SUPPORTING INFORMATION TO THE ARTICLE

METHANE CONCENTRATIONS IN WATER WELLS UNRELATED TO PROXIMITY TO EXISTING OIL AND GAS WELLS IN NORTHEASTERN PENNSYLVANIA

By: Donald I. Siegel, Nicholas A. Azzolina, Bert J. Smith, A. Elizabeth Perry, and Rikka L. Bothun

This Supporting Information provides details about the statistical data analysis methods and results that were not incorporated into the main body of the paper.

ADDITIONAL STATISTICAL DATA ANALYSIS METHOD DETAILS

Sample Counts and Detection Frequencies

We used count functions in Excel to summarize the number of samples at 100-meter increments from 0 to 10,000 meters (10 km) from the nearest oil/gas well, coding non-detect measurements using a binary "censor indicator" of 0 for detected measurements and 1 for non-detect measurements. Non-detect measurements were defined as a reported result less than the MRL, which was either <0.005 mg/L (for 9.9% of the dissolved methane samples) or <0.026 mg/L (for 90.1% of the dissolve methane samples). We then used this field to compute the detection frequency within the 100-meter increments by dividing the number of detected measurements by the total number of samples (Equation S1).

$$\text{Detection Frequency} = \left(\frac{\text{Number of Detected Samples Above MRL}}{\text{Number of Samples}}\right) \times 100\ , \qquad \text{(S1)}$$

We calculated the 95% confidence interval for the detection frequency using a normal approximation for the confidence interval of a population proportion according to Equation S2 (Baron, 2007).

$$\hat{p} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \text{(S2)}$$

We also calculated the proportion of samples above threshold concentrations of 1, 5, and 10 mg/L dissolved methane using Equation S3. The upper value of 10 mg/L was selected because it was the lower bound used by Jackson et al. (2013) to show the "action level for hazard mitigation" in their work. We used the additional threshold values of 1 and 5 mg/L dissolved methane to assess threshold levels between the MRL and 10 mg/L. We calculated the 95% confidence interval for the proportion of samples exceeding 1, 5, and 10 mg/L dissolved methane using a normal approximation for the confidence interval of a population proportion as shown above.

$$\text{Proportion Above Threshold} = \left(\frac{\text{Number of Measurements} \geq 1, \geq 5, \text{or} \geq 10 \text{ mg/L}}{\text{Number of Samples}}\right) \times 100, \quad \text{(S3)}$$

Statistical Tests

1. Discrete *y* / Discrete *x* – Test of Proportions

The test statistic for a two-sample *Z*-test comparing proportions of two populations of independent sample sizes *n* and *m* is (Baron, 2007):

$$\frac{\hat{p}_1 - \hat{p}_2 - D}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}, \quad \text{(S4)}$$

where p_1 and p_2 are the proportions for the near and far group, respectively, D is defined as $p_1 - p_2$, n is the number of samples in Group 1, and m is the number of samples in Group 2.

We implemented the test of proportions in Minitab, which uses Fisher's exact test to compute the *p*-value (Minitab Statistical Software, 2010) and used four different threshold concentration criteria for dissolved methane:

- Detected >MRL (regardless of whether the MRL was 0.005 or 0.026 mg/L);
- Measured ≥ 1 mg/L;
- Measured ≥ 5 mg/L; and
- Measured ≥ 10 mg/L.

We used three different distance grouping in the test of proportions in contrast to Jackson, et al. (2013) which used only one distance group (1 km):

- Group 1 ≤ 500 meters (0.5 km) vs. Group 2 > 500 meters (0.5 km);
- Group 1 ≤ 1000 meters (1 km) vs. Group 2 > 1000 meters (1 km); and
- Group 1 ≤ 1500 meters (1.5 km) vs. Group 2 > 1500 meters (1.5 km).

The null (H_0) and alternative (H_A) hypotheses for the tests of proportions were:

- H_0: $p_{Group1} - p_{Group2} = 0$
- H_A: $p_{Group1} - p_{Group2} > 0$

A *p*-value greater than 0.05 (the *p*-value used in this work to define "significance") means that you cannot reject H_0 and therefore believe that there is no difference in dissolved methane concentrations between the two groups (i.e., the difference is equal to zero).

