Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

journal homepage: www.elsevier.com/locate/jconhyd

Searching for anomalous methane in shallow groundwater near shale gas wells

Zhenhui Li [a,*], Cheng You [b], Matthew Gonzales [c], Anna K. Wendt [c], Fei Wu [a], Susan L. Brantley [c,*]

[a] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, United States
[b] Department of Statistics, The Pennsylvania State University, University Park, PA 16802, United States
[c] Earth and Environmental Systems Institute and Department of Geosciences, The Pennsylvania State University, University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 22 June 2016
Received in revised form 22 October 2016
Accepted 26 October 2016
Available online 28 October 2016

Keywords:
Hydraulic fracturing
Methane
Groundwater
Data mining

ABSTRACT

Since the 1800s, natural gas has been extracted from wells drilled into conventional reservoirs. Today, gas is also extracted from shale using high-volume hydraulic fracturing (HVHF). These wells sometimes leak methane and must be re-sealed with cement. Some researchers argue that methane concentrations, *C*, increase in groundwater near shale-gas wells and that "fracked" wells leak more than conventional wells. We developed techniques to mine datasets of groundwater chemistry in Pennsylvania townships where contamination had been reported. Values of *C* measured in shallow private water wells were discovered to increase with proximity to faults and to conventional, but not shale-gas, wells in the entire area. However, in small subareas, *C* increased with proximity to some shale-gas wells. Data mining was used to map a few hotspots where *C* significantly correlates with distance to faults and gas wells. Near the hotspots, 3 out of 132 shale-gas wells (~2%) and 4 out of 15 conventional wells (27%) intersect faults at depths where they are reported to be uncased or uncemented. These results demonstrate that even though these data techniques do not establish causation, they can elucidate the controls on natural methane emission along faults and may have implications for gas well construction.

1. Introduction

In the U.S.A., the usage of natural gas has increased markedly due to new techniques in developing gas directly from shale. Since 2014, this so-called "unconventional" resource has been estimated to comprise about 50% of the total proven U.S.A. gas reserves (U.S. Energy Information Administration, 2014). Extraction of gas from shale has become possible due to improvements in directional drilling and high-volume hydraulic fracturing (HVHF) (Vidic et al., 2013a). The rapid development in the use of HVHF in the U.S.A. since the 1990s has sometimes led to water quality impacts that have caused concern, including leakage of methane out of gas wells due to well integrity problems (Vidic et al., 2013a). Such problems have been particularly controversial in the Marcellus gas play because this shale formation underlies 8 highly populous northeastern states. One state regulator, the Pennsylvania Department of Environmental Protection (PA DEP), reported, for example, that the most common type of water quality impact related to oil/gas activity by companies developing "unconventional" wells – i.e. wells completed with HVHF – is methane contamination (Brantley et al., 2014). The frequency of well integrity problems (Brantley, 2014) for wells completed with or without HVHF – i.e., "fracking" – is important given that leakage into drinking water resources entails explosion hazards when concentrations approach 10 ppm and methane in groundwater can result in secondary contamination (Vidic et al., 2013a). In addition, eventual release of methane into the atmosphere increases greenhouse warming (Howarth et al., 2011). According to PA DEP records, 3.4% of gas wells were cited for well construction problems before 2013 (Vidic et al., 2013a). Of these, 16 wells (0.24%) were cited for allowing gas to migrate into groundwater. This methane leakage rate in the Marcellus play may have changed with time as operators learned better practices (Brantley, 2014). However, the leakage rate, which is difficult to quantify, has become controversial for HVHF because some have claimed that natural gas leaks more readily from wells in unconventional formations than from "conventional" wells (Howarth et al., 2011; Ingraffea et al., 2014).

One way to investigate leakage is to determine if the concentration of methane in groundwater, *C*, varies with distance from gas wells. However, such studies depend on how many water wells are investigated. For example, an early investigation concluded that *C* increased in ~60 waters sampled within 1 km of unconventional wells located in Pennsylvania (U.S.A.) (Osborn et al., 2011). This claim has been disputed (Davies, 2011; Molofsky et al., 2011; Schon, 2011; Jackson et al., 2011; Molofsky et al., 2013) at least partly because *C* can be high in groundwater due to natural processes (Reese et al., 2014; Baldassare et al., 2014). In a second investigation of the same area with 141 samples, *C* once again was observed to increase near gas wells (Jackson et al., 2013a). However, both the original and extended studies included observations

[*] Corresponding authors.
E-mail addresses: jessieli@ist.psu.edu (Z. Li), sxb7@psu.edu (S.L. Brantley).

http://dx.doi.org/10.1016/j.jconhyd.2016.10.005
0169-7722/© 2016 Elsevier B.V. All rights reserved.

around Dimock PA where investigators have concluded a few gas wells contaminated 18 water supply wells in the early days of shale gas development (Brantley et al., 2014). In contrast, an analysis of >11,000 water samples in northeastern PA revealed no correlation between *C* and proximity to unconventional wells (Siegel et al., 2015). Apparently, higher *C* values may have been present in the larger dataset but not detectable because of the large number of non-impacted groundwater samples. None of these datasets have been released in entirety because of concerns about homeowner confidentiality.

In this paper we analyze a newly published data set from the PA DEP (1690 water samples from shallow, private water wells) from Bradford County, Pennsylvania (Shale Network, 2015) to learn how to interpret environmental datasets of different size. We hypothesized that large datasets, on average, might mask contamination that could be observed in smaller datasets. We also sought to understand the importance of conventional versus unconventional wells and the effect of local geology on methane emission. Our analysis focused on five townships where impacts to groundwater from methane were reported. We developed strategies to use large groundwater datasets to highlight and understand possible sites of methane emission with respect to local conditions. Our new technique relies on the use of large datasets and should be broadly applicable to other environmental data where patterns in distribution of contamination may allow for better environmental practices.

2. Analyzed data

To determine environmental patterns using data mining requires the availability of large numbers of analyses. Large datasets generally require that environmental data be pooled from many sources. The strategy of using large datasets and data mining is therefore predicated on the assumption that fundamental patterns can be gleaned from large datasets even though such sets may be characterized by variable data quality. We implicitly test that proposition here.

The water samples we analyze were collected by independent environmental consultants paid by gas companies before drilling and measured in commercial analytical laboratories that support extensive quality control and assurance measures (see Suppl. Information). The analyses are released to the state regulator to protect the gas company from future liability if water issues are reported. Given this end use, biasing samples or analyses toward lower methane concentrations (for example by allowing volatilization) is likely to be counter-productive.

