

MULTIFIDELITY INFORMATION FUSION ALGORITHMS FOR HIGH-DIMENSIONAL SYSTEMS AND MASSIVE DATA SETS*

PARIS PERDIKARIS[†], DANIELE VENTURI[‡], AND GEORGE EM KARNIADAKIS[§]

Abstract. We develop a framework for multifidelity information fusion and predictive inference in high-dimensional input spaces and in the presence of massive data sets. Hence, we tackle simultaneously the “big N” problem for big data and the curse of dimensionality in multivariate parametric problems. The proposed methodology establishes a new paradigm for constructing response surfaces of high-dimensional stochastic dynamical systems, simultaneously accounting for multifidelity in physical models as well as multifidelity in probability space. Scaling to high dimensions is achieved by data-driven dimensionality reduction techniques based on hierarchical functional decompositions and a graph-theoretic approach for encoding custom autocorrelation structure in Gaussian process priors. Multifidelity information fusion is facilitated through stochastic autoregressive schemes and frequency-domain machine learning algorithms that scale linearly with the data. Taking together these new developments leads to linear complexity algorithms as demonstrated in benchmark problems involving deterministic and stochastic fields in up to 10^5 input dimensions and 10^5 training points on a standard desktop computer.

Key words. multifidelity modeling, response surfaces, uncertainty quantification, high dimensions, big data, machine learning, Gaussian processes

AMS subject classifications. 62G08, 68T37, 60G15

DOI. 10.1137/15M1055164

1. Introduction. Decision making in data-rich yet budget-constrained environments necessitates the adoption of a probabilistic machine learning [1] mindset that combines versatile tools, ranging from experiments to multifidelity simulations to expert opinions, ultimately shaping new frontiers in data analytics, surrogate-based modeling, design optimization, and beyond.

Ever since the pioneering work of Sacks et al. [2], the use of surrogate models for the design and analysis of computer experiments has undergone great growth, establishing regression methods with Gaussian processes (GPs) [3] as a general and flexible tool for building inexpensive predictive schemes that are capable of emulating the response of complex systems. Furthermore, the use of GPs within autoregressive stochastic models, such as the widely used scheme put forth by Kennedy and O’Hagan [4] and the efficient recursive implementation of Le Gratiet and Garnier [5], allows for exploring spatial cross-correlations between heterogeneous information sources. In [6] the authors argue that this offers a general platform for developing multifidelity information fusion algorithms that simultaneously accounts for variable fidelity in models (e.g., high-fidelity direct numerical simulations versus low-fidelity empirical formulae)

*Submitted to the journal’s Computational Methods in Science and Engineering section January 4, 2016; accepted for publication (in revised form) April 21, 2016; published electronically July 7, 2016. This work was supported by DARPA grant HR0011-14-1-0060, AFOSR grant FA9550-12-1-0463, and the resources at the Argonne Leadership Computing Facility (ALCF) through the DOE INCITE program.

<http://www.siam.org/journals/sisc/38-4/M105516.html>

[†]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 (parisp@mit.edu).

[‡]Department of Applied Mathematics and Statistics, University of California Santa Cruz, Santa Cruz, CA 95064 (venturi@ucsc.edu).

[§]Division of Applied Mathematics, Brown University, Providence, RI 02912 (george.karniadakis@brown.edu).

as well as variable fidelity in probability space (e.g., high-fidelity tensor product multielement probabilistic collocation [7] versus low-fidelity sparse grid quadratures [8]). Although this construction is appealing to a wide range of applications, it is mainly limited to low-dimensional input spaces and moderately sized data sets.

A common strategy for constructing autocorrelation models for GPs in high dimensions is by taking the product of one-dimensional autocorrelation kernels. This typically results in an anisotropic covariance model, which assumes that all dimensions actively interact with each other. However, as the dimensionality is increased, one would hope to find sparsity in the input space, i.e., dimensions with negligible or very weak pairwise interactions [9, 10]. This observation has been widely studied in the literature and has motivated the use of additive models in [11]. Durrande et al. [12] have recently adopted this approach in the context of GPs, advocating versatility in constructing custom autocorrelation kernels that respect the structure in the observed data. This suggests that, having a way to quantify the active interactions in the data, one can tailor an autocorrelation model that closely adheres to those trends. To this end, Muehlenstaedt et al. [13] have employed functional analysis of variance (ANOVA) decompositions to compute the degree to which each input dimension, and their pairwise interactions, contribute to the total variability in the observations and used the corresponding sensitivity indices to construct an undirected graph that provides insight into the structure of possible additive autocorrelation kernels that best suit the available data. Although this approach is evidently advantageous for scaling GPs to high-dimensional problems, it may still suffer from computational tractability issues in the presence of big data sets, unless sparse approximations are employed [14, 15].

In general, the design of predictive inference schemes in high dimensions suffers from the well-known curse of dimensionality, as the number of points needed to explore the input space in its entirety increases exponentially with the dimension. This implicit need for big data introduces a severe deadlock for scalability in machine learning algorithms as they often involve the repeated inversion of covariance matrices that quantify the spatial cross-correlations in the observations. This defines the so-called big N problem—an expression used to characterize the demanding operational count associated with handling data sets comprising N observations ($N > 1000$). The implications of such large data sets on learning algorithms are well known, leading to an $\mathcal{O}(N^3)$ scaling for implementations based on maximum likelihood estimation (MLE). Addressing this challenge has received great attention over the last decades and several methods have been proposed to alleviate the computational cost [16, 17, 18]. Here, we will focus our attention on the frequency-domain learning approach recently put forth by De Baar, Dwight, and Bijl [19] that entirely avoids the inversion of covariance matrices at the learning stage and is applicable to a large class of wide-sense stationary autocorrelation models. This essentially enables the development of $\mathcal{O}(N)$ algorithms, hence opening one path to predictive inference on massive data sets.

Here, we overcome the $\mathcal{O}(N^3)$ scaling by employing frequency-domain learning [19] that entirely avoids costly matrix inversions. Scaling to high dimensions is accomplished by employing hierarchical functional decompositions that reveal structure in the data and inspire a graph-theoretic approach for constructing customized GP priors that exploit sparsity in the input space. To this end, we propose a new data-driven dimensionality reduction technique based on local projections—justified by the Fourier projection-slice theorem [20]—to decompose the high-dimensional supervised learning problem into a series of tractable, low-dimensional problems that can be solved in parallel using $\mathcal{O}(N)$ -fast algorithms in the frequency domain.

The paper is structured as follows. In sections 2.1 and 2.2 we provide a brief overview of GP regression and multifidelity modeling via recursive GPs. In section 2.3 we present the basic steps in performing learning via MLE and highlight the bottlenecks introduced by high dimensions and big data. In section 2.3.2 we provide an overview of the frequency-domain approach of De Baar, Dwight, and Bijl [19] that bypasses the shortcomings of MLE and enables fast learning from massive data-sets. Subsequently, in section 3 we elaborate on kernel design in high dimensions. In particular, we outline the a data-driven hierarchical functional decomposition based on random sampling high-dimensional model representation (RS-HDMR) expansions [9] and describe a graph-theoretic approach inspired by [13] for tailoring structured GP priors to the data. Moreover, we discuss how to decompose the global high-dimensional learning problem to a series of local solves using a projection-based dimensionality reduction technique in conjunction with the Fourier projection-slice theorem [20]. In section 3.1 we conclude with a summary of the proposed workflow, underlining key implementation aspects. The capabilities of the proposed methodology are demonstrated through three benchmark problems. First, in section 4.1 we employ a multifidelity modeling approach for constructing the mean field response of a stochastic flow through a borehole. Second, in section 4.2 we use the proposed methodology for estimating the probability density of the solution energy to a stochastic elliptic problem in 100 dimensions. Last, in section 4.3 we present an extreme case of performing GP regression in up to 100,000 input dimensions and 10^5 data points.

2. Multifidelity modeling via recursive GPs. The basic building block of the proposed multifidelity information fusion framework is GP regression. One way of viewing the use of GPs in regression problems is as defining a prior distribution over functions, which is then calibrated in view of data using an appropriate likelihood function, resulting in a posterior distribution with predictive capabilities. In what follows we provide an overview of the key steps in this construction, and we refer the reader to [3] for a detailed exposition to the subject.

