Metamodelling in Multidisciplinary Design Optimization: How Far Have We Really Come?

Felipe A. C. Viana  
University of Florida, Gainesville, Florida 32611  
Timothy W. Simpson  
Pennsylvania State University, University Park, Pennsylvania 16802  
Vladimir Balabanov  
The Boeing Company, Seattle, Washington 98275  
and  
Vasilli Toropov  
University of Leeds, West Yorkshire, England LS2 9JT, United Kingdom  
DOI: 10.2514/1.J052375

The use of metamodelling techniques in the design and analysis of computer experiments has progressed remarkably in the past 25 years, but how far has the field really come? This is the question addressed in this paper, namely, the extent to which the use of metamodelling techniques in multidisciplinary design optimization have evolved in the 25 years since the seminal paper on design and analysis of computer experiments by Sacks et al. (“Design and Analysis of Computer Experiments,” *Statistical Science*, Vol. 4, No. 4, 1989, pp. 409–435). Rather than a technical review of the entire body of metamodelling literature, the focus is on the evolution and motivation for advancements in metamodelling with some discussion on the research itself; not surprisingly, much of the current research motivation is the same as it was in the past. Based on current research thrusts in the field, multifidelity approximations and ensembles (i.e., sets) of metamodels, as well as the availability of metamodels within commercial software, are emphasized. Design space exploration and visualization via metamodels are also presented as they rely heavily on metamodels for rapid design evaluations during exploration. The closing remarks offer insight into future research directions, mostly motivated by the need for new capabilities and the ability to handle more complex simulations.

Felipe A. C. Viana received his first Ph.D. degree in mechanical engineering from the Federal University of Uberlandia in 2008. He earned a second doctorate, in aerospace engineering, from the University of Florida in 2011. Currently, Dr. Viana is a member of the Probabilistics Laboratory at GE Global Research. His responsibilities include delivering state-of-the-art probabilistic design methods and tools, and driving applications in aero thermal and mechanical systems from new designs to fielded product quality improvement across GE businesses. The research of Dr. Viana has produced several papers in the top journals and conferences in the field of probabilistic design and optimization. Dr. Viana is a Member AIAA.

Timothy W. Simpson is a Professor of mechanical and industrial engineering at Pennsylvania State University with affiliations in Engineering Design and the College of Information Sciences and Technology. He received his B.S. in mechanical engineering from Cornell University in 1994 and his M.S. and Ph.D. degrees in mechanical engineering from Georgia Institute of Technology in 1995 and 1998. He has received over $15 million in funding to support his research in product family and product platform design, multidisciplinary design optimization (MDO), metamodelling and trade space exploration, and additive manufacturing, and he has published over 250 peer-reviewed papers to date. He teaches courses on product family design, concurrent engineering, mechanical systems design, and product dissection, and he serves as the Director of the Product Realization Minor in the College of Engineering. He is a recipient of the American Society for Engineering Education Fred Merryfield Design Award and a National Science Foundation Career Award. He has received several awards for outstanding research and teaching at Pennsylvania State University, including the 2007 Penn State University President’s Award for Excellence in Academic Integration. He is a Fellow of the American Society of Mechanical Engineers (ASME) and an Associate Fellow AIAA. He currently serves on the ASME Design Education Division Executive Committee and is a former Chair of both the ASME Design Automation Executive Committee and the AIAA MDO Technical Committee. He is also a Department Editor for IEEE Transactions: Design and Manufacturing and serves on the editorial boards for *Research in Engineering Design*, *Journal of Engineering Design*, and *Engineering Optimization*.

Presented as Paper 2008-5802 at the AIAA Multidisciplinary Analysis and Optimization Conference, Victoria, British Columbia, Canada, 10–12 September 2008; received 19 January 2013; revision received 31 July 2013; accepted for publication 1 August 2013; published online 31 January 2014. Copyright © 2013 by the American Institute of Aeronautics and Astronautics, Inc. All rights reserved. Copies of this paper may be made for personal or internal use, on condition that the copier pay the $10.00 per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923; include the code 1533-385X/14 and $10.00 in correspondence with the CCC.
**Nomenclature**

\[
\begin{align*}
F(x) & = \text{high-fidelity model at a current design point} \\
F(x) & = \text{approximation to the high-fidelity model at a current design point} \\
f(x) & = \text{low-fidelity model at a current design point}
\end{align*}
\]

**I. Introduction**

The phrase “design and analysis of computer experiments,” or DACE as it has become known, was introduced 25 years ago by Sacks et al. [1]. DACE embraces the set of methodologies for generating a metamodel, or “model of the model” as defined by Kleijnen [2], used to replace a generally expensive computer code. Metamodels are also used to facilitate the integration of discipline-specific computer analyses and can provide better insight into the relationships between design (input) variables and system performance (output) responses. In structural and multidisciplinary optimization, this approach is much older. In fact, approximation methods have their remote origins in structural synthesis [3,4] and have been applied to a wide variety of structural design problems [5].

In the general scope of metamodeling, the response may be evaluated via a physical experiment or a computer simulation at a number of points in the domain. However, our review focuses on work dealing primarily with deterministic computer simulations. This does not mean that computer experiments are always noise free. Instead, it only says that we will focus on those simulations for which the same set of inputs generates the same set of outputs. Thus, we differentiate DACE from traditional design of experiments (DOE), which was developed primarily for performing physical experiments [6] that are inherently noisy. To the interested reader, the literature has highlighted that some metamodeling techniques offer the benefit of smoothing numerically “noisy” data [7–9], which can hinder the convergence of many optimization algorithms.

In this paper, we consider metamodeling to be the general process of creating a computationally inexpensive abstraction through the form of either an approximation or an interpolation of data gathered over a certain domain (with a well-defined set of inputs and outputs). Popular metamodeling techniques include polynomial response surface, neural networks, kriging, radial basis functions, spline, support vector regression, and moving least squares, among others. The common feature of all these approaches is that the actual response is known at a finite number of points. However, the metamodel is created to be used as a surrogate for the original model over a certain domain (i.e., it provides a substitute for and is used in lieu of the original computer model). In the literature (and also in this work), surrogate model is synonymous with metamodel. Metamodeling techniques are often classified as being either global or local [5]: global approximations are valid throughout the entire design space (or a large portion of it), whereas local approximations are only valid in the vicinity of a particular point. Midrange approximations also exist for creating local approximations with global qualities [10].

Our goal in this review is to investigate how far metamodeling techniques have evolved since the introduction of DACE more than two decades ago. We do not aim to provide a comprehensive review given the literature reviews that have appeared recently as summarized in the next section. Also, our focus is primarily on the surrogate modeling techniques themselves, not the experimental designs used to generate sample data; interested readers are referred to recent overviews and texts on the topic [11,12]. We begin with a historical perspective in Sec. II to explore how the research, and more importantly how the motivation for the research, has evolved these past 25 years. We then discuss four research directions that have benefited from metamodeling while also driving research in the area: 1) multifidelity approximations (Sec. III), 2) the use of multiple surrogates and metamodel ensembles (Sec. IV), 3) metamodeling capabilities in commercial software packages (Sec. V), and 4) metamodel-based design space exploration and visualization (Sec. VI). Section VII highlights future work, much the same as what has been motivating us for the past two decades.

**II. History of Development**

A. Origins and Early Uses

The prohibitive computational cost of the direct combination of finite element models with methods of mathematical programming.

---

Vladimir Balabanov received an M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology and an M.S. degree in aerospace engineering from Virginia Polytechnic Institute and State University. He holds a Ph.D. in aerospace engineering from Virginia Polytechnic Institute and State University. Dr. Balabanov worked at the Central Aerohydrodynamic Research Institute at Vanderplaats Research and Development, Inc., and currently works at The Boeing Company. He is also an Adjunct Professor at the University of Colorado at Colorado Springs. Dr. Balabanov organized, chaired, and gave presentations at numerous technical conferences. He published more than 60 papers in the conference proceedings and refereed journals. He is an Associate Fellow AIAA and serves on the AIAA Design Engineering Technical Committee and Multidisciplinary Design and Optimization Technical Committee. Dr. Balabanov has worked in the fields of aerodynamic, commercial optimization software design and development, and methods development for design and analysis of aircraft composite structures. His interests are in the field of general-purpose optimization, structural optimization, probabilistic optimization, design of experiments, response surface approximation methods, and their application in industry. Vladimir Balabanov is an award-winning public speaker.

