Assessment of groundwater well vulnerability to contamination through physics-informed machine learning

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

Citation: Mario A Soriano Jr et al 2021 Environ. Res. Lett. 16 084013
As Published: 10.1088/1748-9326/ac10e0
Publisher: IOP Publishing
Persistent URL: https://hdl.handle.net/1721.1/133069
Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context
Terms of use: Creative Commons Attribution 4.0 International license
Assessment of groundwater well vulnerability to contamination through physics-informed machine learning

To cite this article: Mario A Soriano Jr et al 2021 Environ. Res. Lett. 16 084013

View the article online for updates and enhancements.

You may also like

- Probing the Small-scale Structure in Strongly Lensed Systems via Transdimensional Inference
  Tansu Daylan, Francis-Yan Cyr-Racine, Ana Diaz Rivero et al.

- Evolutionary algorithm based optimization of hydraulic machines utilizing a state-of-the-art block coupled CFD solver and parametric geometry and mesh generation tools
  S Kyriacou, E Kontoleontos, S Weissenberger et al.

- Increasing the efficiency of designing hemming processes by using an element-based metamodel approach
  C Kaiser, K Roll and W Volk
Assessment of groundwater well vulnerability to contamination through physics-informed machine learning

Mario A Soriano Jr 1, Helen G Siegel ⃣, Nicholas P Johnson ⃣, Kristina M Gutchess ⃣, Boya Xiong ⃣, Yunpo Li ⃣, Cassandra J Clark ⃣, Desiree L Plata ⃣, Nicole C Deziel ⃣ and James E Saiers ⃣

1 School of the Environment, Yale University, New Haven, CT 06511, United States of America
2 School of Public Health, Yale University, New Haven, CT 06512, United States of America
3 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America
4 Department of Civil, Environmental and Geo-Engineering, University of Minnesota, Minneapolis, MN 55455, United States of America

E-mail: mariojr.soriano@yale.edu and james.saiers@yale.edu

Keywords: groundwater contamination risk assessment, drinking water quality, metamodeling, physics-informed machine learning, unconventional oil and gas development

Supplementary material for this article is available online

Abstract

Contamination from anthropogenic activities is a long-standing challenge to the sustainability of groundwater resources. Physically based (PB) models are often used in groundwater risk assessments, but their application to large scale problems requiring high spatial resolution remains computationally intractable. Machine learning (ML) models have emerged as an alternative to PB models in the era of big data, but the necessary number of observations may be impractical to obtain when events are rare, such as episodic groundwater contamination incidents. The current study employs metamodeling, a hybrid approach that combines the strengths of PB and ML models while addressing their respective limitations, to evaluate groundwater well vulnerability to contamination from unconventional oil and gas development (UD). We illustrate the approach in northeastern Pennsylvania, where intensive natural gas production from the Marcellus Shale overlaps with local community dependence on shallow aquifers. Metamodels were trained to classify vulnerability from predictors readily computable in a geographic information system. The trained metamodels exhibited high accuracy (average out-of-bag classification error <5%). A predictor combining information on topography, hydrology, and proximity to contaminant sources (inverse distance to nearest upgradient UD source) was found to be highly important for accurate metamodel predictions. Alongside violation reports and historical groundwater quality records, the predicted vulnerability provided critical insights for establishing the prevalence of UD contamination in 94 household wells that we sampled in 2018. While <10% of the sampled wells exhibited chemical signatures consistent with UD produced wastewaters, >60% were predicted to be in vulnerable locations, suggesting that future impacts are likely to occur with greater frequency if safeguards against contaminant releases are relaxed. Our results show that hybrid physics-informed ML offers a robust and scalable framework for assessing groundwater contamination risks.

1. Introduction

Groundwater contamination by human activities remains a persistent challenge in the quest towards sustainable resource utilization and equitable access to safe drinking water [1–3]. From a management perspective, these contamination risks are often conceptualized through a source-pathway-receptor approach. Sources include any type of activity or industry that can release contaminants into the environment. Once released, contaminants are transported and potentially influenced by biogeochemical reactions along hydrologic pathways that connect the source to groundwater receptors. The receptor may be defined as the water table, a water supply well, a groundwater-fed stream, or some other point within...
the hydrogeologic system. Following this framework, groundwater vulnerability to contamination is ‘the tendency or likelihood for contaminants to reach a specified position in the groundwater system after introduction at some location above the uppermost aquifer’ [4]. Predicting vulnerability is often operationalized by evaluating contaminant travel times or concentrations at the receptors of interest or estimating contaminant residence time distributions within the groundwater system [5–7].

Given that groundwater is ‘hidden’ beneath the Earth’s surface, contamination risk and groundwater vulnerability are often assessed using models. Physically based (PB) models, which encapsulate the current scientific understanding of hydrogeological processes, have a long history in groundwater contamination risk assessment [8], vulnerability analysis [7], and decision support [9]. At present, however, the application of PB simulation models at large regional scales remains difficult due to a trade-off between process complexity, spatial resolution, and computational demand. In particular, applications involving contaminant transport simulation often need to adopt simplified process representation or coarse spatial discretization to circumvent prohibitively long run times.

Spurred by the ‘big data’ revolution, machine learning (ML) approaches have emerged as popular alternative tools for prediction across the geosciences [10]. In particular, the well-known random forest (RF) algorithm has been applied from local to global scales to predict the probability of exceeding threshold concentrations and vulnerability to a variety of contaminants [11–14]. ML models have several advantages, including high accuracy in matching observations, low computational requirements relative to PB models, and capacity to generate new predictions almost instantaneously. Nevertheless, the large number of observations needed by purely data-driven ML approaches may be too expensive, time-consuming, or difficult to acquire, especially when the events of interest are rare. The prediction-centric nature of ML has also raised questions over model interpretability, with models often characterized as black boxes [15]. In such models, spurious relationships or unreasonable predictions violating conservation laws may not be readily recognized.

