Exposure and Health
https://doi.org/10.1007/s12403-023-00581-w

ORIGINAL PAPER

Objective Cost-Informed Cutoff Criteria Improve the Utility of Machine Learning Models of Environmental Hazards: A Case Study of Groundwater Arsenic Distribution in India

Ruohan Wu[1] · **David A. Polya**[1]

Received: 1 June 2023 / Revised: 1 June 2023 / Accepted: 5 June 2023

Abstract
Although there are an increasing number of artificial intelligence/machine learning models of various hazardous chemicals (e.g. As, F, U, NO_3^-, radon) in environmental media (e.g. groundwater, soil), these most commonly use arbitrarily selected cutoff criteria to balance model specificity and sensitivity. This results in models of hazard distribution that, whilst often of considerable interest and utility, are not designed to optimize cost benefits of the mitigation of those hazards. In this case study, building upon recent machine learning modelling of the geographical distribution of groundwater arsenic in India, we show that the use of objective cost-informed criteria not only results in (i) different cutoff values for the classification of areas as of high or low groundwater arsenic hazard but also, more importantly, (ii) a reduction of overall potential (mitigation + testing + health impacts) costs. Further, we show that the change in optimal cutoff values and the reduction in overall costs vary from state to state depending upon locally specific classification-dependent costs, the prevalence of high arsenic groundwaters, the heterogeneity of the distribution of those high arsenic groundwaters, and the extent to which inhabitants are exposed to the hazard. It follows more generally that using cost-optimized criteria will result in different, more objective, and more cost-relevant appropriate balances being made between specificity and sensitivity in modelling environmental hazard distribution in different regions. This indicates also the utility of developing machine learning models at an appropriate local (e.g. country, state, district) scale rather than more global scales in order to better inform local-scale mitigation strategies.

Keywords Environmental contaminants · Machine learning · Cost optimization · Groundwater · Arsenic · India

Introduction

Machine learning models are increasingly widely used to predict or interpolate the spatial distribution of hazardous chemical constituents of a wide variety of environmental media, including groundwater (e.g. Amini et al. 2008; Winkel et al. 2008; Park et al. 2016; Tesoriero et al. 2017; Podgorski et al. 2018, 2022; Podgorski and Berg 2020; Chakraborty et al. 2020; Tan et al. 2020; Mukherjee et al. 2021; Erickson et al. 2021; Lombard et al. 2021; Wu et al. 2020, 2021a, 2021b; Perović et al. 2021; Connolly et al. 2021; Cao et al. 2022; Kumar and Pati 2022; Ottong et al. 2022; Knierim et al. 2022; Dhamija and Joshi 2022) and soils (Lado et al. 2008; Li et al. 2022; Hengl et al. 2017; Mikkonen et al. 2018; de Menezes et al. 2020; Jia et al. 2021; Kebonye et al. 2021).

Whilst the distribution of environmental chemical hazards has been determined by logistic regression models (e.g. arsenic in groundwater (Wu et al. 2020); fluoride in groundwater (Podgorski et al. 2018) or boosted regression trees (e.g. arsenic in groundwater (Tan et al. 2020); and nitrate in groundwater (Sajed-Hosseini et al. 2018)) there is an increasing preponderance of supervised random forest models, involving assemblies of decision trees. Because of the widely acknowledged (Millot et al. 2011; Bhattacharya et al. 2017; Bretzler et al. 2017b; UNICEF/WHO 2018; Polya et al. 2019) importance of arsenic in groundwater utilized as drinking water in contributing to massive detrimental public health outcomes, such random forest

✉ Ruohan Wu
wrhruohan@hotmail.com

✉ David A. Polya
david.polya@manchester.ac.uk

[1] Department of Earth and Environmental Sciences, School of Natural Sciences and Williamson Research Centre for Molecular Environmental Sciences, University of Manchester, Manchester M13 9PL, UK

Published online: 10 July 2023

models have been widely used to predict the distribution of arsenic contamination in groundwater. These models have been rendered at various spatial scales, including the global scale: (Podgorski and Berg 2020); the country scale: India (Podgorski et al. 2020; Wu et al. 2021b; Mukherjee et al. 2021), China (Rodríguez-Lado et al. 2013); Burkino Faso (Bretzler et al. 2017a); Uruguay (Wu et al. 2021a); and more local scales: Gujarat state (Wu et al. 2020); Varanasi (Uttar Pradesh) (Kumar and Pati 2022), and Puralia (West Bengal) (Ruidas et al. 2022).

