Petroleum Science 20 (2023) 2232–2243

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A machine learning-based study of multifactor susceptibility and risk control of induced seismicity in unconventional reservoirs

Gang Hui [a, b], Zhang-Xin Chen [a, b, c, *], Hai Wang [b], Zhao-Jie Song [a], Shu-Hua Wang [d], Hong-Liang Zhang [e], Dong-Mei Zhang [f], Fei Gu [g]

[a] National Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum, Beijing, 102249, PR China
[b] Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, AB, T2N1N4, Canada
[c] Eastern Institute for Advanced Study, Ningbo, 315200, Zhejiang, PR China
[d] Computer Modeling Group Ltd. Calgary, AB, T2L2M1, Canada
[e] CNOOC Research Institute, Beijing, 100028, PR China
[f] China University of Geoscience (Wuhan), Wuhan, 430074, Hubei, PR China
[g] Research Institute of Petroleum Exploration and Development, Beijing, 10083, PR China

ARTICLE INFO

Article history:
Received 6 June 2022
Received in revised form
3 February 2023
Accepted 4 February 2023
Available online 4 February 2023

Edited by Yan-Hua Sun

Keywords:
Induced seismicity
Hydraulic fracturing
Seismicity susceptibility
Risk control
Machine learning

ABSTRACT

A comprehensive dataset from 594 fracturing wells throughout the Duvernay Formation near Fox Creek, Alberta, is collected to quantify the influences of geological, geomechanical, and operational features on the distribution and magnitude of hydraulic fracturing-induced seismicity. An integrated machine learning-based investigation is conducted to systematically evaluate multiple factors that contribute to induced seismicity. Feature importance indicates that a distance to fault, a distance to basement, minimum principal stress, cumulative fluid injection, initial formation pressure, and the number of fracturing stages are among significant model predictors. Our seismicity prediction map matches the observed spatial seismicity, and the prediction model successfully guides the fracturing job size of a new well to reduce seismicity risks. This study can apply to mitigating potential seismicity risks in other seismicity-frequent regions.
© 2023 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The induced seismicity during the development of unconventional reservoirs has notably increased in North America, West Europe, and East Asia in the last decade (Atkinson et al., 2016; Yang et al., 2017; Schultz et al., 2020). Within the Western Canadian Sedimentary Basin (WCSB), the induced seismicity has been attributed to wastewater disposal in the Brazeau River zone (Schultz et al., 2014), hydrocarbon production in the Strachan D-3A Field (Baranova et al., 1999), enhanced oil recovery in the Rocky Mountain House region (Wetmiller, 1986), and hydraulic fracturing (HF) near Fox Creek (Schultz et al., 2017). Notably, several earthquakes with a local magnitude (M_L) > 3.0 in 2013–2019 were spatiotemporally linked to fracturing stimulation of horizontal wells in the Duvernay Formation in the WCSB (Eaton et al., 2018). Statistically, ~6% of wells in the Duvernay Formation are associated with M_L > 3 earthquakes (Ghofrani and Atkinson, 2020).

The Duvernay shale formation, a primary shale play in the Fox Creek region, is estimated to be 41 m in thickness and mainly consists of shale lithology (Fig. 1). The formation porosity and permeability are averaged to be 0.065 and 394 nanodarcies (3.94×10^{-7} mD) (Hui et al., 2021b). Multistage hydraulic fracturing techniques have been employed to develop shale resources, showing practical field applications. Statistics of fracturing wells suggest that the total pumped volume of fracturing fluids and total mass of proppant placed have a mean value of 56313 m^3 and 7213 t per well, respectively. However, accompanied by a large fracturing job size, a notable increase in induced seismicity rates has been reported simultaneously (Hui et al., 2021a, 2022) (Fig. 1c). Interestingly, although the fracturing parameters of associated horizontal wells were similar, the distribution and magnitude of induced seismicity caused by hydraulic fracturing vary widely. This

* Corresponding author. National Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum, Beijing, 102249, PR China.
E-mail address: zhachen@ucalgary.ca (Z.-X. Chen).

https://doi.org/10.1016/j.petsci.2023.02.003
1995-8226/© 2023 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

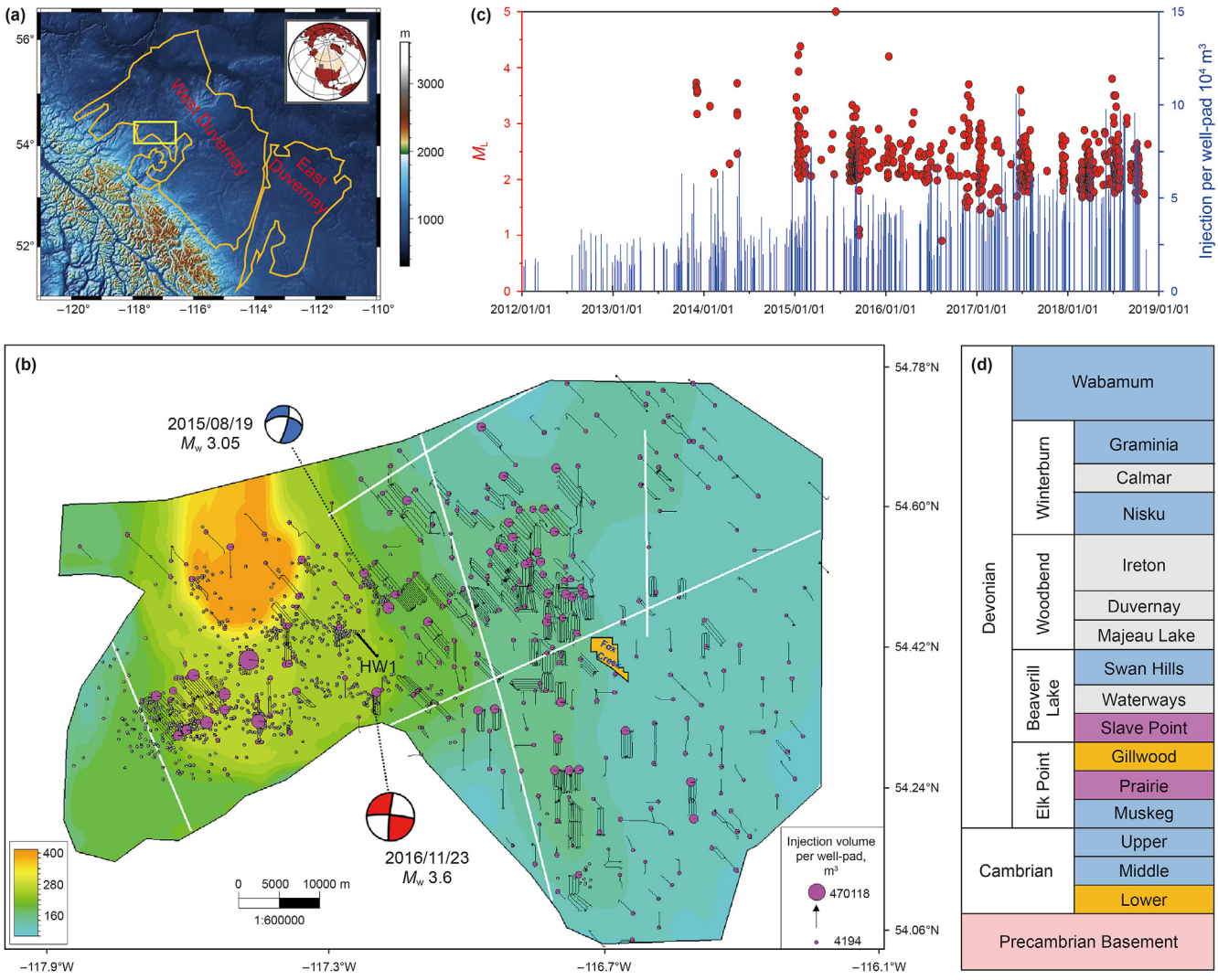

Fig. 1. (**a**) Location of the Duvernay Shale in the WCSB and the studied Fox Creek region. (**b**) Map view of recorded seismicity and fracturing horizontal wells near Fox Creek. The base map shows the ratio of initial formation pressure to the distance to the Precambrian Basement under a unit of MPa/km. The grey circle denoted the recorded seismicity of $M_L \geq 1.3$ up to 2018/12/31. The cumulative fluid injection per well-pad is shown in magenta circles. The white line shows the known faults (Pawley et al., 2018). The black tadpole represents well trajectory, and the black bar labeled HW1 marks the new fracturing well mentioned in Section 3.2. (**c**) Daily observation of monitored seismicity and fluids injection per well-pad. (**d**) The stratigraphy and lithology developed in the studied region (Hui et al., 2021b). Pink, orange, blue, magenta, and grey represent crystalline rock, sandstone, limestone, evaporites, and shale lithology.

variety in the seismicity position and magnitude raises a question regarding the triggering mechanisms of HF-induced seismicity in seismicity-frequent regions.