2.1. GP regression. The main idea here is to model N scattered observations y of a quantity of interest $Y(\mathbf{x})$ as a realization of a Gaussian random field $Z(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$. The observations could be deterministic or stochastic in nature and may well be corrupted by modeling errors or measurement noise $\mathcal{E}(\mathbf{x})$, which is thereby assumed to be a zero-mean Gaussian random field, i.e., $\mathcal{E}(\mathbf{x}) \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$. Therefore, we have the following observation model:

$$(1) \quad Y(\mathbf{x}) = Z(\mathbf{x}) + \mathcal{E}(\mathbf{x}).$$

The prior distribution on $Z(\mathbf{x})$ is completely characterized by a mean $\mu(\mathbf{x}) = \mathbb{E}[Z(\mathbf{x})]$ and covariance $\kappa(\mathbf{x}, \mathbf{x}'; \theta)$ function, where θ is a vector of hyper-parameters. Typically, the choice of the prior reflects our belief about the structure, regularity, and other intrinsic properties of the quantity of interest $Y(\mathbf{x})$. However, our primary goal here is not just drawing random fields from the prior but to incorporate the knowledge contained in the observations y in order to reconstruct the field $Y(\mathbf{x})$. This can be achieved by computing the conditional distribution $\pi(\hat{y}|y, \theta)$, where $\hat{y}(\mathbf{x}^*)$ contains the predicted values for $Y(\mathbf{x})$ at a new set of locations \mathbf{x}^* . If a Gaussian prior is assumed on the hyper-parameters θ , then $\pi(\hat{y}|y, \theta)$ is obviously Gaussian and provides a predictive scheme for the estimated values \hat{y} . Once $Z(\mathbf{x})$ has been trained on the observed data (see section 2.3), its calibrated mean $\hat{\mu}$, variance $\hat{\sigma}^2$, and noise variance $\hat{\sigma}_\epsilon^2$ are known and can be used to evaluate the predictions \hat{y} , as well as to quantify the prediction variance v^2 as (see [21] for a derivation)

$$(2) \quad \hat{y}(\mathbf{x}^*) = \hat{\mu} + r^T (R + \hat{\sigma}_\epsilon^2 I)^{-1} (y - \mathbf{1}\hat{\mu}),$$

$$(3) \quad v^2(\mathbf{x}^*) = \hat{\sigma}_\epsilon^2 \left[1 - r^T (R + \hat{\sigma}_\epsilon^2 I)^{-1} r + \frac{[1 - r^T (R + \hat{\sigma}_\epsilon^2 I)^{-1} r]^2}{\mathbf{1}^T (R + \hat{\sigma}_\epsilon^2 I)^{-1} \mathbf{1}} \right],$$

where $R = \kappa(\mathbf{x}, \mathbf{x}'; \theta)$ is the $N \times N$ correlation matrix of $Z(\mathbf{x})$, $r = \kappa(\mathbf{x}, \mathbf{x}^*; \theta)$ is a $1 \times N$ vector containing the correlation between the prediction and the N training points, and $\mathbf{1}$ is a $1 \times N$ vector of ones. This is a linear regression scheme known as the best linear unbiased predictor in the statistics literature [22]. Note, that for $\sigma_\epsilon^2 = 0$ the predictor exactly interpolates the training data y , returning zero variance at these locations.

2.2. Multifidelity modeling via recursive GPs. Multifidelity stochastic modeling entails the use of variable fidelity methods *and* models both in physical *and* probability space [6]. Efficient information fusion from diverse sources is enabled through recursive GP schemes [5] combining s levels of fidelity and producing outputs $y_t(\mathbf{x}_t)$, at locations $\mathbf{x}_t \in D_t \subseteq \mathbb{R}^d$, sorted by increasing order of fidelity, and modeled by GPs $Z_t(\mathbf{x})$, $t = 1, \dots, s$. Then, the autoregressive scheme of Kennedy and O'Hagan [4] reads as

$$(4) \quad Z_t(\mathbf{x}) = \rho_{t-1}(\mathbf{x}) Z_{t-1}(\mathbf{x}) + \delta_t(\mathbf{x}), \quad t = 2, \dots, s,$$

where $R_t = \kappa_t(\mathbf{x}_t, \mathbf{x}'_t; \hat{\theta}_t)$ is the $N_t \times N_t$ correlation matrix of $Z_t(\mathbf{x})$ and $\delta_t(\mathbf{x})$ is a Gaussian field independent of $\{Z_{t-1}, \dots, Z_1\}$, distributed as $\delta_t \sim \mathcal{N}(\mu_{\delta_t}, \sigma_{\delta_t}^2 R_t(\theta_t))$. Also, $\{\mu_{\delta_t}, \sigma_{\delta_t}^2\}$ are mean and variance parameters, while $\rho(\mathbf{x})$ is a scaling factor that quantifies the correlation between $\{Z_t(\mathbf{x}), Z_{t-1}(\mathbf{x})\}$. The set of unknown model parameters $\{\mu_{\delta_t}, \sigma_{\delta_t}^2, \rho_{t-1}, \theta_t\}$ is typically learned from data using MLE.

The key idea put forth by Le Gratiet [5] is to replace the Gaussian field $Z_{t-1}(\mathbf{x})$ in (4) with a Gaussian field $\tilde{Z}_{t-1}(\mathbf{x})$ that is conditioned on all known observations $\{y_{t-1}, y_{t-2}, \dots, y_1\}$ up to level $(t-1)$, while assuming that the corresponding experimental design sets D_i , $i = 1, \dots, t-1$, have a nested structure, i.e., $D_1 \subseteq D_2 \subseteq \dots \subseteq D_{t-1}$. This essentially allows us to decouple the s -level autoregressive problem to s independent kriging problems that can be efficiently computed and are guaranteed to return a predictive mean and variance that is identical to the coupled Kennedy and O'Hagan scheme [4]. To underline the advantages of this approach, note that the scheme of Kennedy and O'Hagan requires inversion of covariance matrices of size $\sum_{t=1}^s N_t \times \sum_{t=1}^s N_t$, where N_t is the number of observed training points at level t . In contrast, the recursive approach involves the inversion of s covariance matrices of size $N_t \times N_t$, $t = 1, \dots, s$.

Once $Z_t(\mathbf{x})$ has been trained on the observed data $\{y_t, y_{t-1}, \dots, y_1\}$ (see section 2.3), the optimal set of hyper-parameters $\{\hat{\mu}_t, \hat{\sigma}_t^2, \hat{\sigma}_{\epsilon_t}^2, \hat{\rho}_{t-1}, \hat{\theta}_t\}$ is known and can be used to evaluate the predictions \hat{y}_t as well as to quantify the prediction variance v_t^2 at all points in \mathbf{x}_t^* (see [5] for a derivation),

$$(5) \quad \hat{y}_t(\mathbf{x}_t^*) = \hat{\mu}_t + \hat{\rho}_{t-1} \hat{y}_{t-1}(\mathbf{x}_t^*) + r_t^T (R_t + \hat{\sigma}_{\epsilon_t}^2 I)^{-1} [y_t(\mathbf{x}_t) - \mathbf{1}\hat{\mu}_t - \hat{\rho}_{t-1} \hat{y}_{t-1}(\mathbf{x}_t)],$$

$$(6) \quad v_t^2(\mathbf{x}_t^*) = \hat{\rho}_{t-1}^2 v_{t-1}^2(\mathbf{x}_t^*) + \hat{\sigma}_t^2 \left[1 - r_t^T (R_t + \hat{\sigma}_{\epsilon_t}^2 I)^{-1} r_t + \frac{[1 - r_t^T (R_t + \hat{\sigma}_{\epsilon_t}^2 I)^{-1} r_t]^2}{\mathbf{1}_t^T (R_t + \hat{\sigma}_{\epsilon_t}^2 I)^{-1} \mathbf{1}_t} \right],$$

where $R_t = \kappa_t(\mathbf{x}_t, \mathbf{x}'_t; \hat{\theta}_t)$ is the $N_t \times N_t$ correlation matrix of $Z_t(\mathbf{x})$, $r_t = \kappa_t(\mathbf{x}_t, \mathbf{x}_t^*; \hat{\theta}_t)$ is a $1 \times N_t$ vector containing the correlation between the prediction and the N_t training

points, and $\mathbf{1}_t$ is a $1 \times N_t$ vector of ones. Note that for $t = 1$ the above scheme reduces to the standard GP regression scheme of (2)–(3). Also, $\kappa_t(\mathbf{x}_t, \mathbf{x}'_t; \theta_t)$ is the auto-correlation kernel that quantifies spatial correlations at level t .

We recognize that such recursive autoregressive schemes can provide a rigorous and tractable workflow for multifidelity information fusion. This suggests a general framework that targets the seamless integration of surrogate-based prediction/optimization and uncertainty quantification, allowing one to *simultaneously* address multifidelity in physical models (e.g., direct numerical simulations versus experiments) as well as multifidelity in probability space (e.g., sparse grids [8] versus multi-element probabilistic collocation [7]). The reader is referred to [6] for a detailed presentation of this paradigm.

2.3. Parameter estimation.

2.3.1. Maximum likelihood estimation. Estimating the hyperparameters requires learning the optimal set of $\{\mu_t, \sigma_t^2, \sigma_{\epsilon_t}^2, \rho_{t-1}, \theta_t\}$ from all known observations $\{y_t, y_{t-1}, \dots, y_1\}$ at each inference level t . In what follows we will confine the presentation to MLE procedures for the sake of clarity. However, in the general Bayesian setting all hyper-parameters are assigned with prior distributions, and inference is performed via more costly marginalization techniques, typically using Markov chain Monte Carlo integration [3].

