Environmental Science & Technology

pubs.acs.org/est

Article

Statistical Mapping of PFOA and PFOS in Groundwater throughout the Contiguous United States

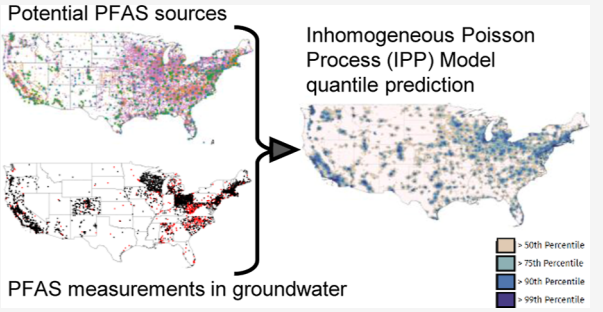
Bumjun Park,* Hyunseung Kang, and Christopher Zahasky

Cite This: *Environ. Sci. Technol.* 2024, 58, 19843−19850

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Per-and polyfluoroalkyl substances (PFAS) are synthetic chemicals that are increasingly being detected in groundwater. The negative health consequences associated with human exposure to PFAS make it essential to quantify the distribution of PFAS in groundwater systems. Mapping PFAS distributions is particularly challenging because a national patchwork of testing and reporting requirements has resulted in sparse and spatially biased data. In this analysis, an inhomogeneous Poisson process (IPP) modeling approach is adopted from ecological statistics to continuously map PFAS distributions in groundwater across the contiguous United States. The model is trained on a unique data set of 8910 PFAS groundwater measurements, using combined concentrations of two PFAS analytes. The IPP model predictions are compared with results from random forest models to highlight the robustness of this statistical modeling approach on sparse data sets. This analysis provides a new approach to not only map PFAS contamination in groundwater but also prioritize future sampling efforts.



KEYWORDS: *PFAS, groundwater, inhomogeneous Poisson process model, random forest model*

1. INTRODUCTION

Per- and polyfluoroalkyl substances (PFAS) have emerged as a growing concern in recent years due to their widespread presence in the environment and risk to ecosystems and human health.[1,2] These synthetic chemicals, many of which exhibit resistance to heat, water, and oil, have found extensive use in a variety of industrial and consumer products.[3] PFAS can be traced back to applications as diverse as nonstick cookware, firefighting foam, waterproof clothing, and food packaging. As PFAS use has become ubiquitous, concerns about their persistence, mobility, and adverse health effects have risen.[4] As drinking water is a primary vector for negative health impacts, understanding and forecasting the distribution of PFAS in groundwater is essential for mitigating negative health impacts.

Mapping and monitoring PFAS contamination in groundwater is key for enabling regulatory and groundwater management decisions to mitigate human exposure. Despite this importance, nationwide groundwater sampling and data aggregation remains a challenge. PFAS compounds originate from a diverse array of point and nonpoint sources. Point sources include applications such as firefighting foam, industrial discharges, and landfills, while nonpoint sources include activities such as biosolid spreading and wet deposition.[5−7] The array of pathways for PFAS to enter the environment makes it difficult to uniquely identify the origins of PFAS contamination.[8−10] A second barrier to widespread monitoring is the expense associated with extensive sampling and analytical quantification of PFAS in water down to parts-per-trillion (ppt) concentrations. The cost-intensive and highly technical nature of collecting and analyzing groundwater samples across a broad geographic area has limited comprehensive and spatially resolved assessments.[11]

These challenges are further amplified in the existing monitoring framework in the United States where PFAS testing and management is primarily overseen by state environmental and natural resource agencies, resulting in stark disparities in sampling and data availability. This introduces sampling biases due to opportunistic testing. Widespread groundwater sampling surveys for PFAS may occur as a result of focused studies (e.g.,[8,12,13]), however PFAS groundwater sampling generally occurs more frequently in time or space in areas with greater population, or in high-risk areas such as near industrial sites or airports.[6] If unaccounted for, a model built upon data with such sampling biases may exaggerate the PFAS risk of densely populated cities and underestimate the risks of nonpoint sources. State-level testing also leads to strong variability in data density. States such as Massachusetts, California, and New Jersey currently have vast

Received: June 5, 2024
Revised: October 15, 2024
Accepted: October 16, 2024
Published: October 23, 2024

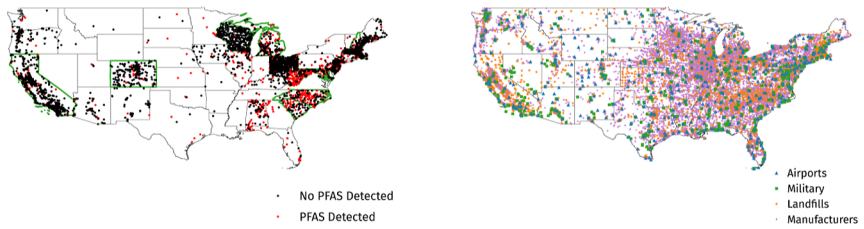


Figure 1. (Left) locations of all the 8910 PFAS observations. States with at least 2 observations per 1000 square miles, and at least one observation within every 1 by 1° pixel are considered presence-absence data and are highlighted in green (additional details in Supporting Information 1 Figure S2). (Right) compilation of potential PFAS point sources.

publicly available data sets, while others have minimal publicly available data on groundwater testing. Furthermore, for regulatory reasons, states often focus primarily on public water supply testing, introducing an additional layer of complexity when water supplies are sourced from both surface water and groundwater and may include some form of treatment.[14] These monitoring and groundwater sampling disparities result in data sparsity, sampling bias, and regional imbalances, collectively impeding efforts to synthesize a cohesive nationwide PFAS risk map.

