Plasticity Model

Here, VPSC plasticity model is used to predict the creep strain rate in face-centered cubic (FCC) steel (Lebensohn et al. 2010). In VPSC, the plastic deformation mechanism is established considering both climb and glide dislocation at the single-crystal level. VPSC fully accounts for the anisotropic properties and response of the constituent single crystals. For polycrystalline aggregates, VPSC supplies a non-linear homogenization-based polycrystal model while fully accounting for aggregate subjected to external strain-rate or stresses. The fundamental equation dominating the plastic strain rate in a single crystal *r*, deforming by climb and glide is given in Eq.2.4:

$$\dot{\varepsilon}_{ij}^{(r)} = \dot{\gamma}_o \left[\sum_{s=1}^{N_s} m_{ij}^{s(r)} \left(\frac{\left| m^{s(r)} : \sigma^{(r)} \right|}{\tau_o^{gl,s(r)}} \right)^{n^{gl}} \operatorname{sgn}\left(m^{s(r)} : \sigma^{(r)} \right) + \sum_{s=1}^{N_s} k_{ij}^{s(r)} \left(\frac{\left| k^{s(r)} : \sigma^{(r)} \right|}{\tau_o^{cl,s(r)}} \right)^{n^{cl}} \operatorname{sgn}\left(k^{s(r)} : \sigma^{(r)} \right) \right] \tag{2.4}$$

In Eq. 2.4, $\dot{\varepsilon}_{ij}^{(r)}$ denotes the plastic strain-rate induced by climb and glide dislocation, while $\sigma_{ij}^{(r)}$ denotes the stress tensor applied to the crystal *r*. The plastic strain rate is calculated by summing the strain for all active slip systems N_s. The threshold

resolved shear stress for glide is denoted with τ_o^{gl} and the threshold normal stress for climb is denoted with τ_o^{cl} associated with system *s*. Terms n^{gl} and n^{cl} are the rate-sensitivity exponents of glide and climb dislocation, respectively. In Eq. 2.4, m_{ij} is the symmetric glide tensor while k_{ij} is the symmetric climb tensor. The products $m:\sigma$ denote the resolved shear stresses, which must reach the predefined threshold value for slip activation. $\dot{\gamma}_o$ is the normalization factor.

VPSC is used to calculate the strain-rate in the grains and the aggregate for a given stress input. The climb dislocation orientation and deviatoric stress input define the domain of applicability (recall control parameters *x* in Eq. 2.2). Climb dislocation orientation loosely defines the importance of climb phenomena in the crystallographic thermal creep. It varies between $0\text{-}90^0$. Climb dislocation remains inactive for a 0^0 angle, while it is fully activated for a 90^0 angle. Deviatoric stress is the stress input of the specimen to induce creep strain. The upper and lower bounds of the control parameters, which define the operational domain, are given in Table 2.1.

Table 2.1 The range of control parameters also known as domain of applicability

Control Parameters	**Min Value**	**Max Value**
Climb Dislocation Orientation	0.1 rad	0.6 rad
Deviatoric stress input	900 MPa	1100 MPa

In Eq. 2.4, rate-sensitivity exponents of glide and climb dislocation, n^{gl} and n^{cl}, and the ratio of threshold resolved shear stress for glide and the threshold normal stress

for climb, τ_0^{cl}/τ_0^{gl}, are uncertain. These three parameters are calibrated by comparing the VPSC predictions with the experimental measurements (recall calibration parameters θ in Eq. 2.2). A uniform prior distribution is assigned for each calibration parameter between upper and lower bounds determined according to expert judgment (see Table 2.2). Note that the rate-sensitivity exponents are powers in Eq. 2.4, exercising significant influence on the predictions and thus resulting in a very difficult inference problem for discrepancy bias.

Table 2.2 Range of calibration parameters

Calibration Parameters	**Min Value**	**Max Value**	**True Value**
Rate sensitivity exponent for glide	2	4	3
The ratio of threshold stress for glide and for climb	8000	12000	11000
Rate sensitivity exponent for climb	2	4	3

2.6 Batch Sequential Calibration

Executing the VPSC code with theoretical "true values" of the three calibration parameters, a synthetic representation of "truth" is generated. First, we investigate the selection of optimal experiments with an exact numerical model (i.e. the representation of physics or engineering principles is complete), in which the only difference between the model and the truth is the experimental variability. Specifically, we focus on the dispersion of the experiments within the domain. Next, we study an (artificially) inexact version of VPSC model, where the synthetic experiments are obtained by adding not only experimental variability but also an artificial discrepancy bias to the truth. The variance

and the complexity of discrepancy bias are varied to investigate the generality of the BSD approach (see Table 2.3). Though both exact and inexact models are calibrated considering four levels of experimental uncertainty (see Table 2.3), only the results for the minimum (0.1%) and maximum (5%) experimental uncertainty are presented here.

Table 2.3 Analysis Case Configurations

Analysis Case Configurations	**Variations**
Experimental Uncertainty	[0.1%; 1.0%; 3.0%; 5.0%]
Variance in discrepancy	[5.0%; 10.0%]
Complexity in discrepancy	Low[.05 .05]; Med-I[.05 .5];Med-II[.5 .05]; High[.5 .5]

Our parametric analysis includes every combination in Table 2.3 for every BSD selection criterion introduced earlier in the manuscript. However, for brevity, we present the findings for one index-based criterion, Expected Improvement for Predictive Stability (EIPS); and one distance-based criterion, Euclidean Distance Criterion (EDIST) and later supply a separate discussion for all investigated selection criteria.

The BSD algorithm is initiated with a starting set of validation experiments selected using a space-filling, Latin hypercube maxi-min sampling (Table 2.4) (Rennen et al. 2010). This initial set of physical experiments, i.e. the starting point for BSD algorithm, provides only a 5% initial coverage and is kept identical for all cases evaluated herein. The potential influence of initial coverage on PMI is investigated later in the

manuscript, in which the BSD procedure is repeated with a higher initial coverage. The BSD augmentation is completed in ten batches with two new experiments in each batch yielding a total of 23 experiments. The procedure is repeated five times to assure the repeatability of findings.

Table 2.4 Settings for the initial three physical experiments, i.e. starting point for BSD

Test No.	Climb Dislocation Orientation (rad)	Deviatoric Stress (MPa)
1	0.1556	922
2	0.3222	1100
3	0.2111	1056

2.7 Exact Model: Coverage of the Domain

The dispersion of EIPS and EDIST selected optimal experiments throughout the domain is investigated considering solely the calibration of the three uncertain parameters of the VPSC code (also known as parameter estimation).

Expected Improvement for Predictive Stability (EIPS)

Figure 2.2 is a representative plot for one of the five repeats showing the distribution of EIPS selected validation experiments in the operational domain as ten new batches are added to the starting experiments. In Figure 2.2, the location of all 23 experiments and the domain coverage, ΩCH corresponding to the addition of every other batch are also indicated. While the initial stage has coverage of only 5%, with the addition of ten EIPS selected batches, the coverage steadily increases to 80%. As is

clearly evident in Figure 2.2 however, the EIPS criterion has a tendency to clump the experiments (see for instance experiments 6, 9, 11 and 16 in Figure 2.2).

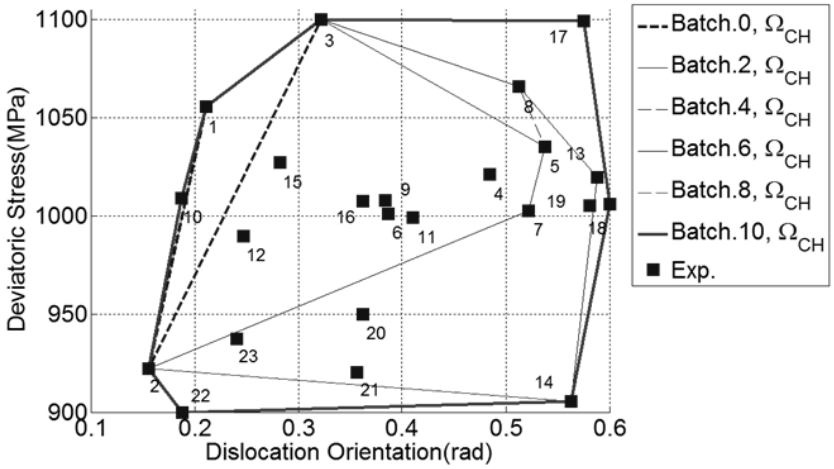

Fig. 2.2 Coverage and domain of applicability of experiments with EIPS (a representative plot, one of the five repeats) with 5% initial coverage

Euclidean Distance Criterion (EDIST)

Compared to EIPS, EDIST selects experiments that explore a greater percentage of the domain and has a tendency to distribute the experiments more uniformly without any noticeable clumping (Figure 2.3). With EDIST, the coverage attribute of PMI immediately increases as the experiments selected for the first batch (experiments 4 and 5) are at significantly distant points from the three starting experiments, resulting in an increase in coverage from 5% to 60% in a single step. After the last batch, the convex hull is nearly equal to the entire domain of applicability, reaching coverage of 99%. Such rapid improvement in coverage is consistently observed for all repeats of the EDIST criterion.

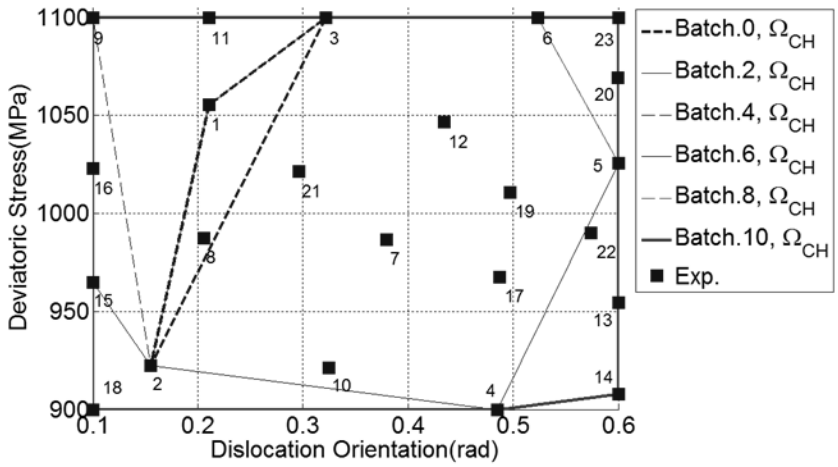

Fig. 2.3 Coverage and domain of applicability of experiments with EDIST

2.8 Inexact Model: Considering Discrepancy

In this section, optimal experiments are selected considering both variance reduction (i.e. parameter calibration) and bias correction (i.e. discrepancy inference). In practical applications, the discrepancy bias assumes a myriad of complexities (smoothly varying, i.e. long correlation length, versus rapidly varying, i.e. short correlation length) and variances depending upon the nature of the missing physics phenomena. We are concerned with the effect of discrepancy variance and complexity on the performance of selection criteria for BSD optimization. Thus, various possible forms of discrepancy bias are represented through a parametric analysis of complexity and variance of the discrepancy function. Recall Eq. 2.2, in which the discrepancy is defined as a function of control parameters i.e., climb dislocation orientation and deviatoric stress input; therefore, the complexity of discrepancy must be defined separately for each of the two control parameters of the VPSC code, resulting in four distinct combinations as given in Table 2.3.

Discrepancy bias of the calibrated model is calculated with reference to the "truth" at 26x26 grid points evenly distributed in the domain of applicability (see Table 2.1 for the bounds of the domain of applicability). The discrepancy estimated for all grid points is then normalized with respect to corresponding mean predictions and evaluated as percentage values.

Expected Improvement for Predictive Stability (EIPS)

Figure 2.4 presents the PMI values for four combinations of discrepancy complexity and variance for 0.1% experimental uncertainty. Visually, the PMI exhibits a convergent behavior towards a value of 90% at the 10^{th} batch for all cases. Figure 2.5 illustrates the improvements in discrepancy and coverage attributes of PMI as ten new EIPS selected batches become available for model calibration. In all four cases of discrepancy variance and complexity, the discrepancy is improved to a level below 6% through ten BSD selected batches. Simultaneously, the coverage of the domain is increased from 5% at the starting set of experiments to above 80%. Generally speaking, the variance and complexity of discrepancy is observed to have minimal influence on the coverage and scaled discrepancy attributes.