2. Discrete *y* / Continuous *x* – Logistic Regression

We extended the test of proportions described above to a continuous *x*-variable (distance) using logistic regression. The binary logistic regression model is described using the following equation (Helsel and Hirsch, 2002; Helsel 2005; Gelman and Hill 2007):

$$p = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}, \quad \text{(S5)}$$

where *p* is the probability of being detected or the probability of being measured above a threshold value (e.g., MRL, 1, 5, or 10 mg/L dissolved methane), β_0 is the intercept parameter, β_1 is the slope parameter, and *x* is the natural log of the distance to the nearest oil/gas well.

We implemented logistic regression in Minitab (Minitab Statistical Software, 2010) and used the same four threshold dissolved methane concentration criteria for the logistic regression that were used in the tests of proportions:

- Detected >MRL (regardless of whether the MRL was 0.005 or 0.026 mg/L);
- Measured ≥ 1 mg/L;

- Measured ≥ 5 mg/L; and
- Measured ≥ 10 mg/L.

3. Continuous *y* / Discrete *x* – Survival Analysis

We implemented survival analysis in Minitab using the reliability/survival module (Minitab Statistical Software, 2010), which uses the Kaplan-Meier (K-M) method to compute percentiles by determining how many observations, detects, and non-detects are above, at, and below each detected observation. The technique makes no assumptions about the distribution of data (e.g., whether they follow a lognormal or other distribution); therefore, the K-M method is non-parametric (Helsel, 2012). The log-rank and generalized Wilcoxon tests can then be used to compare the areas under the K-M-derived empirical cumulative distribution functions (ecdfs) and test for significance between groups. Therefore, survival analysis estimates a continuous ecdf of *y*, but uses a discrete grouping of samples based on *x* (distance from the nearest oil/gas well). Survival analysis can accommodate significant numbers of non-detect measurements and multiple MRLs, thus it is well-suited for this particular data set.

We used the same three distance grouping in the survival analysis as in the tests of proportions:

- Group 1 < 500 meters (0.5 km) vs. Group 2 ≥ 500 meters (0.5 km);
- Group 1 < 1000 meters (1 km) vs. Group 2 ≥ 1000 meters (1 km); and
- Group 1 < 1500 meters (1.5 km) vs. Group 2 ≥ 1500 meters (1.5 km).

In all tests, the null (H_0) and alternative (H_A) hypotheses were:

- H_0: Group 1 = Group 2; and
- H_A: Group 1 ≠ Group 2.

A *p*-value greater than 0.05 (the *p*-value used in this work to define "significance") means that you cannot reject H_0 and therefore believe that there is no difference in dissolved methane concentrations between the two groups.

4. Continuous *y* / Continuous *x* – Correlation

Neither Pearson's *r* nor Spearman's *rho* are appropriate for this data set due to the presence of two MRLs. However, for comparison purposes with Jackson et al. (2013), we calculated the following correlation coefficients in Minitab using the basic statistics/correlation module (Minitab Statistical Software, 2010):

- Pearson's *r* on original units where non-detects were substituted with the MRL (either 0.005 or 0.026 mg/L);
- Pearson's *r* on log-transformed units where non-detects were substituted with the MRL (either 0.005 or 0.026 mg/L);
- Spearman's *rho* on original units where non-detects were substituted with the MRL (either 0.005 or 0.026 mg/L);
- Spearman's *rho* on detected measurements only (discarding all non-detect measurements); and
- Spearman's *rho* on original units where all measurements ≤0.026 mg/L dissolved methane were substituted with 0.026 mg/L.

Alternatively, Kendall's *tau* correlation coefficient can be computed for data with multiple MRLs (Helsel, 2005). Thus Kendall's *tau* is the appropriate statistical measure of correlation for this data set. We used the "Ckend" macro in Minitab, written by Helsel (2012), to compute the nonparametric Kendall's *tau* correlation coefficient. Given the high proportion of non-detect measurements in the data set and the multiple MRLs, the Kendall's *tau* result is preferred over the other correlation coefficients.

ADDITIONAL STATISTICAL DATA ANALYSIS RESULTS

Sample Counts

Figure S1 shows the ecdfof the number of groundwater samples as a function of distance to the nearest oil/gas well from 0 to 10,000 meters (10 km). Sixty-seven percent (7,608 samples); 77% (8,691 samples); and 85% (9,625 samples) of the ground-water samples were collected at distances less than 500 meters (0.5 km), 1000 meters (1 km), and 1500 meters (1.5 km) of an oil/gas well, respectively.

