Delineating the Controlling Factors of Hydraulic Fracturing-Induced Seismicity in the Northern Montney Play, Northeastern British Columbia, Canada, With Machine Learning

Bei Wang[*,1,2] , Honn Kao[1,2] , Ramin M. H. Dokht[2] , Ryan Visser[2,3], and Hongyu Yu[2]

Abstract

Recent studies confirm that the distribution of injection-induced earthquakes (IIE) can be related to both natural (e.g., tectonic, geological, and hydrological) settings and operational details. However, the relative importance of operational factors with respect to the natural ones has not been fully understood for the western Canada sedimentary basin. In this study, we train the eXtreme Gradient Boosting (XGBoost) machine-learning algorithm to comprehensively evaluate six geological and seven industrial operational factors suspected to be correlated with the distribution of IIE in the northern Montney play (NMP), British Columbia. We then derive the Shapley Additive Explanations values to quantitatively interpret the outputs from XGBoost. Our results reveal that operational and geological factors have comparable contributions to the IIE distribution. The top four features that contribute most to the seismicity pattern are horizontal distance to the Cordilleran deformation front, cumulative injected volume, shut-in pressure and vertical distance to the Debolt formation (with respect to the hydraulic fracturing [HF] depth). Features with secondary influence are the thickness of the Montney formation, breakdown pressure, cumulative fault length per unit area, and vertical distance to the basement (with respect to the HF depth). Other remaining features (e.g., the average treating pressure and injection rate) appear the least related. Our results provide critical information to establishing a comprehensive susceptibility model that includes key geological and operational factors affecting the IIE distribution in the NMP area.

Introduction

The Montney play in northeastern British Columbia (NEBC) is a major unconventional gas play in the western Canada sedimentary basin (WCSB), which can be further divided into the northern Montney play (NMP) and southern Montney play (SMP, Fig. 1). Although the foreland tectonics of the Canadian Cordillera mainly controls the structural development in NMP, the main geological structure for SMP is the Dawson Creek graben complex. In NEBC, the Montney formation is predominantly formed by siltstone and gray shale and exhibits a northeastward thinning wedge (Davies *et al.*, 1997). It is the main target of hydraulic fracturing (HF) stimulations due to its enriched unconventional oil and gas resources.

As the shale gas and tight oil development increases, the number of injection-induced earthquakes (IIE) also started to surge (Atkinson *et al.*, 2016; Mahani *et al.*, 2017; Schultz *et al.*, 2020), drawing serious public concerns. The largest HF-related IIE in the WCSB occurred on 17 August 2015, in NMP, with **M** 4.6 (Wang, Harrington, *et al.*, 2020). As the development shifted to the south, several **M** >3.5 earthquakes have occurred in SMP since 2018. For example, an **M** 4.2 earthquake occurred on 30 November 2018, near the Dawson Creek area Peña Castro *et al.* (2020). To monitor the increasing rate of seismicity, more than twenty seismic stations operated by various institutions have been installed.

1. School of Earth and Ocean Sciences, University of Victoria, Victoria, British Columbia, Canada, https://orcid.org/0000-0002-8908-8891 (BW); https://orcid.org/0000-0001-9150-9416 (HK); 2. Pacific Geoscience Centre, Geological Survey of Canada, Sidney, British Columbia, Canada, https://orcid.org/0000-0002-3127-363X (RMHD); 3. Geoscience BC, Vancouver, British Columbia, Canada

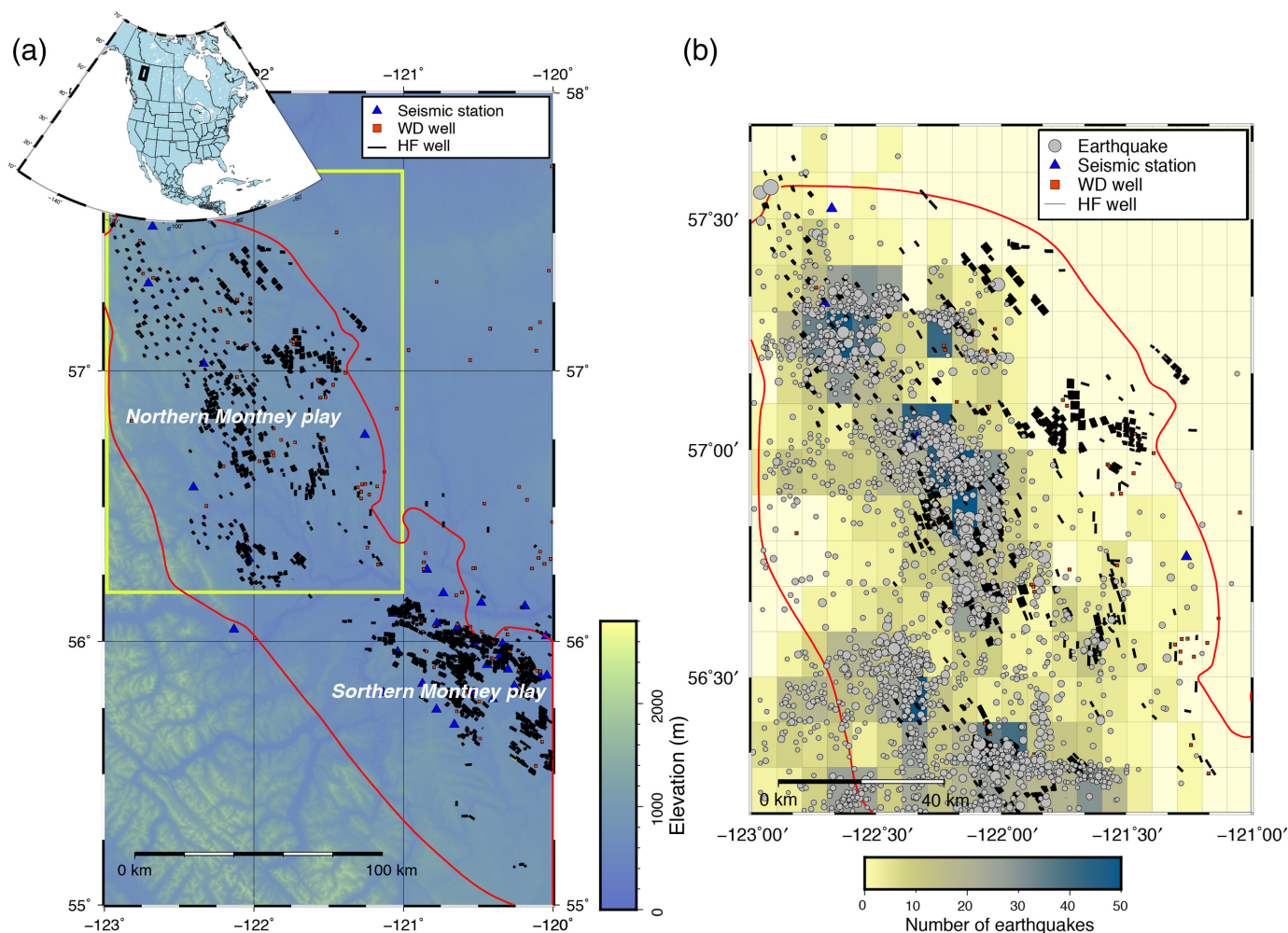*Corresponding author: bei.wang@mail.mcgill.ca

Figure 1. Maps showing the distribution of regional seismicity in northeastern British Columbia, Canada. (a) Injection wells and seismic stations in the northeastern British Columbia (NEBC). The red line marks the outline of the play. The red squares represent wastewater disposal wells. The thin black lines denote hydraulic fracturing wells. The blue triangles are seismic stations. The yellow square marks the study area. The inset shows the study area within North America. (b) Seismicity in the northern Montney play. The gray circles represent earthquake epicenters in our enhanced catalog detected and located by the Source-Scanning Based on Navigated Automatic Phase-Picking method. Earthquakes with **M** >1 are plotted. The background color in each cell represent the number of earthquakes. The color version of this figure is available only in the electronic edition.

