Fusing Models for Microstructure Performance: A Correlation Exploiting Multi-Information Source Approach

Seyede Fatemeh Ghoreishi,
Graduate Research Assistant
Dept. of Mechanical Eng.
Texas A&M University
College Station, TX 77843
Email: f.ghoreishi88@tamu.edu

Abhilash Molkeri
Graduate Research Assistant
Dept. of Materials Science and Eng.
Texas A&M University
College Station, TX 77843
Email: abhilashmolkeri@tamu.edu

Ankit Srivastava
Assistant Professor
Dept. of Materials Science and Eng.
Texas A&M University
College Station, TX 77843
Email: ankit.sri@tamu.edu

Raymundo Arroyave
Professor
Dept. of Materials Science and Eng.
Texas A&M University
College Station, TX 77843
Email: rarroyave@tamu.edu

Douglas Allaire*
Assistant Professor
Dept. of Mechanical Eng.
Texas A&M University
College Station, TX 77843
Email: dallaire@tamu.edu

Integrated Computational Materials Engineering (ICME) has been put forward as the *aspirational framework for the acceleration of materials development, at least within the metallic alloys realm. Often times, however, the "I" in ICME is difficult to realize, particularly because of the perceived lack of explicit interfaces between computational (or experimental) tools. This paper seeks to address this problem by implementing an information fusion framework that exploits correlations among sources and between each individual source and the 'ground truth'. The framework is demonstrated on the prediction (via a set of different reduced-order models under different assumptions) of the mechanical response of composite microstructures consisting of two phases with very different stress-strain responses. It is shown how the proposed framework is capable of producing fused models that are much more accurate, relative to the ground truth, than each of the individual sources taken independently, demonstrating its applicability to the* integration *of ICME tools.*

*Address all correspondence to this author.

1 Introduction

Recently, awareness of the need to accelerate the materials discovery, development, and deployment cycle has increased significantly. Thus, novel approaches have been sought to address some of the shortcomings of conventional frameworks for materials design. In the context of alloy design, for example, Olson's pioneering work [1, 2] framed the general problem of materials design as one where materials are considered as complex hierarchical systems in which connections along the processing-structure-properties-performance paradigm (PSPP) are exploited to achieve optimal performance. In the past decade, Integrated Computational Materials Engineering (ICME) has emerged as an *aspirational* framework that *"integrates science and engineering as well as the results of theory, experiments, and simulations into computational tools that can be used directly in engineering of new products or manufacturing processes"* [3]. ICME thus seeks to exploit PSPP relationships, arrived at through the *integration* of diverse computational tools and experimental data, in order to accelerate materials design. We use here the term "aspirational" because while ICME prescribes the *integration* of multiple computational tools (or information sources) as a necessary step towards the

acceleration of the materials development cycle, it remains agnostic with regards to the means by which such integration can be done in practice.

In fact, a very recent report [4] suggests that one of the key roadblocks to realizing the goals of ICME is the lack of effective strategies to integrate models capable of describing the behavior of materials at multiple scales within a single unified framework. While the integration of models (or information sources) across multiple-scales has already been recognized as important, an associated problem remains largely overlooked: namely, *the integration of multiple information sources describing the behavior of materials at a given scale*. To be more specific, we claim that there are multiple information sources that, while differing in their fidelity, underlying assumptions and structure, describe the same underling connections along the PSPP paradigm. Their common connection to a particular linkage along the PSPP relationship in a given material thus means that such sources must have a certain degree of correlation between them. At the same time, it is to be expected that different sources will have a different level of fidelity (or distance from the *ground truth*), uncertainty and cost.

In this work, we propose a framework that addresses the problem of integrating multiple information sources describing the behavior of materials at a given scale. We put forward a statistical framework that places multiple sources of information about PSPP connections on an equal basis. The framework explicitly exploits the (statistical) correlations that *must* exist between the different sources as all of them have an underlying connection to the *ground truth* that they seek to describe. We deploy our *correlation-exploiting information fusion* approach to the integration of several models that connect microstructural features to the mechanical properties of ductile multi-phase materials. More specifically, we demonstrate the framework by predicting strength normalized strain-hardening rate (an important manufacturing-related attribute) of ductile dual-phase (consisting of two ductile phases: a soft and a hard phase) materials by exploiting three approximate or reduced-order models. Each of the (reduced-order) models is constructed under different assumptions regarding the partitioning of the macroscopic stress/strain among the microstructural constituents.

The remainder of the paper is as follows: First, we proceed to motivate the work in the context of microstructure-sensitive modeling and design of dual-phase ductile materials, e.g., advanced high strength steels. These advanced structural alloys are one of the most technologically sought after materials used in lightweight structural applications, such as automotive manufacturing. Next, we describe a microstructure-based finite element model for predicting the stress-strain response of ductile dual-phase materials. The microstructure-based finite element model is considered to be the ground truth, while three different reduced-order models are used as the sources of information to fuse. After the description of the different models for the mechanical response of the dual-phase microstructure, the proposed information fusion approach is described. Specifically, it is shown how we exploit statistical correlations between information sources to arrive at a fused model with significantly better fidelity with respect to the ground truth than any individual source (reduced-order model). The statistical correlations among the multiple sources are elicited through the so-called *reification* method.

In a companion paper, we show how this information fusion approach can be integrated with a Bayesian sequential design optimization framework to arrive at optimal microstructures that maximize the strength normalized strain-hardening rate by identifying and exploiting optimal sequential queries of different information sources.

2 Mechanical Behavior of Dual-Phase Microstructures

Ductile dual-phase materials, in particular dual-phase advanced high strength steels, primarily consist of hard martensite (phase) islands dispersed in a soft ferrite (phase) matrix [5]. Both these phases of dual-phase steels undergo non-linear elastic-plastic deformation with strikingly different strength levels and strain hardenability [6,7]. The overall mechanical properties, such as strength and strain-hardening of dual-phase steels, are thus determined partly by the mechanical properties of the phases themselves, and partly by microstructural features, such as the volume fraction of the phases. In principle, the properties of the phases and the microstructural features can be tuned and optimized to achieve a particular performance matrix. However, without the ability to predict the overall mechanical properties of dual-phase materials, reliance of conventional material design processes on extensive trial-and-error experimentation is often quite time consuming. The overall mechanical properties of dual-phase materials can be predicted by high-fidelity microstructure-based finite element models but these come at considerable computational cost that precludes their use to carry out search in the microstructure space for regions of optimal performance. There exist low-fidelity reduced-order models based on isostrain [8], isostress [9] or isowork [10] approximations of the partitioning of the macroscopic strain, stress or work, respectively, among the microstructural constituents. However, these models only approximate the bounds on the mechanical properties of the highly non-linear dual-phase materials. In this study, we carryout microstructure-based finite element modeling of dual-phase microstructures and consider it to be the ground truth. The three lower fidelity reduced-order models are used as sources of information that we seek to fuse to achieve higher fidelity prediction capability. For this, we exploit statistical correlations between the reduced-order models to construct a fused model that is a significant improvement over any reduced-order model for approximating ground truth.

2.1 Microstructure-based Finite Element Model

We employ microstructure-based finite element modeling to calculate the overall mechanical response of the dual-phase microstructures. To this end, we generate 3D representative volume elements (RVEs) of the dual-phase microstructures following the procedure described in detail in

[11]. The RVE is a composite dual-phase microstructure with two discretely modeled phases: a soft phase representative of ferrite and a hard phase representative of martensite. Two realizations of a 3D RVE of a dual-phase microstructure with 54.22% volume fraction of the hard phase are shown in Fig. 1. The RVEs shown in the figure consist of 27,000 C3D8 brick elements of the ABAQUS/standard element library [12], and have a dimension of $100\mu\text{m} \times 100\mu\text{m} \times 100\mu\text{m}$. Note, the volume fraction of the phases in the RVE is always an integral multiple of the volume of one element, which is $3.7 \times 10^{-5}\mu\text{m}^3$. The RVEs are subjected to fully periodic boundary conditions on all six faces and monotonically increasing uniaxial tensile deformation. This allows for the calculation of the overall uniaxial tensile stress-strain response of the dual-phase microstructure.

