

## PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.

<http://hdl.handle.net/2066/133224>

Please be advised that this information was generated on 2018-01-01 and may be subject to change.

## Predictions for water clusters from a first-principles two- and three-body force field

Urszula Góra, Wojciech Cencek, <sup>Rafał</sup> Podeszwa<sup>1</sup>, Ad van der Avoird, and Krzysztof Szalewicz

Citation: *The Journal of Chemical Physics* **140**, 194101 (2014); doi: 10.1063/1.4875097

View online: <http://dx.doi.org/10.1063/1.4875097>

View Table of Contents: <http://aip.scitation.org/toc/jcp/140/19>

Published by the *American Institute of Physics*

---

### Articles you may be interested in

[On the accuracy of the MB-pol many-body potential for water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice](#)

*The Journal of Chemical Physics* **145**, 194504194504 (2016); 10.1063/1.4967719

---

**COMPLETELY**

**REDESIGNED!**



**PHYSICS  
TODAY**

*Physics Today* Buyer's Guide  
Search with a purpose.

# Predictions for water clusters from a first-principles two- and three-body force field

Urszula Góra,<sup>1,2</sup> Wojciech Cencek,<sup>2</sup> Rafał Podeszwa,<sup>1,2,a)</sup> Ad van der Avoird,<sup>3</sup> and Krzysztof Szalewicz<sup>2</sup>

<sup>1</sup>*Institute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland*

<sup>2</sup>*Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716, USA*

<sup>3</sup>*Theoretical Chemistry, Institute for Molecules and Materials, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands*

(Received 16 March 2014; accepted 24 April 2014; published online 15 May 2014)

A new rigid-monomer three-body potential has been developed for water by fitting it to more than 70 thousand trimer interaction energies computed *ab initio* using coupled-cluster methods and augmented triple-zeta-quality basis sets. This potential was used together with a modified form of a previously developed two-body potential and with a polarization model of four- and higher-body interactions to predict the energetics of the water trimer, hexamer, and 24-mer. Despite using the rigid-monomer approximation, these predictions agree better with flexible-monomer benchmarks than published results obtained with flexible-monomer force fields. An unexpected finding of our work is that simple polarization models predict four-body interactions to within a few percent, whereas for three-body interactions these models are known to have errors on the order of 50%. © 2014 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4875097>]

## I. INTRODUCTION

Water has always been attracting significant attention of theorists due to its abundance and importance for life, but also since the water monomer is a relatively small molecule so that reasonably accurate *ab initio* calculations of interaction energies could be performed. The water dimer in particular was the benchmark system for comparing the performance of various theoretical methods<sup>1–7</sup> at selected points on the potential energy surface. To connect to experiments, one needs a complete potential in order to perform nuclear dynamics calculations. Such calculations for water clusters and for condensed phases of water predict observable properties such as spectra, dissociation energies, radial distribution functions, etc. The Holy Grail of theory is to develop a universal potential that can correctly predict the properties of all forms of water: from dimer to condensed phases, i.e., the predictions should be sufficiently accurate to be meaningfully confronted with experiment.

One way of obtaining such a potential for water is to fit it in molecular dynamics simulations to reproduce as closely as possible experimental data. The potentials of this type are called empirical potentials and some well-known examples are the TIP4P<sup>8</sup> and SPCE<sup>9</sup> potentials. Since these potentials include three- and higher-body interactions in an effective way via pairwise-only terms, such empirical potentials do not work well for small clusters.<sup>10</sup> Modern empirical potentials are also fitted simultaneously to *ab initio* data on clusters<sup>11,12</sup> and have a polarization term which partly accounts for pairwise nonadditive interactions. Thus, such potentials may perform better on small clusters, but still will probably not be

able to produce sufficiently accurate results since simple polarization approximations can recover only about half of the three-body interaction energy for liquid water.<sup>13</sup> Another type of empirical potential can be obtained by fits to water dimer spectra.<sup>14–17</sup> Such potentials represent the water dimer very well, but since the fitting is done purely to the dimer properties, these potentials cannot provide any information about pairwise nonadditive interactions critical for water clusters and condensed phases.<sup>18</sup> Potentials of this type have been used with the polarization model of nonadditive effects, but this model had to be added *post factum* to the two-body potentials. One can also add an *ab initio* three-body nonadditive potential, but to our knowledge this option has not been tried yet.

The other way of developing a potential for water is to fit it to *ab initio* computed interaction energies. Pioneering work of this type was performed by Clementi and collaborators,<sup>1,19,20</sup> but since only a few hundred grid points could be computed for the water dimer and trimer at that time, the sampling of the surface was hardly adequate. The number of grid points was significantly increased in the next generation of first-principles potentials<sup>21–28</sup> to 2.5 thousand points for the dimer<sup>25</sup> and 7.5 thousand points for the trimer.<sup>28</sup> With further improvements of the dimer potential, the goal of correctly predicting properties of all forms of water, from the water dimer to liquid water, was achieved in Refs. 29–31. A good example of how theory can lead experiment in this field is the dissociation energy,  $D_0$ , of the water dimer. In 2000, the first *ab initio* prediction<sup>26</sup> gave the value of 1067 cm<sup>-1</sup>, while later improved calculations in 2008<sup>31</sup> and 2009<sup>32</sup> gave 1111 and 1104 cm<sup>-1</sup>, respectively. This quantity was accurately measured only in 2011<sup>33</sup> and the result of 1105 ± 10 cm<sup>-1</sup> agreed very well with prior theoretical predictions. Despite this striking agreement on  $D_0$ , there is still

<sup>a)</sup> Author to whom correspondence should be addressed. Electronic mail: rafal.podeszwa@tiger.chem.uw.edu.pl

need for improvements of water potentials, in particular in the pairwise nonadditive part. An especially challenging subject are the anomalous properties of liquid water such as the high boiling temperature, anomalous density-temperature dependence, high dielectric constant, and many others. Another possible peculiarity to investigate is the existence of a liquid-liquid critical point in supercooled water.<sup>34</sup> Molecular simulations aimed at predicting these properties are very sensitive to the quality of the applied intermolecular potentials.

First-principles potentials are commonly based on the many-body expansion of the  $N$ -body interaction energy

$$E_{\text{int}} = E_{\text{int}}[2, N] + E_{\text{int}}[3, N] + \dots + E_{\text{int}}[N, N], \quad (1)$$

where  $E_{\text{int}}[2, N]$  is the sum of pair interactions in the  $N$ -body cluster and  $E_{\text{int}}[K, N]$  with  $K > 2$  are the pairwise-nonadditive  $K$ -body contributions. This expansion utilizes the so-called vertical interaction energies, i.e., the energies relative to the monomer energies at the same geometries as in the  $N$ -body cluster. One can add to this expansion the one-body term which is the difference between these monomer energies and the sum of the energies of monomers at their equilibrium-geometries ( $r_e$ ). The extended expansion defines then the so-called relaxed interaction energy.<sup>35</sup> One should note that the latter quantity is essentially the total electronic energy of the complex, only shifted by a constant. All the interaction energies in the present paper will be vertical unless noted otherwise. The expansion of Eq. (1) converges sufficiently fast to eliminate, except in very rare cases, any need to deal with more than a few initial terms for any given system size  $N$ . A fundamental advantage of *ab initio* potentials is the straightforward separation of the interaction energies into the  $K$ -body contributions. Empirical models, on the other hand, are fitted to bulk physical properties and if the higher-body terms are somehow effectively approximated through a two-body potential, properties such as second virial coefficients that depend only on the physical pair interactions, are described poorly. In water, three-body effects are absolutely essential and contribute as much as 16% to the liquid energy.<sup>36</sup> In some clusters, this percentage is even larger, for instance, 23% in the 24-mer.<sup>37</sup> Also the four-body effects cannot be neglected in high-accuracy calculations, as their contribution is about 1% in the hexamer and 3% in the 24-mer.<sup>37</sup> Higher than four-body effects typically account for only a few tenths of a percent, but still can become relevant in problems such as establishing the relative energetic ordering of close-lying cluster structures. Most of the *ab initio* calculations leading to water potentials have been devoted to the first term in the expansion (1), recent work are Refs. 22–26, 29–32, 38–52. The number of papers devoted to the second term is much smaller, only Refs. 20, 28, 36, 46, 47, and 53. No four-body potentials exist for any system, even an atomic one.

The reason that it is very difficult to develop higher terms in the many-body expansion from first principles is the so-called “dimensionality curse.” The successive terms  $E_{\text{int}}[K, N]$  of the expansion of Eq. (1) are functions of  $3KL - 6$  relative coordinates, where  $L$  is the number of atoms in the monomer. If these functions are to be obtained by an analytic fit to calculated *ab initio* energies, the problem quickly becomes intractable because the number of dimensions pre-

cludes any reasonable coverage of the total space with calculated data points. A solution is to calculate the interaction energies on-the-fly for any geometry generated in nuclear dynamics simulations, instead of producing global analytic functions. Obviously, only low-level *ab initio* methods are sufficiently fast for this purpose, which limits the predictive power of this approach. Therefore, the current state-of-the-art in accurate predictions for water is to employ two- and three-body potentials fitted to *ab initio* interaction energies and approximate the higher-body terms by polarization models which account for the asymptotic induction component (by computing the Coulomb interactions between the permanent multipole moments and the induced multipole moments of the monomers, see Sec. V). To our knowledge, there were no published investigations of how well this approximation works for four- and higher-body effects (the present work will provide such information). As already mentioned, simple polarization models recover only about 50% of the total three-body contribution in liquid water.<sup>13</sup> This shows that one cannot avoid construction of first-principles three-body potentials for accurate water simulations.

A two-body potential with rigid monomers is six-dimensional, whereas the inclusion of the intramonomer degrees of freedom results already for triatomic monomers in a 12-dimensional potential. With current computational power, it is possible to represent reasonably well the 12-dimensional surface by a set of grid points,<sup>41,44,48,52</sup> although as many as 250 000 such points may be required.<sup>41</sup> On the other hand, three-body flexible-monomer potentials are 21-dimensional. Generation of such fits would require calculations of  $2.1 \times 10^6$  data points using a mere 2 points per dimension and as many as  $10 \times 10^9$  data points with 3 points per dimension. The flexible-monomer three-body potentials of Wang *et al.*, fitted to 30 000 points<sup>46</sup> or 40 000 points,<sup>47</sup> use only about 1.6 points per dimension. For the six-dimensional rigid-monomer water dimer, this number of points per dimension would result in a total of only 20 grid points, clearly an inadequate number. Since it is unclear whether 30 000–40 000 points are adequate for the water trimer with flexible monomers, we have decided to use the rigid-monomer approximation, as assumed in Ref. 28. With the 12 resulting degrees of freedom, the about 70 thousand grid points that we have used in our computations correspond to about 2.5 points per dimension. Whereas such sampling still seems barely adequate, this number of points is one order of magnitude larger than used in the development of the rigid-monomer SAPT-3B nonadditive three-body water potential in Ref. 28. We will analyze our results to shed light on the question whether it is more beneficial, with a given number of grid points, to obtain a more accurate rigid-monomer potential or a possibly less accurate flexible-monomer one. It should be noted that the choice of an optimal rigid-monomer geometry is crucial: it has been established that the use of the average geometry in the lowest rovibrational state leads to much more accurate results than the use of the equilibrium geometry of the monomers.<sup>4,54</sup> The rigid-monomer potentials have some obvious limitations, for example, one cannot use such potentials to predict the shifts of intramonomer rovibrational frequencies upon complexation. However, the class of problems where such potentials work

well is broad and we will demonstrate this in particular for the structure and energetics of water clusters.

The rigid-monomer SAPT-3B three-body potential of Mas *et al.*<sup>28</sup> was based on 7533 data points computed using symmetry-adapted perturbation theory (SAPT)<sup>55–59</sup> at the level equivalent to the Hartree-Fock (HF) method and a moderate-size  $[5s3p2d1f/3s2p]$  basis set. The computed non-additive interaction energies were fitted to a physically motivated analytic formula containing representations of the short-range exchange contributions and damped induction terms of the same form as in polarization models. This three-body potential was initially combined with the two-body SAPT-5s potential from Ref. 25 (this combination was denoted as SAPT-5s+3B) and then with the CC-pol-5s potential of Refs. 29 and 30 (CC-pol-5s+3B). The latter two-body potential was fitted to dimer interaction energies computed using the coupled-cluster method with single, double, and non-iterative triple excitations [CCSD(T)] extrapolated to the complete basis set (CBS) limit. The CC-pol-5s+3B potential was used to predict trimer spectra in Ref. 60 and achieved very good agreement with experiment, much better than in the case of other potentials. An early version of the three-body potential, restricted to the trimer tunneling path, was used to predict trimer spectra in Ref. 26. These strictly two- plus three-body potentials were extended by adding a polarization model describing four- and higher-body nonadditive effects (such potentials were denoted by SAPT-5s+NB and CC-pol-5s+NB) and used in simulations of liquid water in Refs. 29, 31, and 36. As already mentioned, the work of Refs. 29 and 31 provided a uniformly accurate description of all forms of water.

The project described in the present paper started from an application of the CC-pol-8s potential of Ref. 43 combined with the three-body potential of Ref. 28 and higher-body polarization effects (CC-pol-8s+NB) in water-cluster calculations of Ref. 37. CC-pol-8s was fitted to the same set of interaction energies as CC-pol-5s, but uses a more elaborate functional form with 8 rather than 5 symmetry-unique sites per monomer. CC-pol-8s is still the most accurate rigid-monomer two-body water potential available. The abbreviation “pol” reflects the fact that a self-consistent two-body polarization term is explicitly included in the potential. When predictions of the CC-pol-8s+NB potential were compared to *ab initio* decompositions of cluster energies, it was found that the three-body contribution clearly dominates the overall error (with respect to benchmark *ab initio* results). This was not surprising in view of the fact that a rather limited number of data points and a moderate level of theory were used in Ref. 28. Therefore, the primary aim of the present project was to develop a significantly more accurate rigid-monomer three-body potential by calculating an order of magnitude more points at a much higher level of theory. This development is described in Secs. II, III, and VI. The functional form of the fit has also been significantly changed compared to that of Ref. 28. In particular, a more sophisticated polarization model was developed and optimized to partly reproduce four-body effects, see Sec. V. To use consistently the same polarization model for all  $K$ -body terms, we have refitted the CC-pol-8s<sup>43</sup> two-body potential. Furthermore, as reported in Sec. IV, we used additional *ab initio* data points computed in Ref. 13 to

improve the accuracy in the repulsive wall region and introduced a very short-distance damping of site-site functions. Section VII describes applications of the complete new  $N$ -body model to the water trimer, hexamer, and 24-mer. The new potential was also used to calculate the trimer spectrum, this work will be described in a separate paper.

The combination of the two-body and three-body potentials described above could be performed in several ways. In fact, since the CC-pol- $l$ s potentials are polarizable, one could iterate the polarization model over all monomers in an  $N$ -body cluster, thus approximating the pairwise nonadditive interaction energies by polarization terms only. Another option is to use a straight sum of the two-body and three-body potentials, CC-pol- $l$ s+3B, equivalent to a truncation of the expansion of Eq. (1) after the second term. Finally, one can add to the CC-pol- $l$ s+3B potential higher than three-body effects by iterating the polarization model over all  $N$  monomers and subtracting from the result the two- and three-body polarization components, which leads to the CC-pol- $l$ s+NB potentials. To avoid confusion with previous work and to simplify the notation, we introduce here a new nomenclature for the potentials developed in this work. The names are composed of the stem “CCpol” followed by the digits “2” and/or “3” (depending on which  $K$ -body terms are present), optionally followed by a plus sign if higher-body effects are treated by the polarization model. Thus, “CCpol2” stands for the two-body potential only (including the two-body polarization) and is the only possible choice in the case of the water dimer or it can be used to determine purely two-body effects in larger clusters or in the bulk. Similarly, “CCpol3” stands for the three-body pairwise nonadditive potential. “CCpol2+” adds higher-than-two-body polarization effects (but not the complete three-body potential). Similarly, “CCpol23” will denote a pure two-plus three-body potential, whereas “CCpol23+” includes polarization effects beyond the three-body level and is our most complete force field for systems larger than the trimer.