Hybrid approaches that combine the strengths of both PB and ML models are rapidly developing under a new paradigm known as physics-informed ML [16], or more broadly, theory guided data science [17]. Physics-informed ML methods enforce scientific theory from PB models while preserving the computational efficiency and scalability of data-driven ML models [18]. One method of implementing such integration is metamodeling [19], wherein ML models are trained to learn functional input–output relationships from PB models. Metamodeling and other physics-informed ML approaches have been shown to learn generalizable relationships and successfully make predictions for conditions not covered by the training dataset [19–24]. Metamodels are therefore potentially valuable tools for studying large unmonitored areas, alternative scenarios (e.g. future climate change or environmental disturbance), and rare events (e.g. transient groundwater contamination). However, applications of metamodels to regional scale contaminant transport problems are limited [25], and their use in quantifying vulnerability to contamination from multiple sources has yet to be investigated [26].

The objective of this research is to test the appropriateness of the metamodeling approach for groundwater vulnerability assessment. In particular, we employ a novel combination of a source-pathway-receptor PB model [27] and ML to evaluate the vulnerability of drinking water wells to contamination from unconventional oil and gas development (UD). We trained metamodels to classify the vulnerability of receptors using spatial metrics readily computable in a geographic information system. The trained metamodels, which exhibit high accuracy, were subsequently used to predict the vulnerability of domestic water wells sampled in 2018 that were distributed across a large geographic region (∼2900 sq km) that hosts UD. We utilized these predictions alongside geochemical indicators of contamination, violation reports, and historical groundwater quality records to elucidate the potential influence of UD in our study sites. Our findings demonstrate the utility of physics-informed ML for vulnerability assessment and contaminant source attribution. Because the approach is scalable and readily deployable, metamodeling is a practical tool to scientists and water resource managers engaged in informing policies intended to safeguard drinking water supplies against UD and other sources of contamination.

2. Background: groundwater quality in the context of unconventional oil and gas development

The risks of groundwater contamination from UD continue to concern communities that rely on shallow aquifers for drinking water [28]. Local groundwater impairment has been linked to UD-related incidents [29]. Surface spills of drilling fluids, hydraulic fracturing fluids, flowback and produced waters are of particular concern [30–32], especially because these fluids contain chemicals that are detrimental to human health [33]. Produced water is continuously generated throughout the operational lifespan of a UD well, and risks of spills are exacerbated by limited options for proper treatment and disposal [34].

Details on the materials released, volumes, and durations of spill events remain sparse, thus precluding the comprehensive assessment of groundwater contamination risks and potential public health
implications [35]. Produced waters typically contain high concentrations of B, Ba, Br, Ca, Cl, Li, Na, Ra, and Sr, and hence their co-occurrence at anomalously elevated concentrations in shallow groundwater may be a useful indicator of contamination by UD [36–39]. Organic chemicals used in hydraulic fracturing fluid additives and derived from hydrocarbon-bearing formations are also present in UD wastewaters and have been detected in groundwaters near UD sites [40–45]. Many inorganic and organic contaminants associated with UD have natural or other anthropogenic sources. Mass (or molar) ratios of inorganic ions, as well as isotopes and noble gas composition, are informative for differentiating UD-impacted waters from non-impacted waters or waters contaminated from alternative sources, but the isotopic and noble gas analyses can be costly and are thus not routinely performed [46–50]. With UD projected to play a significant role in the U.S. economy and energy portfolio for the next few decades [51], new approaches are needed to better characterize the environmental risks and public health implications from industry activities. In this study, we explore the potential role of UD in groundwater contamination through a novel combination of physics-informed ML modeling and geochemical analysis.

3. Materials and methods

3.1. Study area, field sampling, and related data collection

Northeast Pennsylvania is a UD center for natural gas production from the Marcellus Shale. Land use in this region is predominantly forested, but the majority of the local watersheds contain some form of UD infrastructure [52]. More than half of the rural population relies on domestic groundwater wells that typically tap into fractured bedrock aquifers composed of interbedded sandstones, siltstones, shales and mudstones of Upper Devonian formations [53]. Groundwater flow is topographically dominated, with recharge from uplands discharging to adjacent valleys along local flow systems and to major rivers along intermediate and regional systems [54].

We collected groundwater samples from 94 homes (90 domestic wells and 4 private springs) in Bradford and Tioga counties (figure 1) in the summer of 2018 (PA2018 samples) as part of the Water and Energy Resources Study. Homeowner participation in the study was voluntary. Groundwater samples were collected upstream of any household treatment system upon stabilization of pH, temperature, dissolved oxygen and specific conductance. Sample storage and laboratory analyses, which included a suite of inorganic and organic analytes, followed standard protocols [55]. A summary of the laboratory results was mailed back to the homeowners.

To identify potential UD impacts, we utilized six inorganic mass ratio frameworks to identify samples exhibiting geochemical signatures consistent with UD produced water. For these samples, we analyzed violation records [56] for indications of nearby contaminant releases. We also examined past groundwater quality records near these sampling locations at three periods: 1980s [53], 2010s [57], and 2016 [58]. The 2010s data are ‘pre-drill’ records collected prior to the drilling of new UD wells to establish a baseline in case of future alleged impairment. However, most of the pre-drill samples were already surrounded by older UD wells at the time of sampling [59]. Despite these multiple data sources, attribution of impacts to UD contamination remains ambiguous. Models are required to elucidate whether hydrologic pathways connect contaminant sources with well water receptors.