Geostatistical modeling approaches have been widely used to produce distribution maps of groundwater contaminants—including PFAS—ranging from classical regression to machine learning methods. For contaminants such as atrazine and nitrate, there are vast nationwide groundwater sampling data sets spanning decades. These massive data sets enable robust mapping over much of the United States. For example, atrazine has been mapped over the contiguous U.S. using a Tobit linear regression model,[15] while the distribution of nitrate has been mapped using approaches including nonlinear regression,[16] logistic regression,[17] random forest,[18] and extreme gradient boosting.[19]

Unlike these contaminants where data has been collected in groundwater for decades, limitations on existing data and model development have restricted PFAS mapping to relatively confined geographic areas. The most widespread method for these smaller regional studies utilize random forest models (RFs), a machine learning method that combines multiple decision trees.[20−23] RFs have demonstrated high predictive accuracy for estimating PFAS in water resources and ecosystems at the regional scale, however when cast to a national scale, the data sparsity and sampling bias can present challenges for these models. Specifically, RFs were not designed to handle heterogeneity in sampling of the training data; RFs typically assume that each datum is an independent draw from the same underlying population to justify the bagging procedure inside RFs.[24] Previous work has also shown that when applied to groundwater data, RFs can perform poorly when the dependent variable is close to zero,[25] a characteristic of PFAS data at a national scale. Other conventional linear regression models such as the LASSO or Ridge regression model similarly fail when observations are not independent or identically distributed,[26] neither of which is true for the sparse and opportunistically sampled PFAS data.

In this work, a new approach is developed for modeling national PFAS risks in groundwater inspired by models employed for mapping plant and animal species in ecological statistics.[27−29] Issues of PFAS data sparsity and opportunistic sampling have strong parallels to the field of ecological statistics. For instance, when constructing a model for the spatial distribution of a particular bird species, it is difficult to conduct a meticulous grid search over a region. Therefore, models must rely on observation locations that are irregularly scattered across the region of interest. These models are developed to handle intrinsic sampling bias as species are more likely to be observed where there are more potential observers, such as areas of higher human density or wildlife sanctuaries. Such opportunistic data is referred to as presence-only data in that it is gathered only from sightings of presences of a species. Its counterpart, presence-absence data, is compiled from a thorough grid search examining each spatial unit for presences and absences of an organism.[30] When creating a species distribution map using both presence-only and presence-absence data, ecological statisticians apply various methods of adjustment. This study adopts one such method for mapping combined perfluorooctanoic acid (PFOA) and perfluorooctanesulfonic acid (PFOS) distributions in groundwater.

2. MATERIALS AND METHODS

The specific ecological statistics model is called the inhomogeneous Poisson process (IPP) model.[27] An IPP model is governed by the intensity function $\lambda(p)$ where p represents the set of spatial observations or events in a region. This intensity function represents the expected number of observations within a small area around a point. Or, it could be interpreted as the relative probability of observing an event at one given point compared to another. For example, if the intensity at one point is twice the intensity at another, it is twice as likely to observe an event at that point compared to the other. An IPP model built upon the intensity function $\lambda(p)$ is denoted as $\mathrm{IPP}(\lambda(p))$.

In addition to the intensity function $\lambda(p)$, a bias function is used to account for a mixture of presence-only data and presence-absence data. Presence-only data is when the vast majority of PFAS sampling in groundwater occurs in known or suspected contaminated aquifers. Presence-absence data is considered PFAS data from states that conducted more systematic groundwater testing. Areas with greater bias are more likely to be sampled compared to other areas. When the bias function is considered, the IPP model is represented as $\mathrm{IPP}(\lambda(p)b(p))$ where the bias function $b(p)$ either intensifies or attenuates the intensity function.

Both the intensity and the bias functions depend on a set of spatial covariates. This study assumes a set of 15 covariates that may affect both the intensity and the bias. Such an assumption is made on the basis that PFAS sampling is more likely in areas that are suspected to have higher PFAS intensity, causing $b(p)$ and $\lambda(p)$ to be affected by similar covariates. The covariates that are considered are the distances to suspected PFAS sources such as airports large enough to require firefighting training drills (part 139) or military bases that regularly train with firefighting foam containing PFAS, landfills, and various

https://doi.org/10.1021/acs.est.4c05616
Environ. Sci. Technol. 2024, 58, 19843−19850

manufacturers that produce or frequently use PFAS. Potential PFAS contamination from these types of sources has been well documented.[2,6,31] These locations are compiled using resources such as the North American Industry Classification System (NAICS) and the Facility Registry Service, and illustrated in the right pane of Figure 1. The specific list of covariates, their units of measurement, sources, relevant summary statistics, and reasons for inclusion are detailed in Tables S1, S2 and S3 of the Supporting Information 1. In addition to potential PFAS sources, covariates include hydrologic process information such as precipitation, and human geography information such as population density and median income. Using these covariates, we use maximum likelihood estimation to estimate the parameters of the model. Additional details about the model are available in the Supporting Information 1 Section S6.

