Detecting Anomalous Methane in Groundwater within Hydrocarbon Production Areas across the United States

Tao Wen[1]*, Mengqi Liu[2], Josh Woda[3], Guanjie Zheng[2], Susan L. Brantley[3,4]

[1]Department of Earth and Environmental Sciences, Syracuse University, Syracuse, New York 13244, United States
[2]College of Information Sciences and Technology, Pennsylvania State University, University Park, Pennsylvania 16802, United States
[3]Department of Geosciences, Pennsylvania State University, University Park, Pennsylvania 16802, United States
[4]Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania 16802, United States

*Corresponding author: Department of Earth and Environmental Sciences, Syracuse University, Syracuse, New York 13244, United States
Phone: 734-730-8814; E-mail: twen08@syr.edu

Abstract

Numerous geochemical approaches have been proposed to ascertain if methane concentrations in groundwater, [CH_4], are anomalous, i.e., migrated from hydrocarbon production wells, rather than derived from natural sources. We propose a machine-learning model to consider alkalinity, Ca, Mg, Na, Ba, Fe, Mn, Cl, sulfate, TDS, specific conductance, pH, temperature, and turbidity holistically together. The model, an ensemble of sub-models targeting one parameter pair per sub-model, was trained on groundwater chemistry from Pennsylvania (*n*=19,086) and a set of 16 analyses from putatively contaminated groundwater. For cases where [CH_4] > 10 mg/L, salinity- and redox-related parameters sometimes show that CH_4 may have moved into the aquifer recently and separately from natural brine migration, i.e., anomalous CH_4. We applied the model to held-out data for Pennsylvania (*n*=4,786) and groundwater data from three other gas-producing states: New York (*n*=203), Texas (*n*=688), and Colorado (*n*=10,258). The applications show that 1.4%, 1.3%, 0%, and 0.9% of tested samples in these four states, respectively, have high [CH_4] and are >50% likely to have been impacted by gas migrated from exploited reservoirs. If our approach is indeed successful in flagging anomalous CH_4, we conclude that: i) the frequency of anomalous CH_4 (# flagged water samples / total samples tested) in the Appalachian Basin is similar in areas where gas wells target unconventional as compared to conventional reservoirs, and ii) the frequency of anomalous CH_4 in Pennsylvania is higher than in Texas + Colorado. We cannot, however, exclude the possibility that differences among regions might be affected by differences in data volumes. Machine learning models will become increasingly useful in informing decision-making for shale-gas development.

Keywords

Shale gas; machine learning; groundwater; methane, redox, salinity

Introduction

If natural gas production in the United States (US) doubles from 2000 to 2050 as expected, shale gas production will account for more than three-quarters of this natural gas production (U.S. Energy Information Administration, 2018). This development is largely driven by the combined usage of high-volume hydraulic fracturing and horizontal drilling (U.S. Energy Information Administration, 2018). Other countries and regions (e.g., China, Canada, Middle East) are also ramping up their natural gas production from newly discovered technically recoverable shale reserves (U.S. Energy Information Administration, 2017). In the meantime, putative incidents of deterioration of water quality caused by shale gas development have been investigated throughout the US (Brantley et al., 2014; Darrah et al., 2014; Guo et al., 2014; Nicot and Scanlon, 2012; Vidic et al., 2013; Wen et al., 2019b, 2019a; Woda et al., 2018; Yang et al., 2013), where it is not rare to see that hydrocarbon wells show signs of compromised well integrity (Lackey et al., 2021). Proving that contamination was caused by shale gas development activities is always difficult, and estimating the frequency of problems is even more problematic. However, for the public to make decisions about granting social license for shale gas development requires some information about the frequency of problems. In this paper we explore the use of a machine learning model on large groundwater datasets from one region to see if such data-driven models can be used in other regions to assess incidents of contamination.

We developed the model by application to data from Pennsylvania (PA), the state in the US with the longest history of conventional oil/gas (OG) drilling and coal mining. PA has become a leading area for shale gas production through use of horizontal drilling and high-volume hydraulic fracturing to extract gas and liquids from the Marcellus and Utica shales. The contaminant that has been identified as the most common problem in PA is methane (CH_4), the chief component of natural gas (Brantley et al., 2014). Many studies investigating groundwater quality in shale gas basins have focused on contamination by such anomalous CH_4 (Hammond, 2016; Hammond et al., 2020; Li and Carlson, 2014; Molofsky et al., 2013; Nicot et al., 2017c; Osborn et al., 2011; Sherwood et al., 2016; Siegel et al., 2015a; Wen et al., 2017, 2016). Here 'anomalous CH_4' is used to denote dissolved or free-phase CH_4 found in groundwater that putatively derived from the effects of recent human activities related to shale gas development. In general, naturally occurring CH_4 may have long affected an aquifer whereas anomalous CH_4 represents a new source at the impacted site.

The identification of new (anomalous) CH_4 is difficult at least partly because naturally occurring CH_4 is common in many basins, including in PA groundwater (Baldassare et al., 2014; Siegel et al., 2015b; Wen et al., 2019b). In a limited number of areas, the causes of elevated CH_4 in groundwater have been attributed to OG production activities (conventional or unconventional) (Darrah et al., 2014; Grieve et al., 2018; Hammond, 2016; Hammond et al., 2020; Jackson et al., 2013; Osborn et al., 2011; U.S. Environmental Protection Agency, 2015; Warner et al., 2012; Wen et al., 2019b; Woda et al., 2018). Identification of anomalous versus natural CH_4 in groundwater often relies on measurement of carbon isotopes in the gas (e.g., Baldassare et al., 2014; Jackson et al., 2013), but this is somewhat expensive and seldomly definitive. Researchers have also sometimes used other hydrogeochemical patterns to identify anomalous CH_4. For example, the migration of natural CH_4 from deeper formations into shallow groundwater in the Appalachian Basin is often accompanied by salt-containing waters, which leads to slightly elevated salt contents in the shallow aquifer and distinctive chemical signatures (Siegel et al., 2015a, 2015b; Warner et al., 2012; Wen et al., 2019b). Therefore, major and trace elements (e.g., calcium, sodium, and bromide) (Brantley et al., 2014; Cantlay et al., 2020b; Grieve et al., 2018; Lu et al., 2015;

McMahon et al., 2017b; Schout et al., 2018; Warner et al., 2012; Wen et al., 2019b; Woda et al., 2018) have been proposed to distinguish varying sources of subsurface fluids. In addition, electron acceptors can be reduced (e.g., SO_4 can be reduced to sulfide) by bacteria that use CH_4 as an electron donor (McMahon et al., 2017b; Schout et al., 2018; Van Stempvoort et al., 2005; Woda et al., 2018). The concentration of sulfate in waters has sometimes been used to detect anomalous CH_4 as it is easier to collect and analyze water samples for sulfate than sulfide. Other tracers were also proposed recently to detect anomalous CH_4 in groundwater, e.g., strontium isotopes (Warner et al., 2012; Woda et al., 2018), lithium and boron isotopes (Warner et al., 2014), and noble gas isotopes (Darrah et al., 2015, 2014; Wen et al., 2017, 2016; Woda et al., 2018). These latter isotopic tools require a high level of expert knowledge and advanced (and expensive) analytical facilities to ensure proper data acquisition and interpretation. Such characteristics so far have prevented these tracers from being applied widely for monitoring of baseline water quality in the area of shale gas production.

Among these hydrogeochemical tracers, no single parameter is characteristic of a particular source for anomalous CH_4 (Cantlay et al., 2020b). In past studies determining the attribution of anomalous CH_4, bivariate plots were often used, in which the dependent and independent variables are generally concentrations of hydrogeochemical parameters (e.g., chloride) and/or concentration ratios (e.g., Ca/Na). No single bivariate plot has been suggested that provides definitive answers about whether anomalous CH_4 is present. Instead, the source of elevated CH_4 is more likely to be identified through a combination of parallel bivariate plots (Cantlay et al., 2020b; Wen et al., 2019b; Woda et al., 2018).

We wanted to test machine learning approaches to identifying anomalous CH_4 in multiple shale gas plays. Logistic Regression (LR), introduced in the late 1980s (McCullagh and Nelder, 1989), is a type of machine learning algorithm used to make dichotomous prediction (e.g., whether the dissolved CH_4 concentration is higher or lower than a threshold value). LR has been used in environmental science to predict the presence of certain redox-sensitive contaminants in groundwater, e.g., arsenic (Ayotte et al., 2016), nitrate (Nolan et al., 2002), and the redox state (Tesoriero et al., 2015). LR was also used to predict the probability of CH_4 occurrence in aquifers in Alberta (Canada) (Humez et al., 2019). In these studies, a LR model was built based on a few selected categorical or continuous variables. No interactions between variables were explicitly considered. For example, Humez et al. (2019) used sulfate and total dissolved solids as input variables to predict the presence of CH_4, but no concentration ratios of these variables were ever considered. Recognizing that the concentrations of chemical constituents in groundwater are prone to differences in the baseline effects of local geology, topography, and hydrology, we suspected that the exclusion of concentration ratio variables in these previously used models largely hampered investigators' ability to generalize the models and findings to other areas. In addition, in the above-mentioned studies, all variables were fed into a single LR model to generate the final dichotomous prediction, and this made it difficult to completely understand how each variable contributed to the prediction.

In this study, we developed an ensemble LR model trained on data (Shale Network, 2015) from the northern Appalachian Basin in PA. Each sub-model focusses on a pair of hydrogeochemical parameters. More complex classifiers like Random Forest or XGBoost (Chen and Guestrin, 2016) were not used because their outputs are harder to explicitly interpret and visualize. We considered predictor features themselves and their interactions. We then explored use of the refined ensemble model on additional held-out hydrogeochemistry data from PA and other data from New York, Texas, and Colorado (Figures 1). To the best of our knowledge, this study is the first to apply an ensemble LR model to predict whether CH_4 in groundwater is anomalous or natural. With the improved availability of hydrogeochemistry data, such a prediction model could be potentially beneficial for other areas in the US and even for other countries/regions with existing or planned OG production. Although we do not have definitive evidence for which groundwater samples in each basin are affected by anomalous methane, we validated the model against a dataset of 13 sites with high [CH_4] presumed to be caused by gas development activity and 1 site with high [CH_4] thought to be unaffected by anthropogenic activities. We also explored the model by testing it against the following hypotheses we derived from the literature: i) the frequency of detection of anomalous methane (# impacted water samples / total water samples tested) in the Appalachian Basin is greater where gas wells target unconventional as compared to conventional reservoirs (Brantley et al., 2014; Ingraffea et al., 2014), and ii) the frequency of detection of anomalous methane in groundwater in the Marcellus/Utica shale play is higher than in western shale plays such as those in Texas and Colorado (e.g., Brantley et al., 2014; Hammond et al., 2020; McMahon et al., 2017a; Sherwood et al., 2016; Woda et al., 2018).