The triggering mechanisms of induced seismicity mainly include either the increase in pore pressure due to fluid injection or the poroelastic stress perturbation due to injection stimulation (Ellsworth, 2013; Lei et al., 2017). However, when investigated on a local scale, the underlying mechanisms of induced seismicity exhibit a high degree of intricacy. Increasing lines of evidence show that HF-induced seismicity in the Duvernay Shale is susceptible to the combining control of site-specific geological, geomechanical, and operational factors (Pawley et al., 2018; Hui et al., 2021b, 2023a). Such factors contributing to induced seismicity have been found to be proximity to the Precambrian Basement and Swan Hills reef margins (Pawley et al., 2018), initial formation pressure (Shen et al., 2019), critically stressed state of faults (Zhang et al., 2019), regional minimum principal stress (Pawley et al., 2018), and injection fluid volumes (Schultz et al., 2020). Although some factors above could partly explain the susceptibility of HF-induced seismicity in some cases, the quantitative impacts of all related factors on the induced seismicity magnitude remain unclear. In addition, without considering site-specific conditions, the injection volume-magnitude relationship may not be directly utilized to estimate the maximum moment magnitude (M_{wmax}) in response to fluid injection (McGarr, 2014; Wang et al., 2020). Therefore, a novel approach is required to quantify the effects of all related factors on induced seismicity and faithfully forecast the M_{wmax} during HF operations.

Understanding the quantified impact of various factors on induced seismicity requires a comprehensive analysis of large volumes of data (Zhang et al., 2021; Luo et al., 2022). An incomplete or biased investigation of researchers on these factors might hinder the correct understanding of the susceptibility of HF-induced seismicity. Machine learning has been a practical approach to identifying the potential linkages among various parameters and determining the controlling parameters of one event. Pawley et al. (2018) employed the machine learning method to investigate the geological susceptibility of induced seismicity and estimate the

seismogenic activation potential accordingly in the Duvernay play of Western Canada. Perol et al. (2018) applied the neural network method for earthquake detection and prediction in Oklahoma, USA, providing a more robust model and detecting earthquakes 17 times more than the previous model. The machine-learning approaches have also been documented in other prior works for susceptibility analysis and seismicity prediction (Wozniakowska and Eaton, 2020; Asim et al., 2020). Moreover, some researchers also adopted machine learning models in analyzing the fracture dynamics during HF (Zhao et al., 2022). Although accomplishing the machine learning-based data analysis and forecast, these studies failed to integrate different geological, geomechanical, and operational parameters (only one or several) to establish a solid prediction model.

This work develops an integrated machine-learning approach to assess the susceptibility and forecast the magnitude of HF-induced seismicity in the Duvernay Shale. An integrated dataset is collected to obtain the geological, geomechanical, and fracturing parameters as candidate variables, while seismicity magnitude is considered the target variable. The data-mining process determines the primary factors controlling induced seismicity, and the prediction model forecasts the seismicity magnitude. Accordingly, a mitigation strategy based on the prediction model is proposed to guide the fracturing operations of horizontal wells and reduce seismicity risks.

2. Material and methodology

2.1. Datasets description and quality control

An incomplete or biased data compilation might hinder a correct understanding of the susceptibility of HF-induced seismicity via machine learning. Hence, the preparation of comprehensive datasets is significant for data mining (Wozniakowska and Eaton, 2020). To obtain solid results via machine learning, it is essential to consider all relevant data contributing to induced seismicity. We collected the datasets from the potential geological, geomechanical, and operational factors. For a dataset without direct measurements, we employ the proxies that have similar contributions to induced seismicity. The reasons for using selected datasets and their data source are listed below.

Distance to basement. It has been documented that the closer to the Precambrian Basement, the higher the probability of induced earthquakes (Hincks et al., 2018). Additionally, faults in the crystalline basement are more continuous based on the ant tracking results, increasing the likelihood of fault reactivation and nucleation of large-magnitude induced seismicity (Kozłowska et al., 2018). Based on stratigraphic correlations of straight wells, the distance to the Precambrian Basement is measured from Duvernay's bottom to the top Precambrian Basement. For wells without drilling the Precambrian Basement, such a distance is derived from the interpretation of a 3D seismic survey (Pawley et al., 2018). Finally, the Sequential Gaussian Simulation (SGS) method is used to contour its spatial distribution (Fig. 2a) (Hui et al., 2021b; Hui and Gu, 2022).

Distance to reef margins. Statistical analysis suggests that locations of seismicity epicenters were spatially correlated to the margins of the Swan Hills carbonate reefs (Schultz et al., 2017). This coincidence indicates a reef nucleation preference for paleobathymetric highs linked to Precambrian Basement tectonics, representing a proxy for hydrologically conductive faults (Pawley et al., 2018). The distance to the reef margin is calculated by the lateral distance from the fracturing well to the reef margins (Schultz et al., 2017; Pawley et al., 2018). Fig. 2b illustrates its distribution developed by the SGS method.

Distance to known faults. Generally, the reactivation of pre-existing faults is accompanied by the occurrence of HF-induced seismicity. Therefore, locations near known faults (inferred from geophysical or stratigraphic information) indicate potential regions with increased seismic susceptibility (Anyim and Gan, 2020). The distance to a known fault is computed by the shortest distance between the fracturing well and the nearest fault (Fig. 2c). The distribution of known faults in this area has been interpreted from the geophysical or stratigraphic information (Pawley et al., 2018).

Initial formation pressure and minimum principal stress. A close correlation has been found between the initial formation pressure and induced events throughout the Duvernay shale reservoirs (Eaton et al., 2018). The increasing pore pressure can reduce the effective normal stress applied to the associated faults, increasing the likelihood of fault slip (Liu et al., 2011). The state of stress is essential for its influence on the Mohr-Coulomb failure criteria. Thus, the minimum horizontal stress is included to investigate the seismicity susceptibility. We adopted Shen's Matlab program for regional stress and pore pressure calculations (Shen et al., 2019). The available steady end pressure and the instantaneous shut-in pressure (ISIP), derived from a pressure decline analysis of fracturing wells, have also been supplemented to improve pressure and stress data quality (Fig. 2d and e).

Total fluid injection and number of stages, total proppant mass, and horizontal length. Statistical analysis indicates that induced seismicity is associated with fracturing operations that use larger injection volumes (~10^4 to ~10^5 m^3) (Schultz et al., 2020). An increase in the injection volume and the number of fracturing stages can stimulate more hydraulic fractures (Chen et al., 2021), promoting the generation of underground fracture networks and the occurrence of fault reactivation. The enlarging proppant mass and horizontal lengths usually lead to a larger stimulated reservoir volume (SRV). A large SRV increases the likelihood of a hydraulic connection between a stimulated well and pre-existing faults, possibly reactivating such faults and triggering the induced seismicity. These fracturing parameters are available in the GeoSCOUT database, providing real-time information on these candidate parameters.

Wellbore orientation. Statistical results suggest that a safe well-fault distance (i.e., the distance to the farthest detected events from a stage) is 879 m for NS-oriented wells and 749 m for NW-SE-oriented wells in the Fox Creek region (Hui et al., 2021c, 2021d). The wellbore orientation determines the hydraulic fracture orientation, influencing a hydraulic pathway between fracturing wells and pre-existing faults. These data are sourced from the well completions database of GeoSCOUT.

Based on spatial distributions of candidate parameters, induced events favor a short distance to the basement and the reef margins, high initial formation pressure, high minimum principal stress, N–S-oriented horizontal wells, and a large injection volume (Fig. 2). Such observations can be used to corroborate the correctness of the machine-learning-based prediction model.

For other data used in this work, the HF database is compiled of all (~594) horizontal HF well completions in the Duvernay Formation up to 2020/12/31 from the GeoSCOUT database (Fig. 1) (https://www.geologic.com/products/GeoSCOUT/, last accessed on 2022/05/01). The seismicity catalog, including 996 events for 2012–2020, is sourced from the Canadian Composite Seismicity Catalogue (https://www.inducedseismicity.ca/catalogues/, last accessed on 2020/09/21). The magnitude of completeness has been documented to be $M_L = 1.3$ in the Fox Creek region (Schultz et al., 2020). And ten events with a magnitude lower than 1.3 are eliminated in the datasets.