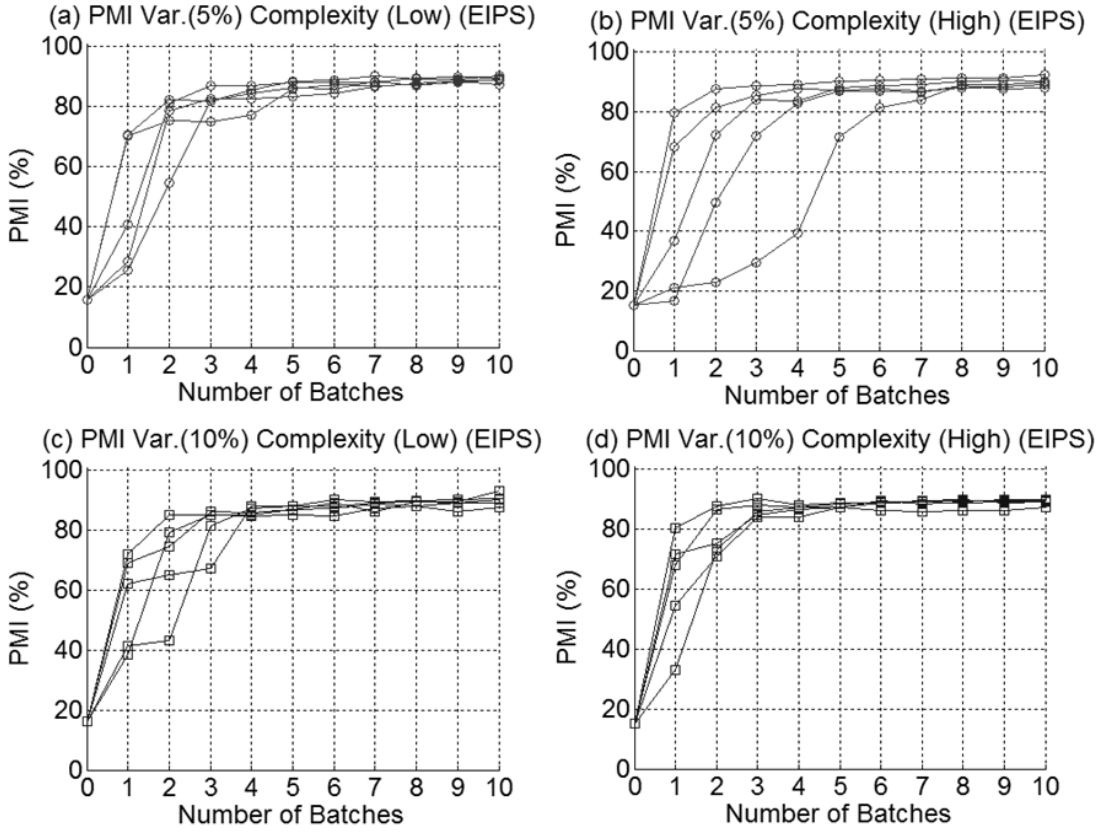

Fig. 2.4 PMI with EIPS for inexact model with 0.1% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

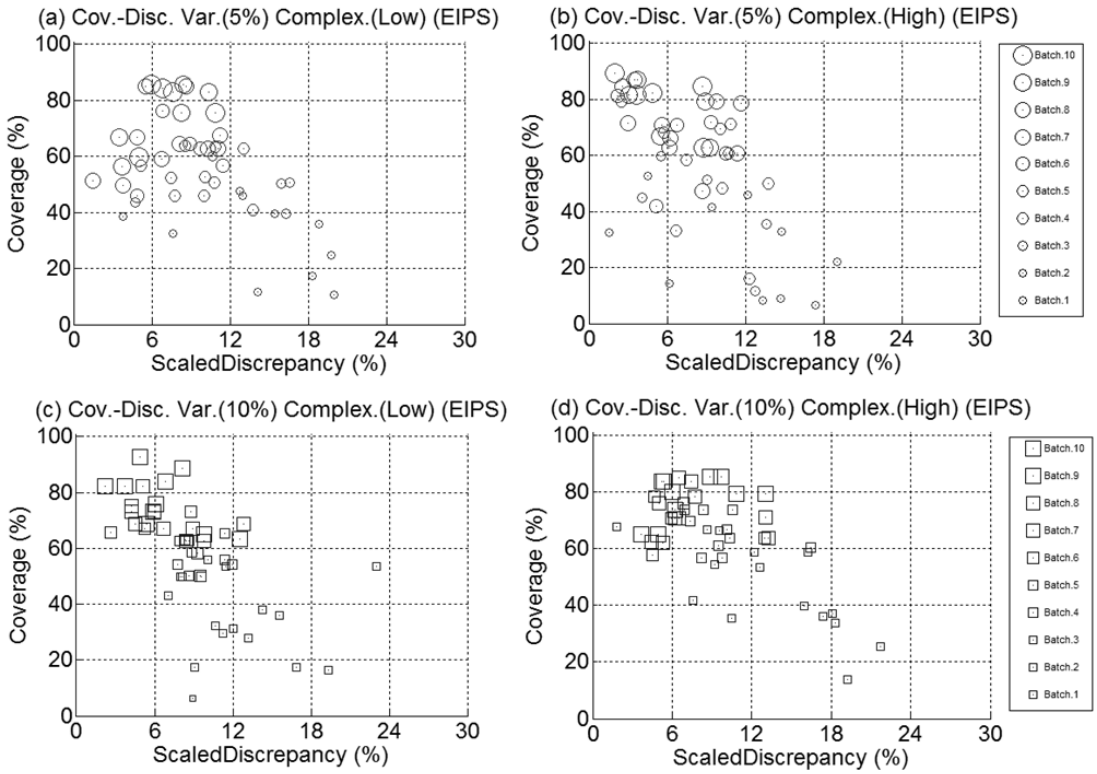

Fig. 2.5 Normalized discrepancy vs. coverage for EIPS for inexact model with 0.1% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

Figure 2.6 illustrates the PMIs calculated for four combinations of discrepancy variance and complexity when the experimental uncertainty is 5%. The improvement in PMIs is non-monotonic with less clear stabilization compared to Figure 2.4. The five repeats on average reach 85% PMI value at the 10^{th} batch. From comparing Figures 2.6 and 2.8, it is evident that the PMIs obtained with the EIPS selected experiments are influenced significantly by experimental uncertainties. Poor PMI values of Figure 2.6 can be explained by the high scaled discrepancy values (approximately 150% at the early

batches) shown in Figure 2.7. The EIPS criterion provides a consistent improvement in discrepancy with every batch and a coverage level consistently around 80% after the 10^{th} batch.

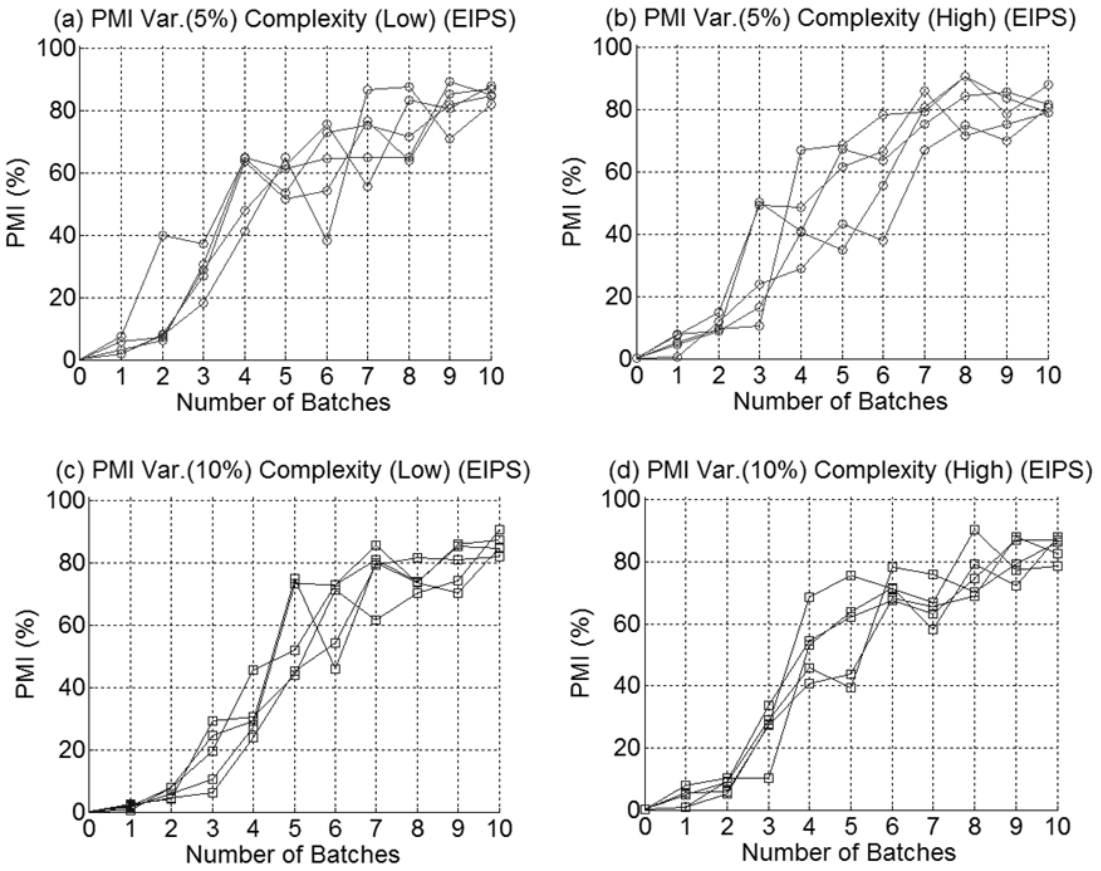

Fig. 2.6 PMI with EIPS for inexact model with 5% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

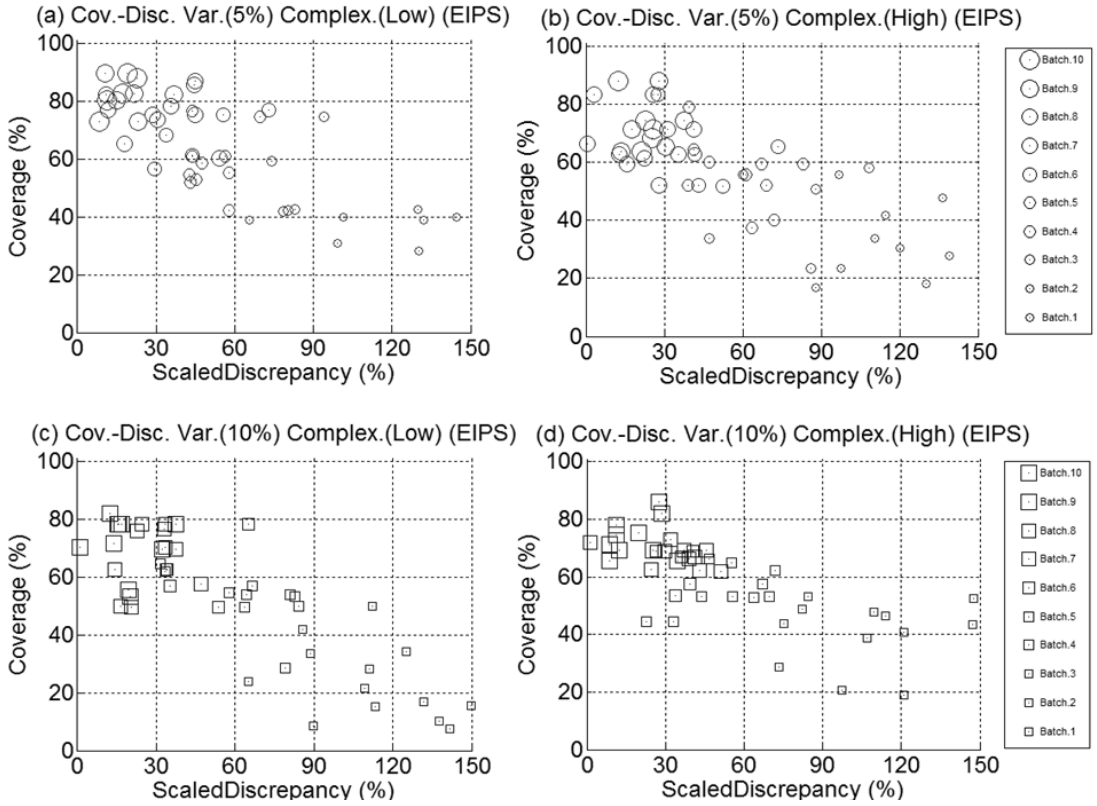

Fig. 2.7 Normalized discrepancy vs. coverage for EIPS for inexact model with 5% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

Euclidean Distance Criterion (EDIST)

Figure 2.8 illustrates the PMI obtained by the EDIST criterion for the 0.1% experimental uncertainty. Regardless of the discrepancy variance and complexity, the PMI consistently reaches a level of 80-85% after the 1st batch and a level of 90%-95% after the 10th batch. In Figure 2.9, we observe that EDIST provides coverage around 60% immediately after the 1st batch and around 95% after the 10th batch. Discrepancy attribute

however, is variable between 3% and 12% for EDIST after the 10^{th} batch. These observations are consistent for every level of complexity and variance of discrepancy.

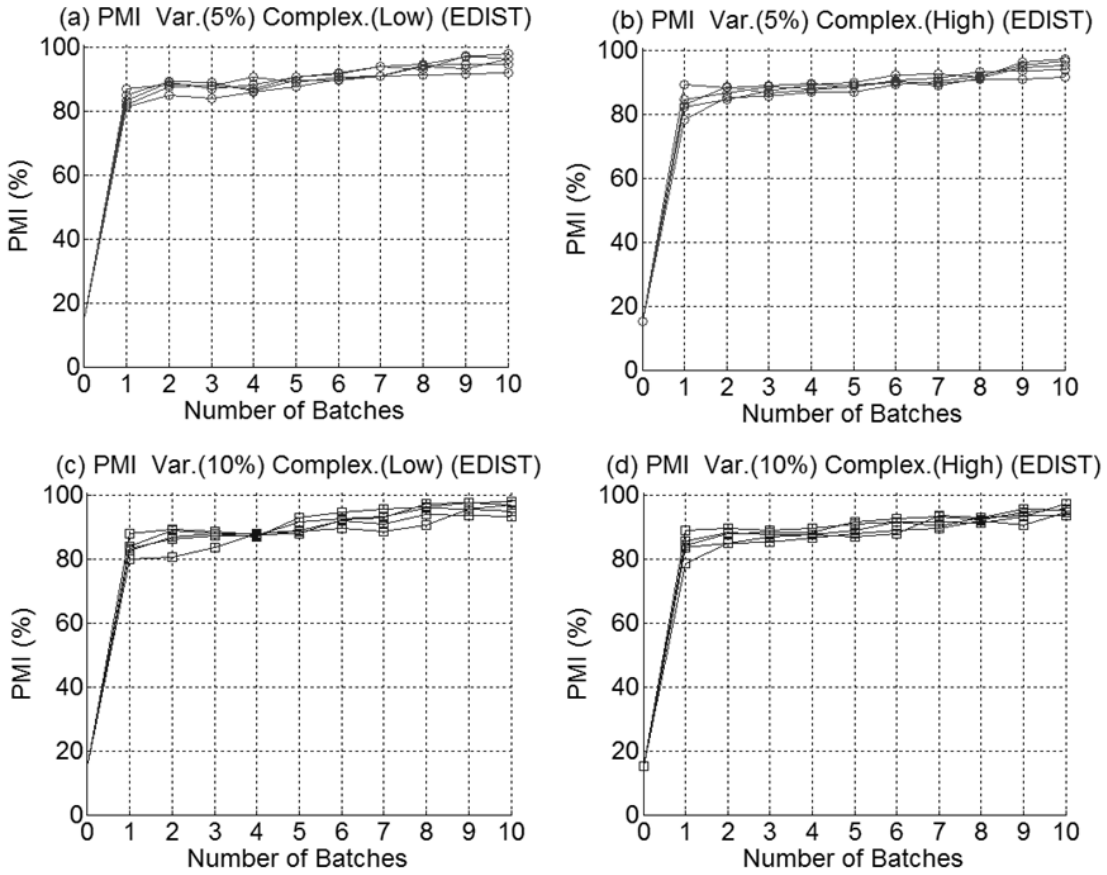

Fig. 2.8 PMI with EDIST for inexact model with 0.1% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

