Acta Geochim
https://doi.org/10.1007/s11631-024-00680-8

ORIGINAL ARTICLE

Machine learning prediction of methane, ethane, and propane solubility in pure water and electrolyte solutions: Implications for stray gas migration modeling

Ghazal Kooti[1,2] · **Reza Taherdangkoo**[1,3] · **Chaofan Chen**[1,3] · **Nikita Sergeev**[1] · **Faramarz Doulati Ardejani**[4] · **Tao Meng**[5] · **Christoph Butscher**[1,3]

Received: 8 January 2024 / Revised: 11 February 2024 / Accepted: 15 February 2024

Abstract Hydraulic fracturing is an effective technology for hydrocarbon extraction from unconventional shale and tight gas reservoirs. A potential risk of hydraulic fracturing is the upward migration of stray gas from the deep subsurface to shallow aquifers. The stray gas can dissolve in groundwater leading to chemical and biological reactions, which could negatively affect groundwater quality and contribute to atmospheric emissions. The knowledge of light hydrocarbon solubility in the aqueous environment is essential for the numerical modelling of flow and transport in the subsurface. Herein, we compiled a database containing 2129 experimental data of methane, ethane, and propane solubility in pure water and various electrolyte solutions over wide ranges of operating temperature and pressure. Two machine learning algorithms, namely regression tree (RT) and boosted regression tree (BRT) tuned with a Bayesian optimization algorithm (BO) were employed to determine the solubility of gases. The predictions were compared with the experimental data as well as four well-established thermodynamic models. Our analysis shows that the BRT-BO is sufficiently accurate, and the predicted values agree well with those obtained from the thermodynamic models. The coefficient of determination (R^2) between experimental and predicted values is 0.99 and the mean squared error (MSE) is 9.97×10^{-8}. The leverage statistical approach further confirmed the validity of the model developed.

Keywords Gas solubility · Hydraulic fracturing · Thermodynamic models · Regression tree · Boosted regression tree · Groundwater contamination

1 Introduction

Hydrocarbon production from unconventional resources has become the focus of attention in the oil and gas industry in the last decades. Due to the exponential growth of energy demand, advances in horizontal drilling technologies, and multi-stage hydraulic fracturing operations, developing unconventional reservoirs has become highly attractive (Taherdangkoo et al. 2019). Shale gas and tight gas reservoirs as the main unconventional resources have extremely low matrix permeability and even the existence of natural fracture networks could not provide flow paths from formation to the wells (King 2012; Kissinger et al. 2013). Reservoir stimulation techniques such as hydraulic fracturing could effectively increase the ability of natural gas recovery from such reservoirs. Hydraulic fracturing improves access to the larger part of the reservoir by creating artificial fracture networks which increase the reservoir permeability and the contact areas over which fluids flow from the matrix to the fractures (Tatomir et al. 2018; Rice et al. 2018).

Shale gas is typically dry gas composed primarily of methane with traces of ethane and propane (King 2012). The gas is stored in three ways including absorbed in the limited pore spaces of these rocks, adsorbed on the surface of organic material, or confined in the natural fractures and

✉ Reza Taherdangkoo
reza.taherdangkoo@ifgt.tu-freiberg.de

[1] Institute of Geotechnics, TU Bergakademie Freiberg, Gustav-Zeuner-Str. 1, 09599 Freiberg, Germany

[2] Department of Petroleum Engineering, Amirkabir University of Technology, Tehran, Iran

[3] Freiberg Center for Water Research ZeWaF, TU Bergakademie Freiberg, 09599 Freiberg, Germany

[4] School of Mining, College of Engineering, University of Tehran, Tehran, Iran

[5] Taiyuan University of Science and Technology, Taiyuan 030024, China

Published online: 22 March 2024

fissures. Therefore, gas production from shale gas reservoirs is more complex in comparison to conventional gas reservoirs (King 2012; Siegel et al. 2015). Several environmental concerns have emerged surrounding shale gas development, such as groundwater contamination, induced seismicity, climate impact from leaked stray gas into the atmosphere, etc. The main risk difference in comparison with other technologies in the subsurface is that hydraulic fracturing is remunerative, thus it is necessary to distinguish between economic and environmental issues (Cahill et al. 2017; Rice et al. 2018; Taherdangkoo et al. 2020b).

The occurrence of light hydrocarbons in groundwater in the vicinity of oil and gas operations is most commonly associated with leakage from hydrocarbon wells (Jackson et al. 2013; Nowamooz et al. 2015; Taherdangkoo et al. 2020a). Natural and anthropogenic permeable pathways such as leaky oil and gas abandoned wells, fault zones and extensive fracture systems could facilitate stray gas migration and an early gas manifestation in groundwater wells (Cahill et al. 2017; Rice et al. 2018; Tatomir et al. 2018). The presence of low-permeability layers leads to gas migration along higher permeability sediments, either up-dip or in the direction of groundwater flow, delaying the breakthrough to the shallow aquifer system (Taherdangkoo et al. 2020a).

Numerical modelling of subsurface flow and transport of stray gas plays an essential role in better evaluating the relationship between groundwater quality and hydrocarbon development. Modelling of stray gas migration requires phase equilibrium calculations to obtain the concentration of each component in liquid and gas phases. The calculation of gas solubility is usually performed by solving thermodynamic equations, which contributes a significant portion of the overall computational cost because compositions of liquid and gas phases must be calculated at each iteration. Therefore, due to the high computational time, an efficient application of complex thermodynamic models in numerical simulation of coupled multi-physics processes in the subsurface is laborious (Grunwald et al. Grunwald et al. 2020).

As an alternative to complex thermodynamic models, machine learning (ML) can be employed to describe the phase behaviour of gas-water-salt systems (Mohammadi et al. 2022; Taherdangkoo et al. 2021a). ML models can handle complex nonlinear relationships between inputs and outputs and can perform high precision interpolations (Qiao et al. 2020). Optimization algorithms can also be employed to tune hyper-parameters of ML algorithms and thus improve their overall performance (Taherdangkoo et al. 2023). Equations of state and predictive ML models estimate maximum gas solubility assuming perfect mixing and mass transfer between gas and aqueous phases. The stray gas migrating in the subsurface typically does not result in maximum solubility manifesting as the mass transfer from gas to the aqueous phase is limited in porous media (Cahill et al. 2018).

The primary goal of this study is to develop a robust ML model able to determine the solubility of light hydrocarbons (C_1–C_3) in aqueous solutions for further application in higher level numerical modelling. We used regression tree and boosted regression tree tuned with a Bayesian optimization algorithm to perform the regression task. The ML models were developed using a dataset containing 2129 experimental data of methane, ethane, and propane solubilities in pure water and aqueous and electrolyte systems. A comparison analysis was designed to evaluate the performance of the most accurate ML model with experimental data as well as Spivey et al. (2004), Duan and Mao (2006), Chapoy et al. (2004), and Pereda et al. (2009) thermodynamic models.

2 Data

We reviewed publicly available experimental data of light hydrocarbons (C_1–C_3) solubility in aqueous solutions to build a dataset to develop a robust machine learning model able to determine the gas solubility under a wide range of field conditions. The compiled dataset includes experimental gas solubility data measured between 1855 and 2007. There are instances in which experimental solubility data are incorrect or inconsistent with other gas solubility measurements. Following Duan et al. (1992), the inconsistent solubility data, e.g., methane solubility in pure water reported by Michels et al. (1936), were not considered in the dataset. The gas solubility has been reported in different units. Most of the experimental data in our dataset were originally reported in mole fraction, i.e., the mole fraction of the gas component in the liquid phase, and thus the remaining solubility values were converted to mole fraction using appropriate conversion functions. Where feasible, to avoid a conversion, the solubility data were taken from Clever and Young (1987), Hayduk (1982, 1986).

The solubility of methane in pure water and brine has been reported over a broad range of pressure, temperature, and NaCl concentrations (Clever and Young 1987), but limited measurements are available for saline solutions containing other salts. The dataset contains a total of 1912 methane solubility experimental data (Amirijafari et al. 1972; Ben-Naim et al. 1973; Blanco et al. 1978; Blount 1982; Bunsen 1855; Byrne and Stoessell 1982; Chapoy et al. 2004; Claussen and Polglase 1952; Cosgrove and Walkley 1981; Cramer 1984; Crovetto et al. 1982; Culberson et al. 1951; Dhima et al. 1998; Duffy et al. 1961; Eucken and Hertzberg 1950; Kiepe et al. 2003; Krader and Franck 1987; Lannung et al. 1960; Lekvam and Bishnoi 1997; Michels et al. 1936; Mishnina et al. 1962; Morrison and Billett 1952; Moudgil et al. 1974; Muccitelli and Wen 1980; Namiot 1961; O'Sullivan and Smith 1970; Rettich et al. 1981; Stoessell and Byrne 1982; Wang et al. 2003; Wen and Hung 1970; Wetlaufer et al. 1964; Winkler 1901; Yamamoto et al. 1976; Yano et al. 1974; Yarym-Agaev et al. 1985). Temperatures are in the range of 273.15 and 799 K, and pressure ranges from 1 to 2630 bar. The methane solubility in the aqueous phase is a function of pressure and temperature and concentrations of dissociated ions of NaCl, KCl, $CaCl_2$, $MgCl_2$, K_2SO_4, $MgSO_4$, Na_2SO_4, K_2SO_4, Na_2CO_3, K_2CO_3 in aqueous solutions. The statistical analysis of parameter values is summarized in Table 1, and their distributions in terms of histogram plots are presented in Fig. 1.