Methods and Materials

Groundwater Quality Data Used in Model Development and Application

We collated groundwater chemistry analyses from shale plays that are some of the largest in the U.S. and that also are associated with large groundwater datasets: the Appalachian Basin within Pennsylvania (*n*=23,858; Shale Network, 2015); the Fort Worth, TX-LA-MS Salt, and Western Gulf basins within Texas (*n*=688; Darvari et al., 2017; Nicot et al., 2017a, 2017b, 2017c); and the Denver-Julesburg, Raton, San Juan, Paradox, Uinta-Piceance, Greater Green, and North Park basins within Colorado (*n*=10,258; https://cogcc.state.co.us/). We decided not to include some shale plays such as the Fayetteville shale in Arkansas in our study because 1) that play has shown little evidence for gas migration (McMahon et al., 2017a; Warner et al., 2013); 2) we were unaware of a large dataset of groundwater chemistry for analysis; and 3) this region is not one of the major shale gas production areas. The samples we analyzed were compiled from a variety of data sources (Figures 1; Table 1). A relatively small dataset of groundwater quality in the counties of New York (*n*=203; Christian et al., 2016) that neighbor Pennsylvania was also tested for comparison because it represents a region of long-standing exploitation of conventional but not unconventional hydrocarbon reservoirs. More detailed descriptions of these datasets and sources are included in the supporting information. Upon publication of this work, all datasets discussed in this study will be released at: https://doi.org/XXXXXXXX.

We trained the model to a set of 19,102 groundwater sample analyses that included a small set of groundwater quality data compiled from sites that are presumed to have been impacted by anomalous CH_4. We then used an additional subset of analyses from such putatively contaminated samples to validate the model performance. Samples are referred to here as putatively contaminated when they were reported to be contaminated in published journal articles or government reports based on field and laboratory investigations (Llewellyn et al., 2015; U.S. Environmental Protection Agency, 2015; Wen et al., 2019b; Woda et al., 2018). These data were divided into a set for training (*n*=16) [from Granville Road and Paradise Road (Bradford Co.), and Gregs Run (Lycoming Co.)] and a set for validation (*n*=13) [from Paradise Road (Bradford Co.) and Sugar Run (Lycoming Co.)]. An additional sample of a well-known brine spring that is naturally CH_4-rich from northern PA was also used as part of the validation dataset. Sites were largely determined to be putatively contaminated based on temporal changes in methane concentration, carbon isotope data in collected water samples, and geological investigations.

Predictor Features Used in the Model

We strove to develop a machine learning-based ensemble model that could use geochemical analyses to quantify the likelihood that a groundwater sample with dissolved CH_4 had been impacted by anomalous rather than natural CH_4. A total of 118 geochemical parameters were considered as potential "predictor features" in the model. In other words, these features were tested for their ability to predict if CH_4 in a given sample was anomalous or natural. These predictor features divide into three main sub-categories discussed in the next paragraphs. Detailed descriptions of these predictor features can also be found in the supporting information.

Data for many of the concentrations were beneath detection (i.e., censored). For example, many of the CH_4 concentrations (31.1%) were censored (i.e., below reporting limits). To maximize the number of measurements for each feature, we chose the fifteen geochemical parameters for which the percent of non-censored measurements was >31.1%. These 15 features are referred to as 'Single Geochemical Features' or 'SGF' in Table 2. These 15 water quality parameters were reported in the training data from PA (Shale Network, 2015) with over 10,000 measurements each (Bicarbonate Alkalinity – Alk, Calcium – Ca, Chloride – Cl, Magnesium – Mg, Sodium – Na, Sulfate or SO_4^{2-} – SO4, Total Dissolved Solids – TDS, Barium – Ba, Iron – Fe, Manganese – Mn, pH, Hydrogen Ion – H, Specific Conductance – SC, Temperature – Temp, and Turbidity. Specific conductance of a water solution is a measure of its ability to conduct electricity. Turbidity is a measure of the degree to which the water solution has lost transparency due to the presence of suspended particulates. pH was considered in addition to hydrogen ion concentration because pH, the more commonly reported term, is a calculated parameter [pH = -log (hydrogen ion concentration)] that is not directly comparable to the other concentrations.

The second and third sub-categories of predictor features are the reciprocals of the concentrations of all the SGF except for pH and water temperature (*n*=13; 'Reciprocal Feature'), and the ratios of any two SGF ('Ratio Features'). Ratio features were calculated for all features except for pH, H^+, SC, water temperature, turbidity (*n*=90). The inclusion of features in the second and third categories was motivated by the observation that these two types of arithmetic combinations of geochemical parameters have been widely used by geoscientists in tracing the source of solutes in groundwater because they can represent stoichiometric coefficients in the governing geochemical reactions (e.g., Bau et al., 2004; Brantley et al., 2014; Cantlay et al., 2020d, 2020a, 2020c; Tisherman and Bain, 2019). Furthermore, we reasoned that the stoichiometries of these geochemical reactions need not be restricted to one hydrocarbon basin alone, and thus might be more predictive across basins. Subtraction and multiplication of SGF were not considered because, unlike ratios, there are no known fundamental reasons why such functions should be predictive.

A machine learning model using only geochemical predictor features allows us to focus on assessing the interplay of geochemical parameters as well as their relative importance in determining whether the elevated [CH_4] present in groundwater samples is likely to represent anomalous CH_4. Therefore, no non-geochemical features (e.g., land use and bedrock geology) were considered in this study.

Machine Learning-based Ensemble Model

Our workflow included sequential phases: ensemble model development (which included training), validation, and application (Figure 2). Taking two parameters from the 118 parameters at a time without repetition yields a total number of 6,903 pair combinations. We started with these

6,903 possible geochemical pairs that were each tested in a sub-model. The process of model development described in the next paragraphs identified the subset of 6,903 predictor pairs that were most successful in determining the likelihood of a groundwater sample being impacted by anomalous CH_4. The likelihood was calculated as the ratio of the number of sub-models flagging the CH_4 as not being naturally derived (and therefore, putatively, “anomalous”) divided by the total number of models considered. The collection of sub-models is termed here the ensemble model.

At the stage of ensemble model development (green in Figure 2), sub-models using the full list of 118 geochemical features paired into 6,903 predictors were trained on 19,102 analyses referred to as training data (Table 1; Figure 2). Each of the 6,903 sub-models (Figure 2) considers only one pair of predictor features. In effect, each sub-model imitates the traditional procedure of using bivariate plots of geochemical variables to investigate the origin of dissolved CH_4 in groundwater [see, for example, Figure S10 in Woda et al. (2018)]. The performance of each sub-model was evaluated with respect to two tasks. First, it was evaluated for its ability to predict whether the [CH_4] is above or below the threshold identified as potentially problematic, 10 mg/L, in the 19,086 water samples (Task 1). Given that these samples were collected by professional consultants working for gas companies before drilling gas wells within a few kilometers of the water wells, the waters could be contaminated or uncontaminated by CH_4 from previous oil/gas activity, but we have no outside evidence of contamination. We hypothesized that successful sub-models would predict [CH_4] < 10 mg/L for samples reported to have [CH_4] ≥ 10 mg/L. We inferred these samples are impacted by anomalous CH_4. This approach is implicitly built on the assumption that natural high-[CH_4] waters are chemically distinct from anomalous high-[CH_4] waters. However, we acknowledge that our method will not detect anomalous methane in waters with very low values of [CH_4]. In Task 2, each sub-model was additionally evaluated to see if it successfully identified the CH_4 in 16 of the putatively contaminated training samples as anomalous (Task 2). The concentration of 10 mg/L was selected because it is a threshold above which immediate action is needed (Eltschlager et al., 2001; Wen et al., 2019b).

During ensemble model development (green in Figure 2), a linear classifier – logistic regression (LR) – is used in the two tasks with 5-fold cross-validation. LR is generally used to solve binary classification problems. With LR, the relationship between predictor features and the prediction outcome can be described by a sigmoidal function as illustrated by Equation (1) in which x represents predictor features and y is the binary model output of 0 ([CH_4] <10 mg/L) or 1 ([CH_4] ≥ 10 mg/L).

$$y = \frac{e^{wx+b}}{1+e^{wx+b}} \tag{1}$$

In the training, the sub-model allows optimization of w and b to best predict the training data. Once w and b are optimized, each sub-model can be used in Task 1 to predict which groundwater samples are characterized by [CH_4] ≥ 10 mg/L. For this task, Area Under the Curve (AUC) is considered as the performance metric where a higher AUC value indicates higher accuracy of prediction. AUC, in particular, was chosen as the accuracy measure as it holistically evaluates the model performance in predicting both positive (i.e., high methane sample) and negative (i.e., low methane sample) results.

In Task 2 the sub-model was applied to 16 putatively contaminated samples. F1 score was calculated following Equation (2) as the performance metric:

$$F1\ score = \frac{2}{recall^{-1}+precision^{-1}} \tag{2}$$

Here *recall* is defined as the proportion of the 16 samples identified correctly while *precision* is the proportion of predicted anomalous CH_4 samples that are actually contaminated. F1 score was chosen to prioritize maximizing true positives (i.e., known problematic samples correctly identified) as it does not consider true negatives (i.e., known natural samples correctly identified).

Both Tasks 1 and 2 were used to identify the best ensemble model to detect anomalous CH_4 in impacted waters. Samples reported to have high [CH_4] that were incorrectly classified as low [CH_4] samples by sub-models in Task 1 were inferred to have been impacted by anomalous CH_4. We chose sub-models with higher accuracy of prediction in Task 1 to yield a lower number of false positives. To evaluate the performance of each sub-model for both Tasks 1 and 2, we defined a synthesized metric, namely, 0.7 x F1 score + 0.3 x AUC. The selection of coefficients 0.7 and 0.3 is not entirely arbitrary. We assigned the higher weight (0.7) to F1 as that task is more critical. All sub-models (and the associated pair of features; *n*=6,903) were ranked by the final synthesized score. We also completed a sensitivity test of the coefficients (Table S1a and S1b) and determined that feature members in the top feature list (e.g., top 20) are mostly unchanged (e.g., SC, Na, Cl/SO4), although the exact rank of features by frequency changes to a limited extent.

In the phases of model validation and application, the prediction results for the best performing sub-models were used to estimate the likelihood of impact of groundwater samples by anomalous CH_4 (Figure 2). This likelihood was defined as the percentage of considered sub-models that classified the sample as impacted by anomalous CH_4. We calculated four types of likelihood for each water sample using the best *n* sub-models where we explored *n* = 100, 200, 500 as well as 1,000. We ultimately chose *n* = 1,000 to be conservative. The selection of *n* =1,000 is supported by the observation that (1) an ensemble model with fewer sub-models would be more likely to be biased towards a few geochemical parameters (e.g., the ensemble model for *n* = 100 was observed to be highly biased towards SC as discussed in the text below) and (2) an ensemble model with a larger number of sub-models (e.g., *n* = 1,000) was expected to assess water chemistry more holistically, improving model applicability and performance for new datasets with less likelihood of overfitting.