3.2. Physics-informed machine learning approach: vulnerability classification

Metamodels were trained on a PB, high resolution, three-dimensional model of groundwater flow and solute transport in a 190 sq km domain in southeastern Bradford county (dashed box in figure 1) [27]. The PB model used was Hydrogeosphere, a finite element hydrologic simulator [60]. The calibrated model successfully replicated observations of hydraulic head determined from USGS monitoring wells and drillers’ well logs and estimates of groundwater discharge from regression equations [27]. The PB model utilizes the concept of capture probability to operationalize vulnerability. Capture probability is the probability that a parcel of water at a given location will arrive at a receptor within a specified time, and vulnerability is defined as the union of capture probability at potential contaminant source locations (text S1 and figure S1 (available online at https://stacks.iop.org/ERL/16/084013/mmedia); see [27] for a detailed discussion). Note that vulnerability of a receptor (water well) does not necessarily equate to its actual contamination, as the latter is also dependent on the occurrence of releases (e.g. spills) at the source locations (UD well pads) inside the capture zone. Capture probability and vulnerability to contamination at 25 years were evaluated for 316 domestic groundwater wells extracted from the Pennsylvania Groundwater Information System [61] that were within the PB model boundaries. The 25 year time horizon corresponds to the expected operational lifespan of UD wells and was selected to provide insights on risks from the early stages of drilling up to well plugging [51]. Note that the 316 receptors within the PB model domain are distinct from the 94 PA2018 wells that are distributed across a larger geographic region (~2900 sq km).

Because PB modeling with sufficient spatial resolution at this large scale remains computationally intractable, we used metamodeling to gain insights

3
Figure 1. Locations of groundwater sampling and boundaries of the physically based (PB) model in Bradford county, Pennsylvania. PB modeling inside the dashed box was used to train metamodels of groundwater well vulnerability to aqueous phase contaminants from unconventional oil and gas development sites. These metamodels were subsequently used to predict vulnerability at the groundwater sampling locations shown.

on vulnerability and to identify simple predictors that reflect the underlying physics of contaminant transport processes.

For our metamodeling approach, we posed the question of vulnerability to contamination from UD as a classification problem: a domestic well is classified as Vulnerable (Class V) if $V \geq 0.001$ and as Non-vulnerable (Class NV) otherwise. This conservative threshold accounts for the dominance of low capture probability regions in domestic well capture zones arising from uncertainty in the streamlines induced at typical household pumping rates while also minimizing the effect of numerical errors [27, 62, 63]. The objective of the metamodeling is to predict the vulnerability class of a domestic well given a set of input predictors describing the hydrology, terrain, and proximity to UD at a domestic well location (table S1). Interactions between topography and proximity to UD were previously shown to be important in predicting groundwater quality in a large dataset, but the reasons for such associations are poorly understood [64]. We hypothesize that metamodels linking these predictors to the hydrologic concepts of capture probability and well vulnerability will elucidate mechanisms of contamination from UD and explain observed geochemical signatures. The predictors were assigned to wells using ArcGIS 10.7.

Many predictors, such as the various proximity metrics, are inherently correlated with one another. Hence, we used a supervised ML algorithm called conditional inference forest (CIF), a variant of the RF algorithm that employs unbiased recursive partitioning through Chi-square tests in its tree-building process to identify the predictors most strongly associated with the response variable [65]. This approach helps the algorithm learn robust relationships even when predictors are correlated [66]. As in the case of RF, random subsets of predictors and bootstrap samples of the data are used for training each tree. The algorithm uses up to two-thirds of the dataset for training each tree, and the remaining ‘out-of-bag’ sample is used for testing and for computing honest performance measures of the model.

The ML training/testing dataset consisted of PB model simulations for the original 316 domestic wells described in the previous section (figure S2(a)). In order to encompass a wider ensemble of field conditions, we augmented this dataset with simulations for 834 ‘seed’ wells [19] that were equally divided but randomly distributed inside uniform $2 \times 2$ km grid cells superimposed on the PB model domain. One hundred CIF model realizations (random restarts), each consisting of 3000 trees, were generated to assess the stability of the metamodel predictions. The synthetic minority over-sampling technique [67] was used to ensure a balanced dataset for each realization. Dominant predictors were identified using conditional permutation importance [68], and
Figure 2. Conditional permutation importance showing top predictors of vulnerability class across all metamodel realizations, ordered from top to bottom by mean rank. Brackets indicate groups of predictors with statistically similar importance (Kruskal–Wallis and post-hoc multiple comparisons test, \( p < 0.05 \)). The predictors describe the topography, hydrology, and proximity to UD at a given receptor location. Definitions: idups—inverse distance to nearest upgradient UD source, d_elev—difference in surface elevation between nearest UD source and the groundwater well, mtpi—multiscale topographic position index, dnrst—distance to nearest UD source, idw2_1 km/2 km/3 km—inverse distance squared weighted count of UD sources within 1, 2, or 3 km of the groundwater well, idw_2 km/3 km—inverse distance weighted count of UD sources within 2 or 3 km, n_1 km/2 km—number of UD sources within 1 or 2 km, wtd—water table depth, dsd—distance from stream to divide, mvdev—deviation of the cell elevation from the mean elevation, dwetland—distance to nearest wetland. Detailed descriptions are in table S2.
a hillside, with large negative values representing major valleys and large positive values representing hilltops or catchment divides [72]. Common proximity metrics [73, 74] such as the (Euclidean) distance to the nearest UD source, dnrst, and variants of inverse distance-weighted UD well counts within specified radii (1, 2, and 3 km) from groundwater wells are also important predictors of vulnerability class. However, our results demonstrate that proximity metrics alone are insufficient and that hydrologic drivers of solute transport must be considered for accurate assessment of drinking water vulnerability to contamination from UD. Other important predictors are described in table S2.

