



# Bayesian three-point water models



Alfred T. Nordman<sup>1</sup>, Stefan Engblom<sup>2</sup> & David van der Spoel<sup>1</sup> ✉

We introduce a Bayesian framework leveraging synthetic likelihoods to enable uncertainty quantification and robust inference of non-bonded force parameters in three-point water models. The approach integrates multiple experimental observables—enthalpy of vaporization, molecular volume, the radial distribution function, and hydrogen bonding patterns—to explicitly infer model parameters. Beyond parameter estimation, we quantify uncertainty in both inference observables and validation properties, including those that are difficult to target by other means. By systematically analyzing the response of these observables to parameter variations, our method highlights inherent limitations of three-point water models. These findings highlight the utility of our framework in integrating diverse data sources in a principled uncertainty quantification workflow, ultimately improving confidence in the ability of molecular dynamics simulations to reproduce experimental data. Additionally, we evaluate the performance of the mean and the mode of the posterior distribution, demonstrating the limitations of this family of models.

Reliable models are fundamental to molecular dynamics (MD) simulations<sup>1</sup>, and the representation of water molecules plays a crucial role in applications ranging from materials and physics<sup>2,3</sup> to biology<sup>4–6</sup>. Three-point water models, one of the most well-known being TIP3P<sup>7</sup>, are a popular choice due to their computational efficiency but due to their simplicity, they have limited accuracy and often struggle to simultaneously reproduce key experimental properties<sup>8</sup>. Arguably, this underscores the need for rigorous inference frameworks that not only estimate parameters but also help identify the fundamental limitations of simple model classes like TIP3P, guiding the development of more accurate models in the future.

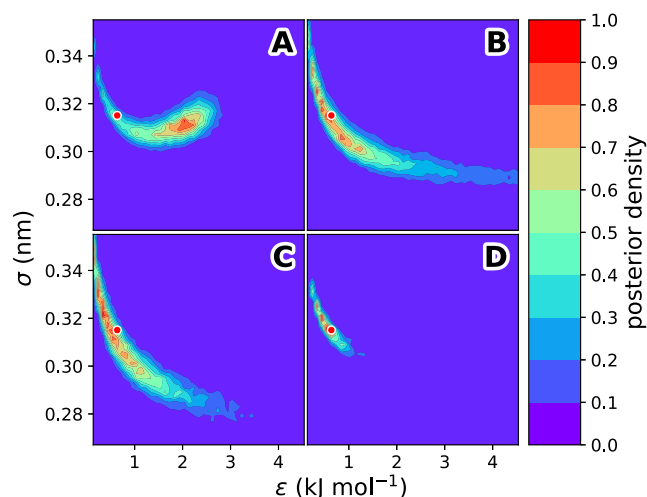
Traditional model calibration methods typically fit parameters to a limited set of experimental observables, which can lead to parameter choices that fail to generalize across different conditions, such as broader temperature ranges<sup>9</sup>. Moreover, parameters fit to reproduce certain properties under specific conditions do not necessarily guarantee accurate predictions of other properties under those same conditions. A robust estimation approach must therefore incorporate multiple experimental restraints to yield parameter sets that generalize across diverse thermodynamic conditions with quantified uncertainty.

Bayesian inference provides a consistent framework for parameter estimation by treating model parameters as probability distributions rather than fixed values<sup>10</sup>. Access to the parameter posterior is essential for quantifying uncertainties, but constructing it is often computationally demanding. To address this, Cailliez et al.<sup>11</sup> developed a surrogate-based approach, using Gaussian process models to approximate simulation outputs and efficiently sample the four-point model TIP4P<sup>7</sup> parameter posterior. This strategy enables uncertainty estimation even when direct

posterior evaluation is infeasible. Ideally, access to the full posterior is preferred; when the simulated observables asymptotically follow a multivariate normal distribution, a synthetic likelihood (SL) can be constructed<sup>12</sup>. In combination with Metropolis-Hastings (MH) sampling<sup>13</sup>, this enables direct exploration of the posterior, albeit at a higher computational cost than relying on surrogate models. Recent advances in GPU computing have made this approach feasible at last, enabling the many repeated simulations required for SL inference to be performed within a practical time frame. Unlike conventional optimization methods that yield point estimates, Bayesian approaches explicitly characterize uncertainty by considering the full posterior, which is particularly advantageous in molecular simulations where small parameter variations can induce large changes in system behavior. By integrating multiple data sources and rigorously accounting for uncertainty, Bayesian frameworks offer a flexible and robust approach to force field (FF) parameterization<sup>14,15</sup>.

In this work, we present a Bayesian framework to estimate the posterior distribution of the non-bonded parameters assigned to the oxygen atom in a TIP3P-like water model—specifically, its Van der Waals  $\epsilon$  and  $\sigma$  parameters, as well as its partial charge  $q$ . We then systematically quantify the variance and bias of both inference observables and those requiring computationally intensive simulations to evaluate, such as the static dielectric constant  $\epsilon_0$  or the diffusion coefficient  $D$ . Force fields differ in the physical properties they are optimized to reproduce, with bulk properties like enthalpy of vaporization  $\Delta H_{vap}$  and density  $\rho$  being the most common. Other properties include features such as the radial distribution function (RDF) and, in particular, hydrogen bonding for water molecules. By directly addressing the competing demands of reproducing multiple bulk properties—each reflecting different physical aspects of the liquid state—our method refines

<sup>1</sup>Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. <sup>2</sup>Division of Scientific Computing, Science for Life Laboratory, Department of Information Technology, Uppsala University, Uppsala, Sweden. ✉e-mail: [david.vanderspoel@icm.uu.se](mailto:david.vanderspoel@icm.uu.se)



**Fig. 1 | Contour plots of scaled SL posteriors.** Plots were obtained from single-point simulations based on 45-by-45 equidistant grid points of  $\epsilon$  and  $\sigma$ . Likelihoods are computed based on (A) enthalpy of vaporization  $\Delta H_{\text{vap}}$  and molecular volume  $V_M$ , (B) the RDF properties  $\tilde{r}_{\text{OO},1}$  and  $\tilde{g}_{\text{OO},1}$ , (C) the average hydrogen bond distance  $\langle r_{\text{HB}} \rangle$  and angle  $\langle \theta_{\text{HB}} \rangle$ , and (D) all six properties combined. The red point marks the parameter values of the original TIP3P model.

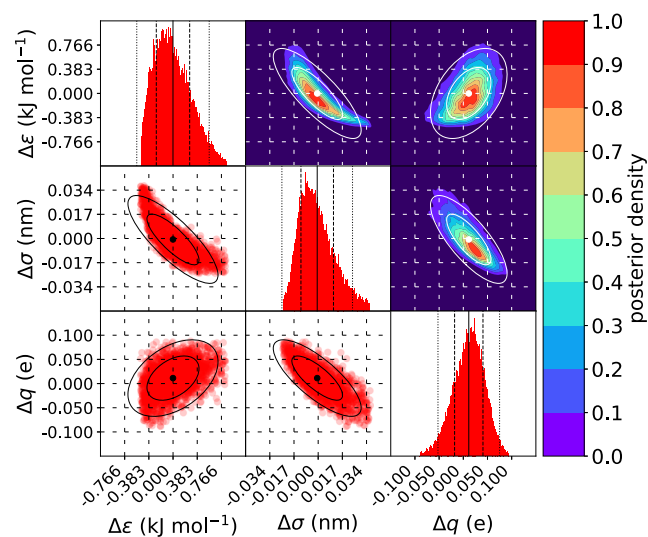
TIP3P parameters while offering broader insights into the inherent limitations of three-point water models at 298.15 K and 1 atm. Finally, we propose two new parameter sets, based on the mode and mean of the posterior distribution respectively<sup>14</sup>, and quantify the uncertainties for both the inference observables and independent validation observables. These estimates establish fundamental bounds on the accuracy achievable within this class of models, offering a benchmark for evaluating properties of other FF models and distinguishing between errors arising from parameter choices and those fundamentally limited by the model and simulation themselves<sup>16,17</sup>.

## Results

### Effect of inference properties on posterior landscape

To investigate how the choice of inference properties affects the SL posterior, single-point simulations over a two-dimensional ( $\epsilon$ ,  $\sigma$ ) grid were performed. The considered properties include the enthalpy of vaporization  $\Delta H_{\text{vap}}$ , molecular volume  $V_M$ , the RDF features  $\tilde{r}_{\text{OO},1}$  and  $\tilde{g}_{\text{OO},1}$ , and the average hydrogen-acceptor distance in a hydrogen bond  $\langle r_{\text{HB}} \rangle$  and average  $\langle \theta_{\text{HB}} \rangle$ . The definition of as well as how the RDF and the hydrogen properties are computed are explained in “Integration of Experimental Observables”.