This set of covariates neglects additional geologic and hydrologic parameters that are known to have an important influence on rates of PFAS loading and transport in groundwater systems such as depth to the water table and the presence of cocontaminants in the vadose zone,[32−37] the extent of aquifer confinement, soil/bedrock properties,[38] and water chemistry conditions.[39,40] While most of this information is not reported or available for PFAS testing locations, neglecting these parameters is expected to only limit the ability of a model to capture rates of spreading in the subsurface. The combination of these transport limitations, the resolution of this national level model, and the lack of constraint on the timing of most PFAS sources into the environment, highlights that this model is intended to represent a relatively static risk map based largely on knowledge of potential source loading into the environment.

The IPP model is trained on a unique data set of 8910 PFAS measurements compiled and preprocessed from state and national databases and web dashboards using the most recently available data at the time of data collection. Other data preprocessing during the compilation stages included removing duplicate locations and missing entries, masking exact locations, and other quality control steps. These steps were necessary to ensure consistency in data management standards and maintaining a coherent data structure. Despite extensive data gathering efforts, there is not yet enough data at a national level to perform a meaningful time-series or spatiotemporal analysis. The exclusion of the temporal aspect is also considered justifiable due to the persistent nature of PFAS in groundwater. Details of each data source, including ones that were excluded from the data, are included in the Supporting Information 2. The full data set is available in an online repository.[41]

In this data set, PFAS observations are defined as the sum of PFOS and PFOA. These two analytes are chosen as they are the most commonly reported and are of specific regulatory focus of state and national agencies. All observations with greater than 8 parts per trillion (ppt) of total PFOA and PFOS detected were considered presence data. For example, if there was 8 ppt of PFOA and PFOS was not detected, the sample would be classified as a presence. On the other hand, if there was 7 ppt of PFOA and 0.9 ppt of PFOS, the sample would not be a presence. While potentially masking individual relationships with each analyte, combining these two analytes allows a more comprehensible investigation of the general relationship between the covariates and PFAS presence in groundwater. Utilizing combined concentrations can account for a wide range of analytical detection and reporting standards while illustrating general PFAS risk. The 8 ppt threshold was selected to align with the sum of the maximum contaminant levels of 4 ppt each for PFOA and PFOS set by the US Environmental Protection Agency (EPA) to establish legally enforceable limits on PFAS in groundwater.[42] The extent and density of these PFAS measurements is illustrated in the left map of Figure 1. The exact distribution of concentrations within the data set are displayed in Figure S1 of Supporting Information 1.

Out of the 8910 observations, this study assumes that data from a select number of states with at least 2 observations per 1000 square miles, and at least one observation within every 1 by 1° pixel, is presence-absence data. The states of California, Colorado, Maryland, Massachusetts, Michigan, New Hampshire, New Jersey, North Carolina, Ohio, Rhode Island, South Carolina, West Virginia, and Wisconsin met these criteria and thus their data were classified as presence-absence. Data collected from other states with more sparse and opportunistic PFAS testing were classified as presence-only data (see additional details in Supporting Information 1 Figure S2). To interpolate the PFAS intensity over the contiguous United States, 50,000 points were randomly generated to establish where the IPP model was evaluated. The number of randomly generated points was chosen to provide adequate model interpolation at the national scale. This number of points was sufficient for producing national maps. However, future users of this methodology may use more points to provide an intensity map with a finer spatial resolution.

In addition to the IPP model, a random forest model was fit using the same set of data to illustrate differences in model robustness to sparse and biased data. Rather than producing estimates of species intensity, the RF model produces a probability estimate for the chances of observing PFAS at each point. To highlight the robustness of RF and IPP approaches to opportunistically collected data, an additional set of models were fit to an intentionally skewed data set. In this comparison, 50% of the lowest PFAS concentration observations were dropped, and the remaining 4455 observations were used to fit the second IPP and RF models. For all cases, the same 50,000 randomly generated points were used for evaluating the models across the contiguous United States.

The RF models were fit in R (R version 4.3.2) using the randomForest (4.7.1.1) package,[43] using 500 decision trees, 3 variables per split, a minimum node size of 1, and no restriction on the number of maximum terminal nodes. The 100% RF model had a sensitivity of 0.64, which means 64% of the actual detects were correctly identified, and specificity of 0.81, which means 81% of the nondetects were correctly identified. The 50% RF model had a sensitivity of 0.97 and specificity of 0.36 (Table 1). The importance of each covariate as weighted by the RF model is included in Supporting Information 1 Table S6.

The fitted IPP and RF models were both evaluated on the same 50,000 randomly generated data points to create maps over the contiguous United States. Triangulated irregular network (TIN) interpolation was used on the modeled points and the 50,000 background points to create a network of triangles between these points to interpolate values in between. The specific details of how each model treats the data is included in Supporting Information 1 Section S6. A comparison with a map created from Kriging, another popular geostatistical method for spatial interpolation, is available in Supporting Information 1 Figure S5. All mapping and

https://doi.org/10.1021/acs.est.4c05616
Environ. Sci. Technol. 2024, 58, 19843−19850

Table 1. Confusion Matrices for 100% RF Model (Left) and 50% RF Model (Right)

		predicted	
		0	1
actual	0	4203	993
	1	1336	2378

		predicted	
		0	1
actual	0	268	473
	1	96	3618

interpolation was performed using the QGIS software[44] in the WGS 84 coordinate reference system. This CRS was chosen as it is the standard system used by the United States Geological Survey for national level maps.[45] The code and data used for analyses are available in the following repository.[41]

