MULTISCALE MODEL. SIMUL.
Vol. 10, No. 4, pp. 1428–1459

UNCERTAINTY QUANTIFICATION IN MD SIMULATIONS. PART I: FORWARD PROPAGATION*

F. RIZZI†, H. N. NAJM‡, B. J. DEBUSSCHERE‡, K. SARGSYAN‡, M. SALLOUM‡, H. ADALSTEINSSON‡, AND O. M. KNIO§

Abstract. This work focuses on quantifying the effect of intrinsic (thermal) noise and parametric uncertainty in molecular dynamics (MD) simulations. We consider isothermal, isobaric MD simulations of TIP4P (or four-site) water at ambient conditions, $T = 298$ K and $P = 1$ atm. Parametric uncertainty is assumed to originate from three force-field parameters that are parametrized in terms of standard uniform random variables. The thermal fluctuations inherent in MD simulations combine with parametric uncertainty to yield nondeterministic, noisy MD predictions of bulk water properties. Relying on polynomial chaos (PC) expansions, we develop a framework that enables us to isolate the impact of parametric uncertainty on the MD predictions and control the effect of the intrinsic noise. We construct the PC representations of quantities of interest (QoIs) using two different approaches: nonintrusive spectral projection (NISP) and Bayesian inference. We verify a priori the legitimacy of the NISP approach by verifying that the MD data satisfy regularity and smoothness conditions in the parameter space. The Bayesian inference approach relies on adaptively sampling the parameter space, based on analyzing the convergence of the PC expansions at different approximation levels. We show that for the present case, the effect of the thermal noise in the atomistic system can be controlled, and the MD predictions for the QoIs can be suitably represented using low-order PC models.

Key words. uncertainty quantification, Bayesian inference, polynomial chaos, molecular dynamics, TIP4P water, adaptive sampling

AMS subject classifications. 60G15, 62F15, 82D15, 62C10, 65C40, 74F05, 82C05, 82C22

DOI. 10.1137/110853169

1. Introduction.

Due to the increasing availability and improvement of computational resources, atomistic simulations are now widely employed in industrial and academic environments to study a large variety of chemical and material systems. Atomistic computations provide a suitable framework for exploring dynamical and thermodynamical properties of a system at the atomic level which, in general, are significantly difficult and expensive to investigate in experimental settings. The success of atomistic simulations, however, is partially limited by the accessible length and time scales due to the large computational cost, thus constraining the analysis to relatively small systems and short times.

We distinguish between two main types of applications. On the one hand, we have a "classical" field of application, where atomistic simulations are used in a self-

*Received by the editors October 27, 2011; accepted for publication (in revised form) August 21, 2012; published electronically December 12, 2012. This research was supported by the U.S. Department of Energy, Office of Advanced Scientific Computing Research, under award DE-SC0002506. Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
http://www.siam.org/journals/mms/10-4/85316.html

†Department of Mechanical Engineering, Johns Hopkins University, Baltimore, MD 21218 (frizzi1@jhu.edu).

‡Sandia National Laboratories, Livermore, CA 94550 (hnnajm@sandia.gov, bjdebus@sandia.gov, ksargsy@sandia.gov, mnsallo@sandia.gov, hadalst@sandia.gov).

§Corresponding author. Department of Mechanical Engineering, Johns Hopkins University, Baltimore, MD 21218. Current address: Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708 (omar.knio@duke.edu).

contained framework focused on analyzing the fundamental properties of a system of interest at the atomistic level. A second field of application that has received growing attention in the recent past concerns hybrid continuum-atomistic simulations, which are computational models suitable for studying systems involving multiple length and time scales. This approach stems from the fact that exploiting fully atomistic simulations for many problems is prohibitively expensive and, at the same time, resorting to all-continuum formulations may be inherently inapplicable (i.e., the underlying assumptions may be too restrictive). The main idea is to couple the particle and continuum formulations, exchanging the information between the two models across suitably defined boundaries. Examples of this type can be found in a variety of fields, from microfluidic devices and heat transfer in nanoflows to solid state systems; see, e.g., [31, 5, 13, 24, 45, 14, 41, 28, 48, 25, 1, 27, 7] and the references therein.

In many situations, uncertainty may have a large impact on observations or predictions. This is particularly relevant in multiscale multiphysics simulations, where the complexity associated with uncertainties increases due to the interaction and exchange of information between the particle and continuum models. We identify two main sources of uncertainty: intrinsic noise and parametric uncertainty. Intrinsic noise is inherently present in atomistic systems due to thermal fluctuations of the atoms. In this context, small systems tend to have large instantaneous fluctuations, and averaging suppresses this noise as $\sim 1/\sqrt{N}$, where N is the number of atoms. It follows that in the limit as $N \to \infty$, the intrinsic noise is increasingly better controlled just by averaging. A major limitation stems from the large computational cost associated with atomistic simulations, making this approach prohibitive for most applications. Salloum et al. [42] presented a methodology to account for the effects of intrinsic noise in the velocity information exchanged between the atomistic and continuum simulation components of a Couette flow model. A different source of uncertainty is parametric uncertainty, which arises when some parameters defining the system are not precisely known, and quantifying the impact of these uncertainties can be particularly challenging in problems involving intrinsic noise.

This work is aimed at uncertainty quantification in multiscale multiphysics simulations. In this paper, we focus on propagating parametric uncertainty in molecular dynamics (MD) simulations. As a test case, we consider isothermal, isobaric MD simulations of TIP4P (or four-site) water at ambient conditions, $T = 298$ K and $P = 1$ atm. Parametric uncertainty is propagated through the system, assuming a subset of the force-field parameters to be uncertain. The intrinsic (thermal) noise present in the atomistic system couples with the parametric uncertainty to yield nondeterministic, noisy MD predictions for the water observables. We develop a framework that enables us to isolate the impact of parametric uncertainty on MD predictions and, thus, quantify the effect of the intrinsic noise.

To this end, we rely on polynomial chaos (PC) expansions [47, 11] of random quantities. Specifically, we adopt a so-called nonintrusive formalism [23, 36], which reduces the problem to one of suitably determining PC coefficients of the stochastic model response on the basis of discrete realizations. Two approaches are pursued within this context. The first is based on the so-called nonintrusive spectral projection (NISP), which generally involves the application of quadrature rules to estimate the coefficients [22]. The second is based on the application of Bayesian inference [10, 43], for which we propose a new strategy based on adaptive sampling of realizations from the nodes of nested Fejér grids [9].

This article is organized as follows. In section 2, we describe the atomistic system and the MD simulations. In section 3, we discuss the stochastic reformulation of

the forward problem and outline the NISP and Bayesian inference approaches. Results and discussion are presented in section 4. Concluding remarks are provided in section 5.

2. Atomistic system and MD computations. In this study, we focus on isothermal, isobaric MD simulations of water at ambient conditions, i.e., $T = 298$ K and $P = 1$ atm, and model the water molecule according to the so-called TIP4P representation [19]. It is also referred to as a four-site model due to the fact that the negative charge of the oxygen is placed on a massless site at a distance d from the oxygen along the bisector of the HOH angle; see Figure 2.1. The TIP4P model belongs to the class of empirical force-field representations, where pairwise-additive functions are fitted to reproduce bulk-phase experimental data in classical Monte Carlo or MD simulations. Other examples of this class are the RWK [37], SPC/E [4], TIP3P [19], TIP4P-Ew [17], and TIP5P [26] representations. As discussed in [19], TIP4P predictions of some bulk water properties are in very good agreement with experimental measurements.

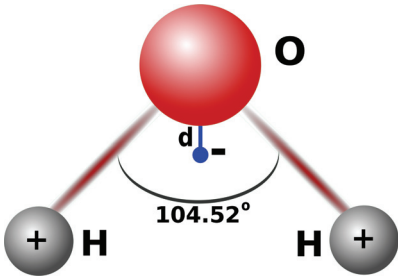


FIG. 2.1. *TIP4P model of the water molecule: H and O label the hydrogen and oxygen atoms, respectively, the + and − signs represent the charges, and d is the distance from the oxygen to the massless point where the negative charge is placed.*

We perform the MD simulations with LAMMPS [35, 21]. The computational domain is a cubic box of side length equal to 37.2 Å with periodic boundary conditions along each axis, containing 1728 water molecules. The potential is defined as the sum of two contributions, namely dispersion (Van der Waals) and electric forces. The Van der Waals interactions are modeled using a Lennard-Jones (LJ) potential, V_{LJ}, while the electrostatic interactions are modeled using Coulomb's law. Note that the TIP4P representation requires the computation of LJ interactions only for oxygen-oxygen interactions. In other words, the interaction between two molecules is modeled through the interaction of their oxygens atoms. On the other hand, Coulombic effects involve both atom types, i.e., hydrogen-hydrogen, hydrogen-oxygen, and oxygen-oxygen pairs. Specifically, we recall that the LJ potential, V_{LJ}, modeling intermolecular interactions is given by

$$V_{LJ}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right], \tag{2.1}$$

where V_{LJ} is the induced potential, ε represents the depth of the potential well, σ is the distance at which the potential becomes zero, and r is the distance between the two oxygen atoms of the interacting molecules. The TIP4P representation also requires three geometric parameters defining the water molecule, namely the angle between the bonds, the oxygen-hydrogen separation, and the distance, d, between

the negative charge and the oxygen atom. It follows that the complete force field is defined by seven parameters, namely three to define the geometric structure of the water molecule, two for the charges of the hydrogen (H) and oxygen (O) atoms, and, finally, the two LJ parameters, ε and σ, defining the dispersion forces between molecules.

Our goal is to characterize how uncertainty in the force-field representation affects the MD predictions for target observables of water. To this end, as discussed in more detail in the following section, we consider the following three parametric uncertainties: the two LJ parameters ε and σ, and the distance, d, from the oxygen to the massless point where the negative charge is placed in the TIP4P model. All the remaining parameters are set to their values commonly used for computational applications of TIP4P water; see, e.g., [19, 18].

Each MD computation is divided into three main parts. First, the system is run for 1500 steps with time step $\Delta t = 0.01$ fs to lose memory of the initial condition. Second, an equilibration stage is run for 15000 steps with $\Delta t = 2$ fs under isobaric-isothermal ensemble controls, $P = 1$ atm and $T = 298$ K. The pressure and temperature of the system are maintained in equilibrium with a heat and mechanical reservoir using a Nose–Hoover/Parrinello–Rahman formalism (see [32, 33, 29, 30, 16]). Dynamic feedback between these reservoirs and the system allows the volume and the kinetic energy to fluctuate about the desired mean values as expected in an equilibrium system. This is meant as a "transient" stage allowing the system to converge towards steady state. Since the present work focuses on uncertain force-field parameters, one concern is that the duration of the transient state might change according to the values of the driving parameters. For consistency, a transient of 15000 steps was chosen to be a suitable choice for all values of the force-field parameters considered in the present study. The system is then simulated for additional 80000 steps with $\Delta t = 2$ fs, under isobaric-isothermal conditions, keeping $P = 1$ atm and $T = 298$ K. This last stage allows us to explore the dynamics under steady-state conditions. The duration of the simulation was chosen as a compromise between a suitable length to explore a meaningful system dynamics and to be computationally feasible for the purposes of the present work. During the simulation, bond lengths and bond angles within each water molecule are held fixed using a "shake" algorithm since the TIP4P model requires a rigid molecule. Note, also, that the system involves long-range Coulombic interactions, which are computed with a P^3M (particle-particle-particle mesh) solver, which maps atom charges to a three-dimensional (3D) mesh, uses a 3D FFT to solve Poisson's equation on the mesh, and, finally, interpolates the electric field on the mesh points back to the atoms [15].

The data for the postprocessing are extracted during the steady state. We focus on the following three observables: density, ρ, self-diffusivity, D, and enthalpy, H. To compute the self-diffusivity, D, we rely on the analysis of the mean-square displacement (MSD) using the Einstein–Smoluchowsky equation (see [8, 46, 34, 6])

$$\left\langle |\mathbf{r}(t) - \mathbf{r}(t_0)|^2 \right\rangle = 6D(t - t_0), \tag{2.2}$$

where (t_0, t) identifies the time interval of interest, the angle brackets $\langle\cdot\rangle$ represent the ensemble average over all atoms, and $\mathbf{r}(t)$ is the displacement vector identifying a particle in the 3D domain at time t, namely $|\mathbf{r}(t) - \mathbf{r}(t_0)|^2 = \sum_{i=1}^{3} |r_i(t) - r_i(t_0)|^2$. It is important to remark that since (2.2) is reliable in the limit as $t \to \infty$, the diffusivity should be extracted based on the asymptotic behavior of the MSD. This further justifies the choice of relying on the MD data collected during the last part of

the steady state of the simulation.

The MD predictions for the observables of interest are filtered during each simulation using a running average. To explore the effect of different time-averaging approaches, we vary the length of the averaging window. Let ω be the number of time steps after which the average is performed. We consider four cases, namely $\omega = 50$, 100, 200, and 400. For each observable, the corresponding average value at the time step t_n is computed using the values at time steps $t_{(n-\omega)}$, $t_{(n-\omega+1)}$, $t_{(n-\omega+2)}$, $\ldots$, $t_{(n-1)}$; i.e., each average is taken over ω values in the preceding part of the simulation. Our goal is to determine the long-term average, doing a reduction by filtering the high-frequency noise. We thus need to select a time window that is long enough to suppress the high-frequency noise without carrying along much history because it would make it difficult to distinguish between transient and steady state. For the purposes of the present work, we choose the averaging based on $\omega = 100$, which is a suitable compromise between high-frequency filtering and long-term averaging. For brevity, we omit from the present manuscript the analysis of time averaging.

The filtering procedure is followed by an additional step needed to extract, for a single simulation, a unique representative value of each observable of interest to be used in the postprocessing. To this end, we rely on the mean of the time-averaged series obtained for $\omega = 100$. This procedure yields, for a given simulation, a single value for each quantity of interest (QoI) that will be used below for the core analysis of this study.

3. Stochastic reformulation of the forward problem. While intrinsic noise is inherently present in any MD system, parametric uncertainty originates from external sources, e.g., when some parameters defining the system are not known exactly. As previously anticipated, in the present study we consider three force-field parameters to be nondeterministic: the two parameters ε and σ defining the LJ potential in (2.1), and the distance, d, from the oxygen to the massless negative charge (Figure 2.1). We assume that these uncertain inputs can be parametrized as

$$\begin{aligned} \varepsilon(\xi_1) &= 0.147 + 0.043\ \xi_1, && \text{kcal/mol}, \\ \sigma(\xi_2) &= 3.15061 + 0.021\ \xi_2, && \text{Å}, \\ d(\xi_3) &= 0.14 + 0.035\ \xi_3, && \text{Å}, \end{aligned} \tag{3.1}$$

where $\{\xi_i\}_{i=1}^3$ are independent and identically distributed (i.i.d.) standard random variables (RVs) uniformly distributed in the interval $(-1, 1)$, i.e., $\{\xi_i\}_{i=1}^3 \sim \mathcal{U}(-1, 1)$. The above parametrization is based on uniform RVs due to the fact that these parameters must be finite. The above reformulation reflects an "uncertain" state of knowledge about these parameters and is based on means and standard deviations extracted from the following sources: [18, 19, 20, 38, 26].