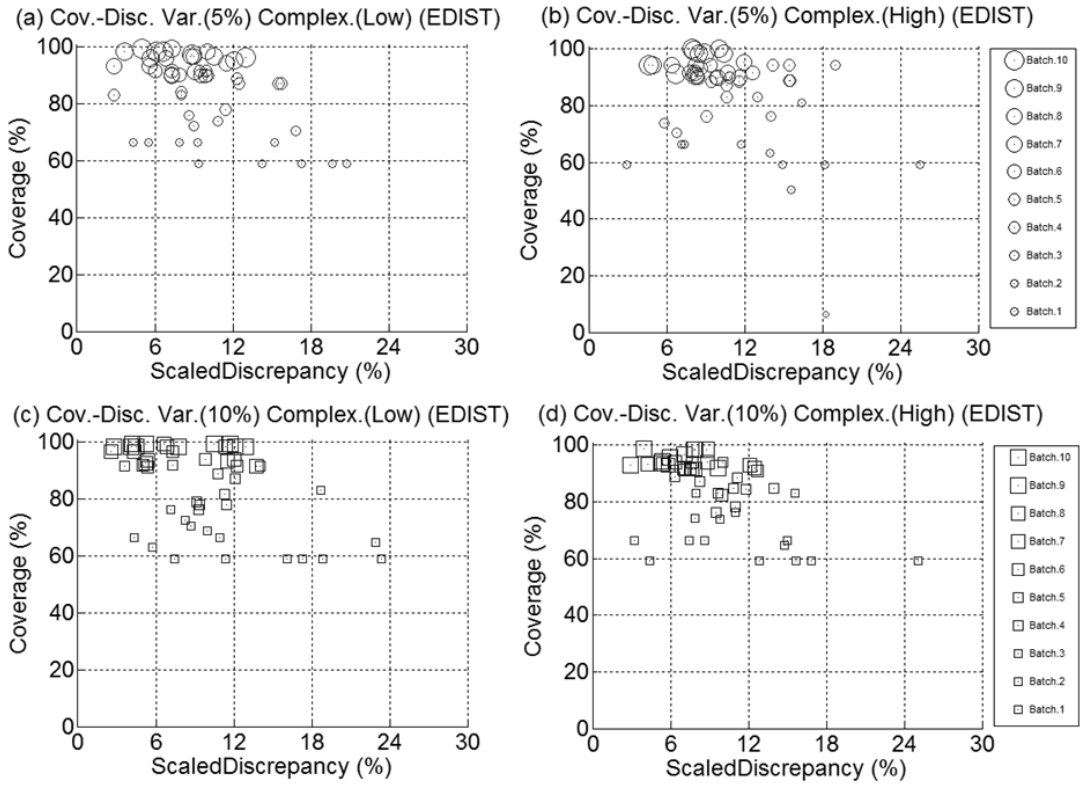

Fig. 2.9 Normalized discrepancy vs. coverage for EDIST for inexact model with 0.1% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

Figure 2.10 illustrates the PMI obtained by BSD when the experimental uncertainty is 5%. PMIs improve in a non-monotonic manner across the ten batches and reach a range between 85%-95% after the 10^{th} batch. The improvement in PMI in this case is slower than the case of 0.1% experimental uncertainty. Generally, PMIs do not exceed 60% until after the 6^{th} batch. Figure 2.11 illustrates that EDIST successfully improves the coverage from 60% to 100% and the discrepancy from 150% to below 15% in four variance and complexity combinations of discrepancy.

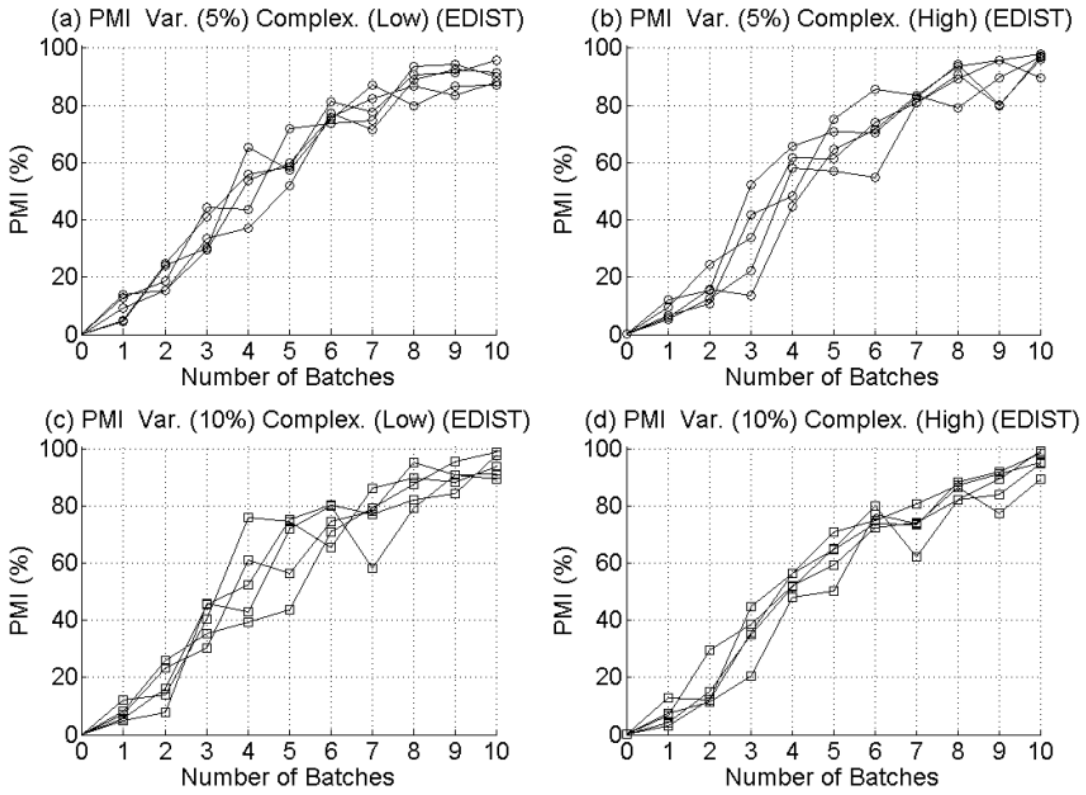

Fig. 2.10 PMI with EDIST for inexact model with 5% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

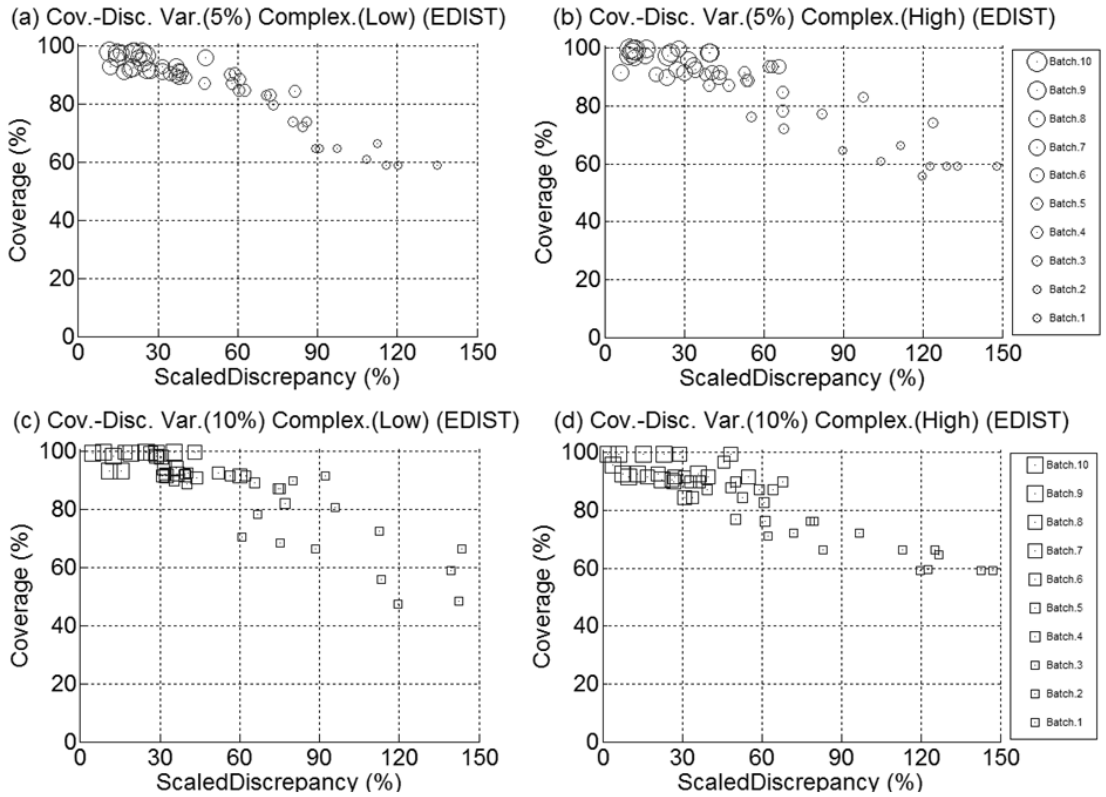

Fig. 2.11 Normalized discrepancy vs. coverage for EDIST for inexact model with 5% experimental uncertainty: (a) Discrepancy with 5% variance and low complexity, (b) Discrepancy with 5% variance and high complexity, (c) Discrepancy with 10% variance and low complexity, (d) Discrepancy with 10% variance and high complexity

Compared to EIPS, the discrepancy bias obtained with EDIST is higher. However, the ability of the EDIST criterion to explore the operational domain is particularly noticeable for all combinations of variance and complexity of discrepancy, and experimental uncertainty.

2.9 Discussion and Findings

In the previous section, BSD optimization proved to be successful in yielding simultaneous improvement in discrepancy and coverage attributes. This is no surprise,

however. Allocating resources to experimentation is expected to result in better coverage and model parameters are typically better conditioned with increased amounts of experimental data. However, the benefit of BSD lies in the efficiency of this improvement. To illustrate the efficiency of BSD selection, Figure 2.12 compares the improvement in PMI obtained by BSD selected experiments with the same number of experiments selected through a single-stage space-filling design strategy. Specifically for comparison, we use maximin Latin-Hypercube design, which is concluded to be comparable to sequential approaches in Williams et al. (2011). As seen, PMI converges with a higher rate and more consistently when BSD is implemented to iteratively select the optimal experiments. This observation is especially obvious at early batches when the number of experiments is inadequate.

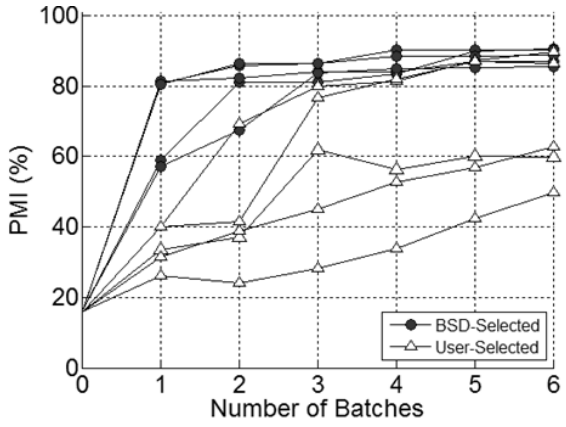

Fig. 2.12 Predictive Maturity achieved by BSD algorithm versus PMI with user-selected test settings

The previous section reveals the strong dependence of PMI values to the experimental uncertainty. In general, for both index-based and distance-based criteria, we observe higher PMI values and a more rapid convergence when the experimental

uncertainty is low. For the 0.1% experimental uncertainty, it is common for PMI to converge as early as three batches; while for the 5% experimental uncertainty; no less than 11 batches are required for convergence (see Figure 2.13). The sensitivity of the PMI values to experimental uncertainty may be explained by two of the three calibration parameters of the VPSC model being exponents (recall Eq. 2.4). The inferred discrepancy bias is therefore very sensitive to the proper calibration of these two parameters, posteriors of which are influenced by the uncertainty in the validation experiments.

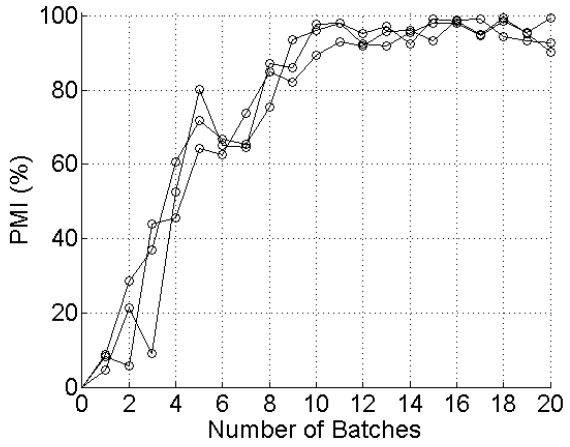

Fig. 2.13 Convergence of PMI through 20 batches by EDIST for exact model with 5% experimental uncertainty

For the discrepancy complexity and variance, trends are less recognizable. For EIPS, the point clouds of coverage tend toward lower values as the discrepancy variance increases (see Figure 2.7, for instance). For EDIST however, no particular trend is recognized in normalized discrepancy and coverage attributes for varying levels of discrepancy complexity and variance.