A quality-control process is conducted through a spatiotemporal association filter (SAF). An HF well is labeled a seismogenic well if events occur within a 5-km distance away from the nearest fracturing stage and within a time window of three months after

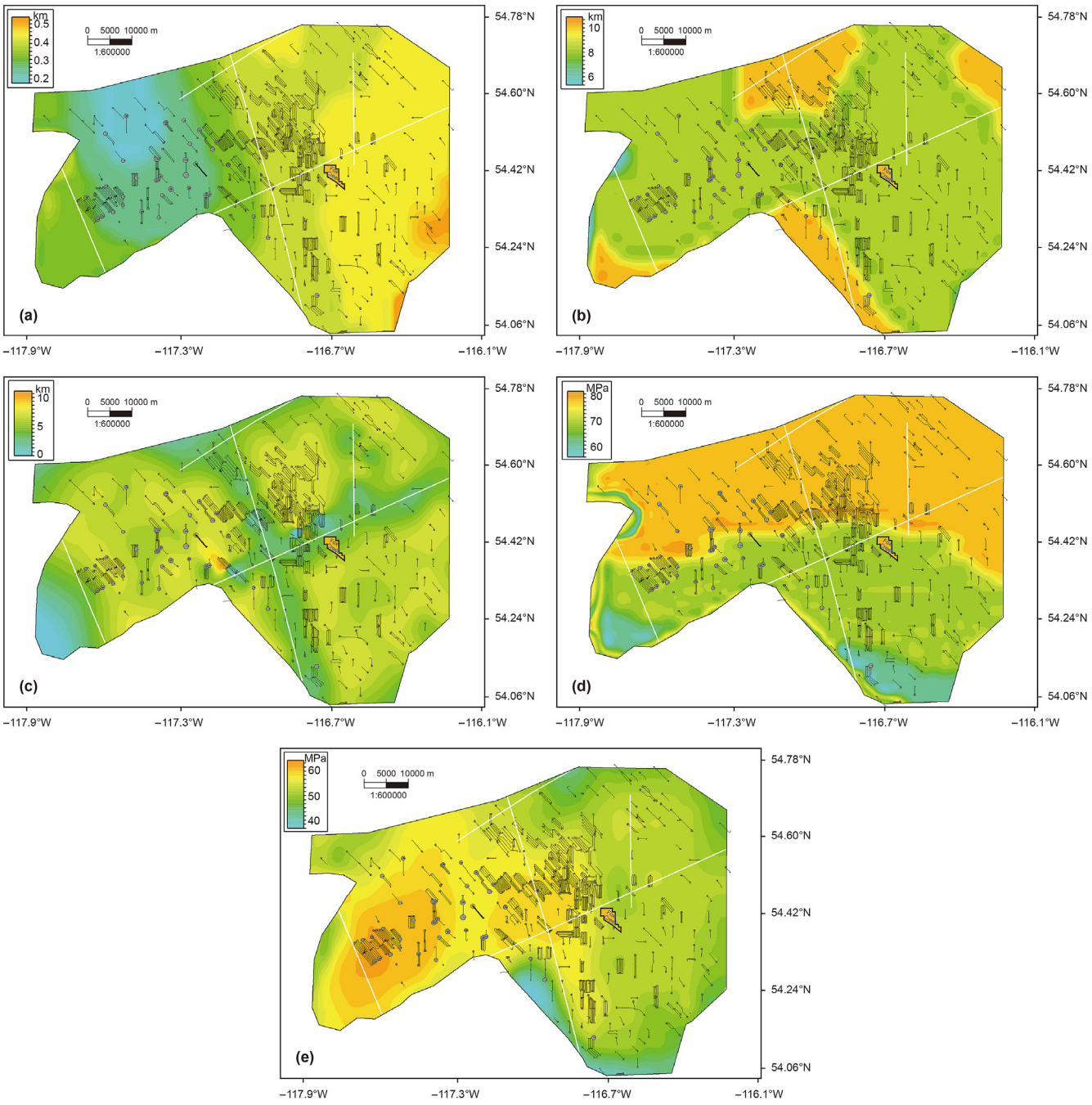

Fig. 2. Distribution of candidate parameters (base map) and target variable M_{wmax} (magnitude-scaled circles). (**a**) Distance to basement (km). (**b**) Distance to reef margins (km). (**c**) Distance to known faults (km). (**d**) Minimum principal stress (MPa). (**e**) Initial formation pressure (MPa).

stimulations (Atkinson et al., 2016). Furthermore, due to the nearly simultaneous fracturing stimulation for different wells within the same well-pad and their cumulative effects on HF-induced seismicity, the related data used in this work are obtained based on a well-pad instead of a single well.

And we consider the maximum magnitudes (M_{max}) of related seismicity clusters as the target parameter. The SAF progress results in a total of 299 well-pads with 299 M_{max}. For computing practicability, aseismic wells (i.e., wells without seismicity) are assigned to zero in the seismicity magnitude. Statistics of these variables are shown in Table 1. It is shown that Table 1 has the information of 299 single data points, and each of them corresponds to a related well pad. Specifically, the input geological, geomechanical and operational parameters of a well pad are derived from Fig. 2a–e. The maximum seismicity magnitude of induced seismicity surrounding a well-pad is regarded as the output variable. It is also noted that 109 out of 299 data points with $M_{wmax} = 0$ (i.e., aseismic HF wells) are incorporated into the datasets.

2.2. Machine learning approach

The machine learning approach identifies the potential linkages among various parameters and determines the controlling parameters contributing to one event. The extra trees (ET) algorithm,

Table 1
Statistics and data source of input and output variables used for the machine learning analysis.

Type	Parameters	Units	Min	Max	Mean	Standard deviation	Source
Geological	Distance to fault	km	0.07	19.5	7.2	4.5	Well completion from GeoSCOUT, Pawley et al., 2018
	Distance to basement	km	0.15	0.55	0.38	0.96	
	Distance to reef	km	5.7	10.9	6.4	1.1	
Geomec-hanical	Initial formation pressure	MPa	36.7	67.1	57.0	5.6	Treatment data and Shen et al. (2019)
	S_{hmin}	MPa	55.7	80.8	74.2	5.8	
Operational	Total fluid pumped	m^3	4194	470118	56811	65497	Well completion and Treatment data from GeoLOGIC
	Total proppant placed	t	249	53366	7148	9382	
	Horizontal length	km	0.368	16.724	2.931	2.91	
	Number of stages		2	330	41	53	
	Well orientation	Degree	0	135	106	54	
Output	Maximum seismicity magnitude		0	4.1	2.68	1.27	Seismicity catalog from CASC

also called extremely randomized trees, is employed in our study to build a computing model of predicted seismicity magnitude in the studied region. The functional diagram of the ET algorithm is shown in Fig. 3. This algorithm generates a large number of base learners in a parallel or serial manner and then uses a voting or averaging strategy to combine these learners, significantly improving the generalization performance and stability of the prediction model (Breiman, 2001). Specifically, ET is an ensemble machine-learning algorithm that combines the predictions from many decision trees (Geurts et al., 2006). It is an ensemble of decision trees related to other ensembles of decision tree algorithms, such as bootstrap aggregation and random forest. The ET approach works by creating numerous unpruned decision trees based on a selected training dataset. Predictions are determined by averaging the predictions of decision trees in a regression case or using majority voting in classification cases. Compared to other algorithms, ET is faster in execution time because it randomly chooses a split point. In addition, ET uses all training samples when constructing a decision tree, and adopts a randomization strategy for the division, showing strong applicability in field cases (Hui et al., 2021a, 2023b). Therefore, ET has a relatively better performance in the data-mining process.

For the data-mining process, the input data are first randomly divided into a training group (60%), a validation group (20%), and a testing group (20%). The ET algorithm is then utilized to run the machine learning models using three data groups. Finally, model prediction performance is assessed by the magnitude of a squared error (MSE) and the determination coefficient (R^2), which are given by:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \tag{1}$$

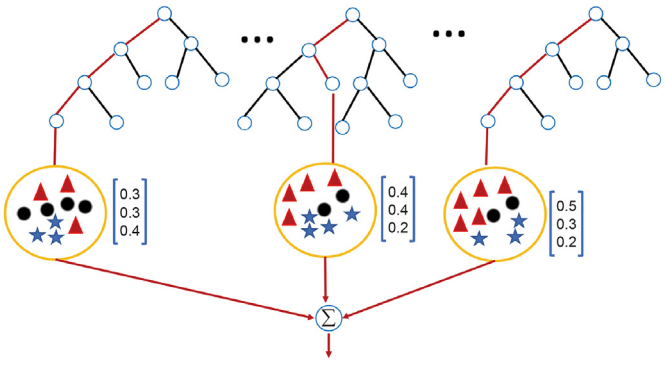

Fig. 3. The functional diagram of the ET algorithm.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \tag{2}$$

where y_i is the ith parameter after normalization; $\hat{y}_i$ is the predicted ith parameter; $\overline{y_i}$ is the mean value of predicted parameters; n is the number of parameters.

During the ET computing process, the random grid search method is employed to optimize the hyperparameters of learning algorithms, and the Adam algorithm is used to minimize a loss (Hui et al., 2021a). Specifically, 1000 independent models are run for each underlying hyperparameter to find the optimal configuration. The details of the computing process using the ET alogrithm are shown in Table 2. Results show that the number of estimators is 75, and the maximum depth is 5. As for the criteria, the mean squared error (MSE), rather than the mean absolute error (MAE), is selected as the indicator to guarantee convergence. The auto, instead of sqrt and log2, is selected in terms of the maximum features. As mentioned before, each model used 60% of the collected data, validated by 20% of the data, and tested 20% randomly. The feature selection process is conducted using data-mining models to determine the controlling factors (Wang et al., 2021, 2023).

