18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Active Data Collection for Inadequate Models

Gabriel Terejanu
Department of Computer Science and Engineering
University of South Carolina
Columbia, South Carolina 29208
Email: terejanu@cec.sc.edu

***Abstract*—Obtaining informative measurements is a fundamental problem when inadequate models are used to guide the design of experiments. A comprehensive approach to experimental design for inadequate physics-based models is proposed by focusing on the coupling between the structural uncertainty modeling and the adaptive data collection process. First, by taking advantage of the structure of physics-based models, unlike current approaches, rigorous structural uncertainty models are created to yield solutions, which satisfy physical constraints such as conservation of mass. Second, new adaptive data collection strategies are proposed by combining two current approaches, model driven and model free experimental design, to optimally trade off between model exploitation and design space exploration. The applicability and feasibility of these new ideas will be demonstrated on dispersion models, which are widely used in practice from regulatory applications to emergency response in chemical, nuclear, biological and radiological releases. These dispersion models are polluted by structural errors due to various assumptions (e.g. diffusion coefficients) that can only be informed using limited experimental data.**

I. Introduction

This study addresses an *open fundamental problem* in the scientific literature, namely how to *obtain informative measurements when models with structural errors are guiding the design of experiments*. Lindley [1] introduced model-driven data strategies based on maximization of information gain in 1956 and fueled a large body of work predominantly on applications to various fields and development of approximate algorithms to speedup the optimization problem [2].

However the *"the Achilles' heel of these methods is that they estimate the utility of a measurement assuming that the model is correct. This might lead to undesirable results. The search for ideal measures of data utility is still open"*, David MacKay [3]. In such situations, conflicting information arises between model predictions and measurements yielding biased estimates and underestimated uncertainties, which undermines the whole experimental design process, as shown by the author and collaborators in Ref. [4].

To address this fundamental problem, the main approach taken here is to develop a basic understanding of the impact of modeling errors on data collection strategies. This will inform the development of novel adaptive methods for experimental design using physics-based models in the presence of *structural uncertainty.* Here, *"physics-based models"* are formulated based on well established physical principles or theories (e.g. conservation of mass) whose applicability to the problem at hand is not questioned, but which also include various less reliable modeling approximations or auxiliary inadequate models, see Fig.1. This is the common structure of computational models and it will be studied in the context of experimental design. The ultimate goal is to devise novel adaptive strategies through coupling of Bayesian optimal design [2] with space filling design [5] to provide an *optimal trade-off between model exploitation and design space exploration.*

Papers on experimental design can be found in geoscience [6], neuroscience [7], biomedical applications [8], [9], [10], engineering [11], systems biology [12], [9], combustion kinetics [13], and electrochemical systems [14] just to name a few. Furthermore, active research is currently done in developing approximate methods to speedup the calculations needed for experimental design [15], [13], [16], [10], [17], [18]. Nonetheless, the application of Bayesian design principles to actual experiments lags behind theoretical advancements [2]. A more efficient learning can be accomplished by using experimental design strategies that tightly couple the computational modeling, experimental endeavors and data analysis. However, *there are research issues that need to be addressed in the design of experiments in the presence of model errors.*

The central challenge in using computational models for scientific discovery, engineering design, or decision support is that the process follows a path contaminated with errors and uncertainties, see Fig.1. One of the key concepts is that the physics-based models in this study are defined as composite models with separable highly reliable theory and less reliable embedded models. This goes beyond constrained estimation of parameters related to physical laws [19]. Here, the focus is on understanding two specific processes along this path, namely structural uncertainty modeling and adaptive data collection.

The starting point is a mathematical model with known structural error, for which uncertainty models need to be constructed. Structural uncertainty is one of the most important and challenging issues in uncertainty quantification and experimental design. Many models used in engineering practice (e.g., Gaussian plume models, RANS turbulence models) are known to be deficient, in that model predictions do not accurately represent experimental observations. *A key innovation proposed in this work is a general formulation based on internal discrepancy models to deal with modeling errors by exploiting the structure of composite models.*

There is no agreement in the scientific community today on the experimental design strategy that can consistently provide informative measurements for inadequate models. In