As shown in Fig. 1, the posterior landscape varies depending on the selected observable set. The  $(\Delta H_{\text{vap}}, V_M)$  joint posterior in Fig. 1A shows a negative correlation for  $\epsilon < 1.2 \text{ kJ mol}^{-1}$  as either  $\epsilon$  or  $\sigma$  accounts for a balanced repulsive force. For  $\epsilon > 1.8 \text{ kJ mol}^{-1}$  there is a positive correlation, suggesting that increasing  $\epsilon$  compensates for the overall weakening of interaction strength—both repulsive and attractive—as molecules are pushed further apart by larger  $\sigma$  values. More notably, there is a probability maximum corresponding to the mode at  $\epsilon \approx 2.04 \text{ kJ mol}^{-1}$ —more than three times the value for TIP3P. The joint posterior of  $(\tilde{r}_{\text{OO},1}, \tilde{g}_{\text{OO},1})$  in Fig. 1B follows a strictly negative trend in correlation, mainly due to the impact of repulsion from molecules within the first solvation shell. The distance to the first peak of the RDF and  $\sigma$  are both correlated with the Van der Waals radius of an atom, which is demonstrated by the convergence of  $\sigma$  for large  $\epsilon$ . A similar trend can be seen in Fig. 1C for  $(\langle r_{\text{HB}} \rangle, \langle \theta_{\text{HB}} \rangle)$ , but even lower values for  $\sigma$  are considered plausible. All three joint pairwise posteriors suggest that nonzero likelihood regions exist for  $\epsilon > 2 \text{ kJ mol}^{-1}$ . However, the regions of high likelihood show limited overlap. Therefore, when the six properties are considered all at once (Fig. 1D), the region of agreement is now smaller, limiting the parameter ranges of the posterior and penalizing



**Fig. 2 | Difference between  $(\epsilon, \sigma, q)$  samples from MCMC and the original TIP3P model.** The center of each plot indicates a zero difference to the TIP3P model. Diagonal plots show marginal distributions of each parameter; off-diagonal plots display pairwise (2D) marginal distributions as scaled kernel density estimates (upper triangle) and scatter plots (lower triangle). Solid lines in the diagonals denote the distribution mean; dashed lines indicate 64.2% and 95.4% confidence intervals. Corresponding points and circles are used for the 2D distributions. Note: y-axes apply only to off-diagonal plots.

higher  $\epsilon$  values. The individual single-observable posteriors are presented in Figs. S2–S7.

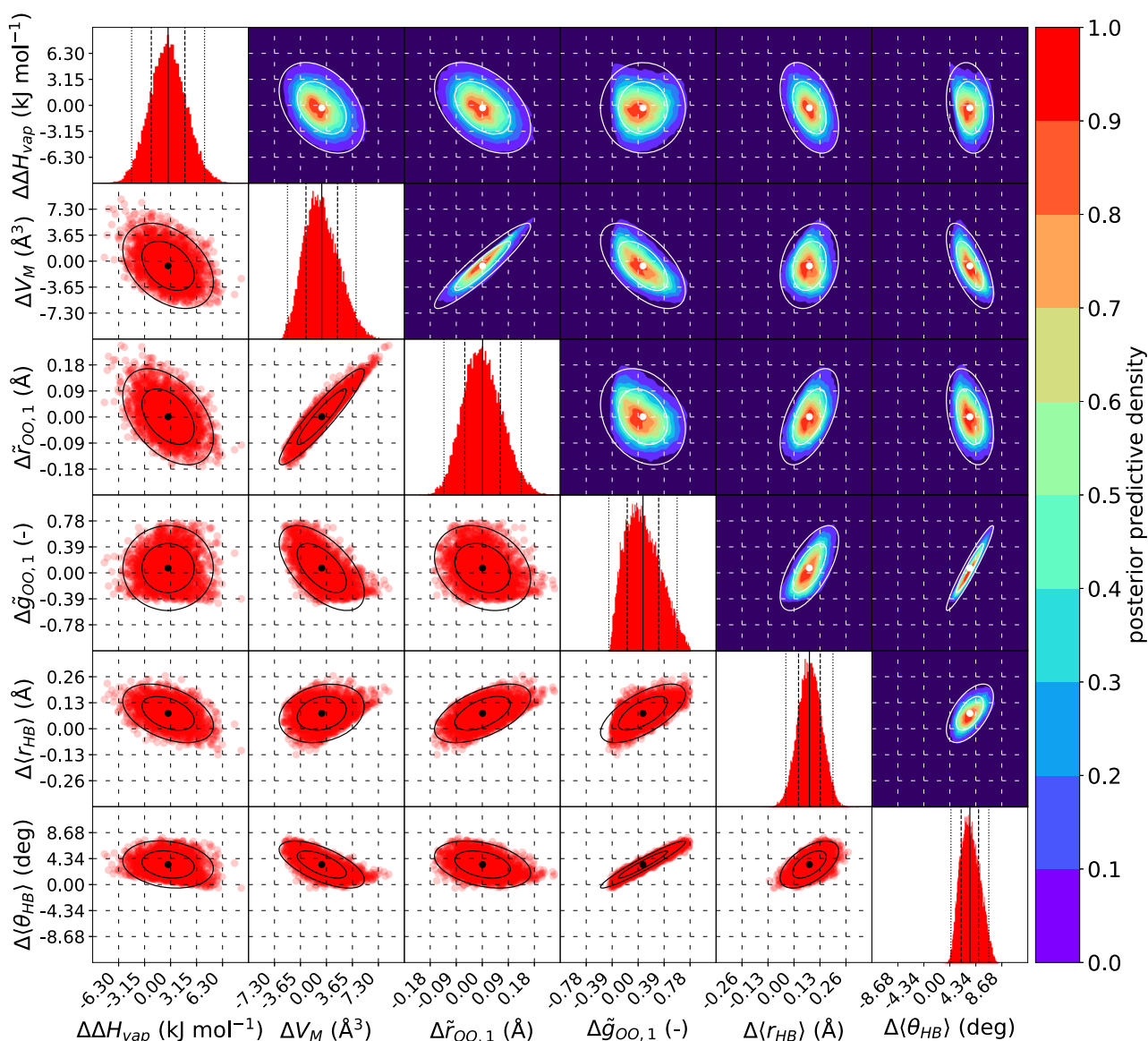
### Validation of synthetic likelihood approximation

The multivariate normality of the synthetic likelihoods across a 45-by-45 grid was assessed using Mardia’s test<sup>18</sup> and the ratio of distributions considered to be normal was 0.9185, 0.9284, and 0.9333 for subplots A, B, and C, respectively in Fig. 1. For the six-property joint posterior in subplot D, the ratio was 0.8489. Given the close agreement between the observed ratios and the theoretical values for multivariate normal distributions, the synthetic likelihoods are considered asymptotically normal, validating the applicability of the SL approximation for this parameter space. For completeness, the univariate normality Shapiro-Wilk tests<sup>19,20</sup> of the individual single-observable posteriors are presented in Figs. S8–S13, and the multivariate normality Mardia’s tests corresponding to the subplots in Fig. 1 are presented in Fig. S14. For comparison, if one were to assume independent variables, the multi-dimensional Shapiro-Wilk tests corresponding to Fig. 1A–D are 0.9057, 0.9007, 0.9052, and 0.7383. The reference value for independent truly normal is  $0.95^2 = 0.9025$  in two dimensions and  $0.95^6 \approx 0.7351$  in six dimensions (see Synthetic Likelihood and Normality Assumption”).

Differences between the two-dimensional posterior and the original TIP3P parameter set are shown in Fig. S15. The sampled posterior reflects a nonlinear, negatively correlated relationship between  $\epsilon$  and  $\sigma$ , consistent with the trends observed in the grid-based likelihoods (Fig. 1). Observables associated with the sampled posterior in Fig. S15 are shown in Fig. S16.  $\Delta H_{\text{vap}}$  and  $V_M$  closely match their experimental reference values. In contrast, the average hydrogen bond distance  $\langle r_{\text{HB}} \rangle$  and angle  $\langle \theta_{\text{HB}} \rangle$  deviate from the reference by approximately one and two standard deviations, respectively.