3. Results

3.1. Machine learning results

The Pearson correlation results of variables are shown in Fig. 4, illustrating a quantitative relationship between two random parameters. It is noted that the calculated correlation coefficient is based on the entire data, including the training, validating, and testing data. It is also worth noting that the target variable (M_{wmax}) increases with the enlargement of initial formation pressure, minimum principal stress, distance to fault, injection volume, proppant mass, horizontal length, and a number of stages. Furthermore, the target M_{wmax} has a negative correlation with the distance to the basement. It is also shown that the distance to a reef edge and a wellbore orientation are not closely related to the target variable compared with other inputs. It is also worth noting that all

Table 2
The details of the computing process using the ET algorithm.

Hyperparameters	Search space	Optimal configuration
Number of estimators	10, 25, 50, 75, 100	75
Maximum depth	3, 4, 5, 6, 7, 8	5
Criterion of split	MSE, MAE	MAE
Maximum features	Sqrt, auto, log2	Auto

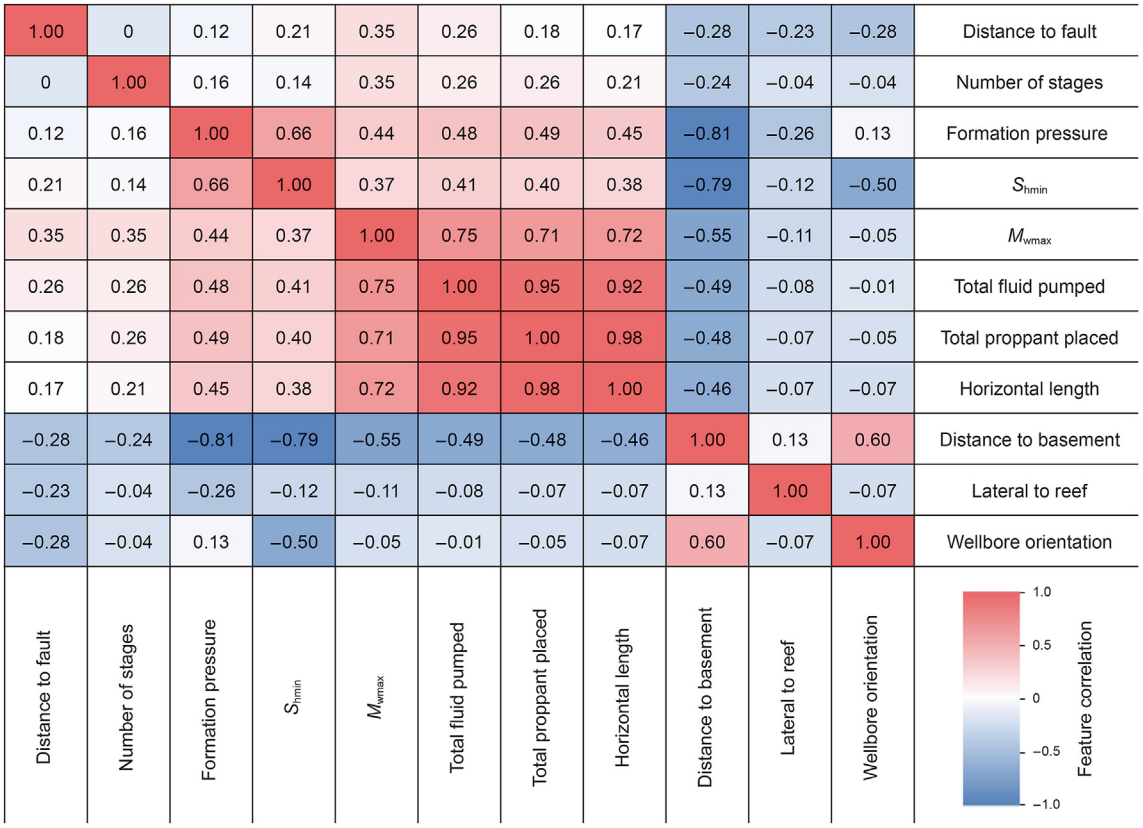

Fig. 4. Pearson correlation matrix of predictive and target variables. The number inside each unit denotes the correlation value of two features. The color label is given to the left of the plot.

four operational parameters positively influence the target variable, demonstrating the role of fracturing parameters on seismicity magnitude (McGarr, 2014).

Moreover, among all geological and geomechanical parameters, the distance to the basement has the largest absolute magnitude in the Pearson coefficient, indicating that this parameter has a more paramount influence on M_{wmax} than other parameters. This result can be explained by a possible hydraulic connection between basement faults and the stimulated well. Once proximity to the basement, the fracturing fluids tend to migrate vertically through permeable fracture or fault networks to the basement faults, increasing the pore pressure within the basement faults to cause a fault to slip (Galloway et al., 2018; Tang et al., 2021).

The results of normalized frequencies for the studied parameters are shown in Fig. 5a, indicating the feature importance in an ET-based computing model. Such parameters are displayed in decreasing order of importance by the distance to a fault, the distance to the basement, minimum principal stress, cumulative injection volume, initial formation pressure, the number of fracturing stages, horizontal length, cumulative proppants placed, wellbore orientation, and the distance to a reef. This result matches the results from Pearson correlations (Fig. 4). However, the total proppants placed are an exception, which ranks the third lowest in feature importance but has a higher rank in Pearson's correlations. This can be explained by the fact that a cumulative injection volume owns a close linear correlation with total proppant mass ($R^2 = 0.95$), hence masking the latter's role in feature importance. Fig. 5b illustrates the ET-based performance of the tested dataset as a function of the number of features. It is noted that the tested dataset with six selected parameters has the largest coefficient of determination ($R^2 = 0.87$) and the lowest mean-squared error (MSE = 0.13), achieving high prediction performance with fewer chosen parameters. Therefore, such top six parameters are used to build the subsequent seismicity prediction model.

Pawley et al. (2018) used the machine learning method to study the HF-induced seismicity throughout the Duvernay Formation (Fig. 1a). He concluded that HF-induced seismicity in the Duvernay Formation was primarily susceptible to a distance to the basement, initial formation pressure, minimum horizontal stress, and proximity to a reef edge, while a distance to known faults is less important. In comparison with our results, both types of research emphasize the role of a distance to the basement, initial formation pressure, and minimum horizontal stress but have the opposite outcomes regarding the proximity to a reef edge and the distance to known faults. One reason for this mismatch is that we incorporate the operational parameters into the candidate datasets, which affects the final feature importance in the prediction model. Another one is attributed to the site-specific features. In other words, the selected well-pads near Fox Creek have a certain distance from these faults and have no relation with the reef edge. Moreover, a wellbore orientation has a low rank in the model. Such poor performance may be explained by the fact that 83% of the selected well-pad have the wellbore orientation of NW-SE. Therefore, the difference of this parameter in this work has relatively less influence on the prediction results. It is also noted that we did not consider some factors in this study, including lithium concentration, dolomite occurrence, and natural seismicity rates. The reasons for neglecting these features are the unavailability of a database related to them and their relatively low importance in prior works (Pawley et al., 2018). Additionally, the faults used in this work are the known large faults, as shown in Fig. 1. It is shown that the west section, away from the known large faults, is the seismicity-

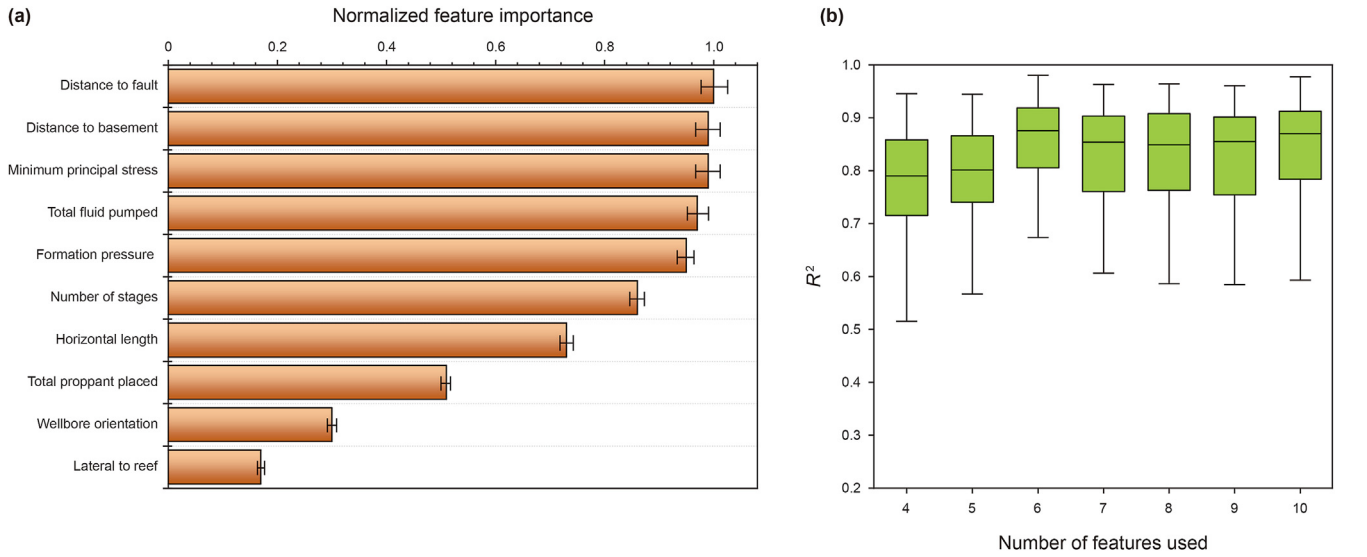

Fig. 5. (**a**) The normalized feature importance for candidate parameters. (**b**) The R^2 of tested datasets with a different number of candidate features.

frequent region. Therefore, the positive correlation between M_{wmax} and the distance to fault follows this observation. However, if we obtain the available 3D seismic data in the examined region, small-scale pre-existing faults could be identified. The closer the formation is to these small-scale faults, the more brittle it may be and the higher the probability of seismicity with higher M_{wmax}. Overall, this work provides insights into how multiple factors influenced seismicity magnitude and successful field application in mitigating potential seismicity risks via machine learning.