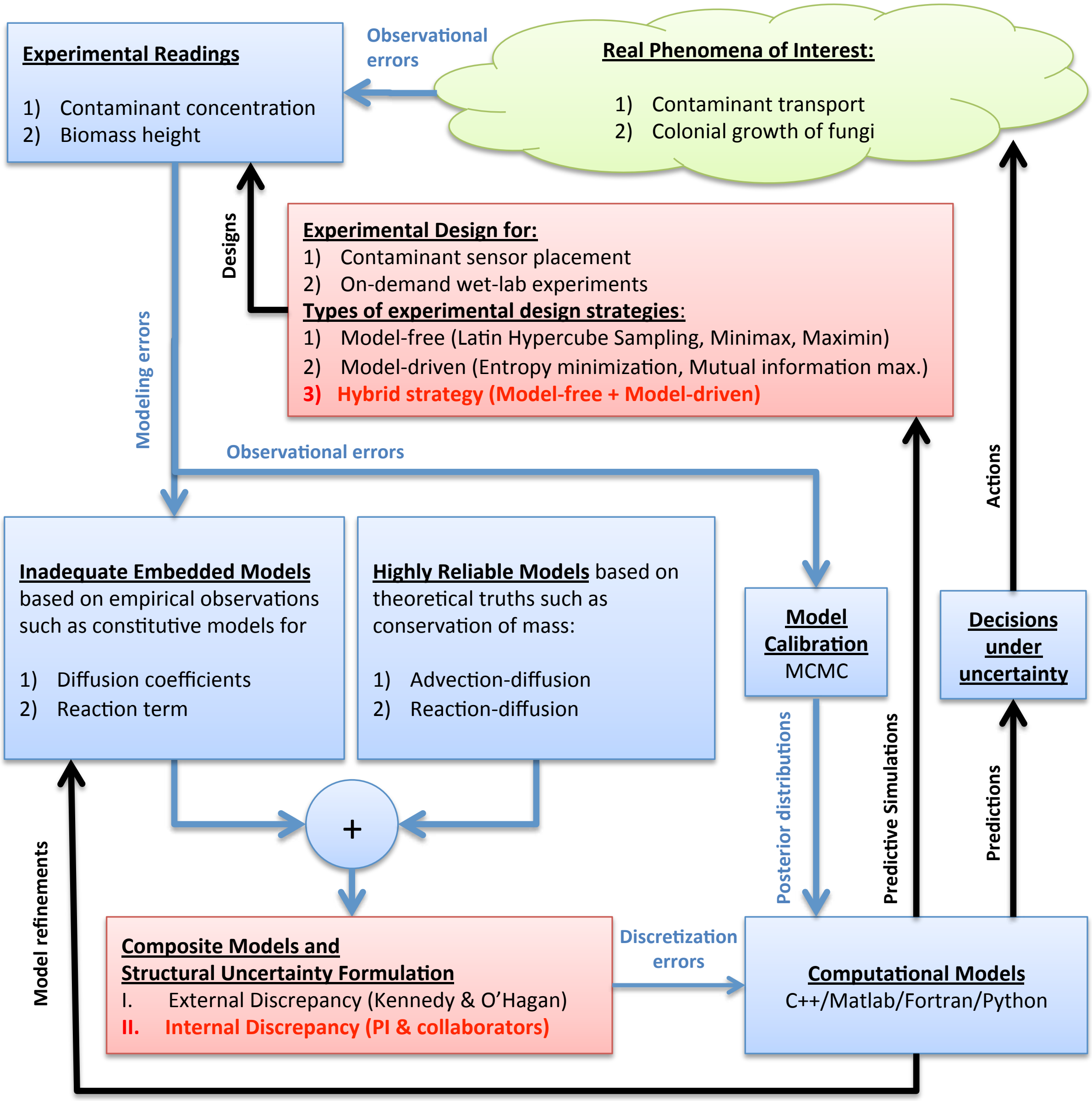


Fig. 1: Active data collection is based on a deep integration of theory, experimentation and computation. Contributions of this paper are in the areas of *structural uncertainty modeling* and *experimental design strategies* with applications to *dispersion models for contaminant transport*.

this work, new adaptive data collection strategies are proposed by combining two current approaches, model driven and model free experimental design to effectively trade off between model exploitation and design space exploration. The view here is that in the presence of model error, the information provided by the structural uncertainty model has to be exploited to inform the data collection process.

To facilitate the development of an experimental design process in the presence of model error and understand the impact of structural error on the data collection process, it is necessary to have a realistic application on which to test new ideas. The proposed framework is tested on Gaussian plume dispersion models [20], which are widely used in practice from regulatory applications to emergency response in chemical, biological, radiological, and nuclear (CBRN) releases [21]. These dispersion models are polluted by structural errors due to various assumptions (e.g. diffusion coefficients) that can only be informed using limited experimental data.

The remainder of the paper begins with an overview of relevant background material on physics-based models in Section II, and continues with proposed approaches contrasted by state-of-the-art structural uncertainty, Section III, and in sequential experimental design, Section IV. Finally, preliminary results are presented in Section V and conclusions and future work in Section VI.