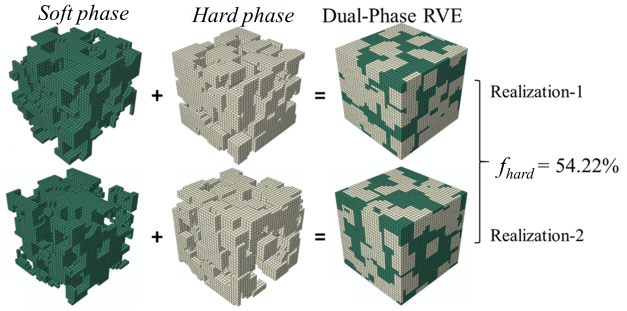

Fig. 1. Two realizations of the representative volume element (RVE) of a dual-phase microstructure consisting of 54.22% hard (martensite) phase and the rest of soft (ferrite) phase. The realizations refers to different distributions of the hard phase particles in the RVE with a fixed phase volume fractions.

In the finite element calculations both the phases are assumed to follow isotropic elastic-plastic stress-strain response. The Young's modulus of both the phases is taken to be, $E = 200$GPa and the Poisson's ratio is taken to be, $\nu = 0.3$. The plastic response of both the phases are modeled using the Ludwik power law constitutive relation,

$$\tau = \tau_o + K(\varepsilon_{pl})^n, \tag{1}$$

where τ is the flow stress, ε_{pl} is the equivalent plastic strain, τ_o is the yield strength, K is the strengthening coefficient, and n is the strain hardening exponent of the phase. The values of τ_o, K, and n for the constituent phases are given in Tab. 1. The parameters are chosen to represent lower initial yield strength of the ferrite (soft) phase compared to the martensite (hard) phase, and higher strain hardenability of the ferrite phase compared to the martensite phase.

2.2 Reduced-order Models

We use three low-fidelity reduced-order models as three sources of information to demonstrate our fusion process in Section 4. These three reduced-order models are:

Table 1. Constitutive parameters for the constituent phases

Constituent Phase	τ_o **[MPa]**	K **[MPa]**	n
Soft (ferrite)	300	2200	0.5
Hard (martensite)	1500	450	0.06

i. The Voigt/Taylor isostrain model, where the basic assumption is that the strain field is uniform among the constituent phases [13]. The effective stress is expressed in terms of the local stress average with respect to both phases weighted by their respective volume fractions. That is, for this model we have

$$\varepsilon_{pl}^T = \varepsilon_{pl}^h = \varepsilon_{pl}^s, \quad \tau^T = f_h\tau^h + (1-f_h)\tau^s. \tag{2}$$

ii. The Reuss/Sachs isostress model, where the basic assumptions is that the stresses among the phases are homogeneous [13]. The effective strain is calculated in terms of the average of the strains in each phase weighted by their respective volume fractions. Thus, for this model we have

$$\tau^T = \tau^h = \tau^s, \quad \varepsilon_{pl}^T = f_h\varepsilon_{pl}^h + (1-f_h)\varepsilon_{pl}^s. \tag{3}$$

iii. The isowork model, which is an approximation based on the principle that work of deformation is equally distributed in all the constituent phases in the dual-phase microstructure at any strain level. That is,

$$\tau^h\varepsilon_{pl}^h = \tau^s\varepsilon_{pl}^s. \tag{4}$$

In Eqs. 2, 3 and 4, ε_{pl}^T is the overall plastic strain, ε_{pl}^h is the plastic strain in the hard (martensite) phase, ε_{pl}^s is the plastic strain in the soft (ferrite) phase, τ^T is the overall stress, τ^h is the stress in the hard (martensite) phase, τ^s is the stress in the soft (ferrite) phase, and f_h is the volume fraction of the hard phase in the microstructure. The stress-strain relations, $\tau = f(\varepsilon_{pl})$, of both phases are assumed to follow, Eq. 1, with the values of the parameters given in Tab. 1.

2.3 Demonstration of Modeling Capabilities

Predictions of the stress-strain relationships for the three low-fidelity reduced-order models are shown in Fig. 2. For comparison, the high-fidelity microstructure-based finite element model prediction for a dual-phase microstructure with 25% volume fraction of the hard phase is also shown on the figure. As can be seen from the figure, compared to the finite element predictions, the isostress model significantly under predicts the stress values at nearly all plastic strain levels. In

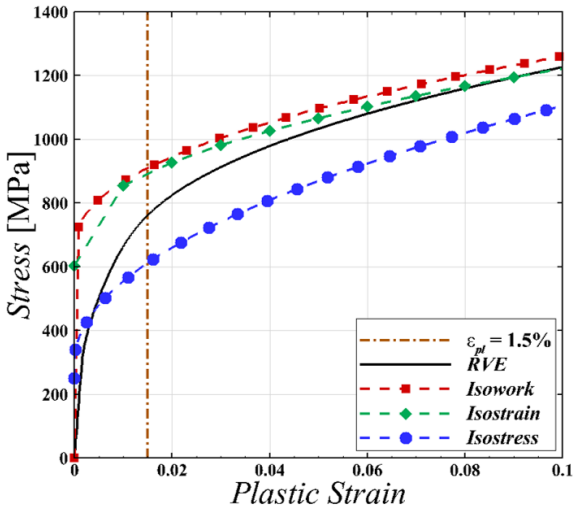

Fig. 2. Comparison of the predictions of the three reduced-order models with the microstructure-based finite element model prediction for a dual-phase microstructure with 25% volume fraction of the hard phase.

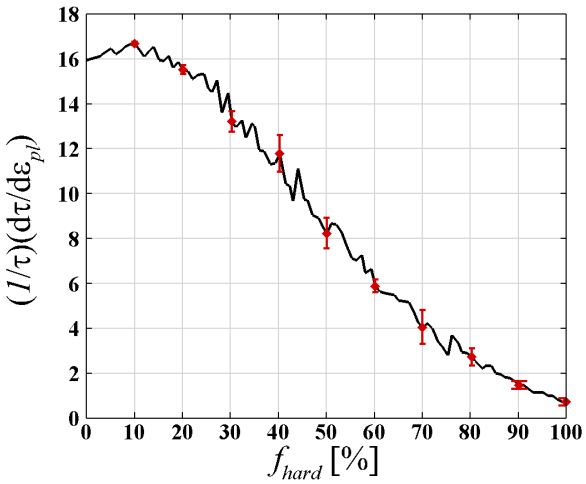

Fig. 3. The variation of the strength normalized strain-hardening rate, $(1/\tau)(d\tau/d\varepsilon_{pl})$ at $\varepsilon_{pl} = 1.5\%$, with the volume fraction of the hard phase, f_{hard}, as estimated from the microstructure-based finite element calculations. The errors bars correspond to seven realizations of the dual-phase microstructure with the same f_{hard} value.

contrast, the isostrain and isowork models at low strain levels over predict the stress values. However, at large strain levels their predictions are comparable with the finite element predictions. In addition, none of the reduced-order models are able to correctly predict the overall strain-hardening response of the dual-phase microstructure. At a plastic strain level of 1.5%, the strain-hardening rate, $d\tau/d\varepsilon_{pl}$, is predicted to be approximately 1.04×10^4MPa by the isostress model, approximately 0.70×10^4MPa by the isostrain model, and approximately 0.77×10^4MPa by the isowork model. These predictions are all significantly lower than the value predicted by the finite element model, which is approximately 1.20×10^4MPa.