In this study, the ensemble model was trained and validated with hydrogeochemistry data in Pennsylvania (i.e., Marcellus shale) and then was tested in other natural gas production regions in Colorado, Texas, and New York. Differences in bedrock geology, regional hydrogeochemistry, topography, and land use between testing and training/validation datasets might in some cases hamper the applicability of the developed ensemble model. We strove to minimize this issue by considering interactions of ratios of concentrations of single geochemical features as these concentration ratios tend to yield less variability across different regions. As more hydrogeochemical data become available, in particular from presumably contaminated sites, the ensemble model can be further validated and even re-trained with new data.

Results and Discussion

Model Development

Two statistical methods – Pearson correlation and Spearman's rank correlation – were adopted to assess the pairwise correlation of the 15 single geochemical features in predicting the target feature, [CH_4]. The Pearson correlation is used to evaluate linear relationships between continuous variables that are assumed to obey normal distributions, while the Spearman's rank correlation is used to evaluate monotonic relationships between two continuous variables without the requirement that variables are normally distributed. Therefore, the Spearman's rank correlation identifies more pairs of features that report statistically significant correlations ($p < 0.05$) compared to the Pearson correlation (Figure 3).

The pairwise correlations led us to note three groups for the fifteen single geochemical features: 1) salinity-related (Alk, Ba, Ca, Cl, Mg, Na, TDS, SC); (2) redox-sensitive (Fe, Mn, and SO4); and (3) other (pH, Temp, Turbidity, and H^+). These groupings are used to categorize the best-performing sub-models in subsequent discussions.

General Observations

The synthesized scores of all sub-models (*n*=6,903) are listed and ranked in Table S1a. The best performing sub-model (synthesized score = 0.9579, F1 = 0.9677, AUC = 0.9350) corresponds to the pair, 1/SC and Ca/Ba (Table S1a). From this observation we inferred that salinity is an effective tool to detect anomalous CH_4 samples (SC, Ca, and Ba are all salinity-related features). Previous studies (e.g., Brantley et al., 2014; Cantlay et al., 2020c; Tisherman and Bain, 2019; Wen et al., 2019b; Woda et al., 2020, 2018) have also identified other salinity-related parameters (Cl, Ca/Na, Mg/Na, Ca/Mg, and Ba/Cl) that are helpful in detecting anomalous CH_4 in groundwater samples or distinguishing chemical signatures of produced waters from shale gas development from other types of contamination. These same parameters, i.e., Cl, Ca/Na, Mg/Na, Ca/Mg, and Ba/Cl, show up in sub-models ranked as high as 8th, 13th, 19th, 82nd, and 191st (Table S1a), respectively. Furthermore, the frequency of inclusion of these features in the top 1000 sub-models are 87, 4, 2, 11, and 8, respectively (Table S1b and Figure 4). In addition to the salinity-related parameters mentioned by previous authors, a few new ones were identified as important (i.e., high frequency) in the top 1,000 performance sub-models, e.g., Ca/TDS (49) and Ba/Alk (81) (Table S1b). These salinity-related features can effectively distinguish natural migration of thermogenic CH_4 from anomalous CH_4 presumably because CH_4 often migrates naturally with salt-containing Appalachian Basin brine into shallow groundwater (e.g., Warner et al., 2012; Wen et al., 2019b;

Woda et al., 2018). Alkalinity may appear because it increases during sulfate reduction, one of the redox processes (see next paragraph) that sometimes couple to methane oxidation (Woda et al., 2018).

In addition to salinity, redox-sensitive features (e.g., Fe/SO4, SO4/Na) are also frequently found in the top performing sub-models (Table S1, and Figure 4). For example, SO4 and Fe concentrations have previously been used to detect anomalous CH_4 in groundwater (Wen et al., 2019b; Woda et al., 2020, 2018) and the frequency of SO4-containing and Fe-containing features (e.g., Fe/SO4) appearing in the top 100 sub-models are 17 and 10, respectively. It is worth noting that many SO4- or Fe-containing features also incorporate salinity-related features, e.g., SO4/Na and Fe/Ba. Redox features are expected because the presence of CH_4 creates an anoxic environment that in turn promotes reduction of redox-sensitive species such as sulfate (SO_4^{2-}) to sulfide (S^{2-}) and Fe(III) to Fe(II). The formation of sulfide leads to precipitation of metal sulfide and a decrease in SO4 and Fe concentrations over time (Woda et al., 2020, 2018). Low concentrations of all of these redox-sensitive parameters are expected when the source of CH_4 is natural and the groundwater has had a relatively long time to reach thermodynamic equilibrium. In contrast, when anomalous CH_4 migrates into shallow, oxygenated groundwater, months can pass before the water reaches chemical equilibrium, allowing the co-existence of transiently high methane, sulfate, and iron in the same water samples that can be detected during the transient (Wen et al., 2019b; Woda et al., 2020, 2018).

Specific Observations

Best performing features are summarized for the top 100, 200, 500, and 1,000 models in Figure 4 and the likelihood of the presence of anomalous CH_4 in groundwater is calculated for all of these four top lists in the following sections. It is clear that in the first two lists (i.e., top 100, 200), the frequency of SC-containing features (SC and 1/SC) dominate over the other top features. So, if the likelihood of groundwater samples being impacted by anomalous CH_4 is calculated from an ensemble model considering only the top 100 (or 200) features, the calculated likelihood will be largely biased towards SC. Two more features (Na and Cl) are reported with high frequency in the top 500 list, but the calculated likelihood of groundwater samples being impacted by anomalous CH_4 still depends strongly on the salinity features SC, Na, and Cl. To holistically assess groundwater chemistry and to be more conservative, we focus the most extensive discussion mainly about the results based on the list for the top 1,000 (the "top 1,000 model", Figure 4).

Model Validation

After training with the use of putatively contaminated sample data (n = 16), the top 1,000 sub-models were validated against another groundwater quality dataset of 14 high-[CH_4] groundwater samples from northeastern or central PA. All were reported with [CH_4] ≥10 mg/L. Of these, 13 were reported for 4 sites putatively impacted by anomalous CH_4 (Llewellyn et al., 2015; Wen et al., 2019b; Woda et al., 2018). The likelihood of contamination was >90% for each of these 13 validation samples (Tables 3 and S2) by the ensemble model using the top 1,000 sub-models. This suggests that the ensemble model has high sensitivity (i.e., high recall rate), i.e., the model is very effective in detecting anomalous CH_4 in groundwater samples (true positives). The 14th sample that we used for validation was groundwater from Salt Spring State Park in northern PA. That spring sample is known to contain naturally occurring CH_4 with [CH_4] > 10 mg/L ('Salt Spring State Park' in Table S2). The predicted likelihood that this groundwater sample is impacted by anomalous CH_4 is 14%, which means that the ensemble model predicts that this sample is likely not contaminated by anomalous CH_4 (i.e., a true negative). Thus, the proposed ensemble model is very effective in distinguishing whether high [CH_4] in groundwater is naturally occurring (e.g., <20% likelihood) or anomalous (e.g., >90% likelihood).

Model Application

We now apply this ensemble model to the hold-out Pennsylvania data (data not incorporated in the training dataset) as well as groundwater data from other states (Figure 5) to seek to detect anomalous CH_4 in other groundwater samples as a way to explore our two hypotheses and to compare to previous research (Christian et al., 2016; Darvari et al., 2017; Nicot et al., 2017a, 2017b, 2017c; Sherwood et al., 2016; Wen et al., 2019b, 2018).

Pennsylvania Hold-out Data

A total of 4,772 groundwater samples from the Shale Network database are in the Pennsylvania hold-out dataset. Among these, only 64 show both [CH_4] ≥ 10 mg/L as well as measurements for all the single predictor features. These 64 groundwater samples are all located in northeastern PA (Figures 5 and 6; Table S3). The top 1,000 model predicts that 29 of these 64 groundwater samples show a >50% likelihood and, of these, 13 show ≥80% likelihood of being impacted by anomalous CH_4 (Figures 5 and 6; Table S3).

To assess the efficacy of the ensemble model for Pennsylvania, we compare the machine learning results to the results based on a more simplified and streamlined approach by Wen et al. (2019b). Based on six selected geochemical parameters (CH_4, Cl, Ca, Na, Fe, and sulfate concentration), Wen et al. (2019b) categorized groundwater samples into five geochemical types. Their type 4 and 5 waters were considered to the most likely to be impacted by anomalous CH_4:

their characteristics include Ca/Na mass ratio ≥0.52, Cl ≤30 mg/L, [CH_4] ≥10 mg/L, and dis-equilibrated Fe concentrations (≥0.3 mg/L) and/or dis-equilibrated sulfate concentration (≥ 6 mg/L). Using this simplified workflow from Wen et al. (2019b), we identified a total of 17 samples of types 4 or 5 in the hold-out PA dataset. Ten of these 17 samples were used to train the ensemble model and five were not considered by the ensemble model because they lacked values for all of the 15 geochemical features. The remaining two analyses (sample IDs are "PADEP_Predrill_CHK_827" and "PADEP_Predrill_Bradford-Burlington_T_005-well"; Table S3) were detected accurately (true positives) by the top 1,000 model as showing a high likelihood of impact by anomalous CH_4.

In addition to the two samples detected by both the ensemble model in this study as well as by Wen et al. (2019b), the ensemble model also identified 27 other samples with a predicted likelihood of >50% (12 samples ≥80%) of anomalous CH_4. The workflow of Wen et al. (2019b) did not detect these 27 samples as impacted by anomalous methane because Cl ≥30 mg/L or sulfate ≤6 mg/L. If the new ensemble model approach is correct, these samples are impacted by anomalous methane and were false negatives in the previous study. Lending credence to the conclusion that these sites were impacted by anomalous methane is the observation that all but 11 of these 27 samples were from four areas that lie near locations with known contamination as reported in the literature (Llewellyn et al., 2015; Wen et al., 2019b, 2018) (Figures 5 and 6; Table S3). The last eleven putatively-false-negatives are not close to any sites previously identified as problematic. Given that all the other 16 sites show some history of local contamination, we conclude that all these putatively-false-negatives were impacted by recent intrusion of anomalous methane (which might be caused by nearby shale gas drilling) at the time of sampling, and should have received in-depth analysis to determine the source and mechanism of the elevated [CH_4].

Although the streamlined workflow from Wen et al. (2019b) is easy to implement and effective in detecting anomalous methane samples, it is limited in the number of geochemical parameters considered and it likely returned at least 27 false negatives for the PA hold-out dataset. In contrast, our ensemble model evaluates a fuller set of groundwater chemistry data and can identify waters that are contaminated that were missed by the streamlined workflow of Wen et al. (2019b).