Progress has been made recently on searching for the underlying factors controlling the occurrence pattern of IIE. In the WCSB, with the deployment of machine-learning algorithms (MLA), Pawley *et al.* (2018) searched for the potential controlling factors affecting the IIE distribution from tectonic, geomechanical, and hydrological proxies within the Duvernay play of Alberta, with a trained linear regression-MLA (LR-MLA). They extracted the feature importance from the LR-MLA and highlighted multiple factors that could strongly influence the distribution of IIE, such as the proximity to fossil reef margins and proximity to the basement. Wozniakowska and Eaton (2020) adopted a similar LR-MLA to investigate the IIE in NEBC, and they found that the most strongly influential factors in NEBC could be the injection depth and the distance to the Cordilleran foreland thrust and fold belt. In Oklahoma, Hincks *et al.* (2018) deployed a Bayesian network to investigate the relative importance of operational and geologic factors controlling the seismogenesis of wastewater-induced earthquakes, and identified the injection depth above the crystalline basement to be the most important. Meanwhile, Ries *et al.* (2020) concluded that HF wells targeting older formations at deeper depths are responsible for the higher seismicity rate in Oklahoma.

However, previous LR-MLA studies mostly focused on geological factors in the WCSB. Operational parameters, for example, treating pressure and cumulative volume of injected fluid that could also be influential factors to the IIE distribution, are not considered in their models. Moreover, although the feature importance from the LR-MLA could be a useful indicator, how to further quantify the causal relationship between each feature and the seismicity rate in a more meaningful way is not completely resolved yet.

Because the regional geological settings in SMP and NMP are dominated by graben and Cordilleran edge, respectively (Mossop and Shetsen, 1994; Davies et al., 2018), simply combining these two regions in the analysis of potential controlling factors may overlook the underlying geological factors. To address the aforementioned questions, in this study, we take a different approach by focusing solely on NMP. In our model, both the geological and operational factors are considered to investigate how these factors contribute to the occurrence patterns of IIE. Specifically, we first build an enhanced catalog for NMP based on a newly developed earthquake detection method. Then we divide the NMP into 0.1° × 0.1° cells and deploy the eXtreme Gradient Boosting-MLA (XGBoost-MLA) to delineate the controlling factors based on the IIE distribution in each cell. Finally, we analyze the Shapley Additive Explanations (SHAP) values to interpret the results from XGBoost-MLA, and to show both the feature importance and the relationships between each feature and the model output. Overall, we find that operational and geological factors have comparable contributions to the seismicity patterns. The top four factors are the horizontal distance to the Cordilleran belt, cumulative injected volume, shut-in pressure, and vertical distance to the Debolt formation (with respect to the HF depth).

Method

In this study, XGBoost is deployed to investigate the potential controlling factors on the IIE distribution. XGBoost is an ensemble tree-based model introduced to speed up and improve the performance of the traditional boosting techniques (Chen and Guestrin, 2016). One of the most noticeable advantages of XGBoost is the effective performance in memory-limited settings on a single machine, achieved through innovative optimization of system resources and algorithms such as scalable end-to-end tree boosting system, justified weighted quantile sketch, sparsity-aware algorithm, and effective cache-aware block structures (Chen and Guestrin, 2016).

A traditional Decision Tree model is formed with a root node, internal nodes, and leaf nodes. The Decision Tree algorithm start the calculation from the root node, then it will branch out through the internal nodes and finally reach the leave nodes. In contrast, XGBoost uses a sequence of decision trees, but each tree learns from the preceding one to improve the overall performance. For a dataset of n samples with m features, $D = \{(\mathbf{x}_i, y_i)\}(i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, and a tree ensemble model of K regression trees, the model output is predicted as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F}, \tag{1}$$

in which $\mathcal{F} = (f(\mathbf{x}) = \omega_{q(\mathbf{x})})(q : \mathbb{R}^m \to T, \omega \in \mathbb{R}^T)$ is the set of all possible regression trees, T is the number of leaves on a tree, f_k represents an independent decision tree with structure q that maps an instance to the corresponding leaf, and ω is the weight of the leaf. The goal is to find a set of functions by minimizing the following regularized objective function

$$J(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2}$$

in which $\Omega(f) = \gamma T + \frac{1}{2}\|\omega\|^2$ is the regularization term to control the complexity of the model, γ is the regularization on the additional leaf, λ is a tuning parameter determining the quantity of penalty, and l is a training loss function measuring the misfit between the input value, y_i, and the prediction value $\hat{y}_i$. An additive training strategy, a sequential algorithm, is adopted to obtain the optimum model parameters. Starting from a constant prediction and building up an additive model by adding a new function at a time, the objective function at the tth iteration can be written as:

$$J^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t). \tag{3}$$

By taking the Taylor expansion approximation of the loss function to the second order and removing the constant term, the objective function can be extended to

$$J^{(t)} \simeq \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t). \tag{4}$$

in which $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ are the first and second order partial derivatives of the loss function, respectively. For a set of indices assigned to the jth leaf $I_j = \{i | q(\mathbf{x}_i) = j\}$, the optimal weight of the leaf ω_j^*, and the corresponding objective reduction J^* can be calculated as:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{5}$$

and

$$J^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{6}$$

A common approach to learn the optimal tree structure is to start from a single leaf and iteratively add branches to the tree (Chen and Guestrin, 2016). After splitting a leaf node into two leaves, the gain of the score will be:

$$\text{Gain} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma. \tag{7}$$

Model Set-up for XGBoost-MLA
Enhanced catalog for NMP
The main goal of this study is to investigate how different features could influence the IIE distribution. Thus, an enhanced catalog is needed to better depict the spatial variance of seismicity distribution. In NMP, an increasing number of seismic stations are deployed to monitor the surged seismicity related to the development of unconventional resources (Fig. 1), allowing us to detect more earthquakes with new earthquake detection and location techniques. Here, we analyze the continuous waveforms recorded at eight seismic stations of the Canadian National Seismograph Network, including NBC4, NBC5, NBC7, NBC8, and MONT1-4 (Incorporated Research Institutions for Seismology network code PQ and 1E, sampling rate: 100 Hz).

We deploy the recently developed Source-Scanning Based on Navigated Automatic Phase-Picking method (Tan *et al.*, 2019) to search for earthquakes from 2014 to 2021. More than 3000 events are detected and located within our study area (Fig. 1b), about 5 times more than that reported in Natural Resources Canada's routine catalog (~650 events). The magnitude of completeness of our enhanced catalog is estimated at **M** ~1 (Fig. S1, available in the supplemental material to this article). In the following analysis, we only include events with magnitude larger than 1 to avoid any detection bias. The 1D velocity model specifically developed by Babaie Mahani *et al.* (2020) for NMP is adopted to better locate these earthquakes. It is also worth mentioning that we only investigate IIE spatially correlated with HF stimulations.