### Influence of partial charge on observables and posterior

To assess the effect of atomic charge on model behavior, we extended the parameter space from two to three dimensions by including the partial charge  $q$  of the oxygen atom as an additional parameter (Fig. 2). This change increases the posterior spread slightly, more for  $\sigma$  than for  $\epsilon$ . While the negative correlation as seen in Fig. S15 between  $\epsilon$  and  $\sigma$  persists,  $q$  exhibits a



**Fig. 3 | Difference between measured observables and experimental reference values.** Differences correspond to the samples from the MCMC ( $\epsilon$ ,  $\sigma$ ,  $q$ ) posterior. The plots are centered at the reference data. Diagonal plots show marginal distributions of each observable; off-diagonal plots display pairwise (2D) marginal

distributions as scaled kernel density estimates (upper triangle) and scatter plots (lower triangle). Solid lines in the diagonals denote the distribution mean; dashed lines indicate 64.2% and 95.4% confidence intervals. Corresponding points and circles are used for the 2D distributions. Note:  $y$ -axes apply only to off-diagonal plots.

positive correlation with  $\epsilon$  and a negative one with  $\sigma$ . Mean parameter values remain close to those of the original TIP3P model in Table 2.

A similar trend to that observed for the parameters is also seen in the observables. When comparing only the diagonal plots of the two-dimensional (Fig. S16) and three-dimensional (Fig. 3) posteriors, the variance increases slightly for  $\tilde{g}_{OO,1}$ ,  $\Delta H_{vap}$ , and  $\langle \theta_{HB} \rangle$ . A more pronounced increase is seen for  $\tilde{r}_{OO,1}$  and  $V_M$ , both of which are strongly correlated with molecular density—consistent with the larger variance observed in the  $\sigma$  parameter. The variance of  $\langle r_{HB} \rangle$  remains unchanged. The limited improvement from including  $q$ , particularly for  $\langle r_{HB} \rangle$  and  $\langle \theta_{HB} \rangle$ , suggests that the original TIP3P model's charge is already near-optimal for this model family and set of inference observables, likely due to the fixed geometry at which the partial charges are positioned.

#### Uncertainty quantification

The coefficient of variation (CV), relative bias (RB), and relative standard error (RSE), as defined in “Quantification of Errors”, are summarized in Table 1. For the properties evaluated during MCMC, parameter uncertainty

—reflected in the CV—is modest, typically between 1% and 11%, and consistently exceeds the RSE, which remains near or below 0.1%. This indicates that uncertainty in the parameters, rather than finite sampling, is the primary source of simulation error in these cases. In contrast, for properties not included during MCMC and evaluated using the unscented transform (UT), the dominant source of error shifts toward model bias. Despite low CVs and RSEs, relative biases can be substantial—for example, the dielectric constant ( $\epsilon_0$ ) and heat capacity ( $C_p$ ) deviate markedly from experimental references, with  $D$  showing an RB of up to 68%. It should be noted that the deviation in  $C_p$  is mainly due to omitted vibrational quantum corrections<sup>17,21</sup>. These findings highlight that well-constrained parameters alone cannot overcome limitations inherent to the functional form of the TIP3P-like model, and emphasize the importance of distinguishing between parameter uncertainty, model inadequacy, and numerical noise.

Estimates obtained via UT, despite relying on a limited number of sigma points, generally align well with those from the full MCMC posterior, validating UT as a practical tool for error propagation in computationally expensive observables. While UT errors tend to be slightly higher than those

**Table 1 | Relative simulation errors for selected observables  $Y$ , including the coefficient of variation  $CV(Y)$ , relative bias  $RB(Y, y)$  with respect to reference values  $y$ , and relative standard error  $RSE(Y)$** 

Property	$y$	ref	$N_{\theta}$	CV( $Y$ ) (%)		RB( $Y, y$ ) (%)		RSE( $Y$ ) (%)	
				$\theta_{\text{MCMC}}$	$\theta_{\text{UT}}$	$\theta_{\text{MCMC}}$	$\theta_{\text{UT}}$	$\theta_{\text{UT}}$	T3B <sub>mean</sub>
$\Delta H_{\text{vap}}$	44.0 $\frac{\text{kJ}}{\text{mol}}$	59	2	5.1	9.9	0.3	0.5	0.0	0.0
			3	5.1	11	0.7	1.1	0.0	0.0
$V_M$	30.0 $\text{\AA}^3$	59	2	4.8	8.1	0.3	-0.2	0.0	0.0
			3	8.2	12	2.2	1.5	0.1	0.0
$\bar{r}_{\text{OO},1}$	2.97 $\text{\AA}$	42	2	1.3	2.4	-0.6	-0.7	0.0	0.0
			3	2.2	3.4	0.0	-0.2	0.0	0.0
$\bar{g}_{\text{OO},1}$	2.14	42	2	11	11	-2.6	-2.6	0.1	0.0
			3	12	12	-3.3	-3.3	0.0	0.0
$\langle r_{\text{HB}} \rangle$	1.93 $\text{\AA}$	60	2	2.8	3.9	-4.1	-4.2	0.0	0.0
			3	2.9	4.3	-3.8	-3.9	0.0	0.0
$\langle \theta_{\text{HB}} \rangle$	14.7 deg	60	2	7.0	7.5	-17	-17	0.0	0.0
			3	8.8	9.0	-19	-19	0.0	0.0
$\epsilon_0$	78.4	59	2	-	14	-	-15	4.7	0.4
			3	-	20	-	-26	6.3	0.4
$D$	$2.30 \times 10^{-5} \frac{\text{cm}^2}{\text{s}}$	61	2	-	38	-	-63	19	1.0
			3	-	41	-	-68	23	1.3
$C_p$	$75.3 \frac{\text{J}}{\text{molK}}$	59	2	-	4.4	-	-44	1.9	0.2
			3	-	6.3	-	-47	2.9	0.1

The observables  $Y(\theta)$  are evaluated using parameter samples  $\theta_{\text{MCMC}}$  from the posterior,  $\theta_{\text{UT}}$  from unscented transform sigma points based on the MCMC posterior mean and covariance, and T3B<sub>mean</sub> denoting the posterior mean. Reference values  $y$  and sources are also listed.

from MCMC, they remain consistent across both the two- and three-dimensional posteriors. One notable exception is  $V_M$ , where MCMC yields slightly higher variability than UT. As expected, the inclusion of atomic charge  $q$  in the three-dimensional posterior increases CV values across properties, reflecting increased parameter sensitivity. For all inference properties except  $\langle \theta_{\text{HB}} \rangle$ , RB remains below 5%. For this particular property, however, the bias is substantial: the reference value is approximately 17% and 19% lower than the samples for both posterior variants respectively. Overall, the decomposition of errors in Table 1 provides a comprehensive view of how parameter uncertainty, model limitations, and sampling precision each contribute to simulation accuracy.

A natural point of comparison for our simulated parameter sets—T3B<sub>mean</sub> and T3B<sub>mode</sub>—is the original TIP3P model. The more recent OPC3 model<sup>22</sup> is also included, which was developed through an exhaustive search in electrostatic parameter space and shown to closely approach the accuracy limits of rigid three-point water models. T3B<sub>mode</sub><sup>2</sup> in particular improves agreement with reference data for nearly all inference properties, with the exception of  $V_M$ , which deviates slightly more from experiment. The OPC3 model, which also incorporates modified molecular geometry, demonstrates the best overall agreement across all properties.

## Discussion

This study applied a Bayesian framework to investigate the uncertainty in model parameters and predictions, as well as the limitations of a TIP3P-like three-point water model for molecular simulations at 298.15 K and 1 atm. This was done by constructing a posterior distribution over the non-bonded parameters governing the intermolecular interactions with synthetic likelihoods. By including multiple experimental inference properties, we examined how well the model can reproduce dynamics as well as structural and thermodynamic properties. We further assessed the extent to which systematic bias and parameter uncertainty each contribute to prediction error, compared to errors stemming from model inadequacies and limited sampling.

The SL posteriors observed in our grids and MCMC samples provide insight into how different experimental observables shape the posterior distribution and thereby constrain the force field parameters. Notably, using  $\Delta H_{\text{vap}}$  and  $V_M$  (Fig. 1A) as inference targets favors a significantly higher  $\epsilon$  value than is used in the original TIP3P model, despite these properties being common targets in FF parameterization. To our knowledge, this shift toward higher  $\epsilon$  values has not been previously reported. In the original design of the TIP3 models by Jorgensen<sup>23</sup>, their parameter tuning included thermodynamics, energy distributions, RDF properties, and hydrogen bonding, similar to the six inference properties in this study (Fig. 1D), explaining the similarities between TIP3P and T3B<sub>mean</sub> variants.