Parameter estimation via MLE at each inference level t is achieved by minimizing the negative log-likelihood of the observed data y_t ,

$$(7) \quad \min_{\{\mu_t, \sigma_t^2, \sigma_{\epsilon_t}^2, \rho_{t-1}, \theta_t\}} \frac{N_t}{2} \log(\sigma_t^2) + \frac{1}{2} \log |R_t(\theta_t) + \sigma_{\epsilon_t}^2 I| \\ + \frac{1}{2\sigma_t^2} [y_t(\mathbf{x}_t) - \mathbf{1}_t \mu_t - \rho_{t-1} \hat{y}_{t-1}(\mathbf{x}_t)]^T [R_t(\theta_t) + \sigma_{\epsilon_t}^2 I]^{-1} [y_t(\mathbf{x}_t) - \mathbf{1}_t \mu_t - \rho_{t-1} \hat{y}_{t-1}(\mathbf{x}_t)],$$

where we have highlighted the dependence of the correlation matrix R_t on the hyperparameters θ_t . Setting the derivatives of this expression to zero with respect to μ_t , ρ_{t-1} , and σ_t^2 , we can express the optimal values of $\hat{\mu}_t$, $\hat{\rho}_{t-1}$, and $\hat{\sigma}_t^2$ as functions of the correlation matrix $(R_t + \sigma_{\epsilon_t}^2 I)$,

$$(8) \quad (\hat{\mu}_t, \hat{\rho}_{t-1}) = [\mathbf{h}_t^T (R_t + \sigma_{\epsilon_t}^2 I)^{-1} \mathbf{h}_t]^{-1} \mathbf{h}_t^T (R_t + \sigma_{\epsilon_t}^2 I)^{-1} y_t(\mathbf{x}_t),$$

$$(9) \quad \hat{\sigma}_t^2 = \frac{1}{c} \{ [y_t(\mathbf{x}_t) - \mathbf{1}_t \hat{\mu}_t - \hat{\rho}_{t-1} \hat{y}_{t-1}(\mathbf{x}_t)]^T [R_t + \sigma_{\epsilon_t}^2 I]^{-1} \\ [y_t(\mathbf{x}_t) - \mathbf{1}_t \hat{\mu}_t - \hat{\rho}_{t-1} \hat{y}_{t-1}(\mathbf{x}_t)] \},$$

where $\mathbf{h}_t = [\mathbf{1}_t \quad \hat{y}_{t-1}(\mathbf{x}_t)]$, and $c = \begin{cases} N_t - 1, & t = 1 \\ N_t - 2, & t > 1 \end{cases}$. Finally, the optimal $\{\hat{\sigma}_{\epsilon_t}^2, \hat{\theta}_t\}$ can be estimated by minimizing the concentrated restricted log-likelihood

$$(10) \quad \min_{\{\sigma_{\epsilon_t}^2, \theta_t\}} \log |R_t(\theta_t) + \sigma_{\epsilon_t}^2 I| + c \log(\hat{\sigma}_t^2).$$

The computational cost of calibrating model hyper-parameters through MLE is dominated by the inversion of correlation matrices $(R_t + \sigma_{\epsilon_t}^2 I)^{-1}$ at each iteration of the minimization procedure in (10). The inversion is typically performed using the Cholesky decomposition that scales as $\mathcal{O}(N_t^3)$, leading to a severe bottleneck in the presence of moderately big data sets. This is typically the case for high-dimensional

problems where abundance of data is often required for performing meaningful inference. This pathology is further amplified in cases where the noise variance $\sigma_{\epsilon_t}^2$ is negligible and/or the observed data points are tightly clustered in space. Such cases introduce ill-conditioning that may well jeopardize the feasibility of the inversion as well as pollute the numerical solution with errors. Moreover, if an anisotropic correlation kernel $\kappa_t(\mathbf{x}_t, \mathbf{x}'_t; \theta_t)$ is assumed, then the vector of correlation lengths θ_t is d -dimensional, leading to an increasingly complex optimization problem (see (10)) as the dimensionality of the input variables \mathbf{x}_t increases. These shortcomings render the learning process intractable for large data sets and suggest seeking alternative routes to parameter estimation. Next, we describe a method that bypasses the deficiencies of MLE and enables the development of fast learning algorithms that scale linearly with the data.

2.3.2. Frequency-domain sample variogram fitting. Following the approach of De Baar, Dwight, and Bijl [19] we employ the Wiener–Khinchin theorem to fit the autocorrelation function of a wide-sense stationary random field to the power spectrum of the data. The latter contains sufficient information for extracting the second-order statistics that fully describe the Gaussian predictor $Z_t(\mathbf{x})$. Therefore, the model hyper-parameters at each inference level t can be learned in the frequency domain by fitting the Fourier transform of the sample variogram as

$$(11) \quad \min_{\{\sigma_{\epsilon_t}^2, \theta_t\}} \sum_{i=1}^{N_t^d} |\log \hat{w}_{t,i}^2 - \log [\hat{a}_{t,i}(\sigma_{\epsilon_t}^2, \theta_t)]|^2,$$

where $\hat{w}_{t,i}^2$ is the amplitude of each of the N_t^d Fourier coefficients in the modal representation of the data $y_t(\mathbf{x})$, $\hat{a}_{t,i}(\sigma_{\epsilon_t}^2, \theta_t)$ are the coefficients of the Fourier transform of the autocorrelation function $\{\kappa_t(\mathbf{x}_t, \mathbf{x}'_t; \theta_t) + \sigma_{\epsilon_t}^2 \delta(\|\mathbf{x}_t - \mathbf{x}'_t\|)\}$, with $\delta(\cdot)$ denoting the Dirac delta function, while $\|\cdot\|$ measures distance in an appropriate norm. The Fourier coefficients $\hat{w}(\xi)$ can be efficiently computed with $\mathcal{O}(N_t \log N_t)$ cost using the fast Fourier transform for regularly spaced samples or the nonuniform fast Fourier transform [23] for irregularly spaced samples. Moreover, for a wide class of autocorrelation functions, the Fourier transform of $\hat{a}(\xi; \sigma_{\epsilon_t}^2, \theta_t)$ is analytically available, whereby each evaluation of the objective function in the minimization of (11) can be carried out with a linear cost, i.e., $\mathcal{O}(N_t)$. This directly circumvents the limitations of hyper-parameter learning using MLE approaches, namely, the cubic scaling associated with inverting dense ill-conditioned correlation matrices, and therefore it enables parameter estimation from massive data sets.

Although the Wiener–Khinchin theorem relies on the assumption of stationarity, modeling of a nonstationary response can also be accommodated by learning a bijective warping of the inputs that removes major nonstationary effects [24]. This mapping essentially warps the inputs into a jointly stationary space, thus allowing the use of standard wide-sense stationary kernels that enable fast learning in the frequency domain. This enables the use of general families of expressive kernels, such as the spectral mixture kernels recently put forth by Wilson and Adams [25] that can represent any stationary covariance function.

A limitation of frequency-domain sample variogram (FSV) fitting is that the summation in (11) is implicitly assumed to take place over all dimensions, i.e., over all N_t^d frequencies in ξ . Although this is tractable for low-dimensional problems, it may easily lead to prohibitive requirements in terms of both memory storage and operation count as the dimensionality increases. In the next section we present an

effective methodology for scalable hyper-parameter learning from massive data sets in high dimensions.

3. Kernel design in high dimensions. While high-dimensional systems may not always be sparse, the geometric law of large numbers [10] states that there is a good probability that a sufficiently smooth multivariate function can be well approximated by a constant function on a sufficiently high-dimensional domain. Empirically, we know that many physical systems are governed by two- or three-body interaction potentials. In such cases, high-dimensional model representations, such as ANOVA and HDMR [9, 26, 27, 13] are proven to dramatically reduce the computational effort in representing input-output relationships. The general form of such representations takes the form

$$(12) \quad y(\mathbf{x}) = y_0 + \sum_{1 \leq i \leq d} y_i(x_i) + \sum_{1 \leq i < j \leq d} y_{ij}(x_i, x_j) + \cdots,$$

where y_0 is a constant, $y_i(x_i)$ are component functions quantifying the effect of the variable x_i acting independently of all other input variables, $y_{ij}(x_i, x_j)$ represents the cooperative effects of x_i and x_j , and higher-order terms reflect the cooperative effects of increasing numbers of variables acting together to impact upon the output of $y(\mathbf{x})$.