II. Physics-based models and Model calibration (Background)

Consider the following mathematical model for a physical system of interest:

$$\mathcal{R}(u, \tau; \boldsymbol{\xi}) = 0, \quad (1)$$
$$\mathbf{y} = f(u) \quad (2)$$
$$\mathbf{q} = g(u) \quad (3)$$

where $\mathcal{R}$ is some operator, u is the solution or state variable, τ is a quantity for which a physical model is required, and $\boldsymbol{\xi}$ is a set of scenario variables needed to precisely define the case being considered, such as boundary conditions. Here, $\mathcal{R}$ would be a partial differential equation expressing conservation of mass, momentum, and energy. For example, it could be advection-diffusion equation, with u being the contaminant mass concentration and τ being the eddy diffusivities. In addition, we require maps from the solution to the observable quantities $\mathbf{y}$ as well as the quantities of interests (QoIs) $\mathbf{q}$. In general, the QoIs will be different from the observable quantities.

If τ were known in terms of u and $\boldsymbol{\xi}$, the system would be closed, and Eq.(1) would implicitly define a mapping from the scenario variables $\boldsymbol{\xi}$ to the solution variables u. Thus, a physical model for τ is needed. *Often, this embedded model for τ is not based on first principles, leading to structural uncertainty* due to model m, which is presumed to depend on the solution and a set of model parameters $\boldsymbol{\theta}$, as follows.

$$\tau \approx m(u, \boldsymbol{\xi}; \boldsymbol{\theta}) \quad (4)$$

In this context, the goal of experimental design is to identify the experimental scenarios $\boldsymbol{\xi}$ that can provide informative measurements d to learn about model parameters $\boldsymbol{\theta}$, or have accurate prediction of either the observable $\mathbf{y}$ or the QoI $\mathbf{q}$. After performing the experiment $\boldsymbol{\xi}$ and obtaining data d the next step is to calibrate the model. In the *Bayesian model calibration*, one seeks a complete statistical description, in the form of a probability density function (pdf), of the parameters that make the model consistent with the experimental data. This pdf is defined by the simple but powerful Bayes' Theorem [22], [23], $p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}; \mathbf{d})$, where $p(\boldsymbol{\theta}|\mathbf{d})$ is the posterior pdf (the solution of the inverse problem), $p(\boldsymbol{\theta})$ is the prior pdf, and $\pi(\boldsymbol{\theta}; \mathbf{d})$ is the likelihood function, which accounts for both *experimental uncertainty* and *structural uncertainty*.

III. Structural Uncertainty Modeling (Current and Proposed Approaches)

The main challenge of model calibration and hence experimental design, is that there is always some discrepancy between the output of the physical model and the values of the real process due to the inadequacy of the embedded model for τ, namely $m(u, \boldsymbol{\xi}; \boldsymbol{\theta})$. *In this section a new formulation is proposed to deal with this structural uncertainty, which is a key contribution of this work.*

A. External Discrepancy (Current Approach)

A common approach for specifying the structural uncertainty model is that of Kennedy and O'Hagan [24]. In this approach, the true (but unknown) value of the observables, $\mathbf{d}_{\text{true}}$, is assumed to be related to the model output by

$$\mathbf{d}_{\text{true}} = f(u(\boldsymbol{\theta})) + \boldsymbol{\epsilon}_{\text{model}},$$

where $\boldsymbol{\epsilon}_{\text{model}}$ is a Gaussian process [25] representing the structural uncertainty (also referred to as model inadequacy or model discrepancy). Note that additive error is not the only option, multiplicative error is possible as well. When coupled with a model of the experimental error, this statement defines the likelihood function and Bayes rule can be used to update knowledge of $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}_{\text{model}}$. Since this approach, called here *external discrepancy formulation*, is formulated in terms of the observable quantity, it is appropriate only when three conditions are satisfied.

(1) Physical constraints on the modeling error should be formulated in terms of the observable, so that the combined model (i.e., $f + \boldsymbol{\epsilon}_{\text{model}}$) does not violate known physical laws. However, if the observable is a concentration field, *blindly adding Gaussian random fields to the observable would lead to a vector random field model where individual realizations do not satisfy conservation of mass.* Since mass is certainly conserved, such a model is inadmissible. Furthermore, due to confounding of model parameters and discrepancy parameters, the calibration results for parameters can be biased and the uncertainty under-estimated. The importance of constraining the discrepancy model has been recently addressed by Brynjarsdottir and O'Hagan [26]. However, given that the constraints are imposed at the observable level, there is no guarantee that the conservation of mass is satisfied.

(2) Predictions of the observable at a scenario of interest should be executed only when the *measurement scenarios are "near" the scenario of interest so that the structural uncertainty model is not extrapolated.* In this formulation, the structural uncertainty is a purely statistical model, and thus, it contains only information extracted from the calibration data. Thus, the validity of its extrapolated predictions is questionable.

(3) Predictions of the QoI should be executed only when the QoI is uniquely defined by the observable, such that the structural uncertainty for the observable can be propagated to the QoI. Otherwise, the predicted uncertainty of the QoI will only account for parametric uncertainty because the structural uncertainty cannot be propagated to the QoI. The last two issues have been recently addressed by the author and collaborators [27].