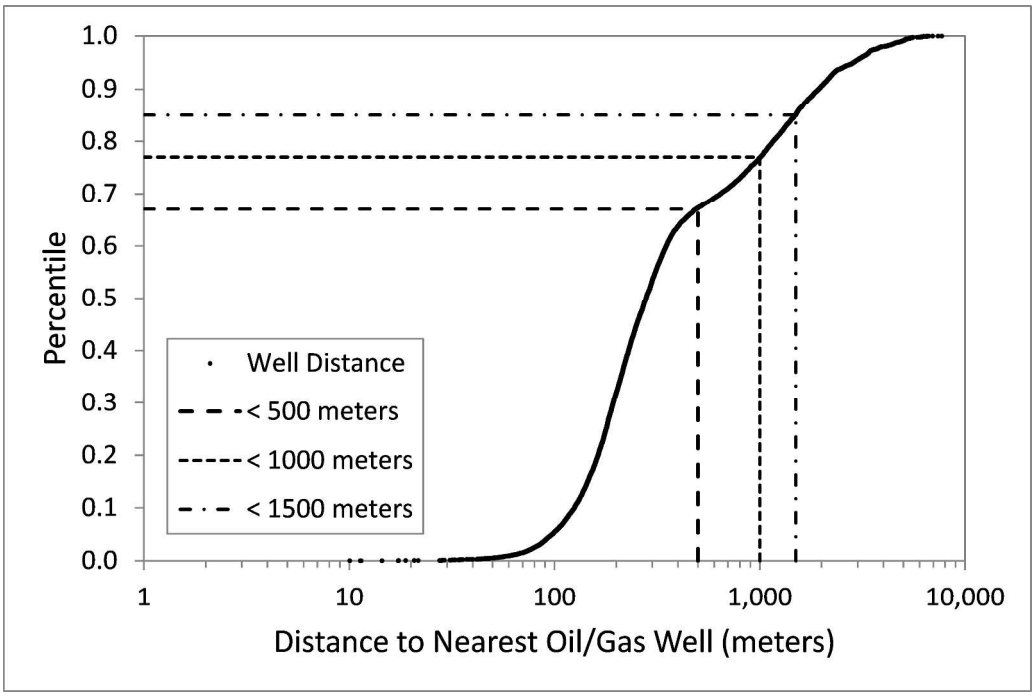

Figure S1. Empirical cumulative distribution function (ecdf) of the number of groundwater samples as a function of distance to the nearest oil/gas well. Dashed lines show the percentages of samples less than 500, 1000, and 1500 meters.

Proportions of Samples Above Threshold Concentrations

Figure S2 shows the proportion of samples detected above the MRL (Figure S2a), ≥1 mg/L (Figure S2b), ≥5 mg/L (Figure S2c), and ≥10 mg/L (Figure S2d) dissolved methane at 100-meter increments from 0 to 10,000 meters (10 km) from the nearest oil/gas well. The square symbols represent either the detection frequency above the MRL (Figure S2a) or the proportion greater than or equal to the concentration threshold (i.e., ≥1 mg/L [Figure S2b]; ≥5 mg/L [Figure S2c]; or ≥10 mg/L [Figure S2d]). The vertical errors bars in all four panels represent the 95% confidence interval. The wide confidence intervals for samples that were collected beyond ~3500 meters reflect the smaller number of samples at these distances.

There were 24.2% detected measurements (2,740 of 11,309 samples) in the entire data record. There is no *visible* increase in the detection frequency closer to oil/gas wells (Figure S2a). There was no *visible* increase in the proportion of samples exceeding 1, 5, or 10 mg/L dissolved methane closer to oil/gas wells (Figures S2b, c, and d). We test our interpretation of visual comparisons using our statistical analysis.