3.2. Seismicity prediction model and application

A significant application of the ET-based models is to establish a seismicity prediction model that can optimize the operational parameters and mitigate potential seismicity risks. The establishment of the seismicity prediction model is as follows. First, we investigate a relationship between fluid injection and placed proppants of fracturing wells within the studied region. The upper and lower bounds of cumulative fluids and proppants pumped are obtained to guide future stimulation operations (Fig. 6a). Next, we run the ET-based seismicity prediction models under different operational parameters as well as site-specific geological and geomechanical parameters. Finally, this prediction model determines the recommended cumulative injection volume and proppant mass with a moderate seismicity magnitude. Given that the magnitude of 2.5 has been demonstrated to be the long-term detection threshold for induced earthquakes in Western Canada (Schultz et al., 2017), here we consider the moderate magnitude as M 2.5. We investigate the fluid and proppants amount to reduce the potential seismic risks. In addition, as the number of stages and horizontal length have positive relations with cumulative fluid injection (Fig. 4), these two factors are assigned to values proportional to cumulative fluid injection in the prediction model.

Fig. 6b and c illustrates the initial and recommended fracturing parameters in the 3.05 and 3.6 cases (Fig. 1b). Table 3 compares the initial and recommended fluid injection and proppants placed in both cases. Based on the ET-based models, both 3.05 and 3.6 earthquakes were triggered due to the larger fracturing job sizes than the recommended ones. To reduce the potential seismicity magnitude, the recommended fluid volume and proppant mass in 3.05 and 3.6 cases would have been designed to be less than approximately 70,000 m^3 and 10,000 t, and 80,000 m^3, and 10,000 t, respectively. In other words, if the total injection volume in both cases is reduced by 58.8% and 55.6%, respectively, the risks of $M > 3$ induced seismicity can be mitigated with a magnitude less than 2.5 instead.

We also studied a new case to demonstrate the application of the ET-based prediction model (Fig. 7). A new horizontal well was drilled in February 2021 near the M 3.6 event (Fig. 1b). Following the procedures of prediction model establishment, the recommended cumulative fluid injection and proppant mass cover a range of 80,000–90,000 m^3 and 10,000–12,500 t (large rectangle in Fig. 7b). The 83,411 m^3 and a total proppant mass of 12,792 t are adopted in the fracturing operations towards this well in March 2021 (small rectangle in Fig. 7b). Interestingly, based on the prediction model, the predicted M_{max} of induced events under such fracturing job size is 2.9, quite similar to that of actual induced events (i.e., 3.05), indicating the robustness of the prediction model in this work.

Based on the seismicity prediction model in the three cases (Figs. 6 and 7), the cumulative fluid injection and proppants placed are recommended at 70,000–90,000 m^3 and 9500–12,500 t, respectively, to avoid larger-magnitude induced events. To create a map of predicted seismicity magnitudes, we employ the mean value (i.e., 80,000 m^3 and 11,500 t) to forecast a seismicity magnitude under site-specific features. The SGS method is used to generate the magnitude map (Fig. 8). It is worth noting that the prediction map matches the observed spatial seismicity. Moreover, 84.6% of the fracturing wells will be seismic wells with a seismicity magnitude lower than the actual one if the fracturing job size is reduced, demonstrating the necessity of this work prior to hydraulic fracturing. This machine learning workflow can also be used to investigate seismicity susceptibility and propose mitigation strategies accordingly in other seismicity-frequent regions. It is extremely practical to guide the operators to implement safe operations in developing unconventional resources.

4. Discussion

This study uses various public data sources to estimate the candidate parameters. The data quality of these parameters depends on the robustness of such public datasets. For example, the

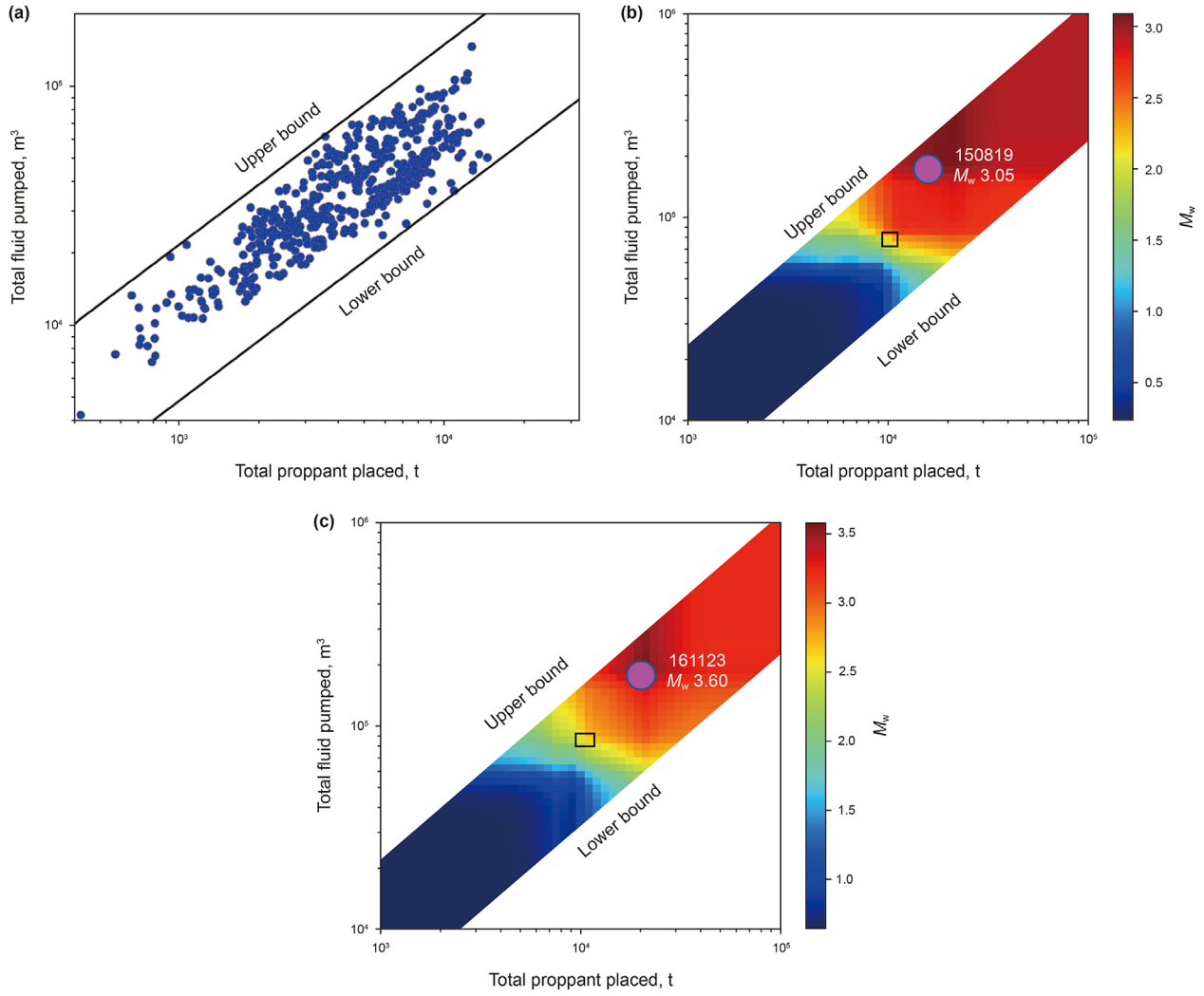

Fig. 6. (**a**) The linkage between the total pumped volume of fracturing fluids and the mass of proppant placed. (**b**, **c**) The initial and recommended fracturing parameters in two cases. The color scale bar represents the expected seismicity magnitude. The pink circles denote the initial operational parameters that triggered the corresponding event. The black rectangular shows the recommended fracturing parameters range that could mitigate the risks of large magnitude seismicity.

Table 3
Comparison of the original and recommended fracturing design in both 3.05 and 3.6 cases.

