International Journal for Uncertainty Quantification, 7(2): 99–115 (2017)

CALIBRATION AND RANKING OF COARSE-GRAINED MODELS IN MOLECULAR SIMULATIONS USING BAYESIAN FORMALISM

Hadi Meidani,[1,*] Justin B. Hooper,[2] Dmitry Bedrov,[2] & Robert M. Kirby[3]

[1]*Department of Civil and Environmental Engineering, 1211 Newmark Civil Engineering Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61822, USA*

[2]*Department of Materials Science & Engineering, 206 Civil and Materials Engineering Building, University of Utah, Salt Lake City, UT 84112, USA*

[3]*School of Computing, 3750 Warnock Engineering Building, University of Utah, Salt Lake City, UT 84112, USA*

*Address all correspondence to: Hadi Meidani, Department of Civil and Environmental Engineering, 1211 Newmark Civil Engineering Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61822, USA, E-mail: meidani@illinois.edu

Original Manuscript Submitted: 2/6/2015; Final Draft Received: 1/6/2017

Understanding and prediction of the performance of complex materials using molecular simulations, as well as the design of a new generation of materials with desired functionality, depend on the predictive capacity of the coarse-grained models that enable reduced computational cost. Depending on what aspect of the behavior is of interest, we often have at our disposal various coarse-grained models with different predictive capabilities. In this work, focusing on coarse-grained water models, we demonstrate how the plausibilities of these models are relatively compared, and how predictions can be made exploiting the ensemble. Using a Bayesian model ranking framework, we will show the plausibility results which are in agreement with experts' expectation on how these models rank in terms of predicting different quantities of interest, and how different models can be mixed and produce an ensemble prediction with higher accuracy.

KEY WORDS: *polynomial chaos, parameter estimation, Bayesian inference, stochastic sensitivity analysis*

1. INTRODUCTION

With the recent advancement in computational power, new methods are being developed to simulate the behavior of physical and chemical processes at larger scales. Coarse-graining is one such method which reduces the degrees of freedom in a computational model and thus enables molecular simulation of materials and phenomena at longer length and time scales. In this approach, it is critical to accurately parameterize the resulting coarse-grained models so that they accurately represent the phenomena of interest. Among the available methods in the literature, Boltzmann inversion [1], reverse Monte Carlo (MC) [2], force-matching [3], and relative entropy minimization [4] are "bottom-up" coarse-graining methods, where inference about the "coarse" interaction potentials are carried out using fine-scale (atomistic) information. These methods typically impose a massive computational burden. The alternative is to pursue a "top-down" approach, where the parameters of the selected coarse-grained interaction potential are the ones that produce coarse-scale output quantities that are closest to the experimentally calculated values.

Independent of which approach in initial parametrization of coarse-grained (CG) model parameters is used, a significant challenge is how to efficiently optimize and infer the sensitivity of desired observables to model parameters.

2152–5080/17/$35.00

These models are prone to uncertainties, of which one type is the parametric uncertainty that results from not knowing the "true" input parameters of the interaction potential that best represent the properties of interest. Uncertainty quantification (UQ) techniques can be used to efficiently assess the impact of the parametric uncertainties on system properties. This task has traditionally been done using MC sampling and its derivatives, which impose significant computational burden, especially in large systems. Numerous techniques have been proposed to improve MC approaches by enhanced sampling recipes [5,6]. In spite of these developments, there is still a need to increase the tractability of these stochastic approaches, particularly for computationally expensive techniques such as molecular simulations where the simulation for one sample point in the parameter space can take a substantial amount of time and resources.

Among the alternatives to MC simulation, the polynomial chaos (PC) methodology consists of spectral approximation of random variables using a series expansion with orthogonal polynomial basis functions [7,8]. Both the model inputs and quantities of interest (QoIs) are represented using truncated expansions in the same vector space, effectively resulting in a finite-dimensional approximation, which leads to significant computational savings. In particular, statistical moments and sensitivity indices can be readily available [7,9]. The calculation of the QoI expansion can be done using the intrusive Galerkin projection [7] or nonintrusive spectral projection [10]. The latter requires drawing samples from the input parameters according to either brute-force MC approach, full tensor product construction based on one-dimensional samples, or the sparse grid construction [10]. PC expansions have been also used to accelerate Bayesian inference [11,12], where the forward model is replaced by a PC surrogate which is analytically evaluated at a substantially lower cost for each Markov chain Monte Carlo (MCMC) sample.

UQ with the discussed polynomial approximation is a powerful tool to evaluate the predictive capacity of atomistic and coarse-grained models. Rizzi et al. have used UQ to quantify the impact of parametric uncertainties on the density, enthalpy, and diffusivity of the TIP4P water model [13]. Rizzi et al. in [14,15] investigated the significance of the intrinsic noise, relative to the parametric uncertainty, on the flux of ions through silica nanopores, where they showed that the thermal noise present in molecular dynamics simulations is relatively negligible. Their findings allow for use of a polynomial approximation for the macroscopic observables in terms of the model parameters. We made use of this finding in using PC surrogates in the Bayesian framework that we will describe later.

To demonstrate the application of Bayesian model ranking and averaging, we focus on coarse-grained models of water. Several coarse-graining paradigms have been proposed for water models, the most popular of which remains to be the standard rigid atomistic water models with electrostatic interactions. This representation can be considered a coarse-grained representation of more detailed flexible or quantum mechanical models of water. The next level of coarse-graining contains models that preserve the all-atom structure of water with modified interactions at reduced length scale [16,17]. Models that treat the water molecule as a single particle represent the next level of coarse-graining [18]. In a recent work [19], the authors used UQ machinery to assess the sensitivity of observables to the parameters of an interaction potential and deterministically calibrated the parameters using an interpolation-based response surface, in a "top-down" approach (see Fig. 1). As QoI, a number of water properties such as density, enthalpy of vaporization, and radial distribution function (RDF) at 300 K were considered.

In this paper, we consider a family of water models. For each model, we treat the parameters to be uncertain and using a surrogate-based Bayesian inference framework find the optimal probability density function for each parameter (see Fig. 2). In Section 2, we describe how the models are formed and how, based on their structural difference we expect different predictive performances from them. Finally, in Section 3, technical background for the Bayesian calibration and surrogate-based Bayesian calibration is discussed, while in Section 4, the basics of Bayesian model ranking and model averaging are included. In Section 5, we demonstrate the Bayesian calibration of one of the three water models in detail and then discuss the Bayesian model ranking results for the three-model ensemble. We highlight the additional insights that this statistical inversion, compared to a deterministic calibration approach, grants on the predictive capacity of different interaction potentials. We will also include the results from an approximate sensitivity analysis, which shed light on how certain observables are more informative for the calibration of certain parameters. Specifically, if the resulting probability distribution from the calibration procedure is not much different from our prior expectation, it means that the particular observed quantity used in the calibration process was not informative.

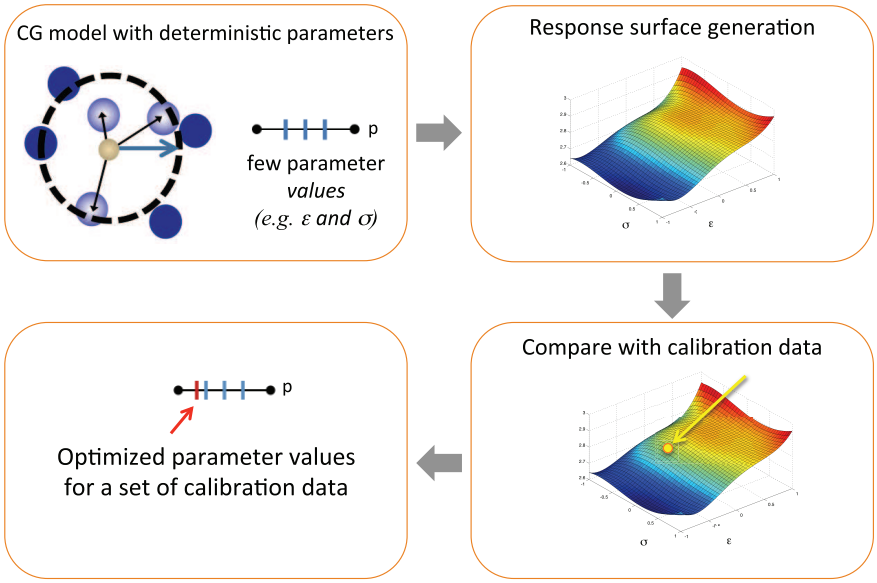


FIG. 1: The deterministic inverse problem framework to calibrate the parameters of the coarse-grained model with data from experiments, or from fully atomistic simulations