The joint distributions in Fig. 2 reveal a pronounced exclusion zone at low  $\epsilon$ , corresponding to unphysical regimes where Lennard-Jones repulsion becomes too weak to counteract singular Coulombic attractions from point charges. The observed negative correlation between  $\epsilon$  and  $\sigma$  aligns with the structure of the Lennard-Jones potential: both parameters influence the interaction energy, but  $\sigma$  enters at higher exponents (6 and 12), making it more sensitive to minor changes than  $\epsilon$ . In three-dimensional inference (Fig. 2),  $\epsilon$  shows a positive correlation with  $q$ , while  $\sigma$  shows a negative one—suggesting that  $\epsilon$  is more involved in balancing repulsion, whereas  $\sigma$  plays a compensatory role in dispersion.

Our results also show that parameter uncertainty is the dominant contributor to simulation error, with inadequate model design explaining the systematic biases observed. In this context, evaluating the adequacy of the model refers to assessing whether the employed functional form and parameter ranges are sufficient to reproduce the selected observables within experimental uncertainty. The remaining systematic deviations indicate that limitations in the functional form, rather than parameter uncertainty, dominate the residual bias, especially for  $\langle \theta_{\text{HB}} \rangle$  and the validation properties, as shown in Table 1. This is particularly evident when comparing the posterior T3B<sub>mode</sub> to the OPC3 model in Table 2; improvements in geometry can apparently reduce the bias of three-point models beyond what non-bonded tuning alone can achieve, as has also been demonstrated for

**Table 2 | Model parameters and simulated properties for the original TIP3P model, the OPC3 model, and posterior samples from the two- and three-dimensional Bayesian inference**

Model	Exp	OPC3	TIP3P	T3B <sup>2</sup> <sub>mean</sub>	T3B <sup>2</sup> <sub>mode</sub>	T3B <sup>3</sup> <sub>mean</sub>	T3B <sup>3</sup> <sub>mode</sub>
Parameter							
$\epsilon$ (kJ/mol)	–	0.68369	0.63627	0.612404	0.473053	0.635323	0.520877
$\sigma$ (nm)	–	0.317427	0.31507	0.316619	0.318831	0.314392	0.315596
$q_O$ (e)	–	–0.89517	–0.8340	–0.8340	–0.8340	–0.82270	–0.82543
$l$ (nm)	0.09572	0.097888	0.09572	0.09572	0.09572	0.09572	0.09572
$\theta$ (deg)	104.52	109.47	104.52	104.52	104.52	104.52	104.52
Observable							
$\Delta H_{vap}$ (kJ/mol)	44.0	44.6*	42.6	41.8	43.3	41.3	43.2
$V_M$ (Å <sup>3</sup> )	30.0	30.1	30.4	30.9	30.5	30.4	29.9
$\bar{r}_{OO,1}$ (Å)	2.97	2.99	3.01	3.02	2.99	3.01	2.98
$\bar{g}_{OO,1}$ (–)	2.14	2.17	2.24	2.22	2.07	2.24	2.12
$\langle r_{HB} \rangle$ (Å)	1.93	1.97	2.03	2.04	1.99	2.04	1.99
$\langle \theta_{HB} \rangle$ (deg)	14.7	15.5	18.1	18.1	17.2	18.4	17.5
$\epsilon_0$ (–)	78.4	78.9	98.3	96.2	92.4	93.1	93.41
$D$ (10 <sup>–5</sup> cm <sup>2</sup> /s)	2.30	2.11	6.76	9.39	4.77	13.0	9.57
$C_p$ (J/molK)	75.3	134	135	136	137	136	137

Posterior samples are summarized by their mean (T3B<sub>mean</sub>) and mode (T3B<sub>mode</sub>). The OPC3 model includes geometry modifications; experimental reference values are provided in the second column. Uncertainty estimates for these properties, including the coefficient of variation, bias, and simulation error, are reported separately in Table 1.

\* corrected according to Berendsen et al.<sup>25</sup> with data from Horn et al.<sup>26</sup>.

four-point models such as GOPAL<sup>24</sup>. It is worth emphasizing, however, that the correction applied to the  $\Delta H_{vap}$  of OPC3 in part relies on quantum mechanical estimates of gas-phase enthalpy<sup>25,26</sup>, which are specific to the system and temperature in question, thereby limiting their general applicability. Importantly, our focus here is not to optimize force field parameters per se, but to quantify the range and source of predictive uncertainty. Extending the present Bayesian framework to include geometric parameters would therefore be a promising next step that could provide further gains in accuracy and insight. A high-dimensional posterior would, however, be tedious to sample and to interpret and is therefore beyond the scope of this work.

The TIP3P model remains one of the most widely used descriptions of liquid water, primarily due to its simplicity and compatibility with biomolecular force fields. However, its empirical design and neglect of polarization and nuclear quantum effects lead to systematic deviations from experimental observables. Comprehensive reviews have shown that earlier rigid, non-polarizable models did not reproduce dielectric properties, diffusion coefficients, and phase-transition behavior accurately, even when tuned to match density and vaporization enthalpy at ambient conditions<sup>2</sup>. More recent parameterizations, such as OPC3<sup>22</sup> and GOPAL<sup>24</sup> have substantially improved the accuracy achievable within this subcategory of models, however. Developments toward many-body and polarizable potentials, such as MB-pol, have achieved higher accuracy at substantially increased computational cost<sup>27</sup>. These studies collectively highlight the trade-off between efficiency and physical realism that still motivates research into error-aware parameterization strategies for simple water models.

It is well established that the TIP3P model does not reproduce certain key properties accurately<sup>28–30</sup>, and our results reflect this limitation (Table 2). These large deviations further underscore the need for error-aware calibration frameworks. When using UT to estimate predictive uncertainty, two assumptions must be acknowledged: (1) the posterior is approximated by its mean and covariance, and (2) the region of posterior mass lies near an optimum in observable space. If the posterior is broad or shifted away from the true optimum, the resulting uncertainty estimates may misrepresent the underlying distribution. Nonetheless, despite these simplifications, the UT approach yields coefficient of variation estimates that are consistent with

those obtained from full MCMC sampling, while requiring significantly fewer simulations. This makes it a practical alternative in high-dimensional, simulation-expensive settings.

Our approach expands on earlier efforts to apply Bayesian methods to molecular simulations. For instance, Cooke and Schmidler<sup>31</sup> emphasized a simulation-as-prediction perspective, treating molecular simulations as statistical models capable of generating experimentally testable predictions, rather than as tools for fitting or visualization alone. Their treatment of FF calibration as a predictive inference task aligns with our aim of describing posterior uncertainty rather than producing a single best-fit parameter set. In parallel, Angelikopoulos et al.<sup>32</sup> introduced a high-performance Bayesian uncertainty quantification framework using transitional MCMC and surrogate models, explicitly quantifying FF uncertainty and its propagation to derived properties—a conceptually similar goal, but relying on surrogate models to approximate simulation outputs, whereas our method samples from a synthetic likelihood based on direct simulations. As Cooke and Schmidler note, using multiple observables “ensures that no particular measurements are well described at the expense of others,” while Angelikopoulos et al. highlight that too few data relative to parameters can lead to “unidentifiable cases”; both points support our finding that diverse observables help constrain the posterior.

In contrast to surrogate-based calibration, our framework shares closer methodological similarities with the BioFF method of Köfinger and Hummer<sup>33</sup>, which also performs inference over FF parameters via iterative simulation and ensemble reweighting. BioFF formulates force field calibration as a Bayesian inference problem and updates parameters by optimizing the agreement between experimental data and reweighted simulation ensembles, using gradient-based optimization to refine the posterior. While BioFF integrates reweighting with gradient optimization, our work avoids ensemble weighting schemes and instead relies on simulation-derived summary statistics that are asymptotically normally distributed, enabling direct likelihood approximation via the SL approach. This tradeoff reflects a different balance between inference tractability and sampling accuracy: while BioFF avoids expensive resampling through ensemble reweighting, its accuracy may suffer if the initial ensemble poorly represents the posterior. In contrast, our method incurs a higher

computational cost by relying on repeated simulations but provides more reliable posterior estimates when summary statistics are asymptotically normally distributed.