Water samples were collected prior to treatment, filtration or water softening using U.S. Geological Survey protocol. Samples were collected and analyzed in accordance with Pennsylvania code § 78.52 which states, "(c) The survey shall be conducted by an independent certified laboratory. A person independent of the well owner or well operator, other than an employee of the certified laboratory, may collect the sample and document the condition of the water supply, if the certified laboratory affirms that the sampling and documentation is performed in accordance with the laboratory's approved sample collection, preservation and handling procedure and chain of custody."

Following a data sharing agreement between PA DEP and Pennsylvania State University, we analyzed data from Bradford County for five townships (Fig. 1). No attempt was made to analyze variation in *C* with time at each location because very few water wells were sampled more than once. The waters were sampled from water wells (average depth 54 m; ranging from 2 to 250 m) before drilling the new gas wells over a period of a few years. However, because the water sampling generally occurred near already-drilled gas wells, the data were investigated here with respect to gas wells that had already been drilled in conventional or unconventional formations. Each water analysis (i.e., sample site) was paired with the closest *previously drilled* unconventional well using data on the PA DEP Oil and Gas Reporting Website as of April 2015 (Murphy, 2012). Distances were determined for the closest already-drilled well (i.e., spud date prior to water sampling) within both the targeted and nearby townships. Of the original 1240 unconventional wells considered for the region, sample sites were paired with 132 unconventional gas wells spudded from June 2008 through July 2012. Likewise, of the 113 conventional gas wells in the overall region, samples were paired with 15 conventional wells: 13 spudded between 1932 and 1983 but now abandoned, and two spudded in 2009 and still active. The number of analyses and wells included in this dataset is intermediate between the previously discussed published datasets (Osborn et al., 2011; Jackson et al., 2013a; Siegel et al., 2015) and this allowed us to test how the size of the dataset affects conclusions about methane migration. In addition, the dataset reported here is the only one published with locations (Shale Network, 2015). More details are described in the Suppl. information (SI).

3. Methods

We analyzed the full dataset and then used increasingly finer spatial resolution by employing the following steps. First, we plotted *C* versus the distance to the nearest already-drilled unconventional or conventional well for the entire dataset. We quantified the correlation between *C* (i.e., dependent variable y) and distance (i.e., independent variable x). However, many statistical measures are not applicable because of the multiple reporting limits (i.e., detection limits) (Siegel et al., 2015; Helsel, 2011). For example, Pearson correlation and linear regression are not suitable; furthermore, Spearman correlation is only suitable for data with one reporting limit. Therefore, we used three measures that are appropriate for censored data with multiple reporting limits: Kendall rank correlation, Akritas-Theil-Sen (ATS) regression, and logistic regression (see SI for more details).

We next subdivided our study area into three subregions (A, B and C), which were selected to produce three clusters largely delineated by townships, each with at least 350 analyses (Fig. 1). The correlation statistics were then re-calculated for samples collected within each subregion.

To learn to analyze subregions of these environmental data randomly, we then developed a new sliding window approach inspired by the spatiotemporal exploratory model (Fink et al., 2010). We scanned the whole region using a "sliding window" of size 5 km × 5 km that was stepped over the map in 200 m increments. For each sliding window observation at each location separated by 200 m, we tested for Kendall rank correlation for the data in the window. The window was marked as $+1$ if the correlation is significantly positive and -1 if significantly negative (significance level of 5%). A spatially-normalized significance value was assigned to each location as defined by the sum of all windows covering the location divided by the total number of windows covering the location. The spatially-normalized significance values, plotted every 200 m, were then used to generate correlation maps showing regions of higher positive or negative correlations.

With the correlation maps, we explored the relationship between hot spots and the underlying geologic structure using maps of known faults in the area. The hot spots are the locations showing negative correlations between methane concentration and distance to well, i.e., higher dissolved methane concentrations closer to the well. For the wells located near hot spots we also investigated the well characteristics (e.g., casing and cementing). Finally, because all the unconventional wells had not been hydraulically fractured by the time of water sampling, we also repeated our methodology on the subset of wells that were completed by HVHF prior to water sampling.

4. Results and discussion

Fig. 2 shows scatter plots of *C* versus distance to the nearest already-drilled unconventional well before and after log transforming the data. These plots are visually misleading because a high percentage of samples cluster near the reporting limits of 1, 5 and 26 ppb (Siegel et al., 2015). A binned plot of the same data (Fig. 3) documents that such