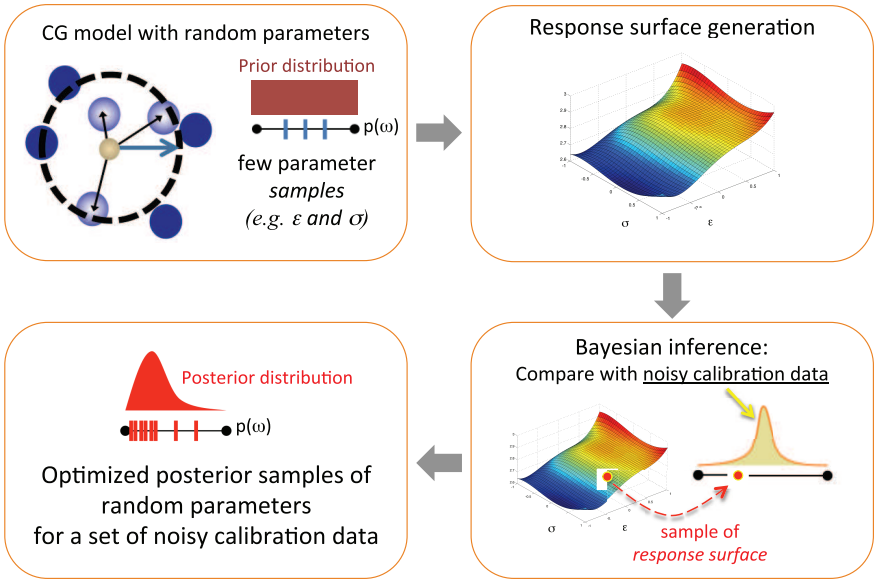


FIG. 2: The statistical inverse problem framework to calibrate the probability model of the random parameters of the coarse-grained model with data from noisy measurement, or from erroneous fully atomistic simulations

2. COARSE-GRAINED MODELS FOR WATER

For this work, we have employed three potentials for water, representing three differing degrees of coarse-graining. In order of complexity; they are the TIP4P/2005 [20], which is an atomically detailed potential, the mW water model [21] which reduces the atomic detail to a single site model while employing an anisotropic potential to regain some degree of local spatial organization; and the isotropic model of Chaimovitz and Shell [22], derived by minimizing the relative entropy of the coarse-grained model to yield an optimal spherically symmetric potential. The most detailed of these

potentials is the TIP4P/2005 model, which has been parameterized to accurately represent the entire phase diagram of condensed water. In this model, each water molecule is comprised of a rigidly bound set of four force centers: two representing hydrogen, one oxygen, and one offset site representing a centralized location for the negative charge carried by oxygen while being parameterized to best represent the overall electrostatic potential of the molecule as a whole. The overall potential between sites in the system can be generalized as

$$U\left(\boldsymbol{r}_{ij}\right) = \frac{A}{\left|\boldsymbol{r}_{ij}\right|^{12}} - \frac{B}{\left|\boldsymbol{r}_{ij}\right|^{6}} + \sum_{\boldsymbol{n}'} \frac{q_i q_j}{\left|\boldsymbol{r}_{ij} + \boldsymbol{n}\right|}, \tag{1}$$

where $\boldsymbol{r}_{ij}$ is the vector originating at site i and terminating at site j, $\boldsymbol{n}$ is the set of all translation vectors comprised of equal integral offsets of the original simulation unit cell along one or more cell axes, and the prime on the $\boldsymbol{n}'$ in the summation indicates that the interaction for $i = j$ with $\boldsymbol{n} = 0$ is omitted. The oxygen atoms of each water molecule interact solely through the first two terms, representing the local repulsion/dispersion (R/D) interaction between the oxygen molecules and setting a local length scale, while the hydrogen atoms as well as the offset charge of the fourth force center interact solely through the third term, which represents the long-range electrostatic interactions between molecules. The structure and thermodynamics of the TIP4P/2005 model are therefore a complex function of both these length scales as well as the energy scales imposed by the R/D and electrostatic interactions. Similarly, local organization is controlled by the complex interplay between the energy length scales and the inherent symmetry of the rigidly bonded force centers, ultimately yielding the tetrahedral organization of associating water molecules which is fundamental in the unique thermodynamic character of liquid water.

Instead of the explicit geometric arrangement of atoms in the TIP4P/2005 model, the mW water model reduces the entire molecule to a single site, which interacts with nearby sites via a modified Stillinger-Weber potential [23]:

$$U\left(\boldsymbol{r}_{ij}\right) = \phi_2\left(\boldsymbol{r}_{ij}\right) + \sum_{k} \phi_3\left(\boldsymbol{r}_{ij}, \boldsymbol{r}_{ik}, \theta_{ijk}\right), \tag{2}$$

$$\phi_2\left(\boldsymbol{r}_{ij}\right) = A\varepsilon \left[B\left(\frac{\sigma}{\left|\boldsymbol{r}_{ij}\right|}\right)^{p} - \left(\frac{\sigma}{\left|\boldsymbol{r}_{ij}\right|}\right)^{q}\right] \exp\left(\frac{\sigma}{\left|\boldsymbol{r}_{ij}\right| - a\sigma}\right), \tag{3}$$

$$\phi_3\left(\boldsymbol{r}_{ij}, \boldsymbol{r}_{ik}, \theta_{ijk}\right) = \lambda\varepsilon\left[\cos\theta_{ijk} - \cos\theta_0\right] \exp\left(\frac{\gamma\sigma}{\left|\boldsymbol{r}_{ij}\right| - a\sigma}\right) \exp\left(\frac{\gamma\sigma}{\left|\boldsymbol{r}_{ik}\right| - a\sigma}\right), \tag{4}$$

with $p = 4$; $q = 0$; $A = 7.049556277$; $B = 0.6022245584$; $\gamma = 1.2$; $\theta_0 = 109.47°$; $a = 1.8$; and σ, ε, and λ are utilized as the fitting parameters. For this interaction, a single length scale is present, implicitly controlled by the values of σ and ε, while the extent of local organization is controlled by λ which sets the relative magnitude of the anisotropic contribution to the overall potential. The net effect of going from a model such as TIP4P/2005 to the mW water model is that of trading off the explicit molecular geometry and long-range interactions (and the emergent structuring and order which they beget) for a shorter-ranged, explicitly ordered potential which is more approximate for intermediate- and long-range orders while also being significantly faster to calculate. Thus, the mW water model effectively coarse-grains the long-range structure of a more explicit model in exchange for faster calculation, under the assumption that most of the thermodynamic and structural characteristics of water arise from nearest-neighbor interactions. Previous work has shown that this coarse-grained approximation does quite well at reproducing various state functions of water [21]. Figure 3 illustrates the difference between the atomistically detailed TIP4P and coarser-grained mW water models.

Finally, the isotropic spherical model of [22] has been investigated as a highly coarse-grained model. Here, the water molecule is again reduced to a single site representation. However, unlike the mW water model which preserves the local tetrahedral structure of water via anisotropic interaction potentials, their proposed potential is isotropic, with the effect of local structural correlations accounted for in a mean-field manner by the presence of a so-called "core-softened" potential which has two overlapping length scales designed to mimic the primary and secondary organization of liquid water correlation shells around a specific water molecule. This potential generates the two requisite length scales by fitting to a potential form comprised of combined Lennard-Jones and Gaussian terms, yielding the Lennard-Jones plus Gaussian (LJG) pair potential [22]:

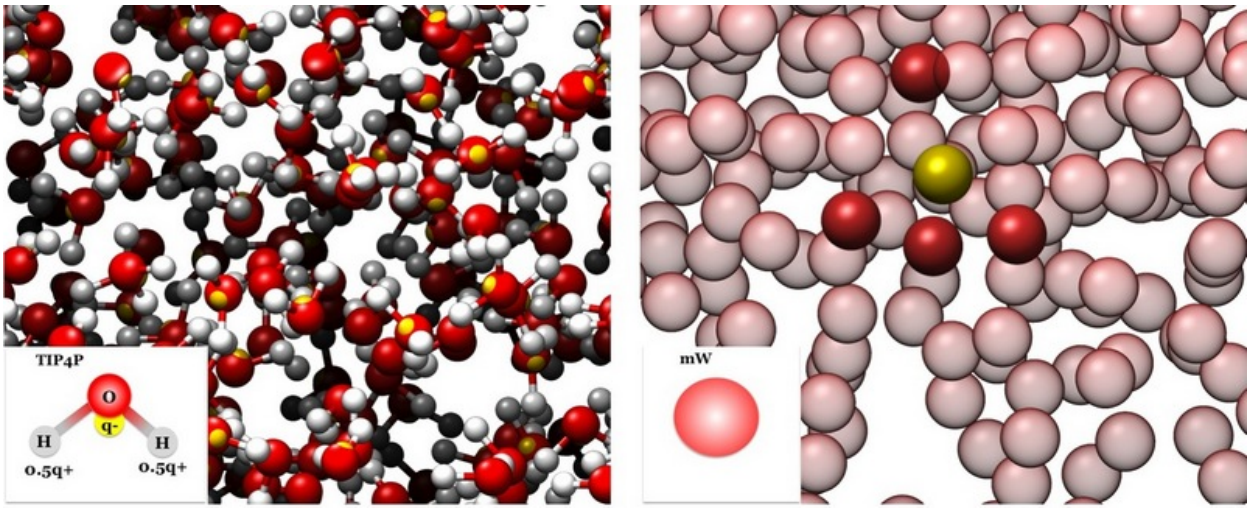


FIG. 3: Schematic illustration of the TIP4P and mW water models and representative snapshots of molecular configurations of bulk water produced by these models in molecular dynamics simulations

$$U_{\mathrm{LJG}}\left(\boldsymbol{r}_{ij}\right) = 4\varepsilon\left[\left(\frac{\sigma}{|\boldsymbol{r}_{ij}|}\right)^{12} - \left(\frac{\sigma}{|\boldsymbol{r}_{ij}|}\right)^{6}\right] + B^{*}\varepsilon\exp\left[-\left(\frac{|\boldsymbol{r}_{ij}|/\sigma - r_0^{*}}{\Delta^{*}}\right)^{2}\right], \tag{5}$$

where σ and ε have their standard meaning of representing the length scale and energy scale of the LJ portion of the interaction, while B^*, r_0^*, and Δ^* represent the relative strength, center, and spread of the Gaussian well. In order to fit this function, Chaimovich and Shell minimize the relative entropy of the functional form with respect to the canonical ensemble configurational probability distribution of an explicit, all-atom model of water. This technique is somewhat complicated (specifically requiring calculation of the free energy of the system which can be a significant effort, depending upon the system). However, the upside of this technique is that the entropy minimization is equivalent to saying that the coarse-grained model is required to reproduce the underlying probability distribution of the ensemble configurations of the explicit, atomistic model as closely as is possible for the reduced degrees of freedom present within the CG model.

For the purposes of this work, the implication is that the ultimate fitting of the CG model to the overall data against which it is compared to other models is strictly a function of the quality of the model, and should contain no additional error due to the quality of the fitting of the CG model to atomistic data. It should be noted that the underlying all-atom model utilized by [22] is different than our atomistic model, which makes it difficult to deconvolute the exact extent of the imprecision injected by the process of coarse-graining the model compared to that introduced by the difference between the different atomistic models. The latter should be minimal, since the SPC/E atomistic model upon which the LJG model is parameterized is generally regarded as an accurate water model.

3. BAYESIAN CALIBRATION

Model calibration is the task of parameter estimation using observation from a real system. Typically, an inverse problem is constructed with the aim of finding the parameter values that produce model outputs that most closely match the observation data. In general, the parameter could be treated as random or deterministic quantities, and the observation could be considered to be noisy or noise-free. Also, the forward model, i.e., the computer simulation that calculates the outputs given actual parameter values, can be subject to errors. Thus, a modeler, on one hand, could be dealing with a fully deterministic case, where observation, model parameters, and model predictions are assumed to be error- and noise-free and therefore considered as deterministic quantities. On the other hand, the modeler may be dealing with cases where at least one of these quantities is random. The water models described in Section 2, as well as other coarse-grained atomistic modelsm are approximations of the true interactions in these systems. Therefore,

it would be logical to consider, instead of a model with deterministic parameters, a stochastic model with multiple possible parameter values. Such model will then lead to predictions which account for multiple potential outcomes with calculated occurrence probabilities. Thus, using these results, one moves away from drawing conclusions based on a potentially invalid model to considering a range of possibilities, and thus increasing the chance of capturing the true behavior. This, of course, comes at the expense of increased computational cost; a potentially prohibitive increase in many systems.

A statistical inverse problem is one in which the model parameters and observation data are considered to be random quantities. The characterization of the randomness in the observation data is given in this problem, often in the form of the probability density function of the observational noise, or also in the form of summary statistics of the data. The objective is then to characterize the randomness in the model parameters using the UQ-enabled solution of the inverse problem.

In order to formulate the statistical inverse problem, the Bayes theorem can be used. Let $\boldsymbol{p} = [p_1, \cdots, p_N] \in \mathbb{R}^N$ denote the vector of N random parameters in the model of interest, and let $\boldsymbol{d} = [d_1, \cdots, d_D] \in \mathbb{R}^D$ denote D observation data, of a single or multiple QoIs. For the sake of notation brevity, we use $\pi(\cdot)$ to denote a generic probability density function. The Bayes theorem, then, states

$$\pi(\boldsymbol{p}|\boldsymbol{d}) = \frac{\pi(\boldsymbol{d}|\boldsymbol{p})\pi(\boldsymbol{p})}{\int \pi(\boldsymbol{d}|\boldsymbol{p})\pi(\boldsymbol{p})\mathrm{d}\boldsymbol{p}}, \tag{6}$$

where $\pi(\boldsymbol{p}|\boldsymbol{d})$ is called the posterior density of parameter p; $\pi(\boldsymbol{p})$ is the prior density of p, prescribed based on the expert or prior knowledge about the parameter; and $\pi(\boldsymbol{d}|\boldsymbol{p})$ is called the likelihood function, which determines the likelihood that the model given a fixed parameter value $\boldsymbol{p}$ produces the output with the value equal to the observation $\boldsymbol{d}$. This likelihood is calculated based on the following general map from parameters $\boldsymbol{p}$ to observation data $\boldsymbol{d}$:

$$\boldsymbol{d} = \boldsymbol{f}(\boldsymbol{p}) + \epsilon_{obs} + \epsilon_{num} + \epsilon_{mod}, \tag{7}$$

where $f : \mathbb{R}^N \to \mathbb{R}^D$ represents the forward model, i.e., the numerical model that produces the QoIs. ϵ_{mod} denotes the error in our understanding and modeling of the actual physical system, ϵ_{num} is the error incurred by the numerical implementation of the proposed models, and ϵ_{obs} is the discrepancy between the model prediction and the observed quantity due to noisy measurement or experimental procedure. In this paper, similar to the majority of Bayesian approaches, we only consider the measurement noise ϵ_{obs} and assume the other two error terms are zero, and henceforth denote it by ϵ. Let us assume that each of the D observation data refers to a distinct quantity of interest, measured with a noise ϵ_i with the given density function π_{ϵ_i}. The likelihood function can then be written as

$$L(\boldsymbol{p}) = \pi(\boldsymbol{d}|\boldsymbol{p}) = \prod_{i=1}^{D} \pi_{\epsilon_i}(d_i - f_i(\boldsymbol{p})). \tag{8}$$

In statistical inverse problems, the objective is to find the posterior density, and there is no direct interest, in general, in computing the integral in the denominator, referred to as the evidence. As a matter of practice, the standard algorithms used to solve these problems are formed based on a proportionality relation which does not involve the evidence, and is as follows:

$$\pi(\boldsymbol{p}|\boldsymbol{d}) \propto \pi(\boldsymbol{d}|\boldsymbol{p})\pi(\boldsymbol{p}). \tag{9}$$

In most of the application areas, due to the nonlinearities in the forward model, the resulting complex form of the likelihood function does not allow for the analytical solution of the posterior probability density function. Instead, algorithms such as MCMC simulation aim to generate samples from this posterior density function. To be able to reliably infer about the uncertain parameters, a sufficiently large number of these samples should be obtained, meaning that the likelihood function, and therefore the forward model $f(\boldsymbol{p})$, should be evaluated for a very large number of parameter samples.