The flow strength and the strain hardenability of a material are two very important mechanical properties. These two properties can be coupled by introducing a strength normalized strain-hardening rate, given as $(1/\tau)(d\tau/d\varepsilon_{pl})$. This quantity is an important manufacturing-related attribute as it dictates the ductility and formability of the material. The variation of $(1/\tau)(d\tau/d\varepsilon_{pl})$ with the volume fraction of the hard phase, f_{hard}, estimated at the strain level, $\varepsilon_{pl} = 1.5\%$, from the microstructure-based finite element calculations, is shown in Fig. 3. As shown in the figure, the value of $(1/\tau)(d\tau/d\varepsilon_{pl})$ at $\varepsilon_{pl} = 1.5\%$, first increases with increasing volume fraction of the hard phase and then starts to decrease. In general a higher value of the quantity $(1/\tau)(d\tau/d\varepsilon_{pl})$ denotes higher formability of the material. Note, in Fig. 3, variation of $(1/\tau)(d\tau/d\varepsilon_{pl})$ with f_{hard} exhibits local perturbations. These perturbations are due to the fact that there are several possible realizations of a RVE of a dual-phase microstructure with a fixed volume fraction of the hard phase. This concept was shown previously for two realizations in Fig. 1. These different realizations result in slightly different values of $(1/\tau)(d\tau/d\varepsilon_{pl})$ for a fixed f_{hard} value. For a few selected volume fractions of the hard phase, seven realizations of the dual-phase microstructures were generated and their mechanical responses were calculated. The standard error on the values of $(1/\tau)(d\tau/d\varepsilon_{pl})$ at $\varepsilon_{pl} = 1.5\%$ due to different realizations of the dual-phase microstructure with fixed volume fraction of the hard phase are also shown in Fig. 3 as error bars.

3 Correlation Exploiting Multi-Information Source Fusion

As is often the case when modeling a physical system, the mechanical behavior of dual-phase steels can be analyzed with many different information sources. In this particular case, these information sources are the isostrain, isostress, isowork, and finite element model based on RVEs for each phase, which is considered ground truth. As shown in Section 4, each of these models provides significantly different estimates of the quantity of interest, $(1/\tau)(d\tau/d\varepsilon_{pl})$, over the domain of the volume fraction of the hard phase.

Working under the hypothesis that each information source brings to bear some useful information regarding a quantity of interest, we seek to systematically fuse available information from each source. Unlike traditional multifidelity methods [14–21], the multi-information source fusion method employed here does not assume a hierarchy of information sources with the goal of efficiently approximating the highest fidelity source. The goal here is to best approximate a ground truth quantity of interest by leveraging all available information. This is achieved by learning the correlations between the discrepancies of the available information sources, which results in the ability to mitigate information source bias and avoid overconfidence that arises from the reuse of dependent information.

In the following subsections, we setup our problem of multi-information source fusion by introducing our use of Gaussian processes as intermediate surrogates and of model

reification as an approach to learning correlations. We then present our correlation exploiting multi-information source methodology. The result is a methodology capable of creating a new fused information source that incorporates and exploits all available information. As shown in Section 4, these fused information sources outperform the individual information sources in terms of mean square error and information fidelity as measured by the Kullback-Liebler divergence from ground truth.

3.1 Problem Setup

To enable the use of correlation exploiting fusion, we create intermediate surrogate models using Gaussian process regression. These intermediate surrogates are constructed from data collected previously for each available source. The use of Gaussian processes provides for normally distributed estimates of a given quantity of interest from a given information source at a particular input configuration. The fusion of information from such distributions is straightforward once correlation is known. We take up the task of learning correlation for fusion in the following subsection. Here we setup our information fusion problem by first constructing the Gaussian process intermediate surrogates. Our use of Gaussian process regression for intermediate surrogate modeling is common in the literature (see, e.g, [22–26]).

We denote by $\bar{f}_i(\mathbf{x})$, $i \in \{1,2,\ldots,S\}$, the available information sources that can be used to estimate the quantity of interest, $f(\mathbf{x})$, at design point $\mathbf{x}$. In the application case considered in this work, these sources are the isostrain, isostress, and isowork models, which can be queried at different values of the percentage volume fraction of martensite. We assume we have some available evaluations from the information sources, and in order to predict the output of each information source at locations that have yet to be queried, an intermediate surrogate using Gaussian process regression [27] is constructed for each information source. These surrogates are denoted by $f_{GP,i}(\mathbf{x})$. To construct these Gaussian processes, we consider their prior distributions as

$$f_{GP,i}(\mathbf{x}) \sim \mathcal{GP}\left(\mathbf{0}, k_i(\mathbf{x},\mathbf{x})\right), \tag{5}$$

where $k_i(\mathbf{x},\mathbf{x})$ is a real-valued kernel function over the input space. Without loss of generality, we employ the commonly used squared exponential covariance function for the kernel function, which is given as

$$k_i(\mathbf{x},\mathbf{x}') = \sigma_s^2 \exp\left(-\sum_{h=1}^{d} \frac{(x_h - x_h')^2}{2l_h^2}\right), \tag{6}$$

where d is the dimension of the input space, σ_s^2 is the signal variance, and l_h, where $h = 1,2,\ldots,d$, is the characteristic length-scale that indicates the strength of the correlation between the points within dimension h. The maximum likelihood method can be used to estimate the parameters σ_s^2 and l_h associated with each information source. For information source i, we assume that N_i evaluations, denoted by $\{\mathbf{X}_{N_i}, \mathbf{y}_{N_i}\}$, where $\mathbf{X}_{N_i} = (\mathbf{x}_{1,i},\ldots,\mathbf{x}_{N_i,i})$ represents the N_i input samples to information source i and $\mathbf{y}_{N_i}$ represents the corresponding outputs from information source i, are available. Then the posterior distributions of $f_{GP,i}(\mathbf{x})$ at any point $\mathbf{x}$ in the input space can be constructed as

$$f_{GP,i}(\mathbf{x}) \mid \mathbf{X}_{N_i}, \mathbf{y}_{N_i} \sim \mathcal{N}\left(\mu_i(\mathbf{x}), \sigma_{GP,i}^2(\mathbf{x})\right), \tag{7}$$

where

$$\mu_i(\mathbf{x}) = K_i(\mathbf{X}_{N_i},\mathbf{x})^T [K_i(\mathbf{X}_{N_i},\mathbf{X}_{N_i}) + \sigma_{n,i}^2 I]^{-1} \mathbf{y}_{N_i}, \tag{8}$$

$$\begin{aligned} &\sigma_{GP,i}^2(\mathbf{x}) = k_i(\mathbf{x},\mathbf{x}) - \\ &K_i(\mathbf{X}_{N_i},\mathbf{x})^T [K_i(\mathbf{X}_{N_i},\mathbf{X}_{N_i}) + \sigma_{n,i}^2 I]^{-1} K_i(\mathbf{X}_{N_i},\mathbf{x}), \end{aligned} \tag{9}$$

where $K_i(\mathbf{X}_{N_i},\mathbf{X}_{N_i})$ is the $N_i \times N_i$ matrix whose m,n entry is $k_i(\mathbf{x}_{m,i},\mathbf{x}_{n,i})$, and $K_i(\mathbf{X}_{N_i},\mathbf{x})$ is the $N_i \times 1$ vector whose m^{th} entry is $k_i(\mathbf{x}_{m,i},\mathbf{x})$ for information source i. Note that the term $\sigma_{n,i}^2$ can be used to model observation error for information sources based on experiments and can also be used to avoid numerical ill-conditioning.