New York

Although New York is almost geologically identical to Pennsylvania in the counties considered here, it is not producing shale gas because high-volume hydraulic fracturing with horizontal drilling is banned. Instead, conventional oil and gas drilling has long been important in the state. Brantley et al. (2014) summarized reports from the PA regulator that showed that the reporting rate of problems of anomalous CH_4 from conventional oil and gas well development in PA was much smaller than that observed for shale gas development after 2004. We thus hypothesized that we would see a very low rate of anomalous CH_4 in the small dataset of groundwater quality available for the region of oil and gas production in New York (Christian et al., 2016). This hypothesis is especially compelling because geologically, northern PA is very similar to NY, but the laws are different, and the use of high-volume high-pressure hydraulic fracturing in NY unconventional reservoirs is not allowed. Therefore, unlike PA samples, all of the New York samples were collected in advance of high-volume hydraulic fracturing but might postdate some of the nearby conventional drilling.

In the New York dataset of 203 groundwater samples, only five are reported with high [CH_4] ($\geq$10 mg/L). Four of these five samples are predicted by the top 1,000 ensemble model as low likelihood (i.e., <30%) of containing anomalous methane (Tables 3 and S4; Figures 5 and 6). However, one ("ST35B"), with a [CH_4] of 13.8 mg/L, yields a 71% likelihood of impact by anomalous CH_4 (Table S4 and Figures 5 and 6). Sample ST35B was identified in the ensemble model because it has a very low [Cl] of 4.1 mg/L, but relatively high [sulfate] and [Fe] of 8.99 and 0.42 mg/L, respectively. It is thus also detected as a contaminated sample by the streamlined test of Wen et al. (2019b) and is likely, based on these tests, to contain CH_4 that has infiltrated the groundwater recently.

Although this sample appears to have been infiltrated with CH_4 recently, other lines of evidence suggest that the migration was not caused by oil or gas development. For example, the distance between site ST35B (sampled in 2013) and the nearest oil or gas well (drilled prior to 1980) is ~6 km. The large distance and long period of time between gas well drilling and water sample collection render the source of methane in water wells unlikely to have originated from the gas wells. But this water well is drilled into the very gas-rich Canadaway Formation near a large number of faults and lineaments, and may nonetheless have been affected by recent migration of naturally occurring CH_4 to the water well. Consistent with this conclusion, Christian et al. (2016) pointed out that values of [CH_4] in the NY groundwater data (including, implicitly, ST35B) do not statistically vary with proximity to gas wells or faults. They argued that elevated [CH_4] in the NY samples is most likely of natural origin. This conclusion was based on the observation that high [CH_4] is often associated with Na-rich waters in valleys in NY that are impacted by natural brines (Christian et al., 2016). However, while ST35B had a high [Na] (30.1 mg/L), it had very low [Cl] (4.1 mg/L). Therefore, the high Na in that sample did not derive from contamination by NaCl-rich brine (and was not flagged as such by the ensemble model). On the other hand, water from ST35B was reported to smell like 'rotten egg' (i.e., H_2S) and is locally associated with brownish red stains (pers. comm., Laura Lautz), perhaps pointing to ongoing sulfate reduction coupled with methane oxidation (Wen et al., 2019b; Woda et al., 2020, 2018). Our prediction that ST35B contains anomalous methane might be a false positive: in particular, it could be water that was recently impacted by natural CH_4 rather than CH_4 migrated from hydrocarbon development activities. In future predictions with the ensemble model, a filter for proximity to oil/gas wells should be included (e.g., waters sampled within 5 km).

A Fisher exact test was used to test our original hypothesis, i.e., to test if the frequency of detection of CH_4-impacted groundwater samples (≥50% likely to be impacted by anomalous methane) differ between PA (i.e., 42/2,983) and NY (i.e., 1/78). Whether we assume that ST35B is an example of anomalous methane caused by oil/gas development or not, the Fisher test results show no statistically significant difference between PA and NY at a confidence level of 95% (p>0.05). Thus, we must reject the hypothesis in that the ensemble model reveals similar frequencies of putative contamination by anomalous methane in the part of the Marcellus shale play utilizing horizontal drilling + hydraulic fracturing as compared to the part with conventional resource development.

Texas and Colorado

We also compiled groundwater quality data from Texas to test the hypothesis that the frequency of identification of anomalous methane in groundwater wells is higher in PA than in other shale plays/states. A total of 688 groundwater samples are included in the compiled Texas data, which cover three major shale gas plays in Texas: Barnett Shale, Eagle Ford Shale, and Haynesville Shale. Among these samples, [CH_4] is larger than 10 mg/L in 34 groundwater samples (Table S5 and Figures 5 and 6). Of these 34 samples, none were identified to have a likelihood of anomalous CH_4 larger than 50%. This is consistent with the findings from previous investigations of the sources of CH_4 in these three shale plays (Darvari et al., 2017; Nicot et al., 2017a, 2017b, 2017c), i.e., no CH_4 in these shallow water wells is associated with recent development of nearby shale gas.

The ensemble model was also applied to a large groundwater quality dataset of 10,258 water samples collected across Colorado mainly within the Denver-Julesburg, Raton, San Juan, and Uinta-Piceance basins (Figure 1). Of these samples, high [CH_4] (≥10 mg/L) samples (*n*=58) with complete measurements for all required single geochemical features exist only in the Denver-Julesburg Basin and Raton Basin (Figure 1). These 58 samples were collected from 29 sites (Table S6), of which only four samples from four sites (i.e., 705739, 755481, 752672, 750143) are associated by the ensemble model with a likelihood >50% for anomalous methane. Three of the sites are in the Denver-Julesburg Basin and one in the Raton Basin. All other samples/sites with high [CH_4] are calculated to be <50% likely to have been impacted by anomalous CH_4. A detailed discussion of the results of model application to Colorado dataset is included in the supporting information.

With these data, we test our second hypothesis, namely that the frequency of detection of anomalous methane in the Marcellus shale play is higher than in Texas + Colorado. Treating the Colorado and Texas measurements as one dataset, the combined frequency of detection of groundwater samples showing >50% likelihood is 4/795, statistically lower than the rate in Pennsylvania (i.e., 42/2,983) at a confidence level of 95% (*p*=0.04) using the Fisher exact test. This finding supports our second hypothesis that the frequency of detection in the Marcellus shale play is higher than in Texas and Colorado. To explain this result, we point to Hammond et al. (2020) who suggested that most of the anomalous CH_4 released from gas wells in the US occurs because primary cementation is not completed along the full lengths of production casings from the target shale to intermediate casings (or to surface casings, if intermediate casings are not used). The higher frequency in the Marcellus play could therefore derive from differences in casings and cementation or differences in gas contents at intermediate depths for boreholes in that play as compared to other plays.

Conclusion

We presented a machine learning ensemble model that shows that salinity-related and redox-related measurements are effective geochemical features that can detect anomalous methane in groundwater. One problem with the application of the model to date is that we cannot exclude the possibility that differences in sample sizes contributed to differences in detection frequencies. Clearly, larger datasets for all shale gas plays could be used to eliminate this problem. Furthermore, a machine learning model that incorporated additional geochemical measurements such as isotopes would presumably be even more adept at finding evidence of anomalous CH_4.

For the regions we studied (PA, TX, CO, NY), the frequency of reported water samples with [CH_4] ≥10 mg/L, the level often considered to be dangerous, was 2.1%, 5.8%, 7.4%, and 3.4%. In contrast, the frequency of identification of anomalous methane by the ensemble machine learning model was 1.4%, 0%, 0.9%, 1.3% (with ST35B) and 0% (without ST35B). These values were determined by application of the ensemble model where we detected 42/2,983, 0/338, 4/457, and 1/78 (or 0/77) of the water samples were >50% likely to be impacted by newly migrated natural gas in these states, respectively. One NY sample, flagged as likely to have newly migrated methane, was sampled ~6 km from the nearest oil/gas well, and thus may be contaminated by new methane from a different source. Fisher exact tests show statistically significant differences between the results for PA versus TX+CO (i.e., with evidence for a higher frequency of migrated methane sites in PA) but show no statistically significant difference within the Marcellus shale play between regions using versus not using high-pressure high-volume hydraulic fracturing. The new machine learning tool appears to be useful in detecting anomalous methane in multiple shale plays. Future work should include additional training and validation of data-driven models as new data from presumably impacted sites in shale plays other than the Marcellus become available.

Acknowledgements

Funding was derived from the National Science Foundation IIS-16-39150 and US Geological Survey (104b award G16AP00079) through the Pennsylvania Water Resource Research Center to S.L.B. and Zhenhui Li. T.W. was supported by the College of Earth and Mineral Sciences Dean’s Fund for Postdoc-Facilitated Innovation at Penn State and by NSF IIS-16-39150. We thank Alison Herman and Marcus Guarnieri for data management. Bruce Lindsey (US Geological Survey), Seth Pelepko and Stew Beattie (PA DEP) are acknowledged for access and help with data, Maurie Kelly and James Spayd (Penn State University) for help with data publication and storage.

References

Ayotte, J.D., Nolan, B.T., Gronberg, J.A., 2016. Predicting Arsenic in Drinking Water Wells of the Central Valley, California. Environmental Science & Technology 50, 7555–7563. https://doi.org/10.1021/acs.est.6b01914

Baldassare, F.J., McCaffrey, M.A., Harper, J.A., 2014. A geochemical context for stray gas investigations in the northern Appalachian Basin: Implications of analyses of natural gases from Neogene-through Devonian-age strata. AAPG Bulletin 98, 341–372. https://doi.org/10.1306/06111312178

Bau, M., Alexander, B., Chesley, J.T., Dulski, P., Brantley, S.L., 2004. Mineral dissolution in the Cape Cod aquifer, Massachusetts, USA: I . Reaction stoichiometry and impact of accessory feldspar and glauconite on strontium isotopes, solute concentrations, and REY distribution 1 1Associate Editor: L. M. Walter. Geochimica et Cosmochimica Acta 68, 1199–1216. https://doi.org/10.1016/j.gca.2003.08.015

Brantley, S.L., Yoxtheimer, D., Arjmand, S., Grieve, P., Vidic, R., Pollak, J., Llewellyn, G.T., Abad, J., Simon, C., 2014. Water resource impacts during unconventional shale gas development: The Pennsylvania experience. International Journal of Coal Geology 126, 140–156. https://doi.org/10.1016/j.coal.2013.12.017

Cantlay, T., Bain, D.J., Curet, J., Jack, R.F., Dickson, B.C., Basu, P., Stolz, J.F., 2020a. Determining conventional and unconventional oil and gas well brines in natural sample II: Cation analyses with ICP-MS and ICP-OES. Journal of Environmental Science and Health, Part A 55, 11–23. https://doi.org/10.1080/10934529.2019.1666561