Given the uncertain inputs above, the MD predictions of the QoIs are random variables with finite variance and can thus be expressed in terms of an orthogonal, Wiener–Legendre (WLe) PC basis [49, 22]. The WLe expansion of a generic QoI, X, is expressed as

$$X = \sum_{l=0}^{\infty} c_l \Psi_l(\boldsymbol{\xi}),$$

where $\{c_l\}_{l=0}^{\infty}$ is the set of PC coefficients, Ψ_l are multidimensional Legendre polynomials, and $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_\alpha\}$ are RVs uniformly distributed on $(-1, 1)$ with α

representing the stochastic dimensionality of X. In practical applications, for a given order, p, of the basis functions $\Psi_l(\boldsymbol{\xi})$, the summation above is truncated after $P+1$ terms, where $P+1 = (\alpha+p)!/\alpha!/p!$. The basis functions Ψ_l can be obtained as products of one-dimensional (1D)-Legendre polynomials using a "multi-index" construction [22].

Below, we outline and apply two approaches for determining the coefficients, c_l, $l = 0, \ldots, P$, in the truncated expansion of X, based on independent realizations of the MD system.

3.1. NISP approach. In this subsection, we discuss the NISP approach used to determine the WLe expansions of density, ρ, self-diffusivity, D, and enthalpy, H.

Let $\boldsymbol{G} = \{\rho, D, H\}$ represent the vector of uncertain observables, with $G_1 \equiv \rho$, $G_2 \equiv D$, and $G_3 \equiv H$. We model each MD observable, G_k, as

$$G_k(\boldsymbol{\xi}, m) = M_k(\boldsymbol{\xi}) + \phi(\boldsymbol{\xi}, m), \qquad k = 1, 2, 3, \tag{3.2}$$

where $\phi(\boldsymbol{\xi}, m)$ is a term accounting for the intrinsic noise which, in general, depends on the force-field parameters through the RVs $\boldsymbol{\xi}$, the number of replicas, m, of the MD system, the system size, and the time-averaging procedure, whereas $M_k(\boldsymbol{\xi})$ is a noise-independent stochastic representation that captures the effect of parametric uncertainty. In other words, $\phi(\boldsymbol{\xi}, m)$ carries the discrepancy between the observable G_k and the model M_k. Each M_k, $k = 1, 2, 3$, is modeled as a WLe expansion of the form

$$M_k(\boldsymbol{\xi}) \simeq \sum_{l=0}^{P} c_l^{(k)} \Psi_l(\boldsymbol{\xi}) \ , \qquad k = 1, 2, 3, \tag{3.3}$$

where $\left\{c_l^{(k)}\right\}_{l=0}^{P}$ denotes the set of PC coefficients for the kth observable, and the approximation sign is used to remark that we are using a truncated representation.

For a stationary MD simulation, one anticipates that as m increases, the error ϕ would decay and become negligibly small, even with fixed time averaging. Thus, as m increases, one would expect that

$$\overline{G}_k(\boldsymbol{\xi}) = M_k(\boldsymbol{\xi}), \qquad k = 1, 2, 3, \tag{3.4}$$

where the overline denotes averaging over m replicas. For finite m, however, time-averaged MD observations are expected to be distributed around $\overline{G}_k$, with small variance. This assumption is verified a priori for the MD system under consideration, as will be discussed in section 4.

Leveraging (3.4) and the orthogonality of the PC basis, the lth PC coefficient for the kth observable can be expressed as

$$\begin{aligned} c_l^{(k)} &= \frac{1}{\langle \Psi_l, \Psi_l \rangle} \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} \overline{G}_k(\xi_1,\xi_2,\xi_3)\Psi_l(\xi_1,\xi_2,\xi_3) \left[\prod_{q=1}^{3} p_f(\xi_q)\right] d\xi_1 d\xi_2 d\xi_3 \\ &= \frac{1}{\langle \Psi_l, \Psi_l \rangle} \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} \overline{G}_k(\xi_1,\xi_2,\xi_3)\Psi_l(\xi_1,\xi_2,\xi_3) \frac{1}{8} d\xi_1 d\xi_2 d\xi_3, \\ & \qquad\qquad k = 1, 2, 3, \quad l = 0, \ldots, P, \end{aligned} \tag{3.5}$$

where $p_f(\xi_q)$ denotes the probability density of ξ_q and we have used the fact that $p_f(\xi_q) = 1/2$ since $\xi_q \sim \mathcal{U}(-1, 1)$ for $q = 1, 2, 3$. The integrals in (3.5) can be suitably

evaluated using a Gauss–Legendre (GL) quadrature. Assuming n quadrature nodes along each stochastic dimension, we obtain

$$c_l^{(k)} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{n} \overline{G}_k(\xi_{1_i}, \xi_{2_j}, \xi_{3_s}) \frac{\Psi_l(\xi_{1_i}, \xi_{2_j}, \xi_{3_s})}{\langle \Psi_l, \Psi_l \rangle} w_i w_j w_s, \tag{3.6}$$
$$k = 1, 2, 3, \quad l = 0, \ldots, P,$$

where $\{(\xi_{1_i}, \xi_{2_j}, \xi_{3_s})\}_{i,j,s=1}^{n}$ are the n^3 GL integration nodes in the space $(-1, 1)^3$, $(w_i w_j w_s)$ is the product of the corresponding GL weights, and the term $\overline{G}_k(\xi_{1_i}, \xi_{2_j}, \xi_{3_s})$ represents the value of the kth observable at the quadrature node $(\xi_{1_i}, \xi_{2_j}, \xi_{3_s})$ obtained by averaging m realizations of the MD system.

Since the formula (3.6) is exact when the integrand is a polynomial of degree $p = 2n-1$ or less, the PC coefficients can be exactly determined if $\overline{G}_k$ can be described by polynomials of degree less than or equal to $(2n-1)/2$. In the present work, we use $n = 7$ quadrature nodes along each dimension and, therefore, we can exactly compute the PC coefficients for polynomials of degree up to $p = 6$. By tensorization of the 1D nodes, we obtain a grid of 343 quadrature points in three dimensions. To quantify the effects of the intrinsic noise in the MD system, we generate $m = 4$ replica simulations at each quadrature node by randomizing the initial velocity field of the atoms. This yields $N_d = n^3 \times m = 1372$ simulations to perform. As discussed in section 4, this grid is suitable for the purposes of the present setup.

In higher-dimensional settings, a full tensorization approach would be unfeasible because of the very large number of realizations that would be required. Sparse tensorization approaches can mitigate the curse of dimensionality, but their application in the context of noisy data may be problematic, namely due to the presence of negative weights in the corresponding quadrature rules [43]. A Bayesian formalism for inferring the PC coefficients provides a suitable alternative to quadrature-based projection methods. First, it allows a certain flexibility in the sampling method of the stochastic space since no specific rule is required a priori; however, the number of sampling points should be commensurate with the selected expansion order. More importantly, the Bayesian approach naturally accommodates the presence of noisy data. Specifically, whereas NISP focuses on expanding (time and replica) averaged QoIs, and a least-squares regression fitting would yield the best fit PC coefficients, the Bayesian framework allows one to capture the full information contained in the noisy data through a PC expansion with uncertain coefficients, namely PC coefficients defined in terms of a joint probability distribution. It follows that this approach not only gives an estimate of the PC coefficients but also quantifies the confidence in those coefficients in terms of the width of their posterior distribution. Such an approach is developed below for the present noisy MD system.

3.2. Bayesian inference approach. The analysis consists of three main steps: collecting a set of the observations of the QoIs, formulating the Bayesian probabilistic model, and, finally, choosing a suitable algorithm for sampling the posterior distribution.

3.2.1. Collection of observations. Given N points $\{\boldsymbol{\xi}^i\}_{i=1,\ldots,N}$ in $(-1, 1)^3$, and considering four realizations of the MD system, we build the following sets of $N \times 4$ observations:

$$\boldsymbol{G}_k = \{G_k^{i,j}\}_{i=1,\ldots,N}^{j=1,\ldots,4}, \qquad k = 1, 2, 3, \tag{3.7}$$

where $\left\{G_k^{i,j}\right\}^{j=1,\dots,4}$ represents the four values obtained for the kth observable at the ith sampling point $\boldsymbol{\xi}^i = (\xi_1^i, \xi_2^i, \xi_3^i)$. A key step in the analysis concerns the selection of the sampling points $\{\boldsymbol{\xi}^i\}_{i=1,\dots,N}$, namely to afford a suitable representation of the response surface while, at the same time, keeping N as small as possible.

We contrast two possible approaches based on GL grids and nested sampling. In the first case, the sampling points coincide with the nodes of the fully tensorized GL quadrature. This choice provides good efficiency for problems with a moderate number of dimensions and is suitable for uniformly covering the domain. Furthermore, since we used the GL points to implement the NISP approach described in section 3.1, the observations for the QoIs would already be available. A key drawback of this approach is that the corresponding grids are not nested; i.e., the grid points at a certain resolution level l' do *not* appear at higher levels $l'' > l'$. Thus, these grids are not well suited for successive or adaptive refinement. Consequently, we explore an adaptive scheme based on nested Fejér grids. Random sampling, e.g., Monte Carlo or Latin hypercube, could alternatively be applied, but a nested grid was preferred, as it provides a simple and effective approach to adaptive refinement.

Originally introduced for numerical integration algorithms, Fejér points [9] yield an increasing degree of accuracy at successive approximation levels. In one dimension, each level, l', is characterized by $n_{l'} = 2^{l'} - 1$ points in the interval $(-1, 1)$, corresponding to the abscissae of the maxima of Chebyshev polynomials of different orders. Extensions to higher-dimensional spaces can be readily obtained by tensorization. Leveraging the nested nature of these grids, we explore an adaptive technique for building a set of observations of the QoIs for the Bayesian inference. More specifically, at a given level l'', we increase the density of sampling points only in the regions of the domain where the convergence of the PC expansions inferred at levels $l' < l''$ is slower. Compared to fully tensored grids, our method yields a considerable reduction in the computational cost without penalizing the accuracy. The details of this approach are discussed in section 4.

3.2.2. Likelihood function, Bayes's theorem, and prior distributions. Assume, for the time being, that we have three sets of MD observations $\{\boldsymbol{G}_k\}_{k=1}^3$, as defined in (3.7), for density ($k = 1$), self-diffusivity ($k = 2$), and enthalpy ($k = 3$). Our goal is to infer the coefficients $\{c_l^{(k)}\}_{l=0}^P$, $k = 1, 2, 3$, appearing in the WLe expansion of each observable:

$$M_k \simeq \sum_{l=0}^{P} c_l^{(k)} \Psi_l(\boldsymbol{\xi}), \quad k = 1, 2, 3.$$

To formulate the likelihood, we express, for each observable, the discrepancy between each data point, $G_k^{i,j}$, and the corresponding model prediction, $M_k(\boldsymbol{\xi}^i)$, as

$$G_k^{i,j} = M_k(\boldsymbol{\xi}_i) + \gamma_k^{i,j}, \qquad k = 1, 2, 3, \;\; i = 1, \dots, N, \;\; j = 1, \dots, 4, \tag{3.8}$$

where $G_k^{i,j}$ represents the jth data point obtained for the kth observable at the ith sampling point, $\boldsymbol{\xi}^i$, $M_k(\boldsymbol{\xi}^i)$ denotes the value of the PC representation of the kth observable evaluated at $\boldsymbol{\xi}^i$, and $\gamma_k^{i,j}$ is an RV capturing their discrepancy.

A key step in the inference analysis [44] concerns the distribution of $\gamma_k^{i,j}$. Since the RVs $\gamma_k^{i,j}$ are introduced to account for the effect of the intrinsic noise in the MD system, and due to the fact that each data point collected corresponds to the mean of many time samples extracted from a running average in the MD simulations,

we expect, based on central limit arguments, that as the number of atoms in the system and the number of time-averaged samples become large, the distribution of G_k, $k = 1, 2, 3$, around the true mean tends to a Gaussian. A suitable and convenient choice is to assume each $\gamma_k^{i,j}$ to be i.i.d. normal RVs with mean zero and variance $\tilde{\sigma}_k^2$, i.e., $\gamma_k^{i,j} \sim \mathcal{N}(0, \tilde{\sigma}_k^2)$, $k = 1, 2, 3$, $i = 1, \ldots, N$, $j = 1, \ldots, 4$. In the present work, the variances $\{\tilde{\sigma}_k^2\}_{k=1}^3$ are treated as hyperparameters. In the above discussion, we made two key assumptions. One is that the dependence of $\tilde{\sigma}_k^2$ on the parameter space is considered negligible. This assumption appears reasonable because, in the present case, the effect of the intrinsic noise is relatively small, and its signature varies slightly with $\boldsymbol{\xi}$, as will be shown later. The second assumption made above is that each model, M_k, properly captures the true behavior of each observable up to a zero-mean error term, and, thus, no model discrepancy term appears in (3.8). This hypothesis can be verified a posteriori by comparing the maximum a posteriori (MAP) estimate of each $\tilde{\sigma}_k^2$ with the variance of the noisy data $\boldsymbol{G}_k$ for the corresponding observable. For a given order, p, of the expansion and a given observable, if the inferred value of the variance is comparable in magnitude with the actual spread in the data, this suggests that our model representation is appropriate. On the contrary, if the two values are significantly different, the order of the expansion must be refined since, in that case, model discrepancies between the approximation and the true behavior are lumped into the data noise.

The above discussion yields the following likelihood function:

$$\mathcal{L}_k\left(\{c_l^{(k)}\}_{l=0}^P, \tilde{\sigma}_k^2 \,;\boldsymbol{G}_k\right) = \prod_{i=1}^{N}\prod_{j=1}^{4} \frac{1}{\sqrt{2\pi\tilde{\sigma}_k^2}} \exp\left(-\frac{[G_k^{i,j} - M_k(\boldsymbol{\xi}^i)]^2}{2\tilde{\sigma}_k^2}\right), \qquad k = 1, 2, 3, \tag{3.9}$$

where $G_k^{i,j}$ is the jth observation obtained at the ith sampling point $\boldsymbol{\xi}^i$ for the kth observable, while $M_k(\boldsymbol{\xi}^i)$ denotes the value of the PC representation of the kth observable evaluated at the ith sampling point $\boldsymbol{\xi}^i$. Using Bayes's theorem, the joint posterior distribution, π_k, of the PC coefficients and noise variance for the kth observable can be expressed as

$$\pi_k\left(\{c_l^{(k)}\}_{l=0}^P, \tilde{\sigma}_k^2 \;\middle|\; \boldsymbol{G}_k\right) \propto \mathcal{L}_k\left(\{c_l^{(k)}\}_{l=0}^P, \tilde{\sigma}_k^2 \,;\boldsymbol{G}_k\right)\; q\big(\tilde{\sigma}_k^2\big)\; \prod_{l=0}^{P} \hat{q}_k\big(c_l^{(k)}\big), \qquad k = 1, 2, 3, \tag{3.10}$$

where $q(\tilde{\sigma}_k^2)$ and $\hat{q}_k\big(c_l^{(k)}\big)$ denote the presumed independent priors of the noise variance and the lth PC coefficient, respectively.