Given a set of randomly sampled scattered observations, the mutually orthogonal component functions are typically computed using Monte Carlo and its variants, or probabilistic collocation methods [28, 29]. From $y_i(x_i)$ and $y_{ij}(x_i, x_j)$ we can directly compute the Sobol sensitivity indices D_i and D_{ij} that quantify the active interactions in the data [30],

$$(13) \quad D_i = \int_0^1 y_i^2(x_i) dx_i \approx \int_0^1 \left[\sum_{r=1}^{k_i} \alpha_r^i \phi_r(x_i) \right]^2 dx_i = \sum_{r=1}^{k_i} (\alpha_r^i)^2,$$

$$(14) \quad D_{ij} = \int_0^1 \int_0^1 y_{ij}^2(x_i, x_j) dx_i dx_j \\ \approx \int_0^1 \int_0^1 \left[\sum_{p=1}^{l_i} \sum_{q=1}^{l'_j} \beta_{pq}^{ij} \phi_p(x_i) \phi_q(x_j) \right]^2 dx_i dx_j = \sum_{p=1}^{l_i} \sum_{q=1}^{l'_j} (\beta_{pq}^{ij})^2,$$

where α_r^i and β_{pq}^{ij} are unknown expansion coefficients determined from data and $\phi_r(x)$ are basis functions of order r . For a set of orthonormal basis functions, the unknown coefficients can be directly determined from the N_t data points at each inference level via Monte Carlo integration or probabilistic collocation methods [28, 29].

These sensitivity indices identify active interactions in high-dimensional data sets. This valuable information can guide the design of correlation kernels that are tailored to the given data set, respecting all significant input-output interactions. To this end, we employ a graph-theoretic approach in which custom correlation kernels can be constructed as an *additive* composition of kernels that describe cross-correlations within each one of the *maximal cliques* of the undirected graph defined by the Sobol sensitivity indices. The first step toward this construction involves assembling the undirected graph $G = (V, E)$ of the computed sensitivity indices, where first-order sensitivities D_i correspond to vertices V , while sensitivity indices of second-order interactions D_{ij} define edges E (see Figure 1). Once the undirected graph is available, a clique \mathcal{C} can be identified as a subset of the vertices, $\mathcal{C} \subseteq V$, such that every two distinct vertices are adjacent. This is equivalent to the condition that the subgraph of

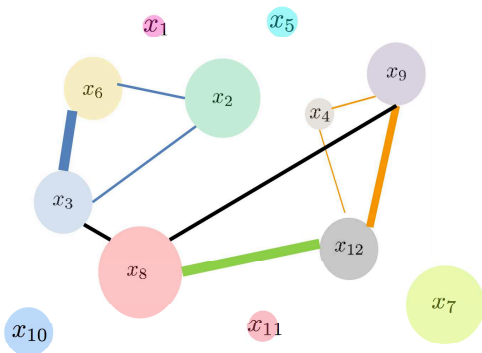


FIG. 1. Sketch of the undirected graph defined by the Sobol sensitivity indices of a 12-dimensional function $y(x_1, x_2, \dots, x_{12})$. The size of the disks corresponds to the magnitude of first-order sensitivity indices, while the thickness of the connecting lines quantifies the magnitude of second-order indices. Here we can identify three maximal cliques of dimensionality 3, $C_1 = \{x_2, x_3, x_6\}$, $C_2 = \{x_8, x_9, x_{12}\}$, and $C_3 = \{x_4, x_9, x_{12}\}$, and one clique of dimensionality 2, $C_4 = \{x_3, x_8\}$. By convention, all remaining five inactive dimensions are grouped together in $C_5 = \{x_1, x_5, x_7, x_{10}, x_{11}\}$.

G induced by \mathcal{C} is complete. A maximal clique is a clique that cannot be extended by including one more adjacent vertex, that is, a clique which does not exist exclusively within the vertex set of a larger clique (see Figures 1, 2(a), 4(a)). Maximal cliques can be efficiently identified from the graph of sensitivity indices using the Bron–Kerbosch algorithm with both pivoting and degeneracy reordering [31].

This procedure reveals the extent to which the observed data encodes an additive structure. The key idea here is to exploit this structure in order to effectively decompose the high-dimensional learning problem into a sequence of lower-dimensional tasks, where estimation of model hyper-parameters can take place independently within the support of each one of the maximal cliques. To this end, recall that fitting the FSV becomes intractable in high dimensions. However, we can utilize the hierarchical representation of (12) in order to exploit the structure encoded in the maximal cliques and efficiently perform FSV fitting locally for each maximal clique. This can be done by constructing an *additive* autocorrelation kernel that reflects the active interactions within each maximal clique $\kappa(\mathbf{x}, \mathbf{x}'; \theta) = \sum_{q=1}^{N_C} \kappa_q(\mathbf{x}_q, \mathbf{x}'_q; \theta_q)$, where N_C is the total number of maximal cliques at each fidelity level. Our goal now is to estimate the hyper-parameters θ_q by fitting the Fourier transform of each autocorrelation kernel $\kappa_q(\mathbf{x}_q, \mathbf{x}'_q; \theta_q)$ to the power spectrum of the data. In order to do so, we first have to identify the contribution of each maximal clique to the power spectrum of the d -dimensional data set. To this end, the hierarchical component functions $y_i(x_i)$ and $y_{ij}(x_i, x_j)$ can be utilized to project data onto the subspace defined by each maximal clique as

$$(15) \quad \mathcal{P}_q y(\mathbf{x}) = y_0 + \sum_{i \in \mathcal{C}_q} y_i(x_i) + \sum_{i,j \in \mathcal{C}_q} y_{ij}(x_i, x_j) + \dots, \quad 1 \leq q \leq N_C,$$

where \mathcal{C}_q is an index set listing all active dimensions contained in the q th maximal clique, and the operator \mathcal{P}_q projects the data onto the subspace defined by all input dimensions that appear in \mathcal{C}_q . Then, by assuming a wide-sense stationary covariance kernel $\kappa_q(\mathbf{x}_q, \mathbf{x}'_q; \theta_q)$ in each maximal clique, we can employ the FSV learning al-

gorithm to estimate θ_q by fitting the power spectrum of the clique-projected data, which typically lives in a subspace of dimension much lower than d . This approach is justified by the Fourier projection-slice theorem [20], which formalizes the equivalence between taking the Fourier transform of a projection versus taking a slice of the full high-dimensional spectrum. The main advantage here is that for high-dimensional cases that admit an additive hierarchical representation, the dimension of the q th subspace is $m = \#\mathcal{C}_q \ll d$, where $\#\mathcal{C}_q$ denotes the cardinality of the set \mathcal{C}_q . Hence, the optimal hyper-parameters $\hat{\theta}_i$ defining the autocorrelation kernel in each clique can be estimated very efficiently using the FSV fitting algorithm. Due to the linearity of the projection this is a distance-preserving transformation of the input space, hence allowing for the consistent estimation of length-scale hyper-parameters. Finally, summing up all clique contributions we construct the global autocorrelation kernel that according to the Fourier projection-slice theorem best captures the power spectrum of the original high-dimensional observations. This allows us to fit GP models to big data in high dimensions by using $\mathcal{O}(N)$ algorithms.

3.1. Implementation aspects. Here we provide an overview of the workflow and discuss some key implementation aspects.

Step 1. Starting from a set of available scattered observations $y_t(\mathbf{x})$ at the inference level $1 \leq t \leq s$, our first task is to compute the RS-HDMR representation. To this end, we adopt the approach of [32] that employs an orthonormal basis of shifted Legendre polynomials (up to order 15), using adaptive criteria for the optimal selection of the polynomial order that approximates each component function, and variance reduction techniques that enhance the accuracy of the RS-HDMR representation when only a limited number of samples is available. For all cases considered, an RS-HDMR expansion with up to second-order interaction terms was sufficient to capture more than 95% of the variance in the observations.

Step 2. Once the RS-HDMR representation is computed, we invoke the Bron-Kerbosch algorithm [31] to identify all maximal cliques in the undirected graph of sensitivity indices. This guides the construction of an additive autocorrelation kernel that comprises all cliquewise contributions. Throughout all recursive inference levels we have assumed an anisotropic product Gaussian autocorrelation function for all corresponding maximal cliques, $1 \leq q \leq N_{\mathcal{C}}$.

$$(16) \quad \kappa_q(\mathbf{x}_q, \mathbf{x}'_q; \theta_q) = \prod_{i=1}^m e^{-\frac{|\mathbf{x}_i - \mathbf{x}'_i|^2}{2\theta_i}},$$

where $m = \text{card}\{\mathcal{C}_q\}$ is the number of dimensions contained in the q th clique, and θ_i is a correlation length hyper-parameter along the i th dimension. In this case, the Fourier transform of the autocorrelation function in (11) is available analytically,

$$(17) \quad \hat{a}(\xi_q; \sigma_{\epsilon_q}^2, \theta_q) = \sigma_{\epsilon_q}^2 + (2\pi)^{\frac{m}{2}} \theta_q e^{-2 \sum_{i=1}^m (\pi \theta_i \xi_i)^2}.$$

Step 3. The next step involves learning the hyper-parameters $\{\sigma_{\epsilon_q}^2, \theta_q\}$ for each maximal clique at inference level t by fitting the power spectrum of the clique-projected data (see (11), (15)). To this end, the data is projected on a regular grid with 128 points along each clique dimension. This corresponds to using 128 Fourier modes for resolving the variability of the shifted Legendre basis functions in the frequency domain. Note that the learning task is directly amenable to parallelization as it can be performed independently for each maximal clique. Once the optimal

values of $\{\hat{\sigma}_{\epsilon_q}^2, \hat{\theta}_q\}$ are known, we can construct the global correlation matrix and factorize it using the Cholesky decomposition. Although this step still scales as $\mathcal{O}(N_t^3)$ it's required to be performed only once for each recursive level. In cases where N_t is extremely large one may employ a preconditioned conjugate gradient solver to approximate $[R_t(\theta_t) + \sigma_{\epsilon_t}^2 I]^{-1} y_t$ without the need for storing the global correlation matrix R_t . Finally, the optimal values of $\{\hat{\mu}_t, \hat{\rho}_{t-1}, \hat{\sigma}_t^2\}$ can be obtained from (8), (9), where $(R_t + \sigma_{\epsilon_t}^2 I)^{-1} y_t$ is either computed via back-substitution of the Cholesky factors or approximated via a gradient descent method.