Previous studies using MLA to delineate controlling factors in the WCSB are often based on individual HF pads (e.g., Pawley *et al.*, 2018; Wozniakowska and Eaton, 2020). In this study, we take a different approach by dividing the NMP region into $0.1° \times 0.1°$ cells. Our grid approach is justified for two main reasons. First, the cumulative effect of injections at multiple nearby pads within the same area cannot be accounted for if the analysis is limited to individual pads. Second, accurate operational parameters may not be always available for all HF wells. A recent study by Hicks *et al.* (2021) also adopted the grid approach to investigate the controlling factors of induced seismicity in the Permian basin, Texas.

Model input
We prepare potential geological and operational factors as the input features to train the XGBoost-MLA model. The results are then used to evaluate the corresponding impacts on the seismicity distribution. Notice that it is unnecessary to normalize the feature inputs as decision trees are not affected by the scaling of data (Chen and Guestrin, 2016).

Geological factors. We consider the distribution of faults as a primary geological factor. We compile all mapped structural lineaments to ensure that every known fault is included in our model (Fig. 2a). The existence of faults could contribute to the occurrence pattern of IIE by playing two different roles, that is, high-permeable pathways and seismogenic structures. For faults acting as high-permeable pathways, they can help transfer the stress via the injected fluid and aseismic slip to a broader and deeper area within a short period of time (Eyre *et al.*, 2019; Peña Castro *et al.*, 2020; Wang *et al.*, 2021; Yu, Harrington, *et al.*, 2021). Similarly, specific geological structures, that is, the fault-related fossil reef margin and karst feature, as suggested by Schultz *et al.* (2016) and Galloway *et al.* (2018), could contribute to the distribution of IIE in Alberta by accelerating the fluid flow to the faults. Faults could also host the seismogenesis of IIE. In theory, more seismogenic structures would imply a higher possibility of having induced earthquakes (Wang, Schmandt, *et al.*, 2020). Therefore, we consider two fault-related proxies (i.e., horizontal distance to the nearest fault from the injection, and the total length of fault traces within one cell) in our analysis of the capacity of inducing IIE (Fig. S2a and S2b, respectively).

The characteristics of local geological units could also be important factors. For instance, features of the Montney formation, its depth and thickness (Fig. 2b), may influence the occurrence pattern of IIE by affecting the fluid migration efficiency (Sanders *et al.*, 2018; Zonneveld and Moslow, 2018; Wang *et al.*, 2021). A recent study based on high-resolution seismicity distribution and seismic reflection results suggests that the Debolt formation below the Montney may play an important role as it may be more seismogenic (Riazi and Eaton, 2020). Moreover, Skoumal *et al.* (2018) suggested that the proximity of the basement may contribute to the likelihood of induced seismicity. Therefore, we incorporate the thickness of the Montney shale, vertical distance to the Debolt formation (with respect to the HF depth; Fig. 2c), and that to the Precambrian basement (with respect to the HF depth; Fig. S2c) into our model inputs.

Another factor in our model is the horizontal distance to the Cordilleran foreland thrust and fold belt (CFTFB; Fig. 2). A recent LR-MLA study shows that this variable is highly correlated with the occurrence of **M** >2.5 earthquakes, although the causal relationship is still unclear (Wozniakowska and Eaton, 2020). Here, we calculate the horizontal distance to the CFTFB from each HF stage within one cell to investigate whether this factor could influence the induced seismicity. A complete list of geological factors (with corresponding values for each cell) is available in Table S1.

Operational factors. As the vast majority of IIE in NMP are related to HF stimulations (Dokht *et al.*, 2021), we focus our analysis to operational parameters submitted by operators in their completion reports to the provincial regulator. Total injection volume, as a widely recognized factor, has been suggested to be associated with seismicity rate in the WCSB. Particularly, Farahbod *et al.* (2015) found that in the Horn river basin, NEBC, IIE started to occur as the monthly HF
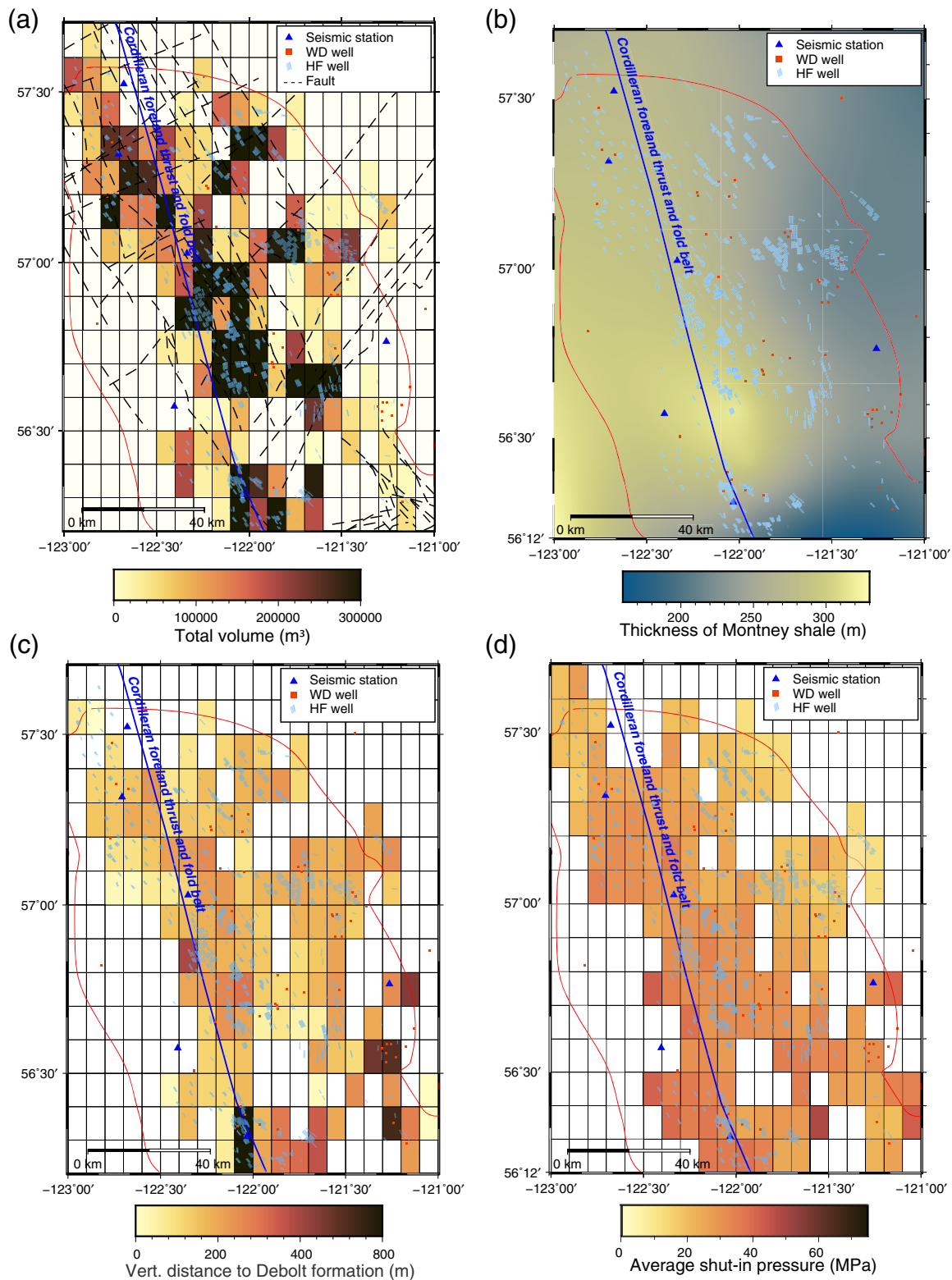
Figure 2. Geological and operational parameters for each 0.1° × 0.1° cell. (a) The total volume of injected fluid within each cell is represented by the background color. The blue line represents the eastern edge of the Cordilleran foreland thrust and fold belt. The black dashed lines represent known faults in our study area. The gray circles mark earthquake epicenters. Short light-blue lines represent hydraulic fracturing (HF) wells, the red squares represent the wastewater disposal (WD) well, and the blue triangles represent seismic stations. (b,c,d) are similar to panel (a), but show the thickness of the Montney shale, vertical distance to the Debolt formation (with respect to the HF depth) and average shut-in pressure, respectively. Cells filled in white have no HF well inside, and are not used in our analysis. The color version of this figure is available only in the electronic edition.

injection volume (summed over the entire basin) exceeding the level of 2.0×10^4 m³. In the Fox Creek area of Alberta, Schultz et al. (2018) reported that seismic productivity could be linearly related to cumulative injected volume when it is on the order of 10^4–10^5 m³ from each injection pad. Therefore, we search for active HF wells in our study area and compile HF injection data to calculate the cumulative volume of fluid injections within each cell, setting it as the primary operational factor (Fig. 2a).