## II. CHOICE OF TRIMER CONFIGURATIONS

The number and choice of the trimer configurations (grid points) that are used in the fitting process significantly impact the quality of the resulting fit. Since the number of points is always limited by the costs of *ab initio* calculations, the optimal selection of such points is critical. The configurations used in Ref. 28 served as an initial guide, and in particular we calculated nonadditive interaction energies for all 7533 grid points from Ref. 28. We have used the same  $\langle r \rangle_0$  (averaged over the ground rovibrational state) rigid-monomer geometry as Ref. 28, which originates from Ref. 4.

Since the set of Ref. 28 contained mostly trimers with intermonomer separations close to those in trimer’s equilibrium structure, we first augmented it with 7315 geometries with larger separations, selected from trimer configurations present in water clusters from the tetramer to the 21-mer taken from the Cambridge Cluster Database.<sup>61</sup> These geometries were optimized using the TIP5P<sup>62</sup> potential. For each cluster of size  $N$ , all possible  $\binom{N}{3}$  trimers were generated. The rigid-monomer geometry in TIP5P is different from ours, therefore we have placed our monomers in these trimers in such a way



that the centers of mass (COM), the bisectors of the HOH angle, and obviously the molecular planes coincide with the TIP5P monomers.

To ensure that we do not completely miss some regions of configuration space, we generated another set of grid points randomly, with the sampling of COM-COM separations  $R$  restricted to the range 2.6–7 Å. To avoid placing grid points close to the existing ones, we created a sorted list of distances between all the atoms in the trimers and compared them one by one. We defined a “distance” between trimers using an extension of formulas (21) and (22) from Ref. 63. We selected 30 000 points that were farthest from the previously chosen ones. Although the algorithm is not fully permutationally invariant, so that the maximum distance criterion is not satisfied, it removes all duplicates and we checked by inspection that it removed most of the trimers similar to the existing ones.

To improve the description of the region relevant for liquid water, 8520 additional points were selected from snapshots of a converged molecular dynamics (MD) simulation at ambient conditions performed in Ref. 13. This was done in the same way as described in Sec. IV of that reference except that 71 snapshots spaced by 5 ps were used.

Since one of the intended applications of our potential are calculations for the water hexamer, a subject of significant recent interest,<sup>53,64,65</sup> a number of trimer configurations were taken from hexamer structures. First, a set of 14 500 trimer configurations was selected from snapshots of quantum diffusion Monte Carlo simulations performed by us for the hexamer with the CC-pol-8s two-body potential, an early version of the three-body potential developed in the present work, and the CC-pol-8s polarization model for higher-body effects. This set should improve the description of regions relevant for the rovibrational motions in the hexamer and other water clusters. To improve the description of the regions near the hexamer local minima, another set of configurations was generated as follows. First, we optimized the geometries of the cage, prism, book, boat, bag, and ring isomers using the CC-pol-8s+NB potential, starting from the configurations taken from Ref. 66. Since the latter configurations included flexible monomers, we “projected” our monomers similarly as in the case of the TIP5P potential, except that our rigid-monomer bisector now coincides with the line connecting the position of the oxygen atom with the midpoint of the segment connecting the two hydrogens. We will refer to cluster geometries produced in this way as geometries with “rigidized” monomers. The geometries were optimized using a simple Powell<sup>67</sup> algorithm, changing all six coordinates (three center-of-mass coordinates and three Euler angles) of a single water molecule at a time and going through all six molecules in cycles until the energy was converged to at least  $10^{-6}$  kcal/mol. The procedure usually converges to the minimum structure closest to the starting point. All the 120 trimers present in the hexamers thus obtained were extracted (these trimers were not included in our data set used for fitting the potential). Then 2400 trimers were created from this set by adding small random increments of either sign to the coordinates: between 0.03 and 1 bohr for the COM-COM distances and between  $1^\circ$  and  $10^\circ$  for the Euler angles.

During the initial fitting of the potential to the set of three-body nonadditive interaction energies computed at the 70 268 grid points described above, we found that the fit was not sufficiently accurate for very small intermonomer separations. Therefore, 1188 points were added in this region as described in Sec. VI B.

Altogether, a total of 71 456 trimer interaction energies were used in the fitting process, almost ten times the number of points used in Ref. 28. As already stated, this corresponds to 2.54 points per dimension. The fit of Wang *et al.*<sup>47</sup> used a comparable number of grid points, about 40 000, but these points had to cover a 21-dimensional space (which amounts to 1.66 points per dimension).

### III. AB INITIO CALCULATIONS

For all calculations in the present paper, the MOLPRO suite of programs<sup>68</sup> was used. We used the MOLPRO’s 1 hartree = 627.5096 kcal/mol conversion factor. For each grid point described in Sec. II, we performed *ab initio* supermolecular calculations of the vertical three-body nonadditive interaction energy. In the counterpoise (CP) corrected approach which removes the basis set superposition error,<sup>35,69</sup> this quantity is defined for a trimer consisting of monomers A, B, and C as

$$E_{\text{int}}[3, 3] = E_{ABC} - E_{AB} - E_{AC} - E_{BC} + E_A + E_B + E_C, \quad (2)$$

where the energies on the right-hand-side are the total energies of the indicated systems, all energies are computed in the full trimer basis set, and the positions of the monomers and of the ghost sites are in all calculations the same as in the trimer. The frozen-core approximation was used in all calculations of the correlation energies unless noted otherwise.

Although the individual calculations of the energies in Eq. (2) are nowadays not very demanding computationally even for fairly large basis sets and high levels of theory, the large number of grid points puts severe restrictions on the basis set size and computational methods. Therefore, we spent some time testing various combinations of these two elements. We performed this testing on 40 water trimers extracted from the cage and prism structures of the hexamer as described in Sec. II. We have used the standard augmented correlation-consistent basis sets,<sup>70</sup> aug-cc-pVXZ (further abbreviated as aXZ) with  $X = 2, \dots, 5$ , as well as the “half-augmented” triple-zeta basis set (haTZ) that includes augmentation only on the oxygen atom. This basis set was recommended for the water hexamer by Bates and Tschumper<sup>71</sup> as a good compromise between size and performance.

The authors of Ref. 28 compared the performance of the HF, MP2 (second-order perturbation theory based on the Møller-Plesset partition of the Hamiltonian), and CCSD(T) methods for the water trimer. Although at that time the conclusion was that the HF level is sufficiently accurate, with the currently desired accuracy we have decided to use the CCSD(T) level as this method is known to provide very reliable interaction energies in virtually all applications. However, due to its  $n^7$  scaling, CCSD(T) is computationally much more expensive than MP2 which scales as  $n^5$ . Therefore, we

also investigated a hybrid approach defined as

$$E_{\text{int}}^{\text{CCSD(T)}}/(X-X') = E_{\text{int}}^{\text{MP2}}(X) + \delta E_{\text{int}}^{\text{CCSD(T)}}(X'), \quad (3)$$

with  $X > X'$ , where

$$\delta E_{\text{int}}^{\text{CCSD(T)}} = E_{\text{int}}^{\text{CCSD(T)}} - E_{\text{int}}^{\text{MP2}}, \quad (4)$$

with both quantities computed in the same basis set.

To construct benchmarks for comparisons, we calculated interaction energies at the CBS level for all trimers. We used an extension of the hybrid method that additionally separates the HF interaction energy resulting in the quantity

$$\delta E_{\text{int}}^{\text{MP2}} = E_{\text{int}}^{\text{MP2}} - E_{\text{int}}^{\text{HF}}. \quad (5)$$

We used the extrapolations to CBS limits tested extensively in Refs. 72 and 73. For the HF interaction energy, the extrapolation formula was

$$E_{\text{int}}^{\text{HF}}(X) = E_{\text{int}}^{\text{HF}}(\text{CBS}) + Ae^{-\alpha X}, \quad (6)$$

where  $\alpha = 1.63$ , as recommended in Ref. 73, and  $A$  is an adjustable parameter. The correlation energies were extrapolated as

$$E_{\text{int}}^{\text{corr}}(X) = E_{\text{int}}^{\text{corr}}(\text{CBS}) + BX^{-3}, \quad (7)$$

where  $E_{\text{int}}^{\text{corr}}(X)$  is given by Eq. (4) or Eq. (5) calculated with the aug-cc-pVXZ basis set. The parameters in the extrapolation formulas can be obtained by solving sets of linear equations resulting from writing these formulas for  $X$  and  $X - 1$ . In cases where it will be relevant to indicate the basis set used to obtain the CBS limit, we will replace “(CBS)” by the cardinal numbers involved, for example, “(TQ)  $\equiv$  (34)” will indicate extrapolations with  $X - 1 = 3$  and  $X = 4$ . Note that the symbols Method( $X'X$ ) with  $X' = X - 1$  and Method/( $X-X'$ ) denote two different computational approaches. At the HF and MP2 levels, we used  $X = 5$ , whereas  $X = 4$  was used at CCSD(T) level. The total benchmark energy was then calculated as the sum of the three CBS values,

$$E_{\text{int}}(\text{CBS}) = E_{\text{int}}^{\text{HF}}(\text{CBS}) + \delta E_{\text{int}}^{\text{MP2}}(\text{CBS}) + \delta E_{\text{int}}^{\text{CCSD(T)}}(\text{CBS}). \quad (8)$$

The results computed in smaller basis sets are compared to the benchmarks in Table I. The maximum absolute error (MAE) of the straightforward CCSD(T) calculations in the aQZ basis relative to the CBS benchmarks is 0.003 kcal/mol. Thus, our CBS results are probably at least that accurate relative to the exact values at the CCSD(T) frozen-core level. Such level of accuracy is actually difficult to reach at the two-body level, see Table III in Ref. 37. The total errors in recovering the values of  $E_{\text{int}}[3, 6]$  are below 0.2% in the aQZ basis. Somewhat surprisingly, the analogous errors in the aDZ basis are below 0.5% only. However, MAEs are as much as 11 (16) times larger for the cage (prism) trimers in the aDZ compared aQZ bases. The aTZ basis would be a good choice in terms of accuracy, but CCSD(T)/aTZ would be too expensive for the calculations on the complete set of grid points. Therefore, the choice is really only between aDZ and haTZ at the CCSD(T) level. For the total three-body contribution, the two bases produce errors very similar in magnitude, however, the MAEs are a factor 3–4 smaller in the latter case. Thus, the choice between these two basis sets would be difficult when

TABLE I. Comparisons of errors of three-body nonadditive interaction energies (in kcal/mol) computed relative to the CBS extrapolated values. The latter energies were obtained in a hybrid way as defined by Eqs. (6)–(8) with  $X = 5$  at the HF and MP2 levels and  $X = 4$  at the CCSD(T) level. All the calculations were performed in trimer-centered basis sets. The trimer geometries were extracted from cage and prism hexamers optimized using the CC-pol-8s+NB potential. There are 20 trimers in each hexamer. The geometries of all trimers are given in the supplementary material.<sup>74</sup> “Sum” corresponds to the sum of the signed errors for individual trimers and is equal to the error in  $E_{\text{int}}[3, 6]$  for the corresponding hexamers. The root mean square errors (RMSE), mean absolute errors (MAE), and maximum absolute errors (MAXE) are also given. The  $E_{\text{int}}[3, 6]$ (CBS) values are equal to  $-8.121$  and  $-7.999$  kcal/mol for cage and prism hexamers, respectively. An extended version of this table containing the contributions for individual trimers is included in the supplementary material.<sup>74</sup>

	MP2	CCSD(T)			Hybrid	
	aTZ	aDZ	haTZ	aTZ	aQZ	aT-aD
			Cage			
Sum	−0.341	0.039	−0.036	0.036	0.013	0.031
RMSE	0.024	0.012	0.004	0.003	0.001	0.004
MAE	0.018	0.010	0.003	0.003	0.001	0.004
MAXE	0.057	0.033	0.009	0.007	0.003	0.009
			Prism			
Sum	−0.405	0.032	−0.040	0.037	0.014	0.028
RMSE	0.028	0.019	0.007	0.004	0.002	0.006
MAE	0.020	0.014	0.005	0.003	0.001	0.004
MAXE	0.073	0.048	0.019	0.008	0.003	0.013

the costs of calculations are taken into account (haTZ with 74 functions per monomer is still much larger than aDZ with 41 functions per monomer). However, it turned out that the hybrid approach, listed in the last column for each hexamer, offers actually the best performance among these three cases, with the smallest magnitude of the total error and with the MAE of 0.009 (0.013) kcal/mol for the cage (prism) trimers. In fact, this approach performs comparably to CCSD(T) in the aTZ basis set. Thus, we have chosen the hybrid approach in aTZ/aDZ bases, i.e.,  $X = T$  and  $X' = D$  in Eq. (3). Since MP2 calculations are much faster than CCSD(T) ones, the hybrid approach saves significant amounts of computer time. Overall, a calculation of  $E_{\text{int}}^{\text{CCSD(T)}}/(T-D)$  is 4 times faster than a calculation of  $E_{\text{int}}^{\text{CCSD(T)}}/aTZ$  and takes about 1 h on a single core of the 2.4 GHz Opteron processor. The root-mean-square error (RMSE) of the  $E_{\text{int}}^{\text{CCSD(T)}}/(T-D)$  values in Table I relative to the CBS results is 0.004 and 0.006 kcal/mol for the cage and prism isomers, respectively. Clearly, our hybrid approach is more than adequate at the current accuracy level of water potentials. For comparison, the calculations of Wang *et al.*<sup>46,47</sup> used the aTZ basis set at the MP2 level. Results at this level of theory are also shown in Table I. As one can see, the RMSEs of the MP2/aTZ approach are 5–6 times larger than in the  $E_{\text{int}}^{\text{CCSD(T)}}/(T-D)$  approach, whereas the corresponding errors in the total three-body contribution are 11–14 times larger in magnitude. Clearly, the addition of the  $\delta E_{\text{int}}^{\text{CCSD(T)}}/aDZ$  contributions, despite the small size of the basis set, dramatically improves the agreement with the benchmarks.