The dominant predictors have nonlinear effects on the log odds of classification into Class V (figure 3). Proximity metrics generally indicate that groundwater wells closer to UD are more likely to be Vulnerable (figures 3(a)–(d)). This insight is consistent with capture probability being highest in the immediate vicinity of a groundwater well and decreasing to zero as one moves further away. The most common UD proximity metric, dnrst, has an inverse relationship with the log odds up to around 2 km, where the effect appears to stabilize (figure 3(b)). This finding implies that domestic well capture zones in the study area can extend up to 2 km and thus some likelihood of impairment by weakly sorbing, recalcitrant UD contaminants is preserved at this distance. Contaminants that are susceptible to degradation or that sorb strongly may be removed, transformed, or retained on aquifer solids at smaller distances. Our results also show that the log odds of being in Class V increase sharply when there is at least one UD source within 1 km (n_1 km) of a groundwater well, but more sources within this distance do not translate to higher odds (figure 3(d)). While seemingly counterintuitive, this finding is consistent with our threshold condition for the classification problem: one source inside the probabilistic capture zone will generally suffice to meet $V \geq 0.001$. The marginal decline in odds is due to interaction between the topography in the training region and the locations of UD well pads. Given the terrain in the PB model domain, areas intercepting more than one well pad in 1 km tend to be located near topographic divides and thus often fail to intercept flow paths (figure S3).

Groundwater wells are more likely to be in Class V if they are proximal to a flow path that extends downhill from a UD source (idups > 0) (figure 3(a)) and if the nearest UD source is at a higher elevation (d_elev > 0) (figure 3(e)), consistent with the topographically driven flow in our study region. Wells located in valleys (mtpi < 0) are also more likely to be in Class V (figure 3(f)), and slightly more so near stream confluences (characterized by small dsd [75]) (figure 3(g)). These areas correspond to discharge zones in the groundwater flow system. Locations with a shallow water table (characterized by small wtd) are more likely to be Vulnerable (figure 3(h)). This behavior is consistent with the popular DRASIC groundwater vulnerability rating system [76], where a shallow water table implies faster infiltration and less degradation of contaminants derived from the surface. These qualitative interpretations collectively suggest that the metamodel predictions are hydrologically sensible.

4.2. Predicted vulnerability classification of the groundwater sampling locations

We used the metamodels to classify the vulnerability of the groundwater receptors where the PA2018 samples were collected. Fifty-seven of the 94 locations (60.6%) were predicted to be in Class V based on the mode across all metamodel realizations (figure S2(b)). This proportion is considerably larger than

![Figure 3](image-url)
the 13.6% of the original 316 domestic wells with V \geq 0.001 simulated inside the PB model domain (figure S2(a)). This discrepancy reflects the greater proximity of our samples to UD sources—the median dnrst for the PA2018 locations was about 1 km less than the median for wells within the PB model domain and Bradford county in general (figure S4). Homeowners who volunteered to join our study tended to live closer to UD, and correspondingly, were more likely to be in locations classified as Vulnerable to contamination. This observation is consistent with research indicating that geographic proximity to UD infrastructure affects one’s level of concern towards UD risks [77]. Studies with similar voluntary recruitment strategies should consider this self-selection when making generalizations about the prevalence of UD impacts.

All 94 sampling locations had \geq60% agreement in their predicted class across all metamodel realizations. Eighty-two had \geq90% and 68 had 100% classification agreement, indicating general robustness of the predictions for majority of our samples (figure S2(b)). Samples near 60% agreement may potentially fall in a different mode class with additional realizations. These locations likely have predictor values with counteracting effects, such that the random predictor subsetting in the CIF algorithm can lead to a different classification with each metamodel run. This underscores the importance of evaluating the stability of the metamodels and their predictions.

4.3. Vulnerability and other lines of inquiry about UD contamination

In this section, we probe the potential influence of UD on the PA2018 samples through multiple complementary approaches: vulnerability metamodeling, analysis of inorganic and organic constituents, and inspection of violation reports and historical groundwater quality records. First, we compared the chemical composition of samples across Vulnerable and Non-vulnerable classes, focusing on inorganic chemicals typically present in high concentrations within produced waters, as well as organic compounds that have been reported in UD wastewaters [36, 38, 43, 45]. Taking the 82 wells with \geq90% agreement, we found elevated median concentrations of several constituents—Ba, Br, Cl, Li, Na, Sr, total dissolved solids (TDS), and gasoline range organics (GRO)—in Class V samples relative to Class NV samples, with Ba, Br, and Li being statistically significant at p < 0.05 (table S3). This result remains even when considering all samples (table S4) or only the 68 samples with 100% classification agreement (table S5), with Cl, Na, and GRO also attaining statistical significance in the latter. This consistency indicates the robustness of these results and supports the need to further assess the Vulnerable sampling locations. Of the constituents evaluated, only Ba has a health-based EPA maximum contaminant level (2000 \mu g l^{-1}). This MCL was exceeded in one Class NV sample (Sample ID 076) and five Class V samples (032, 046, 048, 049, and 065).

Although produced water inputs would raise concentrations of the identified constituents in non-impacted fresh groundwater, our observation of higher median concentrations within Class V samples is insufficient to establish a link with UD activities because these chemicals have other potential sources. Inorganic ion ratios have been used in groundwater source attribution studies and to identify shallow groundwater that have been altered by mixing with produced water [46, 47, 49, 50, 78]. Six frameworks that use inorganic ion ratios show that the majority of our samples plot within non-impacted groundwater, while some samples are geochemically similar to produced waters (figure 4). Eight samples (six Class V: 032, 036, 046, 052, 054, 063, and two Class NV: 076, 092, \geq90% classification agreement) plot within the oil and gas produced water region in three or more frameworks. Note that oil and gas produced waters have inorganic ion ratios similar to Appalachian Basin brines that are known to mix naturally with shallow groundwater beneath valleys of our study area [79, 80]. We therefore explored additional sources of information, including data on organic chemistry, to probe whether the produced water signature in these eight samples could be attributed to UD. Samples 032, 046, 054, and 063 (Class V) exceed the median GRO concentration of the full dataset, with the GRO concentration in 054 more than twice the median GRO (table S6). Samples 046, 054, 063 (Class V), as well as sample 092 (Class NV), exceed the median diesel range organics (DRO), with the 046 and 054 concentrations exceeding the median DRO by at least a factor of two. Sample 046 also exhibits the highest measured DRO in the dataset. Compound-specific analysis of this sample [55] revealed the presence of bis-2-ethylhexyl phthalate, a compound previously detected in Marcellus flowback waters [43], and N-N-dimethyltetradecylamine, a suspected surfactant used in the industry [81].