3. RESULTS AND DISCUSSION

The results of the IPP and RF models with full and skewed data sets are illustrated in Figure 2. The left two maps in Figure 2 show the predicted unbiased intensity of combined PFOA and PFOS produced by the IPP model. The upper map is fitted using the entire data set while the bottom map uses the skewed data set where 50% of the lowest combined PFOA and PFOS observations were not included in the model development. The intensity maps represent the expectation of how often PFAS detects occur in small locations. The intensity is truncated to a maximum intensity of one, the point at which we expect to make at least one observation of PFOA and/or PFOS within a unit pixel (0.01 by 0.01°, approximately 1.1 km by 1.1 km). The right two maps show the predicted probability of observing PFOA and/or PFOS produced by the RF model using 100% and 50% of the data. While the intensity signifies where the detects are more common, probability signifies how likely an event is at a given location.

When comparing the maps using all of the available data, the robustness of the IPP model is striking. Both models correctly highlight areas with known PFAS risk such as St. Paul, MN, Chicago, IL, and other industrialized cities. However, the RF map overemphasizes the risk in areas from which more observational data was available. In the RF maps, the states of MA, NJ, and WV exhibit much higher predicted probabilities than neighboring regions because of the abundance of measurements relative to states with less data. In the RF models another issue is apparent in regions from which very few data points were collected. Despite vast distances from potential sources and the paucity of PFAS observations, rural regions in Nevada, the northern Rocky Mountains, and the northern Great Plains are questionably flagged as high-risk areas by the RF model. In comparison, the IPP model is much more specific in its predictions, as indicated by the higher intensity predictions in localized areas.

In the predicted distributions relying on skewed data, the IPP model robustness becomes even more evident. The RF model was severely challenged by the bias that was introduced, losing most of its specificity and highlighting virtually every region as PFAS probability close to one. On the other hand, the IPP model produced a nearly identical intensity prediction to the model that relied on 100% of the available data. More precise examinations of how the predicted intensities and probabilities shifted as data was removed are included in Supporting Information1 Figure S4.

In addition to these intensity estimates, the IPP model produces an estimate of sampling bias. The bias function $b(p)$ represents the expected proportion of PFAS observations near point p that are included in the presence-only data. In essence, given that two points have identical intensity, the point with greater bias will have more PFAS detections compared to the other. Thus, the estimated bias function estimates the pattern of opportunistic sampling in PFAS testing and can be used to identify regions that would most benefit from additional groundwater sampling. The estimated bias map is included in Supporting Information1 Figure S3.

To further inspect the IPP model and illustrate the potential practical utility, Figure 3 shows the intensity of PFAS colorized based on percentiles. Regions in the darkest shade of blue are in the 99th percentile of expected number of observations of PFAS, while regions in white are below the 50th percentile. Not only does this map correctly identify clusters of observed

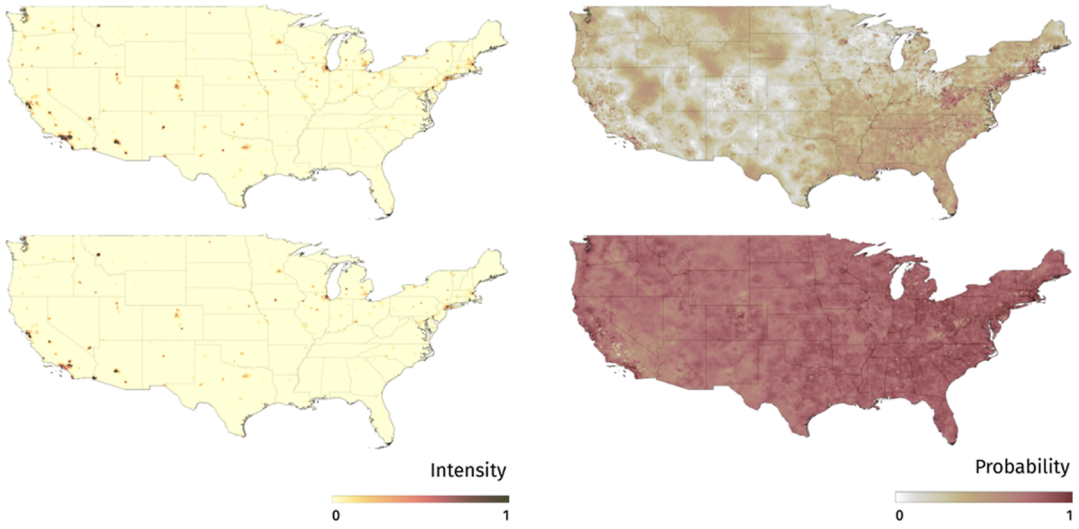


Figure 2. Predicted combined PFOA and PFOS intensity produced by the IPP model (left) and the predicted combined PFOA and PFOS detection probability produced by the RF model (right). Using 100% of the data (top) or using 50% of the data (bottom).

https://doi.org/10.1021/acs.est.4c05616
Environ. Sci. Technol. 2024, 58, 19843−19850