A further comparison of the performance of various methods is provided in Table II. The magnitudes of the errors in the total three-body nonadditive contribution to the

TABLE II. Performance of various methods for computing three-body pairwise nonadditive interaction energies of the trimers extracted from the cage and prism hexamers. The errors are defined as in Table I. CC stands for CCSD(T). The sources of the potentials are: CCpol3—present work; SAPT-3B—Ref. 28; WHBB $n$ —Ref. 47; HBB2-pol—Ref. 53. An extended version of this table containing the contributions for individual trimers is included in the supplementary material.<sup>74</sup>

	CC/(T-D)	HF(45)	MP2(45)	CC(34)	CCpol3	SAPT-3B	WHBB5	WHBB6	HBB2-pol
Cage									
Sum	0.031	-0.427	-0.352	-0.002	0.064	-0.172	-0.343	0.004	-0.187
RMSE	0.004	0.039	0.024	0.000	0.019	0.080	0.045	0.036	0.042
MAE	0.004	0.028	0.018	0.000	0.014	0.062	0.034	0.027	0.035
MAXE	0.009	0.104	0.055	0.001	0.040	0.157	0.101	0.099	0.096
Prism									
Sum	0.028	-0.465	-0.417	-0.002	0.019	-0.813	-0.425	-0.239	-0.291
RMSE	0.006	0.045	0.027	0.000	0.019	0.151	0.065	0.031	0.065
MAE	0.004	0.027	0.021	0.000	0.016	0.112	0.043	0.024	0.052
MAXE	0.013	0.132	0.065	0.001	0.033	0.400	0.188	0.073	0.134

hexamer interaction energies at the HF and MP2 CBS levels are about 0.4–0.5 kcal/mol, or only 1% of the hexamer interaction energy, but are clearly too large for investigations of subtle effects as the cage-prism energy difference that amounts to 0.25 kcal/mol.<sup>37,71</sup> Interestingly enough, the use of MP2 gives only a negligible improvement over the HF level of theory. In contrast, going up to the CCSD(T) level gives a substantial contribution to the hexamer energy. There are no estimates of beyond CCSD(T) contributions to three-body nonadditive energies, but most likely such contributions to the hexamer energies will be on the order of 0.01 kcal/mol. The excellent agreement of the CCSD(T)(34) results with the CBS benchmarks shows that the hybrid approach used in the calculations of benchmarks, with HF and MP2 extrapolations at the (45) level, was not needed (i.e., the CCSD(T)(34) benchmarks would have been sufficiently accurate). The remaining columns of Table II will be discussed in Sec. VII A.

To check the importance of correlating core electrons, we have computed the three-body interaction energies for all trimers extracted from the hexamer with the aug-cc-pCVTZ basis set.<sup>75</sup> The magnitude of the largest correction for a trimer amounted to 0.002 (0.002) kcal/mol and the sum for all trimers was -0.011 (-0.013) kcal/mol for the cage (prism) hexamer. These errors are a few times smaller than the errors of the hybrid approach selected for our work, so the neglect of correlation effects involving the core electrons is justified. However, when a still more accurate future potential will be developed, these effects will have to be included.

#### IV. TWO-BODY FIT

The form of the two-body CCpol2 fit is similar to that of the CC-pol-8s fit of Ref. 43, namely,

$$V_2 = \sum_{a \in A} \sum_{b \in B} \tilde{u}_{ab}(r_{ab}) + V_2^{\text{ind}}(AB), \quad (9)$$

where  $\tilde{u}_{ab}$  are site-site functions depending only on the distances  $r_{ab}$  between sites associated with monomers  $A$  and  $B$  and  $V_2^{\text{ind}}(AB)$  represents the induction interaction. The  $\tilde{u}_{ab}(r_{ab})$  functions can be written as

$$\tilde{u}_{ab}(r_{ab}) = u_{ab}(r_{ab})d_{ab}(r_{ab}), \quad (10)$$

where  $u_{ab}(r_{ab})$  has the form of Eq. (2) in Ref. 43 and  $d_{ab}(r_{ab})$  are (very) short-range damping functions equal to 1 if  $u_{ab}(r_{ab}) > 0$ , and otherwise

$$d_{ab}(r_{ab}) = \{1 + \exp[-\gamma(r_{ab} - \tilde{r})]\}^{-1}, \quad (11)$$

with  $\gamma$  and  $\tilde{r}$  being adjustable parameters. Note that this damping is different from the damping already contained in the asymptotic terms of the functions  $u_{ab}$  and sets in at much smaller  $R$  than the latter damping. The induction interaction  $V_2^{\text{ind}}(AB)$  is calculated with a new polarization model, described in Sec. V, which is more elaborate than that used in CC-pol-8s. After adopting this polarization model, the adjustable parameters of the site-site part in the fit of Eq. (9) were fitted in the same way as described in Sec. II E of Ref. 43, with the damping turned off (i.e., all the functions  $d_{ab}(r_{ab})$  set to one). However, the original set of 2510 data points was enlarged by adding 706 short-separation dimer geometries computed in Ref. 13 in order to improve the description of the repulsive wall. The *ab initio* approach used in these calculations was the same as in Refs. 29 and 30. The RMSE of the new  $V_2$  fit relative to the training set of interaction energies is 0.081 kcal/mol on the whole set of 3216 points and 0.011 kcal/mol for negative interaction energies, i.e., very similar to the error of the fit developed in Ref. 43.

In the next step, we switched on the  $d_{ab}$  damping, keeping the other, previously optimized parameters fixed. We selected this approach instead of performing a simultaneous optimization of all parameters since we did not want to change the known very good behavior of the two-body fit in the physical region. The reason for introducing the additional damping factor  $d_{ab}$  was only to improve the very small  $R$  behavior of the total CCpol23 fit (i.e., with the inclusion of the  $V_3[3, 3]$  part described in Sec. VI). The total interaction energy should be repulsive at very small  $R$ , but the  $V_3[3, 3]$  fits have a tendency to collapse there to unphysical, strongly negative values. We were unable to fully control this behavior of  $V_3[3, 3]$  alone (as described in Sec. VI), so we fixed this problem by accelerating the increase of  $V_2$  for  $R$  going to zero. This acceleration takes place only in the region not relevant for the intended physical applications. In this way, the sum of these terms, i.e., the CCpol23 potential, behaves reasonably. It may seem senseless to work on the behavior of



the potential in an unphysical region, but in molecular simulations this region is occasionally sampled. If the potential found in such sampling is strongly repulsive, as it should be, this sampling has virtually no effect on the simulations. However, if the potential is strongly attractive due to artifacts of the potential functions, the simulation may collapse. At such very small intermonomer separations, the  $V_2^{\text{ind}}$  term is small compared to the sum of the functions  $u_{ab}$  (this means that we do not observe the so-called polarization catastrophe). Therefore, we have not introduced any additional damping in  $V_2^{\text{ind}}$ . The functions  $u_{ab}$  are both positive and negative and despite strong cancellations between them (as it commonly happens for fits with many terms), their sum may become unphysical for very small  $R$  for some intermonomer orientations. We found that damping of the negative contributions in this region has the desired effect on the total CCpol23 fit. With the damping parameters chosen as described below, the functions  $u_{ab}(r_{ab})$  are non-negligibly affected only for  $r_{ab}$  smaller than a fraction of 1 Å, i.e., their behavior is unchanged in the physical region. The damping strength is controlled by the parameters  $\gamma$  and  $\tilde{r}$  in Eq. (11):  $\tilde{r} = 0$  and a very large  $\gamma$  implies no damping, increasing  $\tilde{r}$  turns the damping on at the pertinent site-site distances, while decreasing  $\gamma$  makes the effect more diffuse so that the damping effectively starts at larger distances. The values of these parameters were selected in the following way. First, we set  $\gamma$  to 100 bohr<sup>-1</sup> and increased  $\tilde{r}$  until the RMSE of the  $V_2$  fit for all the points with interaction energies between +10 and +30 kcal/mol started deteriorating. This occurred at  $\tilde{r} = 1.4$  bohrs. Then, we gradually lowered  $\gamma$  with the same RMSE criterion, arriving in this way at  $\gamma = 20$  bohr<sup>-1</sup>. The damped  $V_2$  fit has a RMSE equal to 0.25 kcal/mol between +10 and +30 kcal/mol, compared to 0.16 kcal/mol without damping, while the accuracy below +10 kcal/mol is not affected: both the damped and undamped fits have a RMSE of 0.011 kcal/mol for interaction energies below zero and a RMSE of 0.034 kcal/mol for those in the range from 0 to +10 kcal/mol. The RMSE of the damped fit on all 3216 points is very large, 222 kcal/mol, whereas for the undamped fit it is only 0.081 kcal/mol. Since virtually the whole RMSE in the former case originates from errors coming from the interaction energies above +30 kcal/mol (corresponding to  $kT$  at 15 000 K), which are practically not sampled in simulations near (and below) room temperature, this error is inconsequential for the intended applications of our potential.

We have not tested CCpol2 on the dimer characteristic points and spectra since it is numerically so close to CC-pol-8s in all physically relevant regions that the results computed with CCpol2 should be nearly identical to those computed with CC-pol-8s.<sup>43</sup>

## V. MANY-BODY INDUCTION ENERGY MODEL

We will now define the polarization model used in our potential. This model was applied in the two-body component already described in Sec. IV, the three-body component that will be described in Sec. VI, and alone to represent four- and higher-body effects. The polarization model will be defined

below in the general  $N$ -body context, special cases of two and three bodies follow immediately.

The polarization model represents the asymptotic induction energy of  $N$  molecules. It is often called classical polarization model but in fact the formalism is the same in quantum mechanics. This model can also be damped to account for the charge-overlap effects in induction interactions. In polarization models, the electric field due to the multipole moments of the charge distribution on isolated monomers (called permanent multipole moments) induces multipole moments on each monomer (of course, the permanent moments of a given monomer do not contribute to the field that induces the moments on this monomer). These induced moments, in turn, create an electric field that is added to the original field and induces additional moments. This procedure is iterated until convergence. The converged fields and induced multipole moments (and thus, the polarization energy) can also be found by solving a system of equations. In the simplest case, where only the induced dipole moments are considered and the permanent multipole moments are approximated by a set of distributed (partial) charges, the polarization energy of a system of  $N$  molecules can be written as

$$V_N^{\text{ind}} = -\frac{1}{2} \sum_{i=1}^N \sum_{k=1}^{N_{\text{pol}}^i} \mathbf{E}_{ik}^0 \cdot \boldsymbol{\mu}_{ik}^{\text{ind}}, \quad (12)$$

where  $\mathbf{E}_{ik}^0$  is the electric field generated on the  $k$ th polarizable center of molecule  $i$  by *permanent* distributed charges on all the other molecules, and

$$\boldsymbol{\mu}_{ik}^{\text{ind}} = \alpha_{ik} \mathbf{E}_{ik} \quad (13)$$

is the dipole moment induced on this polarizable center by the *total* electric field  $\mathbf{E}_{ik}$  generated by other molecules. The quantity  $\alpha_{ik}$  is the polarizability (considered here to be isotropic) of the  $k$ th center of molecule  $i$  such that

$$\alpha_i = \sum_{k=1}^{N_{\text{pol}}^i} \alpha_{ik} \quad (14)$$

is the total molecular polarizability. The total electric field on each center is the sum of permanent and induced components

$$\mathbf{E}_{ik} = \mathbf{E}_{ik}^0 + \sum_{j \neq i}^N \sum_{l=1}^{N_{\text{pol}}^j} f_3(\delta_3, r_{ik,jl}) \mathbf{T}_{ik,jl} \boldsymbol{\mu}_{jl}^{\text{ind}}, \quad (15)$$

where

$$\mathbf{T}_{ik,jl} = -\frac{1}{r_{ik,jl}^3} \left( 1 - 3 \frac{\mathbf{r}_{ik,jl} \otimes \mathbf{r}_{ik,jl}}{r_{ik,jl}^2} \right) \quad (16)$$

is the dipole-dipole interaction matrix, vector  $\mathbf{r}_{ik,jl}$  points from the  $k$ th polarization center of molecule  $i$  to the  $l$ th polarization center of molecule  $j$ , and  $\otimes$  denotes the vector direct product. The factor  $f_n(\delta, r)$  in Eq. (15) is the Tang-Toennies damping function<sup>76</sup>

$$f_n(\delta, r) = 1 - e^{-\delta r} \sum_{m=0}^n \frac{(\delta r)^m}{m!}, \quad (17)$$

TABLE III. Comparison of the four-, five-, and six-body nonadditive energies (in kcal/mol) of various structures of the water hexamer (optimized using the CC-pol-8s+NB potential of Ref. 43), calculated with CCSD(T), with one-center polarization model from Ref. 28, and with the current three-center polarization model. The CCSD(T) calculations were performed in the aTZ basis set using full hexamer-centered bases for each cluster. The values in parentheses correspond to  $E_{\text{int}}^{\text{CCSD(T)}}/(\text{T-D})$  calculations in tetramer-centered basis sets (these values were the data used in the optimizations of the polarization potential).

	Prism			Cage			Book		
	4-B	5-B	6-B	4-B	5-B	6-B	4-B	5-B	6-B
CCSD(T)	-0.5737	0.0562	0.00076	-0.4342	0.0027	-0.0014	-0.8692	-0.0311	-0.0028
	(-0.5708)			(-0.4332)			(-0.8667)		
1-center	-0.7037	0.0397	0.00193	-0.4402	0.0081	-0.0018	-0.8167	-0.0349	-0.0037
3-center	-0.6435	0.0354	0.00096	-0.4516	0.0129	-0.0025	-0.8638	-0.0423	-0.0049
	Bag			Ring			Boat		
	4-B	5-B	6-B	4-B	5-B	6-B	4-B	5-B	6-B
CCSD(T)	-0.9204	-0.0141	0.00567	-1.4529	-0.1494	-0.0079	-1.3392	-0.1287	-0.0078
	(-0.9172)			(-1.4487 <sup>a</sup> )			(-1.3347)		
1-center	-0.8427	-0.0314	0.00395	-1.4357	-0.1848	-0.0164	-1.2486	-0.1468	-0.0132
3-center	-0.8386	-0.0402	0.00485	-1.3969	-0.1887	-0.0176	-1.2648	-0.1584	-0.0143

<sup>a</sup>An incorrect value of this energy equal to  $-1.2811$  kcal/mol was used during optimization.

which continuously decays to zero at small  $r$ . The fields from the permanent charges  $q_l$  are also damped,

$$E_{ik}^0 = - \sum_{j \neq i} \sum_{l=1}^{N_q^j} f_2(\delta_2, r_{ik,jl}) q_l \frac{\mathbf{r}_{ik,jl}}{r_{ik,jl}^3}, \quad (18)$$

where  $N_q^j$  is the number of partial charges on molecule  $j$ .

The polarization model defined above is a generalization of that used in the nonadditive three-body,<sup>28,36</sup> CC-pol,<sup>29-31</sup> and CC-pol-8s<sup>43</sup> water potentials, where it was restricted to just  $N_{\text{pol}}^i = 1$ . Note that these past polarization models were damped at the three-body level but not at the two-body level. In the present work, we used three polarization centers located on the atoms of each monomer, with the polarizability values  $\alpha_1 = 6.5186$  a.u. and  $\alpha_2 = \alpha_3 = 1.5507$  a.u. The values are chosen in such a way that the total polarizability  $\alpha$  is equal to 9.62 a.u. (the benchmark CCSD(T) result of Ref. 13), while the ratio of the oxygen to hydrogen values is 4.2036 (the value calculated<sup>77</sup> using the Cam-CASP code,<sup>78</sup> with an asymptotically corrected PBE0 density functional<sup>79,80</sup> and in the doubly augmented daug-cc-pVTZ basis set<sup>81</sup>). The damping constants  $\delta_2$  and  $\delta_3$  were optimized on a training set containing both three-body and four-body nonadditive energies. In the former case, these were pure induction and exchange-induction energies including the overlap effects. In the latter case, we used the complete four-body nonadditive contributions, hoping that in this way our polarization model will effectively improve the description of the four-body interactions. Specifically, let us denote by  $\sigma_i$ ,  $i = 1, \dots, 6$ , the RMSEs of the nonadditive four-body energies for all 15 tetramers contained in each of the hexamer structures: prism, cage, book, bag, boat, and ring, respectively, calculated with the polarization model and relative to the  $E_{\text{int}}^{\text{CCSD(T)}}/(\text{T-D})$  values. Additionally, let  $\sigma_7$  stand for the RMSE of the total nonadditive four-body contributions,  $E_{\text{int}}[4, 6]$ , in the six structures. Finally, let  $\sigma_8$  be the RMSE of the polarization model with respect to the following sum of

three-body SAPT corrections:

$$\begin{aligned} E_{\text{ind}}^{(20)}[3, 3] + E_{\text{exch-ind}}^{(20)}[3, 3] + \delta E^{\text{HF}}[3, 3] \\ = E_{\text{int}}^{\text{HF}}[3, 3] - E_{\text{exch}}^{(10)}[3, 3] \end{aligned} \quad (19)$$

for the 5704 water trimer geometries computed in Ref. 28. [After the fit was completed, we found that we had also erroneously included the 1829 trimers for which  $E_{\text{exch}}^{(10)}$  was not computed and was set to zero in the data set. However, the 1829-point subset consists mostly of large trimers for which  $E_{\text{exch}}^{(10)}$  is very small in magnitude, so we have not corrected this error.] One may question the inclusion of the exchange-induction energies in Eq. (19) since these energies decay exponentially. However, at the two-body level it has been shown that the exchange-induction energy to a large extent cancels the purely exponential overlap component of the induction energy (see Ref. 82) which is also not a part of the polarization model. The values  $\delta_2 = 1.65$  bohr<sup>-1</sup> and  $\delta_3 = 1.55$  bohr<sup>-1</sup> were found by minimizing the sum  $\sum_{i=1}^8 \sigma_i$ . For the final model, the resulting values of  $\sigma_i$  are 0.020, 0.034, 0.018, 0.023, 0.009, 0.045, 0.071, and 0.203 kcal/mol.