It consists of drawing samples based on the product of the prior distribution and likelihood function. Thus, in each step of the algorithm, the value of the likelihood function should be evaluated at a new parameter value. As

mentioned earlier, in order to calculate this likelihood function, one needs to run a computer simulation at the new parameter value. In the case of molecular dynamics or other computationally expensive computer models, several calls to the computer code at each step of the MCMC algorithm impose a significant computational burden, making it infeasible to use the MCMC algorithm for large, complex systems. To alleviate the computational cost associated with the many forward simulations, probabilistic response surfaces have been used as surrogates [11,24,25]. A response surface efficiently relates the model parameters to their corresponding random output. In the most ideal case, such a surface is formed by drawing a sufficiently large number of points in the parameter space, and calculating their associated model outputs. In atomistic systems, for instance, the simulation for each point can take hours or days, depending on thermodynamic conditions. Thus, efforts have aimed at the characterization of an approximate response surface given only a few points. This surface will then serve as a proxy to replace the forward numerical model and thus alleviate the tremendous number of computer simulations. At each step of the MCMC simulation, then, the QoIs are obtained analytically from the approximate surface at the given parameter sample.

In particular, in the Bayesian calibration, we make use of the PC expansion to form the surrogate, an idea first proposed in [11]. The PC expansion consists in representing random variables using an orthogonal series expansion based on polynomial basis functions [7,8]. It has been shown that such representation can significantly reduce the cost associated with uncertainty propagation computation [10,26,27]. The surrogate-based Bayes rule will then be based on the following approximate likelihood: function

$$\hat{L}(\boldsymbol{p}) = \hat{\pi}(\boldsymbol{d}|\boldsymbol{p}) = \prod_{i=1}^{D} \pi_{\epsilon_i}(d_i - \hat{f}_i(\boldsymbol{p})), \tag{10}$$

where $\hat{f}_i(\boldsymbol{p})$ is an approximation of the exact forward model obtained based on the following Pth-order order expansion:

$$\hat{f}_i(\boldsymbol{p}) = \sum_{i=0}^{P} c_i \Psi_i(\boldsymbol{p}), \tag{11}$$

where $\{\Psi_i(\boldsymbol{p})\}$ are the set of polynomial functions that are orthogonal with respect to the probability density function of $\boldsymbol{p}$ [8]. The deterministic coefficients $\{c_i\}$ fully characterize the response surface $\hat{f}_i(\boldsymbol{p})$. These coefficients can be calculated using the stochastic collocation technique where quadrature points are selected based on the density of the input parameters $\boldsymbol{p}$ [10]. In this work, the parameters are assumed to be uniformly distributed, so we select Gauss-Legendre quadrature points. Specifically, in a one-dimensional case with $p \in [-1, 1]$, let $[p^1, p^2, \cdots, p^Q]$ be the Q quadrature points with their corresponding weights $[w^1, w^2, \cdots, w^Q]$. The deterministic coefficients $\{c_i\}$ can then be calculated according to

$$c_i \approx \sum_{q=1}^{Q} \hat{f}_i(p^q)\Psi_i(p^q)w^q. \tag{12}$$

For a multidimensional parameter space with $\boldsymbol{p} \in [-1, 1]^N$, we can use the full tensor product of the collocation points in each dimension to form the multidimensional grid of points, or make use of the Smolyak sparse grid cubature, which exploits a sparse tensor product and is computationally superior, especially in high-dimensional problems [10,28].

It has been shown that if the L^2 distance between the exact forward model and the one approximated by the PC surrogate goes to zero, the Kullback-Leibler distance, which is a distance between between two probability density functions, goes to zero as well [12]. Therefore, it is possible to calculate the posterior density function and statistical moments with guaranteed satisfactory accuracy. The L^2 convergence between the exact forward model and the surrogate can be controlled by the order of the PC expansion and the number of quadrature points.

4. BAYESIAN MODEL RANKING AND AVERAGING

In many applications, including computational chemistry, more than one predictive model is available for the same system. These could be models that are built at different scales, or they could be variations of the same model given

different parameter values or different initial or boundary conditions. Often, the statistical average of the outputs of the ensemble of competing models is more predictive than any of the individual outputs. Bayesian model averaging provides a way to combine the component predictions by each ensemble member. It does so by weighing each prediction according to the plausibility of its associated model. Similar to the task of model averaging, model ranking or selection has also gained attention. The objective in model ranking is to give quantitative scores to competing models based on their predictability with respect to one or multiple observables.

Let us consider a family of K predictive models $\{M_1, \cdots, M_K\}$. Let $\boldsymbol{p}_k$ denote the parameters of model k whose prior distribution is denoted by $\pi(\boldsymbol{p}_k|M_k)$. To compare two statistical models based on the way the evidence provided by data d favors each model, the Bayes factor [29,30] can be used as a criterion defined by

$$B_{k,j} = \frac{\pi(\boldsymbol{d}|M_k)}{\pi(\boldsymbol{d}|M_j)}, \tag{13}$$

where $\pi(\boldsymbol{d}|M_k)$ is referred to as the marginal likelihood of model M_k and can be calculated according to

$$\pi(\boldsymbol{d}|M_k) = \int \pi(\boldsymbol{d}|\boldsymbol{p}_k, M_k)\pi(\boldsymbol{p}_k|M_k)\mathrm{d}\boldsymbol{p}_k. \tag{14}$$

We will discuss the numerical evaluation of this integral later in this section. Bayes factors, by construction, do not depend on prior model probabilities, i.e., any *a priori* perception of the plausibility of each model. In order to incorporate this extra component, we can define the posterior model probability of model M_k as follows:

$$\pi(M_k|\boldsymbol{d}) = \frac{\pi(\boldsymbol{d}|M_k)\pi(M_k)}{\sum_{i=1}^{K}\pi(\boldsymbol{d}|M_i)\pi(M_i)}. \tag{15}$$

A different criterion that includes prior model probabilities can then be defined as the ratio between two posterior model probabilities as follows:

$$\frac{\pi(M_1|\boldsymbol{d})}{\pi(M_2|\boldsymbol{d})} = \frac{\pi(\boldsymbol{d}|M_1)}{\pi(\boldsymbol{d}|M_2)}\frac{\pi(M_1)}{\pi(M_2)}. \tag{16}$$

This criterion is referred to as the posterior odds.

Using the definition of posterior model probabilities, we can also make a prediction for the quantity of interest y based on an ensemble of models using Bayesian model averaging as follows:

$$\pi(y|\boldsymbol{d}) = \sum_{k=1}^{K}\pi(y|M_k, \boldsymbol{d})\pi(M_k|\boldsymbol{d}), \tag{17}$$

where $\boldsymbol{d}$ is the available observation. This equation is essentially a weighted sum of the individual predictions $\{\pi(y|M_k, \boldsymbol{d})\}$ with the weights equal to the posterior model probabilities. Statistical moments of the y can also be calculated using the moments given by individual models. Let $\{\bar{y}_1, \cdots, \bar{y}_K\}$ and $\{\mathrm{Var}(y)_1, \cdots, \mathrm{Var}(y)_K\}$ denote the first and second statistical moments of the QoIs predicted by the K models, i.e.,

$$\begin{aligned} \bar{y}_k &= \mathbb{E}(y|M_k, \boldsymbol{d}), \\ \mathrm{Var}(y)_k &= \mathrm{Var}(y|M_k, \boldsymbol{d}). \end{aligned} \tag{18}$$

Then, the ensemble prediction of these statistical moments can be calculated according to

$$\begin{aligned} \bar{y}_{ens} &= \sum_{k=1}^{K}\bar{y}_k\pi(M_k|\boldsymbol{d}), \\ \mathrm{Var}(y)_{ens} &= \left[\sum_{k=1}^{K}(\mathrm{Var}(y)_k + \bar{y}_k^2)\pi(M_k|\boldsymbol{d})\right] - \bar{y}_{ens}^2. \end{aligned} \tag{19}$$

To evaluate the posterior model probabilities, we need to evaluate the marginal likelihood defined in Eq. (14). In most engineering applications, this integral cannot be evaluated analytically. To numerically evaluate this integral, the most simple way is to use N_{MC} MC samples from the prior distribution, i.e.,

$$\pi(\boldsymbol{d}|M_k) \approx \tilde{\pi}_1(\boldsymbol{d}) = \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} \pi(\boldsymbol{d}|\boldsymbol{p}_k^{(i)}, M_k); \;\; \boldsymbol{p}_k^{(i)} \sim \pi(\boldsymbol{p}_k|M_k). \tag{20}$$