A complementary perspective is offered by Imbalzano et al.<sup>34</sup>, who decompose uncertainty in machine learning-driven MD into sampling noise and model error using committee-based learning. Their approach relies on on-the-fly reweighting combined with a cumulant expansion approximation to efficiently propagate uncertainty from model predictions to thermodynamic observables. While methodologically distinct, this goal aligns with our use of the unscented transform (UT) to approximate predictive uncertainty from a parameter posterior. Their results demonstrate that such strategies can yield reliable uncertainty estimates for structural and thermodynamic properties across diverse systems, and can inform robust active learning strategies.

Finally, our results underscore the importance of rigorous statistical error reporting and convergence diagnostics, as advocated by Grossfield et al.<sup>35</sup>. Their emphasis on best practices—including effective sample size estimation, confidence intervals, and equilibration analysis—reflects principles incorporated in the treatment of posterior sampling and sigma point propagation in this work. By adopting these practices, our Bayesian framework not only quantifies FF uncertainty but also aligns with community standards for reproducibility and transparency<sup>36,37</sup> and TRUE principles proposed by Thompson et al.<sup>38</sup>.

Together, these contributions position our SL approach as a bridge between conventional FF optimization and emerging probabilistic methodologies for molecular simulation. Our parameter posteriors reproduce  $\Delta H_{vap}$  and  $V_M$  within 2–3% of their experimental references (Table 1), which is comparable to the best-performing models like OPC3. However, other observables such as the hydrogen bond angle  $\langle \theta_{HB} \rangle$  show relative biases of 17–19%, and thermal response properties such as  $D$  deviate by up to 68%. These magnitudes are consistent with the findings of Cailliez and Pernot<sup>39</sup>, who demonstrated that parameter uncertainty can dominate over statistical sampling noise in molecular simulations. By decomposing the error into parameter uncertainty, model bias, and numerical noise, our framework reveals how even well-constrained parameters may yield poor predictions when the functional form is insufficient. In this way, probabilistic calibration shifts the focus from merely fitting parameters to critically evaluating the adequacy of the model itself, that is, whether the employed functional form and parameter space are sufficient to reproduce the target observables within their experimental uncertainty.

Recent work by Paliwal and Shirts<sup>24</sup> introduced a reweighting-based optimization framework that combines multistate reweighting and configuration mapping to efficiently explore parameter space across wide temperature and pressure ranges. Reweighting approaches such as the multistate Bennett acceptance ratio (MBAR)<sup>40,41</sup> and configuration mapping<sup>24</sup> provide an efficient means of estimating ensemble-averaged observables across nearby parameter or state points, substantially reducing the computational cost when sufficient configuration-space overlap exists. While their approach can be viewed as an implicit likelihood maximization, the resulting uncertainties stem from frequentist error propagation of reweighted observables. In contrast, our Bayesian framework explicitly samples the full posterior via synthetic likelihood and MCMC, yielding a probabilistic description of uncertainty that captures correlations between parameters and observables. Although our current dataset does not include per-frame energies required for reweighting, integrating reweighting techniques within a Bayesian synthetic-likelihood framework represents a promising direction for future work.

## Methods

In the original TIP3P model, water molecules are kept rigid at the experimental gas-phase geometry, and the only contribution to the energy of a system stems from non-bonded interactions. To model these interactions, the Lennard-Jones 12-6 potential is used to describe van der Waals interactions, while Coulombic interactions account for electrostatics. The Lennard-Jones interaction is applied only between oxygen atoms, whereas

electrostatic interactions act between all atomic partial charges. The total non-bonded potential energy between atoms  $i$  and  $j$  is given by:

$$U_{\text{nonbonded}}(r_{ij}) = \begin{cases} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + 4\epsilon_{OO} \left[ \left( \frac{\sigma_{OO}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{OO}}{r_{ij}} \right)^6 \right], & \text{if } i, j \in \{\text{O}\} \\ \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}, & \text{otherwise} \end{cases} \quad (1)$$

Here,  $r_{ij}$  is the distance between atoms  $i$  and  $j$ ,  $\epsilon_0$  is the vacuum permittivity,  $\epsilon_{OO}$  and  $\sigma_{OO}$  are the Lennard-Jones parameters for the oxygen-oxygen interaction, and  $q_i, q_j$  are the partial charges. In three-point water models like TIP3P, the two hydrogens carry positive partial charges, and the oxygen carries a negative charge equal in magnitude to their sum, ensuring overall charge neutrality.

## Integration of experimental observables

A diverse set of observables are incorporated to constrain the inference of non-bonded FF parameters. These are the enthalpy of vaporization ( $\Delta H_{vap}$ ), molecular volume ( $V_M$ ), structural features of the oxygen-oxygen RDF ( $\tilde{r}_{OO,1}$  and  $\tilde{g}_{OO,1}$ ), average hydrogen-acceptor distance in a hydrogen bond ( $\langle r_{HB} \rangle$ ), and average hydrogen-donor-acceptor angle ( $\langle \theta_{HB} \rangle$ ). Together, these observables capture energetic and structural characteristics of liquid water.

$\Delta H_{vap}$  is calculated from the average potential energy per molecule as:

$$\Delta H_{vap} = RT - \frac{U_{pot}}{N_M}, \quad (2)$$

where  $R$  is the gas constant,  $T$  the temperature,  $U_{pot}$  the total potential energy, and  $N_M$  the number of molecules. This relation holds for rigid water models, where the internal degrees of freedom do not contribute to the enthalpy and the kinetic energy is accounted for by the ideal gas term  $RT$ . According to Berendsen et al.<sup>25</sup> the following was used instead for OPC3:

$$\Delta H_{vap} = RT - \frac{U_{pot}}{N_M} + \frac{pV}{N_M} - E_{pot} + C, \quad (3)$$

where  $p$  is the pressure and  $V$  the volume of the simulation box. The polarization term  $E_{pot}$  is approximated as

$$E_{pot} = \frac{(\mu - \mu_{gas})^2}{2\alpha_{gas}}, \quad (4)$$

where  $\mu$  is the dipole moment of the model geometry,  $\mu_{gas}$  and  $\alpha_{gas}$  are the reference dipole moment and polarizability, respectively of the gas phase. The value of these two references, along with the temperature-dependent correction term  $C$  accounting for the non-ideal gas behavior of intramolecular vibrational modes, is provided by Horn et al.<sup>26</sup>.  $V_M$  is computed as:

$$V_M = \frac{V}{N_M}. \quad (5)$$

The RDF could not be used directly, but rather  $\tilde{r}_{OO,1}$  and  $\tilde{g}_{OO,1}$  was used. Specifically, RDF histograms are converted back to raw binned counts independent of molecule count. These counts are integrated radially until the known coordination number is reached, defined via integration to the first minimum, here estimated to be 4.96 molecules per molecule from an X-ray experimental reference RDF by Skinner et al.<sup>42</sup>. Within this shell, the mean and standard deviation of the counts are estimated assuming Gaussian statistics. The RDF is reconstructed from this Gaussian approximation, and the position and height of the resulting peak yield  $\tilde{r}_{OO,1}$  and  $\tilde{g}_{OO,1}$ , respectively. The tilde notation is adopted to emphasize the

difference between these derived metrics and their traditional RDF counterparts. This approach ensures that the structural features used for inference reflect robust solvation shell properties rather than noise from limited sampling. A visual demonstration of this calculation is shown in Fig S1.

Finally, the hydrogen bond properties  $\langle r_{HB} \rangle$  and  $\langle \theta_{HB} \rangle$  are computed as the mean of their respective distributions, obtained from binned distance and angle data. These observables provide a geometric assessment of the model's ability to reproduce hydrogen bonding behavior.

### Bayesian framework for parameter estimation

A full Bayesian framework was implemented to estimate the posterior distribution of non-bonded force parameters in a TIP3P-like water model. The posterior distribution,  $P(\theta|y)$ , is calculated using Bayes' theorem:

$$P(\theta|y) \propto P(y|\theta)P(\theta), \quad (6)$$

where  $\theta$  represents the FF parameters,  $y$  denotes the experimental observables,  $P(y|\theta)$  is the likelihood function, and  $P(\theta)$  is the prior distribution. The likelihood quantifies how well a given parameter set reproduces the data, while the prior encodes pre-existing knowledge or constraints. In this study, uniform priors are used for all parameters, ensuring equal probability across the defined bounds.