For the PC coefficients, we assume uniform priors of the form

$$\hat{q}_k\big(c_l^{(k)}\big) = \begin{cases} \frac{1}{\beta^{(k)} - \zeta^{(k)}} & \text{for } \zeta^{(k)} \le c_l^{(k)} \le \beta^{(k)}, \\ 0 & \text{otherwise,} \end{cases} \qquad k = 1, 2, 3, \tag{3.11}$$

where the upper, $\beta^{(k)}$, and lower, $\zeta^{(k)}$, bounds are independent of the order of the coefficient and defined solely on the basis of the type of observable and the corresponding observations. In the present study, we leverage the results obtained from the NISP approach to estimate the proper bounds above, while at the same time ensuring them to be wide enough so that no tight constraints are imposed. For problems

where no prior information is available about the ranges of the PC coefficients, the bounds can be made suitably large or, alternatively, one could use an improper prior for the coefficients. For the variances, $\{\tilde{\sigma}_k^2\}_{k=1}^3$, leveraging the fact that their values cannot be negative, we assume a Jeffreys prior [44] of the form

$$q(\tilde{\sigma}_k^2) = \frac{1}{\tilde{\sigma}_k^2}, \qquad k = 1, 2, 3. \tag{3.12}$$

3.2.3. Posterior sampling algorithm. The problem now reduces to simulating from the three posterior densities in (3.10). To this end, we employ an adaptive Metropolis (AM) algorithm (see, e.g., [12, 3, 2, 39, 40]), which is a variation of the common Metropolis Markov chain Monte Carlo (MCMC) algorithm. The main property differentiating these two methods is that while the AM builds the covariance, $\mathcal{C}$, of the proposal distribution at step t taking into account the previous visited states, the simple Metropolis uses the same covariance during the entire MCMC run. In the present work, as the proposal distribution, Q, we adopt a multivariate Gaussian, centered around the current state and whose covariance matrix, $\mathcal{C}$, is built according to

$$\mathcal{C} = \begin{cases} \mathcal{C}_0 & \text{for} \quad t < t_0, \\ \delta\, Cov_{1,2,\ldots,t} & \text{for} \quad t \geq t_0, \end{cases} \tag{3.13}$$

where t_0 defines the step at which the adaptation is triggered, $\mathcal{C}_0$ is the fixed covariance used for the initial steps, $Cov_{1,2,\ldots,t}$ denotes the covariance computed using the samples collected by the chain during all previous steps, and δ is a parameter that must be fixed before running the chain and tuned to achieve good mixing to enable efficient exploration of the target distribution. This leads to a more effective algorithm than Metropolis sampling. In the present work, we assume $t_0 = 500$ steps, and each AM chain is run for 20000 steps. The associated burn-in period, i.e., the number of steps required to locate and converge toward the center of the target distribution, is estimated by analyzing the mixing properties of the chain in terms of the autocovariance function.

4. Results.

4.1. NISP approach. To investigate the impact of the intrinsic noise, we start by computing the variance of the four MD predictions at each Gauss quadrature node in the stochastic space. Since we employ seven points along each dimension, we obtain, for the kth observable, a set of 343 samples $\{\hat{\sigma}_{k,i}^2\}_{i=1}^{343}$. We exploit these samples using kernel density estimation (KDE) to estimate the corresponding probability distribution. Note that we use $\hat{\sigma}_k$ to distinguish it from the notation used to indicate the variance $\tilde{\sigma}_k$ appearing as hyperparameter in the Bayesian inference formulation. The first column of Figure 4.1 shows the KDE of the probability distribution of the normalized variance $\hat{\sigma}_k^2/\Delta_k^2$ computed for (a) density, (c) self-diffusivity, and (e) enthalpy. For the sake of clarity, Figures 4.1(b), (d), and (f) present the same plots graphed using a logarithmic scale for the horizontal axis. We normalize the variance using $\Delta_k^2 = |\max(\overline{G}_k) - \min(\overline{G}_k)|^2$, where $\overline{G}_k$ denotes the value of the kth observable G_k averaged over four realizations of the system at a given quadrature point, and $\max(\overline{G}_k)$ and $\min(\overline{G}_k)$ are the maximum and minimum values of $\overline{G}_k$ in the domain, respectively. For the present problem, the uncertainty range has to become substantially small in order for the noise to dominate because the noise is relatively small. The range of variation of the target observable constitutes a natural scaling factor,

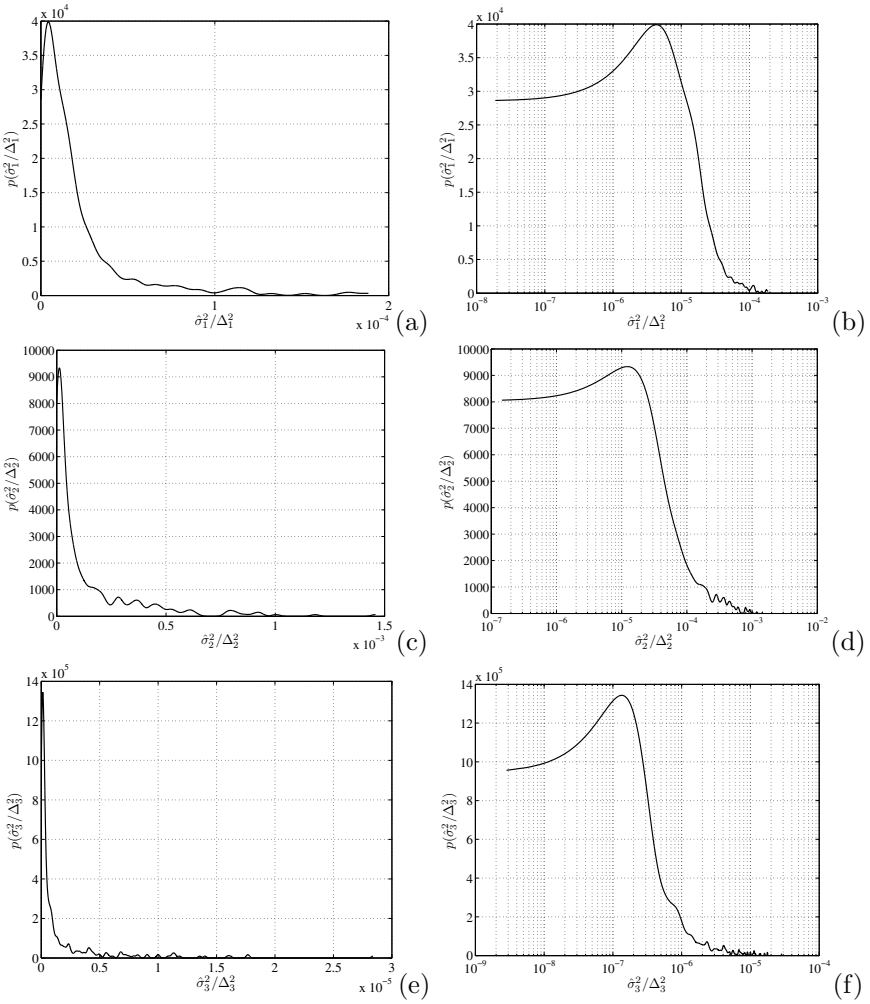


FIG. 4.1. *Left column: lin-lin plots of the probability distributions of the normalized variances* $\hat{\sigma}_k^2/\Delta_k^2$ *of the intrinsic noise for* (a) *density,* (c) *self-diffusivity, and* (e) *enthalpy obtained by applying KDE to the variances computed using the four realizations of the MD system at the* 343 *Gauss nodes. Right column: the same results graphed using a logarithmic scale for the abscissae. For a given observable,* G_k*, the corresponding normalization factor is* $\Delta_k^2 = |\max(\overline{G}_k) - \min(\overline{G}_k)|^2$*, where the overbar denotes averaging over four MD realizations, whereas* $\max(\cdot)$ *and* $\min(\cdot)$ *denote the maximum and minimum value, respectively, of the averaged observable* $\overline{G}_k$ *in the parameter space* $(-1, 1)^3$.

but the above scaling analysis has also been compared with one based on using the square of the mean of the observables, which yielded to a similar conclusion. For brevity, results obtained using this alternative scaling are omitted. Figure 4.1 allows us to estimate the effect of the intrinsic noise on the MD predictions and characterize its main features. Figures 4.1(a), (c), and (e) show that for all three cases, the distributions are peaked around very small values and reveal long tails. The behavior can be better characterized by analyzing Figures 4.1(b), (d), and (f). These plots show that the modes of the distributions for density and self-diffusivity are centered around $\sim 10^{-5}$, whereas for enthalpy they are centered around $\sim 10^{-7}$. This demonstrates

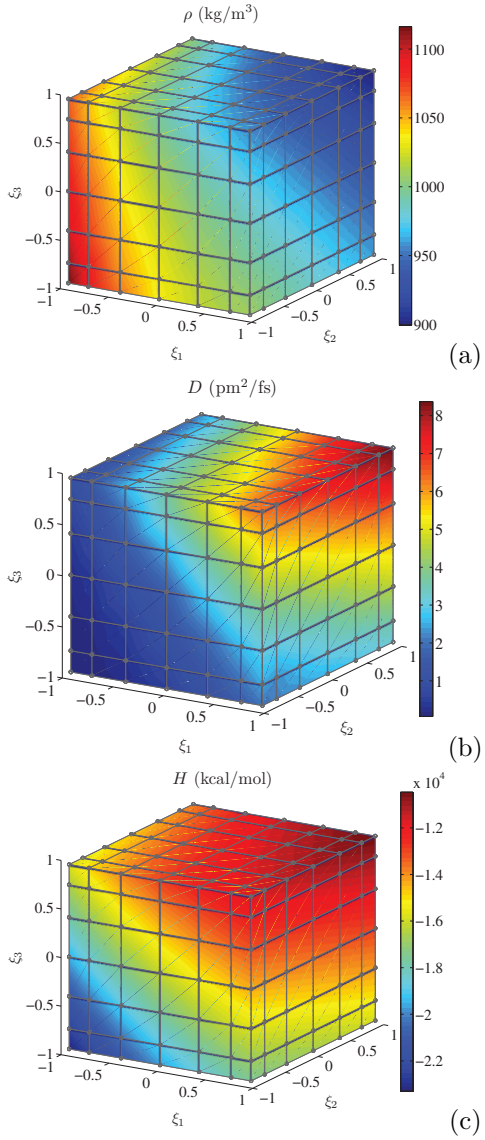


FIG. 4.2. *Functional representation of* (a) *density,* (b) *self-diffusivity, and* (c) *enthalpy in the stochastic space* (ξ_1, ξ_2, ξ_3) *obtained by collecting the corresponding realization-averaged data at the* 343 *GL quadrature nodes (black circles) and interpolating over a denser grid.*

that, for the present case, the effect of the intrinsic noise can be considered small.

The second requirement for grounding the basis for the NISP approach concerns regularity conditions in the data under consideration. This is assessed through the functional representation of the corresponding MD observations. Figure 4.2 presents the 3D plots obtained by interpolating, over a denser mesh, the values of $\overline{G}_k$ computed at the 343 GL quadrature nodes for (a) density, (b) self-diffusivity, and (c) enthalpy. This figure shows that the realization-averaged MD predictions of each observable vary smoothly, with maxima and minima located at the quadrature nodes identifying the top-right and lower-left corners of the volume. We can thus conclude that a global

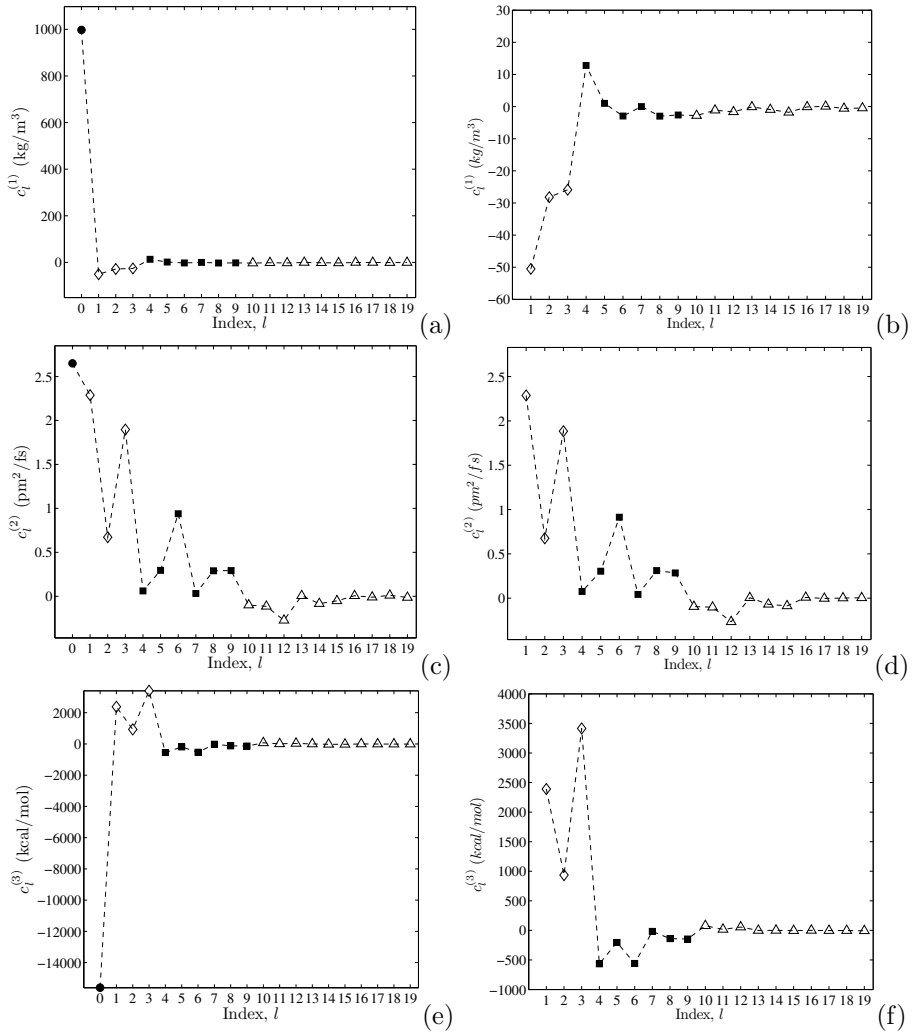


FIG. 4.3. *Left column: spectrum of coefficients for a third-order PC expansion ($p = 3$) of* (a) *density,* (c) *self-diffusivity, and* (e) *enthalpy computed using NISP. Right column: enlarged plots showing the coefficients of order $p \geq 1$ for* (b) *density,* (d) *self-diffusivity, and* (f) *enthalpy. The orders are labelled as follows: zeroth (●), first (◇), second (■), and third (△).*

NISP approach, using a smooth polynomial basis, can be suitably applied to the MD system under study.