As mentioned previously, each of the information sources has an associated uncertainty with respect to the ground truth. We characterize this uncertainty by assigning a probability distribution to the output of each individual information source over the input space. These distributions can be quantified from available ground truth data or expert opinion. We further add the uncertainty of each information source to its Gaussian process surrogate. Specifically, we approximate each information source, with quantified uncertainty, as

$$f_i(\mathbf{x}) = f_{GP,i}(\mathbf{x}) + \delta_{f,i}(\mathbf{x}), \tag{10}$$

where the subscript i again indicates the information source, $f_{GP,i}(\mathbf{x})$ has been estimated via available data for $\bar{f}_i(\mathbf{x})$, and $\delta_{f,i}(\mathbf{x}) \sim \mathcal{N}(0, \sigma_{f,i}^2(\mathbf{x}))$ is the estimated discrepancy of information source i. It is important to note that these discrepancies can vary over the input space. We also note that we set the mean of the discrepancy to be zero to enforce a belief that the discrepancy does not have a known direction. If such a direction is known, presumably it would be built into the model itself. At this point, each of the S available information sources for estimating the ground truth quantity of interest can be written as

$$f_i(\mathbf{x}) = \mu_i(\mathbf{x}) + \delta_i(\mathbf{x}), \tag{11}$$

where $\delta_i(\mathbf{x}) = \delta_{GP,i}(\mathbf{x}) + \delta_{f,i}(\mathbf{x})$, is the total uncertainty, which includes both the uncertainty associated with the

Gaussian process surrogate and the uncertainty associated with the fidelity of the information source. Fig. 4 shows a depiction of total uncertainty, where the darker shaded region represents the uncertainty associated with the Gaussian process and the lighter shaded region represents additional uncertainty added to account for the discrepancy between the information source and certain known ground truth information. Here, ground truth is known at only the input

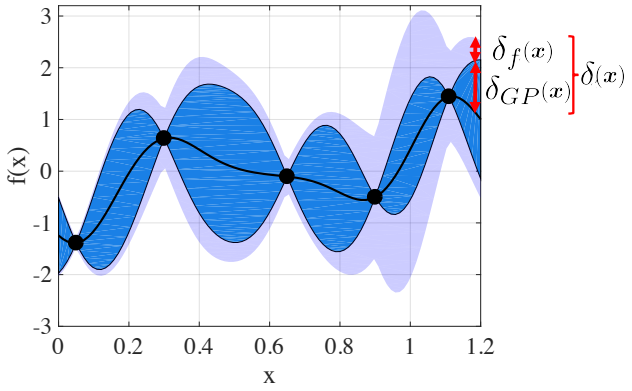

Fig. 4. A depiction of total uncertainty, which includes both the uncertainty associated with the Gaussian process and uncertainty associated with the fidelity of the information source with respect to available ground truth data.

configurations of the five points shown on the figure.

3.2 Fusion of Information Sources

There are many techniques in use for fusing information from multiple sources of information. Among these are approaches such as Bayesian modeling averaging [28–33], the use of adjustment factors [34–37], covariance intersection methods [38, 39], and fusion under known correlation [40–42]. As noted previously, we hypothesize that every information source contains useful information regarding the ground truth quantity of interest. Thus, as more information sources are incorporated into a fusion process, we expect the variance of quantity of interest estimates to decrease. This is not necessarily the case for techniques such as Bayesian model averaging and adjustment factors approaches. For the case of unknown correlations between information sources, recourse must be made to conservative methods, such as covariance intersection. This method fuses information by assuming the worst case correlation information. Thus, there is much to gain from estimating correlation between information sources and incorporating these learned correlations in the fusion process. In the following paragraphs we describe how to fuse information when correlation is known under normal distributions. We then describe how to estimate these correlations. We describe some of the features of this fusion approach in Section 4, with the support of the application case results.

Under the case of known correlations between the model deviations of information sources, the fused mean and variance are shown to be [43]

$$\mathbb{E}[\hat{f}(\mathbf{x})] = \frac{\mathbf{e}^\top \tilde{\Sigma}(\mathbf{x})^{-1}\boldsymbol{\mu}(\mathbf{x})}{\mathbf{e}^\top \tilde{\Sigma}(\mathbf{x})^{-1}\mathbf{e}}, \tag{12}$$

$$\mathrm{Var}\left(\hat{f}(\mathbf{x})\right) = \frac{1}{\mathbf{e}^\top \tilde{\Sigma}(\mathbf{x})^{-1}\mathbf{e}}, \tag{13}$$

where $\mathbf{e} = [1,\ldots,1]^\top$, $\boldsymbol{\mu}(\mathbf{x}) = [\mu_1(\mathbf{x}),\ldots,\mu_S(\mathbf{x})]^\top$ given S models, and $\tilde{\Sigma}(\mathbf{x})^{-1}$ is the inverse of the covariance matrix between the information sources given as

$$\tilde{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1S}\sigma_1\sigma_S \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2S}\sigma_2\sigma_S \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1S}\sigma_1\sigma_S & \rho_{2S}\sigma_2\sigma_S & \cdots & \sigma_S^2 \end{bmatrix}, \tag{14}$$

where ρ is the correlation coefficient between the deviations of information sources at point $\mathbf{x}$, μ_1 and μ_2 are the posterior mean values of Gaussian processes of the information sources computed in Eq. 8, and σ_1^2 and σ_2^2 are the total variances of the information sources computed at point $\mathbf{x}$. Note, the dependence of each term on $\mathbf{x}$ has been suppressed in Eq. 14. We note here also that Eq. 13 is very sensitive at very high values of correlation. Thus, for correlation coefficients greater than 0.99, care should be taken to ensure that numerically the fusion is still well-conditioned. Conservatively, a practitioner could place a maximum allowed value on the correlation coefficient, which would ensure avoidance of ill-conditioning but incur a larger variance in the fused estimate.

In this discussion, it is assumed that the correlations between model deviations are known. To estimate these correlations, we use the reification process defined in [44, 45], which refers to the process of treating each information source, in turn, as ground truth. This means that we assume the data generated by the reified information source represents the ground truth quantity of interest. We use these data to estimate the correlation between the errors of the different information sources. The process is repeated for each information source. The detailed process of estimating the correlation between the errors of two information sources can be found in [44, 45]. We note that here we have extended that work to consider any number of information sources and non-uniformly sampled information source data. Both of these considerations are commonly encountered in practice, and thus, are required advancements for more practical use cases.

Following [45], assuming information source $f_i(\mathbf{x})$ is chosen to reify, the correlation coefficients between the information sources i and j, for $j = 1,\ldots,i-1,i+1,\ldots,S$, are

given as

$$\rho_{ij}(\mathbf{x}) = \frac{\sigma_i^2(\mathbf{x})}{\sigma_i(\mathbf{x})\sigma_j(\mathbf{x})} = \frac{\sigma_i(\mathbf{x})}{\sqrt{(\mu_i(\mathbf{x}) - \mu_j(\mathbf{x}))^2 + \sigma_i^2(\mathbf{x})}}, \quad (15)$$

where $\mu_i(\mathbf{x})$ and $\mu_j(\mathbf{x})$ are the mean values of the Gaussian processes of information sources i and j respectively at design point $\mathbf{x}$, computed according to Eq. (8), and $\sigma_i^2(\mathbf{x})$ and $\sigma_j^2(\mathbf{x})$ are the total variances at point $\mathbf{x}$. The first subscript under the correlation coefficient denotes which model has been reified. Since the only information we have regarding which information source we believe to be more realistic is the total variance of each, in addition to reifying information source i to estimate the correlation, we also reify information source j and estimate $\rho_{ji}(\mathbf{x})$. We then estimate the correlation between the errors as the variance weighted average of the two correlation coefficients as follows

$$\bar{\rho}_{ij}(\mathbf{x}) = \frac{\sigma_j^2(\mathbf{x})}{\sigma_i^2(\mathbf{x}) + \sigma_j^2(\mathbf{x})}\rho_{ij}(\mathbf{x}) + \frac{\sigma_i^2(\mathbf{x})}{\sigma_i^2(\mathbf{x}) + \sigma_j^2(\mathbf{x})}\rho_{ji}(\mathbf{x}). \quad (16)$$

These correlations can then be used to estimate the mean and variance of the fused estimate from Eqs. 12 and 13. We note here that a moving mean estimate of the correlation over the domain can be used to avoid ill-conditioning that can occur when information sources produce identical values at a given point in the domain.