### Synthetic likelihood and normality assumption

In cases like this where the likelihood function  $P(y|\theta)$  is analytically intractable—e.g., when using noisy, simulation-derived observables—an SL approach is employed. This constructs a likelihood using a parametric approximation based on simulated data, under the assumption that the observables are asymptotically normally distributed, as described below. The likelihood is then simply a multivariate Gaussian:

$$P(y|\theta) = \frac{1}{(2\pi)^{N_y/2} |\Sigma_y|^{1/2}} \exp\left(-\frac{1}{2}(y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y)\right), \quad (7)$$

where  $N_y$  is the number of observables,  $\mu_y$  is the vector of proposed means for an estimated parameter set  $\theta$ , and  $\Sigma_y$  is the corresponding covariance matrix. This formulation captures both uncertainty and interdependence among observables.

A key assumption of this framework is the asymptotical normality of the observables for each  $\theta$ . To assess this, the Shapiro-Wilk test<sup>19,20</sup> is applied to the individual observable distributions. At a significance threshold of  $p > 0.05$ , the normality test passes. For  $n$  independent observables, the probability that all marginals pass the test under true normality is approximately 0.95<sup>43</sup>. Since the assumption of independence does not hold for the observables in this study, Mardia's test<sup>18</sup> was used to verify multivariate normality. Unlike univariate tests, Mardia's test provides a single joint evaluation whose significance is based on the asymptotic  $\chi^2$  and normal distributions of its skewness and kurtosis components. While the SL method is known to tolerate mild deviations from normality, strong non-Gaussian features can lead to inaccurate posteriors<sup>12</sup>.

When estimating  $\Sigma_y$ , care must be taken to avoid artificial scaling effects related to system size. Additional adjustments are applied depending on whether the observable is defined on a per-molecule or per-molecular-pair basis. Because this distinction cannot be readily determined from the trajectory output, the variance corresponding to the true per-molecule or per-pair property must be estimated explicitly. Both  $\Delta H_{\text{vap}}$  and  $V_M$  are defined per molecule. For a system with  $N_M$  molecules, where each molecule contributes to the potential energy and occupies a volume while being correlated with itself and  $N_C$  neighbors, the variance scales by a factor of  $\frac{N_M}{1+N_C}$ .

In contrast, properties derived from the RDF or hydrogen bond statistics are inherently defined per molecular pair. If, on average, each molecule interacts with  $N_C$  neighbors, there are approximately  $N_M N_C / 2$  unique pairs. Each such pair is assumed to be correlated with itself and with the remaining  $2N_C - 2$  pairs formed by the atoms within the pair, leading to

a total variance scaling factor of  $\frac{N_M N_C / 2}{2N_C - 1}$ . For covariances between per-molecule and per-pair observables, the scaling factor is taken as the square root of the product of the respective variance scaling factors. These corrections ensure that the covariance matrix  $\Sigma_y$  reflects true statistical uncertainty, avoiding distortions due to finite-size effects or local correlations.

### Metropolis–Hastings sampling

To explore the posterior distribution, the MH algorithm<sup>13</sup> is employed. This MCMC method samples from the posterior without requiring its analytical form. Each iteration consists of proposing a new parameter set and accepting or rejecting it based on the likelihood.

New proposals  $\theta'$  are drawn from log-normal distributions to ensure that the non-bonded parameters remain valid:

$$\theta'_i = \eta_i \theta_i \quad \text{with} \quad \eta_i \sim \exp\left(\mathcal{N}\left(0, \sigma_{\theta_i}^2\right)\right), \quad (8)$$

where  $\sigma_{\theta_i}$  controls the spread of the perturbation. The acceptance probability is given by:

$$A(\theta \rightarrow \theta') = \min\left(1, \frac{P(y|\theta')P(\theta')g(\theta' \rightarrow \theta)}{P(y|\theta)P(\theta)g(\theta \rightarrow \theta')}\right). \quad (9)$$

Because the priors are uniform and proposals outside the prior bounds are immediately rejected, the prior terms cancel. To account for the asymmetry of the log-normal proposal, the Hastings ratio is included in the acceptance probability. For a log-normal proposal, the proposal density is:

$$g(\theta_i \rightarrow \theta'_i) = \frac{1}{\theta'_i \sigma_{\theta_i} \sqrt{2\pi}} \exp\left(-\frac{(\log \theta'_i - \log \theta_i)^2}{2\sigma_{\theta_i}^2}\right), \quad (10)$$

and similarly:

$$g(\theta'_i \rightarrow \theta_i) = \frac{1}{\theta_i \sigma_{\theta_i} \sqrt{2\pi}} \exp\left(-\frac{(\log \theta_i - \log \theta'_i)^2}{2\sigma_{\theta_i}^2}\right). \quad (11)$$

Since the squared log difference is symmetric, the exponentials cancel. The Hastings ratio simplifies to:

$$\frac{g(\theta'_i \rightarrow \theta_i)}{g(\theta_i \rightarrow \theta'_i)} = \frac{1/\theta_i}{1/\theta'_i} = \frac{\theta'_i}{\theta_i} = \eta_i. \quad (12)$$

Thus, the multiplicative factor  $\eta_i$  appears directly in the acceptance probability due to the asymmetry of the log-normal proposal, leading to:

$$A(\theta \rightarrow \theta') = \min\left(1, \frac{P(y|\theta')}{P(y|\theta)} \prod_{i=1}^{N_\theta} \eta_i\right), \quad (13)$$

where  $N_\theta$  is the number of parameters. A proposal is accepted if  $A$  exceeds a uniform random number  $u \sim \mathcal{U}(0, 1)$ ; otherwise, the current parameter set remains.

According to Roberts and Rosenthal on MH, “any algorithm with acceptance rate between say 0.15 and 0.5 will be at least 80% efficient”<sup>44</sup>. Individual values of  $\sigma_{\theta_i}$  are selected such that 88% of proposed steps remain within bounds in the two-parameter case and 92% in the three-parameter case. These choices yield an acceptance rate between 20% and 40%, consistent with Roberts and Rosenthal.

The resulting Markov chain converges to the true posterior after a sufficient number of iterations. To identify the burn-in period, the Geweke diagnostic<sup>45</sup> is used, and initial samples are discarded until stationarity is

reached. After doing so, a total of 101987 parameter samples were left in 2D and 133288 in 3D.

### Unscented transform sigma point calculations

To estimate uncertainties in observables without requiring a full posterior sample, the UT<sup>46</sup> is applied. The UT approximates the propagation of uncertainty through nonlinear models by deterministically generating a set of  $2N_\theta + 1$  sigma points, centered around the posterior mean  $\theta_\mu$  and shaped by the posterior covariance  $\Sigma_\theta$ . These sigma points are propagated independently through the simulation model, and the resulting observable statistics are reconstructed via weighted averaging.

The sigma points  $\{\theta^{(i)}\}_{i=0}^{2N_\theta}$  and their associated weights  $W_m^{(i)}$  and  $W_c^{(i)}$  are constructed as follows. First, a scaling factor is computed:

$$\lambda = \alpha^2(N_\theta + \kappa) - N_\theta, \quad (14)$$

where  $\alpha = 0.3$ ,  $\kappa = 3 - N_\theta$  and  $\beta = 2$  are used in this study.