The first column of Figure 4.3 shows the sets of PC coefficients $\left\{c_l^{(k)}\right\}_{l=0}^{19}$ for a third-order expansion $p = 3$, computed with the NISP approach for density, self-diffusivity, and enthalpy. To highlight the spectral decay characterizing the PC coefficients, we report in the second column of Figure 4.3 enlarged plots showing the behavior of the PC coefficients of order $p \geq 1$. Figures 4.3(a)–(b) and (e)–(f) show that the spectra of coefficients obtained for density and enthalpy are characterized by a rapid decay, where the zeroth-order term dominates, while the higher-order coefficients decrease sharply in magnitude. Starting with the second order, the coefficients are significantly smaller in magnitude than the leading term and settle into an os-

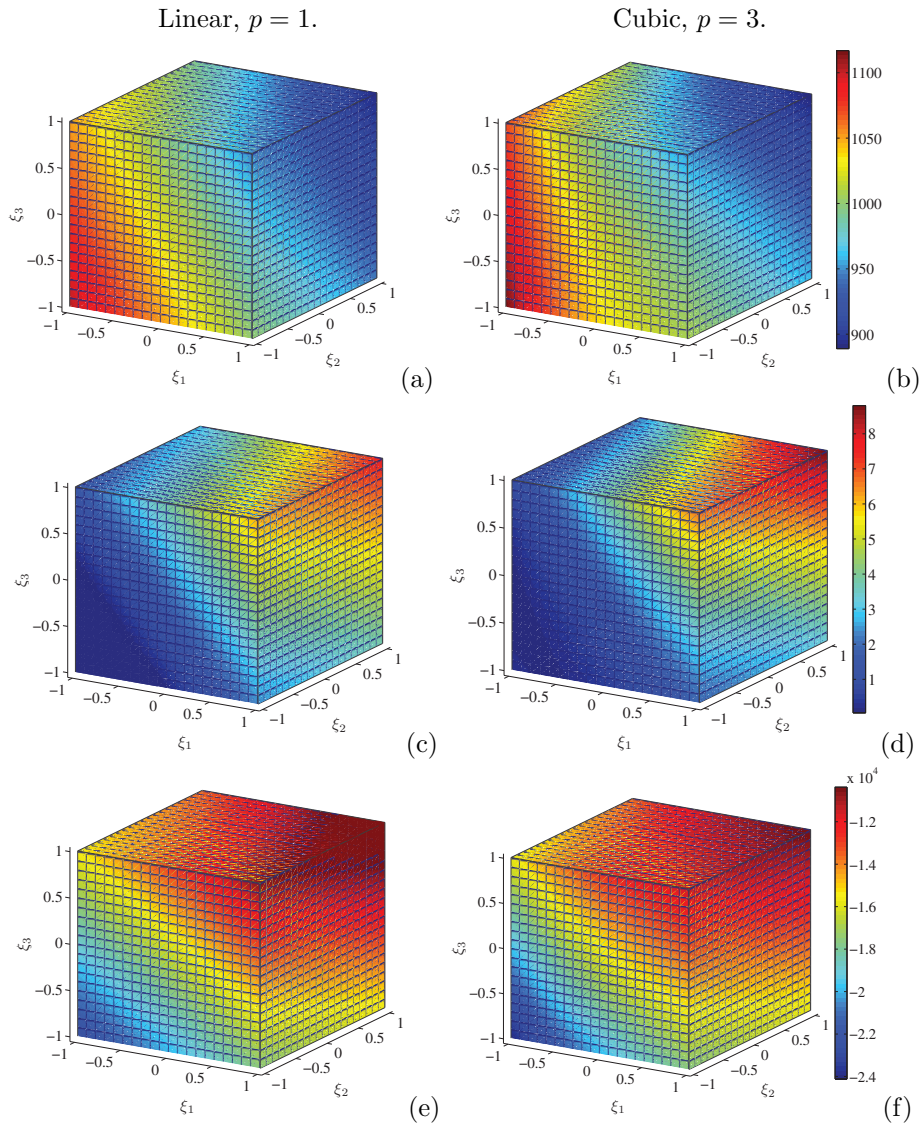


FIG. 4.4. *Plots of density (top), self-diffusivity (middle), and enthalpy (bottom) obtained using first-order (left) and third-order (right) PC expansions computed with the NISP approach. For a given observable, the same color map is used to plot both orders.*

cillatory trend around zero; see Figures 4.3(a)–(b) and (e)–(f). A slightly different behavior is obtained for self-diffusivity. In this case, the spectrum of coefficients does not decay as rapidly as for the previous two cases; see Figures 4.3(c)–(d). The first-order coefficients are comparable in magnitude to the leading term, and the steady fluctuation around zero develops only with the third order.

To investigate the effect of the order of the PC basis, we report in Figure 4.4 the 3D plots of the predictions obtained for all three observables by sampling a *first-* (left column) and *third*-order (right column) PC expansion, $M_k(\boldsymbol{\xi})$, over a dense mesh in the space $\Omega = (-1, 1)^3$. For a given observable, the same color map is used for both orders explored. The main difference is a curvature effect, particularly evident in the

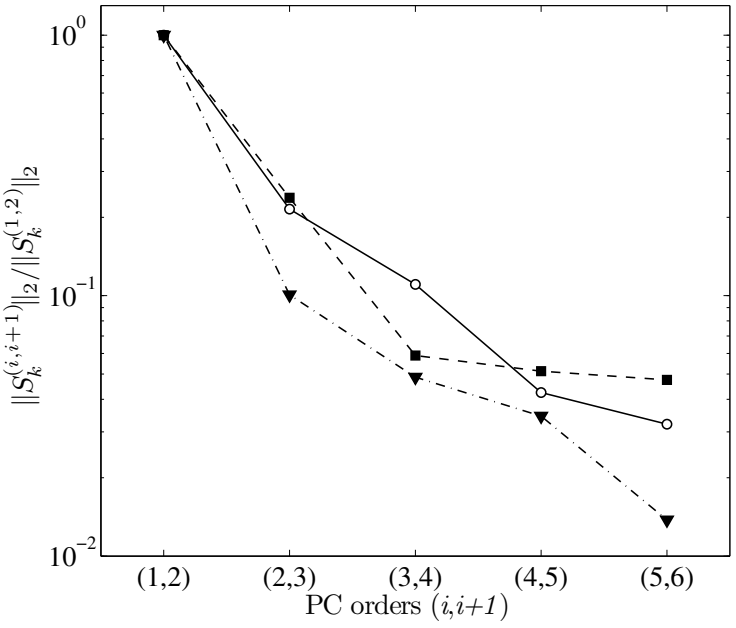


FIG. 4.5. *Decay rate* $\|S_k^{(i,i+1)}\|_2/\|S_k^{(1,2)}\|_2$ *of the discrepancy* $S_k^{(i,i+1)}$ *between the NISP-based PC representations of the* k*th observable at two subsequent orders* i *and* $(i+1)$*, for* $i = 1, 2, 3, 4, 5$*, obtained for density (solid), self-diffusivity (dashed), and enthalpy (dashed-dotted).*

plot for density; see Figures 4.4(a)–(b). For self-diffusivity and enthalpy, increasing the expansion order mainly causes a sharp change in the response surface around the top-right corner of the domain, i.e., in the neighborhood of the point $(1, 1, 1)$; see Figures 4.4(c)–(f). To quantitatively substantiate the above analysis, we define the following function:

$$S_k^{(i,i+1)}(\boldsymbol{\xi}) = |M_k^{(i)}(\boldsymbol{\xi}) - M_k^{(i+1)}(\boldsymbol{\xi})|, \qquad k = 1, 2, 3, \tag{4.1}$$

where $M_k^{(i)}$ represents the ith-order PC expansion for the kth observable. Consequently, $S_k^{(i,i+1)}(\boldsymbol{\xi})$ can be interpreted as a function capturing the pointwise spectral "error" between two expansions of subsequent orders. For a global measure, we rely on the corresponding L_2-norm given by

$$\left\|S_k^{(i,i+1)}\right\|_2 = \left(\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\left|M_k^{(i)}(\xi_1,\xi_2,\xi_3) - M_k^{(i+1)}(\xi_1,\xi_2,\xi_3)\right|^2 \frac{1}{8} d\xi_1 d\xi_2 d\xi_3\right)^{1/2}. \tag{4.2}$$

Since the integrand is at most a polynomial of degree $2(i+1)$, the above expressions can be accurately computed using GL quadrature, where the minimum number of nodes needed is $(2i+3)/2$. As previously anticipated, the NISP approach yields PC representations exact up to sixth order. Consequently, the integrand in (4.2) can be at most a 12th-order polynomial and, therefore, a minimum number of seven quadrature nodes is needed in each dimension.

Figure 4.5 shows the dependence of the ratio $\left\|S_k^{(i,i+1)}\right\|_2/\left\|S_k^{(1,2)}\right\|_2$ on the expansion order i, for $i = 1, \ldots, 5$, computed for density (solid), self-diffusivity (dashed), and enthalpy (dashed-dotted). This figure allows us to evaluate the spectral decay

rate of the PC expansions. For all three observables, the decay rate is monotonically decreasing with respect to the order. However, while the trends obtained for density and enthalpy show a regular and continuous decay for all orders considered, the results for self-diffusivity show a rapid drop for $i < 3$ and a nearly flat trend for $i \geq 3$.

4.2. Bayesian inference. Our objectives in this section are two-fold: (a) to demonstrate the Bayesian estimation of PC coefficients based on noisy data, and (b) to explore whether the PC coefficients can be suitably determined on the basis of sparse computational grids. To this end, we consider the nested Fejér grids and use the corresponding node points as sampling points where MD observables are computed. To address the second objective, we illustrate a strategy based on adaptive selection of quadrature points at successive levels. Specifically, the strategy is based on analyzing, for a given observable, the spectral convergence of the associated PC expansions at successive approximation levels. In practice, the idea behind this method can be summarized with the following two steps. First, we infer, for each observable, the corresponding PC expansion at the resolution levels $l' = 1$ and $l' = 2$ using the *full* grids of Fejér points. Second, rather than also using the full grid at level $l' = 3$, we select only a subset of nodes by identifying the regions of the domain where the spectral convergence of the expansions obtained at levels $l' = 1$ and 2 is slow. This approach can then be extended to higher-order levels $l' > 3$.

Recalling that the 3D Fejér grid at a given resolution level l' comprises $(n_{l'})^3 = (2^{l'} - 1)^3$ points, and that we have four realizations of the MD system, at level $l' = 1$ we have $4 \times (n_{l'=1})^3 = 4$ simulations, level $l' = 2$ comprises $4 \times (n_{l'=2})^3 - 4 = 104$ runs, and at level $l' = 3$ the number of simulations to run becomes $4 \times (n_{l'=3})^3 - 108 = 1264$. These values are calculated accounting for the nested nature of the Fejér grids.

4.2.1. First and second approximation levels. The first approximation level, $l' = 1$, is characterized by one sampling point in the domain, namely the origin $(0, 0, 0)$. For each observable G_k, the associated data set $\boldsymbol{G}_k|_{l'=1}$ consists of four data points corresponding to the realizations of the MD system at the sampling point $(0, 0, 0)$. In short notation,

$$\boldsymbol{G}_k|_{l'=1} = \left\{G_k^{1,j}\right\}^{j=1,\ldots,4}, \qquad k = 1, 2, 3. \tag{4.3}$$

In this case, we can infer only the leading coefficient $c_0^{(k)}$ of the PC expansions, M_k, $k = 1, 2, 3$. For brevity, the results obtained at $l' = 1$ are omitted.

At the second approximation level, $l' = 2$, the sampling grid consists of $N = 27$ points, sharing only the origin with level $l' = 1$. For each observable, G_k, the corresponding set of data points $\boldsymbol{G}_k|_{l'=2}$ comprises 108 observations, namely

$$\boldsymbol{G}_k|_{l'=2} = \left\{G_k^{i,j}\right\}_{i=1,\ldots,27}^{j=1,\ldots,4}, \qquad k = 1, 2, 3, \tag{4.4}$$

where the index i enumerates the grid points, while j enumerates the realizations of the system. Note that given the nested nature of the sampling points, we need to perform, for each observable, 104 new simulations at this level, due to the fact that the four observations at the origin $(0, 0, 0)$ are inherited from level $l' = 1$. With the data sets defined in (4.4), we can infer a *first*- and *second*-order PC expansion for each observable of interest.

As a first step of the analysis, we compare the spectrum of coefficients obtained by inferring a first-order expansion against the corresponding results obtained by inferring a second-order expansion. Figure 4.6 presents the KDE of the marginalized

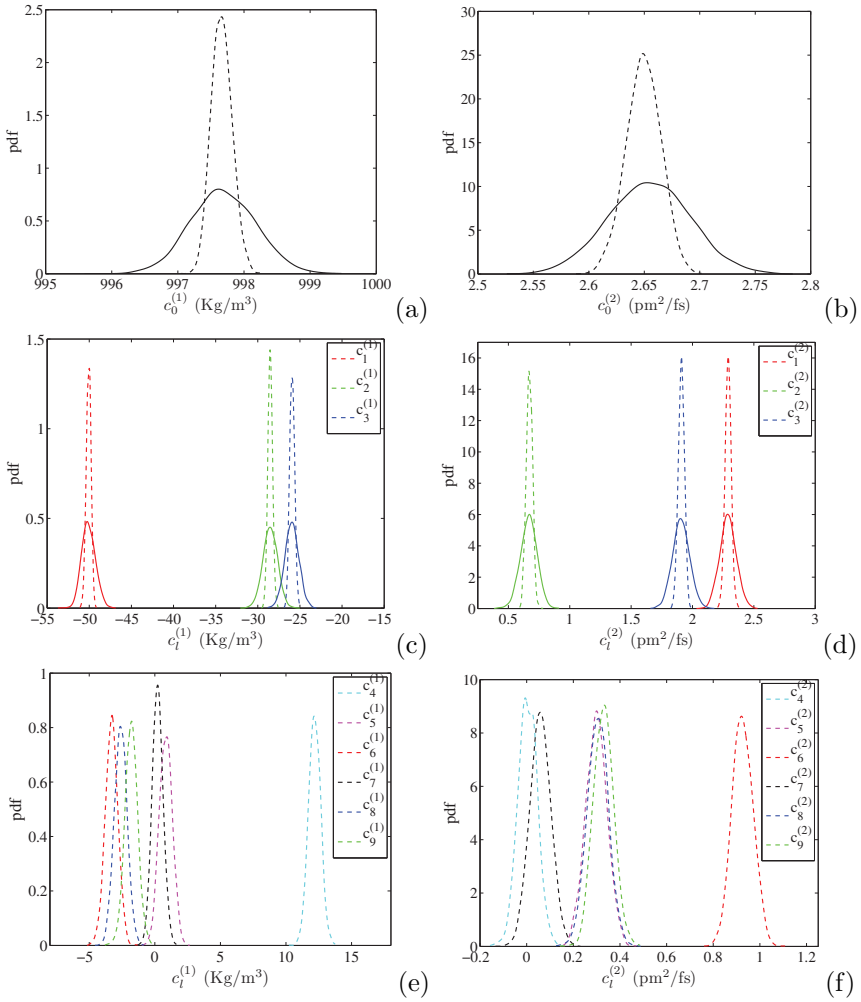


FIG. 4.6. *The solid line depicts the results obtained at level* 2 *by inferring a first-order PC expansion, whereas the dashed line depicts the corresponding results obtained at level* 2 *by inferring a second-order PC expansion. First row: marginalized posteriors of the zeroth-order PC coefficient* $\pi(c_0^{(k)})$*, computed for* (a) *density and* (b) *self-diffusivity. Second row: marginalized posteriors of the first-order PC coefficients* $\{\pi(c_l^{(k)})\}_{l=1}^3$ *for* (c) *density and* (d) *self-diffusivity. Third row: marginalized posteriors of the second-order PC coefficients* $\{\pi(c_l^{(k)})\}_{l=4}^9$ *for* (e) *density and* (f) *self-diffusivity. Different colors are used to distinguish the coefficients of the expansions.*

posteriors of the PC coefficients of ρ (first column) and D (second column) obtained by inferring first- (solid lines) and second-order (dashed lines) expansions. The PC coefficients for enthalpy, H, exhibit trends similar to those for ρ and are omitted for brevity. Specifically, the first row shows the marginalized posteriors of the zeroth-order PC coefficients $\pi(c_0^{(k)})$, $k = 1, 2, 3$. The second row shows the posterior of the first-order coefficients $\{\pi(c_l^{(k)})\}_{l=1}^3$, and the last row presents the results for the second-order PC coefficients $\{\pi(c_l^{(k)})\}_{l=4}^9$. Figures 4.6(a)–(d) show that while the MAP estimates of the zeroth- and first-order PC coefficients are nearly the same when inferred using first- and second-order expansions, major differences arise in the variances of their

distributions. The variance in the zeroth- and first-order coefficients is large when the inference is run using a linear expansion (solid line), and radically decreases when the inferred model is quadratic (dashed line), suggesting that the uncertainty in the coefficients decreases as we increase the order of the expansion. This behavior is expected because, as the order increases, the complexity of the model increases accordingly, improving the resulting representation of the data. Figures 4.6(e)–(f) show that the marginalized posteriors of the second-order coefficients are peaked around small values, close to zero in magnitude. This demonstrates a spectral decay consistent with that found using the NISP approach.