Original			Recommended			Comparison	
M_{wmax}	Cumulative fluid injection, m^3	Cumulative proppant mass, t	M_{wmax}	Cumulative fluid injection, m^3	Cumulative proppant mass, t	Reduction percentage of injection, %	Reduction percentage of proppant, %
3.05	169788	16170	2.5	70000	10000	58.8	38.2
3.6	180206	19806	2.5	80000	10000	55.6	49.5

distances to known faults and Swan Hills are primarily derived from previous work (Pawley et al., 2018). Additionally, the distance to the basement, *in-situ* stress, and initial formation pressure derived from prior works are further calibrated using this region's available logging and treatment data, improving the data quality of the candidate parameters (Fig. 9a) (Pawley et al., 2018; Shen et al., 2019; Jing et al., 2021, 2022). The target variable M_{wmax} has been derived from the focal mechanisms of monitored seismicity. However, because of the monitoring resolution of seismology networks and different methods of focal solutions, the derived seismicity magnitude has underlying uncertainty (Zhu et al., 2017). Additionally, we used a M_{max} of zero for the cases that were aseismic, which is different from the regional detection threshold. However, this setting is used for computing applicability. As shown in Fig. 8, the blue region represents a possible distribution of aseismic fracturing wells to guide future stimulations. Furthermore, we selected 60% of the data for training, 20% for validation, and 20% for testing. This 6:2:2 can solve the hyperparameter problem and obtain a relatively robust result. Regarding the limited 299 data points, we conducted the uncertainty analysis of the ET model via bootstrapping a training set, and the result is shown in Fig. 9b. It is noted that the calculation errors in regions used by the prediction model (blue areas in Fig. 9b) are less than 0.07, indicating the robustness of this prediction model developed by the ET algorithm.

We also compare the machine learning results via ET, gradient boosting decision tree (GBDT), artificial neural network (ANN), and

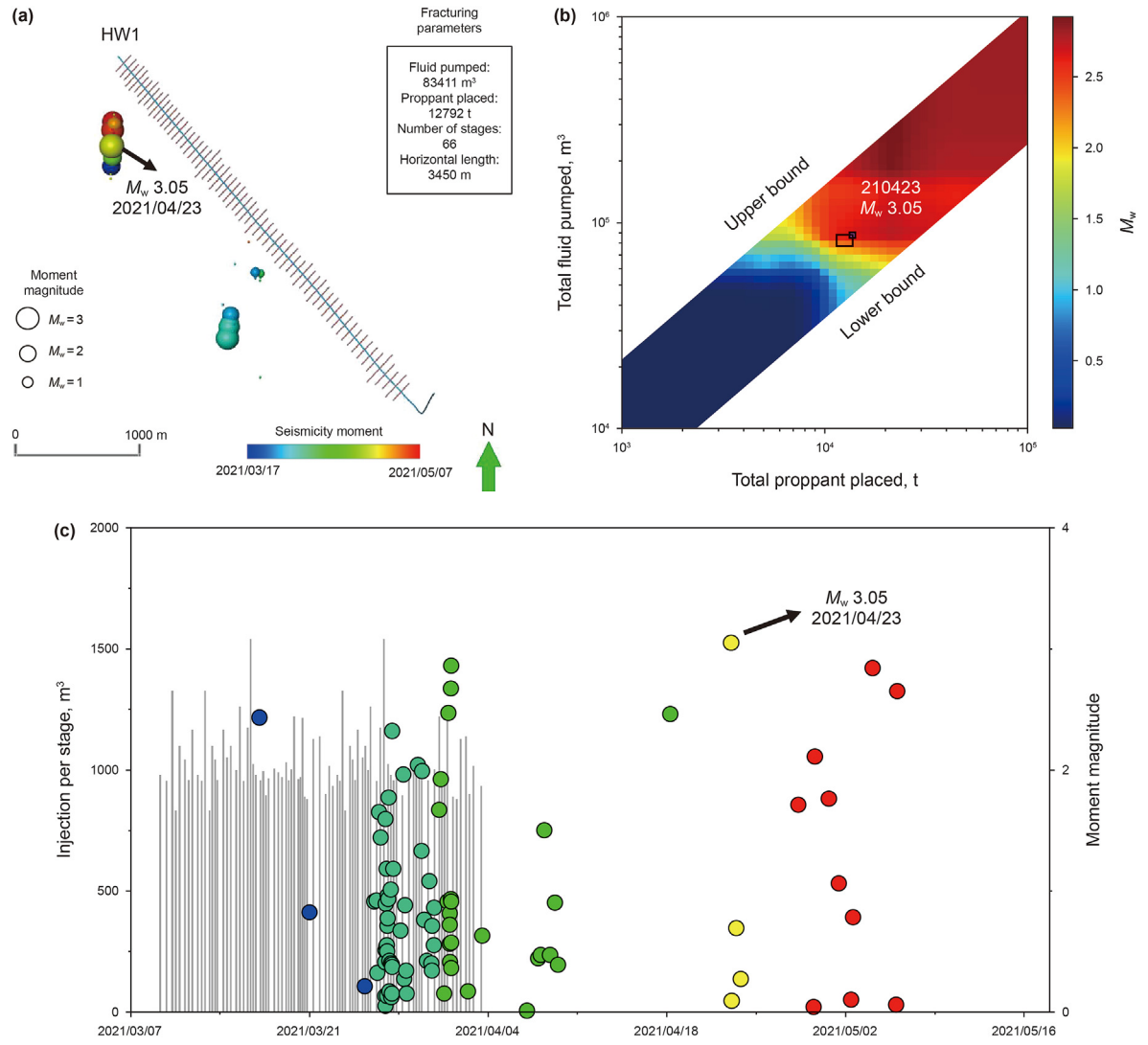

Fig. 7. (**a**) The distribution of new fracturing well and associated induced seismicity. The location of this horizontal well is shown in Fig. 1b. The induced events are scaled by the magnitude and colored by the induced moment. The largest event is the M_w 3.05 event that occurred on 2021/04/23. (**b**) Comparison between the recommended (large rectangular) and actual (small rectangular) parameters. (**c**) Temporal view of induced seismicity and fracturing stages of the new well.

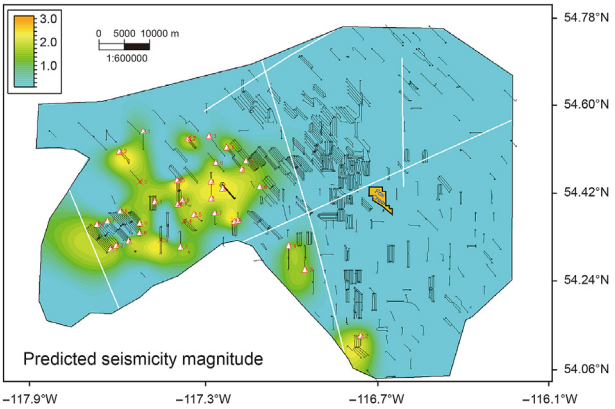

Fig. 8. Predicted seismicity magnitude distribution under site-specific features using recommended fracturing parameters via ET-based prediction model. The triangle marks the wells with an estimated magnitude (red number) lower than the actual one, while the diagonal cross labels the wells with a magnitude (red number) larger than the actual one.

random forest (RF). The prediction performance of the tested dataset using four algorithms as a function of the number of selected parameters is shown in Figs. 5b and 9c–e. It is shown that, for one thing, the tested dataset with six selected parameters has a large coefficient of determination (R^2) in all algorithms, which achieves the goal of high prediction performance with fewer selected parameters. For another, the average R^2 of the tested dataset using six selected parameters gives the value of 0.791 for ANN, 0.799 for RF, 0.870 for ET, and 0.811 for GBDT. Furthermore, the average mean-squared-errors (MSE) are computed to be 0.249 for ANN, 0.189 for RF, 0.130 for ET, and 0.187 for GBDT. Therefore, ET is the best algorithm in this study, with the highest prediction performance and the least calculation errors.