Well-water impairment (i.e. impact) is contingent on both vulnerability and an actual release of contaminants. Thus, we examined UD releases in the vicinity of the eight potentially impacted samples to strengthen our inferences on source attribution. We found violation reports indicating spills, defective casings, and inadequate waste storage facilities at UD well pads within 2 km of five of the six Class V samples (table S7). The 2 km distance is notable because it corresponds to the dnrst predictor that emerged from the metamodeling, further suggesting the pertinence of this distance to the capture zones of domestic wells in the study area. Sample 032 and the
two Class NV samples were linked to violations indicating releases beyond 2 km. The absence of potential UD releases closer to sample 032 and the two Class NV samples may point to a non-UD source of the geochemical signatures in these samples, such as naturally upwelling brines. The absence of known releases may also simply reflect deficiencies and difficulties in documenting such events. None of the violation reports contained sufficient information on the material, volume, and duration of the releases necessary for the detailed modeling of such events. Contamination via mechanisms not represented in our current models, e.g. UD-related roadside truck spills, also cannot be ruled out as possible explanations for the produced water signatures, although transportation spills are estimated to have much lower probabilities than spills at well pads [83].

Temporal data would further reinforce the plausibility of contamination from UD, particularly if samples from a given locale shift from the non-impacted region to the oil and gas produced water region in the mass ratio frameworks. Such change would indicate water quality impairment over the intervening period. On the other hand, if the groundwater quality is consistently similar to produced water, then the sustained geochemical signature is likely derived from a non-UD source, such as mixing with natural brines. Given the absence of continuous
In summary, 57 out of our 94 PA2018 sampling locations were predicted to be Vulnerable to contamination from UD. Eight of the 94 samples exhibited geochemical similarity to oil and gas produced waters. Six of these were predicted to be Vulnerable, and the remaining two were Non-vulnerable. Contamination of the Vulnerable wells from UD is plausible from a hydrologic standpoint, and violations indicating UD releases, together with recent changes in groundwater chemistry, provide further support for a potential UD link (table S9). Non-UD sources of the produced water signatures (e.g. natural upwelling of brines in valleys) or anthropogenic sources not represented in the PB model (e.g. transportation spills) cannot be ruled out, particularly for the two Non-vulnerable sampling locations. We emphasize that the majority of our samples are non-impacted, consistent with previous studies of groundwater contamination by UD in the region [44, 46, 59, 84]. With the exception of Ba concentrations, which can be elevated from natural sources in this region, concentrations of analytes reported in tables S3–S5 did not exceed relevant health-based standards at either Vulnerable or Non-vulnerable sites. Despite the currently low prevalence of observed contamination, the large proportion of Vulnerable wells suggests that drinking water impacts will occur with greater frequency if decisions are made to relax regulatory and operational safeguards intended to prevent environmental releases of UD wastewaters and other fluids.

5. Conclusions

Our analysis elucidates important interactions between groundwater quality, proximity to UD, and topography through the hydrologic concepts of capture probability and well vulnerability. Metamodels successfully emulated predictions from a detailed PB groundwater flow and transport model, and enabled the efficient generation of robust insights on groundwater vulnerability across a larger geographic area. We found that inverse distance to nearest upgradient UD source (idups) was highly informative in classifying vulnerability to contamination in our topographically controlled flow system. This predictor and the other spatial metrics that highly influence vulnerability can serve as functional proxies in initial assessment of groundwater contamination risks. We found that vulnerability clarified the interpretation of samples exhibiting chemical similarity to oil and gas produced waters. Our results show that the physics-informed ML approach of vulnerability classification serves as a useful line of inquiry in investigating suspected cases of contamination, particularly when more extensive and costly methods are infeasible. Metamodeling of vulnerability can also function as a systematic framework for optimizing locations for groundwater quality monitoring and for probing water-related contaminant exposure pathways relevant to public health. Lastly, the approach provides a science-based platform for designing policies to protect groundwater within the context of UD (e.g. by minimizing new development in areas near clusters of vulnerable water wells) and other anthropogenic sources of contamination.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

This research was conducted as part of the Water and Energy Resources Study (https://medicine.yale.edu/water/). The project was developed under Assistance Agreement No. CR839249 awarded by the U.S. Environmental Protection Agency to Yale University. It
has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. M A Soriani was also supported by the Yale Institute for Biospheric Studies Small Grants Program and the Geological Society of America Graduate Student Research Grants program. C J Clark was funded in part by the National Institute of Environmental Health Sciences under the National Institutes of Health [F31ES031441]. We thank the three anonymous reviewers for their thoughtful comments that led to improvements in this manuscript. We are also grateful to Dr Joshua Warren for initial insights on ML, and to Keli Sorrentino, Julie Plano, Livia Nock, Emma Ryan, Benedictis, Nicholas Hoffman, and Nicolette Bugher for project assistance. We also thank the Yale Analytical and Stable Isotope Center and the Cary Institute of Ecosystem Studies for use of laboratory facilities, and the Yale Center for Research Computing for use of the high-performance computing infrastructure.