We now analyze how the inferred values of the variances $\{\tilde{\sigma}_k^2\}_{k=1}^3$ depend on the order of the PC representation. Figure 4.7(a) shows the AM chain of the noise variance $\tilde{\sigma}_1^2$ for density when the inference is run using a first-order (black line) and a second-order (red line) PC expansion. The corresponding marginalized posteriors are shown in Figure 4.7(b). The second and third rows of Figure 4.7 show the corresponding results obtained for self-diffusivity and enthalpy, respectively. For reference, a dashed-dotted line denotes a representative value of the variance directly estimated from the observations. This value is extracted by calculating the variance of the four replicas at each of the 27 nodes sampling the stochastic space and, finally, taking the mean of these 27 samples of variance. For all three observables, increasing the order of the PC representation yields a drop in the inferred variance, which becomes closer to the actual spread in the observations. More specifically, the MAP estimate of $\tilde{\sigma}_1^2$ obtained by inferring a first-order expansion is about 24 $\mathrm{kg}^2/\mathrm{m}^6$ and reduces to 2.81 $\mathrm{kg}^2/\mathrm{m}^6$ when the inference is based on a second-order expansion. This suggests that for density, improving the expansion from linear to quadratic determines a decay of one order of magnitude in the inferred variance, yielding a result comparable to actual spread in the observations, which is about 1.2 $\mathrm{kg}^2/\mathrm{m}^6$. Figures 4.7(c)–(d) show a similar trend for self-diffusivity, D, where the MAP estimate of $\tilde{\sigma}_2^2$ obtained using a first-order expansion is about 0.155 $\mathrm{pm}^4/\mathrm{fs}^2$, and reduces to 0.024 $\mathrm{pm}^4/\mathrm{fs}^2$ when the model is quadratic, becoming nearly of the same order of magnitude as its data-based estimate 0.0098 $\mathrm{pm}^4/\mathrm{fs}^2$. The behavior changes slightly for enthalpy. One the one hand, Figures 4.7(e)–(f) show that increasing the order of the PC expansion from linear to quadratic still yields a consistent reduction in the MAP estimate of $\tilde{\sigma}_3^2$ from 84139 $\mathrm{kcal}^2/\mathrm{mol}^2$ to 663 $\mathrm{kcal}^2/\mathrm{mol}^2$, i.e., a difference of two orders of magnitude. On the other hand, the MAP estimate 663 $\mathrm{kcal}^2/\mathrm{mol}^2$ is still two orders of magnitude larger than the corresponding value estimated using the data, namely 2.8039 $\mathrm{kcal}^2/\mathrm{mol}^2$. The above considerations suggest that while the MD predictions for density and self-diffusivity can be suitably represented by a quadratic expansion, higher-order models are needed to be properly describe the MD predictions of enthalpy.

4.2.2. Third approximation level. At the third approximation level, the *full* Fejér grid consists of $N = 343$ nodes, sharing a subset of 27 points with the second level. Our objective is to operate at high approximation levels ($l' \geq 3$) following a strategy that allows us to reduce the number of sampling points and decrease the computational cost, but without excessively penalizing the accuracy in the results. To this end, we construct an algorithm that quantifies the convergence of the PC expansions inferred at levels 1 and 2 and refines the density of grid points at level 3 only in the regions of the domain where differences in the corresponding representations exceed a user-specified threshold.

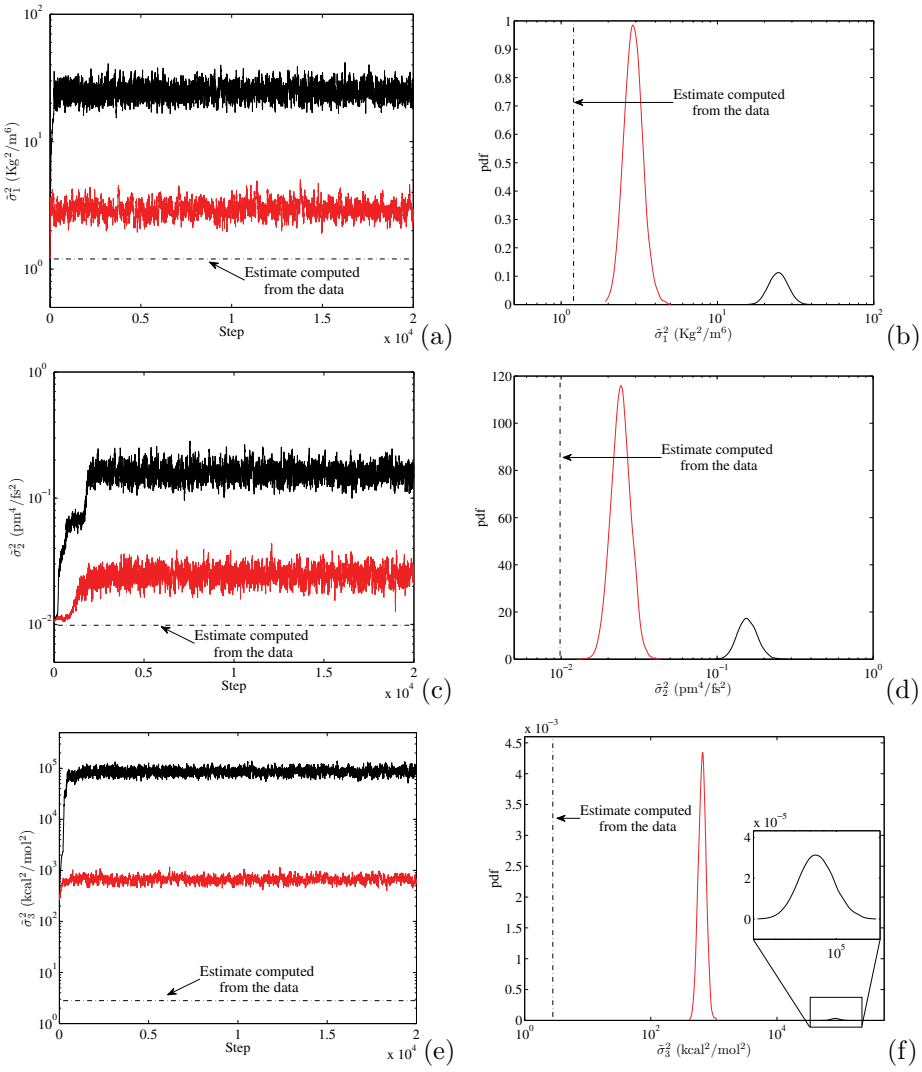


FIG. 4.7. *First row: AM chain* (a) *for the noise variance* $\tilde{\sigma}_k^2$ *computed for density and the corresponding marginalized posterior* (b) *obtained via KDE. Black lines identify the results obtained at level* 2 *by inferring a first-order PC expansion, whereas red lines (distinguishable online only) indicate the results obtained at the same level by inferring a second-order PC expansion. The value of the variance estimated directly from the set of observations used for the inference is denoted by a dashed-dotted line. The results obtained for self-diffusivity and enthalpy are presented in the second row and third row, respectively.*

To this end, we rely on the following function:

$$Z_k^{(l'=1,2)}(\boldsymbol{\xi}) = \left| M_k^{(l'=1,p=0)}(\boldsymbol{\xi}) - M_k^{(l'=2,p=2)}(\boldsymbol{\xi}) \right|, \qquad k = 1, 2, 3, \tag{4.5}$$

where $M_k^{(l'=1,p=0)}$ represents the zeroth-order ($p = 0$) expansion of the kth observable inferred at level 1, while $M_k^{(l'=2,p=2)}$ represents the second-order ($p = 2$) expansion of the kth observable inferred at level 2. As discussed in the previous sections, the PC coefficients defining the expansions $M_k^{(l'=1,p=0)}$ and $M_k^{(l'=2,p=2)}$ are RVs charac-

terized by a well-defined joint probability distribution. In the present work, rather than developing the analysis using the full joint posterior of the coefficients, we simplify the approach and rely on the MAP estimates of their marginalized posteriors. The validity of this approximation is based on the following two arguments. First, the Bayesian analysis yields, for this case, joint posteriors for the coefficients characterized by weak or nearly absent correlation. The weak correlation structure can be verified in Figure 4.8, which shows the matrix of correlation coefficients $-1 \leq \eta(i,j) \leq 1$, $i,j = 1,\dots,P$, computed for the PC spectrum of the second-order ($P = 9$) expansions obtained at level 2 for density, self-diffusion, and enthalpy. Each matrix entry, (i,j), $i,j = 1,\dots,P$, is color coded based on the corresponding value $\eta(i,j)$. The plots show negligible correlation in the off-diagonal entries, which suggests that the correlation structure in the PC spectra can be neglected. Second, one observes that the marginalized posteriors of the PC spectra obtained at levels 1 (not shown) and 2 (Figure 4.6) are symmetrical and tight around the corresponding modes, implying that a Gaussian distribution is a suitable fit for these posterior densities and that their means coincide with their MAP estimates. Moreover, we also verified that, for each observable, the corresponding joint posterior of the PC coefficients closely resembles a multivariate Gaussian, but for brevity we omit the discussion. The above considerations suggest that it is appropriate to treat the PC coefficients in terms of their marginalized posterior. In general, particularly when the above features do not hold, an extended strategy should be applied that exploits the full joint distribution of the PC coefficients.

The function $Z_k^{(l'=1,2)}(\boldsymbol{\xi})$ captures the pointwise "error" between the expansions inferred at levels $l' = 1$ and $l' = 2$. A global measure of this "error" can be evaluated in terms of the L_2-norm according to

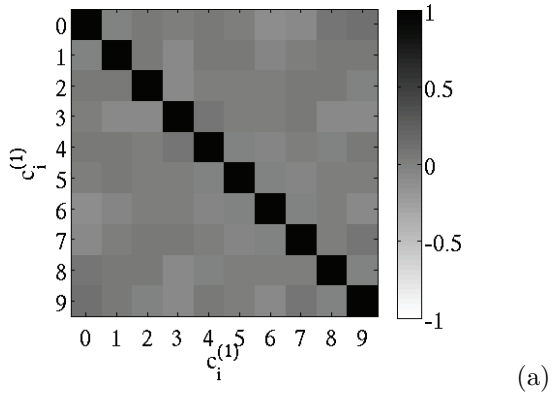
$$\left\| Z_k^{(l'=1,2)} \right\|_2 = \left(\int_\Omega \left| M_k^{(l'=1,p=0)}(\boldsymbol{\xi}) - M_k^{(l'=2,p=2)}(\boldsymbol{\xi}) \right|^2 \frac{1}{8} d\boldsymbol{\xi} \right)^{1/2}, \quad k = 1,2,3, \tag{4.6}$$

where $\Omega = (-1,1)^3$. Exploiting the fact that the integrand is a fourth-order polynomial, the above integrals can be computed exactly using the cubature.

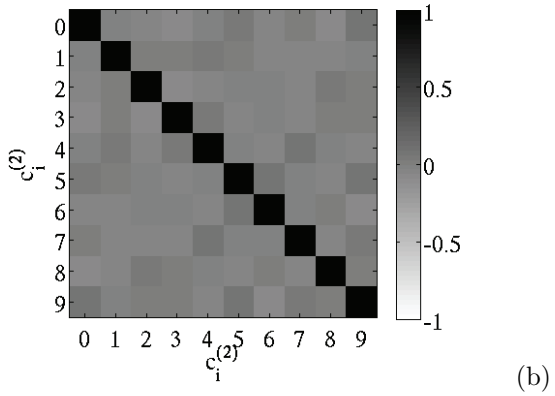
The first and second columns of Figure 4.9 show two different views of the 3D plot of $Z_k^{(l'=1,2)}(\boldsymbol{\xi})$ obtained for density, self-diffusivity, and enthalpy. The minima of $Z_k^{(l'=1,2)}$ identify the central region of the space as the region where there is close agreement between the representations inferred at levels 1 and 2, while the maxima of $Z_k^{(l'=1,2)}$ localize near the corners $(-1,-1,-1)$ and $(1,1,1)$. In particular, Figures 4.9(a)–(b) show that, for density, these two corners are characterized by nearly the same magnitude of the error $Z_1^{(l'=1,2)}$. The result is slightly different for the other two observables. The maximum error is concentrated near the corner $(1,1,1)$ for self-diffusivity and near the point $(-1,-1,-1)$ for enthalpy.

By analyzing the distribution of $Z_k^{(l'=1,2)}$, $k = 1,2,3$, we can identify which points of the full grid at $l' = 3$ are characterized by the highest values of $Z_k^{(l'=1,2)}$. The right column of Figure 4.9 shows the grids of 316 points exclusively belonging to the third approximation level $l' = 3$ (i.e., we omit those shared with levels $l' = 1$ and $l' = 2$), depicted such that the size of the marker associated with the ith node $\boldsymbol{\xi}^i$ is scaled according to the associated value of $Z_k^{(l'=1,2)}(\boldsymbol{\xi}^i)$. These plots give a first visual intuition of which subset will be selected and which will be neglected.