During HF stimulations, multiple treating pressures are reported, including maximum treating pressure, average treating pressure, shut-in pressure, and breakdown pressure. Treating pressures could provide some information on the state of local stress field. For examples, the shut-in pressure could be a good approximation to the horizontal component of the minimum principal stress $S_{H\min}$ (Fig. 2d), whereas in principle, the value of $S_{H\max}$ (the horizontal component of the maximum principal stress) could be derived from the formation breakdown pressure (Breckels and van Eekelen, 1982). To investigate if these pressure conditions could influence the IIE occurrence pattern, we calculate the average treating pressure values for each cell and incorporate them into the model inputs. A complete list of operational factors and the values in each cell are available in Table S1. The spatial distribution of the values of the other five operational factors is shown in Figure S3.

Model fitting and interpretation

Previous studies searching for controlling factors based on MLA often adopted the classification models (e.g., nonseismogenic wells vs. seismogenic wells; Pawley et al., 2018). In our study, the seismicity rate within each cell varies significantly (Fig. 1), a regression model is thus more suitable to evaluate the factors. We deploy a regression-based objective function that minimizes the squared loss between predicted values and ground-truth values. We divide the dataset into one training set (80%), and one test set (20%), and the parameters used for the XGBoost model are listed in Table S2. The model is scored with the coefficient of determination R^2 of the prediction, defined as $(1 - \frac{u}{v})$, in which u is the sum of residual squares, $\Sigma(y_{\text{true}} - y_{\text{pred}})^2$, v is the sum of data variations squares, $\Sigma(y_{\text{true}} - \bar{y}_{\text{true}})^2$, and the best possible score is equal to 1.0.

One drawback of the XGBoost algorithm (similar to Random Forest and Neural networks; Hopfield, 1982; Tin Kam, 1998; Chen and Guestrin, 2016) is that the model is usually a "black-box," which makes it less interpretable and cannot inform how the input variables contribute to the final models. To deal with this limitation, we adopt the SHAP approach to convert the results from XGBoost into values that can be quantitatively interpreted. SHAP is developed to estimate each feature's contribution to the model outputs, based on the coalitional game theory and local explanations (Shapley, 1953; Lundberg and Lee, 2017). If a model N has n features, the contribution ($\varnothing$) of the ith

feature to the output O(N) is calculated based on its marginal contribution using the equation below,

$$\varnothing_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(s)]. \tag{8}$$

On the basis of the following additive feature attribution method, a parameter g, which is a linear function of binary feature, can be calculated via,

$$g(z') = \varnothing_0 + \sum_{i=1}^{M} \varnothing_i z_i', \tag{9}$$

in which $z' \in \{0, 1\}^M$. The value of z' is one when the feature is observed, otherwise $z' = 0$. M represents the number of input features (Lundberg and Lee, 2017).
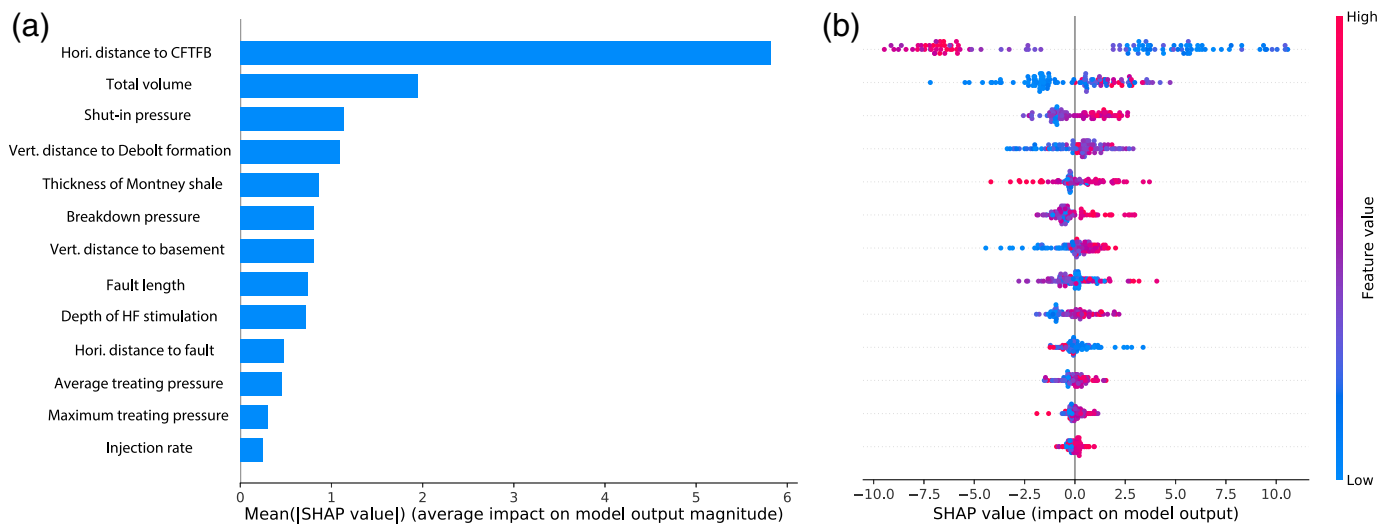
Result and discussion

The R^2 score of our XGBoost model is ~0.79, suggesting that our regression model is an adequate interpretation of how these geological and operational factors contribute to the IIE distribution (Draper and Smith, 1998). Figure S4 shows the predicted and observed numbers of earthquakes for the cells in the test set.

The respective contribution of each controlling factor obtained from SHAP is listed in descending order in Figure 3a. The top four features that contribute most to the seismicity pattern are horizontal distance to the CFTFB, cumulative injected volume, shut-in pressure, and vertical distance to the Debolt formation. Features with secondary influence are the thickness of the Montney formation, breakdown pressure, vertical distance to the basement, and cumulative fault length per unit area. The remaining features appear the least related, including average HF stimulation depth, average horizontal distance to the nearest faults, average treating pressure, maximum treating pressure, and injection rate.

The feature importance from Figure 3a is essential, but the correlation between the IIE distribution and individual factors cannot be quantitatively depicted. To better visualize the relationships, we summarize the distribution of SHAP values in Figure 3b to demonstrate how each factor contributes to the model output. Each sample on the plot corresponds to the SHAP value of a given factor within a cell. Generally, positive and negative SHAP values mean that the factor would lead to more and less IIE in a cell than the average number of IIE of all cells, respectively, and the larger the absolute SHAP value, the higher influence.