Table 1 Range of parameter values in methane solubility dataset

Parameter	Minimum	Maximum	Standard deviation	Mean
P (Bar)	1^+	2630	533.17	485.23
T (K)	273.15	799	112.82	398.85
Solubility (mole fraction)	3.53×10^{-6}	0.0185	0.002	0.002
NaCl (m)	0	5.4	1.71	1.38
KCl (m)	0	4	0.66	0.15
$CaCl_2$ (m)	0	7.35	0.47	0.07
$MgCl_2$ (m)	0	2.16	0.11	0.009
K_2SO_4 (m)	0	0.25	0.006	0.0002
$MgSO_4$ (m)	0	1.5	0.08	0.006
Na_2SO_4 (m)	0	1	0.057	0.004
K_2SO_4 (m)	0	0.5	0.024	0.001
Na_2CO_3 (m)	0	1.5	0.084	0.064
K_2CO_3 (m)	0	2	0.1	0.007

The solubility of ethane and propane in aqueous systems has not been widely examined, and the majority of experiments were conducted before 1980 (Hayduk 1982, 1986). Additionally, only some authors measured the gas solubility under intermediate to high pressure and temperature conditions. The compiled dataset contains 235 ethane solubility in pure water (Anthony and McKetta 1967; Chapoy et al. 2004; Claussen and Polglase 1952; Culberson et al. 1950; Mohammadi et al. 2004; Morrison and Billett 1952; Rettich et al. 1981; Wang et al. 2003; Wen and Hung 1970; Wetlaufer et al. 1964; Winkler 1901; Ben-Naim et al. 1973). Temperatures are in the range of 273.51 and 444.26 K, and pressure ranges from 1 to 685 bar. The statistical analysis of ethane solubility data is presented in Table 2 and Fig. 2.

Table 2 Range of parameter values in ethane solubility dataset

Parameter	Minimum	Maximum	Standard deviation	Mean
P (Bar)	1	685	158.09	83.07
T (K)	273.51	444.26	49.2	327.83
Solubility (mole fraction)	1.47×10^{-5}	0.0041	6.712×10^{-4}	5.211×10^{-4}

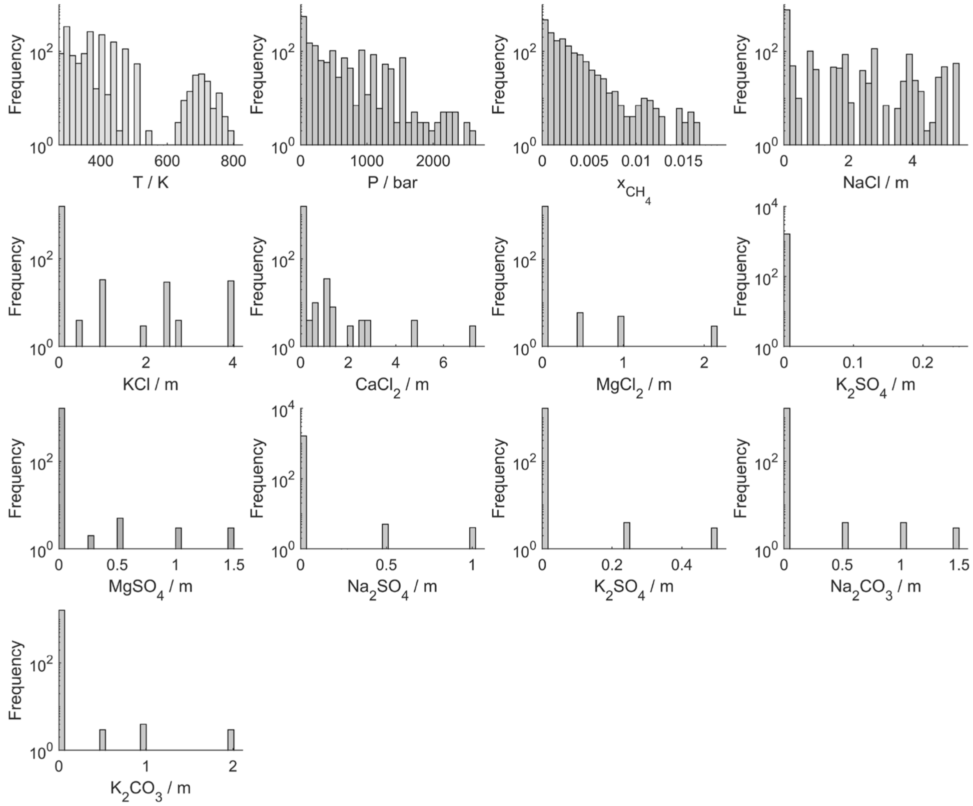
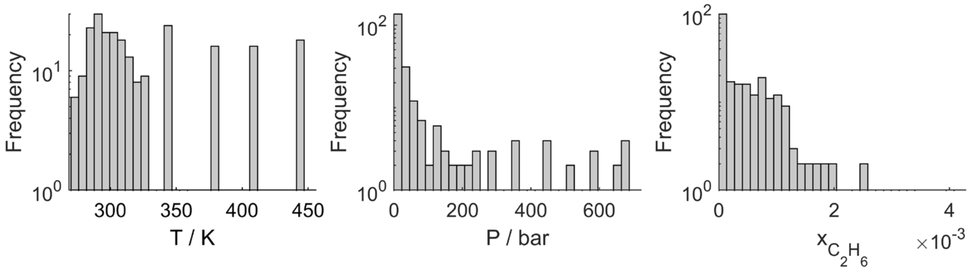


Fig. 1 Distribution of parameter values in methane solubility dataset

Fig. 2 Distribution of parameter values in ethane solubility dataset



The compiled dataset contains 259 propane solubility data in pure water (Azarnoosh and McKetta 1958; Chapoy et al. 2004; Claussen and Polglase 1952; Gaudette and Servio 2007; Kobayashi and Katz 1953; Kresheck et al. 1965; Mokraoui et al. 2007; Morrison and Billett 1952; Umano and Nakano 1958; Wehe and McKetta 1961; Wen and Hung 1970; Wetlaufer et al. 1964; Wishnia 1963). Temperatures are in the range of 273.2 and 422 K, and pressure ranges from 0.103 to 42.69 bar. The statistical analysis of propane solubility data is provided in Table 3 and Fig. 3.

3 Methodology

3.1 Machine learning

A brief description of regression tree and boosted regression tree algorithms is presented. Detailed explanations of mathematical backgrounds and computational procedures can be found in the cited literature. The models were developed using MATLAB 2021b software.

Table 3 Range of parameter values in propane solubility dataset

Parameter	Minimum	Maximum	Standard deviation	Mean
P (Bar)	0.103	42.69	9.8	9.58
T (K)	273.2	422	37.74	322.51
Solubility (mole fraction)	1.309×10^{-6}	3.66×10^{-3}	8.33×10^{-5}	1.26×10^{-4}

3.1.1 *Regression tree*

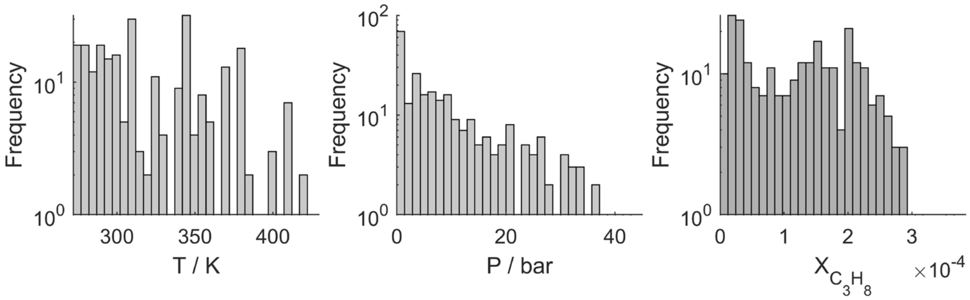
Regression tree (RT) is a supervised technique that uses one or more input (predictor) variables to predict a single output (response) variable. An RT is built through a binary recursive partitioning process, in which the data are split data into partitions or branches, and then splitting of each branch continues further (Leblanc 2006; Loh 2011). All the data in the training set are initially in the same group (root), and then are allocated into two branches (child nodes) using a splitting rule that maximizes homogeneity in the child nodes. The splitting process continues until each node reaches a user-specified minimum node size and becomes a terminal node. The splitting rules are in the internal nodes and the responses are in the leaf nodes (Cichsz 2015; Saha et al. 2015).