Table III compares the predictions of the final model with those of the model from Ref. 28 and with CCSD(T) results for the four-, five-, and six-body interaction energies in the six hexamers. The three-center polarization model leads to a modest improvement over the one-center model in the four-body energies: the RMSE on all hexamers relative to the CCSD(T) benchmarks is 0.059 vs. 0.076 kcal/mol. The relative errors for individual hexamers range between  $-0.6\%$  and  $12.2\%$  with the average magnitude of the relative errors amounting to  $5.8\%$  for the three-center model. This is a surprisingly small relative error, much smaller than in the case of three-body energies<sup>13</sup> where the accuracy of simple polarization models is only about 50%. To our knowledge, this fact has not yet been noted in literature and is of significant importance in developing many-body force fields.

For five- and six-body nonadditive interactions, the RMSEs of the three-center (one-center) models are 0.025 (0.019) and 0.005 (0.004) kcal/mol, respectively. Thus, the

one-center model actually performs slightly better, but these differences are negligible. The overall accuracy is not as good as in the four-body case, but it is reasonable for the contributions that are of more significant size: the magnitudes of relative errors of the three-center model for the contributions larger in magnitude than 0.03 kcal/mol are in the range of 23%–37%.

Interestingly, the sum of the four- to six-body contributions is recovered significantly better than any individual component. The RMSE for the three-center model is only 0.046 kcal/mol and the average magnitude of the relative errors is 3.5%. The many-body effects beyond the polarization interactions are not that small for individual  $K$ -mers within each hexamer, but there are significant cancellations.

We then tested the performance of the polarization model on the set of 600 nonadditive three-body interaction energies for trimer configurations extracted from the ambient-conditions liquid water MD simulations of Ref. 13 (see Sec. II). These energies, ranging from  $-0.9$  to  $1.0$  kcal/mol, were computed at the hybrid CCSD(T)/(T-D) level and are quite well reproduced by our new polarization model. The value of the RMSE relative to *ab initio* benchmarks is 0.073 kcal/mol, much improved compared to the single-center polarization models included in the CC-pol-8s' and CC-dpol-8s' potentials of Ref. 13 which both give a RMSE of 0.107 kcal/mol.<sup>13</sup> This improvement was achieved despite the fact that the new polarization model was partly optimized for four-body effects. This performance indicates that the three-center model is a more physically sound representation of the induction effects in water than the one-center model.

## VI. NONADDITIVE THREE-BODY FIT

### A. Functional form of fit

In analogy with Eq. (18) of Ref. 28, our present three-body fit is the sum of three components,

$$V_3[3, 3] = F_{S^3} + F_{S^2} + V_3^{\text{ind}}[3, 3]. \quad (20)$$

The term  $V_3^{\text{ind}}[3, 3] = V_3^{\text{ind}} - V_2^{\text{ind}}(AB) - V_2^{\text{ind}}(BC) - V_2^{\text{ind}}(AC)$  is the nonadditive induction energy from the new polarization model described in Sec. V. The  $F_{S^2}$  term is the interaction energy contribution of the same form as in Ref. 28 that is due to single two-electron permutations between monomer pairs, while  $F_{S^3}$  represents mostly the interaction energy due to cyclic permutations involving three electrons from three different monomers, but also other residual effects not accounted for by the other terms. Also, nonadditive dispersion energy which is nonnegligible for this system only for short and medium-range intermolecular distances (for the near-equilibrium trimers, this contribution is on the order of only 0.01 kcal/mol<sup>18</sup>) is modeled by this term. Instead of the Legendre polynomial expansion employed in Ref. 28, we used a trimer generalization of the exponential site-site expansion of CC-pol-8s (see Eqs. (1)

and (2) of Ref. 43). Specifically,

$$F_{S^3} = \sum_{a \in A} \sum_{b \in B} \sum_{c \in C} f_{ab}(r_{ab}) f_{ac}(r_{ac}) f_{bc}(r_{bc}) \times \sum_{k=0}^1 \sum_{l=0}^1 \sum_{m=0}^1 c_{klm}^{(abc)} r_{ab}^k r_{ac}^l r_{bc}^m, \quad (21)$$

where

$$f_{ab}(r_{ab}) = \{1 + \exp[-\gamma_{ab}(r_{ab} - \tilde{r}_{ab})]\}^{-1} \exp(-\beta_{ab} r_{ab}) \quad (22)$$

and similarly for  $f_{ac}(r_{ac})$  and  $f_{bc}(r_{bc})$ . As before,  $r_{ab}$  denotes the distance between site  $a$  in molecule  $A$  and site  $b$  in molecule  $B$ . The nonlinear adjustable parameters  $\gamma_{ab}$ ,  $\beta_{ab}$ , and  $\tilde{r}_{ab}$ , as well as the linear ones,  $c_{klm}^{(abc)}$ , can be obtained from the optimization of an appropriate least-squares functional, with linear parameters obtained in each step of the nonlinear optimization by solving the set of equations of the linear least-square problem. The first factor in the function  $f_{ab}(r_{ab})$  is a damping function of the same form as  $d_{ab}(r_{ab})$  in Eq. (11) but, in contrast to the two-body fit, we used parameters  $\gamma_{ab}$  and  $\tilde{r}_{ab}$  that were independently optimized for each site-site pair. The accuracy of the expansion (21) depends critically on the number and location of monomer sites. In Ref. 43, the new sites in the dimer expansion were successively added and their positions optimized until reaching 25 (8 symmetry-unique) sites per molecule. Since a similar optimization procedure would be extremely expensive in the trimer case, we used the first 17 (6 symmetry-unique, including the O and H atoms) site positions obtained in Ref. 43. After grouping the symmetry-equivalent terms as described below, Eq. (21) includes 364 independent linear parameters. For 6 unique sites per molecule, the number of nonlinear fit parameters amounts to 63 (21 for  $\beta_{ab}$ ,  $\gamma_{ab}$ , and  $\tilde{r}_{ab}$  each).

Since the symmetry operations of the water monomer transform a given site into one of its equivalents, the nonadditive three-body potential for the water trimer with rigid monomers should be invariant to permutations of symmetry-equivalent sites within each monomer. In addition, the potential should be invariant to the six (including the identity) permutations of complete monomers A, B, and C. Thus, the terms in Eq. (21) can be separated into groups such that each term in a given group has the same linear coefficient. Also, to realize the symmetry conditions, the functions of Eq. (22) should be identical for all  $a$ 's and all  $b$ 's that are symmetry equivalent. Thus, there are only  $6 \times 7/2 = 21$  different such functions. We used a simple algorithm (executed only during initialization of the potential subroutine) to impose this symmetry. Initially, an integer-valued array  $I(a, b, c, k, l, m)$  is filled with zeros and a counter  $N_{\text{lin}}$  is set to one. In a loop going through all possible  $17^3 \times 2^3 = 39\,304$  terms of the sum in Eq. (21), if  $I(a, b, c, k, l, m)$  is still equal to zero for a given combination of  $a, b, c, k, l, m$ , it is set to  $N_{\text{lin}}$ . If  $I(a, b, c, k, l, m)$  is not equal to zero, this combination is skipped because it has been handled before. Then each of the parameters  $a, b, c$  is identified as belonging to a group of symmetry equivalent sites and  $I(a', b', c', k, l, m)$  are set to  $N_{\text{lin}}$  for all  $a', b', c'$  of the same type as  $a, b, c$ , respectively. At the same time, the ABC permutational symmetry is realized by setting to  $N_{\text{lin}}$

all the six cases resulting from the permutations of the three monomers. Finally, the value of  $N_{\text{lin}}$  is increased by one. In actual calculations using Eq. (21), all terms with the same value of  $I(a, b, c, k, l, m) = n$  are added together and associated with the linear parameter  $c_n$ .

## B. Fitting of three-body potential

The first stage of the fitting procedure used the initial set of 70 268 trimer geometries obtained as described in Sec. II. The data points were unweighted (all weights set to one) and no damping was applied, i.e., the factors  $d_{ab}$  in Eq. (22) were all equal to one. Several hundred different fits were generated by using different (randomly generated) starting values of the nonlinear parameters and then optimizing them using Powell’s algorithm.<sup>67</sup> The induction component was held fixed. Several of the most accurate fits, those with RMSEs (relative to all *ab initio* three-body nonadditive interaction energies) equal to about 0.02 kcal/mol, were tested at short-distance configurations in the following way. For each point on a three-dimensional grid of intermonomer distances from 1.8 to 3.0 Å with steps of 0.2 Å, we generated  $1 \times 10^6$  different trimer geometries by randomly choosing the orientations of the monomers. At each geometry, the two- and three-body energies were evaluated from the fits  $V_2$  (also with  $d_{ab} = 1$ ) and  $V_3[3, 3]$ . The 5940 geometries corresponding to trimer interaction energies that were deemed “most unphysical” were selected. These were not just very negative energies, but we looked in particular at negative energies that were strongly dominated by  $V_3[3, 3]$ . In the next step, we checked if switching on the damping in both fits can eliminate the unphysical behavior at short distances. To this end, a new series of  $V_3[3, 3]$  fits were generated with all parameters optimized (including the damping parameters) and tested on the same 5940 points in the presence of the damped two-body fit. The number and the magnitude of negative interaction energies at short distances were significantly reduced, but not sufficiently enough. Therefore, we decided to enlarge the training data set to encompass a number of very small trimers. To this end, we calculated three-body energies at the 5940 problematic points using the CCSD(T) method and the aDZ basis set. A subset of randomly chosen 1188 points was added to the main data set, while the remaining 4752 points were used for testing purposes. The fitting process was then repeated. However, to prevent deterioration of the fit accuracy in the main, physically relevant region, the following weighting factors were used:

$$w = \begin{cases} 1, & E_{\min} < E < E_{\max} \\ (E_{\min} - E + 1)^{-3}, & E < E_{\min} \\ (E - E_{\max} + 1)^{-3}, & E > E_{\max}, \end{cases} \quad (23)$$

where  $E_{\min} = -1.51$  kcal/mol and  $E_{\max} = 0.69$  kcal/mol are the values of the lowest and highest three-body energies occurring in the 60 trimers present in the three lowest-energy hexamer structures (cage, prism, and book). Such a choice assigns progressively lower weights to trimers with energies far from the physically relevant region (which were included in the training set only to enforce a qualitatively correct behavior in high-energy regions). The addition of these 1188 points to the training set made the fit to behave well enough for all 5940

points. To select a small subset of “finalists” out of a large set of generated fits (differing, again, by the starting values of the nonlinear parameters), we used several criteria: the overall RMSE, the RMSE on the set of 60 hexamer trimers mentioned above (not included in the training set), and the magnitude of the difference between the total three-body energies in the cage and prism structures. The final fit was selected based on the smallest errors on the testing set of 4752 short-distance geometries. Its RMSE on the initial set of 70 268 geometries amounts to 0.0184 kcal/mol and on the 60 trimers extracted from hexamers to 0.0145 kcal/mol. The former RMSE may be compared to the typical values of nonadditive three-body energies in our 70 268 set which range from  $-3.52$  to  $1.94$  kcal/mol and to the  $0.47$  kcal/mol value of their root-mean square. A RMSE of about 0.02 kcal/mol is consistent with that of the two-body potential which is 0.01 kcal/mol (for negative interaction energies) for a single dimer, so it amounts to 0.02 kcal/mol per trimer if the errors are added in squares. Such uncertainties are also consistent with the uncertainties of the *ab initio* calculations estimated in Ref. 43 to be about 0.05 kcal/mol for the water dimer near its van der Waals minimum. Of course, the accuracy of the fit could have been increased easily by using a more elaborate fit function, but such a fit would also take more time in MD simulations. The RMSE of CCpol3 can be compared with that of the fit of Mas *et al.*<sup>28</sup> which was 0.07 kcal/mol and Wang *et al.*<sup>47</sup> which was 0.15 kcal/mol for the 5th-order fit and 0.042 kcal/mol for the 6th-order fit, in all cases relative to the training data set used in a given reference.

A Fortran program evaluating the fit is available in the supplementary material.<sup>74</sup> Note that this program as well as our fitting programs work internally in atomic units and print interaction energies using conversion factor 627.51, different from the one quoted earlier. This is inconsequential since the training data are converted by these programs from kcal/mol to hartrees using the former factor.

## VII. APPLICATION TO CLUSTERS

### A. Water trimer

The main result of this work is a new three-body pairwise nonadditive potential (CCpol3). It was first tested on the 40 trimers selected from the cage and prism hexamers (not included in the training data set) and the results are shown in Table II. As one can see, CCpol3 performs very well on these trimers, with the errors of the fit with respect to the CBS benchmarks about 3–5 times larger than the errors of the *ab initio* results at the CCSD(T)/(T-D) level of theory used to produce the training set. The performance of CCpol3 is still better on the total three-body contribution to each hexamer energy, with the error even slightly smaller for the prism and two times larger for the cage than the CCSD(T)/(T-D) errors. CCpol3 recovers the benchmark energies in Table II much better than any other potential. The SAPT-3B potential of Mas *et al.*<sup>28</sup> works reasonably well for the cage, but gives a large error for the prism (the origin of this error will be discussed below). These errors are consistent with the HF theory level and modest size basis sets used in Ref. 28. The WHBB5



TABLE IV. Trimer stationary-point interaction energies (in kcal/mol) and barriers relative to the energy at the minimum (in  $\text{cm}^{-1}$ ) at geometries optimized with the CCpol23 potential except for the last four columns. The levels of CBS calculations are described in the text and include all-electron correlation in the two-body part. The monomer-flexibility correction  $\Delta E_{R \rightarrow F}$  [in  $\text{cm}^{-1}$ ] relative to its value at the minimum is defined as in Ref. 43. The notation for minima, stationary points (SP), and transition states (TS) follows Ref. 84. The conversion factor from kcal/mol to  $\text{cm}^{-1}$  was 349.7550.