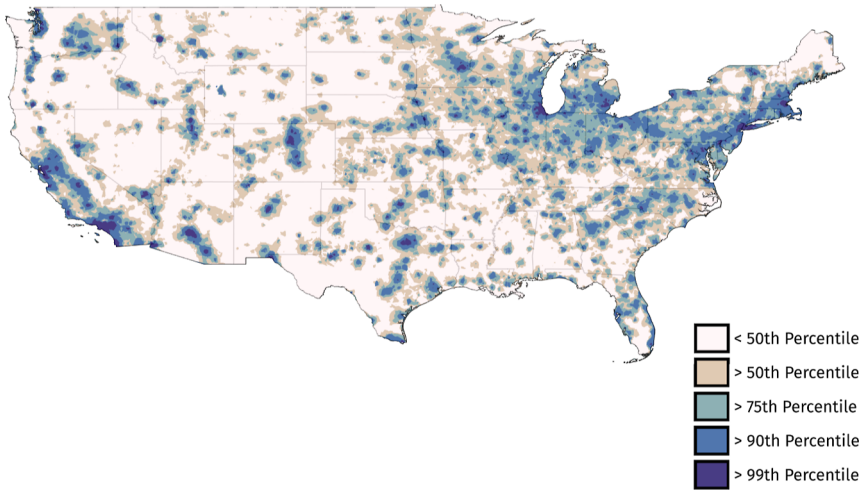


Figure 3. Percentile map of PFAS intensity from IPP model.

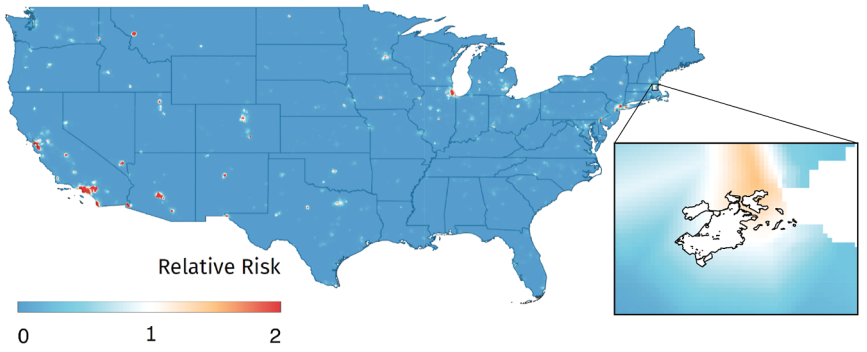


Figure 4. Relative risk of combined PFOA and PFOS detection compared to the average of the city of Boston, MA (top) and a closer look around the city of Boston (inset). Relative risk was truncated at 2.

PFAS in groundwater samples from Figure 1, it also identifies regions from which little, if any, data was provided. Despite the lack of PFAS observations, this map also highlights many regions of elevated PFAS risk that could be used to better inform future sampling campaigns. For example, while central Washington has limited groundwater sampling data in this model data set, the high PFAS intensity could inform and corroborate PFAS ecosystem and bioacculation studies.[20]

Furthermore, an analysis of the specific covariates allows an inspection of PFAS risk even in absence of this map. For example, among the strongest covariates affecting PFAS intensity were distances to military bases, leather manufacturers, or landfills. As the distance from these facilities increases, the intensity, the likelihood of observing PFAS at a certain location, would decrease. Exact analysis on how each of the 15 covariates would affect PFAS intensity and by how much are included in Sections S7 and S8 of Supporting Information 1.

The practicality of the IPP model extends further beyond its robustness to skewed and biased data. The PFAS intensity is easily converted to a more interpretable metric of relative risk by taking the ratio of locations relative to each other. For example, if one pixel has twice the intensity of another, the risk of observing PFAS there would be twice as high. The city of Boston, MA has extensive water sampling and publicly available data in terms of PFAS documentation and management so it could be used as the benchmark to compute a relative risk over the contiguous United States. Figure 4 illustrates this relative risk across the contiguous U.S. relative to Boston, MA. The average intensity across the city boundaries of Boston, as seen in the inset of Figure 4 was set as the reference, with a relative risk of one. The red regions in the map have two times or greater chance of observing PFAS compared to Boston, while the blue regions have a comparatively negligible chance. This example of relative risk can be performed not only at the national level but also at a more local level. This could be especially valuable for the allocation of sampling, remediation, and other technical resources at a local, regional, or national scale.

The application of IPP statistical approaches to forecasting PFAS risk in this study provides a new approach to overcome the inherent issues and limitations of groundwater sampling for PFAS and other emerging contaminants. The resulting PFAS distribution maps highlight known contamination regions and identify regions where future sampling campaigns could be focused. In comparison with random forest model predictions, the IPP model predictions were extremely robust. This is clear from qualitative assessment of national scale predictions, where the RF model forecasts broad regional elevated risks in some of the most undersampled and sparsely populated regions of the United States. These effects were amplified when the data was further skewed so that only the top 50% of PFAS observations were retained. In this case the RF model prediction was significantly degraded while the IPP model remained relatively robust as indicated by minimal changes in the relative intensity

https://doi.org/10.1021/acs.est.4c05616
Environ. Sci. Technol. 2024, 58, 19843−19850

distribution relative to the model results that utilized all of the available data.