4 Demonstrations

In this section, we demonstrate the use of our multi-information source fusion approach to the dual phase steel application. For this demonstration, we fuse information from the three physics-based reduced-order materials information sources with potentially non-uniformly sampled inputs. We compare the results to ground truth data collected from the finite element RVE model. We consider three different cases. The first case involves uniformly sampled data for each information source. The second case involves non-uniformly sampled information from the information sources, with a large region where each information source is only sparsely sampled. The third case involves non-uniform sampling of the information sources where each information source is sampled well over a small region of the input space and sparsely elsewhere. In each case, the multi-information source fusion approach taken here performs well, and is far superior to using any of the information sources in isolation. We conclude this section with a novel analysis of the effective number of information sources used to make the fused estimate at each point in the domain. This analysis provides a clear indication of our ability to exploit correlation for fusion, since without correlation information, only a single information source can confidently be used at a given point in the domain.

4.1 Case 1: Uniform Sampling

In this case, the ground truth is assumed to have been sampled previously at nine uniformly spaced points in the input domain. Each information source, that is, the isostrain, isostress, and isowork models has been evaluated at the points where ground truth information is known. The nine sampled points for each information source are used to construct Gaussian process surrogate models for each. These are shown as black lines through the nine black dots on the bottom three plots of Fig. 5. The dark shaded region on each of these plots represents the uncertainty associated with each Gaussian process respectively. The ground truth data was used to estimate the discrepancy of each information source from ground truth over the domain. This additional uncertainty is the lighter shaded regions in the bottom three plots of the figure. We note here again that we always assume the information sources are unbiased. This assumption allows us to avoid simply fitting each information source to the ground truth data, which would result in eliminating useful information in each information source. On each plot of Fig. 5, the ground truth is represented with the jagged green line and the result of our multi-information source fusion approach is represented by the smooth red line.

From the isostrain, isowork, and isostress subplots, it is clear that no single information source performs well across the domain. Indeed, over much of the domain each source performs poorly. However, as can be seen in the upper left plot, our fused information source is an excellent match to ground truth. This is further evidenced by the data provided in Tab. 2, where the mean squared errors (MSE):

$$\text{MSE}_i = \frac{1}{N}\sum_{j=1}^{N}(g(\mathbf{x}_j) - f_i(\mathbf{x}_j))^2, \quad (17)$$

and mean Kullback-Liebler divergences (MD_{KL}):

$$MD_{KL,i}(g\|f_i) = \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty} p_g(\mathbf{x}_j)\log\frac{p_g(\mathbf{x}_j)}{p_{f_i}(\mathbf{x}_j)}\mathrm{d}g, \quad (18)$$

between the ground truth and each information source are presented. Here, p represents the probability density function of the information source given by the subscript. From the table, we see that the fused source is a significant im-

Table 2. The mean squared errors (MSE) and the mean Kullback-Leibler divergences (MD_{KL}) between the true model (RVE) and the obtained models in Fig. 5.

Model	MSE	MD_{KL}
Fused Model	0.1048	2.4687
Isowork Model	13.2855	5.5577e+03
Isostrain Model	17.2272	7.1609e+03
Isostress Model	193.6599	8.5868e+04

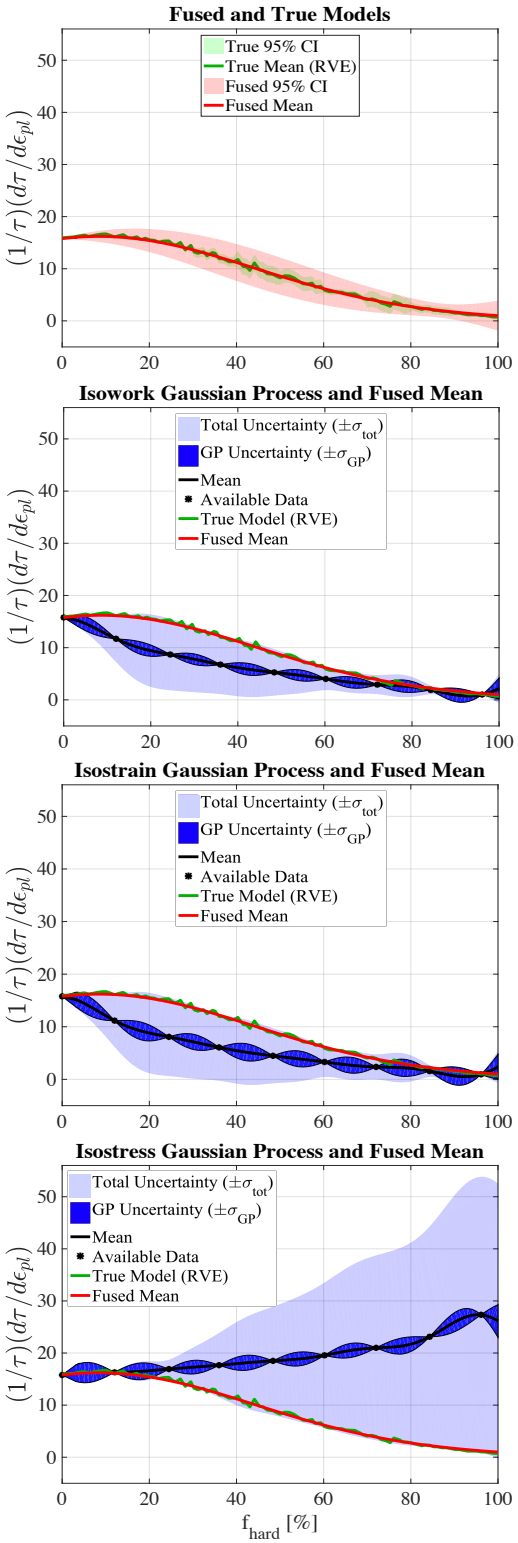

Fig. 5. The fused model and Gaussian processes of the isowork, isostrain and isostress models in comparison with the true (RVE) model.

provement over any of the individual sources and can be used to reliably predict the ground truth over the whole domain.

For this particular example, a subtle but crucial aspect of the use of the reification approach to information source fusion is revealed. For the input region $f_{\text{hard}} \in (95, 100]$, each information source overpredicts the truth. However, as can be seen in the top plot of Fig. 5, the fused information has overcome the bias of each information source to match the ground truth nearly exactly. If the correlation between the information source discrepancies were not learned, this would not be possible. This is due to the fact that the mean of a fused set of normal distributions is greater than the smallest mean and less than the largest mean in the set. That is, the fused estimate can never overcome the bias that occurred here.

4.2 Case 2: Large Sparsely Sampled Region

In this case, each information source has been sampled seven times, with all but one of the seven points being in the region $[0, 50]$. These seven points are not necessarily the same for each information source. We assume we have ground truth information at each of the seven points for each information source. This situation could occur if, for example, different groups have available different sets of ground truth but are unaware of other data or are unwilling to share this information with other groups. The key purpose of this demonstration case is to show the performance of our methodology over a poorly interrogated region of the domain when correlation information has been learned elsewhere over the domain. The results of this demonstration case are shown in Fig. 6. The information on each plot is presented in the same manner as that of Fig. 5. As can be seen from the top plot, the fused information source again performs well, albeit with more predictive uncertainty than Case 1, which is to be expected. The MSE and MD_{KL} values between the ground truth and each information source are given in Tab. 3.