**ORCID iDs**

Mario A Soriani Jr [https://orcid.org/0000-0003-0499-9352](https://orcid.org/0000-0003-0499-9352)

Helen G Siegel [https://orcid.org/0000-0002-3412-5851](https://orcid.org/0000-0002-3412-5851)

Nicholas P Johnson [https://orcid.org/0000-0003-1346-3576](https://orcid.org/0000-0003-1346-3576)

Kristina M Gutchev [https://orcid.org/0000-0002-9745-5049](https://orcid.org/0000-0002-9745-5049)

Boya Xiong [https://orcid.org/0000-0002-7994-3508](https://orcid.org/0000-0002-7994-3508)

Yunpo Li [https://orcid.org/0000-0002-1172-2741](https://orcid.org/0000-0002-1172-2741)

Cassandra J Clark [https://orcid.org/0000-0002-5993-4540](https://orcid.org/0000-0002-5993-4540)

Desiree L Plata [https://orcid.org/0000-0003-0657-7735](https://orcid.org/0000-0003-0657-7735)

Nicole C Deziel [https://orcid.org/0000-0002-5751-9191](https://orcid.org/0000-0002-5751-9191)

James E Saiers [https://orcid.org/0000-0002-0117-3347](https://orcid.org/0000-0002-0117-3347)

**References**

[1] Gleeson T, Cuthbert M, Ferguson G and Perrone D 2020 Global groundwater sustainability, resources, and systems in the anthropocene *Annu. Rev. Earth Planet. Sci.* 48 431–63

[2] Burri N M, Weather L K, Moeck C and Schirmer M 2019 A review of threats to groundwater quality in the anthropocene *Sci. Total Environ.* 684 136–55

[3] Lall U, Josset L and Russo T 2020 A snapshot of the world’s groundwater challenges *Annu. Rev. Environ. Resour.* 45 171–94

[4] NRC 1993 *Ground Water Vulnerability Assessment: Predicting Relative Contamination Potential under Conditions of Uncertainty* (Washington, DC: The National Academies Press) p 224

[5] Focazio M et al 2003 Assessing Ground-Water Vulnerability to Contamination: Providing Scientifically Defensible Information for Decision Makers (Reston, VA: US Geological Survey Circular 1224)

[6] Molson J W and Frind E O 2012 On the use of mean groundwater age, life expectancy and capture probability for defining aquifer vulnerability and time-of-travel zones for surface water protection *J. Contam. Hydrol.* 127 76–87

[7] Wachniew P, Zurek A J, Stump C, Gemitzi A, Gargini A, Filippini M, Rozanski K, Meeks J, Kwarner J and Witczak S 2016 Toward operational methods for the assessment of intrinsic groundwater vulnerability: a review *Crit. Rev. Environ. Sci. Technol.* 46 827–84

[8] Neupauer R M and Wilson J L 2005 Backward probability model using multiple observations of contamination to identify groundwater contamination sources at the Massachusetts Military Reservation *Water Resour. Res.* 41 W02015

[9] Freeze R A, Massmann J, Smith L, Sperling T and James B 1990 Hydrogeological decision analysis: I. A framework *Groundwater* 28 738–66

[10] Tahmasebi P, Kamrava S, Bai T and Sahimi M 2020 Machine learning in geo- and environmental sciences: from small to large scale *Adv. Water Resour.* 142 103619

[11] Kroell I, Breuer L and Bach M 2020 Nation-wide estimation of groundwater redox conditions and nitrate concentrations through machine learning *Environ. Res. Lett.* 15 064004

[12] Podgorski J and Berg M 2020 Global threat of arsenic in groundwater *Science* 368 845

[13] Tesoriero A J, Gronberg J A, Juckem P F, Miller M P and Austin B P 2017 Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification *Water Resour. Res.* 53 7316–31

[14] Rodriguez-Galaino V, Mendes M P, Garcia-Soldado M J, Chica-Olmo M and Ribeiro L 2014 Predictive modeling of groundwater nitrate pollution using random forest and multivariate models related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain) *Sci. Total Environ.* 476–477 189–206

[15] Apley D W and Zhu J 2020 Visualizing the effects of predictor variables in black box supervised learning models *J. R. Stat. Soc. Ser. B* 82 1035–66

[16] Tartakovsky A M, Marrero C O, Perdikaris P, Tartakovsky G D and Barajas-Solano D 2020 Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems *Water Resour. Res.* 56 e2019WR026731

[17] Karpatne A, Alfuth G, Faghmous J H, Steinbach M, Banerjee A, Ganguly A, Shekar S, Samatova N and Kumar V 2017 Theory-guided data science: a new paradigm for scientific discovery from data *IEEE Trans. Knowl. Data Eng.* 29 2318–31

[18] Liang J, Li W, Bradford S and Simunek J 2019 Physics-informed data-driven models to predict surface runoff water quantity and quality in agricultural fields *Water* 11 200

[19] Fienen M N, Nolan B T and Feinstein D T 2016 Evaluating the sources of water to wells: three techniques for metamodeling of a groundwater flow model *Environ. Model. Softw.* 77 95–107

[20] Nolan B T, Malone R W, Gronberg J A, Thorp K R and Ma L 2012 Verifiable metamodels for nitrate loads to drains and groundwater in the Corn Belt, USA *Environ. Sci. Technol.* 46 901–8

[21] Konapala G, Kao S C, Painter S L and Lu D 2020 Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US *Environ. Res. Lett.* 15 104022

[22] Read J S et al 2019 Process-guided deep learning predictions of lake water temperature *Water Resour. Res.* 55 9173–90

[23] Starn J J and Belitz K 2018 Regionalization of groundwater residence time using metamodeling *Water Resour. Res.* 54 6357–73

[24] [Read J S et al 2019 Process-guided deep learning predictions of lake water temperature *Water Resour. Res.* 55 9173–90](https://onlinelibrary.wiley.com/doi/abs/10.1029/2019WR023984)

[25] [Read J S et al 2019 Process-guided deep learning predictions of lake water temperature *Water Resour. Res.* 55 9173–90](https://onlinelibrary.wiley.com/doi/abs/10.1029/2019WR023984)
Starn J J, Kaufman L I, Carlson C S, Reddy J E and Fienen M N 2021 Three-dimensional distribution of groundwater residence time metrics in the glaciated United States using metamodels trained on general numerical simulation models Water Resour. Res. 57 e2020WR027335