The index-based EIPS criterion is observed to favor improving discrepancy over coverage, since even after ten batches, the coverage remains less than 90% for all cases evaluated in Section 2.6. The distance-based EDIST criterion however, exhibits an immediate increase in the coverage after the first batch and consistently reaches 99% coverage. The index-based EIPS criterion, however, consistently yields lower normalized discrepancy compared to the distance-based EDIST criterion. Thus, the EIPS criterion can be considered to be more successful than EDIST in improving the discrepancy at low coverage; while the EDIST criterion to be more successful than EIPS in improving the coverage. To test the performance of EIPS and EDIST criteria from the perspective of discrepancy inference, we provide a mathematical proof-of-concept example:

Discrepancy Comparison: Index- and Distance- Based Criteria

Here, the discrepancy biases inferred from the optimal experiments selected by EIPS and EDIST are compared against the artificially generated "true" discrepancy (i.e., model form error), a smoothly varying analytical function of a sine wave. Similar to the proof-of-concept example discussed in Section 2.2, the "true" discrepancy is known at every point in the domain of applicability. The goal however, is to retrieve this discrepancy bias by exploiting the availability of a sound physics model and validation experiments.

The discrepancy bias is *estimated* twice with fifteen validation experiments selected in six batches by the EIPS and EDIST criteria. Figure 2.14 compares the true values of discrepancy with those that are inferred from the validation experiments. Ideally, this comparison yields a 45° angle; that is if the true

discrepancy is perfectly identified from experiments. For EIPS estimated discrepancy values line up around an angle of 34°, while for EDIST, the angle drops down to 21°. This illustration highlights that EIPS should be favored over EDIST in obtaining a proper inference of discrepancy bias.

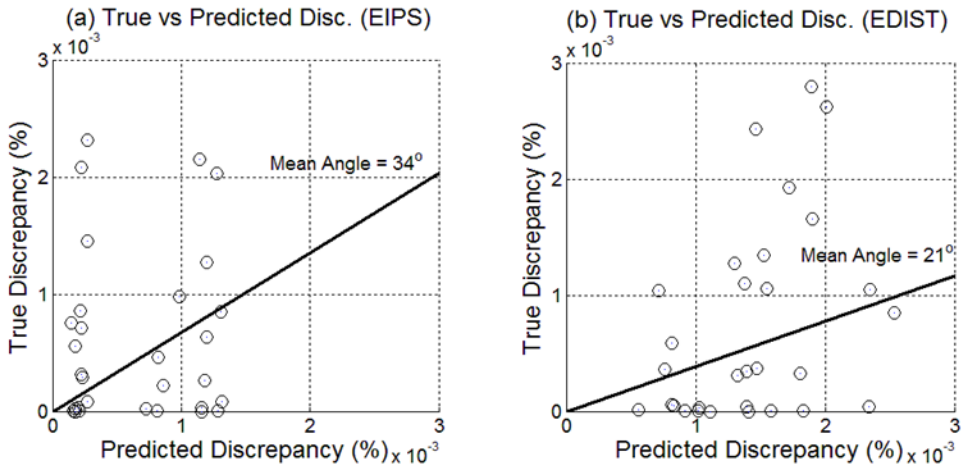

Fig. 2.14 Comparison between true and estimated discrepancy using (a) EIPS, (b) EDIST

The performance of the EIPS criterion in improving the discrepancy inference is observed as a common characteristic of index-based criteria. For instance, on the VPSC polycrystalline plasticity model, the EIGF criterion is observed to yield similar discrepancy levels as EIPS with slightly higher coverage of the domain. Similarly, the IMSE criterion is observed to favor discrepancy over coverage with a lesser performance for both of these attributes compared to EIPS. This is consistent with the observations of Sacks et al. (1989), who reports IMSE's lack of attraction to the boundaries of the domain. While consistent with the EIPS criterion in the reduction of the discrepancy, the ENT criterion, however, is

observed to yield very high coverage (as high as 100%), once again in agreement with the findings of Sacks et al. (1989). In this study, ENT provided the most favorable results among the index-based criteria from the point of view of PMI. However, results in Williams et al. (2011) indicate that ENT has increasing difficulty in stabilizing discrepancy as the dimension of the control variable, x increases, a trend not seen with other BSD criteria.

The index-based EIPS criterion and distance-based EDIST criterion are indicated to provide low discrepancy and high coverage, respectively. To benefit from both of these criteria, the BSD optimization can be performed by mixed strategy; i.e. switching to a different criterion when the discrepancy is reduced below or coverage is increased above a certain threshold. Our investigation of the concept of this mixed strategy is detailed below:

Mixed BSD Strategy: Index- and Distance- Based Criteria

EDIST is first implemented to improve coverage to above 90%, at which point the design criteria is switched to EIPS to reduce the discrepancy. Figure 2.15 illustrates the PMI values through BSD cases for the inexact model solution for the case with 5% experimental uncertainty, 5% discrepancy variance and low complexity. EDIST reaches the 90% coverage after the 6^{th} batch and the EIPS is employed to investigate the further improvement in scaled discrepancy. The mean scaled discrepancy values of the mixed strategy are compared with those that are obtained solely with EIPS and EDIST criteria at the 9^{th} and 10^{th} batch. EIPS and EDIST discrepancy is reduced to 20% and 19% respectively, while the mixed

strategy provides 16% discrepancy. As the low discrepancy values are predicted in the presence of high coverage, the mixed strategy provides improvement in the PMI.

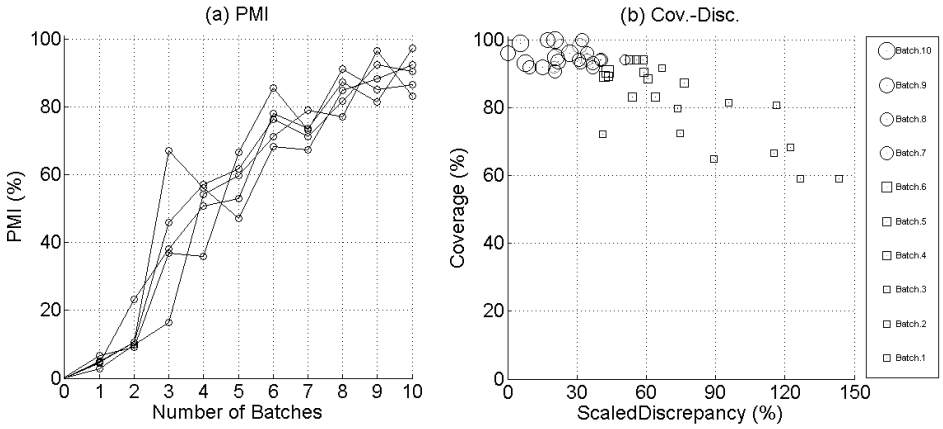

Fig. 2.15 Mixed criteria strategy with EIPS and EDIST for 5% experimental uncertainty in inexact model: (a) PMI vs. number of batches, (b) Normalized discrepancy vs. coverage attributes

A significant difference is observed between the EIPS and EDIST criteria: EDIST provides an immediate increase in coverage after the first batch, while EIPS needs several batches to improve the coverage to a similar level. In all the cases presented in this section, the BSD optimization is initiated with the same initial set of three experiments. As discussed earlier, this initial set provides a very low coverage of the domain, 5%. It is of interest to investigate if the initial coverage of the domain has an impact on the performance of BSD optimization and associated selection criteria.

Effect of Initial Experimental Settings

The starting set of three experiments are set to [0.105*rad*, 910*MPa*], [0.125*rad*, 1100*MPa*] and [0.6*rad*, 910*MPa*] to provide a relatively high coverage level (47%). Figures 2.16a and 2.16b illustrate the PMI values for the EIPS criterion for the inexact model with discrepancy variance of 5% and low complexity, with 0.1% experimental uncertainty and 5% experimental uncertainty, respectively. The high initial coverage improves the PMIs at the early batches for 0.1% experimental uncertainty. When the experimental uncertainty is 5%, the scaled discrepancy becomes higher (150%-200%) at the early batches; and the improved initial coverage is insufficient to yield a significant improvement in PMI values.

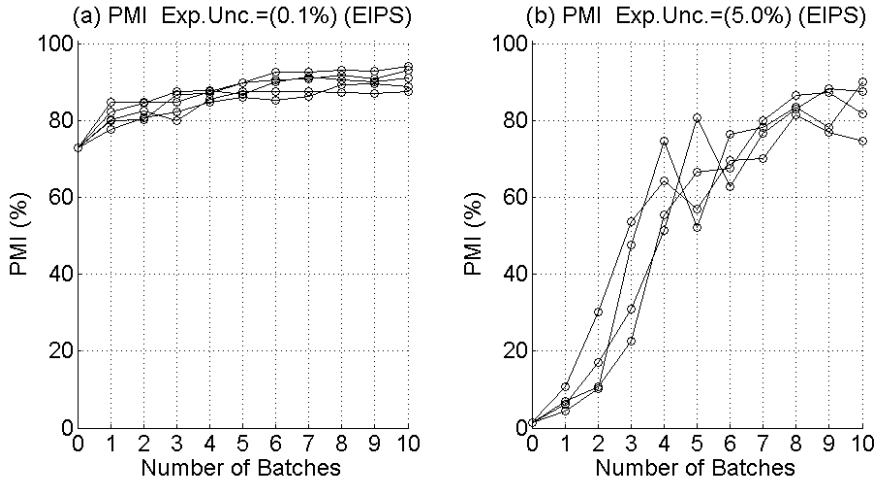

Fig. 2.16 PMI for inexact model by EIPS with 47% initial coverage settings for:

(a) 0.1% experimental uncertainty, (b) 5% experimental uncertainty

Throughout the paper, the PMI values were calculated with equal weight for the coverage and discrepancy attributes. However, the user-defined weighting coefficients of PMI provide flexibility in analysis in cases where coverage or discrepancy may be

assigned higher importance over the other. Next, we investigate the effect of these user-defined gamma values on the performance of EIPS and EDIST criteria:

Selection of Gamma Values

First, we obtain gamma, γ values through an ANOVA-based global sensitivity analysis, which yield two distinct cases: (1) a PMI that is more sensitive to the coverage attribute. (2) a PMI that is more sensitive to the discrepancy attribute. To investigate the influence of the gamma values, the PMIs are compared for EIPS and EDIST for an inexact VPSC model with 5% discrepancy variance, low complexity and 5% experimental uncertainty.

In Figure 2.17a, the gamma values are adjusted to γ_2 =0.5 and γ_3 =5.0 to increase the weight of coverage in the PMI calculations. The slower pace of EIPS in improving coverage becomes more evident in PMI. Similarly, in Figure 17b, the gamma values are adjusted to γ_2 =2.0 and γ_3 =2.0 to increase the weight of discrepancy in the PMI calculations. Here, the PMI values yield comparable results to the default settings.

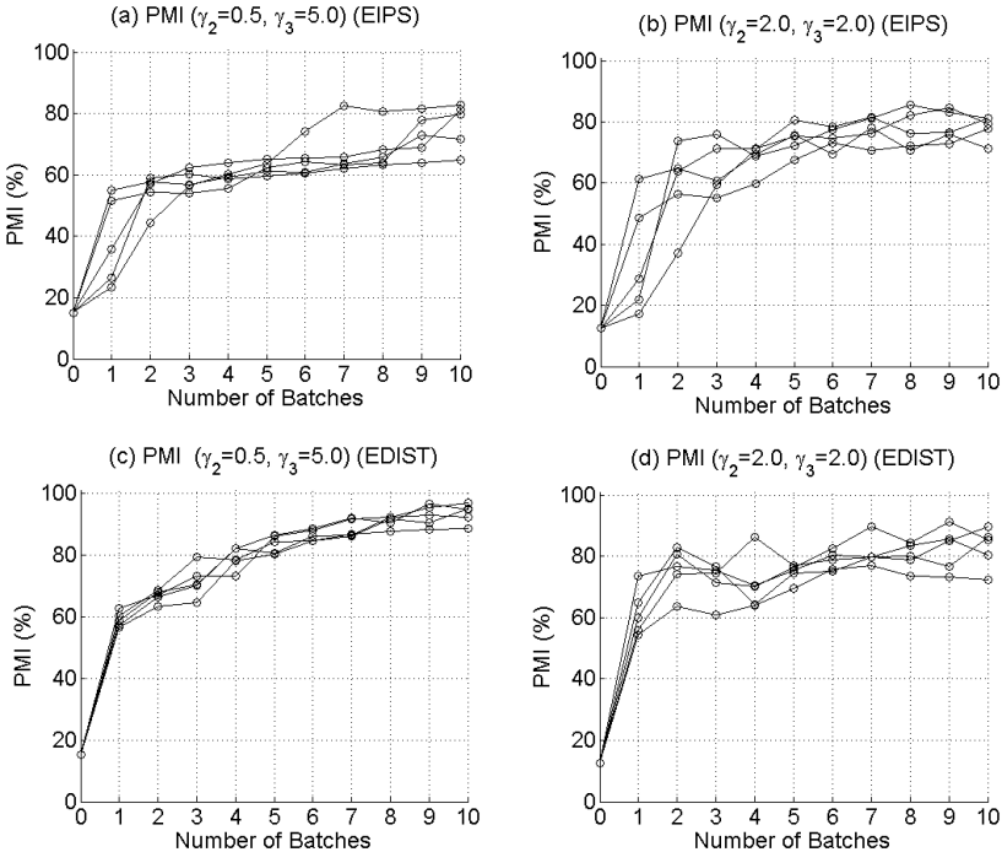

Fig. 2.17 PMI for inexact model by: (a) EIPS with coverage dominant gamma values, (b) EIPS with discrepancy dominant gamma values, (c) EDIST with coverage dominant gamma values, (d) EDIST with discrepancy dominant gamma values

The same procedure is also repeated using the EDIST criterion. In Figure 2.17c, for the coverage dominant case, the PMI exhibits a convergent trend in a generally monotonic manner and reaches as high as 85-90% (in comparison to the non-convergent behavior observed in Figure 2.17a). However, when the weight is shifted towards discrepancy, the improvement in PMI is observed to be less

consistent, due to the fluctuations of the estimated discrepancy bias (see Figure 2.17d).