Vassili Toropov, Professor of Aerospace and Structural Engineering, is a leader of a multidisciplinary optimization research group at the University of Leeds (Leeds). Before joining Leeds in 2006, he spent three years at Altair Engineering engaged in the development of techniques and software for metamodeling, design optimization, and stochastic analysis as well as applying his expertise to industrial problems for aerospace, automotive, offshore, and other clients. Vassili has published over 250 research papers, and his interests include techniques for metamodeling, design of experiments, stochastic problems, composite optimization, evolutionary optimization, and industrial applications. Vassili’s recent and current research projects are funded by Rolls–Royce, plc.; Airbus Industries; EADS Innovation Works; Parker Hannifin; Jaguar Land Rover; and the European Commission ranging from optimization of aerospace structures to optimum flow and thermal management. Professor Toropov is a chartered aerospace engineer, Fellow of the Royal Aeronautical Society, Chairman of the Association for Structural and Multidisciplinary Optimization in the United Kingdom, Executive Committee Member of the International Society for Structural and Multidisciplinary Optimization, a Member of the AIAA Multidisciplinary Optimization and Non-Deterministic Approaches Technical Committees, and a Coeditor of the *Structural and Multidisciplinary Optimization* journal. He as an Associate Fellow AIAA.
stimulated the idea of approximation concepts based on the information from the first-order design sensitivity analysis [3]. Since then, this concept of sequential approximation of the initial optimization problem by explicit subproblems has proven to be very efficient. As examples the methods can be named such as the sequential linear programming used for structural optimization problems by Pedersen [13], the convex linearization method by Fleury and Brabant [14] and Fleury [15,16], and the method of moving asymptotes by Svanberg [17]. All of these methods use the information obtained from response analysis and first-order design sensitivity analysis (i.e., values of functions and their derivatives) at a current point of the design variable space and hence can be classified as single-point approximation methods. Note that all the information from previous design points is discarded. Later on, several first-order approximation techniques have been developed based upon the function value and its derivatives at the current and the previous design points (two-point approximations) [18,19]. The main purpose is to improve the quality of approximations and thus reduce the number of iterations needed to solve the optimization problem and the total optimization time. Rasmussen [20] developed this idea further. His accumulated approximation technique used function values and derivatives at a current design point and the function values obtained at all previous points. Toropov [21] introduced a technique that used function values gained in each iteration at several previous design points (multipoint approximations). The aim was to combine the benefits of the two basic approaches; hence, the technique can be classified as a midrange approximation. Later this technique was expanded to incorporate the design sensitivities (when available) into the approximation building [22].

A different approach to structural optimization [23] is to create approximate explicit expressions by analyzing a chosen set of design points and using response surface methodology. This approach is based on the multiple regression analysis that can use information being more or less inaccurate. They are global in nature and allow designers to construct explicit approximations valid in the entire design space. However, they are restricted by relatively small optimization problems (up to 10 design variables [23]). This approach was used for solving various structural optimization problems by Schoofs [24], Vanderplaats [25], and Rikards [26].

Before 1990, polynomial response surface models, first introduced by Box and Wilson [26] and later detailed in [27], and neural networks [28] were among the most popular approximation methods. Early contributors to approximation method development for multidisciplinary design optimization (MDO) include the research groups at the Virginia Polytechnic Institute and State University (Virginia Tech), the University of Notre Dame, Rensselaer Polytechnic Institute, Old Dominion University, and the NASA Langley Research Center. A review of applications of response surface models, neural networks, and other types of approximations in MDO during this time can be found in [5,29].

Interest in approximation methods and metamodeling techniques grew substantially in the 1990s, particularly within the MDO community. During the first half of the decade, heavy emphasis was placed on response surface methods, which primarily resulted from NASA-funded research related to the High Speed Civil Transport (HSCT). Researchers at the Virginia Tech, University of Notre Dame, Georgia Institute of Technology (Georgia Tech), Rice University, and Old Dominion University continued to advance the state of the art by developing novel methods and uses for response surface models. Many of these efforts were chronicled at the 1995 MDO Workshop sponsored by the Institute for Computer Applications in Science and Engineering/NASA Langley [30–34]. For instance, the variable complexity response surface modeling method developed predominantly at Virginia Tech [32,35,36] uses analyses of varying fidelity to reduce the design space to the region of interest and build response surface models of increasing accuracy. The concurrent subspace optimization (CSSO) procedure from Notre Dame uses data generated during concurrent subspace optimization to develop response surface approximations of the design space, which form the basis of the subspace coordination procedure during MDO [37–39]. Robust design simulation [40] and the robust concept exploration method [41,42] were developed at Georgia Tech to facilitate quick evaluation of different design alternatives and generate robust top-level design specifications. Hafka et al. [43] and Simpson et al. [44] provide extensive reviews of response surface and approximation methods in mechanical and aerospace engineering during this timeframe.

As response surface modeling became more widely used and better understood, its limitations became more apparent, e.g., the “curse of dimensionality” [45,46] and the inability to create accurate global approximations in highly nonlinear design spaces [47]. As a result, some researchers started exploring higher-order response surface models [48] and mixed polynomial models [49], whereas others investigated more efficient experimental designs for sampling the design space using computer analyses [12,50,51]. Other researchers also started investigating the use of gradient information to facilitate metamodel construction [52–54]. Sequential approaches to sampling, building, and optimizing approximation models were also being investigated by many researchers, and the use of move limits [55] and trust region approaches [56,57] were being advocated by many researchers for sequential metamodeling. This led to the development of mathematically rigorous techniques to manage the use of approximation models in optimization such as the surrogate management framework [58], developed collaboratively by researchers at Boeing, IBM, and Rice University. During this time frame, many companies also started to develop software to facilitate the use of approximation methods in design and optimization: iSIGHT [59] by Engineous Software, Inc.; Visual DOC [60] by Vanderplaats R&D, Inc.; Optimus [61] by LMS International; ModelCenter [62] by Phoenix Integration; Design Explorer [58] by The Boeing Company; and DAKOTA [62] by Sandia National Laboratories. Section V gives a more detailed review of the metamodeling capabilities within each of these software packages.

B. Recent Developments

By the end of the 1990s, the vigorous growing of applications motivated collaborative work toward metamodeling frameworks for optimization and approximation. One example is the birth of the efficient global optimization (EGO) algorithm, a joint effort involving General Motors R&D, the National Institute of Statistical Sciences, and the University of Waterloo [64]. As occurred in the early origins, much of the developments also came by tailoring statistical methods to the design and analysis of computer experiments. Another example is the incorporation of Bayesian techniques in the repertoire. By the early 1990s, Bayesian techniques were viewed mostly in statistics outlets, such as the Journal of the American Statistical Association and Statistical Science (see works of Currin et al. [65] and Chaloner and Verdinelli [66]). Along the past decade it has become more and more common to see Bayesian techniques in engineering oriented outlets, such as the American Society of Mechanical Engineers Design Engineering Technical Conference and Engineering and System Safety (for example, [67–69]).

During the past decade, we also saw the emphasis shifting away from the traditional response surface models to alternative approximation methods such as radial basis functions [70–72], multivariate adaptive regression splines [73], kriging [1,74], support vector regression [75,76], and, recently, variations on moving least squares [77–79]. These techniques were the focus in many dissertations [80–83]. As in the past, we did see the importation of technologies and the proper tailoring to the multidisciplinary optimization problems [84]. In the past we imported response surface techniques from statistics and agricultural sciences; now, we import kriging from geostatistics and mining engineering. As a matter of fact, kriging was one of the techniques that attracted a lot of attention during the past decade. Substantial work was done on the tuning of the kriging parameters [85,86], the update of the prediction variance [87,88], the use of kriging in variable fidelity optimization [89,90], and sequential sampling and global optimization [64,91–93].
The vast pool of metamodeling techniques has led to many comparative studies to determine the advantages of different techniques [94–96]. Instead of a clear conclusion, the literature confirms the suspicion that the surrogate performance depends on both the nature of the problem and the DOE. The merits of various metamodeling techniques were vastly discussed at the Approximation Methods Panel held at the Ninth AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization [97]. As discussed in the panel, metamodels are finding a variety of new uses, including optimization under uncertainty, which is receiving considerable attention as of late [98], and robust design and reliability-based design because they provided inexpensive surrogates for Monte Carlo simulation and uncertainty analysis [99,100].

Combining surrogates has also been brought recently to the engineering optimization community. Zerpa et al. [101] used multiple surrogates for optimization of an alkaline-surfactant-polymer flooding processes incorporating a local weighted average model of the individual surrogates. Goel et al. [102] explored different approaches in which the weights associated with each surrogate model are determined based on cross validation. Acar and Rais-Rohani [103] studied weight selection via optimization. The universal kriging can also be thought as a type of “ensemble.” Universal kriging combines a polynomial term with a weighted sum of basis functions (ordinary kriging is the one without any trend function). More recently, a method called blind kriging has been developed [104]. The terms of the polynomial for a universal kriging prediction are chosen using a Bayesian forward selection criterion and cross validation. The results are promising and the method dovetails with the EGO implementation.