To accelerate this MC integration, which is particularly critical in large models, importance sampling, Laplace approximation [31,32], and thermodynamic sampling [33,34] can be used. We used the calculation based on the importance sampling, where the integral is approximated by

$$\pi(\boldsymbol{d}|M_k) \approx \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} w^{(i)} \pi(\boldsymbol{d}|\boldsymbol{p}_k^{(i)}, M_k); \;\; \boldsymbol{p}_k^{(i)} \sim \pi^*(\boldsymbol{p}), \tag{21}$$

where $\pi^*(\boldsymbol{p})$ is the importance sampling function and $w^{(i)} = (\pi(\boldsymbol{p}^{(i)}))/(\pi^*(\boldsymbol{p}^{(i)}))$, where drawing samples from the density function $\pi^*(\boldsymbol{p})$ accelerates the convergence of the sample mean estimator. The posterior harmonic mean estimator was proposed by [35], where $\pi^*(\boldsymbol{p})$ was chosen to be the posterior density function, the integral in Eq. (14) is approximated using posterior parameter samples that have higher likelihood than their prior counterparts. The estimator is then given by

$$\pi(\boldsymbol{d}|M_k) \approx \tilde{\pi}_2(\boldsymbol{d}) = \left(\frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} \frac{1}{\pi(\boldsymbol{d}|\boldsymbol{p}_k^{*(i)}, M_k)} \right)^{-1} ; \;\; \boldsymbol{p}_k^{*(i)} \sim \pi(\boldsymbol{p}_k|d, M_k). \tag{22}$$

The estimator $\tilde{\pi}_2(\boldsymbol{d})$ converges almost surely to the true value of $Pr(\boldsymbol{d}|M_k)$. However, in its current form, it does not satisfy a Gaussian central limit theorem, in general, and samples with small likelihood will result in instabilities and observing large variance on the numerical integration results is possible. We observed these instabilities in our simulations, and used prior samples [i.e., Eq. (20)].

5. NUMERICAL RESULTS

In this section, we demonstrate the Bayesian model calibration and ranking techniques discussed in previous sections on three water models: the TIP4P/2005, mW, and LJG water models. Through numerical results, we highlight the unique capabilities that these Bayesian techniques present. The set of observables (water properties) considered in this study includes water density ρ, enthalpy of vaporization ΔH, and radial distribution function (RDF) $g_{ij}(r)$. These quantities are calculated based on the simulation results according to

$$\begin{aligned} \rho &= \frac{NM}{N_A \langle V \rangle}, \\ \Delta H &= 2.5RT - \langle E + pV \rangle, \\ g_{ij}(r) &= \frac{V}{N^2} \left\langle \sum_i \sum_{j \neq i} \delta(r - r_{ij}) \right\rangle, \end{aligned} \tag{23}$$

where N_A is the Avogadro constant; N is the number of molecules with the molecular mass of water, M, equal to 18.015 g/mol. V, p, and E are the volume, pressure, and total internal energy of the system per mol, R is the gas constant, and T is the temperature. r_{ij} denotes the distance between molecules i and j and $\delta(\cdot)$ is the Dirac delta function. We chose the bulk properties of water at 300 K and 1 atm as the QoIs. The experimental values for these properties are $\rho = 10.52$ g/ml, $\Delta H = 0.997$ kcal/mol. Also, the empirical RDF reported in [36] was used as the reference.

The simulation systems for the mW and TIP4P/2005 models consisted of a periodic cell with 4096 and 2048 water molecules, respectively. Both simulations were performed in an isothermal, isobaric ensemble, where mW simulations had a time step of 10 fs and total simulation time of 2 ns, and TIP4P/2006 simulations were performed for 0.5 ns with a time step of 0.5 fs. The LJG potential was used in an NVT simulation with the time step of 10 fs and total simulation time of 10 ns. The estimates of water density were irrelevant in the case of LJG, given the fixed volume and number of atoms. The uncertain parameters in all the models were considered to be uniformly distributed within a range which will be described in the next sections.

5.1 Bayesian Calibration of mW Models

In a previous study [19], the authors used a UQ approach to estimate the deterministic parameters of mW models. Specifically, they found the values for three parameters, ϵ, σ, and λ, which best predicted the observed values for thermodynamic and structural properties of water. They used the stochastic collocation technique to generate a response surface which facilitated the parameter estimation. The generated response surface replaced the molecular dynamics simulator, leading to significant computational savings.

The distinction in the current work is the treatment of the parameters as random quantities. In the previous work, the optimal parameter value was determined within a given range, whereas in the current work we determine the best probability density function over a prescribed range. We chose the three ranges to be centered around the nominal values of $\epsilon = 6.1890$ kcal/mol, $\sigma = 2.3925$ Å, and $\lambda = 23.1500$, with widths equal to 40% of the associated nominal values. Given such probability density, one could subsequently calculate point estimates, i.e., a single optimal value for each parameter. These point estimates, such as the mean or maximum a posteriori (MAP) estimate, are not necessarily the same as the deterministically calibrated parameter values. To demonstrate this fact, Fig. 4 depicts the results obtained from deterministic and Bayesian (statistical) calibrations. The Bayesian calibration was carried out based on the PC surrogate. The surrogate was the approximate response surface obtained using a level-three sparse grid cubature.

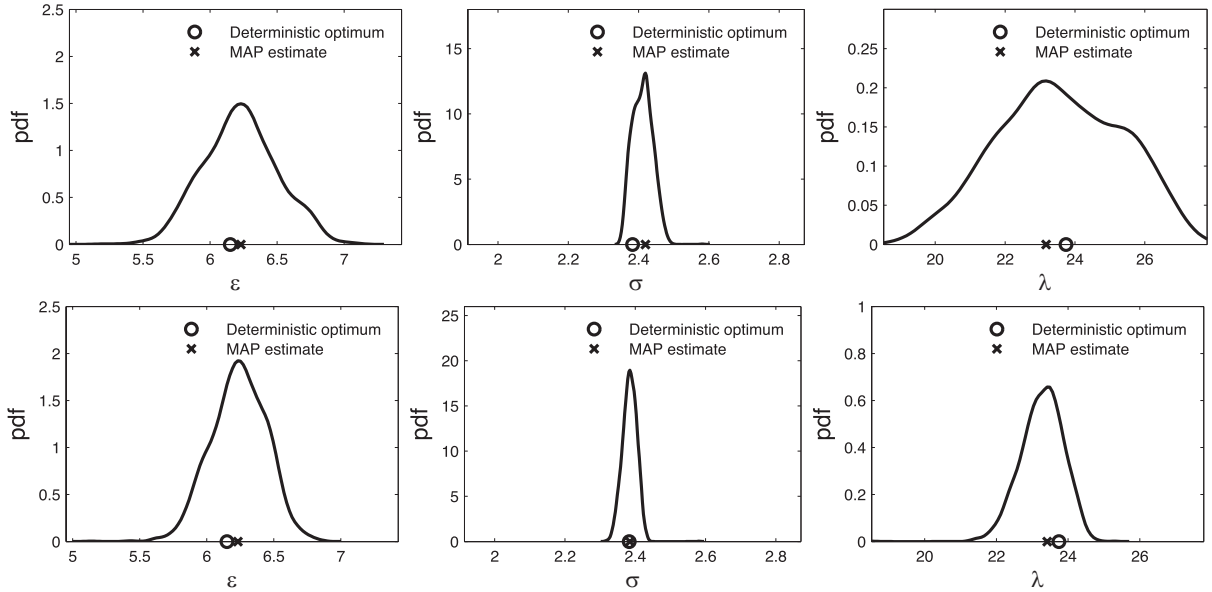


FIG. 4: Comparison between the deterministic and Bayesian calibration of the three mW parameters. Figures show the posterior density functions for parameters ϵ, σ, and λ. Also, the MAP estimates calculated based on the posterior density functions are compared with their deterministic counterparts. The x-range of the plots are set equal to the prior range of the three parameters. The differences between the figures on the first and second row are due to different noise levels in the RDF data. The standard deviations of the noise model for density and enthalpy are 0.05 and 0.5, respectively, in both rows. The standard deviations for the noise in RDF residual are assumed to be 1 for the top row and 0.5 for the bottom one.