Stationary point	$E_{\text{int}}$ CCpol23	$\Delta E_{\text{int}}$ CCpol23	$E_{\text{int}}$ CBS	$\Delta E_{\text{int}}$ CBS	$\Delta E_{R \rightarrow F}$ CBS	$\Delta E_{\text{int}} + \Delta E_{R \rightarrow F}$ CBS	$\Delta E_{\text{int}} + \Delta E_{R \rightarrow F}$ CCpol23	$\Delta E_{\text{int}}$ Ref. 84
Minimum [uud]	-16.061		-16.196					
$C_1 \rightarrow C_1$ TS [udp]	-15.763	104.4	-15.887	108.1	-21.0	87.1	83.4	82
$C_3$ minimum [uuu]	-15.268	277.3	-15.389	282.0	-14.3	267.8	263.0	269
$C_1 \rightarrow C_3$ TS [uup]	-15.216	295.6	-15.336	300.7	-23.5	277.2	272.1	275
$C_{3h}$ SP [ppp]	-14.598	511.8	-14.687	527.8	-70.0	457.8	441.8	440
Bifurcated TS [upbi]	-13.829	780.8	-13.838	824.6	-61.2	763.4	719.7	760
RMSE	0.108 <sup>a</sup>	21 <sup>a</sup>				8 <sup>b</sup>	18 <sup>b</sup>	

<sup>a</sup>With respect to CBS results.

<sup>b</sup>With respect to the barriers of Ref. 84.

potential<sup>47</sup> is of similar accuracy as SAPT-3B, but the more flexible fit used in WHBB6 significantly improves the accuracy. This potential performs overall better than the HBB2-pol potential.<sup>53</sup> WHBB6, the best performing potential from the literature, gives a RMSE relative to the CBS energies for the cage (prism) isomers of 0.036 (0.031), whereas CCpol3 gives 0.019 (0.019) kcal/mol. We can also compare our results to the DPP2 water model of Kumar *et al.*<sup>83</sup> The nonadditive three-body part of this model was fitted to CCSD(T)-level values for the trimers extracted from the cage, prism, book, and ring hexamers. For the two former hexamers, the DPP2 values from Table V of Ref. 83 have errors only of  $-0.22$  and  $-0.19$  kcal/mol, respectively, relative to the sums of three-body nonadditive energies for each hexamer from their training set (however, the latter values are different by 1.5 kcal/mol from our benchmarks). These deviations compare favorably to the performance of most methods in Table II.

Table IV examines the performance of the CCpol23 potential at the stationary points of the water trimer. The geometries of the stationary points were optimized using CCpol23 and are listed in the supplementary material.<sup>74</sup> At each point, CBS interaction energies were computed as a sum of two-body and three-body contributions. The latter contributions were computed as for the trimers investigated in Table I. The former contributions were also computed in the hybrid approach defined by Eqs. (6)–(8), but with  $X = 6$  at the HF and MP2 levels and  $X = 4$  at the CCSD(T) frozen-core level. The midbond  $3s3p2d2f1g$  basis set, the same as used in Ref. 37, was applied in the dimer calculations. The use of such larger basis sets was necessary since the two-body energies converge slower than the three-body energies. In contrast to all other calculations presented here, only a dimer-centered basis set was applied in the two-body calculations, which is justified at the CBS level with such large values of  $X$ .<sup>37</sup> Based on the results in Table III of Ref. 37, one can estimate that such a CBS limit for the total two-body contribution to the trimer interaction energy should be accurate to about 0.01 kcal/mol relative to the exact CCSD(T) frozen-core value. We have then computed the all-electron interaction energies using Eq. (6) at the HF level and Eq. (7) at the MP2 level with bases aug-cc-pCVTZ and aug-cc-pCVQZ.<sup>75</sup> The  $\delta E_{\text{int}}^{\text{CCSD(T)}}$  term was calculated in the aug-cc-pCVTZ basis without any extrapolation. The resulting correction to the frozen-core approxima-

tion ranged from  $-0.118$  kcal/mol for the global minimum to  $-0.092$  kcal/mol for the bifurcated transition state. Thus, the inclusion of this effect is absolutely necessary for predicting the total energies at the 0.01 kcal/mol accuracy. Since the core correction is converged to better than 0.01 kcal/mol in the basis set used, the overall accuracy of our all-electron CCSD(T) results is the same as that of the frozen-core values, i.e., 0.01 kcal/mol. Since the three-body nonadditive energies at the CBS level were estimated in Sec. III to be accurate to at least 0.003 kcal/mol, the overall error of the CBS value is determined by the two-body component.

The *ab initio* interaction energies used as the data set to fit the CCpol2 potential were computed using all electrons but in aug-cc-pVXZ bases which were optimized in frozen core calculations. To check the effects of using such basis sets, we computed the correction to the frozen-core results using the aug-cc-pVTZ and aug-cc-pVQZ bases (i.e., the ones used in Ref. 30) instead of the aug-cc-pCVTZ and aug-cc-pCVQZ ones and found that the results for the six characteristic points were different only by from 0.004 to 0.005 kcal/mol, i.e., negligibly. Thus, the core correction is well reproduced in valence-optimized bases provided that the CBS extrapolations are used.

Table IV shows that the CCpol23 fit recovers the CBS trimer energies with a RMSE of 0.108 kcal/mol and a maximum absolute error of 0.134 kcal/mol. All the CCpol23 interaction energies lie above the CBS ones. The smallest error is for the highest stationary point, which must be fortuitous. One should recall here that the two-body part of this potential was fitted to *ab initio* interaction energies computed at the following levels: HF-aQZ, MP2-(TQ), CCSD(T)-aTZ, i.e., significantly lower than the level of our current benchmarks. For the global minimum, the total two-body contribution at this level is  $-12.613$  kcal/mol, whereas the CBS result with all electrons correlated is  $-12.692$  kcal/mol. The difference of 0.089 kcal/mol constitutes 66% of the difference between CCpol23 and CBS values. The remaining part of the difference, 0.046 kcal/mol, is consistent with uncertainties resulting from the fitting process:  $3 \times 0.011$  kcal/mol from the two-body part and 0.018 kcal/mol from the three-body part.

One can also evaluate the performance of CCpol23 by comparing the barriers on the surface, i.e., the differences between the energies of stationary points and the minimum



energy. Such barriers are very important for the spectra of the trimer. As Table IV shows, the RMSE of CCpol23 relative to the CBS barriers is  $21 \text{ cm}^{-1}$  or  $0.06 \text{ kcal/mol}$ . The error is largest,  $0.126 \text{ kcal/mol}$ , for the barrier to the highest stationary point which is related to the fact that the energy of this point is so well reproduced that there is virtually no cancellations of errors.

Whereas the CCpol23 predictions agree very well with CBS benchmarks, one may ask how close are the stationary-point CCpol23 geometries to those on flexible-monomer potential energy surface. The 21-dimensional optimizations were performed by Anderson *et al.*<sup>84</sup> at the MP2 level. The simplest test for assessing the closeness of geometries was to compute the CCpol23 interaction energies at “rigidized” geometries of Anderson *et al.*, i.e., for each of the six stationary-point trimers from Ref. 84, we have constructed a trimer with monomers at the geometry used in CCpol23, in the way described in Sec. II. The geometries optimized in Ref. 84 were provided to us by the authors of that work (see Ref. 85: note that the original supplementary material of Ref. 84 contains misprints). These rigidized-trimer interaction energies computed using CCpol23 turned out to be very close to the CCpol23 values in the first column of Table IV: the differences range from  $0.009$  to  $0.027 \text{ kcal/mol}$ , and are below  $0.2\%$  of the interaction energies. For the barriers, the differences are between  $1.1$  and  $6.2 \text{ cm}^{-1}$ . These comparisons indicate that the two sets of geometries are indeed very close.

Compared to the small differences in barrier heights at CCpol23-optimized and rigidized geometries of Anderson *et al.*,<sup>84</sup> both the CCpol23 and CBS barrier heights in Table IV are relatively far from the results of Anderson *et al.*,<sup>84</sup> with RMSEs of  $52$  and  $36 \text{ cm}^{-1}$ , respectively. Thus, most of these differences must be due to different monomer geometries in the rigid-monomer and flexible-monomer structures. To check this hypothesis, we have computed a “monomer flexibility correction”  $\Delta E_{R \rightarrow F}$  defined in the same way as in Ref. 43, i.e., as difference between the total electronic energies of the original trimer and the trimer with rigidized monomers. These corrections were computed using Eqs. (6)–(8) with (45) extrapolations at the HF and MP2 levels, and (34) extrapolations at the CCSD(T) level. The frozen-core approximation was used at the correlated levels. Note that one may view the  $\Delta E_{R \rightarrow F}$  correction as describing the energetic effect of trimer geometry optimization starting from a rigid-monomer stationary point and relaxing intramonomer coordinates under the conditions of keeping constant the molecular plane, COM, and the position of the line through oxygen and the midpoint of the segment connecting two hydrogen.

The comparisons utilizing the  $\Delta E_{R \rightarrow F}$  corrections are presented in Table IV for both the CBS and CCpol23 barriers. The barriers from Ref. 84 listed in the table are not the “best estimates” from that work (which include some post-CCSD(T) contributions), but their results at the extrapolated CCSD(T) level. These energies were computed using bases up to  $a6Z$  at the MP2 level and up to  $aQZ$  at CCSD(T) level and performing several different types of extrapolations which led to estimated uncertainties of the barriers between  $6$  and

$16 \text{ cm}^{-1}$ . As the results in Table IV show, the CBS and CCpol23 barriers corrected for monomer-flexibility effects are in excellent agreement with the results from Ref. 84, with RMSEs relative to the values from Ref. 84 almost to within the uncertainties of the latter quantities. These RMSEs ( $0.02$  and  $0.05 \text{ kcal/mol}$ , respectively) are also of the size expected from estimates of the uncertainties of our calculations. This agreement is consistent with the finding discussed above that the rigid-monomer CCpol23 potential predicts the intermolecular geometries of the trimer stationary points very well. The largest discrepancies with the results of Anderson *et al.*<sup>84</sup> shed some light on accuracy of the CCpol23 potential. For the bifurcated transition state where the difference between CCpol23 and CBS barriers is the largest in magnitude, the CBS barrier corrected for monomer-flexibility effects agrees well with value of Anderson *et al.*, showing that despite the discrepancy, the geometry of this transition point is accurate. On the other hand, the CBS barrier corrected for monomer-flexibility effects shows the largest discrepancy with result of Anderson *et al.*<sup>84</sup> for the  $C_{3h}$  structure which is a third-order stationary point. It is possible that for this point our geometry optimization was not completely converged. The other possibility is that this point is sensitive to the  $\delta E^{\text{CCSD(T)}}$  contribution which was not included in the optimizations of Ref. 84.

One can also compare interaction energies in a similar way, although this comparison is less straightforward than for barriers which are just differences of total electronic energies. In the case of interaction energies, the reference points are different: equilibrium monomers in the case of the interaction energies of Anderson *et al.*<sup>84</sup> and  $\langle r \rangle_0$ -geometry-monomers in the case of CCpol23 and our CBS limit benchmarks. The total trimer interaction energies at the global minima are  $-15.89 \text{ kcal/mol}$ <sup>84</sup> (relaxed CP-corrected energy at CCSD(T) level),  $-16.06 \text{ kcal/mol}$  (CCpol23, vertical energy), and  $-16.20 \text{ kcal/mol}$  (CBS, vertical energy). The total deformation energy of structure of Anderson *et al.*<sup>84</sup> is  $0.40 \text{ kcal/mol}$  (computed by us at the CCSD(T)/ $a5Z$  level), so that the corresponding vertical interaction energy is  $-16.29 \text{ kcal/mol}$ . These differences may seem small taking into account that the energy of three  $\langle r \rangle_0$  monomers lies  $0.67 \text{ kcal/mol}$  (Ref. 86) above the energy of three  $r_e$  monomers. One reason is a partial cancellation of contributions. Let us measure the energies from the energy of the global trimer minimum with flexible monomers. The energy of the corresponding trimer with monomers rigidized to the  $\langle r \rangle_0$  geometry is at  $0.39 \text{ kcal/mol}$  (i.e., equal to the magnitude of the monomer-flexibility correction  $E_{R \rightarrow F}$ ). The CBS vertical interaction energy for this geometry is  $-16.17 \text{ kcal/mol}$  (estimated from the value in Table IV and the CCpol23 difference of energies between CCpol23-optimized and structures of Anderson *et al.*<sup>84</sup>), so that three  $\langle r \rangle_0$  monomers are at  $16.56 \text{ kcal/mol}$ . Subtracting  $0.67 \text{ kcal/mol}$  from this value gives  $15.89 \text{ kcal/mol}$ , the magnitude of the relaxed interaction energy of Anderson *et al.*<sup>84</sup> Thus, the partial cancellations in this energetic “cycle” explain to some extent the good agreement of the rigid- and flexible-monomer approaches observed here and later on. The other reason for the closeness of rigid- and flexible-monomer predictions for the equilibrium trimer is that the

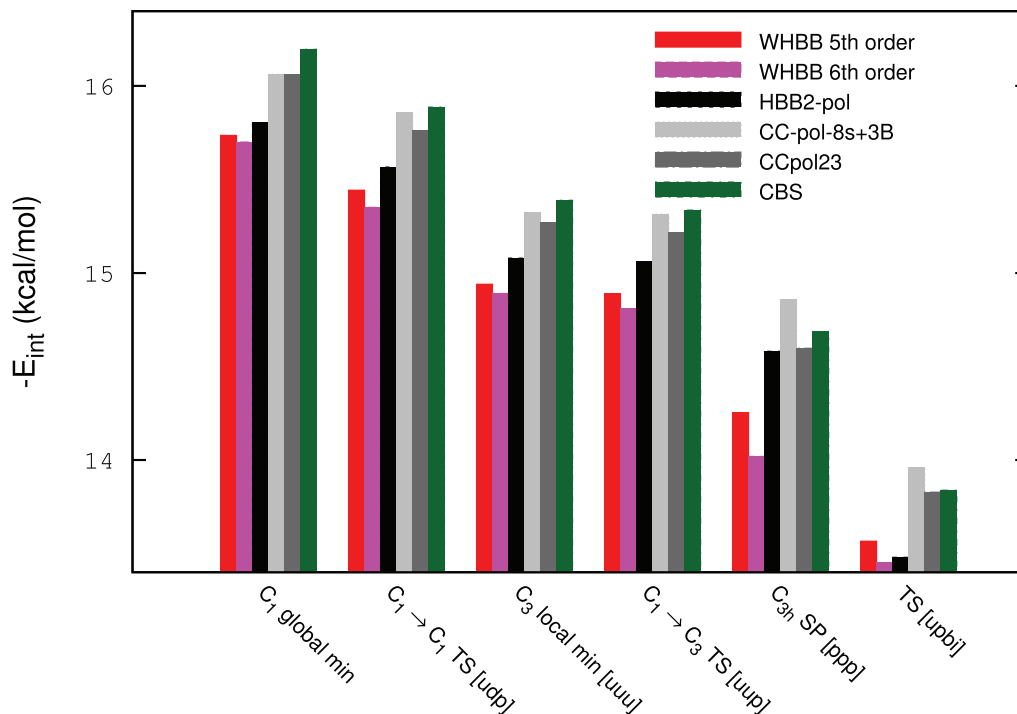


FIG. 1. Comparison of the total interaction energies of the water trimer computed using CCpol23 and potentials from the literature. The geometries are those of stationary points optimized using CCpol23 and are the same for all methods. The CBS results were obtained as described in the text. The WHBB5 and WHBB6 potentials are from Ref. 47, whereas HBB2-pol is from Ref. 53. The vertical interaction energies produced by these potentials were computed by us.

hydrogen-bonded OH bond length in the latter approach, close to 0.972 Å for all monomers, happens to be the same as the  $\langle r \rangle_0$  value. Thus, the “binding ring” is almost identical in the flexible-monomer and rigid  $\langle r \rangle_0$  trimer minima. The free OH bonds in flexible-monomer approach are virtually unchanged from the equilibrium value of 0.959 Å and therefore are quite different than in the  $\langle r \rangle_0$  monomers, but obviously contribute much less to interaction energy. One may finally note that the increase of the length of the OH bond participating in the hydrogen bond amounting to 0.013 Å is much larger than for the dimer where it is only 0.005 Å (Ref. 7).