These conditions severely limit the application of this technique. Thus, *overcoming these drawbacks is a major focus of* *the proposed effort.* The proposed work aims to remove the constraints associated with this approach by formulating the structural uncertainty model wherever known modeling errors are introduced.

B. Internal Discrepancy (Proposed Approach)

In the defined composite models based on physics, structural uncertainties arise from imperfections in the various functionals involved (m, g, and f), as defined in (1) through (3). For simplicity, consider the case where the structural uncertainty is restricted to the model m for the quantity τ. This uncertainty could be represented by introducing an additive or multiplicative error, $\boldsymbol{\epsilon}_{\text{model}}$, into the model for τ, which has been recently proposed by the author and collaborators [27].

$$\mathcal{R}(u, m(u, \boldsymbol{\xi}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_{\text{model}}; r) = 0 \tag{5}$$

$$\mathbf{y} = f(u) \tag{6}$$

$$\mathbf{q} = g(u). \tag{7}$$

By introducing a stochastic model for $\boldsymbol{\epsilon}_{\text{model}}$ to represent incomplete knowledge, this new system becomes a stochastic model governing the state u, which is random as well. While (6) and (7) are formally unchanged from (2) and (3), respectively, $\mathbf{y}$ and $\mathbf{q}$ become random variables because the input to the functions f and g is random.

This *internal discrepancy formulation* promises to remove the constraints associated with the external discrepancy approach. The impact of the structural uncertainty on the QoI can be computed by simply propagating the random solution u through the operator g. Known physical constraints on the model form are either automatically enforced or can be easily checked (e.g., it would be impossible to develop a model, like that described in III-A, that violated conservation of mass). Finally, when informed by data in the calibration phase, this representation is expected to be more generalizable than the external discrepancy approach since the structural uncertainty model represents directly the uncertainty in the physical model, rather than its effect on other quantities.

While the location where uncertainty is introduced and should be modeled is often clear, the most appropriate form of probabilistic model generally is not. *Multiple uncertainty model forms will be explored in the future in the context of the dispersion models, which opens the opportunity to explore experimental design strategies with the goal of model discrimination.*

IV. Sequential Experimental Design (Current and Proposed Approaches)

There are two ways to perform experiments: *batch strategies* that select all designs before experiments are performed, and *sequential strategies* where the selection of experimental conditions are performed in sequence. Given that the focus of this study is the development of design strategies in the presence of model error, sequential design is adopted as it takes advantage of information obtained from previous experiments [28] and allows for the adaptation of the data collection process.

The objective of any experimental design can be grouped in three categories: *calibration, model selection and prediction*, all of which can be subject to additional physical or financial constraints. For *calibration* the aim is to identify the experimental conditions $\boldsymbol{\xi}$ that provide informative measurements to learn model parameters $\boldsymbol{\theta}$ that carry physical meanings. When alternative models are available, the goal of *model selection* is to determine the experimental conditions capable to discriminate the models. And for *prediction*, the aim is to identify the designs that can provide accurate predictions of either the observable $\mathbf{y}$ or the QoI $\mathbf{q}$.

Overall there are two main strategies used to determine experimental conditions, namely *model driven strategies* and *model free strategies*, which are briefly described in the following sections. Nonetheless, *experimental design in the presence of model error is a challenging task seldom mentioned in the literature, and it is argued that neither model driven or model free strategies can consistently provide desired results in this context*. In this study, a couple of investigations are used to develop *adaptive hybrid approaches* (*model driven + model free*) capable to deal with data selection in the presence of model error.

A. Model Driven Strategies - Model Exploitation (Current)

Model driven strategies determine experimental conditions by **exploiting** the information contained in the model. Strategies in this category are mainly given by *Bayesian optimal designs* such as D-optimal design or any alphabetic criteria [29], and information theoretic measures [30] such as maximum entropy [31], maximum mutual information [2], which take advantage of Shannon's measure of information [32]. For illustration purposes consider a model driven strategy given by maximizing the mutual information [1]. It is specifically targeted at reducing the entropy of model parameters at time $k+1$ given all the previous k observations collected, D_k.

$$\boldsymbol{\xi}^*_{k+1} = \arg\max_{\boldsymbol{\xi}_{k+1}} J_{md}(\boldsymbol{\xi}_{k+1}) \tag{8}$$

$$J_{md}(\boldsymbol{\xi}_{k+1}) = \int p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\xi}_{k+1}, D_k) \log \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\xi}_{k+1}, D_k)}{p(\mathbf{y}|\boldsymbol{\xi}_{k+1}, D_k)p(\boldsymbol{\theta}|D_k)} d\mathbf{y}d\boldsymbol{\theta} \tag{9}$$

The *advantage* of this type of strategy is that it targets directly any of the objectives of the experimental design. However, the *challenge* with using a model driven strategy to select experimental conditions in the presence of model error comes from the fact that it inherently uses inadequate simulations, yielding undesirable results [3]. This is due to structural uncertainty models that contain limited information about the overall model error distribution in the design space. This has already been shown by the author and collaborators in Ref. [4], and the current study attempts to formalize an adequate design strategy in this type of situations.