Step 4. Finally, given a set of prediction points \mathbf{x}^* we can employ (5), (6) to evaluate the predictor $\hat{y}_t(\mathbf{x}_t^*)$ and variance $v_t^2(\mathbf{x}_t^*)$. This task is also trivially parallelizable as predictions at different points in \mathbf{x}_t^* can be performed independently of each other.

4. Results.

4.1. Borehole function. The first benchmark illustrates the salient features of the proposed framework, and it involves multifidelity in both physical models and probability space. In particular, we consider two levels of fidelity of functions that simulate stochastic water flow through a borehole and depend on eight input parameters and four random variables. We assume that high-fidelity observations are generated by [33]

$$(18) \quad f_h(\mathbf{x}) = \frac{2\pi T_u(H_u - H_l)}{\log(r/r_w) \left(1 + \frac{2LT_u}{\log(r/r_w)r_w^2 K_w} + \frac{T_u}{T_l}\right)},$$

where $\mathbf{x} = [r_w, r, T_u, H_u, T_l, H_l, L, K_w]$ is a set of parameters defining the model. We also assume that realizations of (18) are perturbed by a non-Gaussian noise term $\eta(\mathbf{z})$ expressed as a function of four normal random variables $\eta(z_1, z_2, z_3, z_4) = z_1 \sin^2[(2z_2 + z_3)\pi] - \cos^2(z_4\pi)$, $z_i \sim \mathcal{N}(0, 1), i = 1, \dots, 4$. This returns stochastic high-fidelity data of the form $y_h(\mathbf{x}; \mathbf{z}) = f_h(\mathbf{x})[1 + 0.2\eta(\mathbf{z})]$. Similarly, stochastic low-fidelity observations are generated by replacing f_h with a lower-fidelity model given by [33]

$$(19) \quad f_l(\mathbf{x}) = \frac{5T_u(H_u - H_l)}{\log(r/r_w) \left(1.5 + \frac{2LT_u}{\log(r/r_w)r_w^2 K_w} + \frac{T_u}{T_l}\right)}.$$

Next, we apply the proposed multifidelity information fusion framework to construct the response surface of the eight-dimensional mean field $S(\mathbf{x}) = \mathbb{E}[y_h(\mathbf{x}; \mathbf{z})]$, given observations $\{y_h(\mathbf{x}; \mathbf{z}), y_l(\mathbf{x}; \mathbf{z})\}$, by employing two methods of different fidelity in probability space. We choose the high-fidelity probabilistic method to be a Gauss–Hermite sparse grid level-5 quadrature rule (SG-L5) [8] using 4,994 sampling points, while the low-fidelity method is a coarser sparse grid level-2 (SG-L2) with just 57 quadrature points. Taking together the available multifidelity information sources yields two models in physical space (y_h, y_l) and two models in probability space (SG-L5, SG-L2). Blending of information is performed by employing a four-level recursive scheme traversing the available models and data in the order $S_{11} \rightarrow S_{12} \rightarrow S_{21} \rightarrow S_{22}$, where $S_{11}(\mathbf{x}) = \mathbb{E}_{SG-L2}[y_l(\mathbf{x}; \mathbf{z})]$, $S_{12}(\mathbf{x}) = \mathbb{E}_{SG-L5}[y_l(\mathbf{x}; \mathbf{z})]$, $S_{21}(\mathbf{x}) = \mathbb{E}_{SG-L2}[y_h(\mathbf{x}; \mathbf{z})]$, and $S_{22}(\mathbf{x}) = \mathbb{E}_{SG-L5}[y_h(\mathbf{x}; \mathbf{z})]$.

First, we compute the hierarchical representation of the data from (12) considering a set of randomly sampled training points $\{1024, 512, 128, 32\}$ that correspond to each one of the four observation models with increasing fidelity. Figure 2(a) shows the

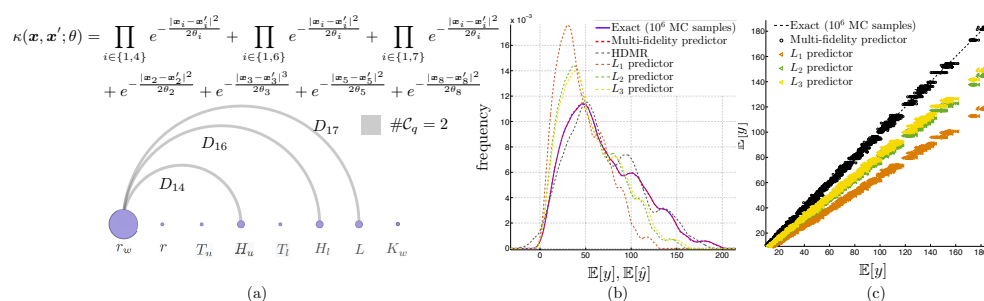


FIG. 2. Borehole function: demonstration of pairwise interactions and multifidelity predictions. (a) Sketch of the undirected graph of the Sobol sensitivity indices generated using 1,024 observations of $S_{11}(\mathbf{x})$, and the resulting additive GP prior. The radius of the purple disks quantifies the sensitivity on each input dimension, while the thickness of the gray arcs indicates the strength of each pairwise interaction. Color coding reveals the dimensionality of the identified maximal cliques. (b) Density plot of the frequency distribution of the exact solution $\mathbb{E}[y]$ (blue solid line), versus the estimated $\mathbb{E}[\hat{y}]$ (dashed lines) resulting from co-kriging and the HDMR representation. The red dashed line corresponds to the final co-kriging predictor accounting for information fusion along the path $S_{11} \rightarrow S_{12} \rightarrow S_{21} \rightarrow S_{22}$. The orange dashed line corresponds to the output of kriging on the lowest-fidelity data (L_1 predictor), while the green and yellow dashed lines correspond to the predictions at each intermediate recursive level, namely, $S_{11} \rightarrow S_{12}$ (L_2 predictor), and $S_{11} \rightarrow S_{12} \rightarrow S_{21}$ (L_3 predictor). (c) Scatter plot of the exact solution $\mathbb{E}[y]$ (black dashed line), versus the co-kriging predictor $\mathbb{E}[\hat{y}]$ (circles) at 2,000 randomly sampled test locations. The black circles correspond to the final co-kriging predictor accounting for information fusion along the path $S_{11} \rightarrow S_{12} \rightarrow S_{21} \rightarrow S_{22}$, while the colored triangles show the predictions of the intermediate recursive levels. (CPU cost: 5 minutes; memory footprint: 3 megabytes.)

resulting undirected graph of Sobol sensitivity indices that characterizes the active interactions in the data. The graph reveals a structure of seven maximal cliques: $C_1 = \{1, 4\}$, $C_2 = \{1, 6\}$, $C_3 = \{1, 7\}$, $C_4 = \{2\}$, $C_5 = \{3\}$, $C_6 = \{5\}$, and $C_7 = \{8\}$. Next, we utilize the available observations to calibrate the autocorrelation hyperparameters by solving $N_C = 7$ independent FSV learning problems. Finally, we sum up all the cliquewise contributions to obtain the global autocorrelation kernel $\kappa(\mathbf{x}, \mathbf{x}'; \theta)$, which is used to construct the correlation matrix R_t at each recursive level, $t = 1, \dots, 4$. Finally, R_t is factorized once using the Cholesky decomposition leading to an optimal set of $\{\hat{\mu}_t, \hat{\rho}_{t-1}, \hat{\sigma}_t^2\}$, and thus enabling the computation of the GP predictive posterior at each level of the recursive algorithm.