Table 3. The mean squared errors (MSE) and the mean Kullback-Leibler divergences (MD_{KL}) between the true model (RVE) and the obtained models in Fig. (6).

Model	MSE	MD_{KL}
Fused Model	0.1799	4.5137
Isowork Model	16.0941	6.3537e+03
Isostrain Model	18.8465	9.7217e+03
Isostress Model	65.8499	1.6965e+04

Here, we see again that the fused information source is far superior to any information source in isolation. We also see that the better sampled situation of the first demonstration case results in a more accurate fused information source. Of additional interest in this case is that the ability of our fusion approach to overcome bias of all information sources is more readily apparent. Particularly, over the region $f_{\text{hard}} \in (60, 85]$, all three information sources overpredict the ground truth. However, the fused estimate has been pushed down towards the ground truth, away from the information source estimates. We stress here that we consider each of the information sources as unbiased in their uncertainty quantification. That is, the direction toward ground

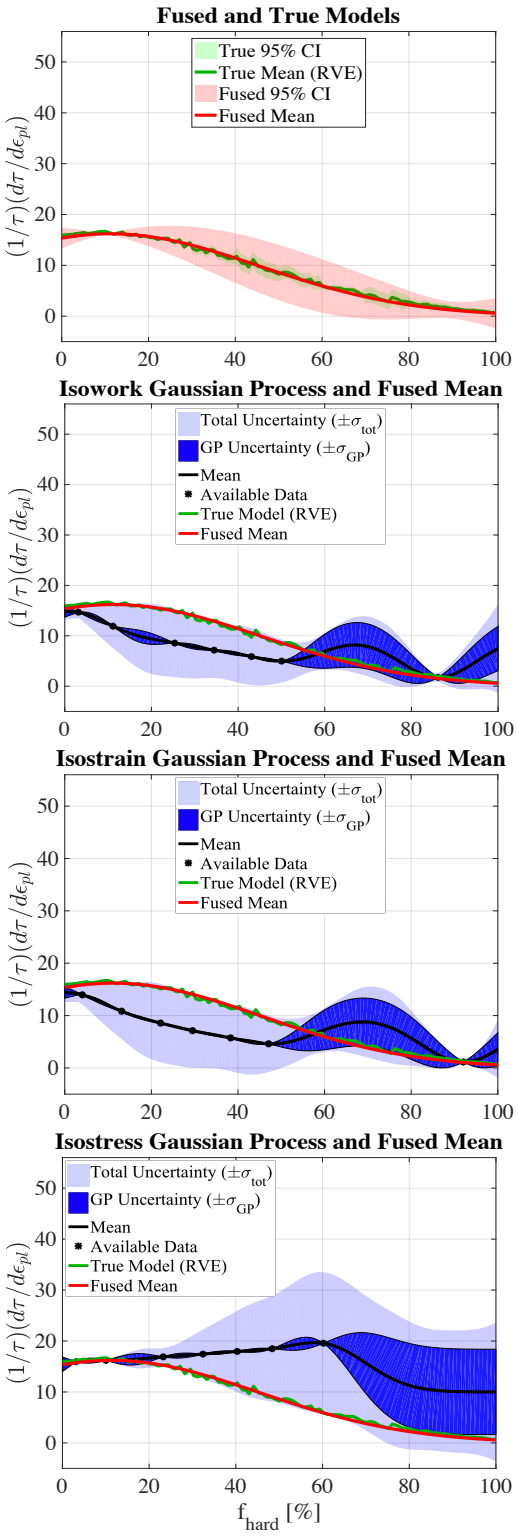

Fig. 6. The fused model and Gaussian processes of the isowork, isostrain and isostress models in comparison with the true (RVE) model when a few number of data is available in one region.

truth was not assumed by fitting each information source discrepancy to the ground truth in a biased fashion. Indeed, there are no ground truth samples in this region. The bias mitigation is due to the exploitation of correlations that have been learned through reification.

4.3 Case 3: Nearly Non-Overlapping Samples of Each Information Source

In this case, each information source has been sampled a few times in a specific region of the domain. The isostrain model was sampled generally in the left half of the domain, the isowork model was sampled generally in the right half of the domain, and the isostress model was sampled generally in the middle of the domain. Ground truth was again used to quantify information source discrepancy but was not shared across information sources. The key purpose of this demonstration case is to show the performance of our methodology when the information sources are essentially disparate in their knowledge of the quantity of interest over the domain. The results of this demonstration case are shown in Fig. 7. The information on each plot is presented in the same manner as that of Fig. 5. As can again be seen by the top plot, our approach performs well. The MSE and MD_{KL} values between the ground truth and each information source are given in Tab. 4.

Table 4. The mean squared errors (MSE) and the Kullback-Leibler divergences (MD_{KL}) between the true model (RVE) and the obtained models in Fig. (7).

Model	MSE	MD_{KL}
Fused Model	1.3049	2.9671
Isowork Model	6.3831	465.7251
Isostrain Model	15.6723	9.0723e+03
Isostress Model	178.4843	4.4007e+04

Of particular interest in this case is the performance of the fused information source in comparison to each individual source where that source was most heavily sampled. We can see clearly from Fig. 7 that the fused information source is a better approximation to ground truth in the left half of the domain than the isostrain model, where the isostrain model was most queried. The same is true when comparing the middle of the domain estimates from the fused information source to the isostress results, and the right half of the domain results from the fused information source and the isowork model. In each case, the fused information source is able to leverage the limited information from the other information sources to significantly outperform the information source that was heavily queried from in a given region.

4.4 Effective Independent Information Sources

To complete the demonstration of our multi-information source fusion approach, we define and present a novel number of effective independent information sources index. The index measures the effective number of independent information sources partaking in the fused estimate at each point of the input domain. To define the index, we first consider the normalized change of variance that occurs when information