B. Model Free Strategies - Design Space Exploration (Current)

Model free strategies are independent of model simulations, and they **explore** the design space well by providing designs that maximize a coverage criteria [33], [34]. These type of strategies have been used in computer experiments with the goal to build simulators to replace the real data generating process. They have been used with *external discrepancy formulation* [35], however their main objective is providing accurate predictions only for the observable. For illustration purposes consider a model free strategy given by maximin distance [36], which spreads the design points uniformly across the whole design space.

$$\boldsymbol{\xi}^*_{k+1} = \arg\max_{\boldsymbol{\xi}_{k+1}} J_{mf}(\boldsymbol{\xi}_{k+1}) \tag{10}$$

$$J_{mf}(\boldsymbol{\xi}_{k+1}) = \min_{\boldsymbol{\xi}'_{k+1}} ||\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}'_{k+1}|| \tag{11}$$

Compared with model driven, the *advantage* of model free strategies is that by throughly exploring the design space one can infer the overall distribution of model error. However, the *drawback* is that by not using the information contained in the model it may yield inefficient designs that provide very little information for the particular objective of the experimental design, increasing this way the overall cost of experimentation.

C. Adaptive Hybrid Strategies - Exploitation vs Exploration (Proposed)

To address the challenges of the previous two approaches, an adaptive hybrid strategy is proposed by combining model driven and model free strategies. This new strategy will *adaptively trade off between exploitation and exploration* to achieve the goal of the sequential experimental design in the presence of model error. One way to combine these two strategies is by creating a new cost function which is a linear combination of the individual cost functions.

$$\boldsymbol{\xi}^*_{k+1} = \arg\max_{\boldsymbol{\xi}_{k+1}} \left[J_{md}(\boldsymbol{\xi}_{k+1}) + \alpha J_{mf}(\boldsymbol{\xi}_{k+1}) \right] \tag{12}$$

Here α is a tunable coefficient. If α is assigned with a large value then model free will play a dominant role. A small α will make the strategy more model driven. The main question then becomes: *How to tune α to achieve the goal of experimental design?* Here are some examples of how to set $alpha$.

1) Set $\alpha = 1$ so that both strategies will have equal weight.
2) First explore the space (α large) to learn as much about the model error as possible, then exploit the model (α small) to further reduce the uncertainty about the parameters or predictions according to the goal of the experiment.
3) Decrease α gradually as more experiments are performed in order to smoothly transition from exploration to exploitation.
4) Make the selection of α adaptive, by monitoring the rate of learning. If the rate of learning is significantly decreased then switch to exploration to open additional opportunities for learning, otherwise exploit the model.

Promising preliminary results point in the direction of adaptively choosing α, see Section V. However, it is not clear at this point the strategy and the factors that will yield desirable and consistent results across models, which motivates a comprehensive future investigation.

V. Preliminary Results

This section describes a preliminary application of experimental design approach to a simple steady-state Gaussian plume model with the goal of identifying a sequence of sensor locations $\{(x_k, y_k)\}$, Fig.2(c), capable to provide information about the release height H.

Two investigations are compared in this section that correspond to modeling the structural uncertainty using both external and internal discrepancy - previously introduced. The true model in Eq. (14) - used to generate synthetic data - involves a physical phenomenon that is not represented in the approximate physical model used for calibration and experimental design, see Fig.2(a,b). Thus model inadequacy is important.

Both the true model and the approximate model are given by the steady-state Gaussian plume solution to the advection-diffusion equation under assumptions of isotropic diffusion and constant wind velocity, $u = 1$, which is sufficiently large such that the longitudinal diffusion term can be neglected [37]. The release mass in both models is $Q = 1$.

The difference between the true and approximate model comes from assumptions regarding the eddy diffusivity K. In the true model the diffusivity is linearly dependent on the downwind distance, while in the approximate model is assumed to be constant. In both cases the same Gaussian plume solution is obtained [38] to calculate the concentration at a specific location $c(x, y)$. Eq. (15) depicts the treatment of structural uncertainty using Kennedy and O'Hagan formulation [24] and the proposed internal discrepancy formulation is given in Eq. (16).