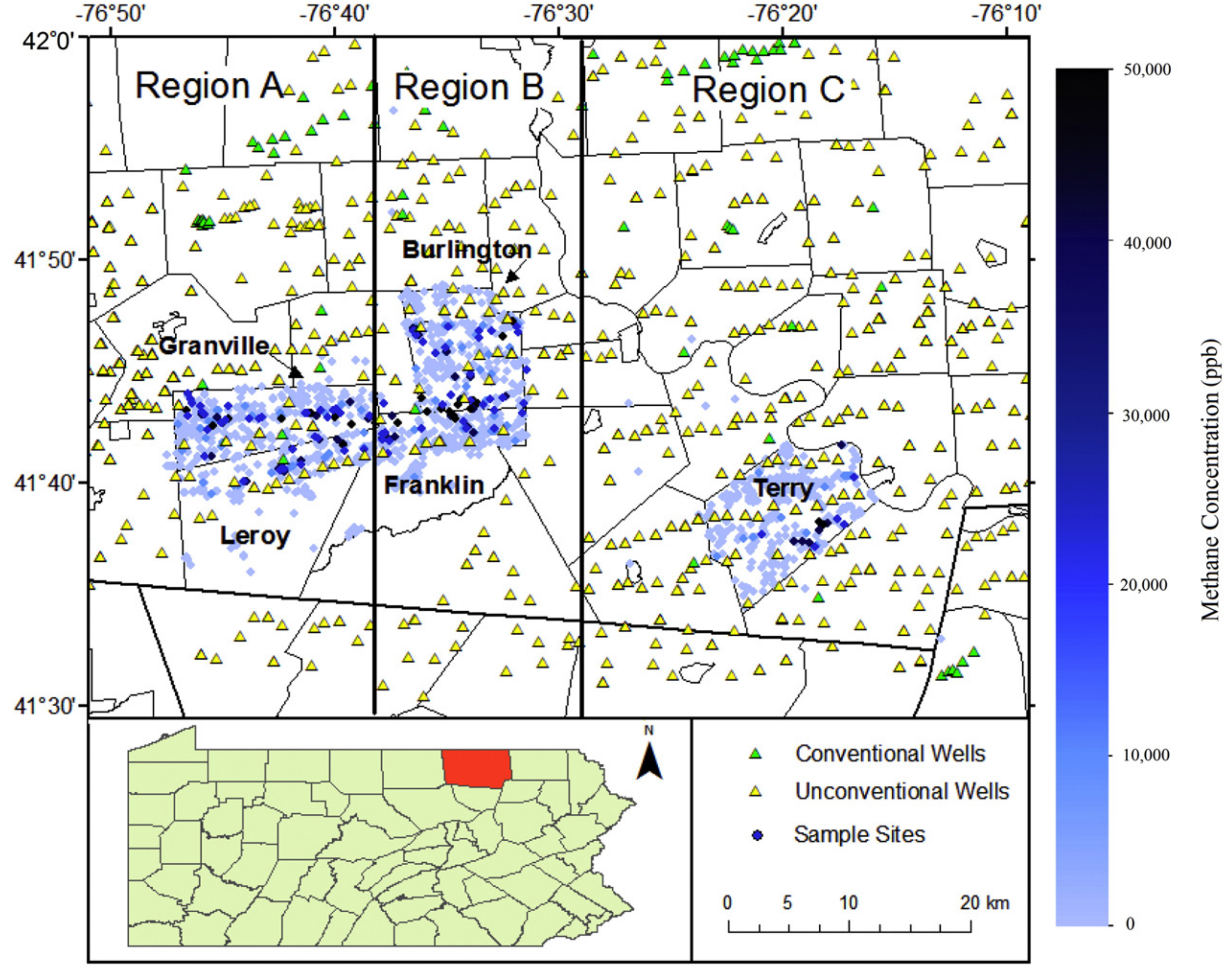

Fig. 1. Location map for the study area in Bradford County in north-central PA (U.S.A.). Site locations for 1690 water samples are shown as blue symbols for the five townships (labelled). Green and yellow triangles represent the locations of conventional (15) and unconventional (1240) gas wells, respectively, as of April 2015. Water samples were collected from 12/20/2010 to 11/23/2012. Symbols range from light blue, indicating *C* not above the reporting limit, to the darkest blue, indicating the highest $C = 46$ ppm CH_4. All data have been published (Shale Network, 2015). The study region was also partitioned into three subregions cut off at longitudes −76.65 (−76 deg 39′) and −76.5 (−76 deg 29′) as shown: 579 samples in Region A, 730 in B, and 381 in C. All data released to us by PA DEP for the townships are shown, including a few water samples released mistakenly for other townships.

low methane concentrations are found at all distances from gas wells but values of $C > 6000$ ppb CH_4 are only identified in sites located within the 3 km perimeter of an unconventional well. The broad spatial distribution of low values of *C* in the study area is consistent with many sites of emission of methane due to natural processes (e.g., Molofsky et al., 2011; Molofsky et al., 2013). As described below and in Methods, we tested several hypotheses to understand the distribution.

4.1. Concentrations versus gas wells for the whole region

All three statistical measures show *C* decreases with distance from an unconventional well – but none of the negative correlations are statistically significant at the 95% confidence level (i.e., p-values > 0.05): Kendall rank correlation value is −0.01537 with p-value = 0.32632; ATS slope is −0.00523 with p-value = 0.32635; the logistic regression fitted coefficient (in distance) is −0.00012 with p-value = 0.05305. Like the conclusions reached previously for such large datasets (Siegel et al., 2015), *C* does not correlate with proximity to unconventional wells when using all 1690 samples.

In contrast to unconventional wells, for the 15 conventional wells associated with water samples, all three statistics yield significant correlations indicating *C* increases with proximity to the well (i.e., p-values < 0.05): Kendall rank correlation value is −0.06398 with p-value = 4.43672e−05; ATS slope is −0.01402 with p-value = 4.43672e−05; the logistic regression fitted coefficient (in distance) is −0.00012 with p-value = 2.14783e−05. This is discussed further below.

4.2. Correlation versus unconventional wells in subregions

After partitioning the study area into three subregions, only subregion B shows significant correlations with proximity to unconventional wells for all three measures (Table 1). The significant negative correlations indicate higher *C* closer to unconventional wells. For every 100 m closer to a pre-existing unconventional well, the dissolved methane concentration is 4 ppb higher in the groundwater samples. While the increase in *C* does not come close to superseding regulatory recommendations for action, correlations with respect to proximity to gas wells are analyzed here to develop machine-learning strategies to find problematic wells. The technique shows that significant correlations exist in subregions that are masked in the larger region.

It is possible that a different regional partitioning might not reveal correlations. This is why we developed a new approach (see Methods) to test this by using sliding windows to calculate correlation maps (Fig. 4). Such “heat maps,” as they are often named, indicate locations with dark coloring (red in Fig. 4a) where the majority of the windows covering those locations show significant correlations. Here, the red coloring indicates correlations that are negative and that indicate *C*