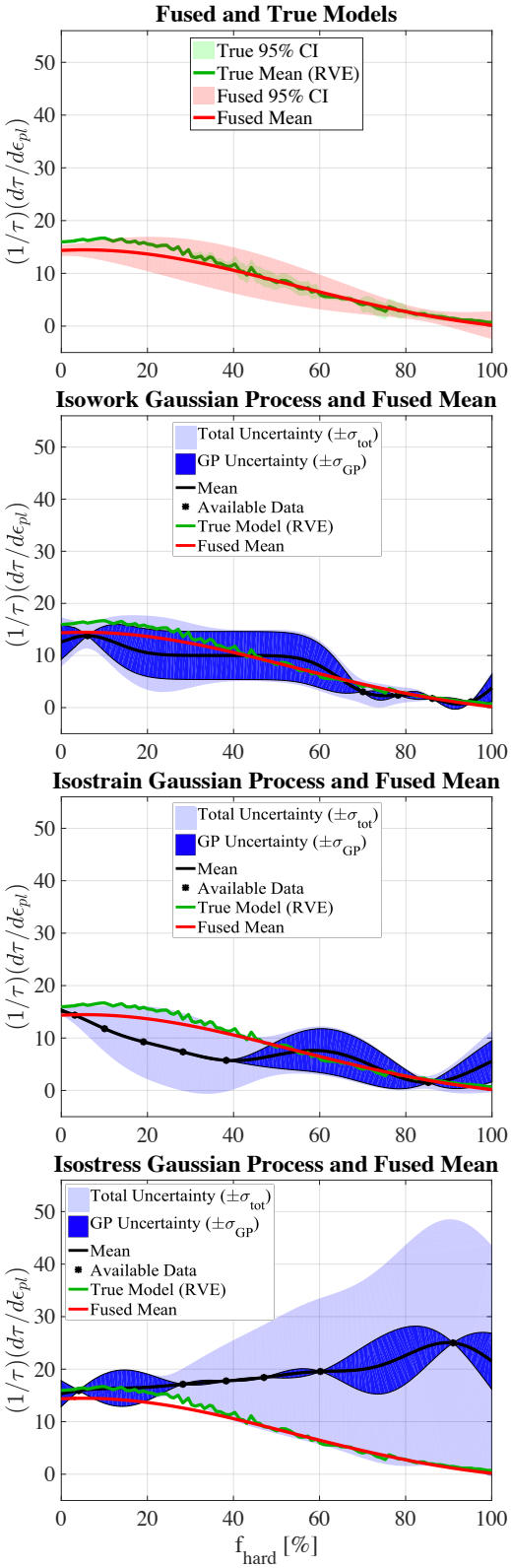

Fig. 7. The fused model and Gaussian processes of the isowork, isostrain and isostress models with data in different regions in comparison with the true (RVE) model.

sources are fused together at a given point. This change can be written as

$$\frac{\Delta\sigma^2(\mathbf{x})}{\sigma_*^2(\mathbf{x})} = 1 - \frac{1}{\sigma_*^2(\mathbf{x})\mathbf{e}^\top\tilde{\Sigma}(\mathbf{x})^{-1}\mathbf{e}}, \tag{19}$$

where $\Delta\sigma^2(\mathbf{x})$ is the variance reduction at $\mathbf{x}$ from the current best information source's variance, $\sigma_*^2(\mathbf{x})$, at that point. Then, for any number, S, of *independent* information sources, each with variance $\sigma_*^2(\mathbf{x})$ at $\mathbf{x}$, we can write

$$\frac{\Delta\sigma^2(\mathbf{x})}{\sigma_*^2(\mathbf{x})} = 1 - \frac{1}{S}. \tag{20}$$

Thus, the number of effective independent information sources with variance $\sigma_*^2(\mathbf{x})$ at the point $\mathbf{x}$ is given as

$$I_{\text{eff}} = \sigma_*^2(\mathbf{x})\mathbf{e}^\top\tilde{\Sigma}(\mathbf{x})^{-1}\mathbf{e}. \tag{21}$$

This index takes the value S when there are S independent sources with the same variance. If any sources have a larger variance, they will not contribute as much to the variance reduction at that point, and I_{eff} will be less than S. Thus, effective independent information sources, as measured by this index, are relative to the best source at a given point. We note here also that for highly correlated information sources, Eq. 13 can result in variance decreases that are larger than would occur with independent information sources. This generally occurs as a result of very similar but biased in the same direction information sources. The ability to exploit this situation is a feature of the reification approach.

The effective independent information source index for demonstration Case 1 is shown in Fig. 8. The figure includes the I_{eff} for the three source case, as well as each pair of two sources. The two source indices are still considered with respect to the best of the three information sources at a given point. This leads to the possible situation where pairs of information sources are contributing less than one effective information source.

For the fused approximation of Case 1, shown in the top plot of Fig. 5, it is interesting to note that the I_{eff} is not large over the input space, as shown in Fig. 8. For this particular problem, there is often one source that is much more uncertain than the other two at each point in the domain. This renders that source's contribution to I_{eff} to be very small. For example, isostrain has large variance for low to medium values of f_{hard} and isostress has large variance for large values.

From the pairwise curves, it is clear that initially the isowork-isostrain pair is driving the fused approximation. It is also clear that in this region of low values of f_{hard}, the fusion process is exploiting the high correlation between these two sources and is performing better than three independent sources could. The isostress model takes over the approximation around $f_{\text{hard}} \approx 10\%$. This holds until $f_{\text{hard}} \approx 30\%$, where all three sources are contributing to the prediction. At

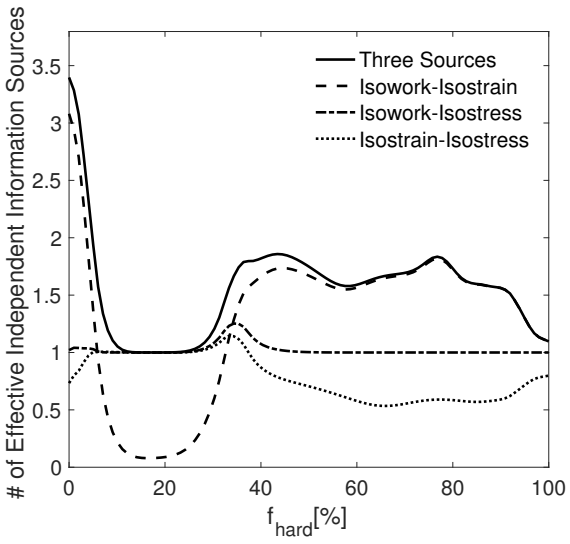

Fig. 8. Number of effective independent information sources, I_{eff} as a function of f_{hard} for demonstration Case 1.

$f_{\text{hard}} \approx 40\%$, the isowork-isostrain pair again drive the prediction. This continues until until the end of the domain is reached. Thus, while all three information sources do not contribute equally over the domain, they are all three required to make the fused approximation presented in Fig. 5.

Though the effective independent information source analysis presented here relied only on the I_{eff}, which was derived through the discrepancy quantification and reification process, the analysis is consistent with the fundamentals of mechanics for these information sources and this application. This provides evidence that such an analysis could be used to aid in the construction of a more sophisticated physics-based model from models using simplified assumptions. That is, this index provides information about when certain assumptions are valid and when they are not. The index also provides a means of valuing a new evaluation of an information source over the domain. For example, this analysis reveals that sampling the isostress model on the interval $f_{\text{hard}} \in [40, 100]$ will provide little value in terms of effective information sources when an isostrain and isostress model are also available. Such a valuation could prove useful in a resource constrained process for estimating a quantity of interest with many possible information sources available.

5 Conclusions and Future Work

In this paper we presented and demonstrated a correlation-exploiting multi-information source fusion approach. The method included new extensions to the fusion of any number of correlated information sources, as well as the creation of a novel effective independent information source index. The fusion methodology was demonstrated on microstructure-sensitive performance prediction for ductile dual-phase materials. In all cases, the proposed fusion approach performed exceptionally well, far exceeding the predictive capabilities of any individual information source. This provides evidence that our approach to information source fusion is highly applicable to the challenge of integration in ICME tools. In future work, the framework developed here will be validated against larger sets of ground truth data and be demonstrated in higher dimensions. The framework will also be extended to handle information sources with misaligned input-output interfaces, which is a key challenge facing the ICME community.

Acknowledgements

The authors would like to acknowledge the support of the National Science Foundation through grant No. NSF-CMMI-1663130, *DEMS: Multi-Information Source Value of Information Based Design of Multiphase Structural Materials*. Arroyave would also like to acknowledge the support of the National Science Foundation through grant No. NSF-CMMI-1534534, *DMREF: Accelerating the Development of Phase-Transforming Heterogeneous Materials: Application to High Temperature Shape Memory Alloys*. Allaire and Arroyave would also like to acknowledge the support of the National Science Foundation through grant No. NSF-DGE-1545403, *NRT-DESE: Data-Enabled Discovery and Design of Energy Materials (D^3EM)*.

References

[1] Olson, G. B., 1997. "Computational design of hierarchically structured materials". *Science,* **277**(5330), pp. 1237–1242.

[2] Olson, G. B., 2000. "Designing a new material world". *Science,* **288**(5468), pp. 993–998.