Figure 4 and Figure S5 show the SHAP values for each controlling factor. According to our results, horizontal distance to the CFTFB is the most important factor influencing the seismicity rate in NMP, that is, a closer distance to the CFTFB corresponds to a higher positive SHAP value, thus a higher seismicity rate (Fig. 4a). This observation is consistent with

Figure 3. Shapley Additive Explanations (SHAP) values of various geological and operational factors. (a) Mean SHAP values of the eXtreme Gradient Boosting (XGBoost) model, representing the relative importance to the model output. The features are ordered based on their importance. (b) Detailed distribution of SHAP values of all cells. Each dot represents one specific feature of a cell, and is colored by the feature's SHAP value. A higher positive SHAP value means this factor would cause more injection-induced earthquakes (IIE) in a cell than the average number of IIE of all cells. The color version of this figure is available only in the electronic edition.

previous studies (Wozniakowska and Eaton, 2020). Although the corresponding physical mechanisms are not clear yet, several tectonic facts may help explain such correlation. First, the tectonic strain rate in the WCSB decreases monotonically with the horizontal distance from the CFTFB. Kao *et al.* (2018) and Dokht *et al.* (2021) reported that the tectonic strain rate is positively correlated with the distribution of IIE. Second, thrust faults are commonly observed along the eastern flank of the southeastern Canadian Cordillera (Price, 1986). It is well established in the literature that, given the same stress state, thrust faults are more prone to seismic failure than strike-slip faults (e.g., Wang *et al.*, 1995; Collettini *et al.*, 2019). Therefore, fluid injections could load the pre-existing thrust fault systems more efficiently to increase the seismogenic capacity. Moreover, when the horizontal distance to the CFTFB exceeds ~10 km, the likelihood of inducing IIE drops sharply (Fig. 4), which could suggest a decreasing density of faults away from the CFTFB.

The total cumulative volume is ranked as the second important controlling factor (Fig. 3). As shown in Figure 4b, the SHAP value is negative when the cumulative volume is less than ~120,000 m³, suggesting that fewer earthquakes (with **M** >1) are expected if the cumulative volume is less than this threshold. Once above, the number of earthquakes surges significantly. However, the relationship between the cumulative volume and the number of earthquakes cannot be simply interpreted from the SHAP values. For example, the SHAP values of the two cells with the largest total volume are much smaller than cells with less total volume (Fig. 4b). Nevertheless, it is not sufficient, at least based on this study, to dismiss the linear correlation, given that the cumulative volume is not the only factor controlling the seismicity rate. Other factors, such as specific geological structures (Schultz *et al.*, 2016; Galloway *et al.*, 2018), may overcome the effects from the cumulative volume, thus affecting the IIE pattern. For example, cells in the northeastern part of NMP are recognized with a considerable amount of injected fluid, whereas the distance to the CFTFB is significantly larger, which may lead to a decreased seismogenic potential and, overall, very few earthquakes were induced.

Findings that differ from previous studies may provide more insights to the previously overlooked controlling factors. Wozniakowska and Eaton (2020) suggest that the vertical distance to the Debolt formation may not be an important factor in inducing **M** >2.5 events; however, in our case, it is one of the top factors affecting the IIE distribution. As shown in Figure 4c, the occurrence of IIE is actually discouraged at places within ~200 m from the Debolt formation. The seismicity rate increases with the vertical distance once it exceeds the ~200 m threshold. In contrast, the vertical distance to the basement seems to be a secondary factor influencing the seismicity pattern (ranked 7th). This is somewhat surprising as in previous studies, the vertical distance to the basement is often conjectured to be an indicative feature that facilitates hydraulic communications with deeper faults (Pawley *et al.*, 2018; Wozniakowska and Eaton, 2020; Yu *et al.*, 2020; Wang *et al.*, 2021), rather than the vertical distance to the Debolt formation. For example, Amini *et al.* (2021) suggest that vertical distance to the Precambrian basement is one of the most important factors for NEBC, correlating negatively with the increasing likelihood of IIE. However, such negative correlation is not observed in our study (Fig. S5). One possible
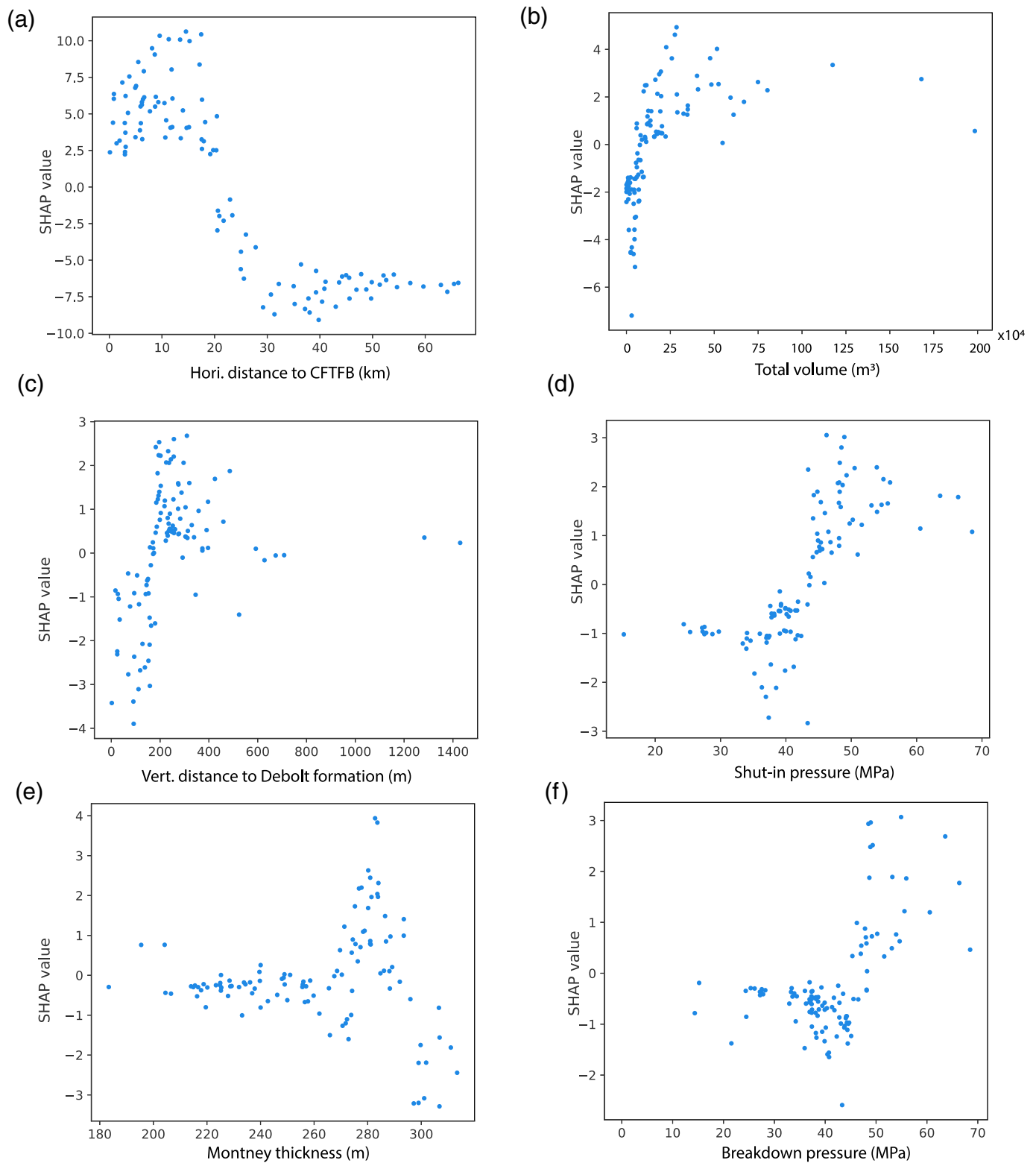
Figure 4. Distribution of SHAP values with respect to six different factors. (a) Horizontal distance to the Cordilleran foreland thrust and fold belt (CFTFB). Each dot represents one cell and the *x* axis shows the corresponding feature value. (b) Total volume of injected fluid. (c) Vertical distance to the Debolt formation (with respect to the HF depth). (d) Shut-in pressure. (e) Thickness of the Montney formation. (f) Breakdown pressure. The color version of this figure is available only in the electronic edition.