Cantlay, T., Bain, D.J., Stolz, J.F., 2020b. Determining conventional and unconventional oil and gas well brines in natural samples III: mass ratio analyses using both anions and cations. Journal of Environmental Science and Health, Part A 55, 24–32. https://doi.org/10.1080/10934529.2019.1666562

Cantlay, T., Bain, D.J., Stolz, J.F., 2020c. Determining conventional and unconventional oil and gas well brines in natural samples III: mass ratio analyses using both anions and cations. Journal of Environmental Science and Health, Part A 55, 24–32. https://doi.org/10.1080/10934529.2019.1666562

Cantlay, T., Eastham, J.L., Rutter, J., Bain, D.J., Dickson, B.C., Basu, P., Stolz, J.F., 2020d. Determining conventional and unconventional oil and gas well brines in natural samples I: Anion analysis with ion chromatography. Journal of Environmental Science and Health, Part A 55, 1–10. https://doi.org/10.1080/10934529.2019.1666560

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 785–794. https://doi.org/10.1145/2939672.2939785

Christian, K.M., Lautz, L.K., Hoke, G.D., Siegel, D.I., Lu, Z., Kessler, J., 2016. Methane occurrence is associated with sodium-rich valley waters in domestic wells overlying the Marcellus shale in New York State. Water Resources Research 52, 206–226. https://doi.org/10.1002/2015WR017805

Darrah, T.H., Jackson, R.B., Vengosh, A., Warner, N.R., Whyte, C.J., Walsh, T.B., Kondash, A.J., Poreda, R.J., 2015. The evolution of Devonian hydrocarbon gases in shallow aquifers of the northern Appalachian Basin: Insights from integrating noble gas and hydrocarbon geochemistry. Geochimica et Cosmochimica Acta 170, 321–355. https://doi.org/10.1016/j.gca.2015.09.006

Darrah, T.H., Vengosh, A., Jackson, R.B., Warner, N.R., Poreda, R.J., 2014. Noble gases identify the mechanisms of fugitive gas contamination in drinking-water wells overlying the Marcellus and Barnett Shales. Proceedings of the National Academy of Sciences 111, 14076–14081. https://doi.org/10.1073/pnas.1322107111

Darvari, R., Nicot, J.-P., Scanlon, B.R., Mickler, P., Uhlman, K., 2017. Trace Element Behavior in Methane-Rich and Methane-Free Groundwater in North and East Texas. Groundwater 1–14. https://doi.org/10.1111/gwat.12606

Eltschlager, K.K., Hawkins, J.W., Ehler, W.C., Baldassare, F., Dep, P., 2001. Technical measures for the investigation and mitigation of fugitive methane hazards in areas of coal mining. Department of the Interior Office of Surface Mining.

Grieve, P.L., Hynek, S.A., Heilweil, V., Sowers, T., Llewellyn, G., Yoxtheimer, D., Solomon, D.K., Brantley, S.L., 2018. Using environmental tracers and modelling to identify natural and gas well-induced emissions of methane into streams. Applied Geochemistry 91, 107–121. https://doi.org/10.1016/j.apgeochem.2017.12.022

Guo, M., Xu, Y., Chen, Y.D., 2014. Fracking and pollution: Can China rescue its environment in time? Environmental Science and Technology 48, 891–892. https://doi.org/10.1021/es405608b

Hammond, P.A., 2016. The relationship between methane migration and shale-gas well operations near Dimock, Pennsylvania, USA. Hydrogeology Journal 24, 503–519. https://doi.org/10.1007/s10040-015-1332-4

Hammond, P.A., Wen, T., Brantley, S.L., Engelder, T., 2020. Gas well integrity and methane migration: evaluation of published evidence during shale-gas development in the USA. Hydrogeology Journal. https://doi.org/10.1007/s10040-020-02116-y

Humez, P., Osselin, F., Wilson, L.J., Nightingale, M., Kloppmann, W., Mayer, B., 2019. A Probabilistic Approach for Predicting Methane Occurrence in Groundwater. Environmental Science & Technology 53, 12914–12922. https://doi.org/10.1021/acs.est.9b03981

Ingraffea, A.R., Wells, M.T., Santoro, R.L., Shonkoff, S.B.C., 2014. Assessment and risk analysis of casing and cement impairment in oil and gas wells in. Proceedings of the National Academy of Sciences 111, 10955–10960. https://doi.org/10.1073/pnas.1323422111

Jackson, R.B., Vengosh, A., Darrah, T.H., Warner, N.R., Down, A., Poreda, R.J., Osborn, S.G., Zhao, K., Karr, J.D., 2013. Increased stray gas abundance in a subset of drinking water wells near Marcellus shale gas extraction. Proceedings of the National Academy of Sciences of the USA 110, 11250–11255. https://doi.org/10.1073/pnas.1221635110/-/DCSupplemental/pnas.201221635SI.pdf

Lackey, G., Rajaram, H., Bolander, J., Sherwood, O.A., Ryan, J.N., Shih, C.Y., Bromhal, G.S., Dilmore, R.M., 2021. Public data from three US states provide new insights into well integrity. Proceedings of the National Academy of Sciences 118. https://doi.org/10.1073/pnas.2013894118

Li, H., Carlson, K.H., 2014. Distribution and Origin of Groundwater Methane in the Wattenberg Oil and Gas Field of Northern Colorado. Environmental Science & Technology 48, 1484–1491. https://doi.org/10.1021/es404668b

Llewellyn, G.T., Dorman, F., Westland, J.L., Yoxtheimer, D., Grieve, P., Sowers, T., Humston-Fulmer, E., Brantley, S.L., 2015. Evaluating a groundwater supply contamination incident attributed to Marcellus Shale gas development. Proceedings of the National Academy of Sciences 112, 6325–6330. https://doi.org/10.1073/pnas.1420279112

Lu, Z., Hummel, S.T., Lautz, L.K., Hoke, G.D., Zhou, X., Leone, J., Siegel, D.I., 2015. Iodine as a sensitive tracer for detecting influence of organic-rich shale in shallow groundwater. Applied Geochemistry 60 IS-, 29–36.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. CRC Press LLC, Boca Raton, UNITED STATES.

McMahon, P.B., Barlow, J.R.B., Engle, M.A., Belitz, K., Ging, P.B., Hunt, A.G., Jurgens, B.C., Kharaka, Y.K., Tollett, R.W., Kresse, T.M., 2017a. Methane and Benzene in Drinking-Water Wells Overlying the Eagle Ford, Fayetteville, and Haynesville Shale Hydrocarbon Production Areas. Environmental Science & Technology acs.est.7b00746. https://doi.org/10.1021/acs.est.7b00746

McMahon, P.B., Belitz, K., Barlow, J.R.B., Jurgens, B.C., 2017b. Methane in aquifers used for public supply in the United States. Applied Geochemistry 84, 337–347. https://doi.org/10.1016/j.apgeochem.2017.07.014

Molofsky, L.J., Connor, J.A., Wylie, A.S., Wagner, T., Farhat, S.K., 2013. Evaluation of Methane Sources in Groundwater in Northeastern Pennsylvania. Groundwater 51, 333–349.

Nicot, J.P., Larson, T., Darvari, R., Mickler, P., Slotten, M., Aldridge, J., Uhlman, K., Costley, R., 2017a. Controls on Methane Occurrences in Shallow Aquifers Overlying the Haynesville Shale Gas Field, East Texas. Groundwater 55, 443–454. https://doi.org/10.1111/gwat.12500

Nicot, J.P., Larson, T., Darvari, R., Mickler, P., Uhlman, K., Costley, R., 2017b. Controls on Methane Occurrences in Aquifers Overlying the Eagle Ford Shale Play, South Texas. Groundwater 55, 455–468. https://doi.org/10.1111/gwat.12506

Nicot, J.-P., Mickler, P., Larson, T., Clara Castro, M., Darvari, R., Uhlman, K., Costley, R., 2017c. Methane Occurrences in Aquifers Overlying the Barnett Shale Play with a Focus on Parker County, Texas. Ground Water 98, n/a-n/a. https://doi.org/10.1111/gwat.12508

Nicot, J.-P., Scanlon, B.R., 2012. Water Use for Shale-Gas Production in Texas, U.S. Environmental Science & Technology 46, 3580–3586. https://doi.org/10.1021/es204602t

Nolan, B.T., Hitt, K.J., Ruddy, B.C., 2002. Probability of Nitrate Contamination of Recently Recharged Groundwaters in the Conterminous United States. Environmental Science & Technology 36, 2138–2145. https://doi.org/10.1021/es0113854

Osborn, S.G., Vengosh, A., Warner, N.R., Jackson, R.B., 2011. Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing. Proceedings of the National Academy of Sciences 108, 8172–8176. https://doi.org/10.1073/pnas.1100682108

Schout, G., Hartog, N., Hassanizadeh, S.M., Griffioen, J., 2018. Impact of an historic underground gas well blowout on the current methane chemistry in a shallow groundwater system. Proceedings of the National Academy of Sciences 115, 296–301. https://doi.org/10.1073/pnas.1711472115

Shale Network, 2015. DOI: 10.4211/his-data-shalenetwork [WWW Document].

Sherwood, O.A., Rogers, J.D., Lackey, G., Burke, T.L., Osborn, S.G., Ryan, J.N., 2016. Groundwater methane in relation to oil and gas development and shallow coal seams in the Denver-Julesburg Basin of Colorado. Proceedings of the National Academy of Sciences 113, 8391–8396.

Siegel, D.I., Azzolina, N.A., Smith, B.J., Perry, A.E., Bothun, R.L., 2015a. Methane Concentrations in Water Wells Unrelated to Proximity to Existing Oil and Gas Wells in

Northeastern Pennsylvania. Environmental Science & Technology 49, 4106–4112. https://doi.org/10.1021/es505775c

Siegel, D.I., Smith, B., Perry, E., Bothun, R., Hollingsworth, M., 2015b. Pre-drilling water-quality data of groundwater prior to shale gas drilling in the Appalachian Basin: Analysis of the Chesapeake Energy Corporation dataset. Applied Geochemistry 63, 37–57. https://doi.org/10.1016/j.apgeochem.2015.06.013

Tesoriero, A.J., Terziotti, S., Abrams, D.B., 2015. Predicting Redox Conditions in Groundwater at a Regional Scale. Environmental Science & Technology 49, 9657–9664. https://doi.org/10.1021/acs.est.5b01869

Tisherman, R., Bain, D.J., 2019. Alkali earth ratios differentiate conventional and unconventional hydrocarbon brine contamination. Science of The Total Environment 695, 133944. https://doi.org/10.1016/j.scitotenv.2019.133944

U.S. Energy Information Administration, 2018. Annual Energy Outlook 2018. https://doi.org/DOE/EIA-0383(2017)

U.S. Energy Information Administration, 2017. International Energy Outlook 2017, International Energy Outlook 2017. https://doi.org/www.eia.gov/forecasts/ieo/pdf/0484(2016).pdf

U.S. Environmental Protection Agency, 2015. Retrospective Case Study in Northeastern Pennsylvania Study of the Potential Impacts of Hydraulic Fracturing on Drinking Water Resources.