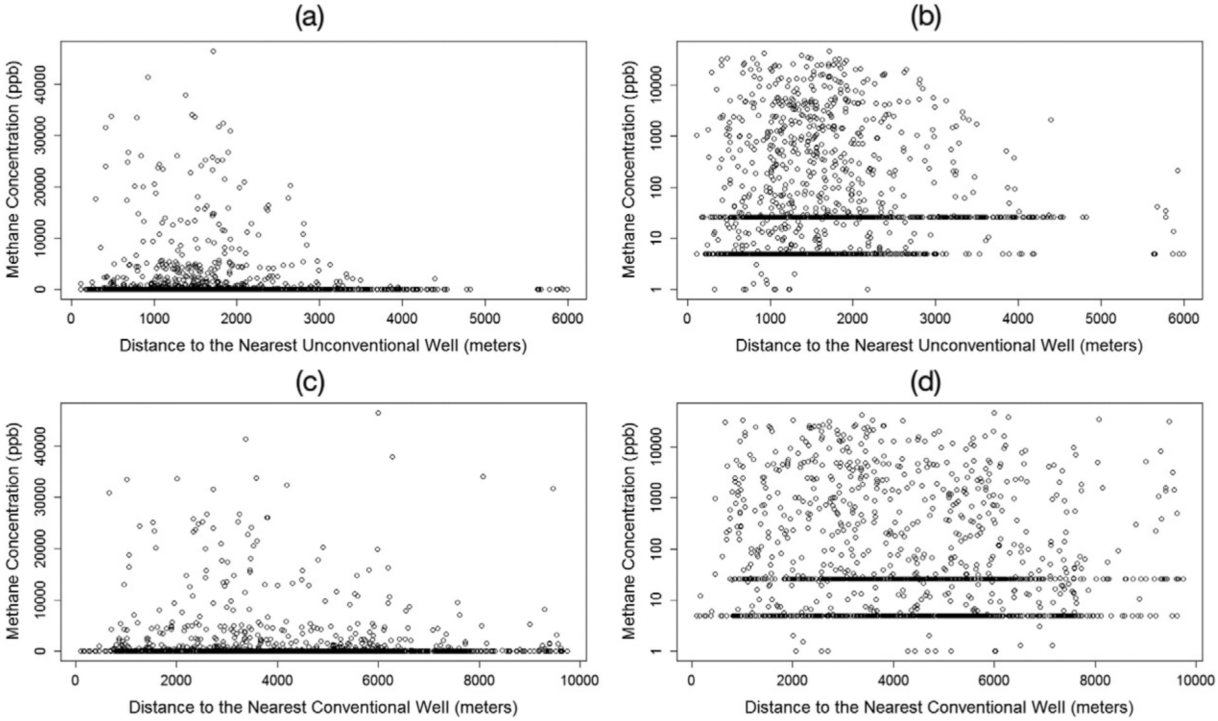

Fig. 2. Plots of methane concentrations vs. distance to the nearest already-drilled unconventional gas well (a, b) or conventional well (c, d). All 1690 methane analyses for samples collected between 12/20/2010 and 11/23/2012 are plotted (see also Fig. 1). Log-transformed data plots (b, d) emphasize that methane was often not detected above one of three reporting limits depending upon the commercial laboratory chosen for analysis: 1 ppb (11 samples), 5 ppb (512), and 26 ppb (575).

increases nearer gas wells. Consistent with Table 1, the maps document that only subregion B shows significant negative correlations.

These maps are not made with a contouring algorithm. Colors show averages for each point located every 200 m: averages were calculated over all the correlations for the 2500 windows that included each point. The colors show correlations only when they are calculated to be statistically significant for a given window. Correlations depend upon both the values of C and their locations as well as the density of data points.

The sliding window technique (Fig. 4a) documents the most likely reason why previous researchers (Jackson et al., 2013a; Siegel et al., 2015) came to discrepant conclusions: they used datasets that were vastly different in size. The correlations between methane and shale gas wells in this sample of data are infrequent and localized and the incidence of problems is not statistically detectable in very large datasets without using a data mining technique.

The sliding window technique was also applied to the ~1600 water analyses to determine spatially significant correlations with the 15

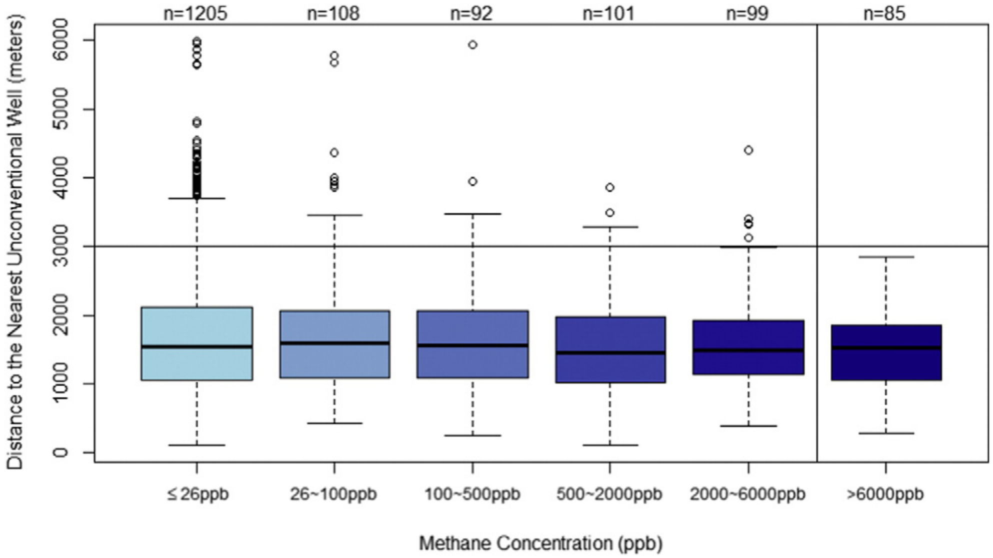

Fig. 3. Boxplots showing binned methane concentrations (x axis) plotted versus distance to the nearest already-spudded unconventional gas well (y axis). Each box shows non-outlier data for the distance to gas well: the minimum, first quartile, median, third quartile and maximum of all distance values are indicated as horizontal bars from bottom to top, respectively. Symbols indicate outliers, i.e., measurements within the range $(Q_3, Q_3 + 1.5 \times IQR)$, where Q_3 is the third quartile and IQR is the interquartile range ($IQR = Q_3 - Q_1$). Very low concentrations (lighter blue boxes) are observed at all distances from gas wells. High concentrations are only observed within 3 km of wells (i.e., the horizontal line).

Table 1
Correlation statistics[a] for subregions.

	Region A	Region B	Region C
Kendall rank correlation	0.01943	−0.06121*	0.01169
Akritas-Theil-Sen (ATS)	0.004864	−0.03912*	0.01156
Logistic regression	−0.00001	−0.00028**	0.00011

[a] Each correlation measures the relationship between distance to the nearest already-drilled unconventional well (the x-variable in meters) and dissolved methane concentration (the y-variable in ppb). For example, the negative ATS correlation coefficient in region B of −0.03912 suggests that on average, for every 100 m closer to a pre-existing unconventional well, the dissolved methane concentration is 4 ppb higher in the groundwater samples. No waters approached recommended action limits (see SI and discussion of Fig. S1).

[*] Indicates 5% significance level.

[**] Indicates 1% significance level.

conventional gas wells. Once again, areas of both positive and negative correlation were observed and are discussed below.

4.3. Concentrations versus faults

Although *C* varies with proximity to gas wells, correlation does not prove causation. In northern PA, for example, faults sometimes channelize naturally emitting methane into groundwater – completely unrelated to oil/gas development (Molofsky et al., 2011; Molofsky et al., 2013; Reese et al., 2014; Baldassare et al., 2014; Llewellyn, 2014). This gas often derives from the underlying Upper Devonian Catskill and Lock Haven Formations (Molofsky et al., 2013). Biological sources also emit methane naturally (Baldassare et al., 2014).

We identified the large faults in the study area (Pohn and Purdy, 1981; Faill, 1998). These faults are low angle, east-west striking, northwest- or southeast-dipping. Negative correlations were calculated between *C* and distance to faults across the entire area (logistic regression: −0.00006, p-value 0.01472; Kendall and ATS have p-values > 5%). Thus, *C* shows a tendency to increase in waters near fault traces but remain well below regulatory recommendations. Since the thrust faults generally outcrop along valleys (Fig. 4c), these results are consistent with other research showing that *C* is higher in the lowlands of northeastern PA (Molofsky et al., 2011; Llewellyn, 2014).