As mentioned earlier, one of the additional capabilities of the Bayesian calibration is the quantification of impact of data quality. Figure 4 also shows the calibration results associated with different assumptions for the measurement inaccuracy. We have in particular demonstrated the data quality impact by changing the error models in the observation of RDFs. As can be seen, as we lower the accuracy level in the RDF observation, our confidence in the calibrated parameters deteriorates accordingly. This change is more apparent in the probability models of λ of σ and less so for ϵ. It could be deduced from these comparative results that the RDF is more sensitive to changes in λ and σ. Also, it could be seen that the MAP estimates get closer to the deterministic optima as the measurement noise decreases. This is consistent with the fact that given an unbiased surrogate model approximating the likelihood function—if we can make a perfect observation, since we are not incorporating any other source of uncertainty or error—then the posterior density function will be a Dirac delta function at the deterministic optimum.

5.2 Sensitivity of the Response to Model Parameters

In the previous section, we observed the impact of the measurement noise on the calibration of the mW parameters. This impact was seen to be different for different parameters. Alternatively, we can perform the sensitivity analysis to quantify the different level of sensitivity for different parameters. In this work, we make use of an approximate derivative-based approach that uses the PC expansion in uncovering the sensitivity character of the system response [37]. Specifically, we approximate the first derivatives of the response with respect to each parameter, as a sensitivity index. These first derivatives are computed at the mean values of the random parameters, using a very small number of simulations. Specifically, given a D-dimensional parameter space, only $D+1$ full simulations are needed to calculate these derivatives. Figure 5 shows the values of these indices for the three observables. As can be seen, RDF residual is most sensitive to σ and λ, which is consistent with the observation made based on Fig. 4.

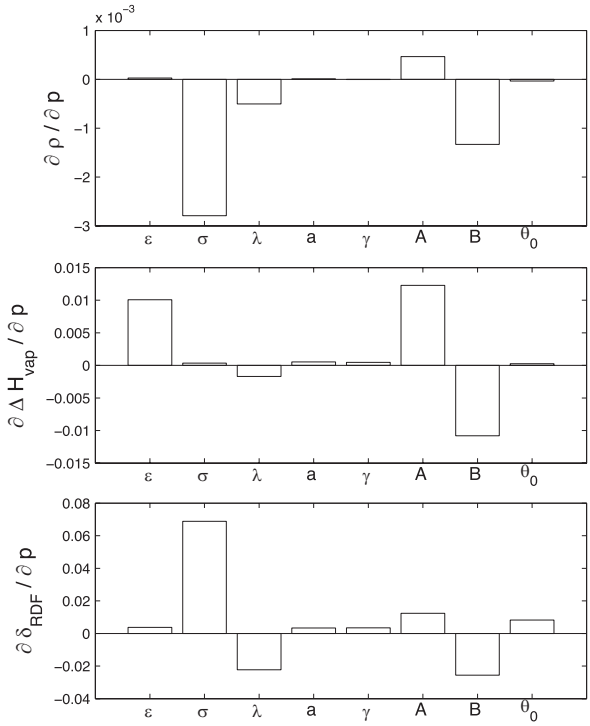


FIG. 5: The approximate values for the first derivatives of the observables with respect to the eight parameters in the mW model. These values can serve as a rough indicator of the global sensitivity character of the system. The range for each parameter was assumed to be centered around its deterministic calibrated value with the width equal to 5% of those nominal values. Note the difference in the derivative values in the three panels. RDF residual is more sensitive to the assumed small variability in the parameters compared to density and enthalpy of vaporization by around one order of magnitude.

5.3 Bayesian Model Ranking

We now demonstrate the quantitative Bayesian framework for model ranking and model averaging given the family of water models, described in Section 2. We aim to assess the plausibility of PC surrogates for each model in predicting the coarse-scale observables. In using the TIP4P/2005 potential of Eq. (1), we set both charge values q_i and q_j equal to a constant q, and considered four random parameters for Bayesian calibration: the two Lennard-Jones parameters A, B, together with the charge q of the "charge center" attached to oxygen and the distance of this "charge center" from the oxygen atom denoted by d_q. The prior distribution for each parameter is chosen to be uniform, which is the maximum entropy distribution when only the range of a random parameter is known. For each parameter, a number of calibrated values are available in the literature forming a family of TIP4 models, each predictive for certain water behaviors. Among the family members, the extrema of parameter values are extracted. The prescribed range for each parameter is one that is 10% larger than the distance of the two extrema, and is centered at the midpoint between the two extrema. Doing so, we consider all the plausible parameter values in the Bayesian calibration procedure. The surrogates for the three observables were constructed using level-two sparse grid points in the four-dimensional parameter space. For the LJG potential of Eq. (5), the parameters ϵ, σ, and B^* are considered to be random. The average values of these random parameters together with the nominal values of the other parameters are set equal to those that minimize the relative entropy, which are reported in [22], i.e., $\epsilon = 4.9$ kcal/mol, $\sigma = 2.43$ Å, and $B = 5.6$ kcal/mol. The random parameters are assumed to be uniformly distributed in ranges that are centered at the corresponding nominal values and are as wide as 10% of the respective average values for ϵ and B and 8% for σ. First, let us visually assess the performance of each model by comparing its posterior prediction versus the observation data (see Fig. 6).

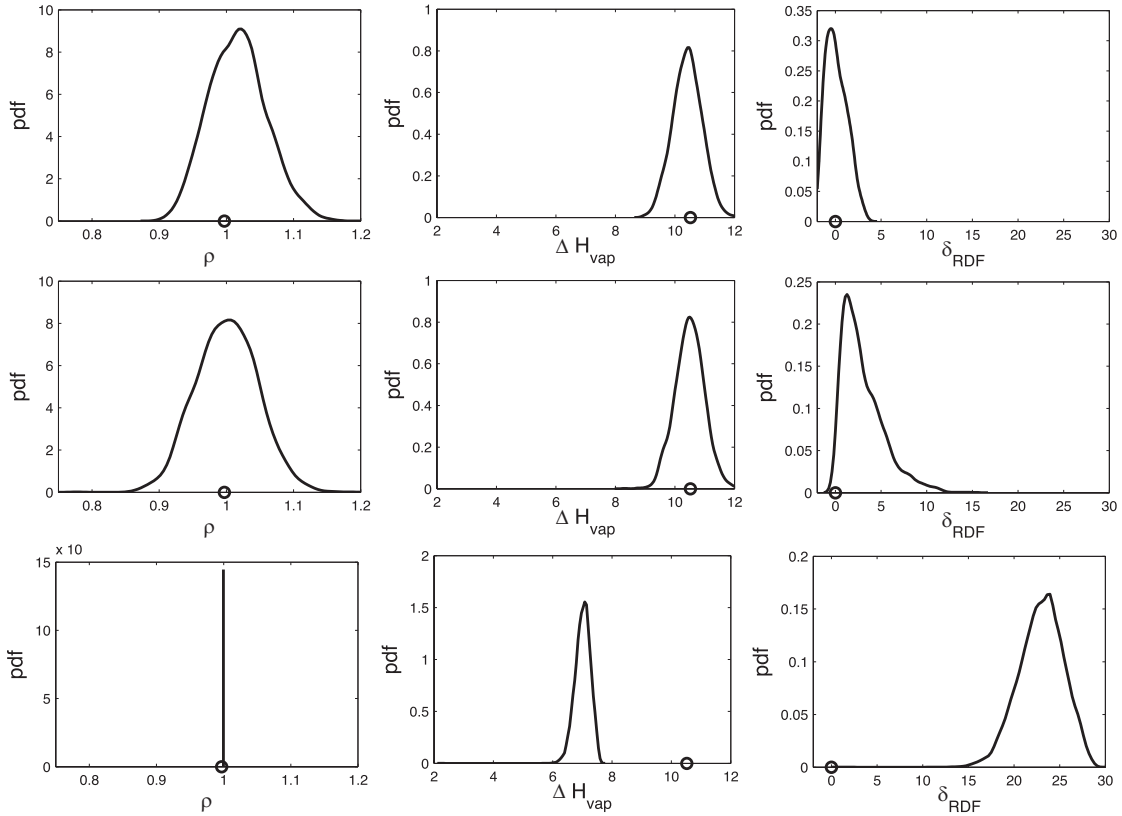


FIG. 6: Prediction of the three observables using the three models: TIP4P/2005 (first row), mW (second row), and LJG (third row). The response surfaces are evaluated at the posterior sample points in the parameter space of each model. The predictions are compared with the observed values (circle markers).