Van Stempvoort, D., Maathuis, H., Jaworski, E., Mayer, B., Rich, K., 2005. Oxidation of fugitive methane in ground water linked to bacterial sulfate reduction. Ground Water 43, 187–199. https://doi.org/10.1111/j.1745-6584.2005.0005.x

Vidic, R.D., Brantley, S.L., Vandenbossche, J.M., Yoxtheimer, D., Abad, J.D., 2013. Impact of Shale Gas Development on Regional Water Quality. Science 340, 1235009–1235009. https://doi.org/10.1126/science.1235009

Warner, N.R., Darrah, T.H., Jackson, R.B., Millot, R., Kloppmann, W., Vengosh, A., 2014. New Tracers Identify Hydraulic Fracturing Fluids and Accidental Releases from Oil and Gas Operations. Environmental Science & Technology 48, 12552–12560. https://doi.org/10.1021/es5032135

Warner, N.R., Jackson, R.B., Darrah, T.H., Osborn, S.G., Down, A., Zhao, K., White, A., Vengosh, A., 2012. Geochemical evidence for possible natural migration of Marcellus Formation brine to shallow aquifers in Pennsylvania. Proceedings of the National Academy of Sciences 109, 11961–11966. https://doi.org/10.1073/pnas.1121181109

Warner, N.R., Kresse, T.M., Hays, P.D., Down, A., Karr, J.D., Jackson, R.B., Vengosh, A., 2013. Geochemical and isotopic variations in shallow groundwater in areas of the Fayetteville Shale development, north-central Arkansas. Applied Geochemistry 35, 207–220.

Wen, T., Agarwal, A., Xue, L., Chen, A., Herman, A., Li, Z., Brantley, S.L., 2019a. Assessing changes in groundwater chemistry in landscapes with more than 100 years of oil and gas development. Environmental Science: Processes & Impacts 21, 384–396. https://doi.org/10.1039/C8EM00385H

Wen, T., Castro, M.C., Nicot, J.-P., Hall, C.M., Larson, T., Mickler, P., Darvari, R., 2016. Methane Sources and Migration Mechanisms in Shallow Groundwaters in Parker and Hood Counties, Texas—A Heavy Noble Gas Analysis. Environmental Science & Technology 50, 12012–12021. https://doi.org/10.1021/acs.est.6b01494

Wen, T., Castro, M.C., Nicot, J.P., Hall, C.M., Pinti, D.L., Mickler, P., Darvari, R., Larson, T., 2017. Characterizing the Noble Gas Isotopic Composition of the Barnett Shale and Strawn

Group and Constraining the Source of Stray Gas in the Trinity Aquifer, North-Central Texas. Environmental Science and Technology 51, 6533–6541. https://doi.org/10.1021/acs.est.6b06447

Wen, T., Niu, X., Gonzales, M., Zheng, G., Li, Z., Brantley, S.L., 2018. Big Groundwater Data Sets Reveal Possible Rare Contamination Amid Otherwise Improved Water Quality for Some Analytes in a Region of Marcellus Shale Development. Environmental Science & Technology 52, 7149–7159. https://doi.org/10.1021/acs.est.8b01123

Wen, T., Woda, J., Marcon, V., Niu, X., Li, Z., Brantley, S.L., 2019b. Exploring How to Use Groundwater Chemistry to Identify Migration of Methane near Shale Gas Wells in the Appalachian Basin. Environmental Science & Technology acs.est.9b02290. https://doi.org/10.1021/acs.est.9b02290

Woda, J., Wen, T., Lemon, J., Marcon, V., Keeports, C.M., Zelt, F., Steffy, L.Y., Brantley, S.L., 2020. Methane concentrations in streams reveal gas leak discharges in regions of oil, gas, and coal development. Science of The Total Environment 737, 140105. https://doi.org/10.1016/j.scitotenv.2020.140105

Woda, J., Wen, T., Oakley, D., Yoxtheimer, D., Engelder, T., Castro, M.C., Brantley, S.L., 2018. Detecting and explaining why aquifers occasionally become degraded near hydraulically fractured shale gas wells. Proceedings of the National Academy of Sciences 115, 12349–12358. https://doi.org/10.1073/pnas.1809013115

Yang, H., Flower, R.J., Thompson, J.R., 2013. Shale-gas plans threaten China's water resources. Science 340, 1288. https://doi.org/10.1126/science.340.6138.1288-a

Figures

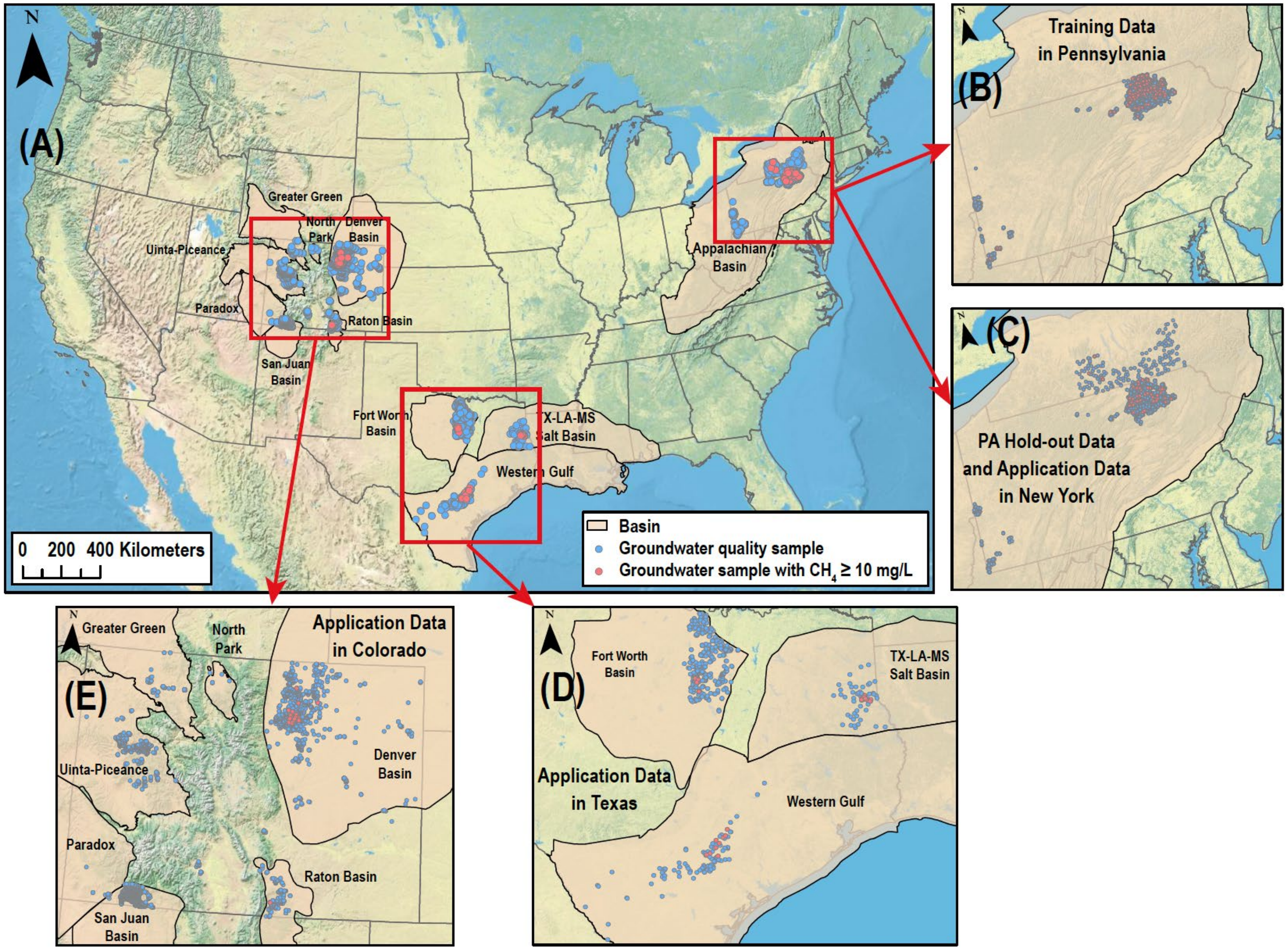

Figure 1. Location of groundwater quality data used in the model training and application in this study. This map layer of basins producing shale gas is adapted from the map of U.S. Energy Information Administration (https://www.eia.gov/maps/maps.htm).