[3] Allison, J., 2011. "Integrated computational materials engineering: A perspective on progress and future steps". *JOM,* **63**(4), pp. 15–18.

[4] The Minerals Metals & Materials Society (TMS), 2015. *Modeling Across Scales: A Roadmapping Study for Connecting Materials Models and Simulations Across Length and Time Scales*. TMS, Warrendale, PA.

[5] Rashid, M., 1981. "Dual phase steels". *Annual Review of Materials Science,* **11**(1), pp. 245–266.

[6] Chen, P., Ghassemi-Armaki, H., Kumar, S., Bower, A., Bhat, S., and Sadagopan, S., 2014. "Microscale-calibrated modeling of the deformation response of dual-phase steels". *Acta Materialia,* **65**, pp. 133–149.

[7] Srivastava, A., Bower, A., Hector Jr, L., Carsley, J., Zhang, L., and Abu-Farha, F., 2016. "A multiscale approach to modeling formability of dual-phase steels". *Modelling and Simulation in Materials Science and Engineering,* **24**(2), p. 025011.

[8] Voigt, W., 1889. "On the relation between the elasticity constants of isotropic bodies". *Ann. Phys. Chem,* **274**, pp. 573–587.

[9] Reuss, A., 1929. "Berechnung der fließgrenze von mischkristallen auf grund der plastizitätsbedingung für einkristalle.". *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik,* **9**(1), pp. 49–58.

[10] Bouaziz, O., and Buessler, P., 2002. "Mechanical behaviour of multiphase materials: an intermediate mixture law without fitting parameter". *Revue de Métallurgie–International Journal of Metallurgy,* **99**(1), pp. 71–77.

[11] Gerbig, D., Srivastava, A., Osovski, S., Hector, L. G., and Bower, A., 2017. "Analysis and design of dual-phase steel microstructure for enhanced ductile fracture resistance". *International Journal of Fracture*, pp. 1–24.

[12] Documentation, A., and Manual, U., 2010. "Version 6.10". *Dassault Systemes*.

[13] Nemat-Nasser, S., and Hori, M., 2013. *Micromechanics: overall properties of heterogeneous materials*, Vol. 37. Elsevier.

[14] Alexandrov, N., Lewis, R., Gumbert, C., Green, L., and Newman, P., 1999. Optimization with variable-fidelity models applied to wing design. Tech. Rep. CR-209826, NASA, December.

[15] Alexandrov, N., Lewis, R., Gumbert, C., Green, L., and Newman, P., 2001. "Approximation and model management in aerodynamic optimization with variable-fidelity models". *AIAA Journal,* **38**(6), November-December, pp. 1093–1101.

[16] Balabanov, V., Haftka, R., Grossman, B., Mason, W., and Watson, L., 1998. "Multifidelity response surface model for HSCT wing bending material weight". In 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization. AIAA 1998-4804.

[17] Balabanov, V., and Venter, G., 2004. "Multi-fidelity optimization with high-fidelity analysis and low-fidelity gradients". In 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference. AIAA 2004-4459.

[18] Choi, S., Alonso, J. J., and Kroo, I. M., 2009. "Two-level multifidelity design optimization studies for supersonic jets". *Journal of Aircraft,* **46**(3), pp. 776–790.

[19] Eldred, M., Giunta, A., and Collis, S., 2004. Second-order corrections for surrogate-based optimization with model hierarchies. 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference. AIAA 2004-4457.

[20] March, A., and Willcox, K., 2012. "Provably convergent multifidelity optimization algorithm not requiring high-fidelity derivatives". *AIAA Journal,* **50**(5), pp. 1079–1089.

[21] March, A., and Willcox, K., 2012. "Convergent multifidelity optimization using Bayesian model calibration". *Structural and Multidisciplinary Optimization,* **46**(1), pp. 93–109.

[22] Chen, S., Jiang, Z., Yang, S., and Chen, W., 2016. "Multimodel fusion based sequential optimization". *AIAA Journal,* **55**(1), pp. 241–254.

[23] Scott, W., Frazier, P., and Powell, W., 2011. "The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression". *SIAM Journal on Optimization,* **21**(3), pp. 996–1026.

[24] Jones, D. R., Schonlau, M., and Welch, W. J., 1998. "Efficient global optimization of expensive black-box functions". *Journal of Global Optimization,* **13**(4), pp. 455–492.

[25] Huang, D., Allen, T. T., Notz, W. I., and Miller, R. A., 2006. "Sequential kriging optimization using multiple-fidelity evaluations". *Structural and Multidisciplinary Optimization,* **32**(5), pp. 369–382.

[26] Moore, R. A., Romero, D. A., and Paredis, C. J., 2014. "Value-based global optimization". *Journal of Mechanical Design,* **136**(4), p. 041003.

[27] Williams, C. K., and Rasmussen, C. E., 2006. "Gaussian processes for machine learning". *The MIT Press*.

[28] Leamer, E., 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. John Wiley & Sons, New York, NY.

[29] Madigan, D., and Raftery, A., 1994. "Model selection and accounting for model uncertainty in graphical models using occam's window". *American Statistical Association,* **89**(428), pp. 1535–1546.

[30] Draper, D., 1995. "Assessment and propagation of model uncertainty". *Journal of the Royal Statistical Society Series B,* **57**(1), pp. 45–97.

[31] Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C., 1999. "Bayesian model averaging: A tutorial". *Statistical Science,* **14**(4), pp. 382–417.

[32] Clyde, M., 2003. Model averaging. In Subjective and Objective Bayesian Statistics, 2nd ed., Chapter 13, Wiley-Interscience.

[33] Clyde, M., and George, E., 2004. "Model uncertainty". *Statistical Science,* **19**, pp. 81–94.

[34] Mosleh, A., and Apostolakis, G., 1986. "The assessment of probability distributions from expert opinions with an application to seismic fragility curves". *Risk Analysis,* **6**(4), pp. 447–461.

[35] Zio, E., and Apostolakis, G., 1996. "Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories". *Reliability Engineering & System Safety,* **54**(2-3), pp. 225–241.

[36] Reinert, J., and Apostolakis, G., 2006. "Including model uncertainty in risk-informed decision making". *Annals of Nuclear Energy,* **33**(4), pp. 354–369.

[37] Riley, M., and Grandhi, R., 2011. "Quantification of modeling uncertainty in aeroelastic analyses". *Journal of Aircraft,* **48**(3), pp. 866–873.

[38] Julier, S., and Uhlmann, J., 1997. A non-divergent estimation algorithm in the presence of unknown correlations. In proceedings of the American Control Conference, pp. 2369-2373.

[39] Julier, S., and Uhlmann, J. General decentralized data fusion with covariance intersection. In D. Hall and J. Llinas, editor, Handbook of Data Fusion. CRC Press, Boca Raton FL, USA, 2001.

[40] Geisser, S., 1965. "A bayes approach for combining correlated estimates". *Journal of the American Statistical Association,* **60**, pp. 602–607.

[41] Morris, P., 1977. "Combining expert judgments: A bayesian approach". *Management Science,* **23**, pp. 679–693.

[42] Winkler, R., 1981. "Combining probability distributions from dependent information sources". *Management Science,* **27**(4), pp. 479–488.

[43] Winkler, R. L., 1981. "Combining probability distributions from dependent information sources". *Management Science,* **27**(4), pp. 479–488.

[44] Allaire, D., and Willcox, K., 2012. "Fusing information from multifidelity computer models of physical systems". In Information Fusion (FUSION), 2012 15th International Conference on, IEEE, pp. 2458–2465.

[45] Thomison, W. D., and Allaire, D. L., 2017. "A model reification approach to fusing information from multifidelity information sources". In 19th AIAA Non-Deterministic Approaches Conference, p. 1949.