Yu X, Cui T, Steekanth J, Mangeon S, Doble R, Xin P, Rassam D and Gillfelder M 2020 Deep learning emulators for groundwater contaminant transport modelling J. Hydrod. 580 125351

Enzenhofer R, Binning P J and Nowak W 2015 Stakeholder-objective risk model (STORM): determining the aggregated risk of multiple contaminant hazards in groundwater well catchments Adv. Water Resour. 83 160–75

Soriano M A, Siegel H G, Gutchess K M, Clark C J, Li Y, Xiong B, Plata D L, Deziel N C and Sayers J E 2020 Evaluating domestic well vulnerability to contamination from unconventional oil and gas development sites Water Resour. Res. 56 e2020WR028005

Clark C J, Warren J L, Kadan-Lottick N, Ma X, Bell M L, Sayers J E and Deziel N C 2021 Community concern and government response: identifying socio-economic and demographic predictors of oil and gas complaints and drinking water impairments in Pennsylvania Energy Res. Soc. Sci. 76 102070

EPA 2016 Hydraulic fracturing for oil and gas: impacts from the hydraulic fracturing water cycle on drinking water resources in the United States Final Report EPA/600/R-16/236F (Washington, DC: Office of Research and Development, US Environmental Protection Agency)

Shanafiel M, Cook P G and Simmons C T 2019 Towards quantifying the likelihood of water resource impacts from unconventional gas development Groundwater 57 547–61

Lefebvre R 2017 Mechanisms leading to potential impacts of shale gas development on groundwater quality WIREs Water 4 e1188

Vengosh A, Jackson R B, Warner N, Darrah T H and Kondash A 2014 A critical review of the risks to water resources from unconventional shale gas development and hydraulic fracturing in the United States Environ. Sci. Technol. 48 8334–48

Wollin K M et al 2020 Critical evaluation of human health risks due to hydraulic fracturing in natural gas and petroleum production Arch. Toxicol. 94 967–1016

Menefee A H and Ellis B R 2020 Wastewater management strategies for sustained shale gas production Environ. Res. Lett. 15 024001

Maloney K O et al 2017 Unconventional oil and gas spills: materials, volumes, and risks to surface waters in four states of the U.S. Sci. Total Environ. 581 369–77

Chapman E C, Capo R C, Stewart B W, Kirby C S, Hammack R W, Schroeder K T and Edlbom H M 2012 Geochemical and strontium isotope characterization of produced waters from Marcellus Shale natural gas extraction Environ. Sci. Technol. 46 3545–53

Mummford A C, Maloney K O, Akob D M, Nettemann S, Proctor A, Ditty J, Ullman E, Lookbill J and Cozzarelli L M 2020 Shale gas development has limited effects on stream biology and geochemistry in a gradient-based, multiparameter study in Pennsylvania Proc. Natl Acad. Sci. 117 3670

Shih J S, Sayers J E, Anisfeld S C, Chu Z, Muehlenbachs L A and Olmstead S M 2015 Characterization and analysis of liquid waste from Marcellus Shale gas development Environ. Sci. Technol. 49 9557–65

Warner N R, Christie C A, Jackson R B and Vengosh A 2013 Impacts of shale gas wastewater disposal on water quality in western Pennsylvania Environ. Sci. Technol. 47 11849–57

Drollette B D et al 2015 Elevated levels of diesel range organic compounds in groundwater near Marcellus gas operations are derived from surface activities Proc. Natl Acad. Sci. 112 13184

Hoelzer K et al 2016 Indications of transformation products from hydraulic fracturing additives in shale-gas wastewater Environ. Sci. Technol. 50 8036–48

Luek J L and Gonsior M 2017 Organic compounds in hydraulic fracturing fluids and wastewaters: a review Water Res. 123 336–48

Maguire-Boyle S J and Barron A R 2014 Organic compounds in produced waters from shale gas wells Environ. Sci. Process. Impacts 16 2237–48

McMahon P B, Lindsey B D, Conlon M D, Hunt A G, Belitz K, Jurgens B C and Varelia B A 2019 Hydrocarbons in upland groundwater, Marcellus shale region, northeastern Pennsylvania and Southern New York, U.S.A Environ. Sci. Technol. 53 8027–35

Oetjen K and Thomas L 2016 Volatile and semi-volatile organic compound patterns in bowflow waters from fracturing sites within the Marcellus Shale Environ. Earth Sci. 75 1043

Brantley S L, Youtheimer D, Arjmand S, Grieve P, Vidic R, Pollak J, Llewellyn G T, Abad J and Simon C 2014 Water resource impacts during unconventional shale gas development: the Pennsylvania experience Int. J. Coal Geol. 126 140–56

Cantlay T, Bain D J and Stolz J F 2020 Determining conventional and unconventional oil and gas well brines in natural samples III: mass ratio analyses using both anions and cations Environ. Sci. Health A 55 24–32

McIntosh J C et al 2019 A critical review of state-of-the-art and emerging approaches to identify fracking-derived gases and associated contaminants in aquifers Environ. Sci. Technol. 53 1063–77

Fisherman R and Bain D J 2019 Alkali earth ratios differentiate conventional and unconventional hydrocarbon brine contamination Sci. Total Environ. 695 133944

Wilson J M, Wang Y and van Briesen J M 2014 Sources of high total dissolved solids to drinking water supply in southwestern Pennsylvania J. Environ. Eng. 140 B4014003

EIA 2019 Annual Energy Outlook (Washington, DC: US Energy Information Administration)

Stonecker E T and Milhein L E 2015 Landscape disturbance from unconventional and conventional oil and gas development in the Marcellus Shale region of Pennsylvania, USA Environments 2 200–20

Williams J H, Taylor L E and Low D J 1998 Hydrogeology and groundwater quality of the glaciated valleys of Bradford, Tioga, and Potter Counties, Pennsylvania Water Resources Report 68 (Harrisburg, PA: Pennsylvania Geological Survey)