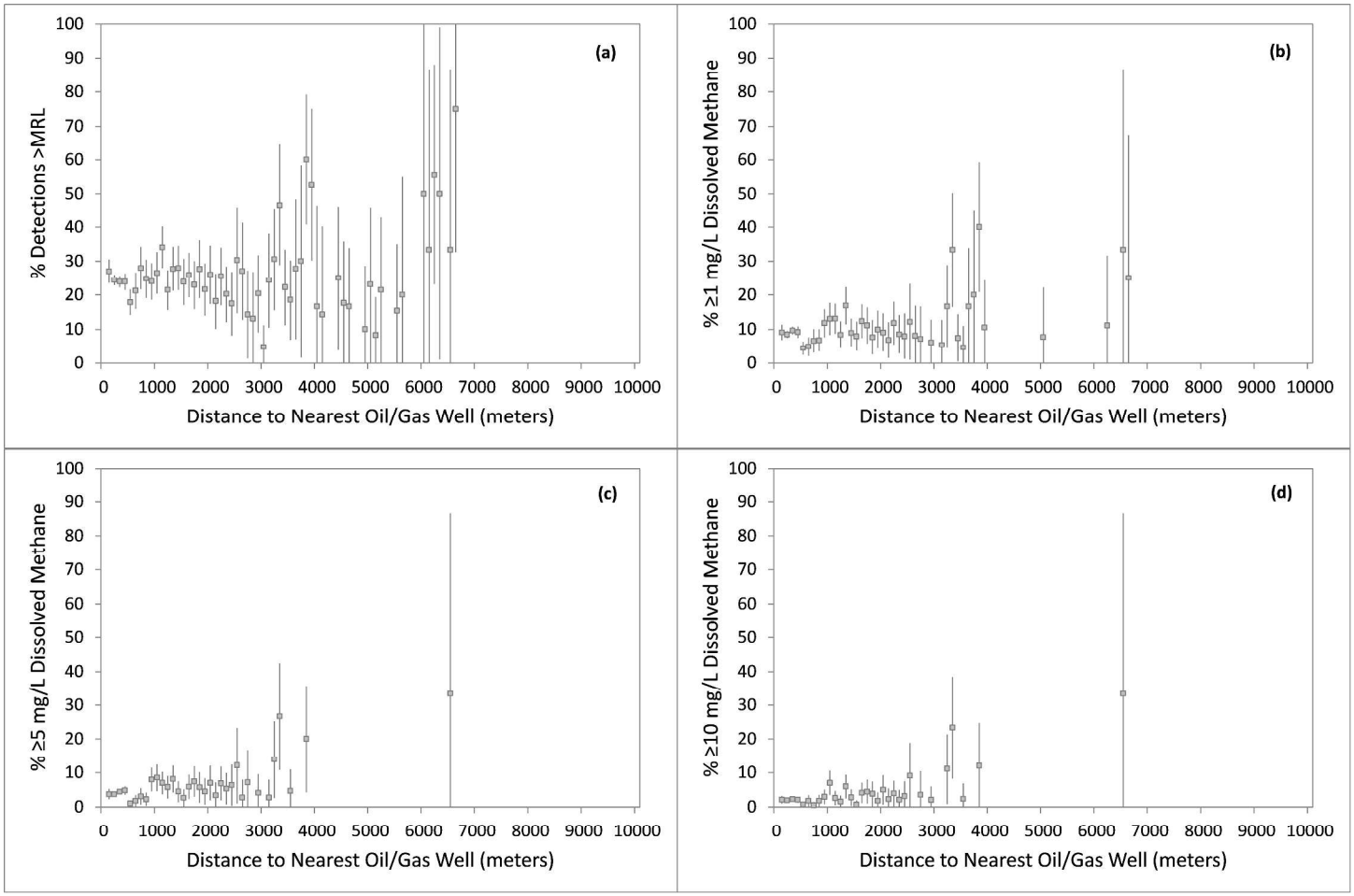

Figure S2. Proportion of samples detected above the MRL (top left [a]), ≥1 mg/L (top right [b]), ≥5 mg/L (lower left [c]), and ≥10 mg/L (lower right [d]) dissolved methane at 100-meter increments from 0 to 10,000 meters (10 km) from the nearest oil/gas well. The square symbols represent the percent greater than or equal to the threshold and the vertical errors bars represent the 95% confidence. The wide confidence intervals for samples that were collected beyond ~3500 meters reflect the smaller numbers of samples at these distances.

Logistic Regression

Table S1 provides the fitted coefficients and standard errors of the four logistic regression models. Figure S3 shows a scatterplot of the binary *y*-variable versus distance from the nearest oil/gas well with the fitted logistic regression curves. There were four different response models:

- Model 1: 0 for non-detected and 1 for detected above the MRL (either 0.005 or 0.026 mg/L) [Figure S3a];
- Model 2: 0 for <1 mg/L and 1 for ≥1 mg/L dissolved methane (Figure S3b);
- Model 3: 0 for <5 mg/L and 1 for ≥5 mg/L dissolved methane (Figure S3c); and
- Model 4: 0 for <10 mg/L and 1 for ≥10 mg/L dissolved methane (Figure S3d).