reason is that their model incorporates data from both SMP and NMP, despite the significant discrepancies in the geological settings between these two regions. A recent study based on high-resolution 3D seismic imaging of subsurface structures documents a positive correlation between the increased seismicity rate and a buried thrust fault system within the Debolt formation (Riazi and Eaton, 2020). This new finding suggests that the Debolt formation may be comparable to the basement in causing IIE, and provides some explanation to our results.

Moreover, our results show that the shut-in pressure and the breakdown pressure are also important factors. As shown in Figure 4, the seismicity rate starts to surge when the shut-in pressure increases from ~45 to ~55 MPa. Similarly, the SHAP value is below 0 when the breakdown pressure is smaller than ~47 MPa. Thus, the contributions from treating pressures, although their effects have not been systematically investigated in the past, should not be overlooked. To explain the seismic responsiveness from treating pressures, one possible scenario is that they could be related to the specific geomechanical condition within the local geological setting. In other words, the range of treating pressures simply reflects the stress required to complete the HF stimulations to a given formation at a given depth. As shown in Figures 2 and 4, cells with higher seismicity rates are generally distributed along the CFTFB. Most of these cells have similar vertical distances to the Debolt formation, similar thickness of the Montney formation, and a comparable background stress regime of $S_{H \max}$ and $S_{H \min}$. Therefore, the observed correlation between a higher seismicity rate and the specific range of treating pressures could imply that certain geological settings are particularly in favor of IIE. It is worth mentioning that only the shut-in pressure is reported by the operators, which could be a proxy to instantaneous shut-in pressure. However, because their relationship cannot be verified, we will not make further inference here.

The spatial distribution and structural geometry of faults are closely linked to the occurrence of IIE (Bao and Eaton, 2016; Galloway et al., 2018; Yu et al., 2019; Schultz et al., 2020; Yu, Kao, et al., 2021; Gao et al., 2022). Therefore, it is not surprising to find out that both the cumulative length of faults and the average distance to the nearest faults from HF stimulations appear to be significant contributors. However, in our model, their contributions (mean SHAP values of ~0.6 and ~0.4 for the former and the latter, respectively) are much less compared to the top two factors (horizontal distance to the CFTFB and total volume of injected fluid with the corresponding mean SHAP values of ~5.7 and ~2.1). One possible explanation is that the fault distribution adopted in our model is incomplete because of many hidden or blind faults missing from the publicly available dataset (Riazi and Eaton, 2020). As a result, it is possible that the contribution from fault-related factors may be somewhat underestimated in our study. More detailed mapping of regional fault systems in NMP would help better define the

relationship between fault-related parameters and the occurrence pattern of IIE.

Despite our best effort, we acknowledge that the 13 factors considered in our model cannot be a complete set controlling the occurrence pattern of IIE. Other possible factors include the stress state of faults (Wang et al., 2018), overpressure of shale formations (Schultz and Eaton, 2018), and the presence of specific geological units facilitating fluid migration (Galloway et al., 2018). At the time of this study, however, we are unable to collect sufficient amount of data to incorporate these factors into our analysis. We expect to expand our collection of features in the future as the quality and quantity of datasets improve.

Generally speaking, collinearity between factors could influence the robustness of models, especially for LR-MLA. However, one key feature of Decision Tree algorithms (including XGBoost) is that they are naturally immune to multicollinearity (Kotsiantis, 2013). To verify the degree of collinearity of our model's input features, we calculate their correlation matrix and the result is shown in Figure S7. As expected, we obtain a high value of cross-correlation coefficient (>0.8) between the maximum treating pressure and average treating pressure. A strong negative correlation (−0.79) is also found between the horizontal distance to the CFTFB and the thickness of the Montney formation, which is reasonable as the Montney shale becomes thinner towards the east, away from the CFTFB (Fig. 2; Fig. S7). No other pairs appear to be highly correlated.

We do not attempt to remove aftershocks from our earthquake catalog prior to the statistical analysis, a process known as declustering (Gardner and Knopoff, 1974), for two practical reasons. First, most earthquakes in NMP are small. During the ~7 yr of our study period, only five **M** >3 events occurred. The lack of big mainshocks implies insignificant/negligible after-[1] shocks (Fig. S8 shows the history of seismicity for cells with **M** >3 events). Second, aftershocks can also be regarded as IIE in our statistical analysis because they can be caused by the combined stress perturbation from fluid injections and nearby earthquakes. Therefore, removing aftershocks from the catalog may actually introduce statistical bias to our analysis.

Finally, our results provide critical information to establishing a comprehensive susceptibility model that includes key geological and operational factors affecting the IIE distribution in the NMP area. This inclusive approach will give regulatory agencies and the energy industry a more effective tool to mitigate the seismic hazard due to IIE.

Conclusion

Distribution of IIE can be related to both natural and operational factors. However, the relative importance of operational factors with respect to the natural ones has not been systematically investigated for the western Canadian sedimentary basin. In this study, we applied a supervised machine-learning algorithm, namely, XGBoost, to systematically evaluate the

contributions of various geological and operational factors on the distribution of IIE in the northern Montney play of northeastern British Columbia, Canada.

We began with the establishment of a much-enhanced earthquake catalog, to better depict the spatiotemporal distribution of the regional seismicity in the study area. Then with XGBoost, we set six geological and seven operational factors as input parameters and trained a regression model to investigate their contributions to the observed seismicity distribution. Next, we adopted the SHAP approach to convert the results from XGBoost into values that can be quantitatively interpreted. Overall, the performance of our XGBoost model is very good with an R^2 score of ~0.79.

Our study reveals that horizontal distance to the Cordilleran deformation front, total injected volume, shut-pressure and vertical distance to the Debolt formation (with respect to the HF depth) are the most significant factors contributing to the IIE distribution. Treating pressures (e.g., shut-in pressure and breakdown pressure) could also contribute to the IIE distribution. Vertical distance to the Precambrian basement (with respect to the HF depth) seems to be a secondary influencing factor, differing from previously studies. One possible explanation is that previous studies simply combined NMP and SMP into the model simultaneously, which may overlook the underlying geological factors. The IIE distribution is less sensitive to fault-related parameters (distance to the nearest faults and the cumulative length of fault traces per unit area). This somewhat surprising result can be an artifact due to the limited knowledge of hidden or blind faults in our study area. Overall, our results provide critical information to establishing a comprehensive susceptibility model that includes both geological and operational factors affecting the IIE distribution in the NMP area. Our models can guide regulatory agencies and the energy industry to effectively mitigate the seismic hazard due to IIE.