C. Observations on DACE-Related Literature

Figure 1 illustrates the volume and impact of DACE-related publications in the past 25 years. The data were obtained using the Publish or Perish software system and Google Scholar set for search occurrences in the subject area of “engineering, computer science, and mathematics.” Our goal is just illustrate the growth of DACE-related literature, instead of a comprehensive and detailed analysis of such literature. Although the specific numbers may vary as the Google database is updated, we see a steady growth in the number of papers on DACE. Figure 1a shows an approximate number of publications the exact phrase “design and analysis of computer experiments.” Figure 1b complements Fig. 1a, showing the number of citations that publications received versus the year of publication. In each year, at least one paper is likely to receive more than 100 citations. In fact, 100 papers have more than 100 citations and 44 more than 200 citations (some of which are exemplified in Fig. 1b). It is clear that the focus on DACE rewards authors with citations.

Figure 2 (similar to one shown in [105]) illustrates the number of publications reporting the use of four types of surrogate techniques. Although other equally popular techniques could be included (e.g., radial basis functions), we believe these four illustrate the diversity of surrogates used in practice. That is, rather than a thorough analysis of the scientific production of works on each type of metamodel, our goal is to illustrate the growth in the reported literature using different examples of surrogate techniques. Data were also obtained using the Publish or Perish software system and Google Scholar. Table 1 details the setup used in the search. These results may also vary due to the update of the Google database. Figure 2a shows a steady growing of publications for both kriging and response surface and an impressive growing for both neural networks and support vector machine. Our observation is that the disparity could be explained by the use of both neural networks and support vector machine in areas like classification and control (besides structural and multidisciplinary optimization). Figure 2b illustrates the results when we also include in the search the expression “structural and multidisciplinary optimization.” We can clearly see that the 1990s trigged the popularity of surrogate techniques. Figure 2c shows the volume of publications in yet another subarea of interest (probabilistic analysis). Here, the search was constrained to include the expression “probabilistic analysis,” and again, we see the growing interest since the 1990s. Response surfaces (or the term “response surface”) still seem to be the most used in structural and multidisciplinary optimization as well as probabilistic analysis. Regardless of the surrogate technique, we believe that 1) the popularity of DACE is interconnected to the developments of individual surrogate modeling techniques and availability in commercial software packages and 2) different surrogates appear to be competitive and equally good given the wide spectrum of regression applications [106].

Now, consider Table 2 [1,5,29,44,97,107–109], which summarizes frequently cited review papers published in the past two decades. We would like to use it to draw few hypotheses about the motivations behind research in metamodeling. The evident common theme in all of these papers is the high cost of computer simulations: despite growing in computing power, surrogate models are still cheaper alternatives to actual simulation models in engineering.
design. This statement alone is weak in the sense that one could argue that, with the computational resources we have today, instead of using surrogate models for approximation purposes, we could rely on low-fidelity codes (or, alternatively, codes that used to be high-fidelity models in the past but that would run much faster on today’s computers). The devil’s advocate would say that both surrogate and the cheap low-fidelity models would bring a level of uncertainty (compared with the current high-fidelity simulations). However, the latter would have the advantage of being already well known and used (besides, it would save one to go to the sampling issues on surrogate modeling).

So what have the advances in computational throughput been used for in different segments of the engineering design? Figure 3 helps us to see that increased computational capabilities have been used to advance optimization and metamodeling as much as they were used to add fidelity to state-of-the-art simulations [110]. Thus, it is likely that one could not use old (legacy) high-fidelity models to obtain the output of some of today’s codes (take the example of the linear static versus transient analyses). On the other hand, we can say that the cost of fitting a given surrogate compared to the cost of simulations has reduced dramatically over time, which helped to make popular more sophisticated surrogates. On top of that, surrogates such as kriging models and even the traditional polynomial response surface also offer information about the prediction error (obviously not found in legacy codes). More than just pointwise uncertainty estimator, recent developments in metamodeling have shown that these structures can be used for rational allocation of the computational resources in the surrogate building and in the optimization process itself. For example, they can guide the refinement of the design space toward regions of high uncertainty or regions in which the optimization is mostly likely to improve upon the present best solution (in terms of the objective function value). This is the case of algorithms such as the efficient global optimization [64] and enhanced sequential optimization algorithms [92] and mode pursuing sampling [111–113]. The bottom line is that the repertoire of design tools has substantially grown over the years and DACE methods help tailoring problem-oriented approaches.

Now, we can go back to fundamental question of this paper “how far we have come, or not.” Table 3 summarizes what the review papers listed in Table 2 reported as successfully achieved and what they pointed as challenges. On the column “How far we have come” (third column), we can see that the community devoted a lot of effort to establishing the basics of design and analysis of computer experiment. There was substantial effort in topics like the suitable sampling schemes. Effort was also substantial in the development of several metamodeling techniques and the assessment of the quality of the fit. On the other hand, the column “Or not” (last column) shows that the definition of the metamodel (either by selecting the metamodeling technique or its parameters) is a theme that constantly attracts interest from the community. Another never-ending problem is the curse of dimensionality. Surprisingly, global optimization, even with the advances driven by kriging, is still one of the topics that will benefit from future research in design and analysis of computer experiments.

### III. Multifidelity Approximations

In the past 25 years, we have observed a sharp increase in the number of papers on metamodeling approaches based on the interaction of high- and low-fidelity numerical models. In such approaches, there is an assumption that a high-fidelity model is more
applications of nonlinear programming methods to large structural design problems could prove cost effective, provided that suitable approximation concepts were introduced.

Simpson et al. [44] 2001 Abstract: “Computer-based simulation and analysis is used extensively in engineering for a variety of tasks. Despite the steady and continuing growth of computing power and speed, the computational cost of complex high-fidelity engineering analyses and simulations maintains pace... Consequently, approximation methods such as design of experiments combined with response surface models are commonly used in engineering design to minimize the computational expense of running such analyses and simulations.”

Queipo et al. [107] 2005 Abstract: “A major challenge to the successful full-scale development of modern aerospace systems is to address competing objectives such as improved performance, reduced costs, and enhanced safety. Accurate, high-fidelity models are typically time consuming and computationally expensive. Furthermore, informed decisions should be made with an understanding of the impact (global sensitivity) of the design variables on the different objectives. In this context, the so-called surrogate-based approach for analysis and optimization can play a very valuable role. The surrogates are constructed using data drawn from high-fidelity models, and provide fast approximations of the objectives and constraints at new design points, thereby making sensitivity and optimization studies feasible.”

Wang and Shan [108] 2007 Abstract: “Computation-intensive design problems are becoming increasingly common in manufacturing industries. The computation burden is often caused by expensive analysis and simulation processes in order to reach a comparable level of accuracy as physical testing data. To address such a challenge, approximation or meta-modeling techniques are often used.”

Forrester and Keane [109] 2009 Abstract: “The evaluation of aerospace designs is synonymous with the use of long running and computationally intensive simulations. This fuels the desire to harness the efficiency of surrogate-based methods in aerospace design optimization.”

accurate but requires a large computing effort whereas a low-fidelity model is less accurate but is considerably less computationally demanding. Such a model can be obtained by simplifying the analysis model (e.g., by using a coarser finite element mesh discretization, a reduced number of the natural modes of the model in dynamic analysis, etc.) or a modeling concept [e.g., simpler geometry, boundary conditions, two-dimensional (2-D) instead of a three-dimensional (3-D) model, etc.]. A low-fidelity model can provide a basis for a high-quality metamodel building resulting in solving an optimization problem to the accuracy of the high-fidelity model at a considerably reduced computational cost. In the metamodel building, a low-fidelity model is corrected (or tuned) using the model response values from a relatively small number of calls for both high-fidelity and low-fidelity models according to a suitable design of experiments. Such tuning can be refined in an adaptive way as optimization progresses. The overall objective of this approach is to attempt to circumvent the curse of dimensionality associated with black-box metamodeling by exploiting domain-specific knowledge [114].

In some cases, there can be a hierarchy of numerical models, e.g., based on Navier–Stokes equations (highest fidelity and most expensive), on Euler equations (lower fidelity and less expensive), linear panel method (lower fidelity and cheaper), etc., down to analytical or empirical formulas (e.g., obtained from the wind-tunnel test data). These can also be exploited in an optimization strategy with hierarchic metamodel building and refinement.