The weights for the mean and covariance are then defined as:

$$W_m^{(0)} = \frac{\lambda}{N_\theta + \lambda}, \quad (15)$$

$$W_c^{(0)} = \frac{\lambda}{N_\theta + \lambda} + (1 - \alpha^2 + \beta), \quad (16)$$

$$W_m^{(i)} = W_c^{(i)} = \frac{1}{2(N_\theta + \lambda)} \quad \text{for } i = 1, \dots, 2N_\theta. \quad (17)$$

To construct the sigma points, the matrix square root of  $(N_\theta + \lambda)\Sigma_\theta$  is computed, e.g., via Cholesky decomposition:

$$\theta^{(0)} = \theta_\mu, \quad (18)$$

$$\theta^{(i)} = \theta_\mu + \left[ \sqrt{(N_\theta + \lambda)\Sigma_\theta} \right]_{:,i} \quad \text{for } i = 1, \dots, N_\theta, \quad (19)$$

$$\theta^{(N_\theta+i)} = \theta_\mu - \left[ \sqrt{(N_\theta + \lambda)\Sigma_\theta} \right]_{:,i} \quad \text{for } i = 1, \dots, N_\theta. \quad (20)$$

Each sigma point corresponds to either two or three modified parameters in the 2D or 3D case respectively. The models with these updated parameters are then simulated independently to compute observables. For each point  $i$ , a set of block-averaged observable vectors  $\mathbf{Y}_b^{(i)}$  is obtained over  $N_b$  time blocks. The final observable mean and variance are estimated by:

$$\hat{\mu}_Y = \sum_{i=0}^{2N_\theta} W_m^{(i)} \left( \frac{1}{N_b} \sum_{b=1}^{N_b} \mathbf{Y}_b^{(i)} \right), \quad (21)$$

$$\hat{\sigma}_Y^2 = \sum_{i=0}^{2N_\theta} W_c^{(i)} \left( \frac{1}{N_b} \sum_{b=1}^{N_b} \mathbf{Y}_b^{(i)} - \hat{\mu}_Y \right)^2. \quad (22)$$

To estimate the statistical error in the mean, the standard deviation of the block-averaged observable is computed (aggregated using the mean weights  $W_m$ ), and multiplied by  $N_b^{-1/2}$ :

$$\text{SE}_Y = N_b^{-1/2} \cdot \text{StdDev} \left( \sum_{i=0}^{2N_\theta} W_m^{(i)} \mathbf{Y}_b^{(i)} \right). \quad (23)$$

This approach provides an efficient estimate of both the mean and uncertainty in the observables without requiring full posterior sampling. The Unscented Transform assumes that the underlying variables are approximately locally normal, allowing mean and covariance to be propagated through nonlinear mappings using a finite set of sigma points. However, this

assumption need not be exact: the transform captures the mean and covariance correctly up to the second order for arbitrary distributions and up to the third order when the variable is Gaussian, with errors entering only at the fourth order and higher in the Taylor expansion of the nonlinear function<sup>46</sup>. Consequently, the UT provides an accurate local linear estimate even when the posterior deviates moderately from perfect normality.

### Quantification of Errors

Beyond estimating parameters, the uncertainty in both the inference observables and additional validation properties is assessed. By using the observables  $Y$  corresponding to the sampled parameters of the posterior, the coefficient of variation of the model can be computed as

$$\text{CV}(Y) = \frac{\sqrt{\text{Var}(Y)}}{E[Y]}, \quad (24)$$

the relative bias as

$$\text{RB}(Y, y) = \frac{y - E[Y]}{E[Y]}, \quad (25)$$

for a point value  $y$ , considered known, and the relative standard error as

$$\text{RSE}(Y) = \frac{s_B / \sqrt{N_B}}{E[Y]}, \quad (26)$$

where  $s_B$  is the block sample standard deviation of  $N_B$  blocks. This allows us to evaluate how sensitive the observables are to parameter variations and to identify systematic deviations.

### Computational details

The grid point as well as the MCMC simulations were performed using OpenMM<sup>47</sup> with a 1 fs integration time step on CUDA-enabled GPUs. Each system was energy-minimized for 1000 steps, followed by equilibration for 1 ns in the NVT ensemble and 1 ns in the NpT ensemble. Production simulations lasted 100 ps, with observables sampled every 1 ps. Simulations for the parameter grid in Fig. 1 employed 2000 water molecules, while those used in MCMC sampling in Figures S15, S16, 2, and 3 were reduced to 500 molecules for computational efficiency. Temperature was maintained at 298.15 K using a Nosé-Hoover chain thermostat<sup>48</sup>, and pressure was controlled at 1 atm using a Monte Carlo barostat<sup>49</sup> applied every 25 steps. Water molecules were treated as rigid<sup>50</sup>. Non-bonded interactions were handled using the particle mesh Ewald<sup>51</sup> method with a 0.8 nm cutoff and an Ewald error tolerance of  $10^{-4}$ . Lennard-Jones interactions were truncated at the same distance, with long-range dispersion corrections enabled. Measurements for  $\Delta H_{\text{vap}}$  and  $V_M$  were extracted directly from OpenMM outputs. RDFs and hydrogen bond statistics were computed using GROMACS<sup>52</sup>. Utilities from the Alexandria Chemistry Toolkit<sup>53</sup> are used to manage force field files.

During MCMC sampling, parameter proposals were applied on-the-fly within a continuous simulation workflow. Each new parameter set was introduced by loading a checkpoint containing the atomic coordinates and velocities from the previously accepted state. This avoided unnecessary repetition of minimization and equilibration, as most proposals resulted in only minor perturbations to the system. To ensure the system remained equilibrated after each parameter update, the final 100 ps of the trajectory were compared to the preceding 100 ps using three statistical tests: the Mann-Whitney  $U$  test<sup>54</sup>, the Kolmogorov-Smirnov test<sup>55</sup>, and the Student's  $t$ -test<sup>56</sup>. These were applied to the distributions of potential energy and box volume. If all  $p$ -values indicated no significant difference ( $p > 0.05$ ), the most recent 100 ps segment was used to compute observable means and covariances for likelihood evaluation. Otherwise, the simulation was extended by an additional 100 ps, repeating the comparison until equilibrium was confirmed. If a proposal was rejected—either due to its posterior probability

or simulation instability—a backup checkpoint of the last accepted state was used to restore the system before continuing the MCMC chain.

Simulations corresponding to sigma points from UT were performed using GROMACS<sup>52</sup>, which provides convenient access to built-in estimators for the validation properties. The consistency between GROMACS and OpenMM implementations was verified (see Fig. S17). Each system was energy-minimized using the steepest descent algorithm, followed by equilibration in the NVT ensemble for 1 ns and in the NpT ensemble for an additional 1 ns. Production simulations were run for 50 ns with a 1 fs time step. Temperature was maintained at 298.15 K using the Nosé-Hoover thermostat with a coupling time constant of 0.2 ps, and pressure was controlled at 1 atm using the isotropic Parrinello-Rahman barostat with a coupling time of 2.0 ps. All bonds involving hydrogen were constrained using the LINCS<sup>57</sup> algorithm, and water molecules were treated as rigid<sup>50</sup>. Non-bonded interactions were handled using the Verlet cutoff scheme with Lennard-Jones and Coulomb cutoffs at 0.8 nm. Long-range electrostatics were treated using the particle mesh Ewald<sup>51</sup> method with a Fourier spacing of 0.12 nm and a real-space tolerance of  $10^{-4}$ . Analytical long-range dispersion corrections were applied to energy and pressure.

RDF features, hydrogen bonding metrics, and validation observables were computed using GROMACS analysis tools. The diffusion coefficient and dielectric constant were obtained from the mean-squared displacement (msd) and total dipole moment (dipoles) analyses of the NpT production trajectories, respectively, rather than from separate NVE simulations. These analyses were performed on the full trajectory and over 10 ns blocks to estimate statistical means and uncertainties. For the posterior mean parameter set, an extended 200 ns simulation was carried out and similarly divided into 10 ns blocks to assess statistical error.

## Data availability

Data and code for Bayesian inference of TIP3P-like water models are available on GitHub: <https://github.com/pastaalfredo/Data>.

## Code availability

Data and code for Bayesian inference of TIP3P-like water models are available on GitHub<sup>58</sup>.

Received: 23 May 2025; Accepted: 17 November 2025;