- True model:

$$c(x, y) = \frac{Q}{2\pi K x} \exp\left(-\frac{u(y^2 + H^2)}{4Kx} \right) \tag{13}$$

$$K = \frac{1}{12} u x \tag{14}$$

- External discrepancy formulation:

$$c(x, y) = \frac{Q}{2\pi K x} \exp\left(-\frac{u(y^2 + H^2)}{4Kx} \right) \epsilon_{\text{model}} \tag{15}$$

- Internal discrepancy:

$$c(x, y) = \frac{Q}{2\pi K \epsilon_{\text{model}} x} \exp\left(-\frac{u(y^2 + H^2)}{4K\epsilon_{\text{model}} x} \right) \tag{16}$$

- Discrepancy model:

$$\epsilon_{\text{model}} \sim \log\mathcal{N}(0, \sigma_\epsilon^2) \tag{17}$$

- Prior distributions:

$$p(H) = \mathcal{U}[0, 10] \tag{18}$$

$$p(K) = \log\mathcal{N}(-0.35, 0.7^2) \tag{19}$$

$$p(\sigma_\epsilon^2) = \log\mathcal{N}(-4, 1) \tag{20}$$

Four different strategies are tested for both formulations. A model-driven strategy given by mutual information maximization (MI), Eq.(8), a model-free strategy given by maximin distance (DIST), Eq.(10), a hybrid strategy (MIXED)