2.10 Conclusions

The accuracy and precision of model predictions can be improved through the availability of validation experiments and in turn, the design of validation experiments can be improved through the availability of sound numerical models. In this case, the central question concerns how to exploit a sound, physics based numerical model for designing validation experiments that are useful for calibrating the model such that a desired level of predictive maturity can be achieved with the least possible number of validation experiments. In this study, we tackle this question through the use of BSD approach. Our particular interest is in the evaluation of various selection criteria that define the desired benefits from future experiments. Depending on the selection criteria, BSD results in designs that either favor exploration of the domain or exploitation of variance and bias.

EIPS is observed to be more favorable for cases where discrepancy is critical, while EDIST is observed to be superior where a high coverage of the domain of applicability is needed. To enhance BSD optimization and benefit from the disparate influences of EIPS and EDIST on the PMI attributes, we recommend mixing these design criteria. The mixed strategy is observed to lower the discrepancy level at the further batches after obtaining a sufficient amount of coverage.

In the application to VPSC, both index- and distance- based criteria are observed to exhibit sensitivity to experimental uncertainty. This can be explained by the significant

influence of the two calibration parameters, climb and glide exponents, have on the inferred discrepancy bias. Both index- and distance- based criteria exhibit negligible sensitivity to the variance and complexity levels of discrepancy considered herein. The comparisons in this study can provide guidance for the analyst selecting the design criteria in the BSD application in future applications.

In this study, it is assumed that experimentation is possible throughout the entire domain. In reality, experimentation may be prohibited in particular regions of the domain due to testing restrictions or infeasibility of recreating extreme operational conditions within a laboratory. In future studies, BSD algorithm will be configured to select validation experiments within a restricted region of the domain. Furthermore, for multivariate models in which different types of experiments may be used in the calibration, the BSD criteria would be needed to be applied to the selection of not only the experimental settings but also the types. For example, in Atamturktur et al. (2014), maturity of the VPSC model is achieved using three types of experiments: stress-strain measurements and two different texture intensities. The application of BSD to select the optimal type as well as settings of validation experiments will be investigated in the future.

References

Atamturktur S, Hegenderfer J, Williams B, Egeberg M, Lebensohn R, Unal C (2014) A Resource Allocation Framework for Experiment-Based Validation of Numerical Models. Journal of Mechanics of Advanced Materials and Structures (Taylor & Francis), DOI 10.1080/15376494.2013.828819

Atamturktur S, Hemez F, Williams B, Tome C, Unal C (2011) A forecasting metric for predictive modeling. Computers & Structures 89:2377-2387

Balci O, Adams RJ, Myers DS, Nance RE (2002) Credibility assessment: a collaborative evaluation for credibility assessment of modeling and simulation applications, In Proceedings of the 34th winter simulation conference: exploring new frontiers, San Diego, California, USA, 214-20

Box GEP, Draper NR (1959) A Basis for the Selection of a Response Surface Design. Journal of the American Statistical Association 54:622-654

Bulutoglu DA, Ryan KJ (2009) D-optimal and near D-optimal 2k fractional factorial designs of resolution V. Journal of Statistical Planning and Inference 139:16-22

Cook RD, Nachtsheim CJ (1980) A Comparison of Algorithms for Constructing Exact D-Optimal Designs. Technometrics 22: 315-324

Dersjö T, Olsson M (2012) Efficient design of experiments for structural optimization using significance screening. Journal of the International Society for Structural and Multidisciplinary Optimization (ISSMO) 45:185–196

Draper D (1995) Assessment and Propagation of Model Uncertainty. Journal of the Royal Statistical Society 57:45–97

Evans J, Manson A (1978) Optimal Experimental Designs in Two Dimensions Using Minimum Bias Estimation. Journal of the American Statistical Association 73:171–176

Farajpour I, Atamturktur S (2012) Error and Uncertainty Analysis of Inexact and Imprecise Computer Models. Journal of Computing in Civil Engineering doi:10.1061/(ASCE)CP.1943-5487.0000233

Fedorov VV (1972) Theory of Optimal Design. New York: Academic

Green LL, Blattnig SR, Hemsch MJ, Luckring JM, Tripathi RK (2008) An uncertainty structure matrix for models and simulations. American Institute of Aeronautics and Astronautics AIAA-2008-2154

Hemez F, Atamturktur S (2011) The dangers of sparse sampling for the quantification of margin and uncertainty. Reliability Engineering & System Safety 96:1220-1231

Hemez F, Atamturktur S, Unal C (2010) Defining predictive maturity for validated numerical simulations. Computers and Structures Journal 88:497-505

Higdon DM, Lee H, Holloman C (2003) Markov chain Monte Carlo–Based approaches for inference in computationally intensive inverse problems. Bayesian Statistics 7:181-197

Higdon D, Gattiker J, Williams B, Rightley M (2008) Computer model calibration using high-dimensional output. Journal of the American Statistical Association 103:482, 570-583

Jacobson JJ, Matthern GE, Piet SJ, Shropshire DE (Aril 2009) Vision: Verifiable Fuel Cycle Simulation Model. Advances in Nuclear Fuel Management IV (ANFM). Hilton Head, South Carolina, USA

Johnson ME, Moore LM, Ylvisaker D (1990) Minimax and maximin distance designs. Journal of Statistical Planning and Inference 26:131-148

Kennedy MC, O'Hagan A (2001) Bayesian Calibration of Computer Models. Journal of the Royal Statistical Society 63: 425-464

Lam CQ, Notz WI (2008) Sequential adaptive designs in computer experiments for response surface model fit. Statistics and Applications 6:207-233

Lebensohn RA, Hartley CS, Tomé CN, Castelnau O (2010) Modeling the mechanical response of polycrystals deforming by climb and glide. Philosophical Magazine 90:5,567-583

Li G, Aute V, Azarm S (2010) An accumulative error based adaptive design of experiments for offline metamodeling. Journal of the International Society for Structural and Multidisciplinary Optimization (ISSMO) 40:137-155

Lindley DV (1972) Bayesian Statistics: A Review. Society for Industrial and Applied Mathematics. Montpelier, Vermont: Capital City Press

Loeppky JL, Moore LM, Williams BJ (2010) Batch sequential designs for computer experiments. Journal of Statistical Planning and Inference 140:1452-1464

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. Journal of Chemical Physics 21:1087-1091

Morris MD, Mitchell TJ (1995) Exploratory designs for computational experiments: two examples. Statistica Sinica 2:359-379

Müller WG, Pötscher BM (November 1989) Batch Sequential Design for a Nonlinear Estimation Problem. Forschungsbericht/Research Memorandum No. 259

Myers RH, Khuri AI, Carter WH (1989) Response Surface Methodology: 1966-1988. Technometrics 31:137-157

Oberkampf WL, Pilch M, Trucano TG (2007) Predictive capability maturity model for computational modeling and simulation. Sandia National Laboratories Report; SAND2007-5948

Ogungbenro K, Graham G, Gueorguieva I, Aarons L (2005) The use of a modified Fedorov exchange algorithm to optimize sampling times for population pharmacokinetic experiments. Computer Methods and Programs in Biomedicine 80:115-125

Rennen G, Husslage B, Van Dam ER, Hertog DD (2010) Nested maximin Latin hypercube designs. Journal of the International Society for Structural and Multidisciplinary Optimization (ISSMO) 41:371-395

Rosner R (2008) Making nuclear energy work How shifting research goals and improving collaboration with industry will help U.S. national labs spur new nuclear energy development. Bulltin of Atomic Scientists 64(1):28-33

Sacks J, Schiller SB (1988) Spatial designs. In Gupta, S. S., Berger, J. O. (Eds.) Statistical Decision Theory and related Topics IV, Springer, Berlin 2:385-399

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Designs and analysis of computer experiments. Statistical Science 4:409-435

Shao, T (2007) Toward a structured approach to simulation-based engineering design under uncertainty. PhD Dissertation, University of Massachusetts, Amherst

Shewry MC, Wynn HP (1987) Maximum entropy sampling. Journal of Applied Statistics 14:165-170

Thompson DE, McAuley KB, McLellan PJ (2010) Design of optimal experiments to improve model predictions from a polyethelene molecular weight distribution model. Macromolecular Reaction Engineering 4(1):73-85

Trucano TG, Swiler LP, Igusa T, Oberkampf WL, Pilch M (2006) Calibration, validation, and sensitivity analysis: What's what. Reliability Engineering & System Safety Journal 91: 1331-1357

Unal C, Williams B, Hemez F, Atamturktur SH, McClure P (2011) Improved best estimate plus uncertainty methodology, including advanced validation concepts, to license evolving nuclear reactors. Nuclear Engineering and Design Journal 241:1813-1833

van Keulen F, Vervenne K (2004) Gradient-enhanced response surface building. Journal of the International Society for Structural and Multidisciplinary Optimization (ISSMO) 27:337-351

Williams B, Higdon D, Gattiker J, Moore L, McKay M, Keller- McNulty S (2006) Combining experimental data and computer simulations, with an application to flyer plate experiments," Bayesian Analysis 1:765-792

Williams BJ, Loeppky JL, Moore LM, Macklem MS (2011) Batch sequential design to achieve predictive maturity with calibrated computer models. Reliability Engineering and System Safety 96(9):1208-1219

Yoshiie T (2005) Factors That Infuence Cascade-Induced Defect Growth in Pure Metals and Model Alloys. Materials Transactions 46(3): 425-432

CHAPTER THREE

DEFINING COVERAGE OF AN OPERATIONAL DOMAIN USING A MODIFIED NEAREST-NEIGHBOR METRIC

3.1 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 (Figure 3.1), which in turn, can result in overconfidence in model predictions [7]. This potential oversight is illustrated in Figure 3.1 by the dashed line suggesting a notional curve, which represents the underestimated predictions of the trained discrepancy bias in untested regions of the domain. To mitigate this problem, it is essential to conduct validation experiments at settings that provide a representation of the entire operational domain. A quantitative measure of the ability of a set of validation experiments to represent the entire domain is referred to as *coverage*.

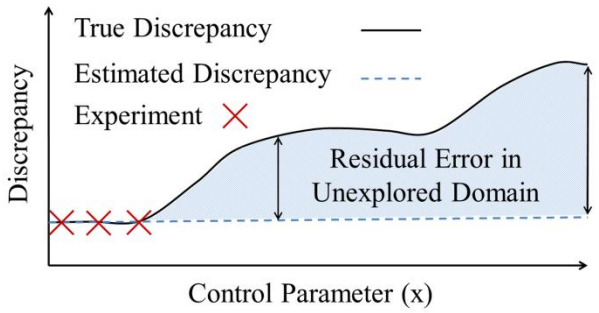

Fig. 3.1 Potential Error in Discrepancy Estimation (reprinted with permission from [23])

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 3.2 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 3.3, we

overview these three abovementioned coverage metrics from the literature and investigate the ability of each metric to meet the identified criteria from Section 3.2.

In Section 3.4 of this paper, a new coverage metric is proposed that alleviates the drawbacks of the existing metrics and satisfies all identified criteria. Section 3.5 demonstrates the applications of the proposed metric on a non-trivial problem of polycrystal plasticity and compares it to existing coverage metrics. In Section 3.6, the effect of dimensionality on the proposed coverage metrics is investigated focusing on a high-dimensional Rosenbrock function. Section 3.7, concludes the paper suggesting alternative uses of the proposed coverage metric.

3.2 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.[4]

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.[5] Experimental design strategies that concentrate runs near the boundaries of the domain, similar to an entropy-based

[4] While conducting experiments at previously sampled settings may provide information about the experimental variability, coverage metrics neglect the benefits of replication.

[5] 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 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.

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.

3.3 Earlier Definitions of Coverage

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

3.3.1 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.[6] Figure 3.2a shows the partitioning of the domain into nearest-neighbor regions. The distance between each grid point and the associated nearest validation experiment is summed for all grid points and normalized by the total number of grid points, as shown in Eq (3.1):

$$nnm = \frac{1}{g}\sum_{i=1}^{g} \min(d_{E,i}) \tag{3.1}$$

where parameter *nnm* represents the nearest-neighbor metric value, *g* represents the total number of grid points, and $min(d_{E,i})$ 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.

[6] Sufficiency of the number of grid points will be discussed later in Section 3.4.

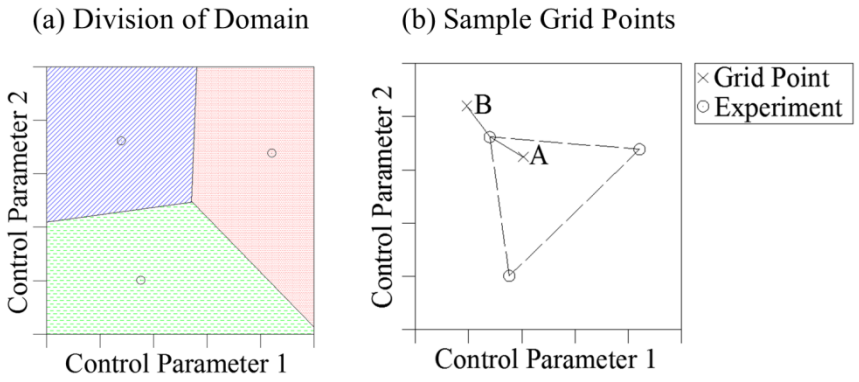

Fig. 3.2 Division of Domain into Nearest-Neighbor Regions

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 Figure 3.2b, 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.

3.3.2 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. 3.2:

$$\eta_c = \frac{V(\Omega_{CH})}{V(\Omega_V)} \tag{3.2}$$

where η_c represents the coverage and $V(\cdot)$ is a function that calculates the volume of a multidimensional domain. Ω_{CH} is the convex hull that surrounds the validation experiments while Ω_V 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 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 Figure 3.3. 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.