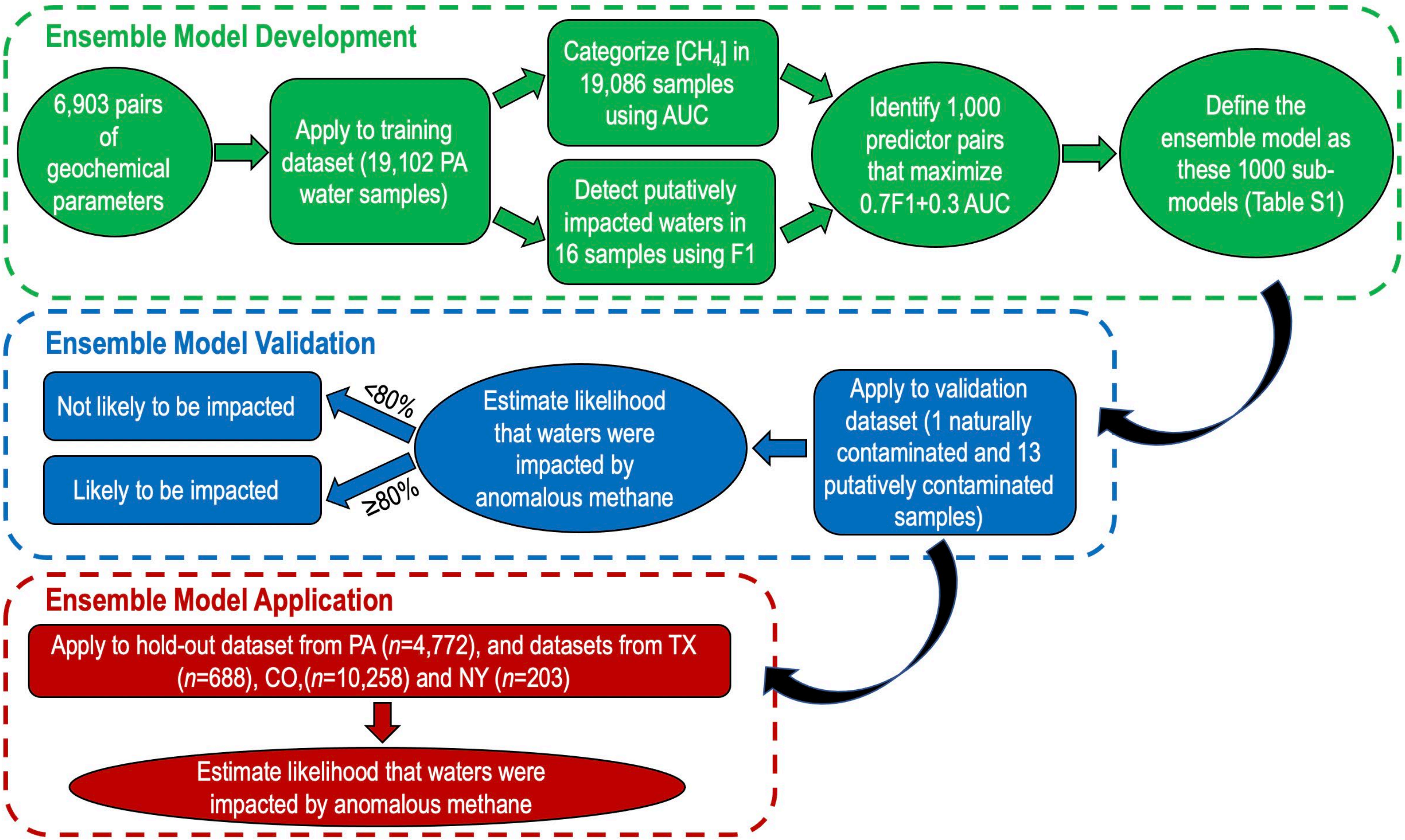
Ensemble Model Development
6,903 pairs of geochemical parameters
Apply to training dataset (19,102 PA water samples)
Categorize [CH_4] in 19,086 samples using AUC
Detect putatively impacted waters in 16 samples using F1
Identify 1,000 predictor pairs that maximize 0.7F1+0.3 AUC
Define the ensemble model as these 1000 sub-models (Table S1)
Ensemble Model Validation
Not likely to be impacted
Likely to be impacted
<80%
≥80%
Estimate likelihood that waters were impacted by anomalous methane
Apply to validation dataset (1 naturally contaminated and 13 putatively contaminated samples)
Ensemble Model Application
Apply to hold-out dataset from PA (*n*=4,772), and datasets from TX (*n*=688), CO,(*n*=10,258) and NY (*n*=203)
Estimate likelihood that waters were impacted by anomalous methane

Figure 2. Workflow of development, validating, and application phases of the ensemble model proposed in this study.

)

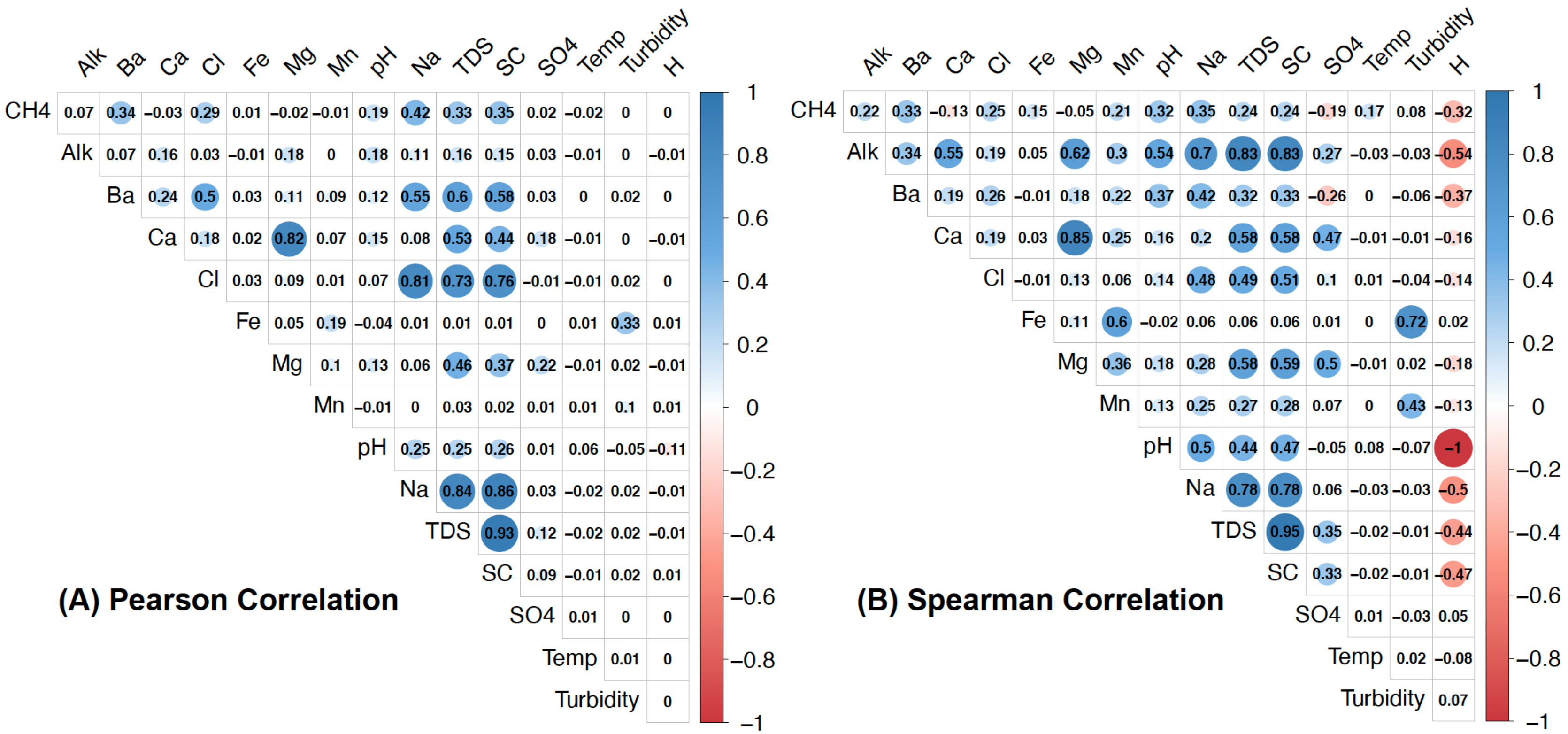

Figure 3. Correlation matrices for the 15 single geochemical features and the target feature (i.e., CH_4) determined for the training data derived from the Shale Network database (*n*= 11,875) for: (A) Pearson correlation and (B) Spearman's rank correlation. The pairwise correlation coefficient is indicated in the corresponding cell. A statistically significant correlation is highlighted by either blue (positive) or red (negative) colors.

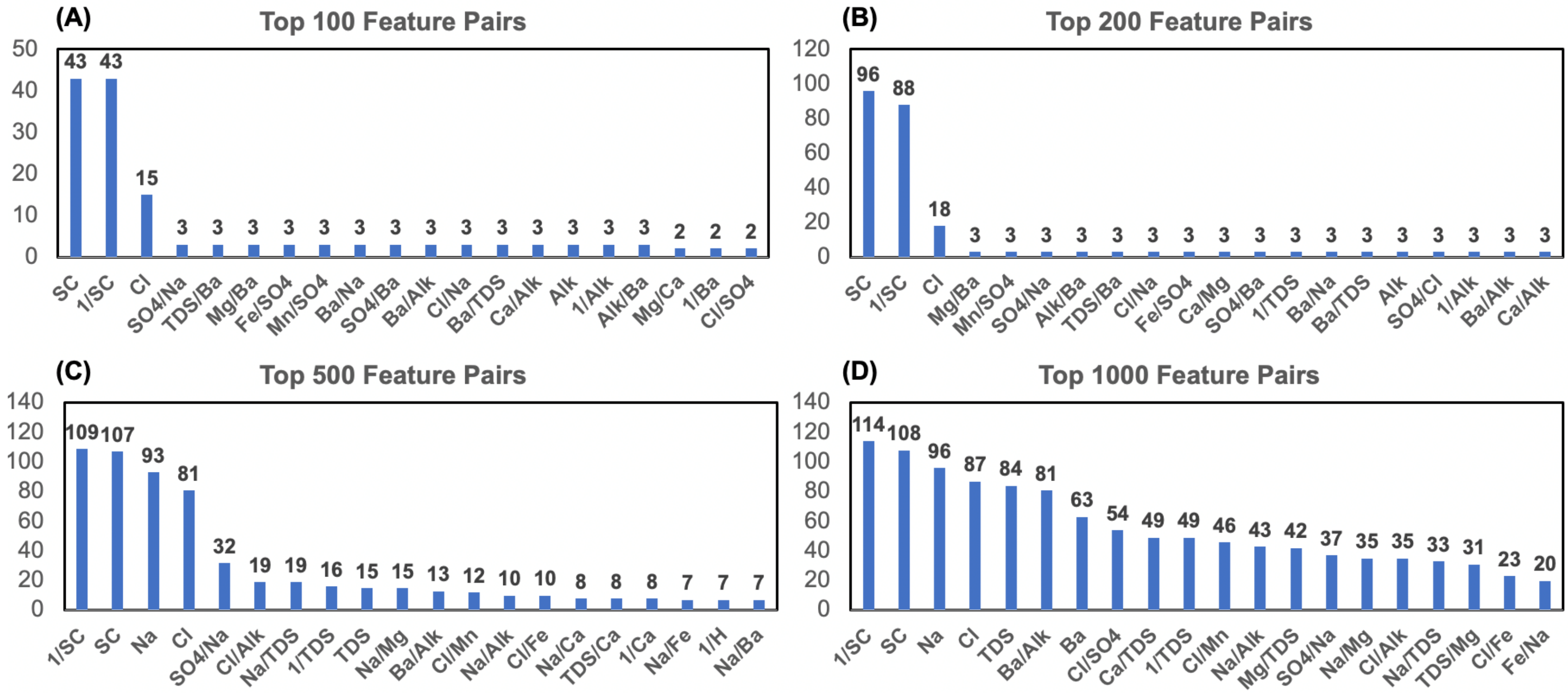

Figure 4. Frequency of the top 20 features in the best-performing models with respect to both the prediction of methane concentration and the detection of putatively contaminated samples in the model training. Feature frequency is shown for (A) top 100 models; (B) top 200 models; (C) top 500 models; and (D) top 1,000 models. The y-axis summarizes the frequency of each feature.