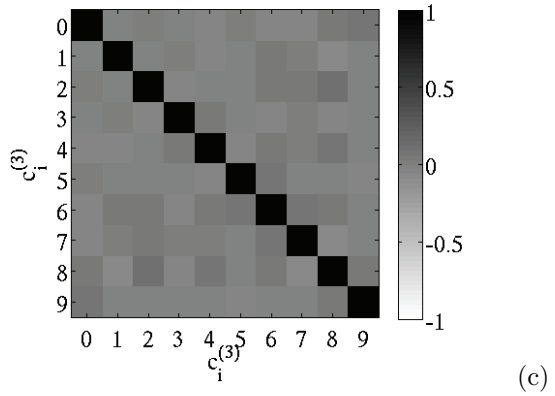
We proceed as follows. We start by nondimensionalizing the "error" $Z_k^{(l'=1,2)}(\boldsymbol{\xi})$



(a)



(b)



(c)

FIG. 4.8. *Matrix of correlation coefficients* $-1 \leq \eta(i,j) \leq 1$, $i,j = 1,\ldots,P$, *computed from the distribution of the PC coefficients of the second-order* ($P = 9$) *PC expansion obtained at level* 2 *for* (a) *density,* (b) *self-diffusion, and* (c) *enthalpy.*

for each observable as

$$\hat{Z}_k(\boldsymbol{\xi}) = \frac{Z_k^{(l'=1,2)}(\boldsymbol{\xi})}{\max_\Omega \left(Z_k^{(l'=1,2)}\right)}, \qquad k = 1,2,3, \tag{4.7}$$

where the normalization factor corresponds to the maximum value of $Z_k^{(l'=1,2)}$ in the

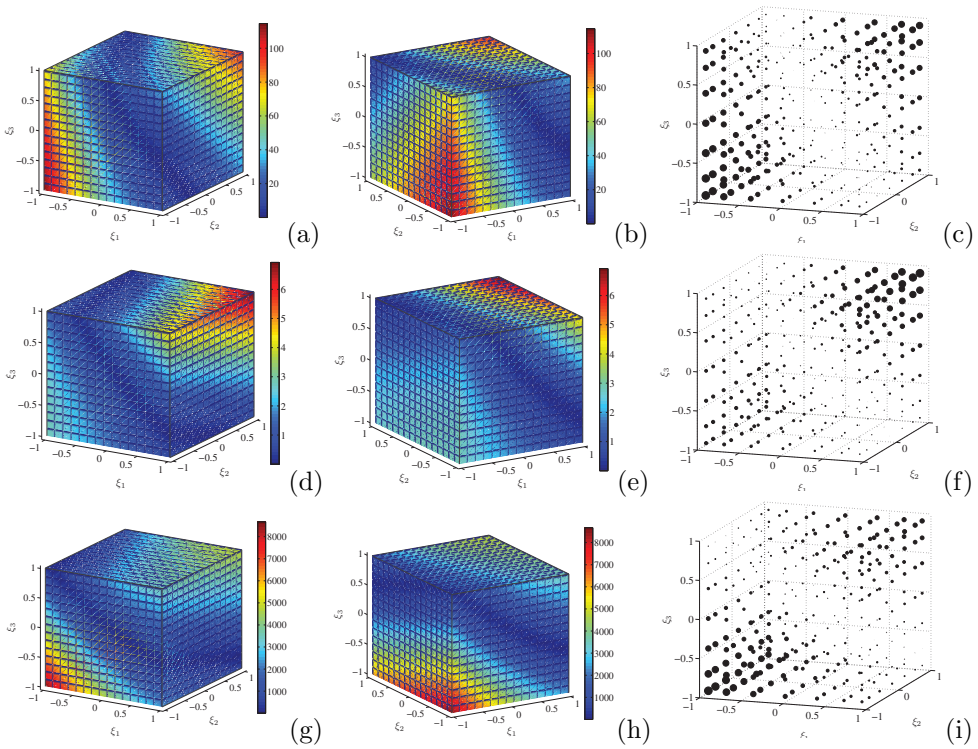


FIG. 4.9. *The first and second columns show two different views of the contours of* $Z_k^{(l'=1,2)}$ *obtained for density* (a), (b), *self-diffusivity* (d), (e), *and enthalpy* (g), (h). *The third column shows the distributions of Fejér grid points* $\{\boldsymbol{\xi}_i\}_{i=1}^{316}$ *at level* 3 *(omitting the subset shared with levels* 1 *and* 2*) obtained for* (c) *density,* (f) *self-diffusivity, and* (i) *enthalpy, where each point* $\boldsymbol{\xi}_i$ *is represented by a marker whose size is scaled according the corresponding value of* $Z_k^{(l'=1,2)}(\boldsymbol{\xi}_i)$.

domain $\Omega = (-1,1)^3$. Using a tolerance λ $(0 \le \lambda \le 1)$, we define $A_k^{(\lambda)}$ to be the set of *new* sampling nodes for the kth observable at level 3 according to

$$A_k^{(\lambda)} = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^{N_k^{(\lambda)}}\} \doteq \{\boldsymbol{\xi}^i : \quad \hat{Z}_k(\boldsymbol{\xi}^i) \ge \lambda, \quad i = 1, \ldots, 316\}, \quad k = 1,2,3, \tag{4.8}$$

where $N_k^{(\lambda)}$ represents the number of points in the resulting reduced grid, which depends on the type of observable and on the value of λ, while the index i enumerates the 316 points that belong exclusively to the full grid at $l' = 3$. Evidently, for any given λ, we must have $N_k^{(\lambda)} \le 316$, and $N_k^{(\lambda)} = 316$ when $\lambda = 0$. In other words, for a given observable, we select at level $l' = 3$ only the subset of points where the error is larger than the desired tolerance.

As before, considering four realizations of the MD system, each $A_k^{(\lambda)}$, $k = 1,2,3$, yields a set of *new* observations $\{G_k^{i,j}\}_{i=1,\ldots,N_k^{(\lambda)}}^{j=1,\ldots,4}$ at $l' = 3$ comprised of $N_k^{(\lambda)} \times 4$ data points. These new observations can be combined with those collected at levels 1 and 2 to form a single set of data points for the inference according to

$$\boldsymbol{G}_k^{(\lambda)}\Big|_{l'=3} \equiv \{G_k^{i,j}\}_{i=1,\ldots,27}^{j=1,\ldots,4}\Big|_{l'=2} \cup \{G_k^{i,j}\}_{i=1,\ldots,N_k^{(\lambda)}}^{j=1,\ldots,4}, \qquad k = 1,2,3, \tag{4.9}$$

where $\{G_k^{i,j}\}_{i=1,\ldots,27}^{j=1,\ldots,4}\Big|_{l'=2}$ represents the sets of observations collected for the kth

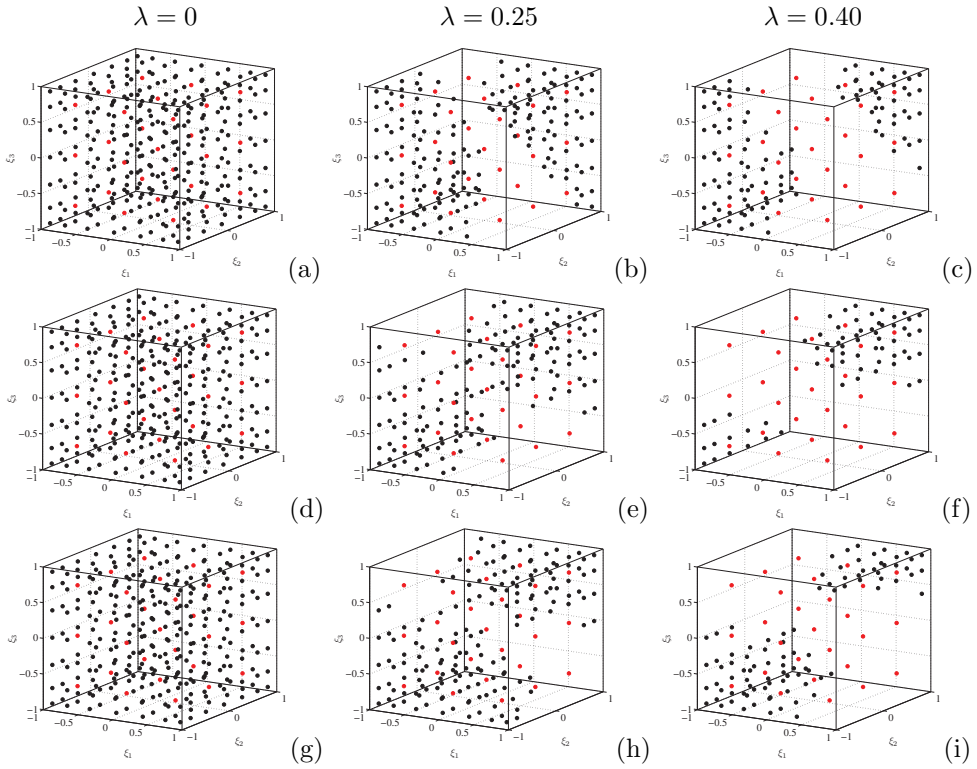


FIG. 4.10. *First row: sampling grids based on Fejér nodes obtained at level* 3 *for density using* (a) $\lambda = 0$, (b) $\lambda = 0.25$, *and* (c) $\lambda = 0.40$. *The corresponding results for self-diffusivity and enthalpy are shown in the second and third rows, respectively. We color code red the markers identifying the points shared with levels* 1 *and* 2.

observable at level 2, which already includes those at level 1, while $\{G_k^{i,j}\}_{i=1,\dots,N_k^{(\lambda)}}^{j=1,\dots,4}$ represents the set of $N_k^{(\lambda)} \times 4$ observations collected for the kth observable at level 3 for a given value of the tolerance λ, as discussed above. As a consequence, the total number of data points at level 3 obtained for the kth observable for a given λ is given by $\{4 \times (27 + N_k^{(\lambda)})\}$.

In the present study, we explore the following values of the tolerance: $\lambda = 0$, 0.25, and 0.4. Figure 4.10 shows the resulting grids of points obtained at level 3 for density, self-diffusivity and enthalpy, and all three λ values. We color code red the subset of points shared with levels 1 and 2, whereas black is used for those that belong exclusively to $l' = 3$. As the value of λ increases, the number of new points selected decreases, but their concentration remains high near the corners $(1, 1, 1)$ and $(-1, -1, -1)$, where, as discussed before, the "error" $Z_k^{(l'=1,2)}$ is large. Note that for $\lambda = 0$ the grid is independent of the observable and consists of $N = 343$ points in all three cases; see the first column of Figure 4.10. For $\lambda = 0.25$, the *total* number of points at $l' = 3$ reduces to $N = 207$ for density, $N = 172$ for self-diffusivity, and $N = 193$ for enthalpy; see Figures 4.10(b), (e), and (h). Finally, the case $\lambda = 0.40$ yields $N = 136$ for density, $N = 92$ for self-diffusivity, and $N = 119$ for enthalpy; see Figures 4.10(c), (f), and (i).

The sets of observations at level $l' = 3$ defined in (4.9) for different values of λ,

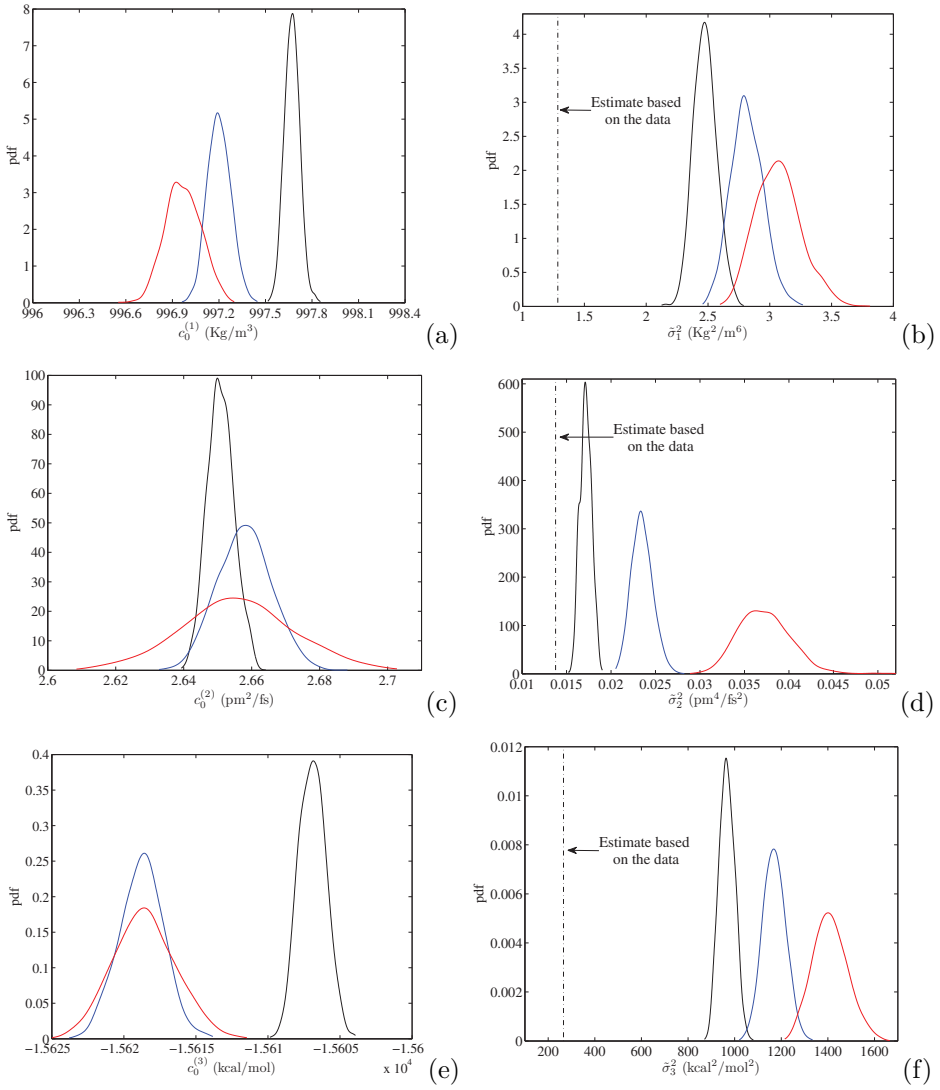


FIG. 4.11. *First row: marginalized posteriors of the zeroth-order PC coefficient* $\pi(c_0^{(k)})$ (a) *and of the noise variance* $\pi(\tilde{\sigma}_k^2)$ (b) *computed for density by inferring a third-order PC expansion at level* 3 *using the observations obtained with* $\lambda = 0$ *(black),* $\lambda = 0.25$ *(blue), and* $\lambda = 0.40$ *(red). The corresponding results for self-diffusivity and enthalpy are presented in the second and third rows, respectively.*

i.e., $\lambda = 0$, 0.25, and 0.40, can be exploited within the Bayesian inference framework discussed in section 3.2. We focus our attention on inferring for each observable a *third-order* PC expansion. Figures 4.11(a)–(b) show the marginalized posteriors of the zeroth-order PC coefficient $\pi(c_0^{(k)})$ and variance $\pi(\tilde{\sigma}_k^2)$ computed for density by inferring at level $l' = 3$ a third-order PC expansion using the set of observations obtained for $\lambda = 0$, 0.25, and 0.40. The corresponding results for self-diffusivity and enthalpy are shown in the second and third rows, respectively. Figures 4.11(a), (c), and (e) show that the MAP estimate of the zeroth-order coefficient varies insignifi-

cantly as λ changes from 0 to 0.40. On the contrary, the width of the posterior is affected noticeably, becoming wider as λ increases. This can be explained by recalling that the total number of sampling points at $l' = 3$ decreases as λ increases from 0 to 0.40, and therefore a larger uncertainty arises as λ increases.