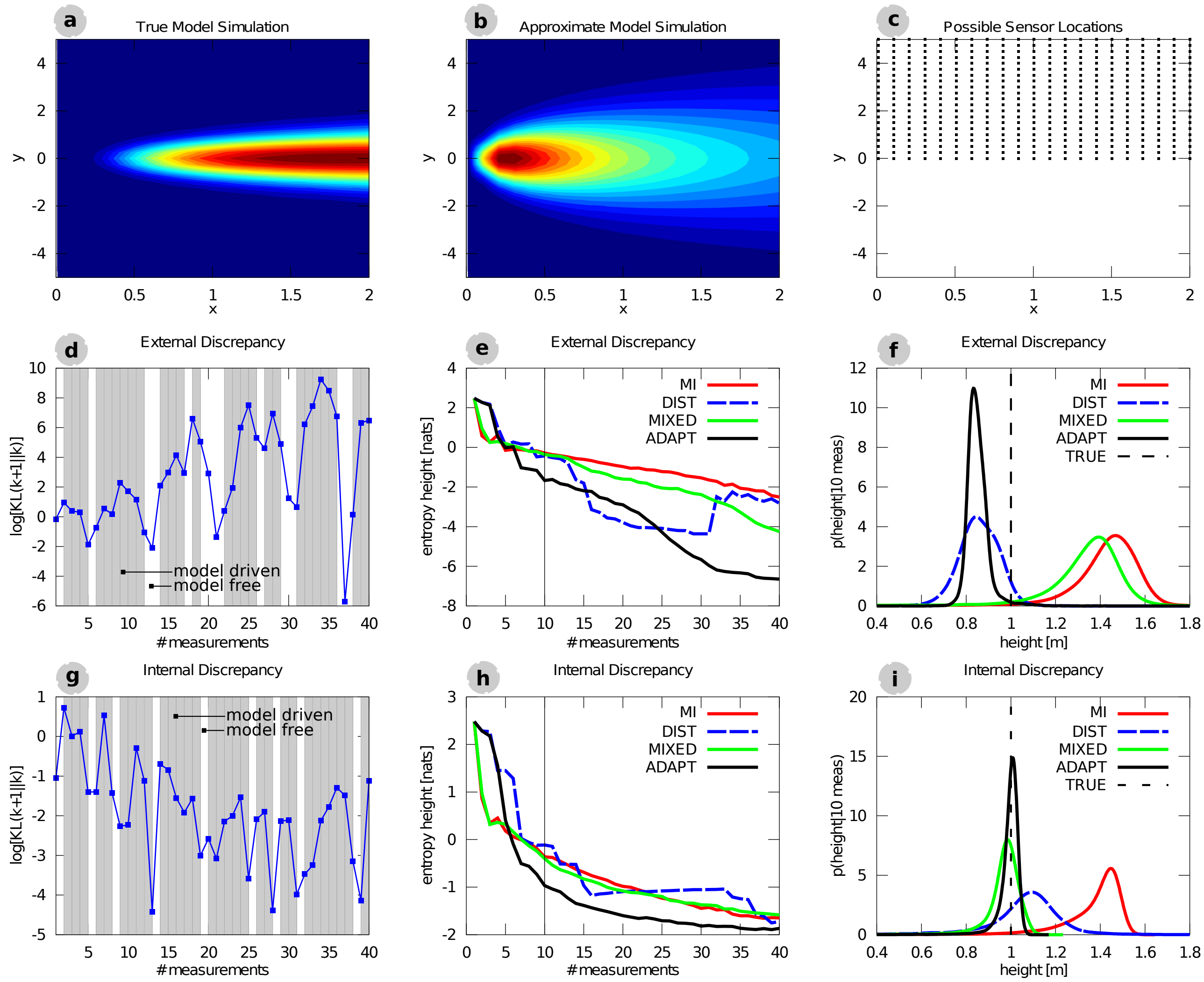


Fig. 2: Preliminary results: (a) Simulation using the true model with eddy diffusion coefficient as a linear function of downwind distance, (b) Simulation using the approximate model which assumes constant eddy diffusion coefficient (here $K = 1$), (c) The complete set of possible sensor locations to be chosen by the design strategies, (d/g) External/Internal discrepancy: the evolution of the Kullback-Leibler divergence between subsequent posterior distributions and the switching between model-driven (α small) and model-free (α large) as it is produced by the adaptive strategy (e/h) External/Internal discrepancy: the entropy of the release height after each new measurement (results averaged over 30 runs), (f/i) External/Internal discrepancy: posterior pdf of the release height after 10 observations (one sensor provides just one observation).

obtained by setting $\alpha = 1$ in Eq.(12), and an adaptive hybrid strategy (ADAPT) where α is set to a high value when the Kullback-Leibler divergence between consecutive posteriors is significantly decreased, see Fig.2(d,g). In other words, if the rate of learning has significantly decreased then switch to design space exploration to find new opportunities for learning, otherwise exploit the model.

In both formulations the adaptive hybrid strategy obtains overall a faster reduction in the uncertainty of the release height, see Fig.2(e,h). In contrast to the external discrepancy formulation, the true height is better captured by the internal discrepancy formulation as it satisfies the physical constraints imposed by the advection-diffusion equation (conservation of mass), see Fig.2(f,i).

VI. Conclusions

The focus of this work is to develop a basic understanding of the impact that modeling errors have on experimental design strategies. These strategies are used to select experimental conditions that provide critical observations to reduce uncertainties in computational models. Through a rigorous modeling of structural errors, new adaptive experimental design strategies can be obtained by exploiting structural uncertainty.

This is significant in applications where physical and financial constraints impede exhaustive data collection.

While preliminary results are promising, a number of questions need to be answered to refine the proposed methodology: (1) Will this adaptive strategy provide consistent and better performance for various dispersion models? (2) The example only addresses the calibration goal. Does it also meet the model selection and prediction goals? and (3) Along the same lines, is one adaptive strategy suitable for both discrepancy formulations in general? The answer to all these questions is planned to be investigated next.

ACKNOWLEDGMENTS

This work is partially supported by an ASPIRE grant from the Office of the Vice President for Research at the University of South Carolina.

REFERENCES

[1] D. V. Lindley, "On a measure of the information provided by an experiment," *Ann. Math. Statist.*, vol. 27(4), pp. 986–1005, 1956.

[2] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statistical Science*, vol. 10(3), pp. 273–304, 1995.

[3] D. J. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.

[4] G. Terejanu, R. R. Upadhyay, and K. Miki, "Bayesian experimental design for the active nitridation of graphite by atomic nitrogen," *Experimental Thermal and Fluid Science*, vol. 36, pp. 178–193, Jan. 2012.

[5] T. J. Santner, B. J. Williams, , and W. I. Notz, *The Design and Analysis of Computer Experiments*. Springer Verlag, New York, 2003.

[6] T. Guest and A. Curtis, "Iteratively constructive sequential design of experiments and surveys with nonlinear parameter-data relationships," *Journal of Geophysical Research*, vol. 114, p. B04307, 2009.

[7] L. Paninski, "Asymptotic theory of information-theoretic experimental design," *Neural Computation*, vol. 17, pp. 1480–1507, 2005.

[8] M. Clyde, P. Muller, and G. Parmigiani, *Bayesian Biostatistics*. Dekker, New York, 1996, ch. Inference and Design Strategies for a Hierarchical Logistic Regression Model, pp. 297–320.

[9] M. Chung and E. Haber, "Experimental design in biological systems," *SIAM Journal on Control and Optimization*, vol. 50(1), p. 471489, 2011.

[10] L. Horesh, E. Haber, and L. Tenorio, *Large-Scale Inverse Problems and Quantification of Uncertainty*. Wiley, 2011, ch. Optimal Experimental Design for the Large-Scale Nonlinear Ill-posed Problem of Impedance Imaging, pp. 273–290.

[11] M. A. A. Tucker, "Application of design of experiments to flight test: A case study," in *AIAA*, Los Angeles, California, February 2008, pp. 2008–847.

[12] C. Kreutz and J. Timmer, "Systems biology: experimental design," *FEBS Journal*, vol. 276, no. 4, pp. 923–942, 2009.