Published online: 30 November 2025

## References

- Frenkel, D. & Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*. Elsevier, Amsterdam (2023)
- Vega, C. & Abascal, Jose L. F. Simulating water with rigid non-polarizable models: a general perspective. *Phys. Chem. Chem. Phys.* **13**, 19663–19688 (2011).
- Ahmadabadi, I., Esfandiari, A., Hassanali, A. & Ejtehadi, M. R. Structural and dynamical fingerprints of the anomalous dielectric properties of water under confinement. *Phys. Rev. Mater.* **5**, 024008 (2021).
- Bellissent-Funel, M.-C. et al. Water determines the structure and dynamics of proteins. *Chem. Rev.* **116**, 7673–7697 (2016).
- Jia, R. et al. Hydrogen-deuterium exchange mass spectrometry captures distinct dynamics upon substrate and inhibitor binding to a transporter. *Nat. Commun.* **11**, 6162 (2020).
- Mustafa, G., Nandekar, P. P., Mukherjee, G., Bruce, N. J. & Wade, R. C. The effect of force-field parameters on cytochrome p450-membrane interactions: structure and dynamics. *Sci. Rep.* **10**, 7284 (2020).
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
- Kadaoluwa Pathirannahalage, S. P. et al. Systematic comparison of the structural and dynamic properties of commonly used water models for molecular dynamics simulations. *J. Chem. Inf. Model.* **61**, 4521–4536 (2021).
- Jorgensen, W. L. & Jenson, C. Temperature dependence of TIP3P, SPC, and TIP4P water from npt Monte Carlo simulations: Seeking temperatures of maximum density. *J. Comp. Chem.* **19**, 1179–1186 (1998).
- Box, G. E. & Tiao, G. C. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, Hoboken, NJ, USA (2011)
- Cailliez, F., Bourasseau, A. & Pernot, P. Calibration of forcefields for molecular simulation: Sequential design of computer experiments for building cost-efficient kriging metamodels. *J. Comp. Chem.* **35**, 130–149 (2014).
- Wood, S. N. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**, 1102–1104 (2010).
- Chib, S. & Greenberg, E. Understanding the Metropolis-Hastings algorithm. *Am. Stat.* **49**, 327–335 (1995).
- Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL (1995)
- Shanks, B. L., Sullivan, H. W., Shazed, A. R. & Hoepfner, M. P. Accelerated Bayesian inference for molecular simulations using local Gaussian process surrogate models. *J. Chem. Theory Comput.* **20**, 3798–3808 (2024).
- Mobley, D. L. Let's get honest about sampling. *J. Comput. -Aided Mol. Des.* **26**, 93–95 (2012).
- van der Spoel, D., Zhang, J. & Zhang, H. Quantitative predictions from molecular simulations using explicit or implicit interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **12**, 1560 (2022).
- Mardia, K. V. Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530 (1970).
- Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611 (1965).
- Villasenor Alva, J. A. & Estrada, E. G. A generalization of Shapiro-Wilk's test for multivariate normality. *Commun. Stat. - Theory Methods* **38**, 1870–1883 (2009).
- Berens, P. H., Mackay, D. H. J., White, G. M. & Wilson, K. R. Thermodynamic and quantum corrections from molecular dynamics for liquid water. *J. Chem. Phys.* **79**, 2375–2389 (1983).
- Izadi, S. & Onufriev, A. V. Accuracy limit of rigid 3-point water models. *J. Chem. Phys.* **145**, 074501 (2016).
- Jorgensen, W. L. Transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *J. Am. Chem. Soc.* **103**, 335–340 (1981).
- Paliwal, H. & Shirts, M. R. Gopal: A water model with improved thermodynamic properties over a large pressure and temperature range optimized using multistate reweighting and configuration space mapping. *J. Chem. Theory Comput.* (2025)
- Berendsen, H. J. C., Grigera, J.-R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
- Horn, H. W. et al. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **120**, 9665–9678 (2004).
- Paesani, F. Getting the right answers for the right reasons: Toward predictive molecular simulations of water with many-body potential energy functions. *Acc. Chem. Res.* **49**, 1844–1851 (2016).
- van der Spoel, D., van Maaren, P. J. & Berendsen, H. J. C. A systematic study of water models for molecular simulation. *J. Chem. Phys.* **108**, 10220–10230 (1998).
- Mark, P. & Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *J. Phys. Chem. A* **105**, 9954–9960 (2001).
- Vega, C., Abascal, Jose L. F., Conde, M. M. & Aragoes, J. L. What ice can teach us about water interactions: a critical comparison of the performance of different water models. *Faraday Discuss.* **141**, 251–276 (2009).

31. Cooke, B. & Schmidler, S. C. Statistical prediction and molecular dynamics simulation. *Biophys. J.* **95**, 4497–4511 (2008).
32. Angelikopoulos, P., Papadimitriou, C. & Koumoutsakos, P. Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework. *J. Chem. Phys.* **137**, 144103 (2012).
33. Köfinger, J. ürgen & Hummer, G. Empirical optimization of molecular simulation force fields by Bayesian inference. *Eur. Phys. J. B* **94**, 245 (2021).
34. Imbalzano, G. et al. Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **154**, 074102 (2021).
35. Grossfield, A. et al. Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1. 0]. *Living J. Comput. Mol. Sci.* **1**, 5067 (2018).
36. The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **16**, 670–673 (2019).
37. Amaro, R. E. et al. The need to implement fair principles in biomolecular simulations. *Nat. Methods* **22**, 641–645 (2025).
38. Thompson, M. W. et al. Towards molecular simulations that are transparent, reproducible, usable by others, and extensible (true). *Mol. Phys.* **118**, 1742938 (2020).
39. Cailliez, F. & Pernot, P. Statistical approaches to forcefield calibration and prediction uncertainty in molecular simulation. *J. Chem. Phys.* **134**, 054124 (2011).
40. Paliwal, H. & Shirts, M. R. Using multistate reweighting to rapidly and efficiently explore molecular simulation parameters space for nonbonded interactions. *J. Chem. Theory Comput.* **9**, 4700–4717 (2013).
41. Naden, L. N. & Shirts, M. R. Rapid computation of thermodynamic properties over multidimensional nonbonded parameter spaces using adaptive multistate reweighting. *J. Chem. Theory Comput.* **12**, 1806–1823 (2016).
42. Skinner, L. B. et al. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide Q-range. *J. Chem. Phys.* **138**, 074506 (2013).
43. Razali, N. M. et al. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Anal.* **2**, 21–33 (2011).
44. Roberts, G. O. & Rosenthal, J. S. Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.* **16**, 351–367 (2001).
45. Geweke, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Federal Reserve Bank of Minneapolis (1991)
46. Julier, S., Uhlmann, J. & Durrant-Whyte, H. F. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Trans. Autom. Control* **45**, 477–482 (2000).
47. Eastman, P. et al. OpenMM 8: molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B* **128**, 109–116 (2023).
48. Martyna, G. J., Klein, M. L. & Tuckerman, M. Nosé–hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635–2643 (1992).
49. Åqvist, J., Wennerström, P., Nervall, M., Bjelic, S. & Brandsdal, Bjørn O. Molecular dynamics simulations of water and biomolecules with a monte carlo constant pressure algorithm. *Chem. Phys. Lett.* **384**, 288–294 (2004).
50. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
51. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. A smooth particle mesh ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
52. Abraham, M. J. et al. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
53. van der Spoel, D. et al. Evolutionary machine learning of physics-based force fields in high-dimensional parameter-space. *Digit. Discov.* **4**, 1925–1935 (2025).
54. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 50–60 (1947)
55. Smirnov, N. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* **19**, 279–281 (1948).
56. Student. The probable error of a mean. *Biometrika*, 1–25 (1908).
57. Hess, B., Bekker, H., Berendsen, H. J. & Fraaije, J. G. LINC: A linear constraint solver for molecular simulations. *J. Comp. Chem.* **18**, 1463–1472 (1997).
58. Nordman, A. T. Data and Code for Bayesian Inference of TIP3P-like Water Models. <https://github.com/pastaalfredo/Data>. Accessed: 2025-10-24 (2025)
59. Haynes, W. M. CRC Handbook of Chemistry and Physics. CRC press, Boca Raton, FL (2016)
60. Modig, K., Pfrommer, B. G. & Halle, B. Temperature-dependent hydrogen-bond geometry in liquid water. *Phys. Rev. Lett.* **90**, 075502 (2003).
61. Mills, R. Self-diffusion in normal and heavy water in the range 1–45. deg. *J. Phys. Chem.* **77**, 685–688 (1973).

## Acknowledgements

This research was supported financially by the project AI4Research at Uppsala University, Sweden, and by the Swedish Research Council (grants 2020-05059, 2024-04314). Funding from eSSENCE - The e-Science Collaboration (Uppsala-Lund-Umeå, Sweden) is gratefully acknowledged. Additional funding and support were provided by the Centre for Interdisciplinary Mathematics (CIM) at Uppsala University. Computer resources provided by the National Academic Infrastructure for Supercomputing Sweden at the PDC Center for High Performance Computing, KTH Royal Institute of Technology, Sweden, partially funded by the Swedish Research Council through (grant 2022-06725). We also acknowledge the Molecular Biophysics program at Uppsala University for providing access to local computational resources.

## Author contributions

Dvd.S., S.E. and A.N. designed the study. A.N. coded the software, performed all calculations, compiled the results and wrote the manuscript draft. Dvd.S. and S.E. contributed to analysis and writing.

## Funding

Open access funding provided by Uppsala University.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41524-025-01879-w>.

**Correspondence** and requests for materials should be addressed to David van der Spoel.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025