Originally, the idea of improving the quality of an approximation by endowing it with some discipline-related properties of the underlying numerical function stems from the empirical model-building theory. For instance, Box and Draper [115] showed that a mechanistic model, i.e., the one that is built upon some knowledge about the system under investigation, can provide better approximations than general ones, e.g., polynomials. An example of application of such a model to a problem of material parameter identification (formulated as an optimization problem) was given by Toropov and van der Giessen [116], in which the structure of the metamodels of response quantities (torque and elongation) for a solid bar specimen in torsion, obtained by a nonlinear finite element (FE) simulation, was derived from a simpler functional relationship based on material parameters for a tubular specimen. The simplified (thin-walled tubular) model, analyzed by solving an ordinary differential equation, was the basis for the metamodel describing the behavior of a solid model that was analyzed by a much more complex numerical simulation. The radii of the artificial tube specimen were treated as metamodel parameters used to match the two models at sampling points.

A different route to introducing approximations based on the interaction of high- and low-fidelity models was taken by Hafkka...
Table 3  Perspective from previous review papers on how far we have come

<table>
<thead>
<tr>
<th>Paper</th>
<th>Year</th>
<th>How far we have come</th>
<th>Open questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sacks et al. [1]</td>
<td>1989</td>
<td>1) Introduced kriging to computer experiments</td>
<td>1) Raised the question of the sensitivity of the estimated kriging parameters with respect to the data set</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Discussed on the sampling schemes for computer experiments</td>
<td>2) What is the choice of optimality criterion for design of computer experiments?</td>
</tr>
<tr>
<td>Barthelemy and Haftka [5]</td>
<td>1993</td>
<td>1) Made distinction between local, medium-range, and global approximations</td>
<td>Found no answer for the question of which approximation technique is better suited for a given problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Put emphasis on local function approximations and use of intermediate variable and response quantities</td>
<td></td>
</tr>
<tr>
<td>Sobieszczanski-Sobieski and Haftka [29]</td>
<td>1997</td>
<td>Established metamodeling as a way to facilitate the interaction between disciplines and computer codes in multidisciplinary optimization (especially in aerospace design)</td>
<td>Found that little interaction between the aerospace multidisciplinary optimization research community and other engineering research communities makes propagation of the technologies slow</td>
</tr>
<tr>
<td>Simpson et al. [44]</td>
<td>2001</td>
<td>1) Reported the development of several metamodeling techniques and sampling schemes for computer experiments</td>
<td>Raised the problem imposed by the high dimensionality and nonlinearity (typically found in computer experiments)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Discussed first attempts of metamodeling technique selection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1) Discussed the design and analysis of computer experiments as an engineering tool; experts from industry, universities, and national research labs debated the impact of this technology</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Reported applications in industry</td>
<td></td>
</tr>
<tr>
<td>Queipo et al. [107]</td>
<td>2005</td>
<td>1) Discussed surrogate selection and construction (including loss function and regularization criteria)</td>
<td>1) Identified the need of research in quantification of uncertainty in the surrogate models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Summarized sensitivity analysis with emphasis on the Sobol’s approach</td>
<td>2) Stressed that model selection remains very challenging</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Reviewed surrogate-based optimization with discussions on multiple surrogates, surrogate and approximation model management frameworks, and optimization convergence</td>
<td>3) Pointed out the need for efficient global sensitivity methods for screening (particularly useful in large dimensional problems)</td>
</tr>
<tr>
<td>Wang and Shan [108]</td>
<td>2007</td>
<td>1) Developments on “optimal” design of computer experiments</td>
<td>1) Large-scale problems: is decomposition always the key? what extent does visualization alleviate the curse of dimensionality?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Methods for assessing accuracy of metamodeling techniques</td>
<td>2) Raised the idea that the benefits of sequential sampling for global approximation are not clear yet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Benefits of metamodeling for visualization and design space exploration</td>
<td>3) Identified the importance of uncertainty quantification in metamodeling (especially for constrained optimization)</td>
</tr>
<tr>
<td>Forrester and Keane [109]</td>
<td>2009</td>
<td>1) Showed different metamodeling techniques readily available for computer experiments (even in the presence of numerical noise)</td>
<td>1) Stressed that intelligent strategies for surrogate selection are still a topic that demands research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Reported literature on the use of gradient information for kriging</td>
<td>2) Identified the difficulties in precisely assessing the merit of another cycle in surrogate-based optimization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Recapitulated different global optimization strategies based on metamodeling (including those that take advantage of the error estimation, e.g., the efficient global optimization algorithm [64])</td>
<td></td>
</tr>
</tbody>
</table>

[117] while aiming at extending the range of applicability of a local derivative-based approximation of the high-fidelity response $F(x)$ at a current design point. A global approximation is also introduced that is considered to be a simple-model approximation (i.e., a low-fidelity model) $f(x)$. A scaling factor $C(x_0) = F(x_0)/f(x_0)$ can be calculated at a current design point $x_0$ and its approximation built at a current design point. A global approximation is also introduced that is less expensive than a high-fidelity model, including low-fidelity numerical models, response surfaces, kriging metamodels, etc. Several examples are given, including optimization of a 3-D wing parameterized with 15 design variables in which both high- and low-fidelity (eight times less expensive) models were based on Euler simulation (incorporating an automatic differentiation tool) of different grid refinement achieving a threefold savings in computing effort. An application of this framework, renamed approximation and model management framework, to the optimization of a two-element airfoil using a Reynolds-averaged Navier–Stokes code and an Euler code as high- and low-fidelity simulation tools, respectively [121], resulted in the run time ratio of 55:1 (excluding sensitivity analysis performed by an adjoint approach) and achieved a five-fold savings in computing effort. Second-order correction methods that require second derivatives of the high- and low-fidelity response were introduced by Eldred et al. [122]. They are implemented in the DAKOTA software [63] and compared to other metamodeling techniques [123].

Knill et al. [124] stated that the gradient-based low-fidelity model correction procedure was effective in reducing the computational cost, but it was adversely affected by the presence of numerical noise in aerodynamic and structural response values. To circumvent this, Giunta et al. [125] suggested first performing a thorough design space exploration using a low-fidelity model and then identifying and excluding “nonsense” regions arriving at a much-reduced ribbonlike...
The bending moment evaluated in the high-fidelity model can be either a linear function \( C(x, a) = a_0 + a_1 x_1 + a_2 x_2 \) or a multiplicative function \( C(x, a) = a_0 x_1^{c_1} x_2^{c_2} \). Any other metamodeling approach can also be used to build the approximation of the difference between the high- and low-fidelity function values over the DOE points (linear form) or of the ratio of the high- to low-fidelity function values over the sampling points (multiplicative form). The main difference from the direct application of a conventional metamodeling approach to approximate the values of the high-fidelity response is that the behavior of the correction function is assumed to be simpler (hence requiring fewer sampling points to approximate well) than that of the original high-fidelity response function. It should be noted that the dimension of the tuning parameter space has to be higher than that of the design variable space, at least by one in a case of either linear or multiplicative correction function.

C. Type 3: Use of Low-Fidelity Model Inputs as Tuning Parameters

The third approach consists of considering some physical parameters of the low-fidelity model to be tuning parameters themselves:

\[
\tilde{F}(x, a) \equiv f(x, a)
\]  

(5)

In this case, the physical parameters of the low-fidelity model that are used as tuning parameters will serve as the means of bringing the response from the low-fidelity model as close as possible to those of the high-fidelity model as calculated over a DOE. It should also be noted that this is an implicit approach; the tuning parameters are not assumed to contribute to the low-fidelity model in any particular way, and hence, the dimension of the tuning parameter space can be much lower than that of the design variable space.

Once the metamodel type is chosen, the tuning parameters \( a \) included in the metamodel are obtained by minimizing the sum of squares of the errors over the \( P \) sampling points at which both high- and low-fidelity models have been run:

\[
\minimize \sum_{p=1}^{P} w_p (F(x_p) - \tilde{F}(x_p, a))^2
\]  

(6)

where \( x_p \) is \( p \)th sampling point and \( w_p \) is a weight that determines the relative contribution of the information at that point.