The advantage of tree-based models is that they are scalable to large problems and can handle smaller datasets than neural networks (Cichsz 2015; Shehab et al. 2024). RT is flexible and has the ability to adjust in time, can easily handle outliers, has an easy implementation on different types of data structures, and is computationally cheap. As with any model, regression tree has its own weaknesses; a single tree model tends to be unstable, which can negatively influence the accuracy of the response variable (Breiman 1996; Loh 2011).

3.1.2 *Boosted regression tree*

Boosted regression tree (BRT) is an ensemble of boosting and regression tree algorithms. In boosting, multiple trees are fitted to the training data, and are then sequentially combined

Fig. 3 Distribution of parameter values in propane solubility dataset



to improve the predictive performance that can be obtained from a single tree (Elith et al. 2008; Taherdangkoo et al. 2022). Boosting emphasizes poorly modeled observations, i.e. observations with high deviations from the mean, in the existing trees to produce a strong prediction and improve the model accuracy. The predictive performance of RTs has been improved by the boosting algorithm (Buhlmann and Hothorn 2007; Saha et al. 2015).

BRT uses a stepwise forward procedure, which means that the existing trees remain unchanged. A new tree is trained at each iteration using the original features and is added to the current tree sequence. Then, residuals of each observation are updated to represent the contribution of the new tree. Once the process is complete, the final predictions are determined by the weighted sum of the predictions of individual trees (Elith et al. 2008; Saha et al. 2015). To minimize the loss function, Friedman (2001) introduced gradient boosting by applying the steepest descent method to the stepwise forward estimate. Later, the gradient boosting method was modified by using a random subsampling of the training data to improve the predictive performance and reduce over-fitting potential and the computation time (Friedman 2002).

3.1.3 Bayesian optimization

The performance of a tree model is dependent on the choice of its hyper-parameters values (Bergstra et al. 2011). We employed a Bayesian optimization algorithm to tune hyper-parameters of the RT and BRT models. Bayesian optimization is suitable for optimizing computationally expensive objective functions, and tolerates stochastic noise in function evaluations. This method is characterized by two features (i) a surrogate model of an objective function, and (ii) an acquisition function computed from the surrogate model, which is used to define the next evaluation point. We employed the expected improvement acquisition function, which was used to construct a utility function from the model posterior to direct sampling to areas where improvement over the current optimum can be expected (Bergstra et al. 2011; Hutter et al. 2019).

3.1.4 Accuracy assessment

We partitioned the data into k randomly groups (or folds) of roughly equal size using k-fold cross-validation. Models were trained using k-1 groups of the dataset and validated on the remaining group. The average error over all k groups was used to obtain the total effectiveness of the ML models. Herein, a fivefold cross-validation (k = 5) was used.

There are various statistical approaches to assess the accuracy of ML models. We used the coefficient of determination (R^2), the mean squared error (MSE), and the median absolute deviation (MAD) metrics. We also used the absolute residual distribution plot to evaluate the model's accuracy and check possible residual trends. Additionally, we employed a leverage statistical approach and sketched William's plot (Narmandakh et al. 2023) to detect outliers.

3.2 Thermodynamic models

The performance of the most accurate ML model was compared with four thermodynamic models as they can effectively describe the phase behavior of the system when conditions fall within their applicability domain. We used Spivey and McCain (2004) to calculate methane solubility, Mao et al. (2005) to calculate ethane solubility, and Chapoy et al. (2004) and Pereda et al. (2009) to calculate propane solubility. Herein, we present each thermodynamic model in its original form.

(1) Spivey model

Spivey and McCain (2004) developed an empirical correlation to calculate methane (CH_4) solubility in pure water, and used a modification of Duan et al. (1992) method to account for salinity. Spivey and McCain (2004) model, simply referred to as "Spivey model", is valid for temperatures from 293.15 to 633.15 K, pressures from 9 to 2000 bar, and NaCl concentrations of up to 6 m. The solubility of methane in pure water and NaCl solutions can be calculated as follows:

$$C_{mCH_4,H_2O} = exp\left(A(T)\left[ln\left(P-P_v\right)\right]^2 + B(T)ln\left(P-P_v\right) + C(T)\right) \tag{1}$$

$$C_{mCH_4,brine} = C_{mCH_4,H_2O}exp\left[-2\lambda_{CH_4,Na}(T,P)C_{mNaCl} - \xi_{CH_4,NaCl}(T,P)C_{mNaCl}^2\right] \tag{2}$$

where $C_{mCH_4H_2O}$ and $C_{mCH_4,brine}$ [mol kg^{-1}] are the solubility of methane in pure water and brine solutions, respectively. T [K] is temperature, P [MPa] is pressure, and P_v [MPa] is vapor pressure of pure water. A(T), B(T), and C(T) are temperature-dependent functions. C_{mNaCl} [mol kg^{-1}] is the NaCl concentration, and $\lambda_{CH_4,Na}(T,P)$ and $\xi_{CH_4,NaCl}(T,P)$ are coefficients.

(2) Mao model

Mao et al. (2005) developed a thermodynamic model based on an equation of state and the theory of Pitzer (Pitzer 1973) to calculate ethane (C_2H_6) in pure water and aqueous NaCl solutions. Mao model is valid within the range of temperature 273 to 444 K, and pressure of 0 to 1000 bar. The solubility of methane in pure water can be calculated using the following equation:

$$ln\frac{y_{C_2H_6}P}{m_{C_2H_6}} = \frac{\mu^{l(0)}_{C_2H_6}(T,P) - \mu^{v(0)}_{C_2H_6}(T)}{RT} - ln\varphi_{C_2H_6}(T,P,y) + ln\gamma_{C_2H_6}(T,P,m) \quad (3)$$

where $m_{C_2H_6}$ [mol kg^{-1}] is the solubility of C_2H_6 in pure water, y_i is the mole fraction of C_2H_6 in the gas phase, P [bar] is pressure, T [K] is temperature, R [bar cm^3 mol^{-1} K^{-1}] equal to 83.14467 is the universal gas constant, $\mu^{l(0)}_{C_2H_6}$ is the chemical potential of C_2H_6 in the liquid phase, and $\varphi_{C_2H_6}$ is the fugacity coefficient.

(3) Chapoy model

Chapoy et al. (2004) developed a thermodynamic model based on uniformity of the fugacity of each component in all phases to calculate propane (C_3H_8) solubility in the liquid phase. The Valderrama modification of the Patel–Teja equation of state (VPT-EoS) with non-density dependent (NDD) mixing rules was used to calculate fugacities in the fluid phases. They acquired experimental propane solubility data to adjust the binary interaction parameters between propane and water. Henry's constants for propane in water can be calculated as:

$$ln\left(H_{iw}\right) = 552.64799 + 0.078453T - \frac{21334.4}{T} - 85.89736lnT \quad (4)$$

where H_{iw} [KPa] is Henry's constant and T [K] is temperature.

(4) Pereda model

Pereda et al. (2009) used a group contribution plus association equation of state (GCA-EoS) to describe the phase behavior of water + hydrocarbon (C_2 to n-C_6, cy–C_6, i–C_4 and i–C_8) system. They acquired experimental solubility data on the solubility of n-hexane, cyclo-hexane and iso-octane in pure water to adjust the parameters of GCA-EoS. The following equation was presented for the hard sphere diameter of water to take into account the temperature dependency of the hydrocarbon solubility in water and the vapor pressure of water.

$$d_W = d_{CW}\{0.554\left[exp\left[\frac{-2T_{CW}}{3T}\right]\right]^2 - 0.543exp\left[\frac{-2T_{CW}}{3T}\right] + 1.097\} \quad (5)$$

where d_W [cm mol^{-1}] is the hard sphere diameter of water, d_{CW} [cm mol^{-1}] is the hard sphere diameter of water at the critical temperature, T [K] is temperature, and T_{CW} [K] is the critical temperature of water.

4 Results

4.1 Model performance evaluation

We employed a Bayesian optimization algorithm with an expected improvement acquisition function to optimize the hyper-parameters of the RT and BRT algorithms. The iteration number for running each algorithm was set to 300. Table 4 summarizes the optimum hyper-parameter values of the models obtained after the optimization process.

The input parameters of the RT-BO and BRT-BO models are pressure [bar], temperature [K], and concentration [m] of NaCl, KCl, $CaCl_2$, $MgCl_2$, K_2SO_4, $MgSO_4$, Na_2SO_4, K_2SO_4, Na_2CO_3, K_2CO_3, and the corresponding gas solubility (mole fraction) is the output parameter. In the case of gas solubility in pure water, the salt concentrations are set to zero. The regression plots (Fig. 4) of predicted gas solubility values from the RT-BO and BRT-BO versus experimental ones, show accumulation of data points close to the 45-degree reference line. The deviations from the reference line are more evident in RT-BO model, indicating its lower accuracy.