Figures 4.11(b), (d), and (f) show the marginalized posteriors $\pi(\tilde{\sigma}_k^2)$ for density, self-diffusivity, and enthalpy. For reference, we identify in the figures with a red line the estimate of the actual spread in data. The results show that the marginalized posterior of the noise variances tends to shift towards high values as λ increases from 0 to 0.40, thus departing from the representative value estimated from the observations. This means that increasing λ yields an overestimation of the noise variance. Comparing the MAP estimate of the inferred noise variance against its value estimated from the data allows us to assess whether a third-order PC representation is a suitable choice for the given λ. To this end, we report in Table 4.1 the values of the ratio

$$r_{(k,\lambda)} = \frac{MAP_\lambda(\tilde{\sigma}_k^2)}{(\tilde{\sigma}_k^2)_{data}}, \tag{4.10}$$

where $MAP_\lambda(\tilde{\sigma}_k^2)$ is the MAP estimate of the inferred noise variance computed for the kth observable and a given value of λ, while $(\tilde{\sigma}_k^2)_{data}$ is the value estimated directly from the data. For all three observables, Table 4.1 shows that $r_{(k,\lambda)} \sim O(1)$ regardless of the value of the tolerance λ considered. The above considerations suggest that, for the present case, a third-order expansion is suitable.

TABLE 4.1

Values of the ratio $r_{(k,\lambda)}$ (see (4.10)) computed for density, self-diffusivity, and enthalpy using $\lambda = 0$, 0.25, and 0.40.

	Density	Self-diffusivity	Enthalpy
$\lambda = 0$	$r = 1.923$	$r = 1.242$	$r = 3.625$
$\lambda = 0.25$	$r = 2.172$	$r = 1.695$	$r = 4.397$
$\lambda = 0.40$	$r = 2.391$	$r = 2.631$	$r = 5.270$

The first column of Figure 4.12 shows the MAP estimates of the spectra of coefficients $\{c_l^{(k)}\}_{l=0}^P$ for a third-order expansion ($p = 3$, $P = 19$) of density, self-diffusivity, and enthalpy, inferred at level 3 using $\lambda = 0$. The observed behavior is similar to that found using the NISP approach (Figure 4.3). The spectra of coefficients computed for density and enthalpy are characterized by a rapid decay; namely, the zeroth-order term is dominant, and the higher-order coefficients decrease in magnitude very rapidly toward zero; see Figures 4.12(a) and (e). The spectrum of self-diffusivity does not converge as rapidly as for the other two cases; see Figure 4.12(c). In this case, the coefficients up to second order are comparable in magnitude to the zeroth-order term, and a significant drop occurs only at third order.

To investigate the dependence of the MAP estimates of the coefficients on λ, we present in the second column of Figure 4.12 the normalized difference $|c_l^{(k)} - c_{l,\lambda}^{(k)}|/|c_0^{(k)}|$, $l = 0, \ldots, 19$, where $c_l^{(k)}$ represents the MAP estimate of the lth PC coefficient for the kth observable inferred at $l' = 3$ using the full grid, i.e., $\lambda = 0$, whereas $c_{l,\lambda}^{(k)}$ is the MAP estimate of the corresponding coefficient inferred at $l' = 3$ using $\lambda = 0.25$ or $\lambda = 0.40$. On the one hand, panels (b) and (f) show that for density and enthalpy the differences are minimal, of the order $\sim O(10^{-4})$ for density and $\sim O(10^{-3})$ for enthalpy. For self-diffusivity, the discrepancy increases slightly, becoming $\sim O(10^{-2})$;

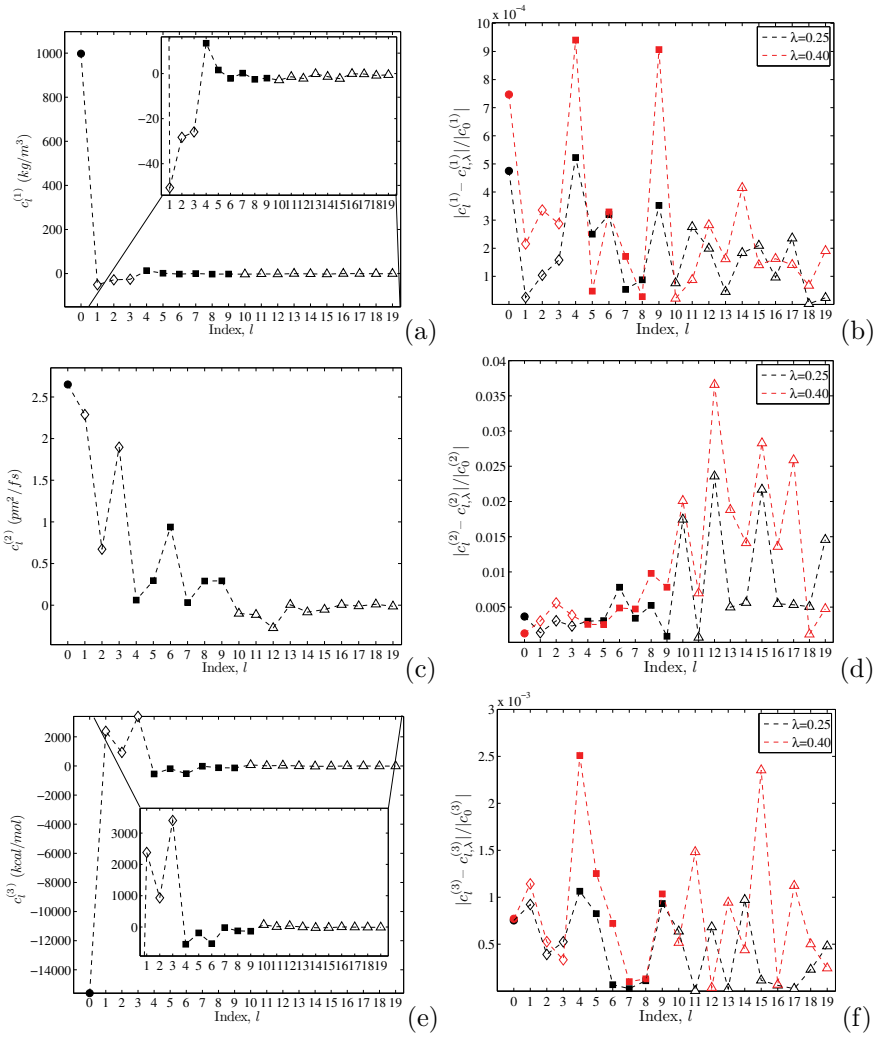


FIG. 4.12. *Left column: MAP estimates of the PC coefficients* $\{c_l^{(k)}\}_{l=0}^{P}$ *for a third-order expansion of* (a) *density,* (c) *self-diffusivity, and* (e) *enthalpy obtained at the level* 3 *using the full Fejér grid, i.e.,* $\lambda = 0$. *Right column: plots of the normalized "discrepancy"* $\{|c_l^{(k)} - c_{l,\lambda}^{(k)}|/|c_0^{(k)}|\}_{l=0}^{19}$, *for* (b) ρ, (d) D, *and* (f) H, *where* $c_{l,\lambda}^{(k)}$ *is the MAP estimate of the* l*th PC coefficient inferred at* $l' = 3$ *using the set of observations derived for* $\lambda = 0.25$ *or* $\lambda = 0.40$, *while* $|c_l^{(k)}|$ *is the absolute value of the MAP estimate of the corresponding coefficient obtained using* $\lambda = 0$. *Subsequent orders are identified by the following markers: zeroth- (●), first- (◇), second- (■), and third-order coefficients (△).*

see Figure 4.12(d). An interesting observation concerns the dependence of the error on the order of the coefficients. Panels (b) and (f) show that for density and enthalpy the discrepancy does not substantially vary with the order of the coefficients. A different trend is observed for self-diffusivity, where a large discrepancy characterizes the high-order coefficients $l > 10$, while smaller deviations occur for low-order coefficients; see panel (d). The difference might be due to the fact that the spectra of density and enthalpy decay rapidly, whereas the spectrum of self-diffusivity exhibits slower decay.

We now analyze how the error depends on λ. The second column of Figure 4.12

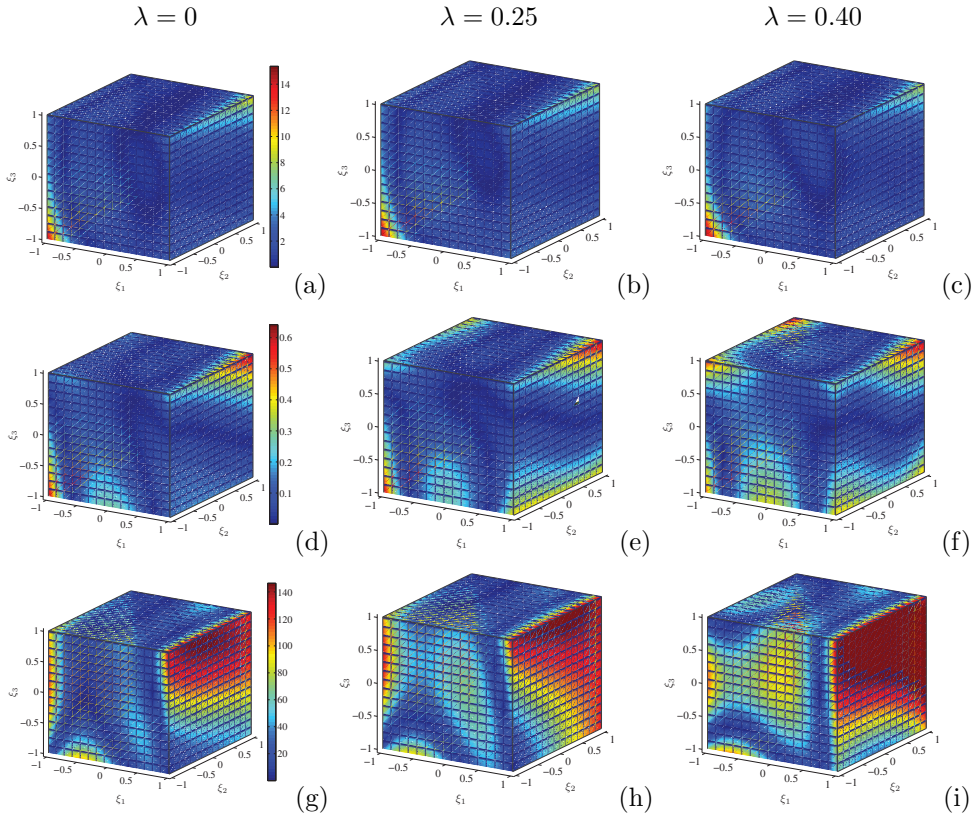


FIG. 4.13. *First row: contour plots of* (a) $Z_{1,\lambda=0}^{(l'=2,3)}$, (b) $Z_{1,\lambda=0.25}^{(l'=2,3)}$, *and* (c) $Z_{1,\lambda=0.40}^{(l'=2,3)}$ *showing the discrepancy in the PC representations of density inferred at levels* 2 *and* 3 *for different values of* λ. *The corresponding results for self-diffusivity are shown in panels* (d)–(f), *whereas those for enthalpy are presented in panels* (g)–(i). *For a given observable, the same color map is applied to all three figures.*

shows that, for a given observable, the discrepancy between the MAP estimate of the coefficients obtained at $l' = 3$ using $\lambda = 0$ and the corresponding result obtained for $\lambda \neq 0$ increases slightly as λ increases from 0.25 to 0.40. This is expected because as λ decreases from 0.40 to 0.25, the set of sampling nodes is larger and approaches the full grid. To further characterize the dependence on λ, we contrast the PC coefficients estimated at level 2 with those obtained at level 3 using $\lambda = 0$, 0.25, and 0.40. To this end, we rely on the following function:

(4.11)
$$Z_{k,\lambda}^{(l'=2,3)}(\boldsymbol{\xi}) = \left| M_k^{(l'=2,p=2)}(\boldsymbol{\xi}) - M_{k,\lambda}^{(l'=3,p=3)}(\boldsymbol{\xi}) \right|, \quad k = 1,2,3, \quad \lambda = 0, 0.25, 040,$$

where $M_k^{(l'=2,p=2)}$ represents the second-order ($p = 2$) PC expansion inferred for the kth observable at level 2, while $M_{k,\lambda}^{(l'=3,p=3)}$ represents the corresponding third-order ($p = 3$) expansion inferred at level 3 using the set of observations built for a given value of λ. As representative values of the PC coefficients of $M_k^{(l'=2,p=2)}$ and $M_{k,\lambda}^{(l'=3,p=3)}$ we rely on the MAP estimates computed from the corresponding marginalized posterior distributions.

The first row of Figure 4.13 shows the contour plots of $Z_{1,\lambda}^{(l'=2,3)}$, illustrating the

discrepancies for density between levels 2 and 3 with different λ. The corresponding results for $Z_{2,\lambda}^{(l'=2,3)}$ (self-diffusivity) and $Z_{3,\lambda}^{(l'=2,3)}$ (enthalpy) are presented in the second and third rows, respectively. Figures 4.13(a)–(c) show that for density, as λ increases from 0 to 0.40, the behavior of $Z_{1,\lambda}^{(l'=2,3)}$ does not change drastically. The space is dominated mainly by small values of $Z_{1,\lambda}^{(l'=2,3)}$, while its maxima are confined in a small neighborhood of the two corners $(-1,-1,-1)$ and $(1,1,1)$. For self-diffusivity, the trend is similar, but a nonnegligible error spreads along the side of the domain. The contour plots change substantially for enthalpy, where high values are not only limited to the corners $(-1,-1,-1)$ and $(1,1,1)$ but spread all over the sides of the domain.

To quantify the observations above, we compute the following L_2-norms:

$$\begin{aligned}\left\|Z_{k,\lambda}^{(l'=2,3)}\right\|_2 &= \left(\int_\Omega \left|Z_{k,\lambda}^{(l'=2,3)}\right|^2 \frac{1}{8}d\boldsymbol{\xi}\right)^{1/2} \\ &= \left(\int_\Omega \left|M_k^{(l'=2,p=2)}(\boldsymbol{\xi}) - M_{k,\lambda}^{(l'=3,p=3)}(\boldsymbol{\xi})\right|^2 \frac{1}{8}d\boldsymbol{\xi}\right)^{1/2}, \\ & \qquad\qquad k=1,2,3, \quad \lambda = 0, 0.25, 040, \end{aligned} \tag{4.12}$$

where $\Omega = (-1,1)^3$. Table 4.2 shows the values of $\|Z_k^{(l'=1,2)}\|_2$ (4.6) and $\|Z_{k,\lambda}^{(l'=2,3)}\|_2$ (4.12), nondimensionalized by the absolute value of the MAP estimate of $c_0^{(k)}$ inferred at $l'=3$ using $\lambda = 0$. The results in Table 4.2 reveal that, for a given observable, increasing the order of the PC expansion from quadratic to cubic yields a substantial reduction in these global measures. Moreover, the estimates obtained for $\lambda = 0.25$ and $\lambda = 0.4$ are nearly the same as those obtained using the full grid. These trends are also illustrated in the left column of Figure 4.14, which depicts the norms using a logarithmic scale. On the right column of Figure 4.14, we plot as a function of the approximation level, l', the corresponding total number of sampling points, N, used for the Bayesian inference for density, self-diffusivity, and enthalpy. The number of sampling points, N, rapidly increases as l' increases from 1 to 3, where the full grid ($\lambda = 0$) comprises 343 points. Choosing $\lambda = 0.40$ or 0.25 allows us to reduce the total number of points at level 3 to less than half the corresponding value of the full grid, yielding a significant reduction in the computational cost. For instance, considering density with $\lambda = 0.40$, the total number of sampling points at level 3 is 136. Therefore, given four realizations of the MD system, the corresponding number of simulations to perform reduces from $343 \cdot 4 = 1372$ for the full grid to $136 \cdot 4 = 544$ for the adapted grid. A comparison between the left and right columns of Figure 4.14 allows us to conclude that considerable reduction in the computational cost can be accomplished by adaptively refining the grid, with only a small increase in the estimated measures.