Equation S5 can be used to solve for *p* (the proportion) given *x* (the distance from an oil or gas well) using the fitted parameter estimates for β_0 and β_1 from Table S2. The process is analogous to the fitted intercept and slope from linear regression; however, in logistic regression Equation S5 ensures that the model is linear for a proportion, which is limited to the bounds between 0 and 1. For example, at 500, 1000, and 1500 meters (0.5, 1.0, and 1.5 km, respectively) the probability of detection above the MRL would be:

$$p_{x=500} = \exp(-1.097 - 0.0073\times100) / [1 + \exp(-1.097 - 0.0073\times100)] = 0.24$$

$$p_{x=1000} = \exp(-1.097 - 0.0073\times1000) / [1 + \exp(-1.097 - 0.0073\times1000)] = 0.24$$

$$p_{x=1500} = \exp(-1.097 - 0.0073\times1500) / [1 + \exp(-1.097 - 0.0073\times1500)] = 0.24$$

The probabilities of detection of dissolved methane above the MRL *do not* increase as you get closer to the oil/gas well.

At 500, 1000, and 1500 meters (0.5, 1.0, and 1.5 km, respectively) the probability of exceeding 10 mg/L dissolved methane would be:

$$p_{x=500} = \exp(-4.678 + 0.1578\times100) / [1 + \exp(-4.678 + 0.1578\times100)] = 0.02$$

$$p_{x=1000} = \exp(-4.678 + 0.1578\times1000) / [1 + \exp(-4.678 + 0.1578\times1000)] = 0.03$$

$$p_{x=1500} = \exp(-4.678 + 0.1578\times1500) / [1 + \exp(-4.678 + 0.1578\times1500)] = 0.03$$

The probabilities of exceeding 10 mg/L dissolved methane *do not* increase as you get closer to the oil/gas well from these equations. Similar results can be shown for the ≥1 and ≥5 mg/L dissolved methane concentration thresholds (calculations not shown).

Table S1. Fitted parameters of a logistic regression for each of the four models as a function of ln(Distance) from the nearest oil/gas well in meters.

Model	Term	Coefficient	Standard Error	Approximate 95% Confidence Interval	
				Lower	Upper
% Detects	B_0 (constant)	-1.10	0.13	-1.35	-0.85
	B_1 (ln[Distance])	-0.01	0.02	-0.05	0.03
% ≥1 mg/L	B_0 (constant)	-2.44	0.19	-2.81	-2.07
	B_1 (ln[Distance])	0.02	0.03	-0.04	0.08
% ≥5 mg/L	B_0 (constant)	-3.74	0.26	-4.26	-3.22
	B_1 (ln[Distance])	0.11	0.04	0.03	0.19
% ≥10 mg/L	B_0 (constant)	-4.68	0.35	-5.37	-3.98
	B_1 (ln[Distance])	0.16	0.06	0.05	0.27