A comparison of the performance of the CCpol23 potential with literature potentials on the trimer stationary points optimized using the CCpol23 potential is shown in Fig. 1. The quantities compared are the total trimer interaction energies and the CBS benchmarks are those described above. As we already know from Table IV, the CCpol23 predictions are remarkably close to the benchmarks. By contrast, the potentials from the literature give predictions with RMSEs relative to the benchmarks (at CCpol23-optimized geometries) between 0.31 and 0.53 kcal/mol, several times larger than the 0.11 kcal/mol RMSE of CCpol23. Surprisingly, CC-pol-8s+NB performs better than other published potentials, in fact as well as CCpol23 as its RMSE is also 0.11 kcal/mol, probably because trimer tunneling paths were well represented in its training data base. Also surprisingly, WHBB5 predicts the trimer interaction energies slightly better than WHBB6 despite the simpler form of the fitting function and despite the opposite performance on the trimers extracted from the hexamer. Note that the shape of the diagram is very similar for each of the six stationary points, which stems from the fact that all the structures have a similar, near-equilateral triangle

oxygen skeleton and differ mainly (except for one case) by flipping of the non-bonded hydrogen atoms.

We next tested the performance of CCpol3 on the set of 600 random geometries from MD simulations of Ref. 13 which were used already in Sec. V. These geometries were part of our training set, but constituted only a small fraction of the total number of points. The comparisons are made here again on the nonadditive energies only, rather than on the trimer interaction energies. The results for CCpol3 and literature fits are summarized in Table V. This table shows that CCpol3 performs very well in this test, with a RMSE relative to  $E_{\text{int}}^{\text{CCSD(T)}/(\text{T-D})}$  values of 0.0154 kcal/mol, very close to the RMSE on the training set and almost 3 times improved relative to the SAPT-3B potential of Ref. 28. Surprisingly, the HBB2-pol potential<sup>53</sup> produces a RMSE only slightly better than that of SAPT-3B, whereas the RMSE of the WHBB potentials<sup>47</sup> is about 1.5 times larger than that of SAPT-3B and very close to that of our pure polarization model described in Sec. V.

TABLE V. Comparison of RMSEs (in kcal/mol) of nonadditive three-body energies on 600 trimers selected from snapshots of MD simulations in Ref. 13.

Polarization model (Ref. 13)	0.107
Polarization model (present work)	0.0734
SAPT-3B (Ref. 28)	0.0418
WHBB5 (Ref. 47)	0.0735
WHBB6 (Ref. 47)	0.0642
HBB2-pol (Ref. 53)	0.0374
CCpol3	0.0154

To appreciate the significance of these RMSE values, we should compare them to the root mean square value of the three-body energy for the 600 trimers which amounts to 0.187 kcal/mol.<sup>13</sup> Thus, the simple polarization models considered in Ref. 13 result in roughly 50% errors. The use of the current three-center polarization model reduces this error to 39%, whereas the use of CCpol23 reduces it to 8%. Most likely the 39% value is close to how well one can reproduce the three-body nonadditive energies with a polarization model based only on asymptotic information. One should mention here, however, that an effective polarization model can be constructed to better reproduce the total three-body nonadditive energies if it is fitted to these values. Such a model is a part of the DPP2 water potential of Kumar *et al.*<sup>83</sup> Its functional form is fairly close to that in our three-center model, but the parameters were fitted simultaneously to the water monomer polarizability and to a set of *ab initio* computed total three-body nonadditive energies for trimers extracted from low-energy isomers of the hexamer. As discussed before, this model reproduces very well the sums of the three-body energies for the isomers from their training set. Obviously, this is achieved in a partly unphysical manner as the nonadditive first-order exchange terms, which are significant for all nonasymptotic separations, have a different functional dependence than the polarization terms for which the form of the model is valid.

Scatter plots of WHBB6, HBB2-pol, and our new three-body fit energies as functions of the 600 benchmark energies are presented in Fig. 2. The horizontal band at the WHBB6 fit energies equal to zero results from the fact that the WHBB potentials neglect the nonadditive three-body energy if any of the distances between oxygen atoms in the trimer is larger than 6 Å (this occurs for 211 points out of the 600). It is worth noting that, among the 600 geometries, the largest absolute error of the CCpol3 fit amounts to 0.099 kcal/mol and there are only 8 points with errors larger than 0.05 kcal/mol.

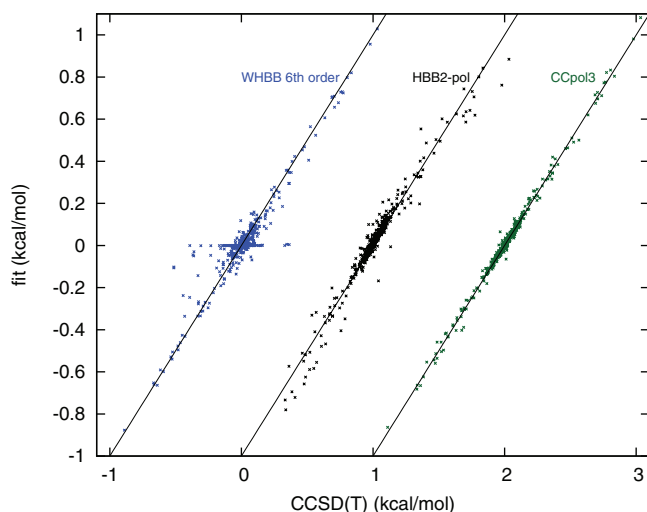


FIG. 2. Nonadditive three-body energies of 600 trimer configurations selected from snapshots of MD simulations in Ref. 13 calculated from WHBB6,<sup>47</sup> HBB2-pol,<sup>53</sup> and CCpol3 fits and compared with CCSD(T)/(T-D) values. The straight lines represent the ideal case (i.e., fit energies equal to CCSD(T) energies). Note that the CCSD(T) energies are augmented by 1 and 2 kcal/mol in the HBB2-pol and CCpol3 plots, respectively.

For WHBB6 and HBB2-pol, the largest absolute errors (the number of points with errors larger than 0.05 kcal/mol) are 0.411 kcal/mol (125) and 0.218 kcal/mol (63), respectively.

## B. Water hexamer

As mentioned before, the hexamer is the smallest water cluster with stable “three-dimensional” forms (in the sense that the oxygen atoms are highly non-planar). It exists in several low-energy local-minima structures, which has led to a long controversy regarding the most stable isomer.<sup>37,53,64–66,71,87,88</sup> Therefore, accurate predictions of the relative energies of various hexamer structures have been recognized as one of the most important tests of water potentials. We investigated six of the lowest structures (prism, cage, book, bag, ring, boat) often considered in the literature, as high-quality comparative benchmarks are available for these configurations. Dahlke *et al.*<sup>66</sup> optimized the geometries at the MP2 level in the haTZ basis set and evaluated the hexamer energies using CCSD(T) in the same basis. Bates and Tschumper<sup>71</sup> used geometries from Ref. 66 and performed analogous calculations for the  $\delta E_{\text{int}}^{\text{CCSD(T)}}$  contributions, but their MP2 energies were calculated using the MP2-R12 method<sup>89</sup> which should give results close to the MP2 CBS limit. Note that there are some nomenclature differences in the literature regarding the water hexamer. We follow the convention adopted in Ref. 66, while in Ref. 71 the names “book-1,” “cyclic-boat-1,” and “cyclic-chair” are used for book, boat, and ring, respectively, and two more structures are considered (“book-2” and “cyclic-boat-2”) differing just by the orientation of the free hydrogen atoms at some monomers.

A comparison of the performance of various potentials on the hexamer isomers is presented in Fig. 3. The total hexamer interaction energies are given relative to the energy of the prism configuration obtained with a given method. The results for the WHBB5, WHBB6, and HBB2-pol potentials were taken from supplementary material of Refs. 47 and 53. The energies of hexamer isomers were optimized in these references varying all coordinates (i.e., with flexible monomers) using the appropriate potentials. The interaction energies plotted for these potentials as well as the benchmark energies taken from Refs. 66 and 71 are the relaxed ones.<sup>35</sup> Geometry optimizations with the CC-pol-8s+NB and CCpol23+ potentials were performed by us and the interaction energies corresponding to these potentials are the vertical ones. Hence, each interaction energy in Fig. 3 is self-contained, i.e., was obtained completely within the given approach (except for the benchmark results which were computed at geometries optimized in smaller bases and not including the  $\delta E_{\text{int}}^{\text{CCSD(T)}}$  contribution).

Figure 3 shows that the interaction energies from CCpol23+ and both benchmarks are very close in all cases, although CCpol23+ is noticeably closer to the CBS-level benchmarks of Bates and Tschumper<sup>71</sup> (0.01–0.03 kcal/mol) than to the haTZ results of Dahlke *et al.*<sup>66</sup> (0.03–0.06 kcal/mol). This is not accidental, as both the two-body and the three-body parts of CCpol23+ were fitted to energies obtained with basis sets much larger than haTZ. This excellent

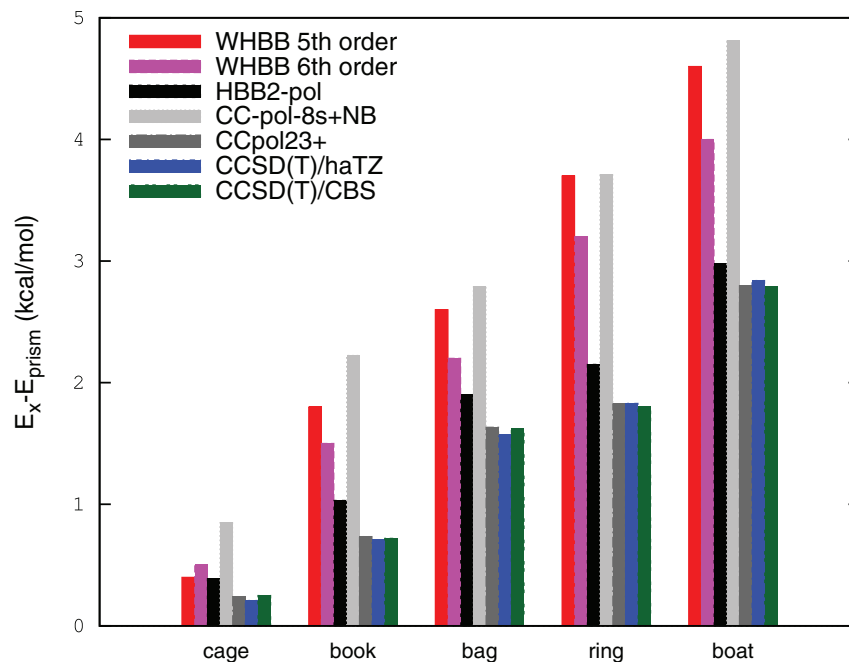


FIG. 3. Energies of various structures of the water hexamer relative to the lowest structure (prism). In each case, the geometry optimization was performed with the same method as the subsequent energy calculation, except for the CCSD(T) energies which were obtained at MP2-optimized geometries. The WHBB and HBB2-pol results are taken from the supplementary material of Refs. 47 and 53, respectively. The CCSD(T) results in the haTZ basis are from Ref. 66. The results denoted as CCSD(T)/CBS are from Ref. 71 and were computed using the MP2-R12 method plus  $\delta E_{\text{int}}^{\text{CCSD(T)}/\text{haTZ}}$ . The flexible-monomer results include monomer-relaxation corrections.

performance partly reflects the fact that several trimers similar to those present in hexamer structures were used in the development of CCpol23+, as described in Sec. II (the trimers from the actual hexamer structures used in Fig. 3 were not part of our fit data set: it contained trimers from TIP4P hexamers and distorted trimers from CC-pol-8s+NB optimizations of the hexamers). This level of agreement is remarkable since we compare here relaxed interaction energies of the benchmarks at flexible-monomer geometries with vertical interaction energies of CCpol23+ at rigid-monomer geometries. As discussed in detail for the trimer case, one reason is that the lowering of the total energy resulting from accounting for the monomer-flexibility effects is partly canceled by the subtracting the equilibrium isolated monomer energies.

The HBB2-pol potential<sup>53</sup> works best among the published potentials. It predicts consistently somewhat too large gaps between the prism and the other structures, with errors (relative to the CBS benchmark) in the range of 0.14–0.35 kcal/mol, about an order of magnitude larger than the errors of CCpol23+. The errors of the WHBB6 potential<sup>47</sup> are in the range of 0.25–1.40 kcal/mol, i.e., a few times larger than in the case of HBB2-pol. The errors of the WHBB5 potential<sup>47</sup> are still larger, although for the cage it performs slightly better than WHBB6.

The CC-pol-8s+NB potential gives predictions generally of similar quality to WHBB5, except for the cage structure where it performs worse. Thus, the CC-pol-8s+NB potential fares much worse here than in the trimer tests. Since the two-body components of CC-pol-8s+NB and of CCpol23+ are almost identical and four- and higher-body effects are too small to account for this effect, the bulk of the difference must stem from the lower quality of the three-body fit of Ref. 28.

We were able to understand this behavior by analyzing the results for the trimers extracted from the cage and prism hexamers (see the extended version of Table II in the supplementary material<sup>74</sup>). CC-pol-8s+NB performs well on stationary states of the trimer since all such states are similar to the minimum structure that was well represented in the training set of Ref. 28 and most of them lie on the tunneling paths also extensively explored in Refs. 24 and 28. By contrast, some of the trimers present in hexamer structures were virtually absent from the training set of Ref. 28. In particular, the two trimers forming the top and the bottom of the prism hexamer contain one water molecule that is a double-donor of hydrogen bonds. These trimers are poorly predicted by SAPT-3B, with errors with respect to the CBS benchmarks of  $-0.40$  and  $-0.32$  kcal/mol, respectively. These two trimers alone make the prism energy computed from the CC-pol-8s+NB potential significantly too negative and—with the cage energy not suffering of this problem—lead to an excessive cage-prism gap.