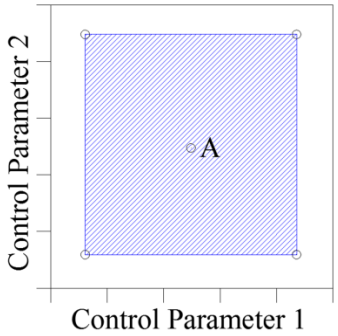

Fig. 3.3 Convex Hull Encompassing Validation Experiments

3.3.3 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. This is defined as:

$$\eta_c = \frac{\sum_{k=1}^{N}(-1)^{k+1}\left(\sum_{1\le i_1<\ldots<i_k\le N}\left|V(\Omega_{E,i_1})\cap\ldots.\cap V(\Omega_{E,i_k})\right|\right)}{V(\Omega_V)} \tag{3.3}$$

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. 3.3, the intersecting convex hulls that double count the coverage are accounted for 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.[7] Furthermore, with this metric, a validation experiment could be added without improving the coverage, provided that the existing convex hulls completely engulf the convex hull of an additional validation experiment, as shown in Figure 3.4 with experiment A.

[7] 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.

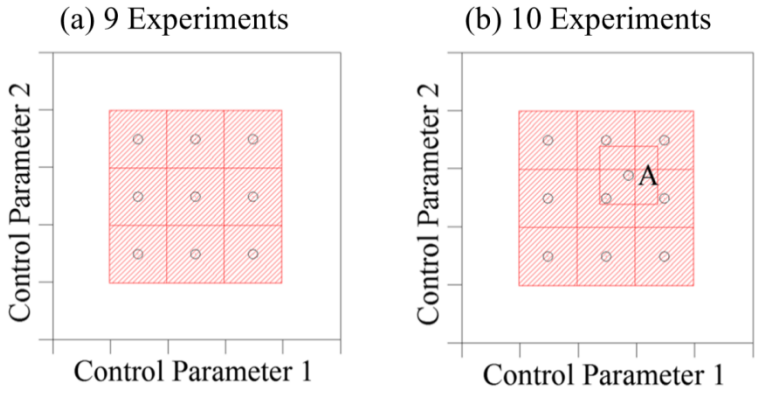

Fig. 3.4 Possible Effect of Adding Validation Experiments on Coverage Metric Proposed by Stull et al. [9]

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 Figures 3.5 and 3.6.

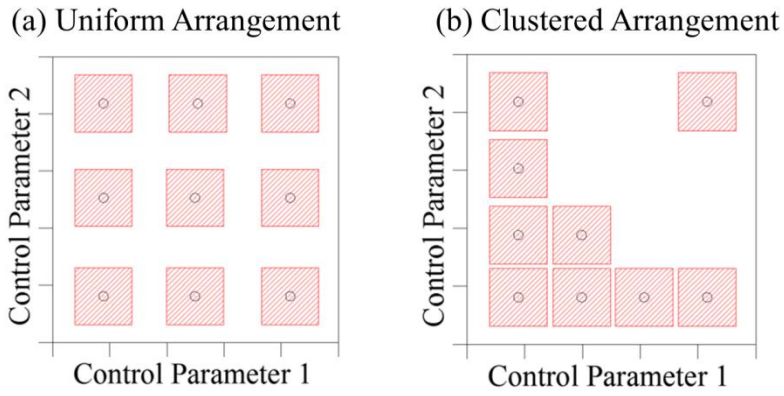

Fig. 3.5 Coverage of Clustered Versus Uniform Arrangement of Validation Experiments

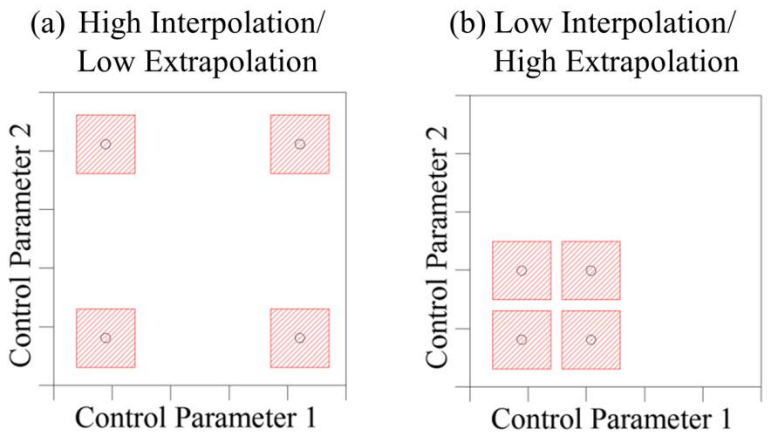

Fig. 3.6 Effect of Interpolation/Extrapolation Ratio on Coverage

The metric proposed by Stull et al. [9] should be given more credit than Figures 3.4, 3.5, and 3.6 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 heavily on expert opinion and increases subjectivity.

The discussion presented in this section is summarized in Table 3.1. Note that each metric fails at least one criterion but each criterion is passed by at least one metric.

Table 3.1 Criterion Satisfaction for Atamturktur et al. [8], Hemez et al. [3], and Stull et al. [9]

Criterion	Atamturktur et al. [8]	Hemez et al. [3]	Stull et al. [9]
1	Pass	Fail	Improved but Imperfect
2	Pass	Fail	Improved but Imperfect
3	Fail	Pass	Fail
4	Pass	Pass	Fail

3.4 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.

3.4.1 Penalizing Extrapolative Predictions in the Coverage Metric

The nearest-neighbor metric is first modified to account for extrapolative predictions. A convex hull encompasses the validation experiments as in Hemez et al. [3], dividing zones of interpolation and extrapolation, as shown in Figure 3.7. 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. 3.4:

$$nnm = \frac{1}{g}\sum_{i=1}^{g} \min(d_{E,i}) + d_{ZI,i} \tag{3.4}$$

where $d_{ZI,i}$ is minimum distance between the i^{th} grid point and the zone of interpolation. Distances d_E and d_{ZI} are shown in Figure 3.7. 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.

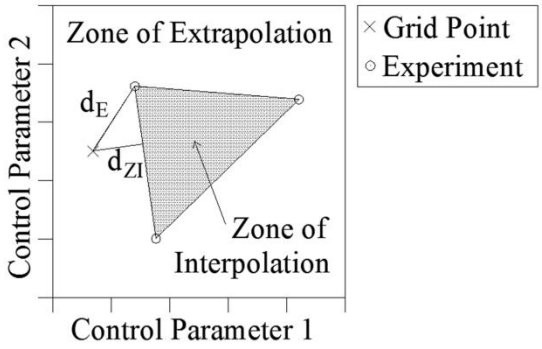

Fig. 3.7 Example Zone of Interpolation and Extrapolation for a Two Dimensional Domain

3.4.2 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 (nnm_{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 (nnm_{max}) is achieved using only one validation experiment. For a rectangular domain, defined by the minimum and maximum values 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 the validation experiment. The average distance between the 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:

$$nnm_{\max} = \frac{\int_{\Omega} \sum_{i=1}^{n} d_{1,i} \; d\Omega}{\int d\Omega} \tag{3.5}$$

where Ω 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. 3.5. This is demonstrated in Figure 3.8 for a two-dimensional domain with the total number of grid points increasing from four to 40,000. In Figure 3.8, 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 Figure 3.9).

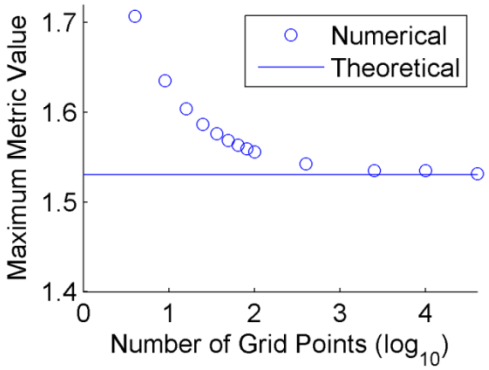

Fig. 3.8 Convergence of maximum metric value to the theoretical value as the number of grid points increases

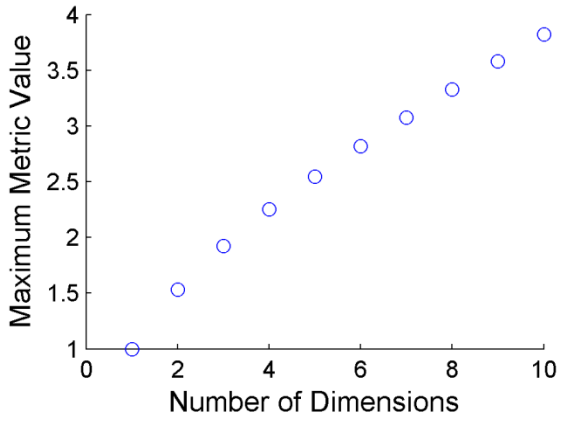

Fig. 3.9 Maximum metric value as a function of dimensionality (for unit sensitivity in each direction)

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:

$$\eta_c = \frac{1}{nnm} - \frac{1}{nnm_{\max}} \tag{3.6}$$

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.

3.4.3 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 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, η_C, robustness to model parameter uncertainty, α_S, scaled discrepancy bias, δ_S, and model complexity, N_K, as described in Eq. 3.7:

$$PMI\left(\eta_C; N_K; \delta_S; \alpha_S\right) = \prod_{i=1}^{5} \Psi_i \tag{3.7}$$

where ψ_i are shown in Table 3.2, with positive, user-defined coefficients γ_1, γ_2, γ_3, and γ_4. The purpose of the gamma values is to weigh the effect of each attribute on the PMI. Note N_R in Table 3.2 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 3.2 are used to provide asymptotic limits between 0 and 1.

Table 3.2 PMI Term Definitions [9]

Term	Definition
Ψ_1	$\begin{cases} \tanh(\gamma_1 \times \eta_c) & \eta_c < 1 \\ 1 & \eta_c \geq 1 \end{cases}$
Ψ_2	$\tanh\left(\left(\frac{N_R}{N_K}\right)^{\gamma_2}\right)$
Ψ_3	$(1-\delta_S)^{\gamma 3}$
Ψ_4	$\left[1-\tanh\left(\frac{\gamma_4}{\alpha_S}\right)\right]$
Ψ_5	$e^{-[e^{-\eta_C} \times \delta_S \times e^{-\alpha_S}]}$

The functional terms shown in Table 3.2 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.[8]

3.5 Demonstrating the use of Coverage Metric

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.

3.5.1 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]:

$$\frac{d\varepsilon}{dt} = \gamma_o \sum_{s=1}^{N_s} \left\{ m^s \left(\frac{\left| m^s : \sigma \right|}{\tau_o^s} \right)^{n_g} \times \operatorname{sgn}\left(m^s : \sigma \right) + c^s \left(\frac{\left| c^s : \sigma \right|}{\sigma_o^s} \right)^{n_c} \times \operatorname{sgn}\left(c^s : \sigma \right) \right\} \tag{3.8}$$

where $\frac{d\varepsilon}{dt}$ denotes the strain rate, and σ represents the stress applied to the crystal. The terms c^s and σ_o^s are the climb tensor and critical stress associated with climb, respectively. Similarly, m^s and τ_o^s are the Schmid tensor and critical resolved shear stress associated with glide. In Eq. 3.8, n_g is the glide stress exponent and n_c is the

[8] 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 Ψ_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 Ψ_1 equals 1 if the coverage equals or exceeds 1 is removed.

climb stress exponent. The single crystal equation is summed over all active slip systems, N_s. Finally, γ_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 (τ_o^s and σ_o^s) 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 [21]. As a result, the VPSC model possesses ten total calibration parameters. In [21], the ten calibration parameters are calibrated against physical validation experiments measuring stress at 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.

3.5.2 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 sensitivity-weighted, 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 Figure 3.10. 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, a fast-running Gaussian Process Model (GPM) emulator [2, 5] is trained to replace the VPSC code. Of course, in the use of an emulator it is necessary to validate the adequacy of the emulator. In this study, closely following the approach taken in [5], hold-out experiments are used to validate that the GPM is trained sufficiently well.

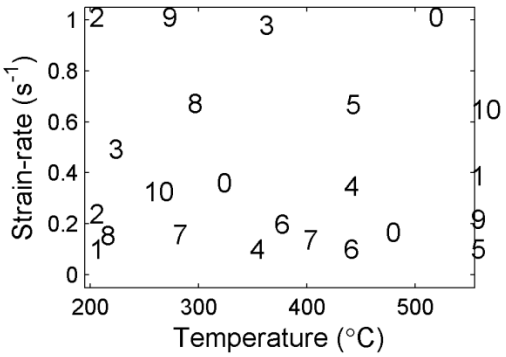

Fig. 3.10 Experimental Settings Selected through BSD (marker number denotes batch number)

3.5.3 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 [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; therefore, a control parameter with greater influence on the output yields a larger β value than a control parameter with less influence. In this application, this sensitivity is determined after the tenth batch.

The coverage obtained using the proposed metric for each batch is shown in Figure 3.11. The coverage of each individual output as well as the average coverage monotonically improves as the number of batches increases.

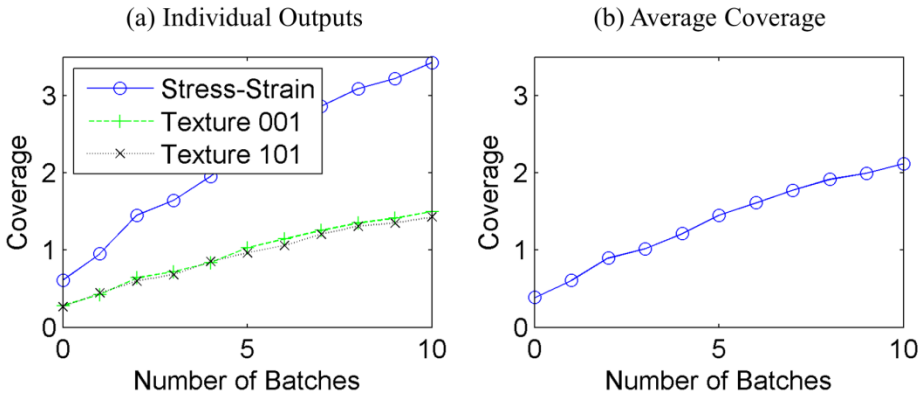

Fig. 3.11 Proposed Coverage vs. Number of Batches

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 Figure 3.12. Between the second and third batches, as well as between the 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 (Figure 3.10). In contrast, the coverage metric proposed herein yields improvement of coverage between every batch (Figure 3.11), recognizing the experiments located inside the convex hull.