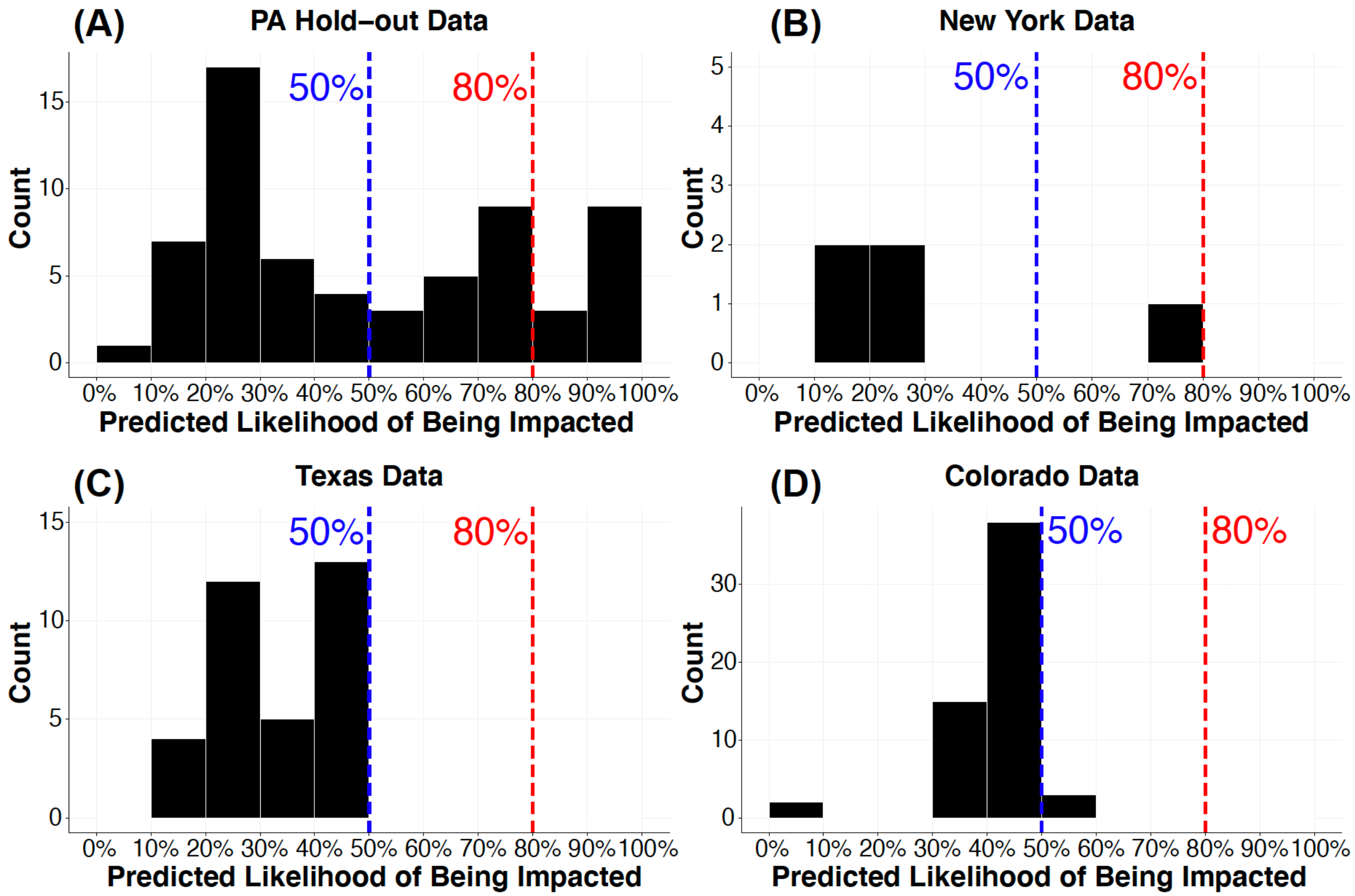

Figure 5. Distribution of predicted likelihood of being impacted by anomalous methane for high methane samples (≥ 10 mg/L) across the U.S. by considering the top 1000 sub-models: (A) Pennsylvania hold-out data, (B) New York data, (C) Texas data, and (D) Colorado data.

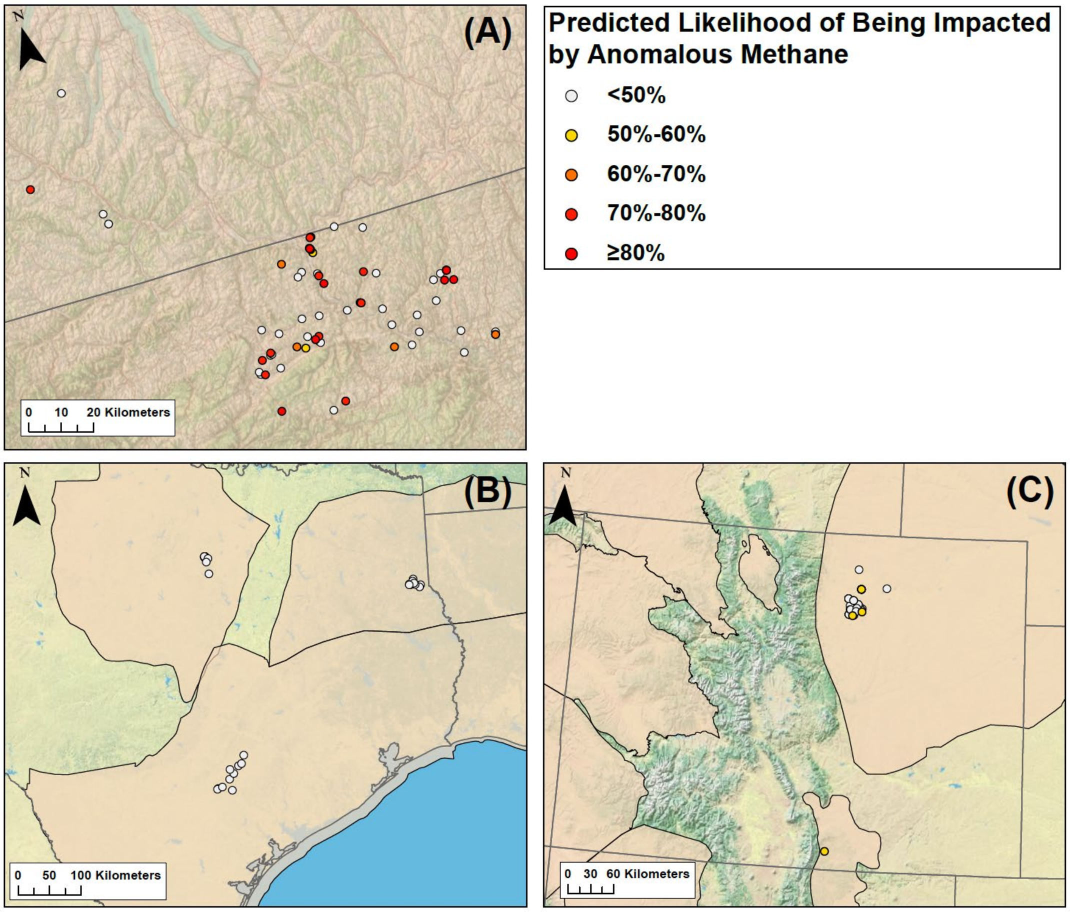

Figure 6. Maps showing locations and the range of predicted likelihood of being impacted by anomalous methane for high methane samples (≥ 10 mg/L) across the U.S. by considering the top 1000 sub-models: (A) Pennsylvania and New York, (B) Texas, and (C) Colorado.

Tables

Table 1. Overview of the datasets for model development and application

Dataset	**N_1[a]**	**N_2[b]**	**N_2/N_1**	**N_3[c]**	**N_3/N_2**	**N_4[c]**	**N_4/N_1**
Model Development							
Shale Network - training	19086	11875	62.22%	258	2.17%	390	2.04%
Known problematic sites - training	16	16	100%	14	87.50%	14	87.50%
Model Validation							
Known problematic sites	13	13	100%	13	100%	13	100%
Model Application							
Shale Network - held-out	4772	2969	62.22%	64	2.16%	98	2.05%
New York	203	78	38.42%	5	6.41%	7	3.45%
Colorado	10258	457	4.46%	58	12.69%	756	7.37%
Texas	688	338	49.13%	34	10.06%	40	5.81%

[a] N_1 refers to the number of groundwater samples with reported values for at least one predictor feature

[b] N_2 refers to the number of groundwater samples with reported values for all predictor features

[c] N_3 refers to the number of groundwater samples with methane concentration at least 10 mg/L and also meet the definition of N_2

[c] N_4 refers to the number of groundwater samples with methane concentration at least 10 mg/L regardless of whether they had all the analytes

Table 2. List of geochemical features used in machine learning models to predict methane concentration in groundwater

Determinand		**Single Geochemical Feature (Y/N)**	**Reciprocal Feature (Y/N)**	**Ratio Feature (Y/N)**
Bicarbonate Alkalinity	Alk	Y	Y	Y
Calcium	Ca	Y	Y	Y
Chloride	Cl	Y	Y	Y
Magnesium	Mg	Y	Y	Y
Sodium	Na	Y	Y	Y
Sulfate	SO4	Y	Y	Y
Total Dissolved Solids	TDS	Y	Y	Y
Barium	Ba	Y	Y	Y
Iron	Fe	Y	Y	Y
Manganese	Mn	Y	Y	Y
pH	pH	Y	N	N
Hydrogen Ion [a]	H	Y	Y	N
Specific Conductance	SC	Y	Y	N
Temperature	T	Y	N	N
Turbidity	Turbidity	Y	Y	N

[a] H^+ concentration is calculated from pH

Table 3. Predicted likelihood of groundwater samples being impacted by anomalous methane for four scenarios: (1) top 100, (2) top 200, (3) top 500, and (4) top 1000 sub-models

Dataset	Likelihood	Top 100	Top 200	Top 500	Top 1000
Model Validation					
Putatively contaminated sites - testing (n=13)	0%–20%	0	0	0	0
	20%–40%	0	0	0	0
	40%–60%	0	0	0	0
	60%–80%	0	0	0	0
	80%–100%	13	13	13	13
Model Application					
Shale Network - hold-out (n=64)	0%–20%	34	35	29	5
	20%–40%	5	2	4	26
	40%–60%	10	14	9	6
	60%–80%	6	0	9	14
	80%–100%	9	13	13	13
Texas (n=34)	0%–20%	22	34	17	4
	20%–40%	12	0	17	17
	40%–60%	0	0	0	13
	60%–80%	0	0	0	0
	80%–100%	0	0	0	0
Colorado (n=58)	0%–20%	41	57	23	2
	20%–40%	17	1	34	15
	40%–60%	0	0	1	41
	60%–80%	0	0	0	0
	80%–100%	0	0	0	0
New York (n=5)	0%–20%	4	4	3	2
	20%–40%	0	0	1	2
	40%–60%	1	1	0	0
	60%–80%	0	0	1	1[a]
	80%–100%	0	0	0	0

[a] Refer to the text for details. No oil or gas wells found within 5 km of this sample.