Accuracy is tested against a test set of 2,000 observations corresponding to an “exact” solution constructed computing $\mathbb{E}[y]$ using 10^6 Monte Carlo samples of the highest-fidelity observation model $f_h(\mathbf{x}^*)$. Figure 2(b) shows a density plot of the frequency distribution of the exact solution $\mathbb{E}[y]$, versus the estimated $\mathbb{E}[\hat{y}]$ resulting from the predictors at each level, as well as the prediction of the hierarchical representation of (12) meta-model denoted by HDMR (equation 12). The output of the Gaussian predictors also has been plotted in the scatter plot of Figure 2(c). Evidently, the response surface of the mean field is captured remarkably well by just using 32 observations of the highest-fidelity model S_{22} , supplemented by a number of inaccurate but mutually correlated low-fidelity observations from (S_{11}, S_{12}, S_{21}) .

4.2. Stochastic Helmholtz equation in 100 dimensions. We consider the following elliptic problem subject to random forcing and homogeneous Dirichlet boundary conditions in two input dimensions:

$$(20) \quad \begin{cases} (\lambda^2 - \nabla^2)u(\mathbf{x}; \omega) = f(\mathbf{x}; \omega), & \mathbf{x} = (x, y), \quad \mathbf{x} \in \mathcal{D} = [0, 2\pi]^2, \\ u(\mathbf{x}; \omega)|_{\partial\mathcal{D}} = 0, \\ f(\mathbf{x}; \omega) = \frac{2}{d} \left\{ \sum_{i=1}^{d/4} [\omega_i \sin(ix) + \omega_{i+d/4} \cos(ix)] + \sum_{i=1}^{d/4} [\omega_{i+d/2} \sin(iy) + \omega_{i+3d/4} \cos(iy)] \right\}, \end{cases}$$

where $d = 100$ is the total number of random variables representing the forcing term, and $\lambda^2 = 1$ is the Helmholtz constant, the value of which has been chosen in order to sustain high-frequency components in the unknown solution field u . The additive forcing is represented by a collection of independent random variables $\omega = (\omega_1, \omega_2, \dots, \omega_{100})$, each of them drawn from the uniform distribution $\mathcal{U}(0, 1)$.

Our goal here is to utilize the proposed multifidelity framework to get an accurate estimate of the probability density of the kinetic energy

$$(21) \quad E_k(\omega) = \frac{1}{2} \int_0^{2\pi} u^2(x, t; \omega) dx.$$

To this end we consider blending the output of an ensemble of variable fidelity models in physical space, by employing different resolutions of a spectral/ hp element discretization of (20) [34]. Higher-fidelity models are obtained by either increasing the number of quadrilateral elements that discretize the two-dimensional physical domain \mathcal{D} (h -refinement) or by increasing the polynomial order of the numerical approximation within each spectral element (p -refinement). In this context, a sample solution to (20) is approximated in terms of a polynomial expansion of the form

$$(22) \quad u(\mathbf{x}) = \sum_{i=1}^{N_{dof}} w_i \Phi_i(\mathbf{x}) = \sum_{e=1}^{N_{el}} \sum_{m=1}^M w_m^e \phi_m^e(\mathbf{x}_e(\xi)),$$

where N_{dof} is the total number of degrees of freedom, $M = (P+1)^2$ is the total number of modes in each quadrilateral spectral element, ξ defines a mapping from the physical space to the *standard* element, and $\phi_p^e(\mathbf{x}_e(\xi))$ are local to each element polynomials of order P , which when assembled together under the mapping $\mathbf{x}_e(\xi)$ result in a C^0 continuous global expansion $\Phi_p(\mathbf{x})$ [34]. In Figure 3 we present representative samples of the random forcing field and the numerical solution to (20).

We consider three levels of fidelity corresponding to different discretization resolutions in physical space. In particular, the highest-fidelity observations are obtained by solving (20) on a grid of $n_e^{(3)} = 144$ uniformly spaced spectral elements using a polynomial expansion of order $P^{(3)} = 10$ in each element. This discretization is fine enough to resolve all high-frequency components in the forcing term and return accurate solution samples of $u(\mathbf{x}; \omega)$. At the intermediate fidelity level S_2 , we have chosen a discretization consisting of $n_e^{(2)} = 64$ spectral elements of polynomial order $P^{(2)} = 8$. Similarly, the low-fidelity data S_1 is generated by a discretization of $n_e^{(1)} = 16$, and $P^{(1)} = 4$. Neither the intermediate nor the low-fidelity levels can resolve the high-frequency forcing term in (20), and, consequently, they return solutions that are contaminated with aliasing errors. However, the computational effort required to obtain a solution sample with the low-fidelity discretization is one order of magnitude smaller compared to the intermediate-fidelity level, and two orders of magnitude smaller compared to the high-fidelity level.

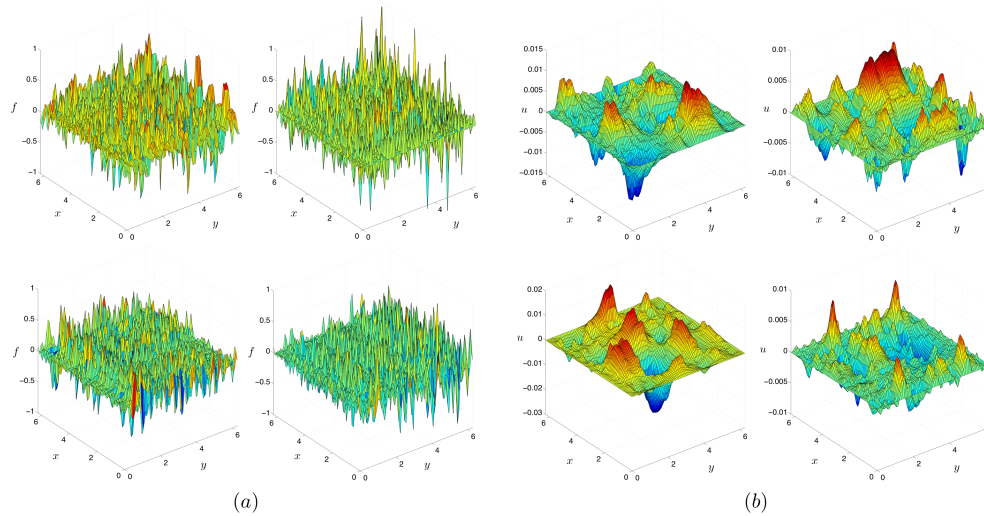


FIG. 3. *Stochastic Helmholtz equation: Representative samples of the random forcing term $f(\mathbf{x}; \omega)$ (left) and the numerical solution $u(\mathbf{x}; \omega)$ (right), obtained using a high-fidelity spectral element discretization of (20) with 144 elements and a 10th-order polynomial basis expansion in each element.*

We train a multifidelity predictor that can accurately emulate the kinetic energy of the solution to (20) for any given random sample ω . Nested training sets are constructed from 10^4 low-fidelity, 10^3 intermediate-fidelity, and 10^2 high-fidelity realizations of (20) by sampling the random forcing term in $[0, 1]^{100}$ using a space filling Latin hypercube strategy. With this training data set we compute the corresponding hierarchical expansion of (12) up to second order and identify the active dimension interactions that contribute in the variance decomposition of $y = E_k(\omega)$. The resulting undirected graph of first- and second-order Sobol sensitivity indices is depicted in Figure 4(a), indicating that all input variables are equally important and revealing very complex conditional dependency patterns between them. Interactions can be grouped in 135 maximal cliques, each containing 1 to 6 active dimensions. This information is then encoded to a structured GP prior by employing the additive autocorrelation kernel suggested by the computed Sobol indices (see Figure 4(a)).

Employing the steps outlined in section 3.1 we optimize the clique-wise kernel hyper-parameters for each level of the recursive information fusion algorithm to arrive at a predictive Gaussian posterior for the kinetic energy $\hat{y}(\omega)$. The mean of such Gaussian distribution over one-dimensional functions yields an estimate for the probability density function $\pi(E_k(\omega))$, while the variance quantifies our uncertainty with respect to that prediction. To assess the quality of the multifidelity predictor we compare the estimated probability density against a reference solution obtained by Monte Carlo averaging over 10^6 uniformly distributed solution samples that were obtained using the highest-fidelity discretization describe above. Figure 4(b) illustrates that comparison, along with the predicted densities resulting from considering only the low- and intermediate-fidelity training sets. Also, in Figure 4(c) we demonstrate the ability of the multifidelity model to generalize to unobserved inputs. Specifically, we test the predictions of $E_k(\omega)$ for unobserved inputs ω against the values obtained using the highest-fidelity discretization in a set of 2,000 randomly chosen test loca-