We also applied the sliding window method to faults (Fig. 4c). Almost all the colors and locations of hotspots on the fault map are also seen on the unconventional well map, although with consistently higher values of significance in Fig. 4c than Fig. 4a. One possible explanation for this observation is that methane emits naturally along faults, and that gas wells are somewhat aligned along fault-parallel valleys and ridgelines. In this case, the explanation for the overall stronger correlations in Fig. 4c as compared to Fig. 4a is that *C* slightly increases in proximity to unconventional wells not because gas wells leak but because gas wells are preferentially located near faults that are natural zones of gas emission.

As a partial test of this idea, we analyzed correlations between methane and inorganic water chemistry. Often, natural thermogenic methane in PA is accompanied by dissolved salts (Molofsky et al., 2011;

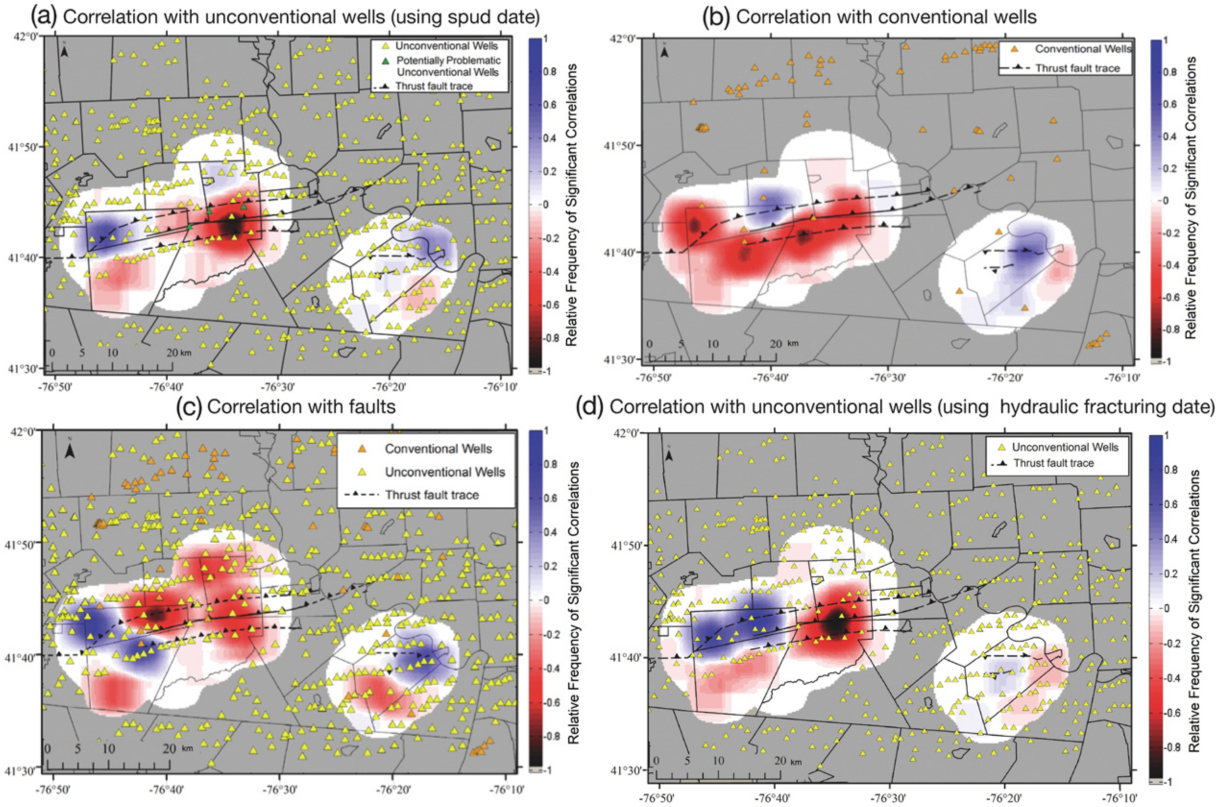
(a) Correlation with unconventional wells (using spud date)
Unconventional Wells
Potentially Problematic Unconventional Wells
Thrust fault trace
(b) Correlation with conventional wells
Conventional Wells
Thrust fault trace
(c) Correlation with faults
Conventional Wells
Unconventional Wells
Thrust fault trace
(d) Correlation with unconventional wells (using hydraulic fracturing date)
Unconventional Wells
Thrust fault trace
Relative Frequency of Significant Correlations
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
42°0'
41°50'
41°40'
41°30'
-76°50'
-76°40'
-76°30'
-76°20'
-76°10'
0 5 10 20 km

Fig. 4. Maps showing statistically significant correlations calculated for *C* in sliding windows for distance to: (a) unconventional wells, (b) conventional wells, (c) faults, (d) completed unconventional wells. Water samples were included in the analysis in (a) when sampling date was after spud date for an unconventional gas well, in (b) for all sampling dates, in (c) for all sampling dates because sampling was always after the spud date of the conventional wells, and in (d) when sampling date was after date of completion by HFHV. Color indicates locations with significant negative (red) or positive (blue) correlations between *C* and the distance to the nearest location of the independent variable. Regions with no water samples are colored grey. White indicates no significant correlations or # of significant positive correlations (blue) ≈ # of significant negative correlations. On the map, locations were defined as points separated by 200 m and correlations were calculated for 2500 sliding windows that included each point; therefore no contouring algorithm was used for coloration. The color intensity at a location shows the relative frequency of windows with significant correlation values that cover the location (see text and SI). In (a), three unconventional wells are shown in green that lack cemented casing at the depth interval where they intersect the nearby fault. Black solid lines indicate observed surface traces of thrust faults; dashed lines indicate faults extrapolated from depth. Triangles on lines indicate direction of dip. From north to south, the faults are the Bridge Street, Towanda, and three without names: a putative unnamed fault that runs along Towanda Creek in Bradford County, an unnamed thrust fault described in the literature (Llewellyn, 2014), and an additional unnamed fault (Pohn and Purdy, 1981; Woodrow, 1968; Berg et al., 1980; Vidic et al., 2013b).

Molofsky et al., 2013; Llewellyn, 2014; Warner et al., 2012) because thermogenic methane is dissolved in fault-channelized porewaters which are chemically similar to highly diluted Appalachian Basin brines (Dresel and Rose, 2010; Poth, 1962). The fault waters are also higher in pH and alkalinity and lower in dissolved oxygen (Reese et al., 2014; Llewellyn, 2014). As discussed in the SI, we observed significant correlations consistent with methane co-existing with salt-containing fluids along the faults.