The linear and multiplicative functions of type 1 and type 2 correction functions have been successfully used for a variety of design optimization problems [10]. The linear form of the type 1 metamodel (2) was later expanded by Umakanth et al. [131] to higher-order polynomials and examples were given for the use of quadratic functions \( \tilde{F}(x, a) = a_0 + a_1 f(x) + a_2 f(x)^2 \). Sharma et al. [132] also used the linear form of the metamodel (2), but instead of using the low-fidelity metamodel directly, it was replaced by a cubic response surface built on a sufficiently large number of runs of the low-fidelity model. A comparison of such a metamodel was made with the metamodels obtained by the type 2 correction functions, and it was shown that different metamodel types may perform better than others in different problems. The type 3 metamodel, similarly to mechanistic models, is based on deeper understanding of a process being modeled, which can be useful but is problem dependent. Markine et al. [133] gave an illustrative example of a four-link mechanism optimization in which the cross-sectional areas of the three movable links are chosen as the design variables. The optimization problem is to minimize the total mass of the system subject to constraints on the maximum values of the bending stresses in the links over a fixed period of time. The bending stresses are evaluated using the values of the bending moment \( \sigma(x, t) = (4/\pi^2 f(x)^{1/3})M_r \), where \( M_r \) is the bending moment evaluated in the 7th link of the link section. The following explicit mechanistic approximations of the constraint functions have been suggested:

\[
\tilde{F}_1(x) = a_1 (4x_2 + x_3)x_1^{1/2}, \quad \tilde{F}_2(x) = a_2x_2^{1/2}, \quad \tilde{F}_3(x) = a_3 x_3^{1/2}.
\]
The dynamic analysis of a flexible mechanism (high-fidelity model) is a time-consuming procedure as it requires integration of a system of nonlinear differential equations of motion. A model with rigid links for which the dynamic analysis is two orders of magnitude faster was then used as a lower-fidelity model. Because the stress distribution along the link depends on its inertia properties, the mass densities of the links have been chosen as the tuning parameters to construct the approximations of the third type (5); namely, \( \alpha_i = \rho_i \), \( i = 1, 2, 3 \). The results of optimization indicated that the quality of mechanistic approximations was the best, followed by (from best to worst) type 2 in multiplicative form (4) with multiplicative correction function, type 3 (5), type 2 in linear form (3) with linear correction function, and type 1 in multiplicative form (2). The quality of the metamodel of type 1 in linear form (1) was unacceptably low. A comprehensive comparison of the three types of metamodels (2–5) on an aerelastic optimization problem was presented by Gano et al. [89] and demonstrated the approach by a wing aerodynamic design problem. Balabanov and Venter [145] used gradient-based optimization in which the one-dimensional search points are evaluated using high-fidelity analysis, and the gradients are evaluated using low-fidelity analysis. The result is a multimodeling optimization scheme that does not require correlation between the results of the high- and low-fidelity analyses.

An alternative approach termed space mapping also uses high- and low-fidelity models. Gano et al. [90] demonstrated a high-accuracy response surface-based optimization scheme that does not require correlation between the results of the high- and low-fidelity simulations. We refer the interested reader to [144,158–161] for details) and a trust region methodology was incorporated [151]. The main difference between the previously discussed approaches, in which the results of the low-fidelity models are in some way tuned to match those of the high-fidelity model, and the space mapping approach is that in the latter a design variable space distortion is applied to the design variables of the low-fidelity model to cause its optimum point to match that of the high-fidelity model [152]. A simple analogy, offered by Keane and Nair [114], is that the space mapping approach is similar to drawing the low-fidelity model on a rubberized sheet that can be then distorted to agree topologically with the high-fidelity results. This seems to be a natural approach for solving inverse problems in which the main objective is to find parameters of the model, whereas a design optimization problem primarily aims at achieving the best performance characteristics (responses) of the system. This technique has been used extensively in microwave circuit design (see the review by Bandler et al. [152]) with fewer applications in other engineering fields. Leary et al. [153] demonstrated the use of space mapping in structural optimization on a simple beam problem. Ignatovich and Diaz [154] used space mapping in crashworthiness applications using a specially developed truss structure as a low-fidelity model. Redhe and Nilsson [155] used a multipoint version of space mapping in which a high-fidelity response evaluation is done in each iteration to improve the mapping function and combined it with the response surface methodology. The technique is compared to the correction surface-based approach and applied to a vehicle crashworthiness structural optimization problem. Space mapping has also found applications in sheet metal forming [156,157]. Another recent trend is to use Gaussian process models and kriging to handle integration of variable fidelity models. In summary, the covariance matrix of the Gaussian process is augmented to accommodate the information coming from two different sources, namely, the high- and low-fidelity simulations. We direct the interested reader to [144,158–161].

Until now, relatively little attention has been paid to a case when the number of design variables in a low-fidelity model differs from that in a high-fidelity model, particularly when a different modeling concept is used; e.g., a 3-D model is considered instead of a 2-D model. In such cases, some form of mapping between the spaces of the design variables is required. Robinson et al. [162] developed two new mapping methods, corrected space mapping and proper orthogonal decomposition (POD) mapping, that are used in conjunction with trust region model management. It is reported that on a wing design problem the use of POD mapping achieved 53% savings in high-fidelity function calls over optimization directly in the high-fidelity space. A hybrid of POD mapping and space
IV. Using Multiple Surrogates and Ensembles of Metamodels

A. Why So Many Surrogates? And How to Generate Multiple Surrogates?

The diversity of surrogate modeling techniques is based primarily on a combination of three components [106]: first are the statistical model and its assumptions. For example, although response surface techniques assume that the data are noisy and the model obtained with the basis functions is exact, kriging usually assumes that the data are exact, but the function is a realization of a Gaussian process. Second, for response surfaces, the basis functions are usually polynomials, but other functions have been occasionally used. Kriging allows different trend functions that are usually monomials. In support vector regression, the basis functions are specified in terms of a kernel. Third, although most surrogates are based on minimizing the mean square error, there are alternative measures that could be used for the loss function. For example, minimizing the average absolute error (L1 norm) would lead to surrogates that are less sensitive to outliers. Figure 4 illustrates the differences when surrogates are fitted to five data points. Four of them lie on a straight line, and the fifth represents erroneous data. The surrogates are a kriging model (ordinary kriging with Gaussian correlation fitted by minimizing the mean square error) and two versions of linear polynomials (one using the L1 loss function and the other using the mean square error). It is seen that, although the erroneous point substantially affects the surrogates fitted with the mean square error, it has no effect on the one fit with the L1 norm. With this figure and example, by no means are we suggesting that people should shift to L1 norm. We are just illustrating that there are cases in which other loss functions are more robust than the conventional mean square error.

It is clear that one can generate multiple surrogates by simply using different statistical models. One simple scenario was exemplified in Fig. 4. However, some techniques also allow different instances to be created. This is possible to accomplish using response surfaces (e.g., different implementations of stepwise regression), but larger changes would require increasing the order of the polynomial (often limited by the availability of data). Surrogates like kriging or support vector regression are much more flexible. With kriging, one could easily change the correlation function (see [164] for different implementations of correlation functions) and end up with a totally different surrogate model. With support vector regression, one could change the kernel and loss functions (see [75]) for different implementations of kernel and loss functions. Figure 5 illustrates this idea. We can see that the resulting models are substantially different and although they originated from the same data set and used variations of kriging models (i.e., same surrogate technique but different correlation models).

We consider two scenarios in which multiple surrogates are interesting. The first is when the surrogate is used to replace the actual (expensive) simulations for prediction over the design space (e.g., in global sensitivity analysis). In this case, the overall accuracy of the surrogate model is the main objective. The other scenario is the use of surrogates in optimization. Here, the global accuracy is less important than the ability to lead to the global optimum.

B. Using Multiple Surrogates for Prediction

Historically, as discussed in Sec. II, most applications employ a single instance of a particular metamodeling technique based on past experience, personal preference, or simply availability of a software package. Recently, however, there has been interest in the simultaneous use of multiple surrogates rather than a single one [165–168]. This makes sense because no single surrogate works well for all problems, and the cost of constructing multiple surrogates is often small compared to the cost of simulations. In cases in which the considered application requires a single model (e.g., coupling of different computer codes), the use of multiple surrogates can reduce the risk associated with poorly fitted models.

Creating a set of surrogates can be as simple as creating surrogate models based on different techniques (polynomial response surface, kriging, neural network, support vector regression, just to mention a few) or a more elaborate task, in which one considers variations on the same technique (such as kriging models with different correlation functions). The number of surrogates to be created is obviously a
difficult question to answer, but it might be a number that reflects the computational budget and the software capabilities. Once the set is created, one has to apply a selection criterion that is suitable to all surrogate models. Although test points (not used in the fitting process) allow a much more accurate assessment of accuracy [169], in practice, they are not always affordable. Because of that, the practitioners tend to favor criteria that are based solely on the sampled data, such as cross validation, the Akaike information criterion (AIC) and the Bayesian information criterion [170,171]. Here, we will focus the discussion on the use of cross-validation errors and the estimation of the mean square error (i.e., the prediction sum of squares, or PRESS, which is the mean square of the cross-validation errors). Cross validation is attractive because it does not depend on the statistical assumptions of a particular surrogate technique, and it does not require extra expensive simulations (test points). Nevertheless, cross validation should be used with caution because the literature has reported problems such as bias in error estimation [172]. 