The statistical indices (R^2, MSE, and MAD) indicate the superior performance of the BRT-BO model (Table 5). In this case, the MSE equals 9.97×10^{-8} and the MAD is 1.72×10^{-4}. The RT-BO model also has a relatively high predictive capability; MSE and MAD equal 2.15×10^{-7} and 2.33×10^{-4}, respectively.

The predictive capabilities of the models are further illustrated in Fig. 5. The experimental and predicted solubility values of the BRT-BO model mostly cover each other, indicating its good performance. The results show that the model is slightly less accurate in determining high gas solubility values, especially where the solubility in the aqueous phase is higher than 0.01 mol fraction because most of the data points have lower solubility values (see Sect. 2). The maximum and mean values for methane solubility are 0.0185, and 0.0024. These values for ethane are 0.0041 and 0.00052, respectively, and for propane are 0.00366 and 0.00012. The

Table 4 Optimal hyper-parameter values of the ML models and their range of variations

Model	Hyper-parameter	Range of variation	Optimum value
RT-BO	Minimum leaf size	1–1064	1
BRT-BO	Minimum leaf size	1–1064	22
	Number of learners	10–500	492
	Learning rate	0.001–1	0.19136
	Number of predictors to sample	1–15	15

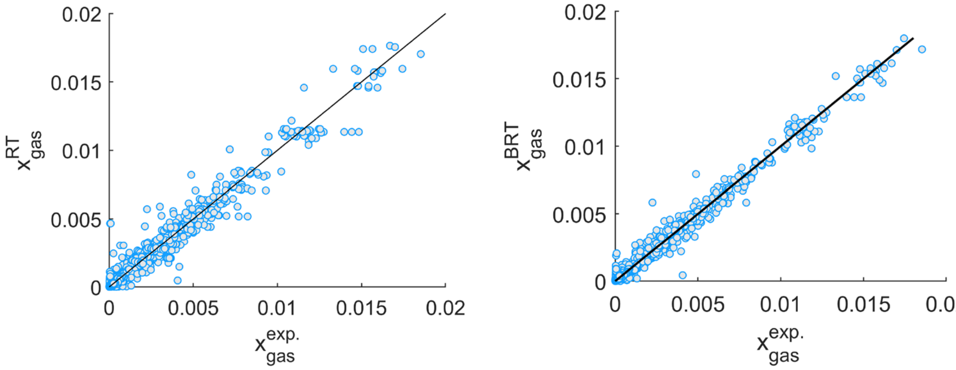


Fig. 4 Regression plots of the RT-BO and BRT-BO models, showing predicted versus experimental solubility values of methane, ethane, and propane gases. x_{gas} is the gas solubility in mole fraction

Table 5 Summary of statistical indices for the ML model's performance

Model	R^2	MSE	MAD
BRT-BO	0.99	9.97×10^{-8}	1.72×10^{-4}
RT-BO	0.97	2.15×10^{-7}	2.33×10^{-4}

statistical analysis shows that the BRT-BO model's deviations from the experimental data are minor, except for some data points.

We conducted more analysis to evaluate the performance of the BRT-BO model since it is more accurate than the RT-BO. The empirical cumulative distribution function (eCDF) of the BRT-BO (Fig. 6) indicates the high accuracy of the model because the curve is close to the Y-axis; the residuals are mostly distributed near zero. Furthermore, 70 % of the precipitated gas solubility values have an absolute error of lower than 1.5×10^{-4}, and 90 % of the predicted values have an error of lower than 3.9×10^{-4}.

We calculated the partial dependence between the predictor variables and gas solubility in aqueous phase using the BRT-BO model. Figure 7 displays the two-variable partial dependence of gas solubility on joint values of pressure and temperature. The solubility of light hydrocarbons has a strong partial dependence on pressure; gas solubility increases with increasing of pressure. The strong partial dependence of the gas solubility on the temperature is evident. These outcomes confirm that the developed model is reliable following the gas solubility behavior observed during the laboratory testing.

We sketched the Williams plot (Taherdangkoo et al. 2021b) on the basis of the standardized residuals and Hat values (diagonal elements of the Hat matrix) for the

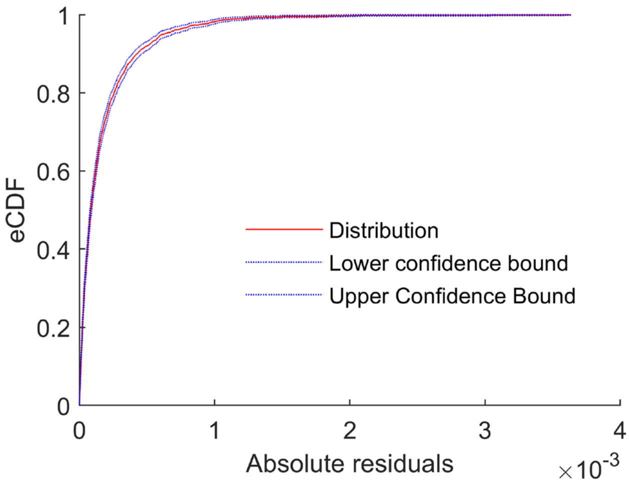


Fig. 6 Cumulative frequency of absolute residuals obtained from the BRT-BO model

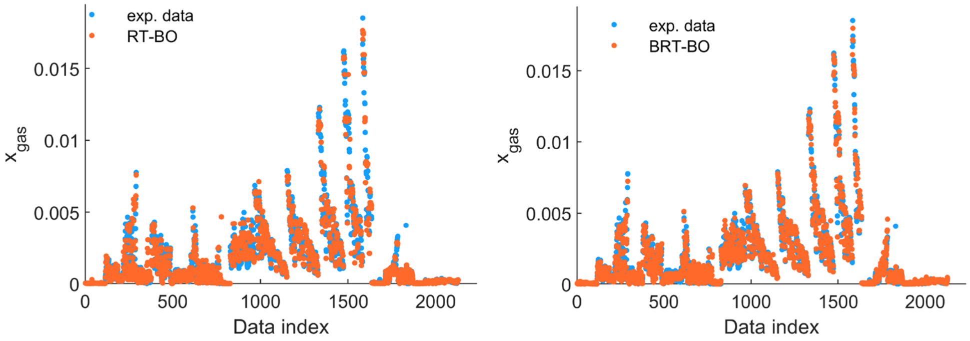


Fig. 5 Comparing the RT-BO and BRT-BO calculated light hydrocarbon solubility values with experimental values versus corresponding data index

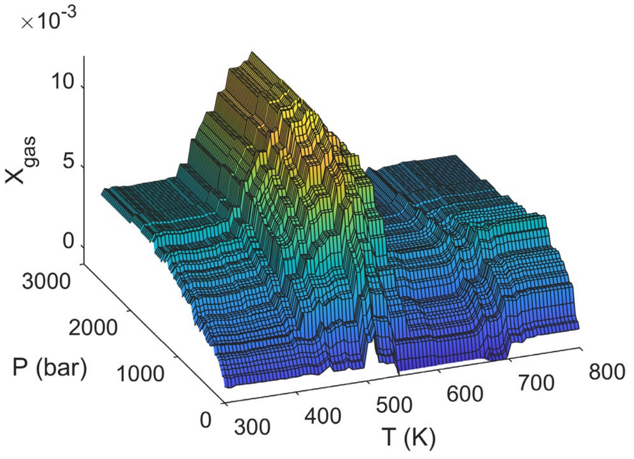


Fig. 7 Partial dependence of the gas solubility in aqueous phase obtained from the BRT-BO on pressure and temperature

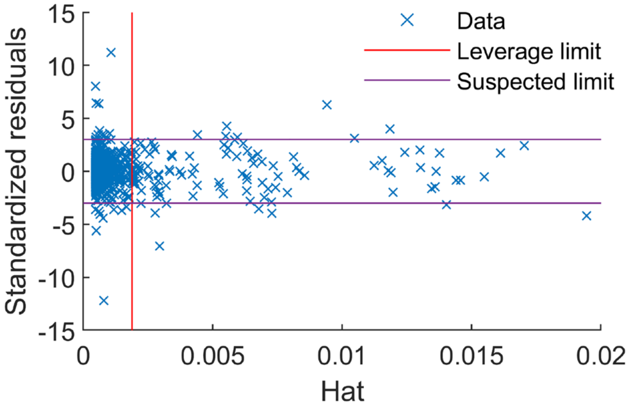


Fig. 8 Williams plot of gas solubility dataset for the BRT-BO model

BRT-BO model to detect suspected experimental data and high leverage points, which are outliers falling outside of the applicability domain of the model. The analysis (Fig. 8) shows that bulk of the gas solubility data falls in the valid domain, 0 ≤ Hat ≤ 0.0019 and -3 ≤ standardized residuals ≤ 3, indicating the reliability of the compiled dataset and the statistical validity of the model.