4.4. Concentrations versus well completion

Potential issues related to correlations on Fig. 4a, b, c could be due to drilling rather than HVHF. In fact, in our analysis of unconventional wells so far, we paired each water sample with the closest already-drilled well (sampling date after spud date). However, it is not uncommon for an unconventional well to be drilled (i.e., spud date) long before completion with HVHF (completion date). For example, for the 1690 total water samples plotted in Fig. 4a, b, c, we found the dates of HVHF for 111 of the 132 gas wells (Fink et al., 2010; Llewellyn, 2014). To analyze our data with respect to completion date, we therefore limited our analysis to a water database that contained only 1497 of the 1690 originally reported water samples. Of these, 392 of these waters were sampled before HVHF.

With this smaller dataset of 1105 waters sampled after HVHF, we recalculated correlations and correlation maps (see SI and Fig. 4d). If hotspots of correlation observed in Fig. 4d are indicators of methane leakage, this leakage could have occurred after HVHF.

4.5. Using data mining to highlight potentially problematic wells

In several hotspots on the unconventional (Fig. 4a, d) and conventional (Fig. 4b) well maps, the significance of correlations is higher than on the fault map (Fig. 4c). We hypothesize that methane from gas wells near these hotspots may be leaking. The most prominent of these hotspots is in subregion B in the southern part of the intersection of Franklin and Burlington townships (Fig. 4a). To test whether the data mining technique might be useful in identifying problematic practices by gas companies, we inspected company reports of wells in those hotspots. Because we hypothesized that the correlations in Fig. 4b can be attributed to movement of methane along faults, we particularly sought to understand how the wells were constructed where the boreholes crossed the fault near the hotspot in subregion B.

Well construction issues with respect to casing and cementing are the most common reason for subsurface fugitive gas migration (Vidic et al., 2013a; Gorody, 2012). Three of the four operators in the study region (Fig. 5) have been cited for such well issues (Breiman et al., 1984). Casings are metal tubes that line a well from the surface to various depths, each with progressively smaller diameter. Conductor casing stabilizes the unconsolidated sediment. Surface casing, cemented along its entire length, prevents contamination of shallow groundwater. Intermediate casing isolates the borehole at intermediate depths where wells sometimes intersect hydrocarbon-bearing formations, abnormally pressured, or fractured zones (American Petroleum Institute, 2009). Production casing is cemented in the zone of gas production. Regulations for cement and casing vary around the world: in PA prior to 2011, intermediate casing was only required if aquifers were encountered below the depths of surface casing (Harvey Consulting LLC, 2009) or for some cases of blowout preventer support. If the borehole is left uncemented or uncased, thermogenic gas at intermediate depths (Molofsky et al., 2013; Baldassare et al., 2014) can enter or leave a section of a borehole and travel vertically to aquifers (Llewellyn et al., 2015).

Different operators use different techniques for casing and cementation. The operator that drilled the largest number of wells over the whole study region mostly did not emplace intermediate casings (Schoell, 1980). Three other operators in the area (Fig. 5) used surface (to ~600 ft (~180 m)), intermediate (~2000 ft (~610)) and production casings (~9550 ft (~2900 m)). We used online company-recorded data (Schoell, 1980) to find if any of the boreholes were reported to lack cemented or cased intervals at depths of fault intersection. We assumed

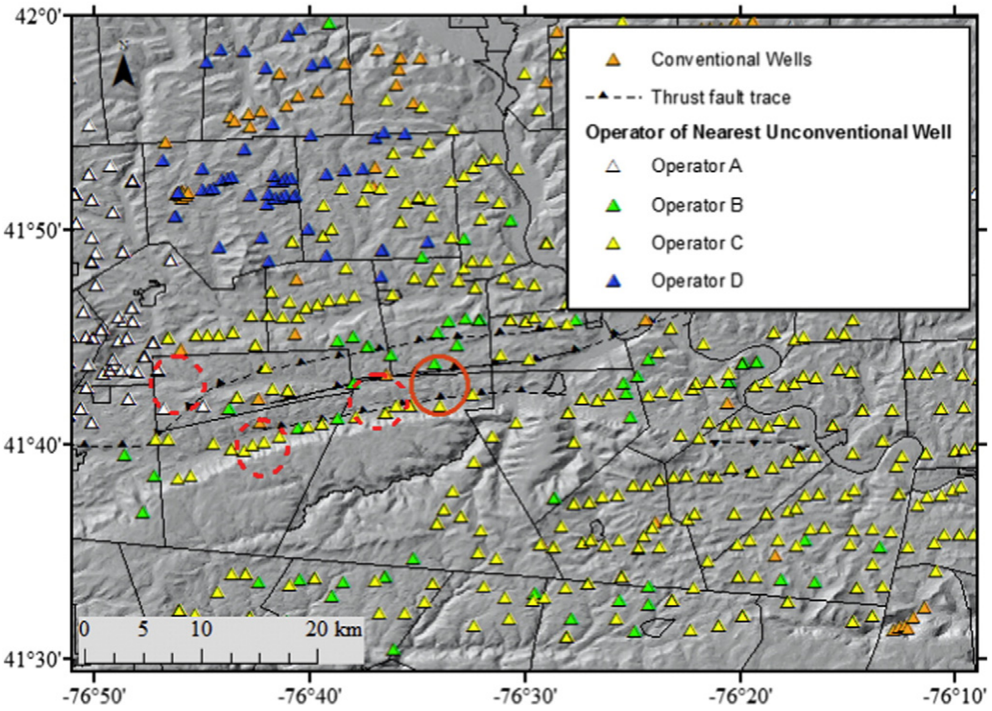

Fig. 5. Wells colored to indicate operators. The prominent hotspot from Fig. 4a is mapped onto a digital elevation map (10) as a red solid-line circle (at value 0.5) along with the three most intense hotspots in Fig. 4b in dashed red. Orange triangles show conventional wells. Unconventional wells drilled by different operators are indicated as described in the legend: operator A (associated with 165 water samples, 12 gas wells), B (400, 17), C (1115, 98), and D (10, 2).

the faults in Fig. 4 dip at 20–25° (Pohn and Purdy, 1981). This assumption was corroborated for one fault in Terry township where we could compare our geometric calculations based on the assumed dip (Table S3) with seismic data (Llewellyn et al., 2015).

Using this geometric analysis (Table S3), we identified 30 unconventional wells that are likely to intersect the faults in subregions A and B (Fig. 1). For 6 of these 30 wells near the most intense hotspot in Fig. 4a, we found no reports of casing or cementing across the depth of intersection with a fault. Of these, 3 are associated with water samples in blue areas on Fig. 4a near Granville Summit (see discussion in SI). Although they could be related to problems, the sliding window approach is unlikely to be helpful in pinpointing potentially problematic wells in such an area where C decreases near gas wells.