A cross-validation error is the error at a data point when the surrogate is fitted to a subset of the data points not including this point. When the surrogate is fitted to all the other \( p - 1 \) points (so-called leave-one-out strategy), the process has to be repeated \( p \) times to obtain the vector of cross-validation errors. Alternatively, the \( k \)-fold strategy can also be used for computation of the PRESS vector. According to the classical \( k \)-fold strategy [173], after dividing the available data (\( p \) points) into \( p/k \) clusters, each fold is constructed using a point randomly selected (without replacement) from each of the clusters. Of the \( k \) folds, a single fold is retained as the validation data for testing the model, and the remaining \( k - 1 \) folds are used as training data. The cross-validation process is then repeated \( k \) times with each of the \( k \) folds used exactly once as validation data. Note that 1) \( k \) fold turns out to be the leave one out when \( k = p \) and 2) the computational cost associated with PRESS depends on the strategy used (\( k \)-fold or leave-one-out) and on the cost of fitting the surrogate model itself.

With the set of surrogates and the respective vectors of PRESS errors, two approaches involving multiple surrogates are 1) selecting one based on PRESS [174–176] and 2) using cross-validation errors to create a weighted surrogate. The first approach is as simple as the straightforward use of the surrogate with the lowest PRESS. Exemplified by Roecker [174], the literature on response surface models shows the application of such approach for the selection of the coefficients of the polynomial. The same can be found in the neural network literature. For example, Utans and Moody [175] applied the same concept for selecting neural network architectures. Recent works such as [102] and [105] extend the idea to selection of models built using different metamodeling techniques. This way, instead of choosing among different possible polynomials (or neural networks), the set of surrogates would contain, for example, polynomials, neural networks, and kriging models. In the second approach, weights are computed via minimization of the integrated square error [103,177,178]. A weighted average surrogate intends to take advantage of a set of surrogates in the hope of canceling errors in prediction through proper weighting selection in the linear combination of the models. Weights defined according to the individual PRESS errors have also been explored [178].

The advantages of combination over selection have never been clarified as discussed by Yang [177]. The article was literally entitled “Regression with Multiple Candidate Models: Selecting or Mixing?” Yang [177] suggested that selection can be better when the errors in prediction are small, and combination works better when the errors are large. Viana et al. [105] recently pointed out that 1) the potential gains from using weighted surrogates diminish substantially in high dimensions and 2) the poor quality of the information given by the cross-validation errors in low dimensions makes the gain very difficult in practice. Nevertheless, it is clear that creating multiple surrogates is an easy and cheap way to avoid poorly fitted models.

C. Using Multiple Surrogates for Optimization

Optimization is one of those cases in which there is no restriction to the simultaneous use of multiple surrogates. For instance, Mack et al. [165] employed polynomial response surfaces and radial basis neural networks to perform global sensitivity analysis and shape optimization of bluff body devices to facilitate mixing while minimizing the total pressure loss. (We would like to point out that global sensitivity analysis is a very important topic. To limit the scope of the paper, we refer the reader to [107,179,180] for more information.) They showed that, due to small islands in the design space in which mixing is very effective compared to the rest of the design space, it is difficult to use a single surrogate model to capture such local but critical features. Kammer and Alvin [181] create “master” response surfaces from linear weighted combinations of individual metamodels to create super metamodels that are valid over the entire input parameter space. Glaz et al. [168] used polynomial response surfaces, kriging, radial basis neural networks, and weighted average surrogate for helicopter rotor blade vibration reduction. Their results indicated that multiple surrogates can be used to locate low-vibration designs, which would be overlooked if only a single approximation method was employed. Sanchez et al. [167] presented an approach toward the optimal use of multiple kernel-based approximations (support vector regression). They reported that, in their set of analytical functions as well as in the engineering example of a six-bladed wind turbine, the use of individual surrogates was difficult in practice. Nevertheless, it is clear that creating multiple surrogates is an easy and cheap way to avoid poorly fitted models.
The capabilities of the computer models and the designer method combines information from multiple sources (including qualitative and quantitative sequential sampling technique. The interested reader is referred to [186]. There is also significant literature on the so-called infill criteria. Recent developments in sequential sampling are exemplified by Pronzato and Müller [182] points to be sampled. Forrester and Keane [109] present a very good technical overview on the topic, and we would be more confident that the present best solution is converging to the desired one. If they are not clustered, then we confirmed that using multiple surrogates was a smart decision that improved the chances of success of the optimization task. Validation of the five candidate solutions could take the same 1 h if they can be run in parallel or, in the case of serial computation, 5 h. One should keep in mind that optimizing the surrogates offers no guarantees that the surrogate optimum is even close to the true optimum. In practice, surrogate-based optimization proceeds sequentially with the point predicted by the surrogate being evaluated with the high-fidelity model and then refitting the surrogates and performing optimization again until convergence is achieved. That explains why frameworks for sequential sampling have been the target of considerable research lately (e.g., efficient global optimization [64] and its derivatives).

The previous framework only uses the prediction capabilities of the surrogate models. This might be useful when the computational budget allows only very few optimization cycles (perhaps only one or two). Nevertheless, we would like to point out that sequential sampling strategies (such as the variants of the efficient global optimization [64] and enhanced sequential optimization algorithms [92]) have been heavily studied recently. In this case, optimization is certainly conducted through more than two cycles with the uncertainty in surrogate modeling being used to decide the next points to be sampled. Forrester and Keane [109] present a very good technical overview on the topic, and Pronzato and Müller [182] provide insights on sequential sampling for metamodel generation. Recent developments in sequential sampling are exemplified by Rai and Campbell [183], Turner et al. [184], and Gorissen et al. [185]. [There is also significant literature on the so-called infill criteria (functional optimized to decide the next point or points to be added in sequential sampling). The interested reader is referred to [186–188] for more detail.] Specifically, Rai and Campbell [183] introduced a qualitative and quantitative sequential sampling technique. The method combines information from multiple sources (including computer models and the designer’s qualitative intuitions) through a criterion called “confidence function.” The capabilities of the approach were demonstrated using various examples including the design of a bistable micro-electromechanical system. Turner et al. [184] proposed a heuristic scheme that samples multiple points at a time based on nonuniform rational B splines (NURBs). The candidate sites are generated by solving a multi-objective optimization problem. They used four objectives: 1) proximity to existing data, 2) confidence in control point locations (in NURBs, the greater the distance between a control point and its nearest data points, the less confidence there is in the location of the control point), 3) slope magnitude (the authors argue that rapid change could be due to the presence of multiple local optimal and this might be of interest to a designer), and 4) model (used to search for minima, maxima, or extrema in the model). The effectiveness of the algorithm was demonstrated for five trial problems of engineering interest. Gorissen et al. [185] brought multiple surrogates to adaptive sampling. The objective is to be able to select the best surrogate model by adding points iteratively. They tailored a genetic algorithm that combines automatic model type selection, automatic model parameter optimization, and sequential design exploration. They used a set of analytical functions and engineering examples to illustrate the methodology. Another two ways of using multiple surrogates in sequential sampling might be through 1) blind kriging [104,189], which can be seen as an ensemble, and 2) the use of surrogates to provide multiple points per cycle of the EGO algorithm, such as in [190].