In order to quantitatively assess the plausibility of each surrogate, we need to calculate the integral in Eq. (14). This is done by evaluating the PC surrogates at the parameter samples drawn from the prior uniform distributions. The reason is that calculation of marginal likelihood according to Eq. (14) using posterior samples suffers from instabilities caused by the occurrence of small likelihood posterior samples. We observed these instabilities in using the harmonic mean estimator of Eq. (22) due to the occasional occurrence of low likelihood posterior samples and used samples from the prior distributions [Eq. (20)]. We could compare the predictive capacity of these models with respect to different observables. One can evaluate the capability of these models to simultaneously reproduce all three observables, namely, density, enthalpy of vaporization, and RDF, or their capability in reproducing each observable separately. We quantify these capabilities using Bayes factors.

Table 1 includes the relative Bayes factors between different pairs of models corresponding to difference observable cases. Because of the way the LJG potential was formed and calibrated, it is expected that it can predict the structural properties, namely, the RDF, rather accurately. The quantified Bayes factors support this fact, where it can be seen that the RDF predictive capacity of the LJG model is three times that of the mW model, and is almost the same as that of the computationally expensive TIP4P/2005. The overall predictive capacity is, however, extremely poor, as seen in the table and as evident in the LJG prediction of the enthalpy of vaporization shown in Fig. 6. However, in Fig. 6, one notices a very poor RDF prediction, measured by the RDF residual, obtained from the LJG model. The reason is that these predictions were made based on samples from the posteriors that were inferred from all three observables. In Fig. 7, we show the posterior prediction of the RDF residual, when we conduct an isolated inference procedure using only RDF residual as the single observables. In this figure, it is shown qualitatively that the LJG model has the superior performance in reproducing the RDF, compared to the other two models, as supported by the Bayes factors in Table 1.

5.4 Model Averaged Prediction of Higher-Dimensional Structural Properties

In this section we investigate the performance of an ensemble of surrogates over predictions made by individual surrogates. In particular, we demonstrate a case where the ensemble prediction obtained by two surrogates aggregated

TABLE 1: Computed Bayes factors between each pair in the three-model family. The marginal likelihood integral was numerically calculated using prior samples [Eq. (20)]

	ρ	ΔH_{vap}	δ_{RDF}	Total
TIP4P/mW	2.28	0.71	3.12	2.13
TIP4P/LJG	0.25	5E+12	1.09	6E+16
mW/LJG	0.11	7E+12	0.35	3E+16

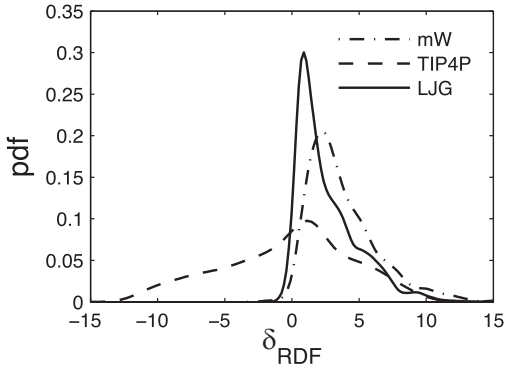


FIG. 7: The pdf of the predictions for RDF residual for the three models. The parameter samples used for this prediction are from the posteriors inferred from the empirical RDF as the only observable.

using Bayesian model averaging weights is more accurate than the prediction made by each individual surrogate. Let us consider the PC-based surrogates for mW and LJG, each generated using level-three sparse grid points. We denote these two models (surrogates) by M_{mW} and M_{LJG}. Each surrogate consists of three response surfaces for the three observables ρ, ΔH_{cap}, and δ_{RDF}, each constructed upon the three-dimensional parameter space of each model.

Let us focus on the predictive capacity for the estimation of four structural properties: the magnitudes and locations of the extrema of the first peak and valley of the radial distribution function, $g(\cdot)$. The empirical RDF of water used in this study, which is reported in [36], is shown in Fig. 8, where the two extrema are highlighted based on which of the four structural QoIs are computed. Let $\boldsymbol{y}_k \in [\mathbb{R}^+]^4$ denote the vector of four QoIs obtained from model α, where $k \in \{\text{mW}, \text{LJG}\}$. The Bayesian model averaged prediction can be given by

$$\pi(\boldsymbol{y}|d) = \sum_{k\in\{\text{mW,LJG}\}} \pi(\boldsymbol{y}|M_k, d)\pi(M_k|d), \tag{24}$$

where d is taken to be the RDF residual, as it is the informative observation for the structural QoIs, and $\pi(M_k|d)$ is given in Eq. (15). The term $\pi(\boldsymbol{y}|M_k, d)$ is the probability of observing $\boldsymbol{y}$ conditioned on M_k and d. This means that the uncertain parameters of model M_k are inferred from data d.

In order to compare the predictions of the two-member ensemble with that of individual surrogates, we calculate the average values for the four QoIs, denoted by $\bar{\boldsymbol{y}}_{ens}$ and $\bar{\boldsymbol{y}}_k$, according to

$$\begin{aligned} \bar{\boldsymbol{y}}_{ens} &= \sum_{k\in\{\text{mW,LJG}\}} \mathbb{E}(\boldsymbol{y}|M_k, d)\pi(M_k|d), \\ \bar{\boldsymbol{y}}_k &= \mathbb{E}(\boldsymbol{y}|M_k, d) \;\; k \in \{\text{mW}, \text{LJG}\}. \end{aligned} \tag{25}$$

The expectation operator is applied on individual predictions corresponding with the posterior parameter samples. A representative response surface for the first peak magnitude corresponding to mW and LJG is shown in Fig. 9. The thickness of the surfaces shows the level of variability, in particular the standard deviation, due to the third uncertain parameter not shown in the figures.

To evaluate the accuracy of each prediction, the root-mean-square errors (RMSE) with respect to the four empirical values are computed. The empirical locations and magnitudes of the first peak and valley are 2.7900, 2.5691, 3.4500, and 0.8386, respectively. Table 2 tabulates the RMSE values for different cases. As can be seen, the ensemble prediction is more accurate than the individual predictions from the two member models.

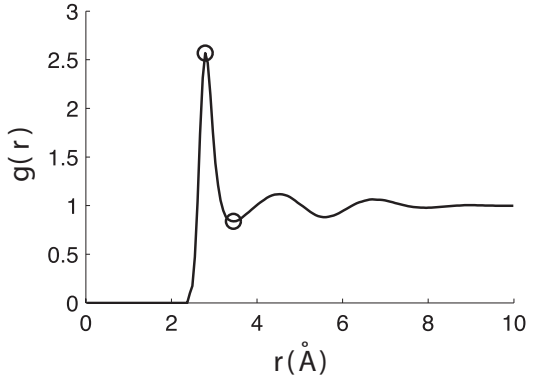
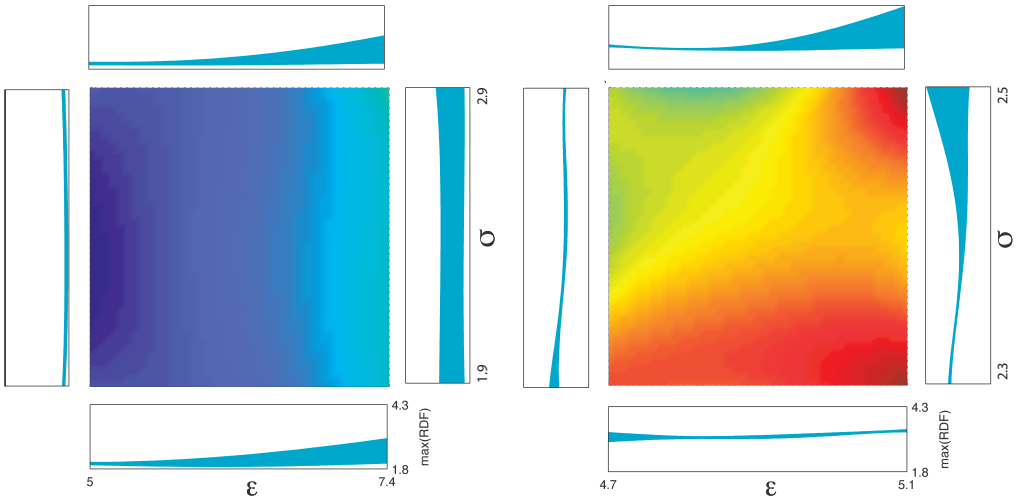


FIG. 8: The empirical radial distribution function of water after [36]. The circle markers show the extrema of the first peak and valley.



(a) (b)

FIG. 9: Comparison between the PC-based response surfaces for the peak magnitude of the RDF obtained by the (a) mW and (b) LJG models.