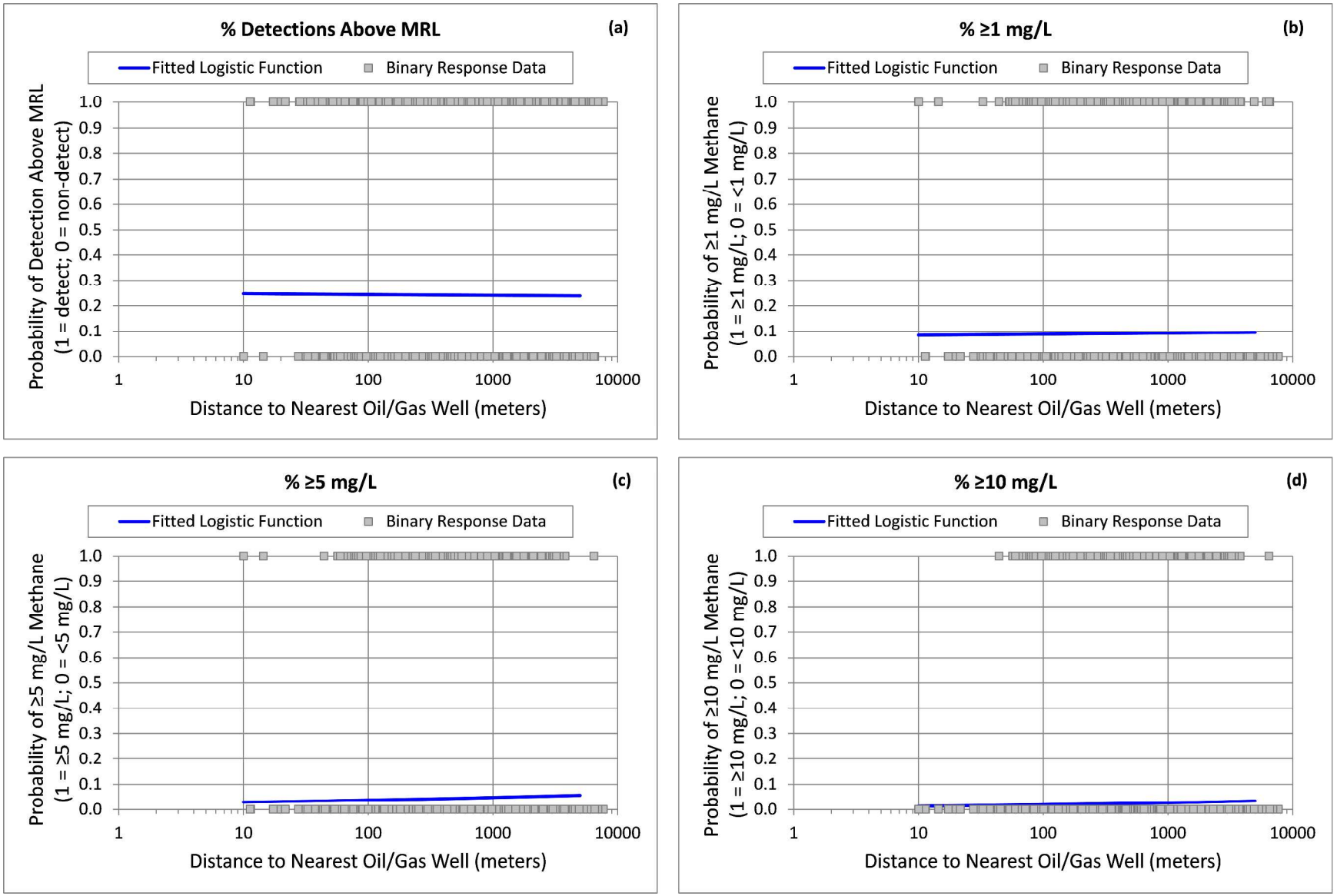

Figure S3. Results of binary logistic regression for percent detections above the MRL (top left [a]); % ≥1 mg/L (top right [b]); % ≥5 mg/L (bottom left [c]); and % ≥10 mg/L dissolved methane (bottom right [d]).

Acknowledgements

The authors would like to thank Chesapeake Energy Corporation for access to their pre-drilling groundwater dataset for NE Pennsylvania and to the landowners who allowed their water wells to be sampled. The opinions and conclusions expressed in this paper are those of the authors and do not necessarily reflect those of Chesapeake Energy. We also thank three anonymous reviewers whose comments help us to make this paper as clear as possible.

References

Baron, M. (2007) *Probability and Statistics for Computer Scientists*. Boca Raton, FL: Chapman & Hall.

Gelman, A., and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York.

Helsel, D.R. and Hirsch, R.M. (2002) *Statistical Methods in Water Resources*. U.S. Geological Survey Techniques of Water Resources Investigations, Book 4, Chapter A3, 512 p.

Helsel, D.R. (2005) *Nondetects and Data Analysis*. New York, NY: Wiley.

Helsel, D.R. (2010) Much ado about next to nothing: Incorporating nondetects in science. *Ann. Occup. Hyg.* 54(3): 257-262.

Helsel, D.R. (2012) *Statistics for Censored Environmental Data using Minitab and R*. Hoboken, NJ: Wiley.

Jackson, R.B.; Avner Vengosha, A.; Darraha, T.H.; Warnera, N.R.; Down, A.; Poredac, R.J.; Osbornd, S.G.; Zhao, K.; and Karra, J.D. (2013) Increased stray gas abundance in a subset of drinking water wells near Marcellus shale gas extraction. *Proc. Natl. Acad. Sci.* USA 110(28):11250–11255.

Minitab Statistical Software (2010) Minitab 17 [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com).