The quantitative analysis shows that the standardized residuals of 36 data points (1.69% of the compiled data) are outside the range of − 3 to 3, which are considered as suspected data. Additionally, 112 data points (5.26%) have high leverage (Hat value ≥ 0.0019). However, the BRT-BO model's performance analysis (Fig. 8) shows that the predicted gas solubility residuals are in an acceptable range.

4.2 Comparison analysis

In Fig. 9, the predictive ability of the BRT-BO model to calculate methane solubility in pure water was compared with experimental data and Spivey model in the temperature range between 298.2 and 344.3 K and pressure between 22.8 and 680 bar. The BRT-BO and Spivey models capture the solubility trend observed in the experimental data; increase of the methane solubility in liquid phase with increasing of pressure. The BRT-BO model can accurately determine methane solubility values at low pressure and temperature conditions. Furthermore, the comparison analysis with O'Sullivan (1970), and Culberson et al. (1951) shows the efficiency of the model in determining the solubility values at high pressures. The modeling deviations, i.e. the difference between experimental and predicted values, are minor showing the potential of the BRT-BO for future applications.

Methane solubility in NaCl solutions was compared at temperatures ranging from 324 to 378 K, and NaCl concentration between 0.88 and 2.5 m. The gas solubility in liquid phase decreases with the increase in salinity, which was effectively modeled (Fig. 10). The BRT-BO is highly accurate in predicting the experimental data at wide ranges of pressure (41.8–1339 bar) demonstrated by the comparison analysis. The overall performance of the BRT-BO model is slightly better than Spivey. The comparison analysis shows that the methane solubility predictions are accurate in the CH_4–H_2O–NaCl and CH_4–H_2O systems. The analysis shows that the BRT-BO can be employed for modeling of two-phase flow and transport of methane in shallow and deep subsurface, e.g. freshwater and saline water aquifers, with an accuracy needed for hydrogeological applications.

The BRT-BO model was compared with Mao's model to calculate ethane solubility in pure water. The predictive ability of both models is satisfactory, showing only minor deviations from the experimental data in some conditions (Fig. 11). For instance, Mao model is slightly more accurate than the BRT-BO to determine experimental ethane solubility values taken from Mohammadi et al. (2004) at temperature of 298.3 K, while the BRT-BO model performs better at 313 K. The BRT-BO model provides better predictions on Wang et al. (2003) dataset. The BRT-BO model demonstrates a good covering of experimental data points and can be applied for ethane solubility prediction in aqueous phase at pressures ranging from 4.39 to 573.6 bar.

The BRT-BO and Chapoy model's calculations regarding the propane solubility in pure water are close to experimental values taken from Chapoy et al. (2004) (Fig. 12). Similar to previous cases, the BRT-BO predictions follow the gas solubility behavior observed in experimental data. Chapoy model is the most accurate model in predicting propane solubility followed by the BRT-BO and Pereda models. The BRT-BO is slightly more accurate than the Pereda model under the conditions studied.

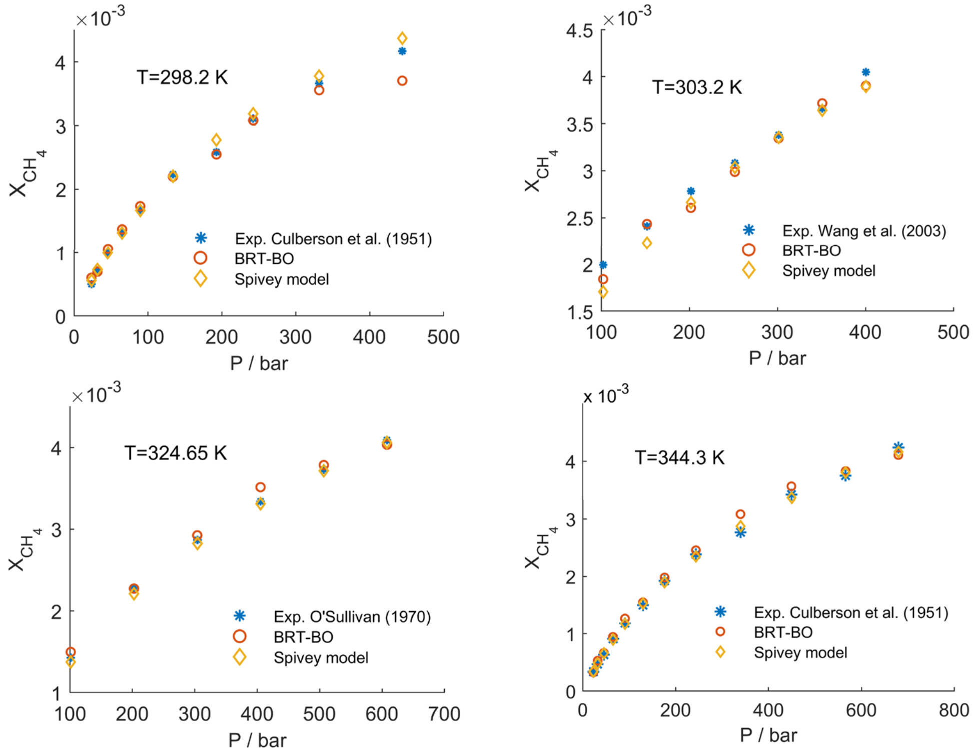


Fig. 9 Comparison of experimentally determined solubility of methane in H_2O at 298.2 K (Culberson et al. 1951), 303.2 K (Wang et al. 2003), 324.65 K (O'Sullivan 1970), and 344.3 K (Culberson et al. 1951) with the calculated values from the BRT-BO and Spivey models

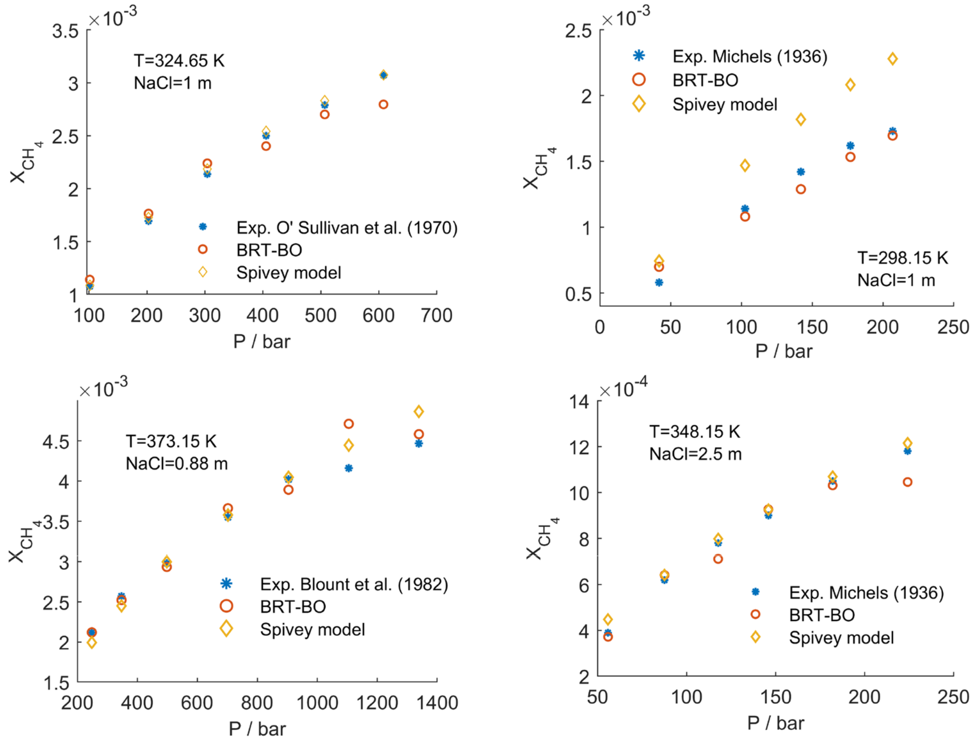


Fig. 10 Comparison of experimentally determined solubility of methane in the system H_2O-NaCl at 324.65 K (O'Sullivan 1970), 298.15 K (Michels et al. 1936), 373 K (Blount et al. 1982), and 348.15 K (Michels et al. 1936) with the calculated values from the BRT-BO model and Spivey model

5 Discussion

The results showed that the BRT-BO model is able to determine solubility of methane, ethane, and propane gases in pure water and electrolyte solutions with sufficient accuracy, highlighting its potential for wide-ranging geological and hydrogeological applications. In general, the model provides accurate outcomes compared to established thermodynamic models such as those by Spivey, Mao, Chapoy, and Pereda models. The BRT-BO model serves as a viable alternative

Fig. 11 Comparison of experimentally determined solubility of ethane in H_2O at 298.3 K (Mohammadi et al. 2004), 303.2 K (Wang et al. 2003), 313.19 K (Mohammadi et al. 2004), and 444.26 K (Culberson et al. 1951) with the calculated values from the BRT-BO model and Mao model