Omission of 50% of the data serves as a proxy for cross-validation, a method commonly used to assess how well a model will perform on new and unseen data by intentionally withholding data, or other machine learning performance metrics. In the case of point processes, the concept of new and unseen data does not apply as the 8910 observations are considered a single observation of 8910 potentially related points, invalidating the purpose of cross-validation. Rather, by intentionally skewing the data by deleting the bottom half of PFAS observations, the stability and reproducibility of each model is tested, as models should be robust to such forms of sampling bias. As detailed in Figure S4 of Supporting Information 1, the IPP model exhibits such stability while the RF model does not.

While this IPP model has shown promising results, this approach has limitations. First, due to the decentralized nature of PFAS sampling and management, the data used in this study lacks uniformity. The data was gathered from multiple sources and state-level databases, all of which had different PFAS quantification, analytes sampled, detection limits, and reporting standards. A nationwide uniform PFAS sampling scheme would improve this and future model efforts. Second, this study simplifies the convoluted structure of PFAS distribution. For the sake of statistical modeling, this study only considered 15 continuous covariates, disregarding other potentially influential factors. Additional studies may use this same framework to evaluate the effect of other possible covariates such as more granular hydrogeologic conditions or additional PFAS source information. Third, for the sake of comprehensibility, this study summed PFOA and PFOS together to observe an overarching relationship between these common PFAS and the covariates. Using this novel methodology to its fullest extent, with more fine-grained data, more detailed relationships between each individual analytes may be examined in future studies. Finally, even if all of the relevant covariates are included, the IPP model is a parametric model. If the relationship between spatial distribution of PFAS and spatial covariates are nonlinear and complex, the IPP model predictions may be inaccurate. Despite these possibly restrictive parametric assumptions, the IPP model outperforms common machine learning approaches, primarily because the proposed model properly adjusts for the sparse sampling and data heterogeneity inherent in the data set. Therefore, the problem of model mis-specification may not be a major concern. However, future work may focus on expanding to semiparametric or nonparametric IPP models.

Despite these limitations, this study provides a unique and extensive data set and model to better understand the state of PFAS contamination in the contiguous United States. In addition, this approach provides new insights and methodologies for devising PFAS sampling schemes. If facing time and cost restrictions, stakeholders may choose to target only the areas of high PFAS intensity since the IPP model inherently accounts for sampling likelihood which in turn provides robustness against opportunistic sampling. Even in the absence of a meticulous presence-absence sampling within a region, this analysis provides an important starting point for gathering more information about the distribution of PFAS in groundwater systems across the United States.

■ ASSOCIATED CONTENT

Data Availability Statement

All code and data used for analyses are included on GitHub (*bpark67/mapping-pfas*) and Zenodo.[41] In addition, maps of Figures 3 and 4 in the manuscript are uploaded on a separate Web site.[46]

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.4c05616.

Supporting Information 1, 12 sections and 11 pages long, containing a list of covariates and data sources, summary statistics and visualizations, various covariate analyses, other maps and visualizations excluded from the manuscript. Supporting Information 2, one spreadsheet detailing each PFAS data source, method of access, date of access, URL (if applicable), and reason for exclusion (if applicable) (PDF) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Bumjun Park − **Department of Biostatistics, University of Washington, Seattle, Washington 98195, United States; Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States;* orcid.org/0009-0008-0361-3810; Email: bpark67@uw.edu

Authors

Hyunseung Kang − *Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States*

Christopher Zahasky − *Department of Geoscience, University of Wisconsin−Madison, Madison, Wisconsin 53706, United States;* orcid.org/0000-0002-3427-5622

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c05616

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based upon work supported in part by the National Science Foundation under grant no EAR 2054263. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

■ REFERENCES

(1) Fenton, S. E.; Ducatman, A.; Boobis, A.; DeWitt, J. C.; Lau, C.; Ng, C.; Smith, J. S.; Roberts, S. M. Per- and Polyfluoroalkyl Substance Toxicity and Human Health Review: Current State of Knowledge and Strategies for Informing Future Research. *Environ. Toxicol. Chem.* **2021**, *40* (3), 606−630.

(2) Evich, M. G.; Davis, M. J. B.; McCord, J. P.; Acrey, B.; Awkerman, J. A.; Knappe, D. R. U.; Lindstrom, A. B.; Speth, T. F.; Tebes-Stevens, C.; Strynar, M. J.; Wang, Z.; Weber, E. J.; Henderson, W. M.; Washington, J. W. Per- and polyfluoroalkyl substances in the environment. *Science* **2022**, *375* (6580), No. eabg9065.

(3) Kebede, M. M.; Terry, L. G.; Clement, T. P.; Mekonnen, M. M. Mapping Per- and Polyfluoroalkyl Substance Footprint from Cosmetics and Carpets across the Continental United States. *ACS ES&T Water* **2024**, *4*, 3882−3892.