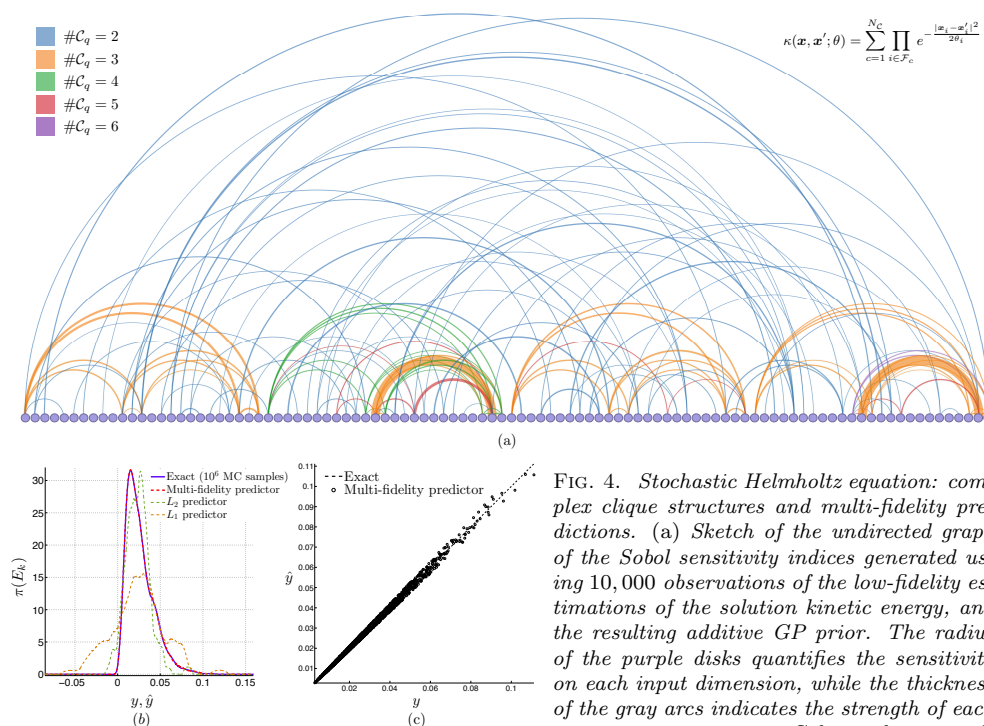


FIG. 4. *Stochastic Helmholtz equation: complex clique structures and multi-fidelity predictions.* (a) Sketch of the undirected graph of the Sobol sensitivity indices generated using 10,000 observations of the low-fidelity estimations of the solution kinetic energy, and the resulting additive GP prior. The radius of the purple disks quantifies the sensitivity on each input dimension, while the thickness of the gray arcs indicates the strength of each pair-wise interaction. Color coding reveals

the dimensionality of the identified maximal cliques. The additive auto-correlation kernel is constructed by summing up all clique-wise contributions where i is a multi-dimensional index accounting for the active dimensions in each of the 135 maximal cliques. (b) Probability density function of the solution kinetic energy $y = E_k(\omega)$ obtained by Monte Carlo averaging of 10^6 high-fidelity samples (blue solid line), versus the estimated $\hat{y} = E_k(\omega)$ (dashed lines) resulting from 3-level recursive co-kriging (red), 2-level recursive co-kriging trained on low- and intermediate-fidelity observations (L_2 predictor, green), and kriging trained on low-fidelity observations only (L_1 predictor, orange). (c) Scatter plot of the reference solution $y = E_k(\omega)$ (black dashed line), versus the 3-level co-kriging predictor $\hat{y} = E_k(\omega)$ (black circles) at 2,000 test locations, randomly sampled in $[0, 1]^{100}$. (CPU cost: 70 minutes, memory footprint: 800 megabytes)

tions in $[0, 1]^{100}$. It is evident that the multifidelity surrogate can correctly emulate the functional relationship that maps values of ω to E_k , at a fraction of the computational cost compared to a brute-force Monte Carlo simulation of the high-fidelity solver.

4.3. Sobol function in 10^5 dimensions. In this last example we consider an extreme demonstration involving the approximation of the Sobol function in $d = 10^5$ input dimensions. The Sobol function is a tensor product function that is routinely used as a benchmark problem in sensitivity analysis [33]. We consider the input space defined by the unit hypercube $[0, 1]^d$ and

$$(23) \quad y(\mathbf{x}) = \prod_{i=1}^d \frac{|4x_i - 2| + a_i}{1 + a_i},$$

where $a_i = i^2$, and for each index i , a lower value of a_i indicates a higher importance of the input variable x_i . Although this tensor product form assumes that all dimensions are actively interacting with each other, Zhang, Choi, and Karniadakis [26] have demonstrated that for this particular choice of a_i , the effective dimension-

ality of the Sobol function is much lower than d , and an additive representation with up to second-order interaction terms is capable of capturing more than 97% of the variance.

This example aims at demonstrating the capability of the proposed framework to simultaneously handle high dimensions and massive data sets, but also to highlight an important aspect of the machine learning procedure, namely, the effect of lack of data in such high-dimensional spaces. This is illustrated by considering two cases corresponding to training a GP surrogate on 10^4 (Case I) and 10^5 (Case II) training points generated by space filling Latin hypercubes in $[0, 1]^{100,000}$. The high dimensionality of the problem introduces a computational burden in constructing the hierarchical decomposition of (12). In particular, we have $\binom{100,000}{2} = 49,950,000$ second-order interaction terms that need to be computed and stored. To reduce the computational cost we use the adaptivity criterion proposed by Zhang, Choi, and Karniadakis [26] that uses information encoded in the first-order component functions to screen the selection process of active second-order interactions. This yields the additive autocorrelation kernel

$$(24) \quad \kappa(\mathbf{x}, \mathbf{x}'; \theta) = \sum_{c=1}^{\#\mathcal{F}} \prod_{j \in \mathcal{F}_c} e^{-\frac{|\mathbf{x}_j - \mathbf{x}'_j|^2}{2\theta_j}} + \sum_{i \in \mathcal{Q}} e^{-\frac{|\mathbf{x}_i - \mathbf{x}'_i|^2}{2\theta_i}},$$

where each member \mathcal{F}_c in the set \mathcal{F} is a tuple of two-dimensional indices corresponding to each one of the active second-order component functions, and \mathcal{Q} is a set of one-dimensional indices that contains all dimensions from 1 to d that do not appear in \mathcal{F} . Here, $\#\mathcal{F} = 147$ for the Case I training set and $\#\mathcal{F} = 769$ for the Case II training set, using an adaptivity threshold of 10^{-5} [26]. This construction helps to highlight the extent to which a purely additive kernel decomposition can capture the full tensor product response of $y(\mathbf{x})$.

Once the cliquewise kernel hyper-parameters are calculated from each training set, we arrive at a predictive GP posterior distribution that aims at emulating the input-output relation encoded in observed Sobol function data with quantified uncertainty. In Figure 5(a) we show the resulting frequency distribution obtained from probing the exact solution of (23) and the GP predictors for both cases in 2,000 randomly chosen test locations that lie within unit hyper-spheres centered at observed locations. Similarly, Figure 5(b) presents a visual assessment of the predictive accuracy of both GP predictors. In particular, we observe that the training data considered in Case II seems adequate for enabling the resulting GP posterior to perform accurate predictions for unobserved inputs, but the same cannot be claimed for Case I. There, the lack of training data hinders the predictive capability of the surrogate model as it affects both the accurate determination of active interactions in the hierarchical expansion of (12) as well as the identification of appropriate kernel hyper-parameters that resolve the correlation lengths present in the Sobol data. A similar deterioration in predictive accuracy is also observed for Case II if the radius of the hyper-sphere within which the test locations reside is increased. This is expected as the Gaussian autocorrelation kernel used here has good smoothing properties but limited extrapolation capacity. In such cases one should explore the use of more expressive kernels such as the family of spectral mixture kernels [25].

Due to the high dimensionality and the large number of observations, the computational cost is dominated by the computation of the component functions in (12) (69% of the computation), followed by the prediction step (25%). In contrast, for problems of lower dimensionality and moderately sized data sets (e.g., the borehole

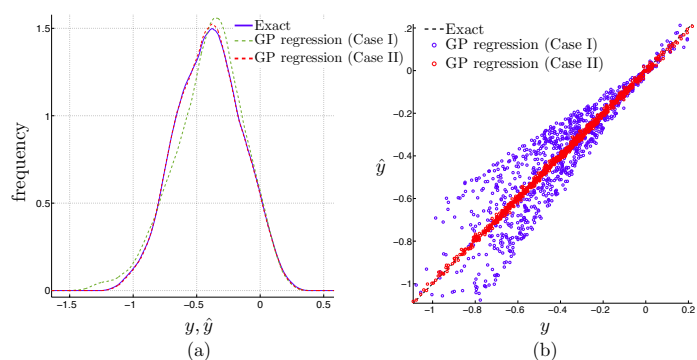


FIG. 5. *Sobol function in 10^5 dimensions: efficient scaling of GP regression to high dimensions and massive data sets.* (a) *Density plot of the frequency distribution of the exact solution $y(\mathbf{x}^*)$ (blue solid line) versus the estimated $\hat{y}(\mathbf{x}^*)$ (dashed lines) resulting from training a GP predictor on 10^4 observations (Case I, green) and 10^5 observations (Case II, red).* (b) *Scatter plot of the exact solution $y(\mathbf{x}^*)$ (black dashed line) versus the GP predictor $\hat{y}(\mathbf{x}^*)$ for Cases I and II. The comparison corresponds to 2,000 randomly chosen test locations living in unit hyper-spheres centered at observations. (CPU cost: 11 hours; memory footprint: 90 gigabytes, due to storing the large training set.)*

function case) the cost is typically attributed to learning the hyper-parameters within each maximal clique.