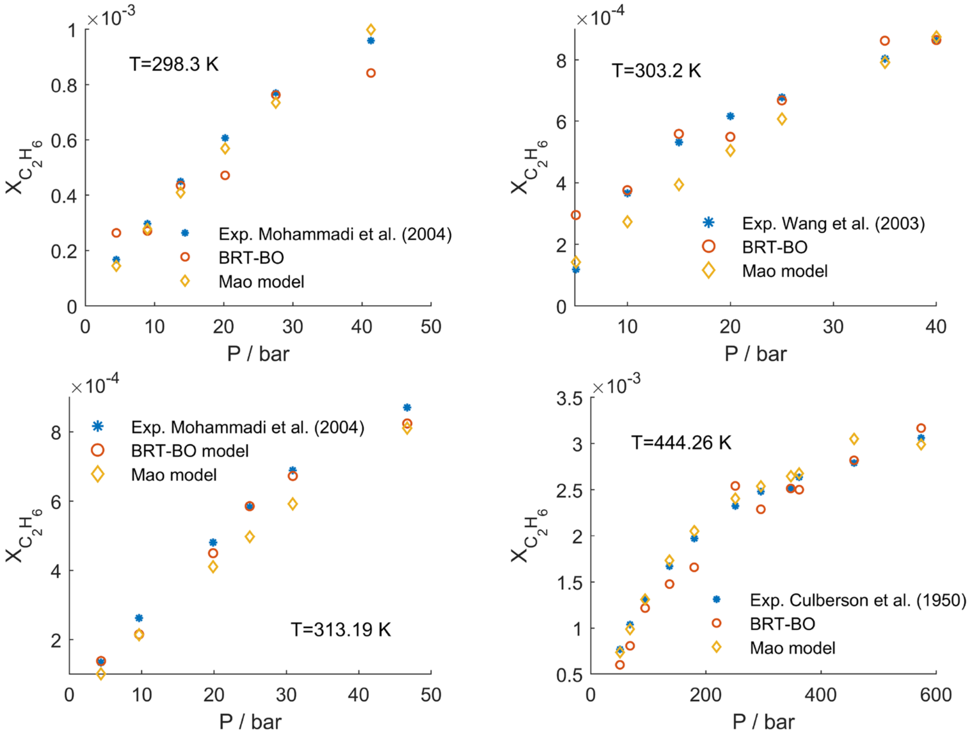
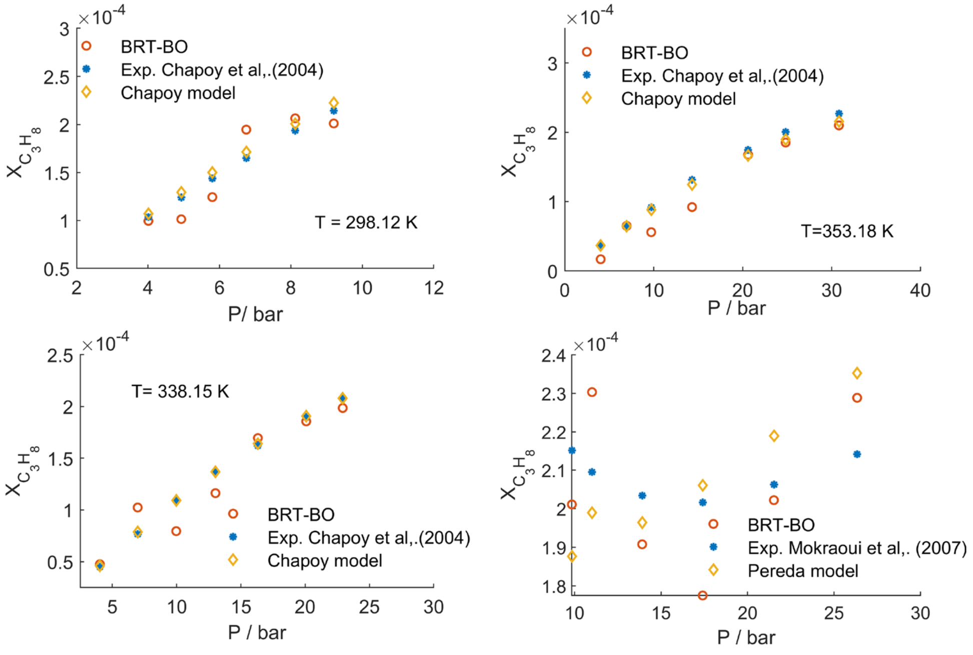


Fig. 12 Comparison of experimentally determined solubility of propane in H_2O at 298.12, 338.15, and 353.18 K (Chapoy et al. 2004), and various temperatures (Mokraoui et al. 2007) with the calculated values from the BRT-BO, Chapoy, and Pereda models



for predicting light hydrocarbon solubility in aquatic systems under diverse conditions.

One of the main advantages of the BRT-BO is its applicability to calculate the solubility of C_1–C_3 gases in a wide range of conditions. Although thermodynamic approaches are highly efficient, they are usually complex and have a limited application domain. For example, the Mao model, which requires many empirical parameters, is limited to the solubility calculations of ethane in aqueous systems. Numerical modeling of the transport of light hydrocarbons between deep gas reservoirs and shallow groundwater aquifers is complex as it involves multi-phase, multi-component flow through different media such as fault zones, fracture networks, and low permeability layers. The BRT-BO model can serve as an alternative to thermodynamic models needed to calculate the phase behavior of various gas components during the transport. This would reduce the complexity of numerical models making groundwater contamination models more efficient. Therefore, the BRT-BO model can be further implemented in numerical modeling frameworks to address issues in the field of science and engineering.

Future studies might explore incorporating the impact of ionic strength and specific ion effects on gas solubility directly within the machine learning algorithm, aiming to further refine its predictive accuracy. Additionally, the application of alternative optimization algorithms could offer improvements in the performance of machine learning models. While the dataset used for model development was adequately extensive, expanding this dataset could further improve the model's accuracy and its ability to generalize across a broader range of conditions.

6 Conclusions

We employed regression tree (RT) and boosted regression tree (BRT) algorithms optimized with a Bayesian optimization algorithm to build a model able to calculate solubility of methane, ethane, and propane in aquatic systems over a wide range of pressure, temperature, and salt concentrations. The RT-BO and BRT-BO are able to determine the solubility of light hydrocarbons in aquatic systems. The BRT-BO is more accurate, evidenced by the MSE value close to zero ($\mathrm{MSE} = 9.97 \times 10^{-8}$) and an R^2 value of 0.99. The predictions of the BRT-BO are in good agreement with the experimental hydrocarbon solubility dataset, which contains 2129 experimental data of methane, ethane, and propane solubility in pure water and various electrolyte solutions. The comparison analysis of the BRT-BO model's predictions with four well-established thermodynamic models confirms the high prediction capability of the ML model. The application of the leverage approach showed that the majority of data points (5.26% outliers) fall in the valid domain, verifying the statistical validity of the model. We conclude that the BRT-BO model is a well-suited and robust tool which can be regarded as an alternative to more classical approaches for light hydrocarbon solubility calculations needed for various scientific and engineering applications such as numerical modeling of stray gas migration in the subsurface environment, development of effective environmental risk management strategies, optimization of gas extraction and processing operations, and development of strategies aimed at mitigating atmospheric emissions of methane and other light hydrocarbons.

Author contributions GK: Modelling, Visualization, Writing—Original Draft. RT: Conceptualization, Methodology, Writing—Original Draft. CC: Formal analysis, Validation, Writing—Review & Editing NS: Data Curation, Data Analysis, Writing—Review & Editing FDA: Writing—Review & Editing TM: Writing—Review & Editing CB: Project administration, Writing—Review & Editing.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The data used for this study is available from the corresponding author upon reasonable request.

Conflict of interest N/A.

Consent to participate Informed consent was obtained from all individual participants involved in the study.

Ethical approval All procedures performed in this study were in accordance with the ethical standards.

References

Amirijafari B, Campbell JM (1972) Solubility of gaseous hydrocarbon mixtures in water. Soc Pet Eng J. 12(01):21–27. https://doi.org/10.2118/3106-PA

Anthony RG, McKetta JJ (1967) Phase equilibrium in the ethylene-ethane-water system. J Chem Eng Data. 12(1):21–28.

Azarnoosh A, McKetta JJ (1958) The solubility of propane in water. Petrol Refiner. 37(11):275–278.

Ben-Naim A, Wilf J, Yaacobi M (1973) Hydrophobic interaction in light and heavy water. J Phys Chem. 77(1):95–102. https://doi.org/10.1021/j100620a021

Bergstra J, Bardenet R, Bengio Y, Kégl B, (2011) Algorithms for hyper-parameter optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. NIPS'11. Granada, Spain. Curran Associates Inc., (pp. 2546–2554). ISBN: 9781618395993.

Blanco LC, Smith NO (1978) The high pressure solubility of methane in aqueous calcium chloride and aqueous tetraethylammonium bromide. Partial molar properties of dissolved methane and nitrogen in relation to water structure. J Phys Chem. 82(2):186–191. https://doi.org/10.1021/j100491a012

Blount CW, Price LC, (1982) Solubility of methane in water under natural conditions: a laboratory study, p. 161.