In contrast to the comparisons for the trimers, the performance of CCpol23+ relative to flexible-monomer literature potentials, demonstrated in Fig. 3, could not have been anticipated. For the trimers, such comparisons were made for rigid monomers (except for some comparisons to the *ab initio* benchmarks at the trimer characteristic points). Therefore, the good performance of CCpol23+ was expected since it is easier to fit a 12-dimensional than a 21-dimensional potential. For the hexamer, the rigid-monomer CCpol23+ potential achieves a better agreement with the flexible-monomer benchmarks than any flexible-monomer potential. This fact leads to the following answer to the question posed in the Introduction: monomer flexibility effects are less important than an



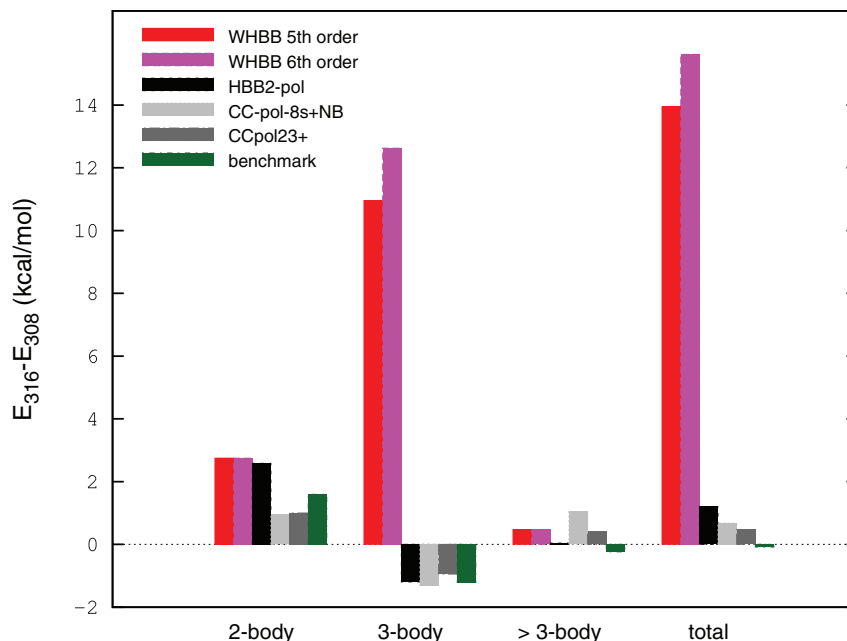


FIG. 4. Differences of the  $K$ -body and total interaction energies between the “316” and “308” structures of the water 24-mer. The benchmark *ab initio* results are taken from Ref. 37. The  $K$ -body contributions shown are in all cases the vertical interaction energies. The  $K > 3$ -body contribution is the pure four-body effect in the case of the benchmark. The total interaction energy differences from the flexible-monomer potentials as well as the *ab initio* benchmarks for this quantity include the one-body term. The WHBB and HBB2-pol values were obtained by us at the geometries of Ref. 91. The CC-pol-8s+NB and CCpol23+ results were obtained for the “rigidized” structures, i.e., with all monomers within the clusters transformed to their  $(r)_0$ -monomer geometries and the total interaction energy differences for these potentials do not include any one-body terms.

accurate description of each  $K$ -body contribution at the rigid-monomer level with the current state-of-the-art of the *ab initio* methods, at least for cluster equilibrium structures.

### C. 24-mer

The  $(\text{H}_2\text{O})_{24}$  cluster was the subject of an extensive CCSD(T) calculation within the conventional supermolecular framework, i.e., it included the calculation of the whole cluster energy, which took 76 years of CPU time.<sup>90,91</sup> The authors of that work identified two energetically low structures, labeled 308 and 316, and found their total energies differing by only 0.01 kcal/mol, with structure 316 being more stable. As impressive as this calculation is from the computational point of view, due to the small basis set used (cc-pVTZ with the  $f$  functions removed) and the lack of a CP correction, its predictive value is rather limited. An alternative way of calculating the energies of large clusters, the “stratified approximation” many-body approach (SAMBA), was proposed and applied to several water clusters by Góra *et al.*<sup>37</sup> The idea of this method consists of calculating  $K$ -body contributions to the interaction energy (with  $K \geq 1$ ) separately, in basis sets limited to  $K$  monomers in each case, thus avoiding the use of the exceedingly large basis sets of the whole cluster. Only low- $K$  contributions need to be considered since the many-body expansion converges fast. In Ref. 37, the consecutive  $K$ -body contributions to the energy difference  $E_{316} - E_{308}$  were found from *ab initio* calculations to be  $-0.23$ ,  $1.60$ ,  $-1.21$ , and  $-0.24$  kcal/mol for  $K = 1, 2, 3$ , and  $4$ , respectively, giving the sum equal to  $-0.08$  kcal/mol. Although the uncertainty of the total energy predictions for each structure was estimated

in Ref. 37 to be 4.8 kcal/mol, the analysis of the respective differences shows that the value of  $-0.08$  kcal/mol is probably accurate to within 0.1–0.2 kcal/mol. The remaining (higher than four-body) contributions to  $E_{316} - E_{308}$  were estimated to be probably around 0.3 kcal/mol and not higher than 0.6 kcal/mol. Thus, the SAMBA method provides a rather accurate estimate of the value of  $E_{316} - E_{308}$  for such a large cluster although, due to the almost equal energies, the question about the most stable structure remains open. Although the total CPU time of the SAMBA calculations for  $(\text{H}_2\text{O})_{24}$  was 200 times shorter than the CPU time needed for the calculations in Ref. 91, it was still a large computational effort. It is therefore an important question how accurate can be the predictions of first-principles water potentials applied to such clusters, as such results can be obtained in mere seconds.

Figure 4 shows the differences of the  $K$ -body contributions and total interaction energies between the two 24-mers. All  $K$ -body contributions shown are vertical whereas the total energy differences are relaxed for flexible-monomer approaches and vertical for rigid-monomers approaches. The total values of  $E_{316} - E_{308}$  are 13.95, 15.62, 1.20, 0.68, and 0.47 kcal/mol from the WHBB5, WHBB6, HBB2-pol, CC-pol-8s+NB, and CCpol23+ potentials, respectively. When we take into account the maximum uncertainty estimate of 0.8 kcal/mol for the benchmark value discussed above, it is seen that both the CC-pol-8s+NB and CCpol23+ potentials are within the benchmark error bars, HBB2-pol is reasonably close, while the WHBB potentials are far from it (again, as for the trimer stationary points, the 6th-degree fit performs worse than the simpler one). Since the  $(\text{H}_2\text{O})_{24}$  cluster contains a large number of trimers (2024), the inadequacies of the



TABLE VI. Many-body contributions (in kcal/mol) to the interaction energies of the 316 and 308 water 24-mers. SAMBA results are from Ref. 37.

	2-B	3-B	4-B	5-B	>5-B	>4-B	$\sum_2^4$	$\sum_1^4$
Isomer 316								
CC-pol-8s+NB	-183.856	-45.675	-6.664	-0.349	0.086	-0.263	-236.195	
CCpol23+	-183.888	-51.251	-6.498	-0.320	0.075	-0.245	-241.637	
SAMBA	-191.532	-59.002	-7.001				-257.534	-238.719
Isomer 308								
CC-pol-8s+NB	-184.800	-44.365	-7.381	-0.611	0.019	-0.592	-236.546	
CCpol23+	-184.889	-50.310	-6.722	-0.461	0.030	-0.431	-241.921	
SAMBA	-193.128	-57.791	-6.765				-257.864	-238.638
316-308								
CC-pol-8s+NB	0.945	-1.310	0.716	0.262	0.067	0.329	0.351	
CCpol23+	1.001	-0.942	0.224	0.141	0.045	0.186	0.284	
SAMBA	1.596	-1.211	-0.236				0.150	-0.081

three-body WHBB fits add up to the values of +12.2 kcal/mol in the case of WHBB5 and +13.8 kcal/mol in the case of WHBB6.

Table VI analyzes the many-body expansion for the 24-mer in more detail, in particular we give the four- and higher-body terms separately for each isomer. We have not included the flexible-monomer potentials from the literature since their polarization models are of similar complexity as the 1-center model included in CC-pol-8s. The many-body expansion was truncated in Ref. 37 at  $K = 4$  (the SAMBA contribution in the set of columns denoted in Fig. 4 as “>3-body” is the pure four-body contribution), so only terms up this  $K$  can be compared. As one can see in Table VI, the two-body energies of both isomers are nearly identical for the two potentials and differ from SAMBA results by about 8 kcal/mol or 4%. The 8 kcal/mol differences cancel to a fraction of kcal/mol in the 316–308 difference quantities. In the three-body case, the predictions of the CCpol23+ agree with SAMBA significantly better than those of CC-pol-8s+NB, the discrepancies are about 7–8 and 13 kcal/mol, respectively, and again a significant cancellation of errors takes place in the calculations of difference quantities. In these comparisons, one should take into account that all the energies in Table VI (except for the last column) are vertical, whereas the geometries are different in the SAMBA and potential calculations.

The last column of Table VI lists the relaxed interaction energies. The average deformation correction per monomer of 0.8 kcal/mol is about six times larger than in the global minimum of the water trimer. One reason is that a large fraction of monomers in the investigated 24-mers have both hydrogens participating in hydrogen bonds, which leads to elongation of both intramonomer OH bonds. In contrast, in the trimer only one OH bond is elongated. Second, this elongation increases with cluster size in compact clusters (the OH bond length in ice is 1.01 Å). Finally, the optimizations of 24-mers were performed<sup>92</sup> at the MP/aDZ level and the length of the isolated monomer OH bond is overestimated at this level by 0.006 Å.

We have already seen the very good performance of both polarization models on the tetramers extracted from hexamers, as discussed in Sec. V. This performance is equally good for the 24-mer, with errors of the three-center model amount-

ing to 7% for isomer 316 and 0.6% for 308. Although the 316–308 difference in the four-body contribution predicted by CCpol23+ is of different sign than predicted by SAMBA, the magnitude of the discrepancy amounts only to 0.4 kcal/mol, i.e., is rather small.

If the terms with  $K = 2-4$  are added together, the differences between the CCpol23+ and SAMBA predictions, i.e., the sums of differences discussed above, are about 16 kcal/mol. However, if the monomer distortion corrections from Ref. 37 are added to the SAMBA results, the discrepancies decrease to only about 3 kcal/mol.

For the  $K > 4$ -body effect, we do not have *ab initio* results to compare with. However, in view of the good performance at the four-body level and the previous observations for the hexamer, the  $N$ -body polarization models included in the CC-pol-8s and CCpol23+ potentials should provide reasonable estimates of these nonadditive many-body effects. Table VI shows that the five- and higher-body effects give a very small contribution to the total interaction energies of both isomers, on the order of 0.1%. Thus, the magnitude of these estimates is in agreement with the estimates made in Ref. 37. In contrast to the lower- $K$  contributions, the high- $K$  terms do not cancel out between the 316 and 308 isomers and the difference is of the same order of magnitude as the contributions for individual monomers. One may be tempted to add these differences to the SAMBA prediction (which was truncated at four-body interactions and with the one-body term included gives  $E_{316} - E_{308} = -0.08$  kcal/mol). Such an addition gives the amended SAMBA value of  $E_{316} - E_{308} = +0.11$  kcal/mol, in even better agreement with the CCpol23+ value of  $E_{316} - E_{308} = +0.47$  kcal/mol.

The results for the water 24-mer show that the performance of the CCpol23+ potential in predictions for clusters such as  $(\text{H}_2\text{O})_{24}$  is competitive with the most advanced applicable *ab initio* methods, while the latter require (in this case) an about seven orders of magnitude larger computational effort. Analogously to the case of the hexamer, it should be pointed out that the inclusion of the monomer-flexibility effects makes little difference in the quality of predictions, despite the large deformations of monomers. The CC-pol-8s+NB and CCpol23+ potentials give such good predictions despite being evaluated at the rigidized geometries of Ref. 91.

### VIII. SUMMARY AND CONCLUSIONS

A new first-principles three-body pairwise-nonadditive interaction energy potential has been developed for water using the rigid-monomer approximation with monomers in their average rovibrational ground-state geometry. This 12-dimensional surface was fitted to 71 456 *ab initio* three-body nonadditive interaction energies obtained using a hybrid approach that combines results computed at the CCSD(T) level using the aug-cc-pVDZ basis with those computed at the MP2 level using the aug-cc-pVTZ basis. This level of *ab initio* theory gives more than sufficient accuracy as the RMSEs relative to CBS benchmarks on trimers extracted from equilibrium hexamers are only 0.004 and 0.006 kcal/mol for the cage and prism hexamers, respectively.

The functional form of the fit was motivated by the SAPT decomposition of the nonadditive energy into physical components. It included a new, three-center damped polarization model. This model alone recovered relatively well the three-body nonadditive effects as its RMSE on 600 trimers selected randomly from snapshots of MD simulations in ambient conditions was 0.073 kcal/mol. This accuracy is similar to that of some recent potentials fitted to *ab initio* calculations such as WHBB5<sup>47</sup> and significantly better than that of single-center polarization potentials such as those used in Ref. 13 which gave a RMSE of 0.107 kcal/mol. This polarization model was combined with terms describing exchanges of electrons between monomers. Significant care was taken to ensure that the fit behaves physically at very short separations between nuclei, a feature important for molecular simulations. The final fit, denoted as CCpol3, uses 63 nonlinear and 364 linear adjustable parameters and its RMSE on the subset of 70 268 grid points that excludes geometries with very small intermonomer separations is 0.0145 kcal/mol. To make our three-body potential consistent with the two-body one, the CC-pol-8s potential from Ref. 43 was refitted using the new polarization model and additional damping functions for very short separations. The accuracy of the new two-body fit, denoted as CCpol2, is unchanged compared to CC-pol-8s in the physical region.

As a byproduct of this work, we have found that the polarization model recovers very well the four-body nonadditive interaction energies, with errors on the order of 10% rather than the 50% errors observed in Ref. 13 in the three-body case. Apparently, this very positive fact has not been noted earlier: it is of significant importance in developing many-body force fields. In the case of water hexamers, we have also compared the performance of the polarization model for five- and six-body contributions. Whereas this performance was significantly worse than for the four-body terms, the worst performance was for contributions of negligible magnitude. The sum of the four- through six-body effects was recovered by the polarization model of CCpol23+ with the average magnitude of the relative error over the six isomers amounting to only 3.5%.

CCpol3 was first tested on the nonadditive interaction energies of water trimers. On the set of trimers contained in the cage and prism hexamers, the RMSE of the fit is 0.019 kcal/mol for each isomer. On the set of 600 hexam-

ers from MD simulations, the RMSE is 0.0154 kcal/mol, 2.4 times smaller than given by the HBB2-pol potential<sup>53</sup> which performed best of the published potentials.

We then tested the sum of CCpol3 and of CCpol2, denoted as CCpol23, on the total trimer interaction energies at the six trimer stationary geometries. The geometries optimized using CCpol23 were found to be close to those of the *ab initio* optimizations of Ref. 84 with flexible monomers after the latter geometries were rigidized. The RMSE of the CCpol23 predictions relative to the CBS benchmarks computed by us is only 0.11 kcal/mol and it is a few times smaller than given by the best potentials from the literature.

The remaining tests were performed on the water hexamer and 24-mer using the CCpol23+ potential, i.e., CCpol23 plus the four- and higher-body interactions represented by the polarization model. In contrast to the tests described so far in this section, we compared to cluster energies computed using flexible-monomer approaches for those methods that do not use rigid-monomer approximation. Somewhat surprisingly, despite the rigid-monomer approximation, the energetic predictions of CCpol23+ are closer to the best *ab initio* benchmarks (which include monomer-flexibility effects) than the predictions of the best published flexible-monomer models. We believe that the main reason for this behavior is the high accuracy of our three-body potential. For hexamers, the energies of the five considered isomers relative to the energy of the lowest-energy prism isomer are within 0.01–0.03 kcal/mol of the Bates-Tschumper<sup>71</sup> benchmarks, i.e., are within the uncertainties of the *ab initio* calculations. The second-best prediction is given by the HBB2-pol potential, but the errors of 0.14–0.35 kcal/mol are larger by an order of magnitude. For the 24-mer, only CCpol23+ predicts the difference between the two lowest-energy “308” and “316” isomers, estimated from *ab initio* calculations to be  $-0.1 \pm 0.8$  kcal/mol, within the error bars. It gives the value of 0.47 kcal/mol for this difference, whereas the best performing flexible-monomer potential, HBB2-pol, predicts 1.20 kcal/mol.

The comparison for hexamers and 24-mers suggests that at the currently possible level of accuracy, the residual errors in intermolecular part of flexible-monomer potentials are larger than the monomer-flexibility effects on cluster energetics. Hence, at this point it is more advantageous from a physical point of view to improve the accuracy of trimer rigid-monomer potentials rather than to develop flexible-monomer potentials which cannot be sufficiently accurate in their dependence on the intermolecular coordinates due to a too low density of grid points. This is in contrast to the two-body case where one can now develop very accurate 12-dimensional potentials.<sup>41,44,48,52,86</sup>

Our work also indicates an issue concerning the accuracy of the potentials used in recent investigations of hexamer structures aimed at determining the energetic order of isomers.<sup>53,64,65</sup> Since the difference in the interaction energies of the prism and cage isomers is as small as 0.25 kcal/mol,<sup>37,71</sup> only CCpol23+ is sufficiently accurate among existing potentials to correctly recover this quantity.