TABLE 2: Computed RMSE values between the predicted and observed QoIs. The four QoIs are the magnitudes and locations of the first peak and valley of the RDF

	mW	**LJG**	**(mW+LJG)**
RMSE	0.9700	0.9537	0.9412

6. CONCLUSION

In this study, we examined the plausibility of three surrogate-based water models. We demonstrated the application of Bayesian model ranking together with PC-based surrogates, and showed how the quantitative ranking provided by this framework is in agreement with the experts' expectation. We also demonstrated how the Bayesian calibration framework for these models can be used to characterize the variabilities in the input parameters that give rise to the observed variability of some measured quantities. In other words, it effectively, and in a quantitative way, takes into account the interplay between the accuracy of the experimentation/instrumentation procedures and the uncertainty level in the parameter values. Also, compared to the deterministic calibration, in this approach, even if the point estimates of the parameters are used for practical purposes, the modeler will have obtained the quantified confidence in these point estimates. We also discussed some insights on how computational acceleration can be achieved and how cautious one should be in interpreting the results. In particular, due to the use of surrogates in lieu of the full simulator, the posterior samples are subject to an error. In the current form, since the numerical error in the use of surrogates is not quantified and is absent in the MCMC procedure, one cannot account for the impact of numerical error on the posterior samples. Therefore, the Bayesian ranking scores are interpreted as the plausibility of the PC surrogates constructed from these physical systems. The approach that we used in this paper can be further extended to assist with the design of experiments. Specifically, because of the analytical form of the approximate response surfaces, they lend themselves readily to relatively fast sensitivity studies, the outcome of which guides the choice of observables that are most informative for the model parameters.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Liam Jacobson for helpful discussions. This research was sponsored by ARL under Cooperative Agreement Number W911NF-12-2-0023. The views and conclusions contained in this document

are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

1. Reith, D., Pütz, M., and Müller-Plathe, F., Deriving effective mesoscale potentials from atomistic simulations, *J. Computat. Chem.*, 24:1624–1636, 2003.
2. McGreevy, R.L. and Pusztai, L., Reverse Monte Carlo simulation: A new technique for the determination of disordered structures, *Mol. Simul.*, 1(6):359–367, 1988.
3. Izvekov, S. and Voth, G.A., A multiscale coarse-graining method for biomolecular systems, *J. Phys. Chem. B*, 109(7):2469–2473, 2005.
4. Shell, M.S., The relative entropy is fundamental to multiscale and inverse thermodynamic problems, *J. Chem. Phys.*, 129(14):144108, 2005.
5. Fox, B., *Strategies for Quasi-Monte Carlo*, Boston: Kluwer Academic, 1999.
6. Niederreiter, H., Quasi-Monte Carlo methods and pseudo-random numbers, *Bull. Am. Math. Soc.*, 84(6):957–1041, 1978.
7. Ghanem, R. and Spanos, P., *Stochastic Finite Elements: A Spectral Approach*, New York: Dover Publications, 2003.
8. Xiu, D. and Karniadakis, G.E., Modeling uncertainty in flow simulations via generalized polynomial chaos, *J. Computat. Physi.*, 187(1):137–167, 2008.
9. Sudret, B., Global sensitivity analysis using polynomial chaos expansions, *Reliab. Eng. Syst. Saf.*, 93(7):964–979, 2008.
10. Xiu, D. and Hesthaven, J., High-order collocation methods for differential equations with random inputs, *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.
11. Marzouk, Y.M., Najm, H.N., and Rahn, L.A., Stochastic spectral methods for efficient Bayesian solution of inverse problems, *J. Computat. Phys.*, 224(2):560–586, 2007.
12. Marzouk, Y. and Xiu, D., A stochastic collocation approach to bayesian inference in inverse problems, *Commun. Computat. Phys.*, 6:826–847, 2009.
13. Rizzi, F., Najm, H., Debusschere, B., Sargsyan, K., Salloum, M., Adalsteinsson, H., and Knio, O., Uncertainty quantification in MD simulations. Part II: Bayesian inference of force-field parameters, *Multiscale Model. Simul.*, 10(4):1460–1492, 2012.
14. Rizzi, F., Jones, R.E., Debusschere, B.J., and Knio, O.M., Uncertainty quantification in MD simulations of concentration driven ionic flow through a silica nanopore. I. Sensitivity to physical parameters of the pore, *J. Chem. Phys.*, 138(19):194104, 2013.
15. Rizzi, F., Jones, R.E., Debusschere, B.J., and Knio, O.M., Uncertainty quantification in md simulations of concentration driven ionic flow through a silica nanopore. ii. uncertain potential parameters, *J. Chem. Phys.*, 138(19):194105, 2013.
16. Remsing, R., Rodgers, J., and Weeks, J., Deconstructing classical water models at interfaces and in bulk, *J. Stat. Phys.*, 145(2):313–334, 2011.
17. Izvekov, S., Swanson, J.M.J., and Voth, G.A., Coarse-graining in interaction space: A systematic approach for replacing long-range electrostatics with short-range potentials, *J. Phys. Chem. B*, 112(15):4711–4724, 2008.
18. Hadley, K.R. and McCabe, C., Coarse-grained molecular models of water: a review, *Mol. Simul.*, 38(8-9):671–681, 2012.
19. Jacobson, L.C., Kirby, R.M., and Molinero, V., How short is too short for the interactions of a water potential? Exploring the parameter space of a coarse-grained water model using uncertainty quantification, *J. Phys. Chem. B*, 118(28):8190–8202, 2014.
20. Abascal, J.L.F. and Vega, C., A general purpose model for the condensed phases of water: TIP4P/2005, *J. Chem. Phys.*, 123(23):144504, 2005.
21. Molinero, V. and Moore, E.B., Water modeled as an intermediate element between carbon and silicon, *J. Phys. Chem. B*, 113(13):4008–4016, 2009.
22. Chaimovich, A. and Shell, M.S., Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy, *Phys. Chem. Chem. Phys.*, 11:1901–1915, 2009.

23. Stillinger, F. and Weber, T., Computer simulation of local order in condensed phases of silicon, *Phys. Rev. B*, 31:5262–5271, 1985.
24. Balakrishnan, S., Roy, A., Ierapetritou, M.G., Flach, G.P., and Georgopoulos, P.G., Uncertainty reduction and characterization for complex environmental fate and transport models: An empirical bayesian framework incorporating the stochastic response surface method, *Water Resour. Res.*, 39(12):1350, 2003.
25. Wang, J. and Zabaras, N., Using Bayesian statistics in the estimation of heat source in radiation, *Int. J. Heat Mass Transfer*, 48(1):15–29, 2005.
26. Eldred, M., *Recent Advances in Non-Intrusive Polynomial Chaos and Stochastic Collocation Methods for Uncertainty Analysis and Design*, AIAA Paper 209-2274, 2009.
27. Ma, X. and Zabaras, N., An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations, *J. Computat. Phys.*, 228(8):3084–3113, 2009.
28. Smolyak, S., Quadrature and interpolation formulas for tensor products of certain classes of functions, *Sov. Math. Dokl.*, 4:240–243, 1963.
29. Jeffreys, H., *Theory of Probability*, Reprint of the 1983 edition, Oxford Classic Texts in the Physical Sciences, Oxford: Oxford University Press, 1998.
30. Kass, R.E. and Raftery, A.E., Bayes factors, *J. Am. Stat. Assoc.*, 90(430):773–795, 1995.
31. Tierney, L. and Kadane, J., Accurate approximations for posterior moments and marginal densitie, *J. Am. Stat. Assoc.*, 81:82–86, 1986.
32. Bernardo, J. and Smith, A., *Bayesian Theory*, Chichester, UK: John Wiley & Sons, 2000.
33. Kirkwood, J.G., Statistical mechanics of fluid mixtures, *The Journal of Chemical Physics*, 3(5):300–313, 1935.
34. Gelman, A. and Meng, X.L., Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Stat. Sci.*, 13(2):163–185, 1998.
35. Newton, M.A. and Raftery, A.E., Approximate Bayesian inference with the weighted likelihood bootstrap, *J. R. Stat. Soc. Ser. B*, 56(1):3–48, 1994.
36. Skinner, L.B., Huang, C., Schlesinger, D., Pettersson, L.G.M., Nilsson, A., and Benmore, C.J., Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide q-range, *J. Chem. Phys.*, 138(7):074506, 2013.
37. Xiu, D., Numerical integration formulas of degree two, *Appl. Numer. Math.*, 58(10):1515–1520, 2008.