Breiman L (1996) Bagging predictors. Mach Learn. 24(2):123–140. https://doi.org/10.1007/BF00058655

Buhlmann P, Hothorn T (2007) Boosting algorithms: regularization, prediction and model fitting. Stat Sci. 22:477–505.

Bunsen RW (1855) Solubility data series (1987). Pergamon Press, Oxford, p 7.

Byrne PA, Stoessell RK (1982) Methane solubilities in multisalt solutions. Geochim Cosmochim Acta. 46(11):2395–2397. https://doi.org/10.1016/00167037(82)90210-1

Cahill AG et al (2017) Mobility and persistence of methane in groundwater in a controlled-release field experiment. Nat Geosci. 10(4):289–294. https://doi.org/10.1038/ngeo2919

Cahill AG, Parker BL, Bernhard M, Ulrich Mayer K, Cherry JA (2018) High resolution spatial and temporal evolution of dissolved gases in groundwater during a controlled natural gas release experiment. Sci Total Environ. 622–623:1178–1192. https://doi.org/10.1016/j.scitotenv.2017.12.049

Chapoy A, Mohammadi AH, Richon D, Tohidi B (2004) Gas solubility measurement and modeling for methane–water and methane–ethane–n-butane–water systems at low temperature conditions. Fluid Phase Equilib. 220(1):113–121. https://doi.org/10.1016/j.fluid.2004.02.010

Cichsz P (2015) Regression trees. Data mining algorithms. Wiley, pp 261–294. https://doi.org/10.1002/9781118950951.ch9

Claussen WF, Polglase MF (1952) Solubilities and structures in aqueous aliphatic hydrocarbon solutions. J Am Chem Soc. 7(19):4817–4819. https://doi.org/10.1021/ja01139a026

Cosgrove BA, Walkley J (1981) Solubilities of gases in H_2O and $2H_2O$. J Chromatogr A 216:161–167.

Cramer SD (1984) Solubility of methane in brines from 0 to 300 degree C. Ind Eng Chem Process Des Dev. 23(3):533–538. https://doi.org/10.1021/i200026a021

Crovetto R, Fern´andez-PriniJapas RML (1982) Solubilities of inert gases and methane in H_2O and in D_2O in the temperature range of 300 to 600 K. J Chem Phys. 76(2):1077–1086. https://doi.org/10.1063/1.443074

Culberson OL, McKetta Jr JJ (1950) Phase equilibria in hydrocarbon-water systems II-the solubility of ethane in water at pressures to 10,000 psi. J Pet Technol 2(11):319–322.

Culberson OL, McKetta JJ Jr (1951) Phase equilibria in hydrocarbon-water systems III-the solubility of methane in water at pressures to 10000 PSIA. J Pet Technol. 3(08):223–226. https://doi.org/10.2118/951223-G

Dhima A, de Hemptinne JC, Moracchini G (1998) Solubility of light hydrocarbons and their mixtures in pure water under high pressure. Fluid Phase Equilib. 145(1):129–150. https://doi.org/10.1016/S0378-3812(97)00211-2

Duan Z, Mao S (2006) A thermodynamic model for calculating methane solubility, density and gas phase composition of methane-bearing aqueous fluids from 273 to 523 K and from 1 to 2000 bar. Geochim Cosmochim Acta. 70(13):3369–3386. https://doi.org/10.1016/j.gca.2006.03.018

Duan Z, Møller N, Greenberg J, Weare JH (1992) Prediction of methane solubility in natural waters to high ionic strength from 0 to 250°C and from 0 to 1600 bar. Geochim Cosmochim Acta. 56(4):1451–1460. https://doi.org/10.1016/0016-7037(92)90215-5

Duffy JR, Smith NO, Nagy B (1961) Solubility of natural gases in aqueous salt solutions—I: liquidus surfaces in the system CH_4-H_2O-$NaCl_2$-$CaCl_2$ at room temperatures and at pressures below 1000 psia. Geochim Cosmochim Acta. 24(1):23–31. https://doi.org/10.1016/0016-7037(61)90004-7

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol. 77(4):802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Eucken A, Hertzberg G (1950) Aussalzeffekt und Ionenhydratation. Z Phys Chem. 195(1):1–23. https://doi.org/10.1515/zpch-1950-19502

Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal. 38(4):367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Statist. 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

Gaudette J, Servio P (2007) Measurement of dissolved propane in water in the presence of gas hydrate. J Chem Eng Data. 52(4):1449–1451. https://doi.org/10.1021/je7001286

Grunwald N, Maßmann J, Kolditz O, Nagel T (2020) Non-iterative phase-equilibrium model of the H_2O-CO_2-NaCl-system for large-scale numerical simulations. Math Comput Simul. 178:46–61. https://doi.org/10.1016/j.matcom.2020.05.024

Hayduk W (1982) IUPAC solubility data series. Ethane, vol 9. Pergamon Press, Oxford.

Hayduk W (1986) IUPAC solubility data series. Propane, Butane and 2-Methylpropane, vol 24. Pergamon Press, Oxford.

Hutter F, Kotthoff L, Vanschoren J (2019) Automated machine learning. The Springer series on challenges in machine learning. Springer, Cham, p 219. https://doi.org/10.1007/978-3-030-05318-5

Jackson RB, Vengosh A, Darrah TH, Warner NR, Down A, Poreda RJ, Osborn SG, Zhao K, Karr JD (2013) Increased stray gas abundance in a subset of drinking water wells near Marcellus shale gas extraction. Proc Natl Acad Sci. 110(28):11250–11255. https://doi.org/10.1073/pnas.122163511

Kiepe J, Horstmann S, Fischer K, Gmehling J (2003) Experimental determination and prediction of gas solubility data for methane + water solutions containing different monovalent electrolytes. Ind Eng Chem Res. 42(21):5392–5398. https://doi.org/10.1021/ie030386x

King GE (2012) Hydraulic fracturing 101: what every representative, environmentalist, regulator, reporter, investor, university researcher, neighbor and engineer should know about estimating frac risk and improving frac performance in unconventional gas and oil wells. In: SPE hydraulic fracturing technology conference. https://doi.org/10.2118/152596-MS

Kissinger A, Helmig R, Ebigbo A, Class H, Lange T, Sauter M, Heitfeld M, Klünker J, Jahnke W (2013) Hydraulic fracturing in unconventional gas reservoirs: Risks in the geological system, part 2. Environ Earth Sci. 70(8):3855–3873. https://doi.org/10.1007/s12665-013-2578-6

Kobayashi R, Katz D (1953) Vapor-liquid equilibria for binary hydrocarbon-water systems. Ind Eng Chem. 45(2):440–446. https://doi.org/10.1021/ie50518a051

Krader T, Franck EU (1987) The ternary systems H_2O–CH_4–NaCl and H_2O–CH_4–$CaCl_2$ to 800 K and 250 bar. Ber Bunsengers Phys Chem. 91:627–634. https://doi.org/10.1002/bbpc.19870910610

Kresheck GC, Schneider H, Scheraga HA (1965) The effect of D2O on the thermal stability of proteins. Thermodynamic parameters for the transfer of model compounds from H_2O to D2O1,2. J Phys Chem. 69(9):3132–3144. https://doi.org/10.1021/j100893a054

Lannung A, Gjaldbæk JC, Rundqvist S, Varde E, Westin G (1960) The solubility of methane in hydrocarbons, alcohols, water, and other solvents. Acta Chem Scand. 14:1124–1128. https://doi.org/10.3891/acta.chem.scand.14-1124

Leblanc M (2006) Regression trees. Encyclopedia of environmetrics. American Cancer Society. https://doi.org/10.1002/9780470057339.var027

Lekvam K, Raj Bishnoi P (1997) Dissolution of methane in water at low temperatures and intermediate pressures. Fluid Phase Equilib. 131(1):297–309. https://doi.org/10.1016/S0378-3812(96)03229-3

Loh WY (2011) Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov. 1(1):14–23. https://doi.org/10.1111/insr.12016

Mao S, Zhang Z, Hu J, Su R, Duan Z (2005) An accurate model for calculating C2H6 solubility in pure water and aqueous NaCl solutions. Fluid Ph Equilib 238(1):77–86.

Michels A, Gerver J, Bijl A (1936) The influence of pressure on the solubility of gases. Physica. 3(8):797–808. https://doi.org/10.1016/s0031-8914(36)80353-x

Mishnina TA, Avdeeva OI, Bozhovskaya TK (1962) Methane: Solubility data series. Inf. Sb. Vses Nauchnissled Geol Inst. 56:137–145.