### V. Metamodeling Capabilities in Commercial Software

Currently, there exists a number of commercial software packages that implement a variety of metamodeling techniques, but for many of these, commercial software systems metamodeling is not a final goal. Instead, in many cases, metamodeling is a companion to optimization and design exploration capabilities, when identifying the best system for given conditions in an automatic fashion remains the main task. Developments in commercial software come from two primary sources: 1) academic research and 2) industrial needs. Academic research offers great flexibility and ability to explore various directions of scientific development, whereas industrial needs are dictated by problems required to be solved in a short period of time. Because their financing depends mostly on industry, software companies aim to accommodate current industry needs first, while trying to anticipate any future requirements. This necessity to anticipate future needs makes them conduct their own research and also requires them to follow trends in academic research. Keeping the balance between efficient research for future needs and immediate day-to-day important industry requests is not an easy task, especially accounting for the need to sell the software. An essential part of this balance is efficient and robust implementation of the methods and algorithms.

| Table 4 Example of wall time for PRESS computation of several surrogates |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Cost                        | Second-order polynomial     | Four surrogates (second-order polynomial, kriging, neural network, and support vector regression) |
| Fitting                     | 0.1 s                       | Approximately 9 min         |
| PRESS                       | 3 s                         | Approximately 1 h 10 min    |
| Total                       | 3.1 s                       |                             |

*These numbers are chosen as illustration (they may change in the real application).*

| Table 5 Comparison of different exercises using actual and surrogate models |
|-------------------------------|-----------------|-----------------|-----------------|
| Application                   | Number of evaluations | High-fidelity model (four variables, one response) | Single surrogate model | Five surrogate models |
| Extra point evaluation        | 1                | 0.05 s          | 0.25 s          |
| Single-run optimization       | 500              | 20 days 20 h    | 25 s            | 2 min 5 s |
| Global optimization           | 10,000           | 416 days 16 h   | 8 min 30 s      | 42 min 30 s |

*Extra point evaluation refers to the computation (or estimation) of the responses for a new set of values of the input variables. By single-run optimization, we mean performing local search based either on the high fidelity or on the surrogates. For simplicity, in both optimization exercises we do not consider refitting the surrogates with the results of the optimization. The number of evaluations in both optimization exercises is somewhat arbitrarily chosen for illustration. We considered 100 iterations in the local search, with each requiring five evaluations for gradient computation through finite difference, and global optimization costing 20 times more than local search. To build the surrogates, we assume that 30 points were sampled (accounting for 30 h). This cost should be taken into consideration when using surrogates.*
The advances in computer technology and graphical user interface (GUI) development as well as conference presentations of competing software vendors led to a situation in which commercial software systems offer similar capabilities regarding integration with the third-party analysis/simulation codes as well as pre/postprocessing and even underlying algorithms. Even running on remote CPUs as well as parallel computation capabilities became a common feature of most software. As a result, on one hand, the GUI of various software products became resembling, even though similar algorithms are called different names due to the lack of established terminology. As a result, learning curves for users may be different depending on the software environments. Because of all these factors, and considering the size of the industries that are customers to these software companies, a common practice is the straight communication between the customers and software engineers during the development phase. A potential user is strongly encouraged to talk to a specific software company regarding industry-specific needs. Also encouraged is personal evaluation of several candidate software according to industry preferences: ease of use, typical computational cost for the specific type of optimization problems, visualization capabilities, metamodeling and optimization algorithms, etc., to make sure that the software offers what is needed. Interestingly enough, the result is that, although many methods in commercial software have similar names, the implementation is quite different due to different customer demands and needs as well as due to different practical experience of software developers.

The same factors that force the user to spend more time evaluating each software (GUI similarity, similarity in postprocessing and integration features, some fuzziness in the actual software capabilities, different names for similar algorithms, etc.) also make it harder to perform unbiased comparison of the software. The nonstandardized terminology may be partially blamed for that (for example, from a user’s perspective, is there any difference between a “metamodel” and an “approximation”?). The main reason for difficulty when making unbiased comparisons is that for most software companies, selling software became a necessity with all its pluses (the need to develop fast and robust algorithms along with nice companion capabilities) and minuses (the need to actually sell the product, rather than to just periodically present advances with detailed description of all new capabilities, updates, and applications). This created an uneven playing field as software evolved.

Table 6 lists some of the popular commercial software products and their capabilities related to metamodeling methods and general-purpose optimization. The software products are presented in alphabetical order. We present neither a complete list of all available software products nor the complete list of capabilities but rather a brief introduction to software available specifically for metamodeling and optimization tasks. We also offer some brief remarks regarding each of the software. We did not include popular software (e.g., MATLAB®, Excel, JMP®, Minitab®) that do not specialize in optimization into the table; however, we do provide some remarks regarding their capabilities.

<table>
<thead>
<tr>
<th>Software product</th>
<th>Metamodelling capabilities</th>
<th>Optimization capabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOSS/Quattro (LMS International)</td>
<td>Least-squares regression for polynomials and posynomials, radial-basis functions, neural networks, kriging</td>
<td>Gradient-based optimization, surrogate-based optimization, genetic algorithm, multi-objective optimization, probabilistic optimization</td>
</tr>
<tr>
<td>IS-OPT (Livermore Software Technology Corporation)</td>
<td>Least-squares regression for polynomials, Taylor series approximation, radial-basis functions, neural networks, kriging, variable-complexity modeling</td>
<td>Self-organizing optimization algorithms specifically targeted for multi-objective and probabilistic optimization</td>
</tr>
<tr>
<td>Model Center (Phoenix Integration)</td>
<td>Least-squares regression for polynomials, Taylor series approximation, radial-basis functions, neural networks, kriging, variable-complexity modeling</td>
<td>Surrogate-based optimization, gradient-based optimization, genetic algorithm, simulated annealing, probabilistic optimization, multi-objective optimization</td>
</tr>
</tbody>
</table>

The same factors that force the user to spend more time evaluating each software (GUI similarity, similarity in postprocessing and integration features, some fuzziness in the actual software capabilities, different names for similar algorithms, etc.) also make it harder to perform unbiased comparison of the software. The nonstandardized terminology may be partially blamed for that (for example, from a user’s perspective, is there any difference between a “metamodel” and an “approximation”?). The main reason for difficulty when making unbiased comparisons is that for most software companies, selling software became a necessity with all its pluses...
inexperienced users may be overwhelmed by a variety of algorithms and the options in each of them. Use of C++ as a core language provides plug-and-play capability of components and natural paths for extensibility. Having several thousands of download registrations from around the world, DAKOTA relies on community discussion forums to enable a distributed support model. This is different from the commercial model and requires some sophistication from the user base to be able to function without commercial-quality support.

Not widely known is the fact that Excel has an optimization tool suitable for solving nonlinear problems. Although lagging behind specialized optimization software in terms of the scale of problems that can be solved, Excel Solver provides nice and easy-to-use optimization capabilities. In addition to that, most of the commercial optimization/metamodeling software packages have specialized interfaces to Excel.

One of the main advantages of HyperStudy is its close ties to the other products by Altair Engineering, especially to HyperMesh, a popular pre-postprocessor for major CAD/CAE, FEA, and CFD software. The integration with HyperMesh enables direct parameterization of FEA/MBD/CFD solver input data and one-step extraction of plot and animation output, thus making the solver integration to HyperStudy efficient. Shape variables can be easily defined using the morphing technology in HyperMesh without the need for CAD data. Advanced data mining techniques in HyperStudy such as redundancy analysis and clustering with principal component analysis simplify the task of studying, sorting, and analyzing results.

IOSO offers unique state-of-the-art optimization algorithms that are based on a self-organizational strategy and efficiently combine traditional response surface methodology with gradient-based optimization and evolutionary algorithms in a single run. The offered algorithms are equally efficient for the problems of complex and simple topology that may include mixed types of variables.

iSight is one of the most widely used commercial optimization packages. It supports direct integration to a large number of third-party analysis/simulation tools and CAD programs. One of the unique tools iSight offers is using the physical dimensions of the parameters to create a smaller number of nondimensional parameters for easier and semi-automatic reduction of the design variables and identifying underlying trends in system designs. iSight is tightly coupled with the plug-and-play engineering workflow environment based on the FiPER architecture, which allows workflow and component sharing as well as web workflow execution.

Although JMP may lack direct optimization capabilities, JMP and SAS software are definitely worth evaluating, as this software is the leader and de facto standard in statistical analysis, design of experiments technique, and response surface modeling. Like JMP, Minitab lacks direct optimization capabilities. However, being one of the most popular statistical packages, Minitab certainly is one of the leaders in generating and performing statistical analysis on various metamodels.

LS-OPT is developed alongside LS-Dyna nonlinear FE code and provides optimization and metamodeling capabilities for operating directly within the bounds of tight coupling with this FE code as well as independently. Tight coupling with LS-Dyna provides unique capabilities of performing various types of optimization as well as tradeoff studies efficiently with nonlinear FE code. Specifically, instability/noise/outlier investigations and variable screening for some structural problems sets LS-OPT apart from the other tools.

Meanwhile, MATLAB’s main focus is not metamodeling or optimization; rather, it is a numerical computing environment and programming language. It is a flexible and widespread tool with almost all specialized metamodeling/optimization software having direct interfaces to it. In addition, MATLAB itself has an optimization/metamodeling toolbox with a variety of algorithms and options available. As MATLAB provides nice programming, pre- and postprocessing capabilities as well as links, methods, and tools from a wide variety of fields, it is in itself an attractive system for optimization and metamodeling.

modeFRONTIER is used in a wide range of applications across all industry sectors but prides itself in advanced engineering fields that use CAE packages. modeFRONTIER provides a range of intuitive yet impressive and powerful data visualization and data filtering tools and charts. Along with a range of statistical data analysis tools, it also offers multidimensional analysis and clustering methods such as self-organizing maps and hierarchical and partitive clustering.