Data and Resources

The seismic data can be downloaded from the Incorporated Research Institutions for Seismology Data Management Center (IRIS DMC; https://ds.iris.edu/ds/nodes/dmc/) for seismic stations of the Canadian National Seismograph Network, including NBC4, NBC5, NBC7, NBC8, and MONT1-4 (IRIS network code PQ and 1E, sampling rate: 100 Hz). Earthquake catalog is available as part of the Geological Survey of Canada Open File reports at Natural Resources Canada's (NRCan's) GEOSCAN database (https://geoscan.nrcan.gc.ca) and available at https://doi.org/10.5281/zenodo.6378188. The operational parameters can be accessed at https://www.bcogc.ca/data-reports/data-centre/. The surface fault traces are published by Hayes et al. (2021). Other geological information could be found at https://ags.aer.ca/. XGBoost package can be downloaded at https://xgboost.readthedocs.io/en/stable/. SHAP is available at https://shap.readthedocs.io/en/latest/index.html. The supplemental material to this article includes eight supplemental figures (catalog completeness, map view of three geological factors, map view of five operational factors, comparison between predicted and observed number of injection-induced earthquakes (IIE) for the test dataset, distribution of Shapley Additive Explanations (SHAP) values of various controlling factors, results of feature importance with different bin sizes, cross-correlation values between these factors, temporal distribution of earthquakes for the five cells with **M** >3 events, and two tables (the input values of geological and operational factors of each cell to the eXtreme Gradient Boosting [XGBoost] model, parameters of XGBoost model). All websites were last accessed in January 2022.

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.

References

Amini, A., A. Mehrabifard, and E. Eberhardy (2021). Development of an induced seismicity susceptibility framework and map for NEBC using an integrated machine learning and mechanistic validation approach, *Geoscience BC Rept. 2021-11*, 117.

Atkinson, G. M., D. W. Eaton, H. Ghofrani, D. Walker, B. Cheadle, R. Schultz, R. Shcherbakov, K. Tiampo, J. Gu, R. M. Harrington, *et al.* (2016). Hydraulic fracturing and seismicity in the western Canada sedimentary basin, *Seismol. Res. Lett.* **87,** no. 3, 631–647.

Babaie Mahani, A., D. Malytskyy, R. Visser, M. Hayes, M. Gaucher, and H. Kao (2020). Well-log-based velocity and density models for the Montney unconventional resource play in northeast British Columbia, Canada, applicable to induced seismicity monitoring and research, *Seismol. Res. Lett.* **92,** no. 2A, 886–894, doi: 10.1785/0220200213.

Bao, X., and D. W. Eaton (2016). Fault activation by hydraulic fracturing in western Canada, *Science* **354,** no. 6318, 1406–1409.

Breckels, I. M., and H. A. M. van Eekelen (1982). Relationship between horizontal stress and depth in sedimentary basin, *J. Petrol. Technol.* **34,** no. 09, 2191–2199, doi: 10.2118/10336-PA.

Chen, T., and C. Guestrin (2016). XGBoost: A scalable tree boosting system, *Paper presented at the Proc. of the 22nd ACM SIGKDD International Conf. Knowledge Discovery and Data Mining*, San Francisco, California, U.S.A., 13–17 August 2016, doi: 10.1145/2939672.2939785.

Collettini, C., T. Tesei, M. M. Scuderi, B. M. Carpenter, and C. Viti (2019). Beyond Byerlee friction, weak faults and implications for slip behavior, *Earth Planet. Sci. Lett.* **519,** 245–263.

Davies, G. R., T. F. Moslow, and M. D. Sherwin (1997). The lower Triassic Montney formation, west-central Alberta, *Bull. Can. Petrol. Geol.* **45,** no. 4, 474–505, doi: 10.35767/gscpgbull.45.4.474.

Davies, G. R., N. Watson, T. F. Moslow, and J. A. MacEachern (2018). Regional subdivisions, sequences, correlations and facies relationships of the lower Triassic Montney formation, west-central Alberta to northeastern British Columbia, Canada—With emphasis on role of paleostructure, *Bull. Can. Petrol. Geol.* **66,** no. 1, 23–92.

Dokht, R. M. H., H. Kao, A. Babaie Mahani, and R. Visser (2021). Spatiotemporal analysis of seismotectonic state of injection-induced seismicity clusters in the western Canada sedimentary basin, *J. Geophys. Res.* **126,** no. 4, e2020JB021362, doi: 10.1029/2020JB021362.

Draper, N. R., and H. Smith (1998). *Applied Regression Analysis*, Vol. 326, John Wiley and Sons, Hoboken, New Jersey.

Eyre, T. S., D. W. Eaton, D. I. Garagash, M. Zecevic, M. Venieri, R. Weir, and D. C. Lawton (2019). The role of a seismic slip in hydraulic fracturing-induced seismicity, *Sci. Adv.* **5,** no. 8, eaav7172, doi: 10.1126/sciadv.aav7172.

Farahbod, A. M., H. Kao, D. M. Walker, J. F. Cassidy, and A. Calvert (2015). Investigation of regional seismicity before and after hydraulic fracturing in the Horn river basin, northeast British Columbia, *Can. J. Earth Sci.* **52,** no. 2, 112–122.

Galloway, E., T. Hauck, H. Corlett, D. Panǎ, and R. Schultz (2018). Faults and associated karst collapse suggest conduits for fluid flow that influence hydraulic fracturing-induced seismicity, *Proc. Natl. Acad. Sci.* **115,** no. 43, E10003–E10012.

Gao, D., H. Kao, B. Wang, R. Visser, R. Schultz, and R. M. Harrington (2022). Complex 3D migration and delayed triggering of hydraulic fracturing-induced seismicity: A case study near Fox Creek, Alberta, *Geophys. Res. Lett.* **49,** no. 2, e2021GL093979, doi: 10.1029/2021GL093979.

Gardner, J., and L. Knopoff (1974). Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian? *Bull. Seismol. Soc. Am.* **64,** no. 5, 1363–1367.

Hayes, B. J., J. H. Anderson, M. Cooper, P. J. McLellan, B. Rostron, and J. Clarke (2021). Wastewater disposal in the maturing Montney play fairway, northeastern British Columbia (NTS 093P, 094A, B, G, H), *Geoscience BC Summary of Activities 2020: Energy and Water, Geoscience BC, Rept. 2021-02*, 91–102.

Hicks, S. P., S. Goes, A. C. Whittaker, and P. J. Stafford (2021). Multivariate statistical appraisal of regional susceptibility to induced seismicity: Application to the Permian basin, SW United States, *J. Geophys. Res.* **126,** no. 12, e2021JB022768, doi: 10.1029/2021JB022768.

Hincks, T., W. Aspinall, R. Cooke, and T. Gernon (2018). Oklahoma's induced seismicity strongly linked to wastewater injection depth, *Science* **359,** no. 6381, 1251–1255.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci.* **79,** no. 8, 2554–2558.

Kao, H., R. Hyndman, Y. Jiang, R. Visser, B. Smith, A. Babaie Mahani, L. Leonard, H. Ghofrani, and J. He (2018). Induced seismicity in western Canada linked to tectonic strain rate: Implications for regional seismic hazard, *Geophys. Res. Lett.* **45,** no. 20, 11,104–11,115.

Kotsiantis, S. B. (2013). Decision trees: A recent overview, *Artif. Intell. Rev.* **39,** no. 4, 261–283, doi: 10.1007/s10462-011-9272-4.

Lundberg, S. M., and S.-I. Lee (2017). A unified approach to interpreting model predictions, *Paper presented at the Proc. of the 31st International Conf. Neural information processing systems*, Long Beach, California, 4–9 December 2017.

Mahani, A. B., R. Schultz, H. Kao, D. Walker, J. Johnson, and C. Salas (2017). Fluid injection and seismic activity in the northern Montney play, British Columbia, Canada, with special reference to the 17 August 2015 M w 4.6 induced earthquake, *Bull. Seismol. Soc. Am.* **107,** no. 2, 542–552.