[13] X. Huan and Y. M. Marzouk, "Simulation-based optimal bayesian experimental design for nonlinear systems," *J. Comput. Phys.*, vol. 232, no. 1, pp. 288–317, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jcp.2012.08.013

[14] F. Ciucci, T. Carraro, W. C. Chueh, and W. Lai, "Reducing error and measurement time in impedance spectroscopy using model based optimal experimental design," *Electrochimica Acta*, vol. 56, no. 15, pp. 5416 – 5434, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0013468611003252

[15] Q. Long, M. Scavino, R. Tempone, and S. Wang, "Fast estimation of expected information gains for bayesian experimental designs based on laplace approximations," *Computer Methods in Applied Mechanics and Engineering*, vol. 259, pp. 24–39, 2013.

[16] E. Haber, Z. Magnant, C. Lucero, and L. Tenorio, "Numerical methods for a-optimal designs with a sparsity constraint for ill-posed inverse problems," *Computational Optimization and Applications*, vol. 52(1), pp. 293–314, 2012.

[17] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, "A-optimal design of experiments for infinite-dimensional bayesian linear inverse problems with regularized ℓ_0-sparsification," *To appear in SISC*, 2014. [Online]. Available: arxiv.org/abs/1308.4084

[18] P. Muller, *Bayesian Statistics 6*. Oxford University Press, 1999, ch. Simulation-based optimal design, pp. 459–474.

[19] N. Rao, D. Reister, and J. Barhen, "Information fusion methods based on physical laws," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 1, pp. 66–77, Jan 2005.

[20] D. Turner, *Workbook of atmospheric dispersion estimates: an introduction to dispersion modeling (2nd Edition ed.)*. CRC Press, 1994.

[21] G. S. Settles, "Fluid mechanics and homeland security," *Annu. Rev. Fluid Mech.*, vol. 38, pp. 87–110, 2006.

[22] R. P. Christian, *The Bayesian Choice*. Springer, 2001.

[23] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[24] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001. [Online]. Available: http://dx.doi.org/10.1111/1467-9868.00294

[25] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[26] J. Brynjarsdottir and A. O'Hagan, "Learning about physical parameters: The importance of model discrepancy," *Submitted to SIAM/ASA Journal of Uncertainty Quantification*, 2013.

[27] T. A. Oliver, G. Terejanu, C. S. Simmons, and R. D. Moser, "Validating predictions of unobserved quantities," *Computer Methods in Applied Mechanics and Engineering*, vol. 283, no. 0, pp. 1310 – 1335, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S004578251400293X

[28] A. S. R Jin, W Chen, "On sequential sampling for global metamodeling in engineering design," in *Proceedings of the ASME Design Automation Conference*, 2002.

[29] A. DasGupta, "Review of optimal bayes designs," Purdue University, Tech. Rep., 1995.

[30] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[31] P. Sebastiani and H. P. Wynn, "Maximum entropy sampling and optimal Bayesian experimental design," *J. R. Statist. Soc. B*, vol. 62, Part 1, pp. 145–157, 2000.

[32] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October, 1948.

[33] L. Pronzato and W. G.Muller, "Design of computer experiments: space filling and beyond," *Statistic and Computing*, vol. 22, no. 3, pp. 681–701, 2012.

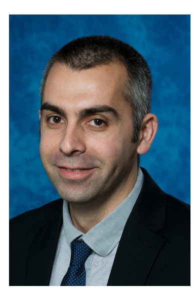
[34] P. Z. G. Qian, "Sliced latin hypercube designs," *Journal of the American Statistical Association*, vol. 107, pp. 393–399, 2012.

[35] J. Sacks, W. J. Welch, T. J. Mitchell, , and H. P. Wynn, "Design and Analysis of Computer Experiments," *Statistical Science*, vol. 4, no. 4, pp. 297–437, 1989.

[36] J. M., M. L., and Y. D., "Minimax and maximin distance designs," *Journal of tatistical planning and inference*, vol. 26, pp. 131–148, 1990.

[37] J. Stockie, "The mathematics of atmospheric dispersion modeling," *SIAM Review*, vol. 53, no. 2, pp. 349–372, 2011.

[38] M. P. Singh and A. K. Yadav, "Mathematical model for atmospheric dispersion in low winds with eddy diffusivities as linear functions of downwind distance," *Atmospheric Environment*, vol. 30, no. 7, pp. 1137–1145, 1996.



Gabriel Terejanu received his B.E. in automation from University of Craiova, Romania in 2004, and M.S. and Ph.D. in computer science and engineering from University at Buffalo, in 2007 and 2010 respectively. In 2010, he was awarded a postdoctoral fellowship for two years at the Institute for Computational Engineering and Sciences (ICES) at University of Texas at Austin. He has developed several algorithms for nonlinear filtering, information fusion, and data collection. His research interests are in information fusion, model validation, uncertainty quantification, and machine learning.