The main focus of Phoenix Integration’s Model Center is integration of various software into a single design environment, when software models can be located across the network or locally. Model Center has specific tight interfaces to many analysis/simulation software from various fields. Optimization is viewed as just one part of this environment. Third-party optimizers may be plugged into this environment.

In addition to offering optimization procedures and specialized interfaces to many CAD/CAE packages and local “legacy” codes, OPTIMUS automates and monitors simulation processes as well as allows the users to automatically visualize and explore the design space. OPTIMUS automates simulation tasks across multiple engineering disciplines and offers more flexibility for process integration by allowing multiple nested workflows. One of the unique features of OPTIMUS is linking the user’s own optimization algorithm with OPTIMUS.

Finally, with development led by Vanderplaats Research & Development, Inc., VisuALDOC offers a simple and intuitive but robust and flexible environment for interfacing with any analysis/simulation software to efficiently solve any general-purpose optimization problem. The main distinction of VisuALDOC is one of the most efficient and robust gradient-based optimization algorithms available.

VI. Metamodel-Driven Design Space Exploration and Visualization

Metamodeling not only reduces the computational costs of optimization but also provides a means for rapid design space exploration and, often even more importantly, visualization. Because metamodels are fast to evaluate, virtually instantaneous, they enable performance analyses to be computed in real time when design (input) variables are changed within a graphical design environment. The importance of having fast response in graphical design interfaces has been corroborated by many studies. Nearly 30 years ago, Goodman and Spence [191] found that response delays as little as 1.5 s in the software can increase the time to complete an interactive design task by approximately 50%. Recent experimental studies have found similar results: the time to find a good design solution using a graphical design interface increased by 33% when response delays were 1.5 s [192]. These recent studies have also found that errors in user performance, the ability to locate the optimum within the graphical design interface, can increase by 150% [193] or nearly twice that amount [192] when response delays of 1.5 s are present, depending on the size and complexity of the problem. Needless to say, rapid analysis capability for effective design space exploration is paramount in today’s computer-mediated design environment.

Within the MDO community, research in this area has proceeded primarily in two fronts: 1) improving software and visualization tools that use metamodels for design space exploration and 2) assessment of visualization strategies that employ metamodels. Examples of the former include Graph Morphing [194,195], Cloud Visualization [196], B&F C’Viz [197], and the Advanced Systems Design Suite [198,199], which use metamodels of various types to allow users to steer and visualize a simulation through real-time interactions. Recent developments have sought to improve methods for visualizing the resulting n-dimensional Pareto frontiers [200,201]. Meanwhile, aerospace researchers at Georgia Tech have been extensively using the metamodeling and visualization capabilities in JMP to perform multidimensional trade studies and explore the design space [40,202]. Ford Motor Company and SGI also joined forces to investigate the use of surrogate modeling and high-performance computing to facilitate rapid visualization of design alternatives during the MDO process [203]. Finally, researchers at
Penn State and the Applied Research Laboratory are investigating the use of visual steering commands that allow designers to explore and navigate multidimensional trade spaces in real time using the rapid analysis capabilities of metamodels [204, 205].

As for the second line of research, assessment of the benefits of metamodel-based visualization is becoming more prevalent now that visual design environments are routinely used by many engineering design teams. For instance, Ligetti and Simpson [206] studied the use of first-order, stepwise, and second-order polynomial regression models for approximating the system responses in a detailed manufacturing simulation and found that using stepwise regression models significantly reduced task completion time and decreased error compared to the first-order and second-order polynomial regression models. These improvements in efficiency and effectiveness, respectively, resulted primarily from having a more parsimonious regression model (i.e., the same level of accuracy with the fewest possible terms) during the one-factor-at-a-time variations that were permitted within the graphical design interface. Meanwhile, in a wing design problem that used second-order response surface models for analysis, Simpson et al. [193] found that problem size significantly affected the users' average error, which doubled each time as it increased from two to four and then to six design variables. These findings, in combination with the aforementioned importance of response delay, have significant implications on the use and development of metamodel-driven visual design environments; the potential benefits are great, but we must be very mindful of the human-computer interaction to avoid the pitfalls that can likewise occur.

VII. Future Research Directions

Despite these advances, recent metamodeling books for engineers [207], and the wide availability of many software packages, several research thrusts continue to push DACE advancements. First, the curse of dimensionality still exists as problems have just gotten larger. There are many factors that make high-dimensional problems inherently difficult, including design space sampling strategies (both all-at-once and sequential) and limited visualization capabilities. Shan and Wang [208, 209] note that new metamodeling techniques are needed to handle the peculiarities of high-dimensional problems. Here, we believe that global sensitivity analysis will play an important role in alleviating the curse of dimensionality, and many are investigating approaches to reduce the computational expense associated with high-dimensional problems [199, 210, 211]. Statistically sound, robust, and scalable (i.e., capable of handling large number of input and outputs) methods will help us 1) understand what are the subset of input variables responsible for most of the output variation, 2) decide whether or not we can safely fix one or more of the input variables without significantly affecting the output variability (screening), and 3) identify the main regions of interest in the input space if additional samples become available, among other things. Second, computational complexity still exists. Problems have just gotten more complex and/or we are trying to do more (e.g., optimization under uncertainty, reliability-based design optimization, and robust design) [110]. Third, there are still issues with numerical noise, which appear to be getting worse due to added computational complexity of many analyses [8] and also poses additional challenge when performing model validation (checking whether the model reproduces well known behavior of the process of interest). Fourth, the challenges of handling mixed discrete/continuous variables still exist, and it may have gotten worse due to the nature of problems now being investigated [110]. Fifth, demands on surrogate-based optimization continue to receive considerable attention in light of these advances. For instance, establishing training boundaries [212], setting targets [213, 214], and assessing the value of another iteration [215, 216], are but a few of the areas under investigation. Finally, validation of metamodels and the underlying model is as critical as before. Traditional measurements of accuracy based on the data set (such as the coefficient of determination, $R^2$, and its adjusted version, $R^2_{adj}$) might not be applicable to all metamodeling techniques and using separate data just for validation might be too expensive. Research has improved the assessment of global accuracy [105, 217] (most of the time, valid for any metamodeling technique). One open field though is the provision and improvement of the local uncertainty estimates for several metamodeling techniques. Seok et al. [218] present a Bayesian approach for estimating prediction intervals for support vector regression. Den Hertog et al. [88] show that the kriging variance is underestimated in expectation and presented a parametric bootstrapping technique to estimate the correct kriging prediction variance. Viana and Haftka [219] proposed to import the prediction variance from one surrogate to the other (e.g., support vector regression models together with kriging uncertainy estimates). A related wide open field for research is the incorporation of the error of the metamodel and the underlying model itself into the problem formulation [220].

VIII. Conclusions

In this paper, the motivations for advancements in the DACE within the MDO community in the past 25 years were examined. The paper began with a historical perspective to better understand the extent to which the use of metamodeling techniques in MDO have evolved since the seminal paper on DACE by Sacks et al. [1]. Based on current thrusts in the field, multifidelity approximations and issues of using ensembles of metamodels, as well as the availability of metamodels within commercial software and for design space exploration and visualization, were delved into deeper.

The goal in the paper was not to review the vast body of metamodeling work that exists in the literature but rather to better understand how developments in DACE are being used today while providing a collection of relevant references that complement existing literature reviews. It is a great to see that books about DACE, written by engineers for engineers, are finally starting to appear. These and related efforts (e.g., freely available surrogate toolboxes for the widely used MATLAB) will continue to broaden the use of DACE in MDO and non-MDO arenas while fostering its acceptance by an even larger community.

It is hoped that this paper will also help industry become more aware of the available capabilities and recent developments in metamodeling and metamodel-based optimization, thus driving the practical application of these techniques. It is the practical use and application of these techniques by engineers that is the ultimate indicator of their success or failure. Metamodeling and optimization have still a long way to go in this respect to become every day and common tool in industry. The main learning from this paper is that, although the big picture may mislead one to conclude that the computational budget is the main or only motivation behind the research on DACE, a closer look reveals that the continuous demand for newer and more practical capabilities is what really motivates these developments. It is hoped that this work can help both new and experienced researchers to provide context for their work and get inspired to continue developing new methods and advancing DACE research.

Acknowledgments

F. A. V. Viana is grateful to the National Science Foundation (NSF) for support of this work through NSF grant No. DMI-0423280. T. W. Simpson acknowledges support from National Science Foundation on NSF grant No. CMMI-0620948. Any opinions, findings, and conclusions or recommendations presented in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

A New Method for Structural Optimization


R. Haftka
Associate Editor