This left 3 gas wells, colored green in Fig. 4a, which fulfill two criteria: i) they are associated with water samples in the red hotspot and ii) they are not associated with reports of casing or cementing in the interval of intersection with the Towanda or next most southern fault (see Table S3: permit numbers 015-20116, 015-21353, 015-20612). The first and last of these were completed with HVHF before water sampling. One of the wells is reported to lack intermediate casing entirely (the middle green well in Fig. 4a (015-20116). For the other two wells, intermediate casing was not reported to extend to the depth of the fault. If the production casing in those two wells is cemented adequately across the fault-borehole intersection, or if management practices are being utilized to manage the casing pressure, then those wells are likely not allowing gas migration to aquifers. We found no public information to assess these possibilities (Schoell, 1980).

For the hotspot at the southern part of the intersection of Franklin and Burlington townships, we thus conclude that if any of the unconventional wells are leaking, the most likely are the wells colored green on Fig. 4a. In particular, gas could be moving in or along the boreholes and then moving into the fault and travelling updip to emit into groundwaters to the south. The unconventional wells highlighted by this analysis have not received well integrity violations from the regulator (Breiman et al., 1984); therefore, if emission is occurring, it could be continuing today. For example, the gas well without intermediate casing was spudded on 1/13/2009 and completed on 5/19/2009. The values of very slightly higher methane ($C > 5000$ ppb ($n = 7$)) were sampled in nearby waters between June 2011 and January 2012, 2 to 3 years after spudding.

We also similarly looked at company data for the very small set of conventional wells near hotspots in Fig. 4b: four were identified that do not report casing or cement at depths where they cross a fault (Table S4). Of these wells, 3 are abandoned. One of these was an orphan well drilled in 1932. Such older wells may be of particular note since methane is known to emit from some wells that were not completed according to modern standards (Kang et al., 2014). Consistent with this, C weakly correlates with time since drilling of these wells (Table S5, Fig. S4). Although the number of conventional wells in our target area is very small, our results point to the need for analysis of such legacy wells. A relatively large fraction of the ~350,000 oil and gas wells drilled in Pennsylvania could lack needed casing or cement (Pennsylvania Department of Environmental Protection, 2011).

4.6. Implications

Using a large publicly available dataset, we showed that methane concentrations in groundwaters in northeastern PA, C, tend to increase with proximity to unconventional wells in small subregions (730 analyses in ~100 km^2) although they always remain much, much lower than regulatory recommendations. The correlation with gas wells is not statistically significant when calculated over a larger region (~1000 km^2) with 1690 analyses. In contrast, C values show a small tendency across the entire region to increase with proximity to conventional wells, as well as to geological faults. Using groundwater datasets to understand the nature and frequency of methane impacts is thus characterized by a “Goldilocks” problem: datasets must be large but not too large since impacts are rare, but the size of the required dataset is not known before screening. This type problem is not unusual when analyzing regional environmental data.

To solve such problems of environmental data analysis, we developed a new “sliding window” technique that allows the researcher to find rare, localized spatial correlations without making biased or ad hoc choices. Using the technique we discovered that maps of correlations between C and proximity to gas wells or faults are similar. Where C appears to increase near gas wells in this study area, we therefore concluded that the correlation is mostly explained by the rough alignment of gas wells along faults where methane naturally emits. Our purely statistical approach therefore corroborates previous researchers who identified natural emission of methane along faults in PA (Molofsky et al., 2011; Molofsky et al., 2013; Reese et al., 2014; Baldassare et al., 2014; Llewellyn, 2014).

However, we also discovered one hotspot where C increases (while remaining well below regulatory recommendations) near unconventional gas wells and which does not appear on the map for faults. This map in turn led us to discover that cemented casing may be lacking at depths where 3 gas wells intersect a large fault near the hotspot. Hotspots on the correlation map for conventional wells are also centered near 4 gas wells that lack casing or cement where they intersect a fault. Three of these gas wells are now abandoned: this may be important because we also observed a very weak but significant increase in C with time since drilling of conventional wells. The fraction of these gas wells that lack cemented casing across faults and that are associated with hotspots is much lower for the 132 unconventional (~2%) than the 15 conventional wells (27%) associated with water analyses.

Further research should target gas wells that are uncased and/or uncemented where they intersect large faults because our statistical analysis shows that such boreholes are associated with slightly higher methane concentrations in a very small number of nearby groundwater wells. Our statistical approach thus highlights small areas that are rare and that have very low C values that statistically correlate with very small numbers of gas wells. A multiple lines of analysis approach (Molofsky et al., 2011; Breen et al., 2005; Révész et al., 2010) is needed to determine if gas migration is indeed occurring in the hotspots. Given the expense of such studies, however, our approach is a significant advance in that it highlights fundamental controls on methane emission (faults) while simultaneously pointing investigators toward potential problems in gas well construction (lack of cement or casing along boreholes intersecting faults).

More generally, our study emphasizes the utility of data mining and the importance of publication of groundwater data in regions of anthropogenic impacts. This is particularly important in cases such as hydraulic fracturing where impacts have created public controversy. New techniques of data mining, including the novel sliding window technique used here, are only possible when datasets are large. Such data mining could also be used for investigation of temporal effects in large sets of data, an analysis beyond the scope of the current work. Data mining can help scientists evaluate low-frequency environmental issues that can result in deleterious but localized impacts.

Acknowledgements

This work was funded from a gift to Penn State for the Pennsylvania State University General Electric Fund for the Center for Collaborative Research on Intelligent Natural Gas Supply Systems and was funded by National Science Foundation IIS Award # 1639150. S. Pelepko, W. Kosmer and J. Lichtinger of the PA DEP are acknowledged for providing the data to the Shale Network through an agreement facilitated by S. Perry (PA DEP) with Pennsylvania State University (managed by S.L.B.). Shale Network is funded by National Science Foundation RCN-SEES funding (OCE-11-40159) to S.L.B., Penn State, and the Consortium of Universities for the Advancement of Hydrologic Science, Inc.

(CUAHSI). Data management was facilitated by J. Williams, J. Ritzman, L. Brazil, and P. Grieve. D. Yoxtheimer, G. Llewellyn, T. Engelder, D. Fisher, M. Arthur, D. Oakley, and R. Slingerland are acknowledged for discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jconhyd.2016.10.005.