5. Conclusions. In data-driven stochastic simulations, as the dimensionality of the system increases there is an increasing need for assimilating more data so that we maintain a reasonable predictive accuracy. This creates a huge computational bottleneck since in addition to the exponentially increasing cost due to dimensionality, we also face the cost due to big data. The present work address this important issue for first time and proposes a computational framework with overall linear complexity. This leads to the possibility of sampling hundreds of thousands of dimensions and using hundreds of thousands of points on a standard desktop computer. The new framework can be used across different fields for probabilistic design, for parameter inference under uncertainty, and in data assimilation for weather prediction.

We have presented a tractable data-driven paradigm for computing response surfaces of high-dimensional deterministic and stochastic dynamical systems. Although the developed multifidelity framework generalizes well beyond the benchmark cases presented here, it does not constitute a panacea for all difficulties. High predictive accuracy can be expected only when the training data lie on a sufficiently smooth manifold; hence the study of regions where the response may present discontinuities (e.g., due to system bifurcations) may be problematic. To some degree this can be addressed by warping the input space [24], although a more elaborate treatment suggests the adoption of computationally demanding deep GP hierarchies [35]. Moreover, even in low-dimensional supervised learning problems, the use of Gaussian priors can be insufficient (e.g., when outliers are present in the data), mandating the use of more robust non-Gaussian prediction schemes that can be trained only with costly marginalization procedures (e.g., Markov chain Monte Carlo sampling). Finally, the merits of employing a multifidelity approach can be exploited only when the available model outputs exhibit some degree of correlation. In absence of such correlations any low-fidelity observations are essentially uninformative, and one can only rely on probing costly high-fidelity models. Despite these limitations, the proposed workflow

provides an efficient and flexible tool for tackling challenging problems in applied and computational science, such as uncertainty quantification, data assimilation, inverse problems, design optimization, and beyond.

REFERENCES

- [1] Z. GHAHRAMANI, *Probabilistic machine learning and artificial intelligence*, Nature, 521 (2015), pp. 452–459.
- [2] J. SACKS, W. J. WELCH, T. J. MITCHELL, AND H. P. WYNN, *Design and analysis of computer experiments*, Statist. Sci., 4 (1989), pp. 433–435.
- [3] C. E. RASMUSSEN, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006.
- [4] M. C. KENNEDY AND A. O'HAGAN, *Predicting the output from a complex computer code when fast approximations are available*, Biometrika, 87 (2000), pp. 1–13.
- [5] L. LE GRATIET AND J. GARNIER, *Recursive co-kriging model for design of computer experiments with multiple levels of fidelity*, Int. J. Uncertain. Quantif., 4 (2014).
- [6] P. PERDIKARIS, D. VENTURI, J. ROYSET, AND G. KARNIADAKIS, *Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields*, Proc. R. Soc. Land. Ser. A Math. Phys. Eng. Sci., 471 (2015).
- [7] J. FOO, X. WAN, AND G. E. KARNIADAKIS, *The multi-element probabilistic collocation method (ME-PCM): Error analysis and applications*, J. Comput. Phys., 227 (2008), pp. 9572–9595.
- [8] E. NOVAK AND K. RITTER, *High dimensional integration of smooth functions over cubes*, Numer. Math., 75 (1996), pp. 79–97.
- [9] H. RABITZ, Ö. F. ALIŞ, J. SHORTER, AND K. SHIM, *Efficient input-output model representations*, Comput. Phys. Commun., 117 (1999), pp. 11–20.
- [10] B. BAXTER AND A. ISELES, *On the Foundations of Computational Mathematics*, Report DAMTP-NA, University of Cambridge Department of Applied Mathematics and Theoretical Physics, 2002.
- [11] T. J. HASTIE AND R. J. TIBSHIRANI, *Generalized Additive Models*, Monogr. Statist. Appl. Probab. 43, CRC Press, Boca Raton, FL, 1990.
- [12] N. DURRANDE, D. GINSBOURGER, O. ROUSTANT, AND L. CARRARO, *Additive Covariance Kernels for High-dimensional Gaussian Process Modeling*, preprint, arXiv:1111.6233, 2011.
- [13] T. MUEHLENSTAEDT, O. ROUSTANT, L. CARRARO, AND S. KUHN, *Data-driven kriging models based on FANOVA-decomposition*, Statist. Comput., 22 (2012), pp. 723–738.
- [14] E. SNELSON AND Z. GHAHRAMANI, *Sparse Gaussian processes using pseudo-inputs*, Advances in Neural Information Processing Systems, 18, MIT Press, Cambridge, MA, 2005, pp. 1257–1264.
- [15] J. HENSMAN, N. FUSI, AND N. D. LAWRENCE, *Gaussian Processes for Big Data*, preprint, arXiv:1309.6835, 2013.
- [16] C. DIETRICH AND G. N. NEWSAM, *Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix*, SIAM J. Sci. Comput., 18 (1997), pp. 1088–1107.
- [17] F. LINDGREN, H. RUE, AND J. LINDSTRÖM, *An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach*, J. R. Stat. Soc. Ser. B, 73 (2011), pp. 423–498.
- [18] M. L. STEIN, J. CHEN, M. ANITESCU, ET AL., *Stochastic approximation of score functions for Gaussian processes*, Ann. Appl. Stat., 7 (2013), pp. 1162–1191.
- [19] J. DE BAAR, R. P. DWIGHT, AND H. BIJL, *Speeding up kriging through fast estimation of the hyperparameters in the frequency-domain*, Comput. Geosci., 54 (2013), pp. 99–106.
- [20] M. LEVOY, *Volume Rendering Using the Fourier Projection-Slice Theorem*, Computer Systems Laboratory, Stanford University, 1992.
- [21] D. R. JONES, *A taxonomy of global optimization methods based on response surfaces*, J. Global Optim., 21 (2001), pp. 345–383.
- [22] N. A. CRESSIE AND N. A. CASSIE, *Statistics for Spatial Data*, Wiley Ser. Probab. Stat. 900, Wiley, New York, 1993.
- [23] L. GREENGARD AND J.-Y. LEE, *Accelerating the nonuniform fast Fourier transform*, SIAM Rev., 46 (2004), pp. 443–454.
- [24] J. SNOEK, K. SWERSKY, R. S. ZEMEL, AND R. P. ADAMS, *Input Warping for Bayesian Optimization of Non-stationary Functions*, preprint, arXiv:1402.0929, 2014.

- [25] A. G. WILSON AND R. P. ADAMS, *Gaussian Process Kernels for Pattern Discovery and Extrapolation*, preprint, arXiv:1302.4245, 2013.
- [26] Z. ZHANG, M. CHOI, AND G. E. KARNIADAKIS, *Error estimates for the ANOVA method with polynomial chaos interpolation: Tensor product functions*, SIAM J. Sci. Comput., 34 (2012), pp. A1165–A1186.
- [27] N. DURRANDE, D. GINSBOURGER, O. ROUSTANT, AND L. CARRARO, *ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis*, J. Multivariate Anal., 115 (2013), pp. 57–67.
- [28] G. LI, S.-W. WANG, AND H. RABITZ, *Practical approaches to construct RS-HDMR component functions*, J. Phys. Chem. A, 106 (2002), pp. 8721–8733.
- [29] J. FOO AND G. E. KARNIADAKIS, *Multi-element probabilistic collocation method in high dimensions*, J. Comput. Phys., 229 (2010), pp. 1536–1557.
- [30] I. M. SOBOLOV, *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates*, Math. Comput. Simul., 55 (2001), pp. 271–280.
- [31] C. BRON AND J. KERBOSCH, *Algorithm 457: Finding all cliques of an undirected graph*, Commun. ACM, 16 (1973), pp. 575–577.
- [32] T. ZIEHN AND A. TOMLIN, *GUI-HDMR—a software tool for global sensitivity analysis of complex models*, Environ. Modell. Softw., 24 (2009), pp. 775–785.
- [33] S. SURJANOVIC AND D. BINGHAM, *Virtual Library of Simulation Experiments: Test Functions and Datasets*, <http://www.sfu.ca/~ssurjano> (2015).
- [34] G. KARNIADAKIS AND S. SHERWIN, *Spectral/hp Element Methods for Computational Fluid Dynamics*, Oxford University Press, 2013.
- [35] A. C. DAMIANOU AND N. D. LAWRENCE, *Deep Gaussian Processes*, preprint, arXiv:1211.0358, 2012.