Heisig P M and Scott T M 2013 Occurrence of methane in groundwater of south-central New York State, 2012-systematic evaluation of a glaciated region by hydrogeologic setting Scientific Investigations Report 2013–5190 (Reston, VA) p 44

Xiong B et al Low organic chemical occurrence in groundwater near hydraulic fracturing activities associated with limited groundwater development transport in northeastern Pennsylvania (in revision)

PADEP 2020 Oil and Gas Compliance Report [online database] (available at: www.depgreenport.state.pa.us/ReportExtracts/OG/OilComplianceReport) (Accessed 01 August 2020)

Brantly S L 2018 Shale Network [online database] (available at: 10.4211/his-data-shalenetwork) (Accessed 01 August 2020)

Clune J W and Cravotta C A I 2019 Drinking water health standards comparison and chemical analysis of groundwater for 72 domestic wells in Bradford County, Pennsylvania, 2016 Scientific Investigations Report 2018–5170 (Reston, VA: U. S. Geological Survey) p 76

Wen T, Niu X, Gonzales M, Zheng G, Li Z and Brantly 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 Environ. Sci. Technol. 52 7149–59
[60] Aquanty Inc 2018 HydroGeoSphere: A Three-dimensional Numerical Model Describing Fully-integrated Subsurface and Surface Flow and Solute Transport (Waterloo: Aquanty Inc)

[61] Pennsylvania Geological Survey 2020 Pennsylvania Groundwater Information System (PaGWIS) (Middletown, PA: PA Department of Conservation and Natural Resources Bureau of Topographic and Geologic Survey)

[62] Henri C V and Harter T 2019 Stochastic assessment of nonpoint source contamination: joint impact of aquifer heterogeneity and well characteristics on management metrics Water Resour. Res. 55 6773–94

[63] Libera A, De Barros F P J and Guadagnini A 2017 Influence of pumping operational schedule on solute concentrations at a well in randomly heterogeneous aquifers J. Hydrol. 546 490–302

[64] Yan B et al 2017 Association of groundwater constituents with topography and distance to unconventional gas wells in NE Pennsylvania Sci. Total Environ. 577 195–201

[65] Hothorn T, Hornik K and Zeileis A 2006 Unbiased recursive partitioning: a conditional inference framework J. Comput. Graph. Stat. 15 651–74

[66] Tyralsis H, Papacharalampous G and Langouis A 2019 Review of random forests for water scientists and practitioners and their recent history in water resources Water 11 910

[67] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique J. Artif. Intell. Res. 16 321–57

[68] Strobl C, Boulesteix A-L, Kneib T, Augustin T and Zeileis A 2008 Conditional variable importance for random forests BMC Bioinformat. 9 307

[69] James G et al 2013 An Introduction to Statistical Learning with Applications in R (Berlin: Springer)

[70] Tarboton D 2015 Terrain Analysis Using Digital Elevation Models (Taudem) (Logan: Utah Water Research Laboratory, Utah State University)

[71] Haitjema H M and Mitchell-Brucker S 2005 Are water tables a subdued replica of the topography? Ground Water 43 781–6

[72] Theobald D M, Harrison-Atlas D, Monahan W B and Albano C M 2015 Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning PLoS One 10 e0143619

[73] Meng Q and Ashby S 2014 Distance: a critical aspect for environmental impact assessment of hydraulic fracturing Extr. Ind. Soc. 1 124–6

[74] Deziel N C, Brokovich E, Grotto I, Clark C J, Barnett-Itzhaki Z, Broday D and Agay-Shay K 2020 Unconventional oil and gas development and health outcomes: a scoping review of the epidemiological research Environ. Res. 182 109124

[75] Belitz K, Moore R B, Arnold T L, Sharpe J B and Starn J J 2019 Multiorder hydrologic position in the conterminous United States: a set of metrics in support of groundwater mapping at regional and national scales Water Resour. Res. 55 11188–207

[76] Aller L et al 1987 DRASTIC: A Standardized System For Evaluating Groundwater Pollution Potential Using Hydrogeologic Settings (Washington, DC: U.S. Environmental Protection Agency)

[77] Zanocco C, Boudet H, Clarke C E, Stedman R and Evansen D 2020 NIMBY, YIMBY, or something else? Geographies of public perceptions of shale gas development in the Marcellus Shale Environ. Res. Lett. 15 074039

[78] Barth-Naftilan E, Sohng J and Saiers J E 2018 Methane in groundwater before, during, and after hydraulic fracturing of the Marcellus Shale Proc. Natl Acad. Sci. 115 6970–5

[79] Siegel D I, Smith B, Perry E, Bothun R and Hollingsworth M 2015 Pre-drilling water-quality data of groundwater prior to shale gas drilling in the Appalachian Basin: analysis of the Chesapeake Energy Corporation dataset Appl. Geochem. 63 37–57

[80] Warner N R, Jackson R B, Darrah T H, Osborn S G, Down A, Zhao K, White A and Vengosh A 2012 Geochemical evidence for possible natural migration of Marcellus Formation brine to shallow aquifers in Pennsylvania Proc. Natl Acad. Sci. 109 11961

[81] EPA 2020 ChemView: chemical data reporting database under the toxic substances control act (available at: https://chemview.epa.gov/chemview) (Accessed 30 September 2020)

[82] Blondes M S et al 2018 U.S. Geological Survey National Produced Waters Geochemical Database (ver. 2.3, January 2018) [data release] (available at: 10.5066/F796W8) (Accessed 30 September 2020)

[83] Rozell D J and Reaven S J 2012 Water pollution risk associated with natural gas extraction from the Marcellus Shale Risk Anal. 32 1382–93

[84] EPA 2015 Retrospective case study in Northeastern Pennsylvania: study of the potential impacts of hydraulic fracturing on drinking water resources EPA/600/R-14/088 (Washington, DC: Office of Research and Development, US Environmental Protection Agency)