(4) Cousins, I. T.; Johansson, J. H.; Salter, M. E.; Sha, B.; Scheringer, M. Outside the Safe Operating Space of a New Planetary Boundary for Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci. Technol.* **2022**, *56* (16), 11172−11179.
(5) Masoner, J. R.; Kolpin, D. W.; Cozzarelli, I. M.; Smalling, K. L.; Bolyard, S. C.; Field, J. A.; Furlong, E. T.; Gray, J. L.; Lozinski, D.; Reinhart, D.; Rodowa, A.; Bradley, P. M. Landfill leachate contributes per-/poly-fluoroalkyl substances (PFAS) and pharmaceuticals to municipal wastewater. *Environ. Sci.: Water Res. Technol.* **2020**, *6* (5), 1300−1311.
(6) Salvatore, D.; Mok, K.; Garrett, K. K.; Poudrier, G.; Brown, P.; Birnbaum, L. S.; Goldenman, G.; Miller, M. F.; Patton, S.; Poehlein, M.; Varshavsky, J.; Cordner, A. Presumptive Contamination: A New Approach to PFAS Contamination Based on Likely Sources. *Environ. Sci. Technol. Lett.* **2022**, *9* (11), 983−990.
(7) Bierbaum, T.; Klaas, N.; Braun, J.; Nürenberg, G.; Lange, F. T.; Haslauer, C. Immobilization of per- and polyfluoroalkyl substances (PFAS): Comparison of leaching behavior by three different leaching tests. *Sci. Total Environ.* **2023**, *876*, 162588.
(8) Silver, M.; Phelps, W.; Masarik, K.; Burke, K.; Zhang, C.; Schwartz, A.; Wang, M.; Nitka, A. L.; Schutz, J.; Trainor, T.; Washington, J. W.; Rheineck, B. D. Prevalence and Source Tracing of PFAS in Shallow Groundwater Used for Drinking Water in Wisconsin, USA. *Environ. Sci. Technol.* **2023**, *57* (45), 17415−17426.
(9) Balgooyen, S.; Remucal, C. K. Tributary Loading and Sediment Desorption as Sources of PFAS to Receiving Waters. *ACS ES&T Water* **2022**, *2* (3), 436−445.
(10) Thompson, K. A.; Mortazavian, S.; Gonzalez, D. J.; Bott, C.; Hooper, J.; Schaefer, C. E.; Dickenson, E. R. V. Poly- and Perfluoroalkyl Substances in Municipal Wastewater Treatment Plants in the United States: Seasonal Patterns and Meta-Analysis of Long-Term Trends and Average Concentrations. *ACS ES&T Water* **2022**, *2* (5), 690−700.
(11) Smalling, K. L.; Romanok, K. M.; Bradley, P. M.; Morriss, M. C.; Gray, J. L.; Kanagy, L. K.; Gordon, S. E.; Williams, B. M.; Breitmeyer, S. E.; Jones, D. K.; DeCicco, L. A.; Eagles-Smith, C. A.; Wagner, T. Per- and polyfluoroalkyl substances (PFAS) in United States tapwater: Comparison of underserved private-well and public-supply exposures and associated health implications. *Environ. Int.* **2023**, *178* (March), 108033.
(12) Hohweiler, K. A.: *Occurrence of Per- and Polyfluoroalkyl Substances (PFAS) in Private Water Supplies in Southwest Virginia*, 2023. [Online; accessed 9-8-2024].
(13) Buszka, P. M.; Mailot, B. E.; Mathes, N. A. *Per- and Polyfluoroalkyl Substances in Groundwater from the Great Miami buried-valley Aquifer, Southwestern Ohio, 2019−20*; Scientific Investigations Report, 2023; .
(14) Sims, J. L.; Stroski, K. M.; Kim, S.; Killeen, G.; Ehalt, R.; Simcik, M. F.; Brooks, B. W. Global occurrence and probabilistic environmental health hazard assessment of per- and polyfluoroalkyl substances (pfass) in groundwater and surface waters. *Sci. Total Environ.* **2022**, *816*, 151535.
(15) Stackelberg, P. E.; Barbash, J. E.; Gilliom, R. J.; Stone, W. W.; Wolock, D. M. Regression models for estimating concentrations of atrazine plus deethylatrazine in shallow groundwater in agricultural areas of the united states. *J. Environ. Qual.* **2012**, *41* (2), 479−494. https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2134/jeq2011.0200
(16) Nolan, B. T.; Hitt, K. J. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the united states. *Environ. Sci. Technol.* **2006**, *40* (24), 7834−7840.
(17) Gurdak, J. J.; Qi, S. L. Vulnerability of recently recharged groundwater in principle aquifers of the United States to nitrate contamination. *Environ. Sci. Technol.* **2012**, *46* (11), 6004−6012.
(18) Pennino, M. J.; Leibowitz, S. G.; Compton, J. E.; Hill, R. A.; Sabo, R. D. Patterns and predictions of drinking water nitrate violations across the conterminous United States. *Sci. Total Environ.* **2020**, *722*, 137661.
(19) Ransom, K. M.; Nolan, B. T.; Stackelberg, P. E.; Belitz, K.; Fram, M. S. Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States. *Sci. Total Environ.* **2022**, *807*, 151065.
(20) DeLuca, N. M.; Mullikin, A.; Brumm, P.; Rappold, A. G.; Cohen Hubal, E. Using geospatial data and random forest to predict pfas contamination in fish tissue in the columbia river basin, united states. *Environ. Sci. Technol.* **2023**, *57* (37), 14024−14035.
(21) Fernandez, N.; Nejadhashemi, A. P.; Loveall, C. Large-scale assessment of pfas compounds in drinking water sources using machine learning. *Water Res.* **2023**, *243*, 120307.
(22) George, S.; Dixit, A. A machine learning approach for prioritizing groundwater testing for per-and polyfluoroalkyl substances (pfas). *J. Environ. Manage.* **2021**, *295*, 113359.
(23) Hu, X. C.; Ge, B.; Ruyle, B. J.; Sun, J.; Sunderland, E. M. A statistical approach for identifying private wells susceptible to perfluoroalkyl substances (pfas) contamination. *Environ. Sci. Technol. Lett.* **2021**, *8* (7), 596−602.
(24) Hastie, T.; Tibshirani, R.; Friedman, J. *Random Forests. in: The Elements of Statistical Learning*; Springer: New York, NY, USA, 2008; pp 587−604..−
(25) Belitz, K.; Stackelberg, P. E. Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environ. Model. Softw.* **2021**, *139*, 105006.
(26) Hastie, T.; Tibshirani, R.; Friedman, J. *Shrinkage Methods. in: The Elements of Statistical Learning*; Springer: New York, NY, USA, 2008; pp 61−79..−
(27) Fithian, W.; Elith, J.; Hastie, T.; Keith, D. A. Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods Ecol. Evol.* **2015**, *6* (4), 424−438.
(28) Renner, I. W.; Elith, J.; Baddeley, A.; Fithian, W.; Hastie, T.; Phillips, S. J.; Popovic, G.; Warton, D. I. Point process models for presence-only analysis. *Methods Ecol. Evol.* **2015**, *6* (4), 366−379.
(29) Warton, D. I.; Shepherd, L. C. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Stat.* **2010**, *4* (3), 1383.
(30) Aarts, G.; Fieberg, J.; Matthiopoulos, J. Comparative interpretation of count, presence−absence and point methods for species distribution models. *Methods Ecol. Evol.* **2012**, *3*, 177−187.
(31) Moody, C. A.; Field, J. A. Perfluorinated Surfactants and the Environmental Implications of Their Use in Fire-Fighting Foams. *Environ. Sci. Technol.* **2000**, *34* (18), 3864−3870.
(32) Brusseau, M. L. Assessing the potential contributions of additional retention processes to PFAS retardation in the subsurface. *Sci. Total Environ.* **2018**, *613−614*, 176−185.
(33) Costanza, J.; Arshadi, M.; Abriola, L. M.; Pennell, K. D. Accumulation of PFOA and PFOS at the Air-Water Interface. *Environ. Sci. Technol. Lett.* **2019**, *6* (8), 487−491.
(34) Brusseau, M. L.; Yan, N.; Van Glubt, S.; Wang, Y.; Chen, W.; Lyu, Y.; Dungan, B.; Carroll, K. C.; Holguin, F. O. Comprehensive retention model for PFAS transport in subsurface systems. *Water Res.* **2019**, *148*, 41−50.
(35) Guo, B.; Zeng, J.; Brusseau, M. L. A Mathematical Model for the Release, Transport, and Retention of Per- and Polyfluoroalkyl Substances (PFAS) in the Vadose Zone. *Water Resour. Res.* **2020**, *56* (2), 1−21.
(36) Brusseau, M. L. Examining the robustness and concentration dependency of PFAS air-water and NAPL-water interfacial adsorption coefficients. *Water Res.* **2021**, *190*, 116778.
(37) Gnesda, W. R.; Draxler, E. F.; Tinjum, J.; Zahasky, C. Adsorption of PFAAs in the Vadose Zone and Implications for Long-Term Groundwater Contamination. *Environ. Sci. Technol.* **2022**, *56* (23), 16748−16758.
(38) Adamson, D. T.; Nickerson, A.; Kulkarni, P. R.; Higgins, C. P.; Popovic, J.; Field, J.; Rodowa, A.; Newell, C.; DeBlanc, P.; Kornuc, J. J. Mass-Based, Field-Scale Demonstration of PFAS Retention within AFFF-Associated Source Areas. *Environ. Sci. Technol.* **2020**, *54* (24), 15768−15777.