To investigate the role of well-pad data for the responding M_{wmax} variable, we also built an ET-based prediction model using the dataset of a single horizontal well, compared to that of the whole HF pad. It is shown that the determination coefficient of the prediction model with a single well is only 0.53, less than 0.87 with the well-pad. The lower prediction performance using a single well leads to nearly identical parameters for single wells within the same HF-pad, reducing the prediction effectiveness in the model

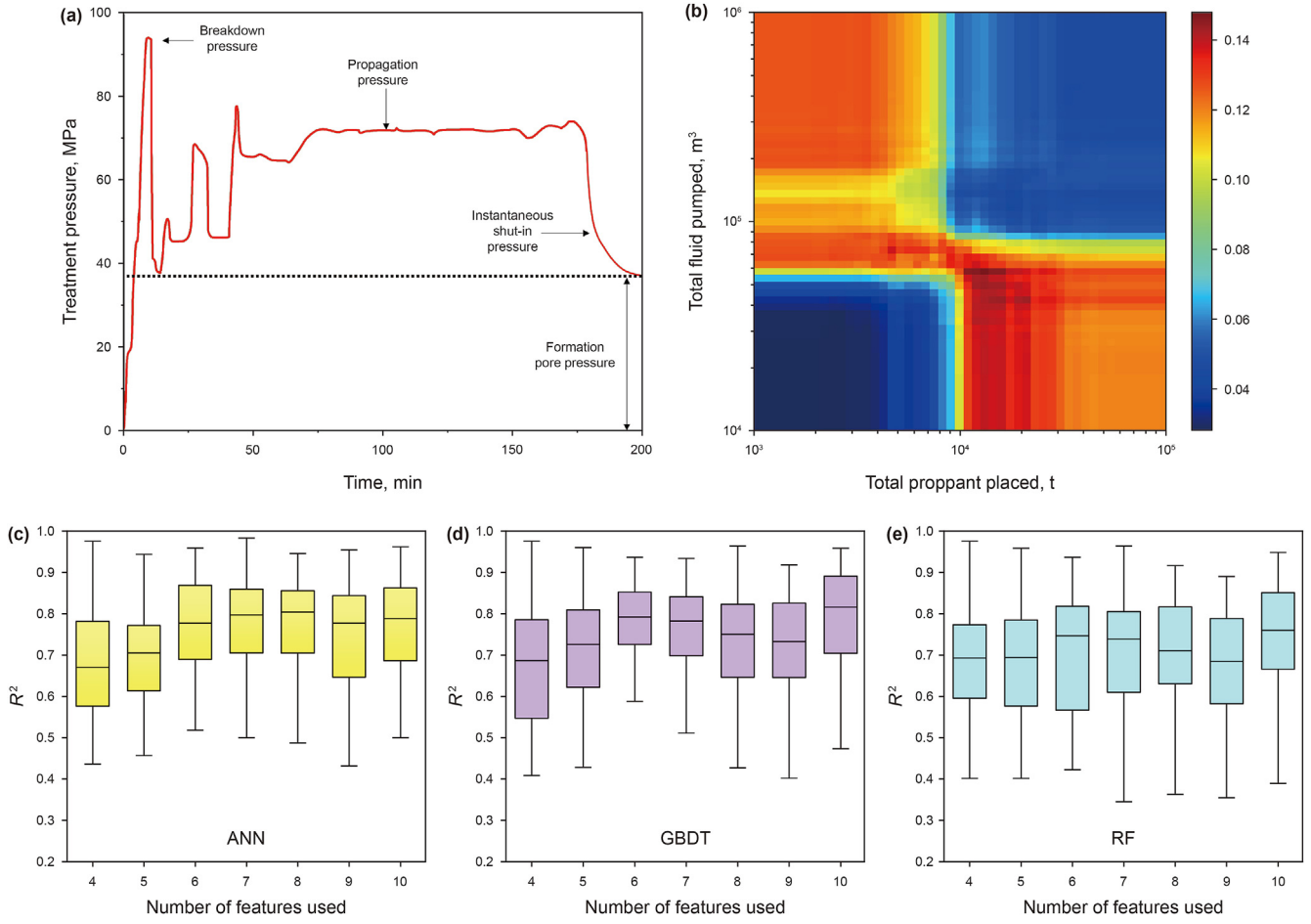

Fig. 9. (**a**) The treatment plot for initial formation pressure and stress estimation. (**b**)The calculation error of the ET-based prediction model. (**c**–**e**) The R^2 of tested datasets with a different number of candidate features using ANN, GBDT and RF, respectively.

(parameter redundancy). Therefore, the well-pad-based data have a better performance in the prediction model via the machine learning method. Moreover, some researchers adopted novel onset time methods to extract the temporal features from 4D seismic surveys (Liu et al., 2020). This "onset time" method could be a tool used for investigating the temporal relation between seismic events.

Furthermore, although 0 seismicity is the target goal in the shale development area, the $M > 0$ induced events are commonly accompanied by hydraulic fracturing in the Fox Creek region (Fig. 1c), especially in the western part of the studied region (Hui et al., 2021e, f). Because the long-term detection threshold for produced earthquakes in Western Canada has been shown to be 2.5 (Schultz et al., 2017), here we refer to the intermediate magnitude as M 2.5 in the examined region (Fig. 6c). For some seismicity-quiescent regions, the magnitude of 0 may be regarded as the target magnitude used for the seismicity prediction model.

5. Conclusions

In this work, an integrated machine-learning approach is developed to assess the susceptibility and forecast the magnitude of HF-induced seismicity in the Duvernay Shale. The dataset is obtained from the field to derive associated candidate and target variables. The machine learning approach determines the factors controlling induced seismicity and forecasts the seismicity magnitude. The mitigation strategy based on a machine learning model is proposed to guide the fracturing operations of horizontal wells and reduce the earthquake risks in the studied region. The primary conclusions of this paper are drawn as follows.

(1) A comprehensive dataset from 594 fracturing wells throughout the Duvernay Formation near Fox Creek, Alberta, is collected. The spatiotemporal association filter results in a total of 299 well-pads with a related maximum magnitude.
(2) Feature importance indicates that a distance to a fault, a distance to the basement, minimum principal stress, cumulative fluid injection, initial formation pressure, and the number of fracturing stages are among significant model predictors.
(3) $M > 3$ induced seismicity can be potentially mitigated if reducing a fluid injection volume and proppant mass per well-pad based on the ET-based model.
(4) ET-based prediction map of seismicity magnitude matches the observed spatial seismicity. 84.6% of the fracturing wells will be triggered with a seismicity magnitude lower than the actual one if reducing the fracturing job size.
(5) This machine learning workflow can be applied to other seismicity-frequent regions to mitigate potential seismicity risks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been made possible by contributions from the Natural Sciences and Engineering Research Council (NSERC)/Energi Simulation Industrial Research Chair in Reservoir Simulation and the Alberta Innovates (iCore) Chair in Reservoir Modeling. This research was supported by the Science Foundation of China University of Petroleum, Beijing (No. 2462023BJRC001) and the National Natural Science Foundation of China Joint Fund Key Support Project (No. U19B6003).

References

Anyim, K., Gan, Q., 2020. Fault zone exploitation in geothermal reservoirs: production optimization, permeability evolution and induced seismicity. Advances in Geo-Energy Research 4 (1), 1–12. https://doi.org/10.26804/ager.2020.01.01.

Asim, K.M., Moustafa, S.S., Niaz, I.A., et al., 2020. Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in Cyprus. Soil Dynam. Earthq. Eng. 130, 105932. https://doi.org/10.3390/rs12020301.

Atkinson, G.M., Eaton, D.W., Ghofrani, H., et al., 2016. Hydraulic fracturing and seismicity in the Western Canada sedimentary basin. Seismol Res. Lett. 87 (3), 631–647. https://doi.org/10.1785/0220150263.

Baranova, V., Mustaqeem, A., Bell, S., 1999. A model for induced seismicity caused by hydrocarbon production in the Western Canada Sedimentary Basin. Can. J. Earth Sci. 36 (1), 47–64. https://doi.org/10.1139/e98-080.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, J., Wang, L., Wang, C., et al., 2021. Automatic fracture optimization for shale gas reservoirs based on gradient descent method and reservoir simulation. Advances in Geo-Energy Research 5 (2), 191–201. https://doi.org/10.46690/ager.2021.02.08.

Eaton, D.W., Igonin, N., Poulin, A., et al., 2018. Induced seismicity characterization during hydraulic fracture monitoring with a shallow-wellbore geophone array and broadband sensors. Seismol Res. Lett. 89 (5), 1641–1651. https://doi.org/10.1785/0220180055.

Ellsworth, W.L., 2013. Injection-induced earthquakes. Science 341, 1225942. https://doi.org/10.1126/science.1225942.

Galloway, E., Hauck, T., Corlett, H., et al., 2018. Faults and associated karst collapse suggest conduits for fluid flow that influence hydraulic fracturing-induced seismicity. Proc. Natl. Acad. Sci. USA 115 (53), E10003–E10012. https://doi.org/10.1073/pnas.1807549115.

Geurts, P., Damien, E., Louis, W., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Ghofrani, H., Atkinson, G.M., 2020. Activation rate of seismicity for hydraulic fracture wells in the western Canada sedimentary basin. Bull. Seismol. Soc. Am. 110 (5), 2252–2271. https://doi.org/10.1785/0120200002.

Hincks, T., Aspinall, W., Cooke, R., et al., 2018. Oklahoma's induced seismicity strongly linked to wastewater injection depth. Science 359 (6381), 1251–1255. https://doi.org/10.1126/science.aap7911.

Hui, G., Gu, F., 2022. An integrated method to mitigate hazards from hydraulic fracturing–induced earthquakes in the Duvernay Shale Play. SPE Reservoir Eval. Eng. https://doi.org/10.2118/210287-PA.

Hui, G., Chen, S., He, Y., et al., 2021a. Production forecast for shale gas in unconventional reservoirs via machine learning approach: case study in Fox Creek, Alberta. J. Nat. Gas Sci. Eng. 94, 104045. https://doi.org/10.1016/j.jngse.2021.104045.