References

American Petroleum Institute, 2009. API Guidance Document HF1: Hydraulic Fracturing Operations-Well Construction and Integrity Guidelines. Washington DC.

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 Bull. 98 (2), 341–372.

Berg T, Edmunds W, Geyer A, Glover A, Hoskins D, MacLachlan D, Root S, Sevon W, Socolow A, & Miles C (1980) Geologic Map of Pennsylvania: Pennsylvania Geological Survey, 4th Series. (Map).

Brantley, S.L., 2014. Drinking water while fracking: now and in the future. Ground Water 53 (1), 21–23.

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. Int. J. Coal Geol. 126, 140–156.

Breen, K.J., Revesz, K., Baldassare, F.J., McAuley, S.D., 2005. Natural gases in ground water near Tioga junction, Tioga County, north-Central Pennsylvania-occurrence and use of isotopes to determine origins. Geological Survey (US) 2007 2328-0328.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Davies, R.J., 2011. Methane contamination of drinking water caused by hydraulic fracturing remains unproven. Proc. Natl. Acad. Sci. 108 (43), E871.

Dresel, P.E., Rose, A.W., 2010. Chemistry and origin of oil and gas well brines in western Pennsylvania. Open-File Report OFOG, p. 01.00.

Faill, R.T., 1998. A geologic history of the north-central Appalachians; part 3, the Alleghany orogeny. Am. J. Sci. 298 (2), 131–179.

Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D., Kelling, S., 2010. Spatiotemporal exploratory models for broad-scale survey data. Ecol. Appl. 20 (8), 2131–2147.

Gorody, A.W., 2012. Factors affecting the variability of stray gas concentration and composition in groundwater. Environ. Geosci. 19 (1), 17–31.

Harvey Consulting LLC, 2009. New York State (NYS) Casing Regulation Recommendations. Report to National Resources Defense Council.

Helsel, D.R., 2011. Statistics for Censored Environmental Data Using Minitab and R. John Wiley & Sons.

Howarth, R.W., Santoro, R., Ingraffea, A., 2011. Methane and the greenhouse-gas footprint of natural gas from shale formations. Clim. Chang. 106 (4), 679–690.

Ingraffea, A.R., Wells, M.T., Santoro, R.L., Shonkoff, S.B., 2014. Assessment and risk analysis of casing and cement impairment in oil and gas wells in Pennsylvania, 2000–2012. Proc. Natl. Acad. Sci. 111 (30), 10955–10960.

Jackson, R., Osborn, S., Warner, N., Vengosh, A., 2011. Responses to Frequently Asked Questions and Comments About the Shale-gas Paper by Osborn et al. Center on Global Climate Change, Duke University, Durham, NC (June 13).

Jackson, R.B., Vengosh, A., Darrah, T.H., Warner, N.R., Down, A., Poreda, R.J., Osborn, S.G., Zhao, K., Karr, J.D., 2013a. Increased stray gas abundance in a subset of drinking water wells near Marcellus shale gas extraction. Proc. Natl. Acad. Sci. 110 (28), 11250–11255.

Kang, M., Kanno, C.M., Reid, M.C., Zhang, X., Mauzerall, D.L., Celia, M.A., Chen, Y., Onstott, T.C., 2014. Direct measurements of methane emissions from abandoned oil and gas wells in Pennsylvania. Proc. Natl. Acad. Sci. 111 (51), 18173–18177.

Llewellyn, G.T., 2014. Evidence and mechanisms for Appalachian Basin brine migration into shallow aquifers in NE Pennsylvania, USA. Hydrogeol. J. 22 (5), 1055–1066.

Llewellyn, G.T., Dorman, F., Westland, J., 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. Proc. Natl. Acad. Sci. 112 (20), 6325–6330.

Molofsky, L.J., Connor, J.A., Farhat, S.K., Wylie, A.S., Wagner, T., 2011. Methane in Pennsylvania water wells unrelated to Marcellus shale fracturing. Oil & Gas Journal 109 (19), 54.

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 (3), 333–349.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

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 (20), 8172–8176.

Pennsylvania Department of Environmental Protection, 2011. Oil and Gas Well Drilling and Production in Pennsylvania.

Pohn, H., Purdy, T.L., 1981. A major (?) thrust fault at Towanda, Pennsylvania: an example of faulting with some speculation on the structure of the Allegheny Plateau. Geology of Tioga and Bradford Counties,Pennsylvania. Guidebook for the 46th Annual. Field Conference of Pennsylvania Geologists, Harrisburg, Bureau of Topographic and Geologic Survey, pp. 45–56.

Poth, C.W., 1962. The occurrence of brine in western Pennsylvania. Pennsylvania Geological Survey Bulletin M47, 1–53.

Reese, S.O., Negoba, V.V., Pelelpko, S., Kosmer, W.J., Beattie, S., 2014. Groundwater and petroleum resources of Sullivan County, Pennsylvania, Water Resource Report 71. In: Resources D.o.C.a.N. (Ed.), Pennsylvania Geological Survey.

Révész, K.M., Breen, K.J., Baldassare, A.J., Burruss, R.C., 2010. Carbon and hydrogen isotopic evidence for the origin of combustible gases in water-supply wells in north-central Pennsylvania. Appl. Geochem. 25 (12), 1845–1859.

Schoell, M., 1980. The hydrogen and carbon isotopic composition of methane from natural gases of various origins. Geochim. Cosmochim. Acta 44, 649–661.

Schon, S.C., 2011. Hydraulic fracturing not responsible for methane migration. Proc. Natl. Acad. Sci. 108 (37), E664.

Shale Network (2015). doi:10.4211/his-data-shalenetwork.

Siegel, D.I., Azzolina, N.A., Smith, B.J., Perry, A.E., Bothun, R.L., 2015. Methane concentrations in water wells unrelated to proximity to existing oil and gas wells in northeastern Pennsylvania. Environmental Science & Technology 49 (7), 4106–4112.

U.S. Energy Information Administration, 2014. Annual Energy Outlook, DOE/EIA-0383.

Vidic, R., Brantley, S., Vandenbossche, J., Yoxtheimer, D., Abad, J., 2013a. Impact of shale gas development on regional water quality. Science 340 (6134), 1235009.

Vidic, R.D., Brantley, S.L., Vandenbossche, J.M., Yoxtheimer, D., Abad, J.D., 2013b. Impact of shale gas development on regional water quality. Science 340, 826 (810.1126/science.1235009).

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. Proc. Natl. Acad. Sci. 109 (30), 11961–11966.

Woodrow, D.L., 1968. Stratigraphy, structure, and sedimentary patterns in the upper Devonian of Bradford county, Pennsylvania. Bureau of Topographic Geologic Survey, General Geology Report.