Since the CCpol23+ accuracy for the hexamer is close to that of the state-of-the-art *ab initio* benchmarks, whereas for the 24-mer it is likely higher, this potential can be used to

generate high-accuracy benchmarks for large water clusters with costs completely negligible compared to any *ab initio* calculations. We also expect this potential to find broad applications in predicting the properties of the condensed phases of water, including the anomalous properties. The results for the clusters suggest that the rigid-monomer character of the potential does not impair the accuracy of the CCpol23+ predictions significantly, although for clusters with particularly large distortions of the monomer geometries, the potential may not work that well.

*Note added in proof:* A recent paper by Babin *et al.* developed a three-body flexible-monomer water potential based on 12 000 interaction energies.<sup>93</sup>

## ACKNOWLEDGMENTS

The authors would like to thank Omololu Akin-Ojo for several discussions and for providing the set of trimers for MD simulations. This research was supported by the National Science Foundations (NSF) Grant No. CHE-1152899. R.P. would like to acknowledge support from Polish National Science Centre Grant No. DEC-2012/05/B/ST4/00086.

- <sup>1</sup>O. Matsuoka, E. Clementi, and M. Yoshimine, *J. Chem. Phys.* **64**, 1351 (1976).
- <sup>2</sup>B. Jeziorski and M. van Hemert, *Mol. Phys.* **31**, 713 (1976).
- <sup>3</sup>K. Szalewicz, S. J. Cole, W. Kolos, and R. J. Bartlett, *J. Chem. Phys.* **89**, 3662 (1988).
- <sup>4</sup>E. M. Mas and K. Szalewicz, *J. Chem. Phys.* **104**, 7606 (1996).
- <sup>5</sup>W. Klopper, J. G. C. M. van Duijneveldt-van de Rijdt, and F. B. van Duijneveldt, *Phys. Chem. Chem. Phys.* **2**, 2227 (2000).
- <sup>6</sup>G. S. Tschumper, M. L. Leininger, B. C. Hoffman, E. F. Valeev, H. F. Schaefer III, and M. Quack, *J. Chem. Phys.* **116**, 690 (2002).
- <sup>7</sup>J. R. Lane, *J. Chem. Theory Comput.* **9**, 316 (2013).
- <sup>8</sup>W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- <sup>9</sup>H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *J. Phys. Chem.* **91**, 6269 (1987).
- <sup>10</sup>K. Szalewicz, C. Leforestier, and A. van der Avoird, *Chem. Phys. Lett.* **482**, 1 (2009).
- <sup>11</sup>L.-P. Wang, T. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, and V. S. Pande, *J. Phys. Chem. B* **117**, 9956 (2013).
- <sup>12</sup>K. T. Wikfeldt, E. R. Batista, F. D. Vilazc, and H. Jonsson, *Phys. Chem. Chem. Phys.* **15**, 16542 (2013).
- <sup>13</sup>O. Akin-Ojo and K. Szalewicz, *J. Chem. Phys.* **138**, 024316 (2013).
- <sup>14</sup>R. S. Fellers, C. Leforestier, L. B. Braly, M. G. Brown, and R. J. Saykally, *Science* **284**, 945 (1999).
- <sup>15</sup>R. S. Fellers, L. B. Braly, R. J. Saykally, and C. Leforestier, *J. Chem. Phys.* **110**, 6306 (1999).
- <sup>16</sup>N. Goldman, R. S. Fellers, M. G. Brown, L. B. Braly, C. J. Keoshian, C. Leforestier, and R. J. Saykally, *J. Chem. Phys.* **116**, 10148 (2002).
- <sup>17</sup>N. Goldman, C. Leforestier, and R. J. Saykally, *Philos. Trans. R. Soc. London, Ser. A* **363**, 493 (2005).
- <sup>18</sup>K. Szalewicz, R. Bukowski, and B. Jeziorski, in *Theory and Applications of Computational Chemistry: The First 40 Years. A Volume of Technical and Historical Perspectives*, edited by C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria (Elsevier, Amsterdam, 2005), Chap. 33, pp. 919–962.
- <sup>19</sup>E. Clementi and P. Habitz, *J. Phys. Chem.* **87**, 2815 (1983).
- <sup>20</sup>U. Niesar, G. Corongiu, E. Clementi, G. R. Kneller, and D. K. Bhattacharya, *J. Phys. Chem.* **94**, 7949 (1990).
- <sup>21</sup>C. Millot and A. J. Stone, *Mol. Phys.* **77**, 439 (1992).
- <sup>22</sup>E. M. Mas, K. Szalewicz, R. Bukowski, and B. Jeziorski, *J. Chem. Phys.* **107**, 4207 (1997).
- <sup>23</sup>C. Millot, J. C. Soetens, M. T. C. M. Costa, M. P. Hodges, and A. J. Stone, *J. Phys. Chem. A* **102**, 754 (1998).
- <sup>24</sup>G. C. Groenenboom, E. M. Mas, R. Bukowski, K. Szalewicz, P. E. S. Wormer, and A. van der Avoird, *Phys. Rev. Lett.* **84**, 4072 (2000).
- <sup>25</sup>E. M. Mas, R. Bukowski, K. Szalewicz, G. C. Groenenboom, P. E. S. Wormer, and A. van der Avoird, *J. Chem. Phys.* **113**, 6687 (2000).
- <sup>26</sup>G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **113**, 6702 (2000).
- <sup>27</sup>M. J. Smit, G. C. Groenenboom, P. E. S. Wormer, A. van der Avoird, R. Bukowski, and K. Szalewicz, *J. Phys. Chem. A* **105**, 6212 (2001).
- <sup>28</sup>E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4386 (2003).
- <sup>29</sup>R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *Science* **315**, 1249 (2007).
- <sup>30</sup>R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 094313 (2008).
- <sup>31</sup>R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 094314 (2008).
- <sup>32</sup>A. Shank, Y. Wang, A. Kaledin, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **130**, 144314 (2009).
- <sup>33</sup>B. E. Rocher-Casterline, L. C. Ch'ng, A. K. Mollner, and H. Reisler, *J. Chem. Phys.* **134**, 211101 (2011).
- <sup>34</sup>O. Mishima and H. E. Stanley, *Nature (London)* **392**, 164 (1998).
- <sup>35</sup>K. Szalewicz and B. Jeziorski, *J. Chem. Phys.* **109**, 1198 (1998).
- <sup>36</sup>E. M. Mas, R. Bukowski, and K. Szalewicz, *J. Chem. Phys.* **118**, 4404 (2003).
- <sup>37</sup>U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, *J. Chem. Phys.* **135**, 224102 (2011).
- <sup>38</sup>C. J. Burnham and S. S. Xantheas, *J. Chem. Phys.* **116**, 1500 (2002).
- <sup>39</sup>C. J. Burnham and S. S. Xantheas, *J. Chem. Phys.* **116**, 5115 (2002).
- <sup>40</sup>G. S. Fanourgakis and S. S. Xantheas, *J. Phys. Chem. A* **110**, 4100 (2006).
- <sup>41</sup>K. Szalewicz, G. Murdachaew, R. Bukowski, O. Akin-Ojo, and C. Leforestier, in *Lecture Series on Computer and Computational Science: ICCMSE 2006*, edited by G. Maroulis, and T. Simos (Brill Academic Publishers, Leiden, 2006), Vol. 6, pp. 482–491.
- <sup>42</sup>R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **125**, 044301 (2006).
- <sup>43</sup>W. Cencek, K. Szalewicz, C. Leforestier, R. van Harrevelt, and A. van der Avoird, *Phys. Chem. Chem. Phys.* **10**, 4716 (2008).
- <sup>44</sup>X. Huang, B. J. Braams, and J. M. Bowman, *J. Phys. Chem. A* **110**, 445 (2006).
- <sup>45</sup>X. Huang, B. J. Braams, J. M. Bowman, R. E. A. Kelly, J. Tennyson, G. C. Groenenboom, and A. van der Avoird, *J. Chem. Phys.* **128**, 034312 (2008).
- <sup>46</sup>Y. Wang, B. C. Shepler, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **131**, 054511 (2009).
- <sup>47</sup>Y. Wang, X. Hunag, B. C. Shepler, B. J. Braams, and J. M. Bowman, *J. Chem. Phys.* **134**, 094509 (2011).
- <sup>48</sup>C. Leforestier, K. Szalewicz, and A. van der Avoird, *J. Chem. Phys.* **137**, 014305 (2012).
- <sup>49</sup>V. Babin, G. R. Medders, and F. Paesani, *J. Phys. Chem. Lett.* **3**, 3765 (2012).
- <sup>50</sup>G. R. Medders, V. Babin, and F. Paesani, *J. Chem. Theory Comput.* **9**, 1103 (2013).
- <sup>51</sup>G. R. Medders and F. Paesani, *J. Chem. Theory Comput.* **9**, 4844 (2013).
- <sup>52</sup>V. Babin, C. Leforestier, and F. Paesani, *J. Chem. Theory Comput.* **9**, 5395 (2013).
- <sup>53</sup>V. Babin and F. Paesani, *Chem. Phys. Lett.* **580**, 1 (2013).
- <sup>54</sup>M. Jeziorska, P. Jankowski, K. Szalewicz, and B. Jeziorski, *J. Chem. Phys.* **113**, 2957 (2000).
- <sup>55</sup>B. Jeziorski, R. Moszyński, and K. Szalewicz, *Chem. Rev.* **94**, 1887 (1994).
- <sup>56</sup>K. Szalewicz, *WIREs: Comput. Mol. Sci.* **2**, 254 (2012).
- <sup>57</sup>V. F. Lotrich and K. Szalewicz, *J. Chem. Phys.* **106**, 9668 (1997).
- <sup>58</sup>V. F. Lotrich and K. Szalewicz, *J. Phys. Chem.* **106**, 9688 (1997).
- <sup>59</sup>V. F. Lotrich and K. Szalewicz, *Phys. Rev. Lett.* **79**, 1301 (1997).
- <sup>60</sup>A. van der Avoird and K. Szalewicz, *J. Chem. Phys.* **128**, 014302 (2008).
- <sup>61</sup>T. James, D. Wales, and J. Hernández-Rojas, *Chem. Phys. Lett.* **415**, 302 (2005).
- <sup>62</sup>M. Mahoney and W. Jorgensen, *J. Chem. Phys.* **112**, 8910 (2000).
- <sup>63</sup>W. Cencek, M. Jeziorska, O. Akin-Ojo, and K. Szalewicz, *J. Phys. Chem. A* **111**, 11311 (2007).
- <sup>64</sup>C. Perez, M. T. Muckle, D. P. Zaleski, N. A. Seifert, B. Temelso, G. C. Shields, Z. Kisiel, and B. H. Pate, *Science* **336**, 897 (2012).
- <sup>65</sup>Y. M. Wang, V. Babin, J. M. Bowman, and F. Paesani, *J. Am. Chem. Soc.* **134**, 11116 (2012).
- <sup>66</sup>E. E. Dahlke, R. M. Olson, H. R. Leverentz, and D. G. Truhlar, *J. Phys. Chem. A* **112**, 3976 (2008).

- <sup>67</sup>M. J. D. Powell, *Comput. J.* **7**, 155 (1964).
- <sup>68</sup>H.-J. Werner, P. J. Knowles, R. Lindh, M. Schütz *et al.*, MOLPRO, version 2009.1, a package of *ab initio* programs, 2009, see <http://www.molpro.net>.
- <sup>69</sup>S. F. Boys and F. Bernardi, *Mol. Phys.* **19**, 553 (1970).
- <sup>70</sup>R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, *J. Chem. Phys.* **96**, 6796 (1992).
- <sup>71</sup>D. M. Bates and G. S. Tschumper, *J. Phys. Chem. A* **113**, 3555 (2009).
- <sup>72</sup>A. Halkier, W. Klopper, T. Helgaker, P. Jørgensen, and P. R. Taylor, *J. Chem. Phys.* **111**, 9157 (1999).
- <sup>73</sup>A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, and J. Olsen, *Chem. Phys. Lett.* **302**, 437 (1999).
- <sup>74</sup>See supplementary material at <http://dx.doi.org/10.1063/1.4875097> for detailed cluster geometries, energies, and a Fortran program for the potential.
- <sup>75</sup>D. E. Woon and T. H. Dunning, Jr., *J. Chem. Phys.* **103**, 4572 (1995).
- <sup>76</sup>K. T. Tang and J. P. Toennies, *J. Chem. Phys.* **80**, 3726 (1984).
- <sup>77</sup>A. J. Misquitta, private communication (2009).
- <sup>78</sup>A. J. Misquitta and A. J. Stone, "CamCASP: A program for studying intermolecular interactions and for calculations of molecular properties in distributed form," University of Cambridge, UK, 2010.
- <sup>79</sup>J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- <sup>80</sup>C. Adamo and V. Barone, *J. Chem. Phys.* **110**, 6158 (1999).
- <sup>81</sup>D. E. Woon and T. H. Dunning, Jr., *J. Chem. Phys.* **100**, 2975 (1994).
- <sup>82</sup>V. F. Lotrich, H. L. Williams, K. Szalewicz, B. Jeziorski, R. Moszynski, P. E. S. Wormer, and A. van der Avoird, *J. Chem. Phys.* **103**, 6076 (1995).
- <sup>83</sup>R. Kumar, F. F. Wang, G. R. Jenness, and K. D. Jordan, *J. Chem. Phys.* **132**, 014309 (2010).
- <sup>84</sup>J. A. Anderson, K. Crager, L. Fedoroff, and G. S. Tschumper, *J. Chem. Phys.* **121**, 11023 (2004).
- <sup>85</sup>G. S. Tschumper, see [http://quantum.chem.olemiss.edu/026\\_DATA](http://quantum.chem.olemiss.edu/026_DATA) for geometries of water trimers of Ref. 84.
- <sup>86</sup>P. Jankowski, G. Murdachaew, R. Bukowski, O. Akin-Ojo, C. Leforestier, and K. Szalewicz, "Ab initio water pair potential with flexible monomers," *J. Phys. Chem. A* (submitted).
- <sup>87</sup>K. Liu, M. G. Brown, C. Carter, R. J. Saykally, J. K. Gregory, and D. C. Clary, *Nature (London)* **381**, 501 (1996).
- <sup>88</sup>S. S. Xantheas, C. J. Burnham, and R. J. Harrison, *J. Chem. Phys.* **116**, 1493 (2002).
- <sup>89</sup>W. Kutzelnigg and W. Klopper, *J. Chem. Phys.* **94**, 1985 (1991).
- <sup>90</sup>E. Apra, R. J. Harrison, W. A. deJong, A. P. Rendell, V. Tipparaju, and S. S. Xantheas, in *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis* (ACM, New York, 2010), Article No. 66.
- <sup>91</sup>E. Apra, R. J. Harrison, W. A. deJong, A. P. Rendell, V. Tipparaju, and S. S. Xantheas, in *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis* (ACM, New York, 2010), Article No. 66 (extended version).
- <sup>92</sup>S. Yoo, M. V. Kirov, and S. S. Xantheas, *J. Am. Chem. Soc.* **131**, 7564 (2009).
- <sup>93</sup>V. Babin, G. R. Medders, and F. Paesani, *J. Chem. Theory. Comput.* **10**, 1599 (2014).