Hui, G., Chen, S., Chen, Z., et al., 2021b. Comprehensive characterization and mitigation of hydraulic fracturing-induced seismicity in Fox Creek, Alberta. SPE J. 26 (5), 2736–2747. https://doi.org/10.2118/206713-PA.

Hui, G., Chen, S., Chen, Z., et al., 2021c. Influence of hydrological communication between basement-rooted faults and hydraulic fractures on induced seismicity: a case study. J. Petrol. Sci. Eng. 206, 109040. https://doi.org/10.1016/j.petrol.2021.109040.

Hui, G., Chen, S., Chen, Z., et al., 2021d. Investigation on two Mw 3.6 and Mw 4.1 earthquakes triggered by poroelastic effects of hydraulic fracturing operations near Crooked Lake, Alberta. J. Geophys. Res. Solid Earth 126, e2020JB020308. https://doi.org/10.1029/2020JB020308.

Hui, G., Chen, S., Chen, Z., et al., 2021e. Role of fluid diffusivity in the spatiotemporal migration of induced earthquakes during hydraulic fracturing in unconventional reservoirs. Energy Fuel. 35 (21), 17685–17697. https://doi.org/10.1021/acs.energyfuels.1c02950.

Hui, G., Chen, S., Chen, Z., et al., 2021f. An integrated approach to characterize hydraulic fracturing-induced seismicity in shale reservoirs. J. Petrol. Sci. Eng. 196, 107624. https://doi.org/10.1016/j.petrol.2020.107624.

Hui, G., Chen, Z., Wang, P., et al., 2022. Mitigating risks from hydraulic fracturing-induced seismicity in unconventional reservoirs: case study. Sci. Rep. 12, 12537. https://doi.org/10.1038/s41598-022-16693-3.

Hui, G., Chen, Z., Lei, Z., et al., 2023a. A synthetical geoengineering approach to evaluate the largest hydraulic fracturing-induced earthquake in the East Shale Basin. Alberta. Pet. Sci. https://doi.org/10.1016/j.petsci.2023.01.006.

Hui, G., Chen, Z., Wang, Y., et al., 2023b. An integrated machine learning-based approach to identifying controlling factors of unconventional shale productivity. Energy 266, 126512. https://doi.org/10.1016/j.energy.2022.126512.

Jing, G., Chen, Z., Hui, G., 2021. A novel model to determine gas content in naturally fractured shale. Fuel 306, 121714. https://doi.org/10.1016/j.fuel.2021.121714.

Jing, G., Chen, Z., Hu, X., Hui, G., et al., 2022. Influence of different shut-in periods after fracturing on productivity of MFHW in Duvernay shale gas formation with high montmorillonite content. Fuel 314, 122719. https://doi.org/10.1016/j.fuel.2021.122719.

Kozłowska, M., Brudzinski, M.R., Friberg, P., et al., 2018. Maturity of nearby faults influences seismic hazard from hydraulic fracturing. Proc. Natl. Acad. Sci. USA 115 (8), E1720–E1729. https://doi.org/10.1073/pnas.1715284115.

Lei, X., Huang, D., Su, J., et al., 2017. Fault reactivation and earthquakes with magnitudes of up to Mw4.7 induced by shale-gas hydraulic fracturing in Sichuan Basin, China. Sci. Rep. 7, 7971. https://doi.org/10.1038/s41598-017-08557-y.

Liu, T., Chen, H., Hetz, G., et al., 2020. Integration of time-lapse seismic data using the onset time approach: the impact of seismic survey frequency. J. Petrol. Sci. Eng. 189, 106989. https://doi.org/10.1016/j.petrol.2020.106989.

Liu, Y., Xu, L., Yang, D., 2011. Pore pressure diffusion characteristics of Longtan reservoir-induced-earthquake. Chin. J. Geophys. 54 (4), 1028–1037. https://doi.org/10.3969/j.issn.0001-5733.2011.04.017 (in Chinese).

Luo, S., Xiao, L., Jin, Y., et al., 2022. A machine learning framework for low-field NMR data processing. Petrol. Sci. 19, 581–593. https://doi.org/10.1016/j.petsci.2022.02.001.

McGarr, A., 2014. Maximum magnitude earthquakes induced by fluid injection. J. Geophys. Res. Solid Earth 119 (2), 1008–1019. https://doi.org/10.1002/2013JB010597.

Pawley, S., Schultz, R., Playter, T., et al., 2018. The geological susceptibility of induced earthquakes in the Duvernay play. Geophys. Res. Lett. 45 (4), 1786–1793. https://doi.org/10.1002/2017GL076100.

Perol, T., Gharbi, M., Denolle, M., 2018. Convolutional neural network for earthquake detection and location. Sci. Adv. 4, e1700578. https://doi.org/10.1126/sciadv.1700578.

Schultz, R., Stern, V., Gu, Y.J., 2014. An investigation of seismicity clustered near the Cordel Field, west central Alberta, and its relation to a nearby disposal well. J. Geophys. Res. Solid Earth 119, 3410–3423. https://doi.org/10.1002/2013JB010836.

Schultz, R., Wang, R., Gu, Y.J., et al., 2017. A seismological overview of the induced earthquakes in the Duvernay play near Fox Creek, Alberta. J. Geophys. Res. Solid Earth 122 (1), 492–505. https://doi.org/10.1002/2016JB013570.

Schultz, R., Skoumal, R.J., Brudzinski, M.R., et al., 2020. Hydraulic fracturing-induced seismicity. Rev. Geophys. 58, e2019RG000695. https://doi.org/10.1029/2019RG000695.

Shen, L., Schmitt, D.R., Haug, K., 2019. Quantitative constraints to the complete state of stress from the combined borehole and focal mechanism inversions: Fox Creek, Alberta. Tectonophysics 764, 110–123. https://doi.org/10.1016/j.tecto.2019.04.023.

Tang, Y., Weng, A., Yang, Y., et al., 2021. Connection between earthquakes and deep fluids revealed by magnetotelluric imaging in Songyuan, China. Sci. China Earth Sci. 64, 161–176. https://doi.org/10.1007/s11430-019-9633-y.

Wang, B., Harrington, R., Liu, Y., et al., 2020. A study on the largest hydraulic-fracturing-induced earthquake in Canada: observations and static stress-drop estimation. Bull. Seismol. Soc. Am. https://doi.org/10.1785/0120190261.

Wang, H., Chen, Z., Chen, S., et al., 2021. Production forecast and optimization for parent-child well pattern in unconventional reservoirs. J. Petrol. Sci. Eng. 203, 108899. https://doi.org/10.1016/j.petrol.2021.108899.

Wang, H., Wang, S., Chen, S., Hui, G., 2023. Predicting long-term production dynamics in tight/shale gas reservoirs with dual-stage attention-based TEN-Seq2Seq model: a case study in Duvernay formation. Geoenergy Science and Engineering 223, 211495. https://doi.org/10.1016/j.geoen.2023.211495.

Wetmiller, R.J., 1986. Earthquakes near Rocky Mountain House, Alberta, and their relationship to gas production facilities. Can. J. Earth Sci. 23 (2), 172–181. https://doi.org/10.1139/e86-020.

Wozniakowska, P., Eaton, D.W., 2020. Machine learning-based analysis of geological susceptibility to induced seismicity in the Montney Formation. Canada. Geophys. Res. Lett. 47, e2020GL089651. https://doi.org/10.1029/2020GL089651.

Yang, H., Liu, Y., Wei, M., et al., 2017. Induced earthquakes in the development of unconventional energy resources. Sci. China Earth Sci. 60, 1632–1644. https://doi.org/10.1007/s11430-017-9063-0.

Zhang, H.L., David, W.E., German, R., Suzie, Q.J., 2019. Source-mechanism analysis and stress inversion for hydraulic-fracturing-induced event sequences near Fox Creek, Alberta. Bull. Seismol. Soc. Am. 109 (2), 636–651. https://doi.org/10.1785/0120180275.

Zhang, Y., Xi, K., Cao, Y., et al., 2021. The application of machine learning under supervision in identification of shale lamina combination types —a case study of Chang 73 sub-member organic-rich shales in the Triassic Yanchang Formation, Ordos Basin, NW China. Petrol. Sci. 18, 1619–1629. https://doi.org/10.1016/j.petsci.2021.09.033.

Zhao, X., Jin, F., Liu, X., Zhang, Z., Cong, Z., Li, Z., Tang, J., 2022. Numerical study of fracture dynamics in different shale fabric facies by integrating machine learning and 3-D lattice method: a case from Cangdong Sag, Bohai Bay basin, China. J. Petrol. Sci. Eng. 218, 110861. https://doi.org/10.1016/j.petrol.2022.110861.

Zhu, Y., Wang, J., Sun, F., et al., 2017. Micro-seismic monitoring and instrument for hydraulic fracturing in the low-permeability oilfield. Chin. J. Geophys. 60 (11), 4282–4293 (in Chinese). https://10.6038/cjg20171116.