(39) Higgins, C. P.; Luthy, R. G. Sorption of perfluorinated surfactants on sediments. *Environ. Sci. Technol.* **2006**, *40* (23), 7251−7256.
(40) Guelfo, J. L.; Higgins, C. P. Subsurface transport potential of perfluoroalkyl acids at aqueous film-forming foam (AFFF)-impacted sites. *Environ. Sci. Technol.* **2013**, *47* (9), 4164−4171.
(41) Park, B. *bpark67/mapping-pfas: Code and Data for "Statistical Mapping of PFOA and PFOS in Groundwater throughout the Contiguous United States*; Online; accessed-21-May-2024, 2024; ..
(42) *U.S. EPA: Per- and Polyfluoroalkyl Substances (PFAS)*. U.S. EPA 2024.
(43) Liaw, A.; Wiener, M. Classification and regression by randomforest. *R News* 2002, *2*(3), 18−22.
(44) QGIS Development Team *QGIS Geographic Information System. QGIS Association*; QGIS Association, 2023. https://www.qgis.org.
(45) *What map projections are used in The National Map tiled base map services and dynamic overlay services?. U.S. Geological Survey*. [Online; accessed 9. Aug. 2024] (2024). https://www.usgs.gov/faqs/what-map-projections-are-used-national-map-tiled-base-map-services-and-dynamic-overlay.
(46) Park, B. *bpark67/webmap-est: Figures 3 and 4 in "Statistical Mapping of PFOA and PFOS in Groundwater throughout the Contiguous United States"*; Online; accessed-21-Aug-2024, 2024; .. https://bpark67.github.io/webmap-est/

https://doi.org/10.1021/acs.est.4c05616
Environ. Sci. Technol. 2024, 58, 19843−19850