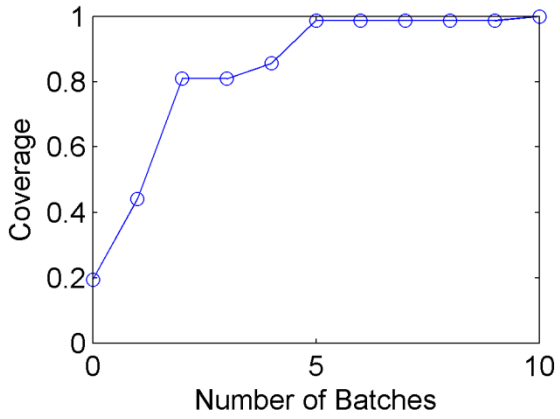

Fig. 3.12 Coverage vs. Number of Batches using Hemez et al. [3] coverage metric

Several alternative coverage values can be obtained using the Stull et al. [9] metric depending on the bound chosen by the expert (Figure 3.13). 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 final batch. In contrast, the coverage metric proposed herein is objective and thus insusceptible to the potential variability between the opinions of two experts.

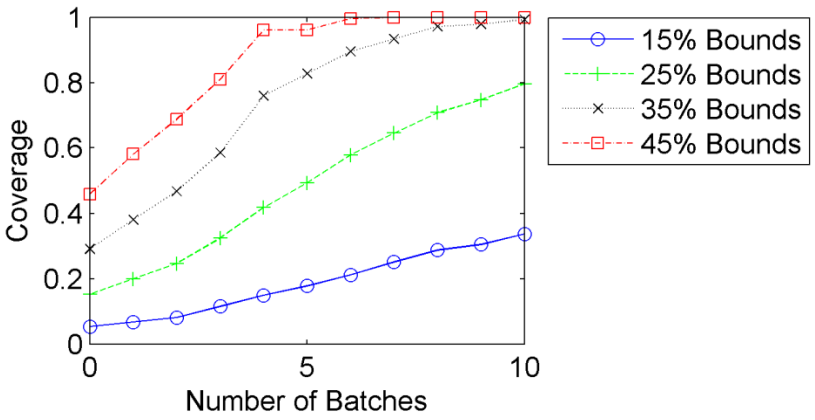

Fig. 3.13 Coverage vs. Number of Batches using Stull et al. [9] coverage metric

3.6 Dimensionality

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

3.6.1 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 equal to one for all dimensions are assumed. The

simulation 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. The results are shown in Figure 3.14.

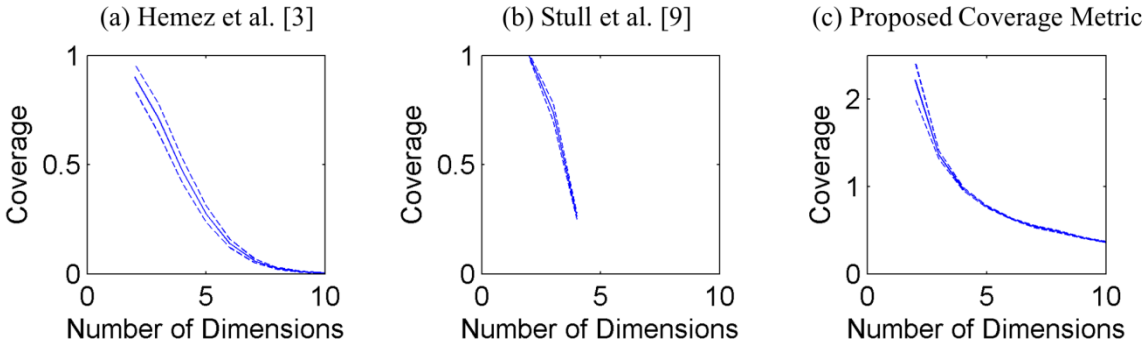

Fig. 3.14 Average Coverage (solid line) ± 3 standard deviations (dashed lines) achieved with 50 simulations of a 100 experiment LHS design

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 high-dimensional problem with 100 dimensions, the coverage achieved using the Hemez et al. [3] metric is nearly negligible, as expected. Similarly, the Stull et al. [9] metric displays a steep decline in coverage as the dimensionality increases and quickly becomes computationally prohibitive.[9] The coverage metric proposed herein is shown in Figure

[9] Data can only be collected for a large sample size for up to four dimensions due to the high computational cost at high 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. A small number of LHS runs yield an average metric value of 0.0686 at five dimensions, consistent with the trend shown in Figure 3.14b.

3.14(c).[10] The rate of decrease in coverage is largest when the number of dimensions is still low. As with the Hemez et al. [3] and Stull et al. [9] metrics, the proposed coverage metric suffers from the curse of dimensionality, displaying a decreasing value as the number of dimensions increases. Therefore, Figure 3.14 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.

In addition to the computational constraints of the Stull et al. [9] metric, the Hemez et al. [3] metric requires a greater number of experiments than the number of dimensions in order to evaluate the metric. Therefore, it would not be possible to evaluate the Hemez et al. [3] metric if there were ten or fewer experiments. Alternatively, the proposed coverage metric may be evaluated using a few as one experiment. Therefore, for a high-dimensional domain with an equal or fewer number of experiments, the proposed coverage metric may be used to evaluate the coverage.

6.2 Coverage of High-Dimensional Domain: Application to the Rosenbrock Function

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

$$Y = \sum_{k=1}^{N-1} (1 + X_k)^2 + C_{k+1}(X_{K+1} - X_k{}^2)^2 \tag{3.9}$$

In Eq. 3.9, 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} = 1,024 runs) design of experiments are analyzed with an analysis-of-

[10] For eight or more dimensions, the proposed coverage metric is evaluated assuming that the entirety of the domain is an extrapolative regime.

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 3.3.

Table 3.3 Coefficients of the Rosenbrock function and statistics for main-effect analysis

Variable (X_k)	**Coefficient (C_k)**	**R^2 Statistic (%)**
1	1.0	22.5%
2	5.0	3.6%
3	2.0	11.2%
4	3.0	10.8%
5	4.0	10.4%
6	5.0	10.0%
7	6.0	6.0%
8	5.0	4.9%
9	2.0	15.4%
10	6.0	5.2%

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 Figure 3.15.

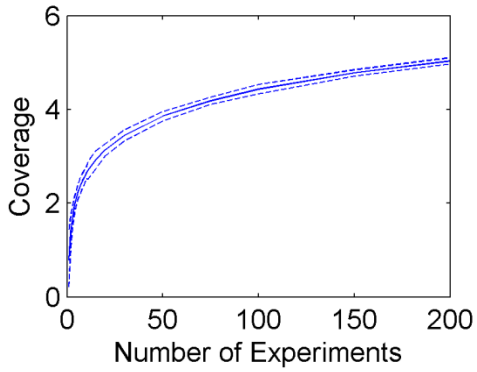

Fig. 3.15 Average Coverage (solid line) ± 3 standard deviations (dashed lines)

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 Figure 3.15 to help determine when the gains in coverage do not justify the cost of further experiments, thus 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.

3.7 Conclusions

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.

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.

References

[1] Draper D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society B* 1995; **57(1):** 45-97.

[2] Kennedy MC, O'Hagan A. Bayesian Calibration of Computer Models. J*ournal of the Royal Statistical Society* 2001; **63:** 425-464.

[3] Hemez F, Atamturktur S, Unal C. Defining predictive maturity for validated numerical simulations. *Computers and Structures Journal* 2010; **88:** 497-505.

[4] Higdon D, Gattiker J, Williams B, Rightley M. Computer model calibration using high-dimensional output. *Journal American Statistical Association* 2008; **103(482):** 570-83.

[5] Higdon D, Nakhleh C, Gattiker J, Williams B. A Bayesian calibration approach to the thermal problem. *Computer Methods in Applied Mechanics and Engineering* 2008; **197(29-32):** 2431-2441.

[6] Farajpour I, Atamturktur S. Error and Uncertainty Analysis of Inexact and Imprecise Computer Models. *Journal of Computing in Civil Engineering* 2013; **27(4):** 407-418.

[7] Atamturktur S, Hemez F, Williams B, Tome C, Unal C. A forecasting metric for predictive modeling. *Computers and Structures* 2011; **89(23,24):** 2377-2387.

[8] Atamturktur S, Hemez F, Unal C, William B. Predictive Maturity of Computer Models Using Functional and Multivariate Output. *In Proceedings of the 27th SEM International Modal Analysis Conference*, Orlando, FL. 2009.

[9] Stull CJ, Hemez F, Williams B, Unal C, Rogers ML. An improved description of predictive maturity for verification and validation activities. *Los Alamos National Laboratory Technical Report* 2011; LA-UR-11-05659.

[10] Johnson ME, Moore LM. Ylvisaker D, Minimax and Maximin Distance Designs. *Journal of Statistical Planning and Inference* 1990; **26(2):** 131-148.

[11] Williams BJ, Loeppky JK, Moore LM, Macklem MS,. Batch sequential design to achieve predictive maturity with calibrated computer models. *Reliability Engineering System Safety* 2011; **96:** 1208-1219.

[12] Shao T. Toward a structured approach to simulation-based engineering design under uncertainty. University of Massachusetts Amherst), *ProQuest Dissertations and Theses* 2007; 265. Retrieved from http://search.proquest.com/docview/304846542?accountid=6167. (304846542).

[13] Sacks J, Welch W, Mitchell T, Wynn H. Designs and analysis of computer experiments. *Statistical Science* 1989; **4:** 409-435.

[14] Fryer RJ, Shepherd JG. Models of codend size selection. *J. Northw. Fish. Sci.* 1996; **19:** 51-58.

[15] Logan RW, Nitta CK, Chidester SK. Risk Reduction as the Product of Model Assessed Reliability, Confidence, and Consequence. *Lawrence Livermore National Laboratory Technical Report* November 2003; UCRL-AR-200703.

[16] Oberkampf WL, Pilch M, Trucano TG. Predictive capability maturity model for computational modeling and simulation. *Sandia National Laboratory Technical Report* 2007; SAND-2007-5948.

[17] Montgomery DC. *Design and analysis of experiments* (5^{th} edn). John Wiley & Sons: New York, NY, 1997; 416-417.

[18] O'Hagan A, Oakley JE. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety* 2004; **85:** 239-248

[19] Hemez F, Atamturktur S, Unal C. Defining predictive maturity for validated numerical simulations. In *Proceedings of the IMAC-XXVII*, Orlando, FL, USA. February 9-12, 2009.

[20] Cameron, Peter J. (1994), Combinatorics: Topics, Techniques, Algorithms, Cambridge University Press.

[21] Atamturktur S, Hegenderfer J, Williams B, Egeberg M, Lebensohn R, Unal C. A Resource Allocation Framework for Experiment-Based Validation of Numerical Models. *Mechanics of Advanced Materials and Structures* 2014; DOI 10.1080/15376494.2013.828819.

[22] Stull C, Williams B, Unal C. Assessing the Predictive capability of the LIFEIV nuclear fuel performance code using sequential calibration. *Los Alamos National Laboratory Technical Report* 2012; LA-UR-12-22712.

[23] Lebensohn RA, Hartley CS, Tomé CN, Castelnau O. Modeling the mechanical response of polycrystals deforming by climb and glide. *Phil Mag* 2010; **90(5):** 567-83.

[24] Atamturktur S, Hegenderfer, J, and Williams B. (accepted, in print), A Selection Criterion Based on Exploration-Exploitation Approach for Batch Sequential Design. *Journal of Engineering Mechanics (ASCE);* (accepted, in print).

[25] Atamturktur S, Williams B, Egeberg M, and Unal C. Batch Sequential Design of Optimal Experiments for Improved Predictive Maturity in Physics-Based Modeling. *Structural and Multidisciplinary Optimization (Springer)* 2013; **48(3):** 549-569.

CHAPTER FOUR

CONCLUSIONS

The two journal articles presented in this thesis aim to reduce the experimental resources required to reach predictive maturity of complex numerical models.

In chapter two, several batch sequential design (BSD) selection criteria are applied to the Visco Plastic Self-Consistent material plasticity model. The Predictive Maturity Index (PMI), influenced herein by discrepancy bias and coverage, is used to evaluate the performance of each selection criteria. Index-based selection criteria such as expected improvement for predictive stability (EIPS) are observed to favor exploitation of variance and bias, making these criteria more favorable when discrepancy is of high importance such as when model fidelity is critical. Meanwhile, distance-based selection criteria such as Euclidean distance (EDIST) favor exploration of the operational domain and are therefore more favorable when coverage of the operational domain is of high importance such as when the underlying physics differ between regions of the operational domain. An effective technique is to use a mixed approach in which a distance-based selection criterion is initially used to provide sufficient coverage of the operational domain, next an index-based selection criterion is used to achieve a desired discrepancy bias. This study provides guidance to analysts when selecting a selection criterion to most efficiently improve the predictive maturity of a given numerical model.

In chapter three, four characteristics of an exemplar coverage metric are identified. Coverage should (i) improve if a new experiment is added at untested settings, (ii) favor a more uniform distribution of experiments over a clustered arrangement, (iii)

distinguish between interpolation and extrapolation, and (iv) be objective. The coverage metrics from the literature are found to be unsuitable for all four criteria, thus a new coverage metric is proposed. The proposed coverage metric is found to exhibit satisfactory performance in all four criteria and shows aptitude when applied to high-dimensional operational domains. Through the refined coverage metric, this study helps decision makers quantify coverage, an important component in determining when a numerical model has reached predictive maturity. This may save unnecessary experimental resources from being used after predictive maturity has been achieved.