Mohammadi AH, Chapoy A, Tohidi B, Richon D (2004) Measurements and thermodynamic modeling of vapor−liquid equilibria in ethane−water systems from 274.26 to 343.08 K. Ind Eng Chem Res. 43(17):5418–5424. https://doi.org/10.1021/ie049747e

Mohammadi MR, Hadavimoghaddam F, Atashrouz S, Abedi A, Hemmati-Sarapardeh A, Mohaddespour A (2022) Modeling the solubility of light hydrocarbon gases and their mixture in brine with machine learning and equations of state. Sci Rep. 12(1):1–25. https://doi.org/10.1038/s41598-022-18983-2

Mokraoui S, Coquelet C, Valtz A, Hegel PE, Richon D (2007) New solubility data of hydrocarbons in water and modeling concerning vapor–liquid–liquid binary systems. Ind Eng Chem Res. 46(26):9257–9262. https://doi.org/10.1021/ie070858y

Morrison TJ, Billett F (1952) The salting-out of non-electrolytes. Part II. The effect of variation in non-electrolyte. J Chem Soc. https://doi.org/10.1039/JR9520003819

Moudgil BM, Somasundaran P, Lin IJ (1974) Automated constant pressure reactor for measuring solubilities of gases in aqueous solutions. Rev Sci Instrum. 45(3):406–409. https://doi.org/10.1063/1.1686640

Muccitelli JA, Wen W-Y (1980) Solubility of methane in aqueous solutions of triethylenediamine. J Solut Chem. 9(2):141–161. https://doi.org/10.1007/BF00644485

Namiot, AY (1961) In HL Clever, & CL Young (Eds.), Methane, Solubility data serie (1987, Vol. 27–28, pp. 14).

Narmandakh D, Butscher C, Ardejani FD, Yang H, Nagel T, Taherdangkoo R (2023) The use of feed-forward and cascade-forward neural networks to determine swelling potential of clayey soils. Comput Geotech. 157:105319.

Nowamooz A, Lemieux JM, Molson J, Therrien R (2015) Numerical investigation of methane and formation fluid leakage along the casing of a decommissioned shale gas well. Water Resour Res. 51(6):4592–4622. https://doi.org/10.1002/2014WR016146

O'Sullivan TD, Smith NO (1970) Solubility and partial molar volume of nitrogen and methane in water and in aqueous sodium chloride from 50 to 125. Deg and 100 to 600 atm. J Phys Chem. 74(7):1460–1466. https://doi.org/10.1021/j100702a012

Pereda S, Awan JA, Mohammadi AH, Valtz A, Coquelet C, Brignole EA, Richon D (2009) Solubility of hydrocarbons in water: Experimental measurements and modeling using a group contribution with association equation of state (GCA-EoS). Fluid Ph. Equilib 275(1):52–59

Pitzer KS (1973) Thermodynamics of electrolytes. I. Theoretical basis and general equations. J Phys Chem. 77(2):268–277. https://doi.org/10.1021/j100621a026

Qiao C, Yu X, Song X, Zhao T, Xu X, Zhao S, Gubbins KE (2020) Enhancing gas solubility in nanopores: A combined study using classical density functional theory and machine learning. Langmuir. 36(29):8527–8536.

Rettich TR, Handa YP, Battino R, Wilhelm E (1981) Solubility of gases in liquids 13 High-precision determination of Henry's constants for methane and ethane in liquid water at 275 to 328 K. J Phys Chem. 85(22):3230–3237. https://doi.org/10.1021/j150622a006

Rice AK, Lackey G, Proctor J, Singha K (2018) Groundwater-quality hazards of methane leakage from hydrocarbon wells: A review of observational and numerical studies and four testable hypotheses. Wiley Interdiscip Rev Water. 5(4):e1283. https://doi.org/10.1002/wat2.1283

Saha D, Alluri P, Gan A (2015) Prioritizing highway safety Manual's crash prediction variables using boosted regression trees. Accid Anal Prev. 79:133–144. https://doi.org/10.1016/j.aap.2015.03.011

Shehab M, Taherdangkoo R, Butscher C (2024) Towards reliable barrier systems: a constrained XGBoost model coupled with gray wolf optimization for maximum swelling pressure of bentonite. Comput Geotech. 168:106132. https://doi.org/10.1016/j.compgeo.2024.106132

Siegel DI, Azzolina NA, Smith BJ, Perry AE, Bothun RL (2015) Methane concentrations in water wells unrelated to proximity to existing oil and gas wells in northeastern Pennsylvania. Environ Sci Technol. 49(7):4106–4112. https://doi.org/10.1021/es505775c

Spivey JP, McCain WD Jr, North R (2004) Estimating density, formation volume factor, compressibility, methane solubility, and viscosity for oilfield brines at temperatures from 0 to 275 °C, pressures to 200 MPa, and salinities to 57 Mole/kg. J Can Petrol Technol. 43(07):10. https://doi.org/10.2118/04-07-05

Stoessell RK, Byrne PA (1982) Salting-out of methane in single-salt solutions at 25 °C and below 800 psia. Geochim Cosmochim Acta. 46(8):1327–1332. https://doi.org/10.1016/0016-7037(82)90268-X

Taherdangkoo R, Tatomir A, Anighoro T, Sauter M (2019) Modeling fate and transport of hydraulic fracturing fluid in the presence of abandoned wells. J Contam Hydrol. 221:58–68.

Taherdangkoo R, Tatomir A, Sauter M (2020a) Modeling of methane migration from gas wellbores into shallow groundwater at basin scale. Environ Earth Sci. 79(18):432. https://doi.org/10.1007/s12665-020-09170-5

Taherdangkoo R, Tatomir A, Taherdangkoo M, Qiu P, Sauter M (2020b) Nonlinear autoregressive neural networks to predict hydraulic fracturing fluid leakage into shallow groundwater. Water. 12:3. https://doi.org/10.3390/w12030841

Taherdangkoo R, Liu Q, Xing Y, Yang H, Cao V, Sauter M, Butscher C (2021a) Predicting methane solubility in water and seawater by machine learning algorithms: Application to methane transport modeling. J Contam Hydrol. 242:103844.

Taherdangkoo R, Yang H, Akbariforouz M, Sun Y, Liu Q, Butscher C (2021b) Gaussian process regression to determine water content of methane: Application to methane transport modeling. J Contam Hydrol. 243:103910.

Taherdangkoo R, Nagel T, Tang AM, Pereira JM, Butscher C (2022) Coupled hydro-mechanical modeling of swelling processes in clay-sulfate rocks. Rock Mech Rock Eng. 55(12):7489–7501.

Taherdangkoo R, Tyurin V, Shehab M, Ardejani FD, Tang AM, Narmandakh D, Butscher C (2023) An efficient neural network model to determine maximum swelling pressure of clayey soils. Comput Geotech. 162:105693.

Tatomir A, McDermott C, Bensabat J, Class H, Edlmann K, Taherdangkoo R, Sauter M (2018) Conceptual model development using a generic features, events, and processes (FEP) database for assessing the potential impact of hydraulic fracturing on groundwater aquifers. Adv Geo Sci. 45:185–192. https://doi.org/10.5194/adgeo-45-185-2018

Umano S, Nakano Y (1958) Solubilities of propane and n-butane in water and common-salt solution. Kogyo Kagaku Zasshi. 61(5):536–542.

Wang L-K, Chen G-J, Han G-H, Guo X-Q, Guo T-M (2003) Experimental study on the solubility of natural gas components in water with or without hydrate inhibitor. Fluid Phase Equilib. 207(1):143–154. https://doi.org/10.1016/S0378-3812(03)00009-8

Wehe AH, McKetta JJ (1961) n-Butane-1-butene-water system in the 3-phase region. J Chem Eng Data. 6(2):167–172. https://doi.org/10.1021/je60010a002

Wen W-Y, Hung JH (1970) Thermodynamics of hydrocarbon gases in aqueous tetraalkylammonium salt solutions. J Phys Chem. 74(1):170–180. https://doi.org/10.1021/j100696a032

Wetlaufer DB, Malik SK, Stoller L, Coffin RL (1964) Nonpolar group participation in the denaturation of proteins by urea and Guanidinium salts. Model compound studies. J Am Chem Soc. 86(3):508–514. https://doi.org/10.1021/ja01057a045

Winkler LW (1901) Solubility of gas in water. Ber Dtsch Chem Ges. 34:1408–1422.

Wishnia A (1963) The hydrophobic contribution to micelle formation: the solubility of ethane, propane, butane, and pentane in sodium dodecyl sulfate solution1. J Phys Chem. 67(10):2079–2082. https://doi.org/10.1021/j100804a027

Yamamoto S, Alcauskas JB, Crozier TE (1976) Solubility of methane in distilled water and seawater. J Chem Eng Data. 21(1):78–80. https://doi.org/10.1021/je60068a029

Yano T, Suetaka T, Umehara T, Horiuchi A (1974) Methane: Solubility Data Series. Pergamon Press, Oxford, pp 27–28

Yarym-Agaev NL, Sinyavskaya RP, Koliushko II, Levinton LY (1985) Phase-equilibria in the water methane and methanol methane binary-systems under high-pressures. J Appl Chem USSR 58(1):154–157