Mossop, G. D., and I. Shetsen (1994). *Geological Atlas of the Western Canada Sedimentary Basin*, Canadian Society of Petroleum Geologists, available at https://ags.aer.ca/reports/atlas-western-canada-sedimentary-basin (last accessed February 2022).

Pawley, S., R. Schultz, T. Playter, H. Corlett, T. Shipman, S. Lyster, and T. Hauck (2018). The geological susceptibility of induced earthquakes in the Duvernay play, *Geophys. Res. Lett.* **45,** no. 4, 1786–1793.

Peña Castro, A. F., M. P. Roth, A. Verdecchia, J. Onwuemeka, Y. Liu, R. M. Harrington, Y. Zhang, and H. Kao (2020). Stress chatter via fluid flow and fault slip in a hydraulic fracturing-induced earthquake sequence in the Montney formation, British Columbia, *Geophys. Res. Lett.* **47,** no. 14, e2020GL087254, doi: 10.1029/2020GL087254.

Price, R. A. (1986). The southeastern Canadian Cordillera: Thrust faulting, tectonic wedging, and delamination of the lithosphere, *J. Struct. Geol.* **8,** no. 3, 239–254.

Riazi, N., and D. W. Eaton (2020). Anatomy of a buried thrust belt activated during hydraulic fracturing, *Tectonophysics* **795,** no. 20, 228640, doi: 10.1016/j.tecto.2020.228640.

Ries, R., M. R. Brudzinski, R. J. Skoumal, and B. S. Currie (2020). Factors influencing the probability of hydraulic fracturing-induced seismicity in Oklahoma, *Bull. Seismol. Soc. Am.* **110,** no. 5, 2272–2282, doi: 10.1785/0120200105.

Sanders, S., C. Etienne, A. Gegolick, D. Kelly, and J.-P. Zonneveld (2018). The middle Montney Altares member: Lithology, depositional setting and significance for horizontal drilling and completion in the Altares field, British Columbia, *Bull. Can. Petrol. Geol.* **66,** no. 1, 318–337.

Schultz, R., and D. W. Eaton (2018). Increased likelihood of induced seismicity in highly over pressured shale formations, *Geophys. J. Int.* **214,** no. 1, 751–757, doi: 10.1093/gji/ggy167.

Schultz, R., G. Atkinson, D. Eaton, Y. Gu, and H. Kao (2018). Hydraulic fracturing volume is associated with induced earthquake productivity in the Duvernay play, *Science* **359,** no. 6373, 304–308.

Schultz, R., H. Corlett, K. Haug, K. Kocon, K. MacCormack, V. Stern, and T. Shipman (2016). Linking fossil reefs with earthquakes: Geologic insight to where induced seismicity occurs in Alberta, *Geophys. Res. Lett.* **43,** no. 6, 2534–2542, doi: 10.1002/2015GL067514.

Schultz, R., R. J. Skoumal, M. R. Brudzinski, D. Eaton, B. Baptie, and W. Ellsworth (2020). Hydraulic fracturing-induced seismicity, *Rev. Geophys.* **58,** no. 3, e2019RG000695, doi: 10.1029/2019RG000695.

Shapley, L. S. (1953). Stochastic games, *Proc. Natl. Acad. Sci.* **39,** no. 10, 1095–1100.

Skoumal, R. J., M. R. Brudzinski, and B. S. Currie (2018). Proximity of Precambrian basement affects the likelihood of induced seismicity in the Appalachian, Illinois, and Williston basins, central and eastern United States, *Geosphere* **14,** no. 3, 1365–1379.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. [2] Salakhutdinov (2014). Dropout: A simple way to prevent neural

networks from overfitting, *J. Machine Learn. Res.* **15,** no. 1, 1929–1958.

Tan, F., H. Kao, E. Nissen, and D. Eaton (2019). Seismicity-scanning based on navigated automatic phase-picking, *J. Geophys. Res.* **124,** no. 4, 3802–3818.

Tin Kam, H. (1998). The random subspace method for constructing decision forests, *IEEE Trans. Pattern Anal. Machine Intell.* **20,** no. 8, 832–844.

Wang, B., R. M. Harrington, Y. Liu, H. Kao, and H. Yu (2018). Remote dynamic triggering of earthquakes in three unconventional Canadian hydrocarbon regions based on a multiple-station matched-filter approach, *Bull. Seismol. Soc. Am.* **109,** no. 1, 372–386, doi: 10.1785/0120180164.

Wang, B., R. M. Harrington, Y. Liu, H. Kao, and H. Yu (2020). A study on the largest hydraulic-fracturing-induced earthquake in Canada: Observations and static stress-drop estimation, *Bull. Seismol. Soc. Am.* **110,** no. 5, 2283–2294, doi: 10.1785/0120190261.

Wang, B., A. Verdecchia, H. Kao, R. M. Harrington, Y. Liu, and H. Yu (2021). A study on the largest hydraulic fracturing induced earthquake in Canada: Numerical modeling and triggering mechanism, *Bull. Seismol. Soc. Am.* doi: 10.1785/0120200251.

Wang, K., T. Mulder, G. C. Rogers, and R. D. Hyndman (1995). Case for very low coupling stress on the Cascadia subduction fault, *J. Geophys. Res.* **100,** no. B7, 12,907–12,918.

Wang, R., B. Schmandt, M. Zhang, M. Glasgow, E. Kiser, S. Rysanek, and R. Stairs (2020). Injection-induced earthquakes on complex fault zones of the Raton basin illuminated by machine-learning phase picker and dense nodal array, *Geophys. Res. Lett.* **47,** no. 14, e2020GL088168, doi: 10.1029/2020GL088168.

Wozniakowska, P., and D. W. Eaton (2020). Machine learning-based analysis of geological susceptibility to induced seismicity in the Montney formation, Canada, *Geophys. Res. Lett.* **47,** no. 22, e2020GL089651, doi: 10.1029/2020GL089651.

Yu, H., R. M. Harrington, H. Kao, Y. Liu, R. E. Abercrombie, and B. Wang (2020). Well proximity governing stress drop variation and seismic attenuation associated with hydraulic fracturing induced earthquakes, *J. Geophys. Res.* **125,** no. 9, e2020JB020103, doi: 10.1029/2020JB020103.

Yu, H., R. M. Harrington, H. Kao, Y. Liu, and B. Wang (2021). Fluid-injection-induced earthquakes characterized by hybrid-frequency waveforms manifest the transition from aseismic to seismic slip, *Nat. Commun.* **12,** no. 1, doi: 10.1038/s41467-021-26961-x.

Yu, H., R. M. Harrington, Y. Liu, and B. Wang (2019). Induced seismicity driven by fluid diffusion revealed by a near-field hydraulic stimulation monitoring array in the Montney basin, British Columbia, *J. Geophys. Res.* **124,** no. 5, 4694–4709.

Yu, H., H. Kao, R. Visser, and B. Wang (2021). From seismic quiescence to surged activity after decades of wastewater disposal: A case study in central-west Alberta, Canada, *Geophys. Res. Lett.* **48,** no. 22, e2021GL095074, doi: 10.1029/2021GL095074.

Zonneveld, J.-P., and T. F. Moslow (2018). Palaeogeographic setting, lithostratigraphy, and sedimentary framework of the lower Triassic Montney formation of western Alberta and northeastern British Columbia, *Bull. Can. Petrol. Geol.* **66,** no. 1, 93–127.