Future work may build from this thesis to further improve BSD techniques. In chapter two, the coverage metric used in the PMI is shown in chapter three to have several shortcomings. A possible investigation would be to repeat the study in chapter two while using the coverage metric proposed in chapter three in the formulation of the PMI. This would provide a more refined evaluation of predictive maturity of the numerical models and allow a more accurate comparison between selection criteria. Also, the coverage metric proposed in chapter three could be used as a BSD selection criterion alone. Furthermore, the coverage metric could be combined with an index-based selection criterion to simultaneously explore the operational domain and exploit the discrepancy bias.

APPENDIX

In chapter two, the findings are presented considering EIPS as representative of index-based criteria. The appendix summarizes the findings for the three other index-based criteria: EIGF, ENT and IMSE. The PMIs along with coverage and scaled discrepancy attributes are given for the exact model solution. Figures A.1a and A.1c illustrate the PMI for the EIGF criterion when the experimental uncertainty is 0.1% and 5%, respectively. For 0.1% experimental uncertainty, EIGF successfully provides a converging PMI value (see Figure A.1a) and the improvement in discrepancy is similar to that obtained by EIPS (see Figure A.2b). The coverage reaches the range of 83%-92%. For the 5% experimental uncertainty, the PMI monotonically increases and the discrepancy is reduced below 20% after the 10^{th} batch (see Figure A.1d). The coverage range after the 10^{th} batch is between 76% and 86%.

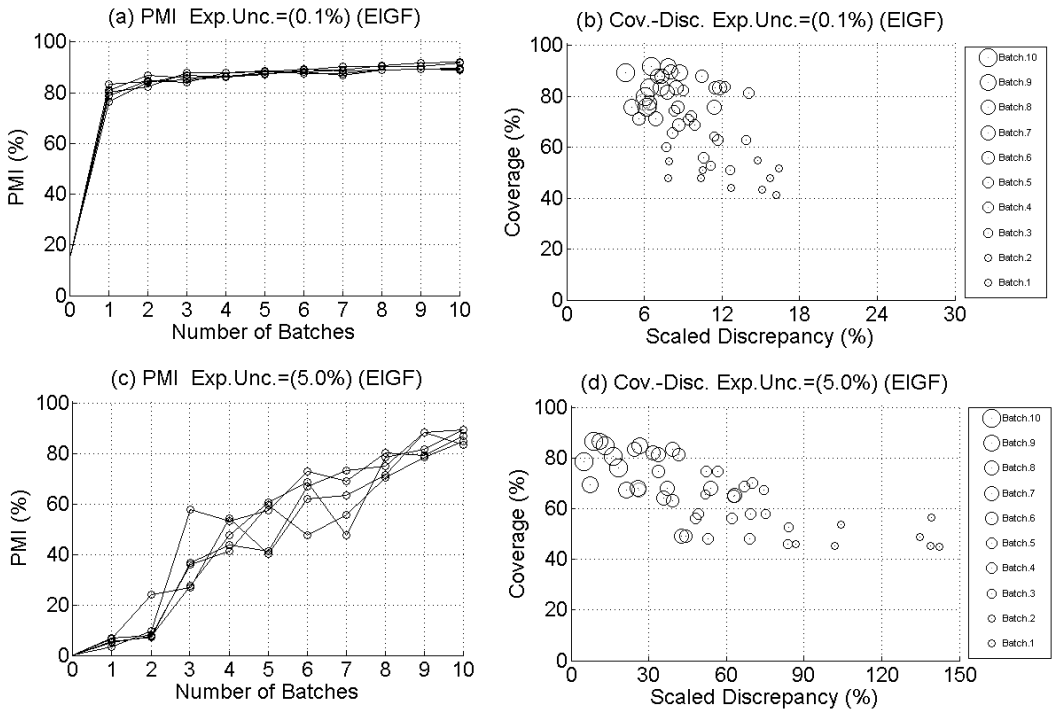

Fig. A.1 Exact model by EIGF: (a) PMI for 0.1% experimental uncertainty: (b) Normalized discrepancy vs. coverage attributes for 0.1% experimental uncertainty, (c) PMI for 5% experimental uncertainty, (d) Normalized discrepancy vs. coverage attributes for 5% experimental uncertainty

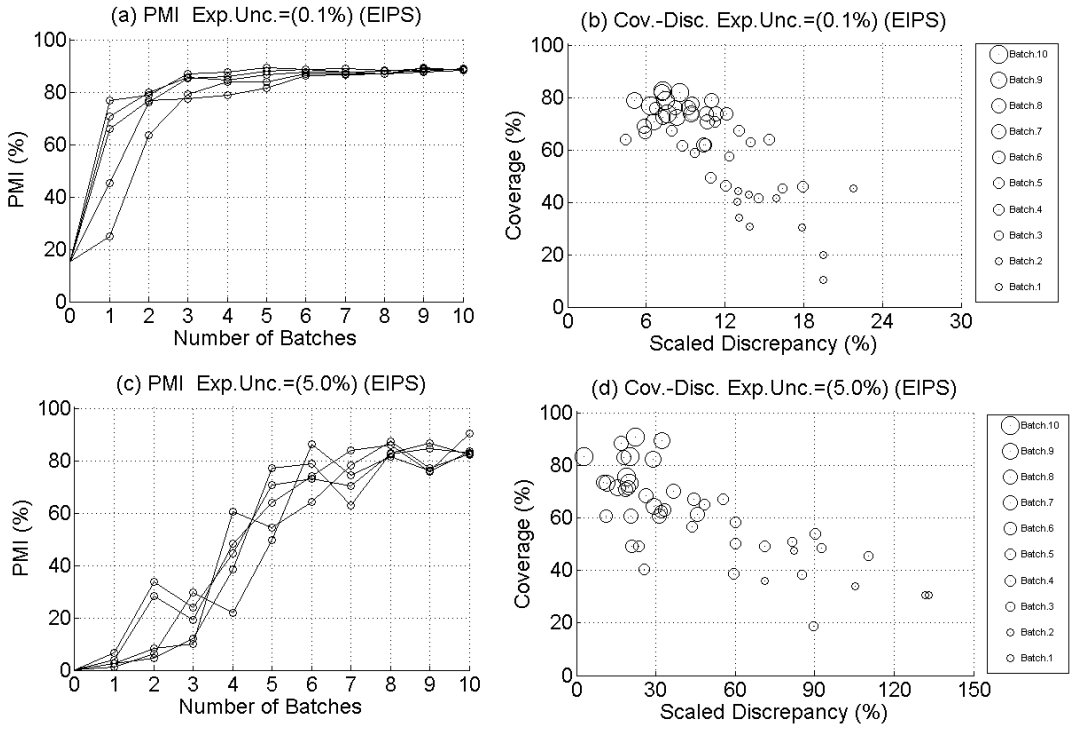

Fig. A.2 Exact model by EIPS: (a) PMI for 0.1% experimental uncertainty: (b) Normalized discrepancy vs. coverage attributes for 0.1% experimental uncertainty, (c) PMI for 5% experimental uncertainty, (d) Normalized discrepancy vs. coverage attributes for 5% experimental uncertainty

Figures A.3a and A.3c illustrate the improvement in PMI through batches using the index- based criterion ENT for 0.1% and 5% experimental uncertainty. The specific characteristic of this criterion is the high coverage when compared to the other index-based criteria. The ENT criterion is observed to select experimental settings at the boundary of the domain in early batches regardless of the level of experimental uncertainty (see Figures A.3b and A.3d). For 0.1% experimental uncertainty, the improvement in discrepancy by ENT, however, is not as high as that of EIPS and EIGF

(above 12%). For 5% experimental uncertainty, after the 10^{th} batch discrepancy is reduced below 30% and PMI has increased to over 95%.

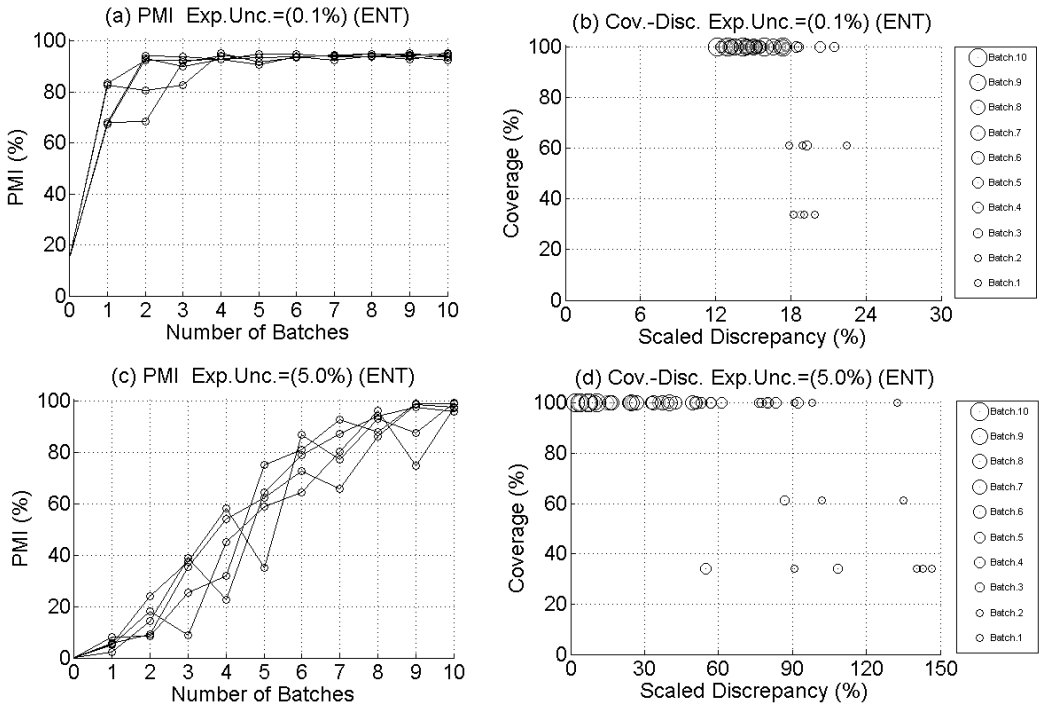

Fig. A.3 Exact model by ENT: (a) PMI for 0.1% experimental uncertainty: (b) Normalized discrepancy vs. coverage attributes for 0.1% experimental uncertainty, (c) PMI for 5% experimental uncertainty, (d) Normalized discrepancy vs. coverage attributes for 5% experimental uncertainty

Figures A.4a and A.4c illustrate the PMI for the IMSE criterion when the experimental uncertainty is 0.1% and 5%, respectively. PMI values converge to a range between 85% and 90%. The cloud of normalized discrepancy and coverage is centered between a range from 10% to 12% for discrepancy and 60%-85% for coverage (see Figure A4b). However, the concentration of the cloud in EIPS in Figure A.2b is denser

between 6%-10% for discrepancy and 75%- 82% for coverage. For 5% experimental uncertainty, the coverage of IMSE (between 65%-80%) is lower compared to EIPS after the 10^{th} batch.

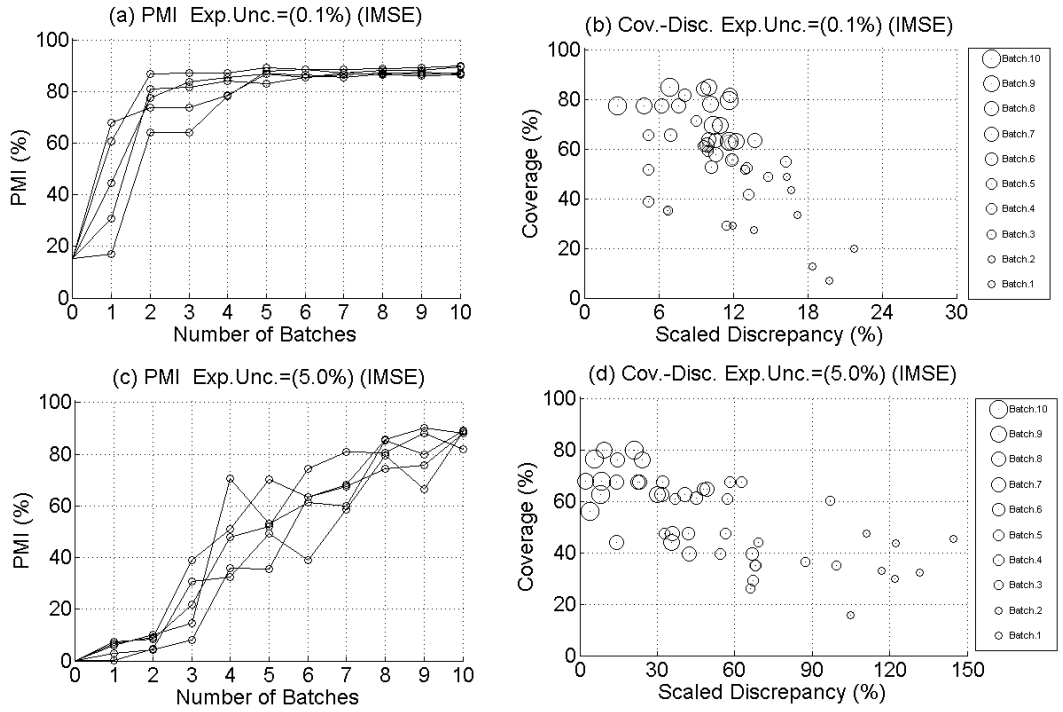

Fig. A.4 Exact model by IMSE: (a) PMI for 0.1% experimental uncertainty: (b) Normalized discrepancy vs. coverage attributes for 0.1% experimental uncertainty, (c) PMI for 5% experimental uncertainty, (d) Normalized discrepancy vs. coverage attributes for 5% experimental uncertainty