The grid adaptation process described above can be further extended to higher levels if necessary. A possible measure for monitoring the impact of refinement can be based on the normalized relative error

$$\frac{\left\|Z_{k,\lambda}^{(l',l'+1)}\right\|_2}{|c_0^{(k)}|}, \qquad k=1,2,3, \tag{4.13}$$

where λ is a given tolerance introduced to optimize the refinement of grid points at high approximation levels, and $|c_0^{(k)}|$ is the leading term of the expansion at level l' or, equivalently, the mean of the data. In the present case, as shown by Table 4.2, the third level, $l'=3$, already provides an acceptable approximation level.

TABLE 4.2

L_2-norms of $Z_k^{(l'=1,2)}$ and $Z_{k,\lambda}^{(l'=2,3)}$ nondimensionalized by the absolute value of the MAP estimate of $c_0^{(k)}$ inferred at level 3 using $\lambda = 0$.

	Density ($k = 1$)	Self-diffusion ($k = 2$)	Enthalpy ($k = 3$)
$\|Z_k^{(l'=1,2)}\|_2/\|c_0^{(k)}\|$	0.10527	1.93094	0.45349
$\|Z_{k,\lambda=0}^{(l'=2,3)}\|_2/\|c_0^{(k)}\|$	0.00484	0.10226	0.00619
$\|Z_{k,\lambda=0.25}^{(l'=2,3)}\|_2/\|c_0^{(k)}\|$	0.00548	0.11245	0.00744
$\|Z_{k,\lambda=0.40}^{(l'=2,3)}\|_2/\|c_0^{(k)}\|$	0.00593	0.12566	0.00876

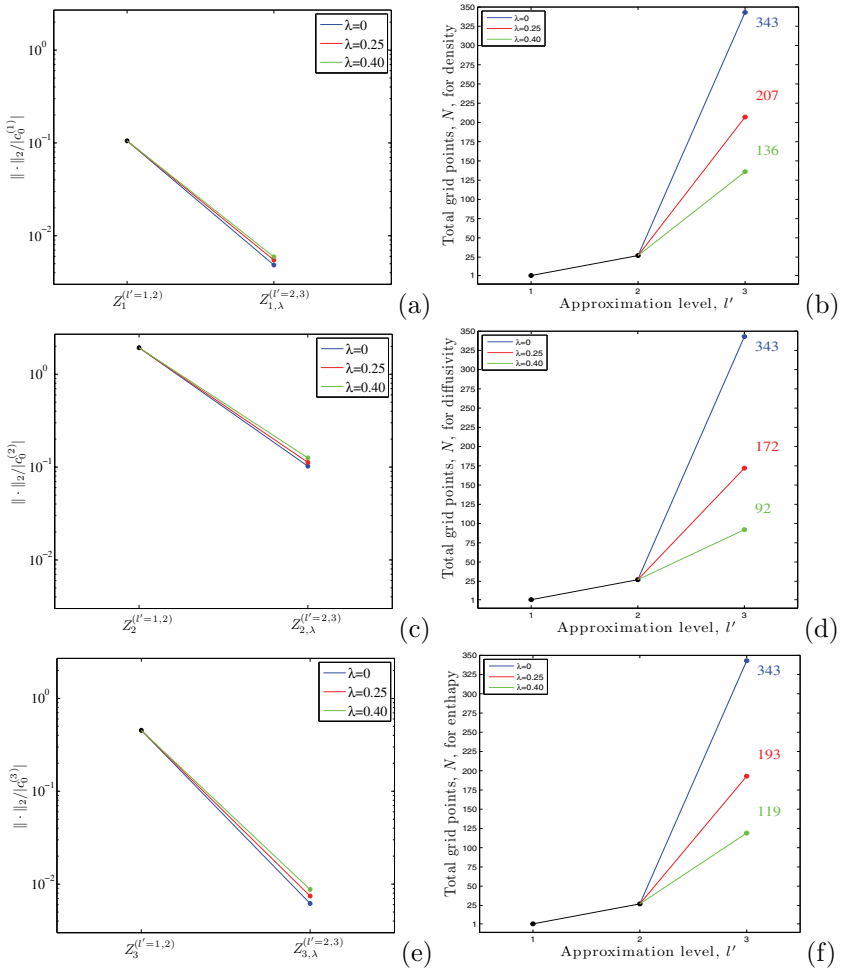


FIG. 4.14. *Left column: L_2-norms $\|Z_k^{(l'=1,2)}\|_2$ (see (4.6)) and $\|Z_{k,\lambda}^{(l'=2,3)}\|_2$ (see (4.12)) nondimensionalized by $|c_0^{(k)}|$ and obtained for* (a) *density,* (c) *self-diffusivity, and* (e) *enthalpy. Right column: total number of sampling points versus the approximation level, l', plotted for* (b) *density,* (d) *self-diffusivity, and* (f) *enthalpy. Curves are generated for $\lambda = 0$ (blue), 0.25 (red), and 0.40 (green). Color is distinguishable online only.*

5. Summary and conclusions. In this study, we focused on quantifying uncertainty in MD simulations of TIP4P water at ambient conditions, accounting for both intrinsic noise and parametric uncertainty.

Parametric uncertainty was introduced by assuming nondeterministic force-field parameters, namely by parametrizing them in terms of i.i.d. uniform RVs. For the present case, we chose as uncertain the two LJ parameters ε and σ and the distance, d, from the oxygen to the massless point where the negative charge is placed in the TIP4P model. The intrinsic (thermal) noise present in the atomistic system combines with the parametric uncertainty to yield nondeterministic, noisy MD predictions for the observables of water. Exploiting PC representations, we developed a framework suitable for isolating the impact of parametric uncertainty on MD predictions. We focused on three water observables, namely density, self-diffusion, and enthalpy, but the same approach can be readily extended to other physical and thermodynamical properties.

To build the PC expansions, we employed both an NISP approach and a Bayesian inference approach. We showed that for the present case, the effect of the intrinsic noise is weak, which yields a smooth dependence of the macroscale observables on the uncertain parameters. This allowed us to apply an NISP approach, in which the intrinsic noise is controlled both in terms of averaging over time and over realizations of the system, yielding well-defined, deterministic PC expansions of the observables of interest. The resulting spectra of PC coefficients for the observables of interest showed a rapid decay. As a second step in the forward propagation, we computed the PC expansions for the observables of interest using a Bayesian inference approach. To this end, we developed a new technique, based on Fejér nested grids, to sample the parameter space in order to collect the observations used for the inference. At a given grid approximation level l', we refined the density of sampling points only in the regions of the space where the differences between the PC expansions inferred at the preceding coarser levels exceed a user-specified threshold. We showed that the PC representations based on adapted grids are consistent with those obtained using the full grid, with the advantage of a substantial reduction in the computational cost. We provided a criterion for monitoring the refinement level of sampling grids and showed that, for the present work, the MD predictions for the water observables can be suitably represented using low-order PC models.

REFERENCES

[1] H. Adalsteinsson, B. J. Debusschere, K. R. Long, and H. N. Najm, *Components for atomistic-to-continuum multiscale modeling of flow in micro- and nanofluidic systems*, Sci. Program., 16 (2008), pp. 297–313.

[2] C. Andrieu and E. Moulines, *On the ergodicity properties of some adaptive MCMC algorithms*, Ann. Appl. Probab., 16 (2006), pp. 1462–1505.

[3] Y. F. Atchadé and J. S. Rosenthal, *On adaptive Markov chain Monte Carlo algorithms*, Bernoulli, 11 (2005), pp. 815–828.

[4] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *The missing term in effective pair potentials*, J. Phys. Chem., 91 (1987), pp. 6269–6271.

[5] J. F. Bourgat, P. Le Tallec, and M. Tidriri, *Coupling Boltzmann and Navier-Stokes equations by friction*, J. Comput. Phys., 127 (1996), pp. 227–245.

[6] K. A. Dill and S. Bromberg, *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*, Garland Science, New York, 2003.

[7] A. Donev, J. B. Bell, A. L. Garcia, and B. J. Alder, *A hybrid particle-continuum method for hydrodynamics of complex fluids*, Multiscale Model. Simul., 8 (2010), pp. 871–911.

[8] A. Einstein, *Investigations on the Theory of the Brownian Movement*, Dover, New York, 1956.

[9] L. Fejér, *On the infinite sequences arising in the theories of harmonic analysis, of interpola-*

tion, and of mechanical quadratures, Bull. Amer. Math. Soc., 39 (1933), pp. 521–534.
[10] R. G. Ghanem and A. Doostan, *On the construction and analysis of stochastic models: Characterization and propagation of the errors associated with limited data*, J. Comput. Phys., 217 (2006), pp. 63–81.
[11] R. Ghanem and P. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.
[12] H. Haario, E. Saksman, and J. Tamminen, *An adaptive Metropolis algorithm*, Bernoulli, 7 (2001), pp. 223–242.
[13] N. Hadjiconstantinou and A. T. Patera, *Heterogeneous atomistic-continuum representation for dense fluid systems*, Int. J. Mod. Phys., C8 (1997), pp. 967–976.
[14] N. G. Hadjiconstantinou, *Hybrid atomistic-continuum formulations and the moving contact-line problem*, J. Comput. Phys., 154 (1999), pp. 245–265.
[15] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*, Taylor & Francis, New York, 1989.
[16] W. G. Hoover, *Canonical dynamics: Equilibrium phase-space distributions*, Phys. Rev. A (3), 31 (1985), pp. 1695–1697.
[17] H. W. Horn, W. C. Swope, and J. W. Pitera, *Characterization of the TIP4P-Ew water model: Vapor pressure and boiling point*, J. Chem. Phys., 123 (2005), 194504.
[18] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. H. Gordon, *Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew*, J. Chem. Phys., 120 (2004), pp. 9665–9678.
[19] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *Comparison of simple potential functions for simulating liquid water*, J. Chem. Phys., 79 (1983), pp. 926–935.
[20] W. L. Jorgensen and J. D. Madura, *Temperature and size dependence for Monte Carlo simulations of TIP4P water*, Mol. Phys., 56 (1985), pp. 1381–1392.
[21] *LAMMPS (Large-Scale Atomic/Molecular Massively Parallel Simulator)*, http://lammps.sandia.gov.
[22] O. P. Le Maître and O. M. Knio, *Spectral Methods for Uncertainty Quantification*, Springer, New York, 2010.
[23] O. P. Le Maître, M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio, *A stochastic projection method for fluid flow.* II. *Random process*, J. Comput. Phys., 181 (2002), pp. 9–44.
[24] J. Li, D. Liao, and S. Yip, *Coupling continuum to molecular-dynamics simulation: Reflecting particle method and the field estimator*, Phys. Rev. E (3), 57 (1998), pp. 7259–7267.
[25] J. Liu, S. Chen, X. Nie, and M. O. Robbins, *A continuum-atomistic simulation of heat transfer in micro- and nano-flows*, J. Comput. Phys., 227 (2007), pp. 279–291.
[26] M. W. Mahoney and W. L. Jorgensen, *A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions*, J. Chem. Phys., 112 (2000), pp. 8910–8922.
[27] K. M. Mohamed and A. A. Mohamad, *A review of the development of hybrid atomistic-continuum methods for dense fluids*, Microfluid. Nanofluid., 8 (2010), pp. 283–302.
[28] X. B. Nie, S. Y. Chen, W. N. E, and M. O. Robbins, *A continuum and molecular dynamics hybrid method for micro- and nano-fluid flow*, J. Fluid Mech., 500 (2004), pp. 55–64.
[29] S. Nosé, *A molecular dynamics method for simulations in the canonical ensemble*, Mol. Phys., 52 (1984), pp. 255–268.
[30] S. Nosé, *A unified formulation of the constant temperature molecular dynamics methods*, J. Chem. Phys., 81 (1984), pp. 511–519.
[31] S. T. O'Connell and P. A. Thompson, *Molecular dynamics—continuum hybrid computations: A tool for studying complex fluid flows*, Phys. Rev. E (3), 52 (1995), pp. R5792–R5795.
[32] M. Parrinello and A. Rahman, *Crystal structure and pair potentials: A molecular-dynamics study*, Phys. Rev. Lett., 45 (1980), pp. 1196–1199.
[33] M. Parrinello and A. Rahman, *Polymorphic transitions in single crystals: A new molecular dynamics method*, J. Appl. Phys., 52 (1981), pp. 7182–7190.
[34] R. K. Pathria, *Statistical Mechanics*, Elsevier, Amsterdam, 1996.
[35] S. J. Plimpton, *Fast parallel algorithms for short-range molecular dynamics*, J. Comput. Phys., 117 (1995), pp. 1–19.
[36] M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio, *Uncertainty quantification in reacting flow simulations through non-intrusive spectral projection*, Combust. Flame, 132 (2003), pp. 545–555.
[37] J. R. Reimers, R. O. Watts, and M. L. Klein, *Intermolecular potential functions and the*

properties of water, Chem. Phys., 64 (1982), pp. 95–114.
[38] S. W. Rick, S. J. Stuart, and B. J. Berne, *Dynamical fluctuating charge force fields: Application to liquid water*, J. Chem. Phys., 101 (1994), pp. 6141–6156.
[39] G. O. Roberts and J. S. Rosenthal, *Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms*, J. Appl. Probab., 44 (2007), pp. 458–475.
[40] G. O. Roberts and J. S. Rosenthal, *Examples of adaptive MCMC*, J. Comput. Graph. Statist., 18 (2009), pp. 349–367.
[41] R. E. Rudd and J. Q. Broughton, *Concurrent coupling of length scales in solid state systems*, Phys. Status Solidi (B), 217 (2000), pp. 251–291.
[42] M. Salloum, K. Sargsyan, R. Jones, B. Debusschere, H. N. Najm, and H. Adalsteinsson, *A stochastic multiscale coupling scheme to account for sampling noise in atomistic-to-continuum simulations*, Multiscale Model. Simul., 10 (2012), pp. 550–584.
[43] K. Sargsyan, B. Debusschere, H. N. Najm, and Y. Marzouk, *Bayesian inference of spectral expansions for predictability assessment in stochastic reaction networks*, J. Comput. Theor. Nanosci., 6 (2009), pp. 2283–2297.
[44] D. S. Sivia, *Data Analysis: A Bayesian Tutorial*, Oxford Sci. Publ., Oxford University Press, Oxford, UK, 2006.
[45] S. Tiwari, *Coupling of the Boltzmann and Euler equations with automatic domain decomposition*, J. Comput. Phys., 144 (1998), pp. 710–726.
[46] M. Von Smoluchowski, *Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen*, Ann. der Physik, 326 (1906), pp. 756–780.
[47] N. Wiener, *The homogeneous chaos*, Amer. J. Math., 60 (1938), pp. 897–936.
[48] H. S. Wijesinghe, R. D. Hornung, A. L. Garcia, and N. G. Hadjiconstantinou, *Three-dimensional hybrid continuum-atomistic simulations for multiscale hydrodynamics*, J. Fluid Eng., 126 (2004), pp. 768–777.
[49] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.