Machine learning random forest models are used to predict binary target dependent variables (e.g. high (value assigned = 1) or low (value assigned = 0) hazard compared to a user-defined hazard concentration). These models generate four different types of prediction: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) with respect to the binary target variable. In the context of classifying an environmental hazard variable as "high" or "low", model sensitivity (also known as the true-positive rate) refers to the proportion of truly "high" hazard values being correctly modelled as being "high"; whilst model specificity (also known as the true-negative rate), refers to the proportion of truly "low" hazard values being correctly modelled as being "low". In general, there is a trade-off between model sensitivity and model specificity.

Thus, all machine learning models which ultimately classify areas as being of "high" or "low" hazard require some criterion to be used to determine the balance to be made between model sensitivity and model specificity or between other measures of model accuracy (see Table 1). Typically, the criteria used in previously published studies of machine learning modelling of the spatial distributions of chemical contaminants in environmental media have been based on either (i) adopting a cutoff value where specificity = sensitivity (Podgorski et al. 2020; Wu et al. 2021a) or (ii) simply using a cutoff value of 0.5 (Wu et al. 2020, 2021b). Although these cutoff criteria are convenient to use, there is no objective reason to choose one criterion over another, other than perhaps intellectual or artistic elegance. Further, as shown in Table 1, there are many other criteria that could and have been used as the basis for determining cutoff values: again, the selection of one of these cutoff criteria in preference to another would largely seem to be matter of convenience or personal preference rather than being based on any documented comprehensive objective reasoning.

In contrast to current published models of environmental chemical hazard distribution, models of public health relevant tests that aim to classify patients for the purposes of indicating preferred methods of treatment (or, indeed, of non-treatment) (Brinati et al. 2020), now widely use cutoff criteria that explicitly take into account the relative utility or costs of false-positive and false-negative classifications. The substantial disparity that may exist between the costs of false positives (e.g. arising from recalling non-diseased patients for unnecessary further diagnostic tests or treatments) and false negatives (e.g. arising from detrimental health outcomes arising from the failure to promptly treat a disease) frequently gives rise to cutoff values that are very different from those that would arise from giving explicit or implicit equal weighting to sensitivity and specificity. This is reflected in the widespread adoption during the early stage of the COVID-19 pandemic of rapid antigen detection tests with relatively low sensitivities compared to specificities (ECDC 2021) and, in contrast, elsewhere, recognition that using cutoff criteria that result in relatively high numbers of false positives can make screening for certain disease states as not being cost-effective (e.g. Sharib et al. 2020). Habibzadeh et al. (2016) amongst others indicates the importance of factoring in the costs of model-based misdiagnosis. Medical decision theory has long emphasized (i) the importance

Table 1 List of commonly used criteria for determining cutoff values for machine learning models. (Modified after Lopez-Raton et al. (2019))

No	Short name for criterion	Description	References
1	Default	Default cutoff of 0.5	Wu et al. (2021a, b); (2020)
2	SpEquSe	Cutoff where sensitivity is equal to specificity	Hosmer and Lemeshow (2000)
3	PPVquNPV	Cutoff where positive prediction value (PPV) is equal to negative prediction value (NPV)	Podgorski et al. (2020); Vermont et al. (1991)
4	Youden	Cutoff based on Youden's Index, defined as $YI_{(C)} = max_c(Se_{(c)} + Sp_{(c)} - 1)$ (Youden 1950)	Youden (1950); Lewis et al. (2008), Greiner (1995, 1996)
5	ROC01	Cutoff on the point on the ROC curve closest to the point (0,1)	Metz (1978)
6	MaxKappa	Cutoff with the max Kappa Index	Cohen (1960); Greiner et al. (2000); see also Feinstein (1975); Galen (1986); Greiner (1995, 1996)
7	PreMatch	Cutoff value with the equality of sample prevalence (p) and predicted prevalence, defined as $pSe_{(c)} + (1-p)(1-Sp_{(c)})$	Manel et al. (2001); Kelly et al. (2008)
8	MaxNPVPPVProd	Cutoff maximizing the product of PPV and NPV	Lopez-Raton et al. (2019)

of utilizing cost-optimized cutoff criteria (e.g. Phelps and Mushlin 1988), (ii) that that would generally give rise to unequal weighting to sensitivity and specificity (e.g. Gail and Pfeiffer 2005), and (iii) that "regret probability", defined as 1—the positive predictive value (PPV) (Maxim et al. 2014), may vary substantially for the same test depending upon the prevalence of the disease state in the population (Grimes and Schulz 2002; Maxim et al. 2014).

Few if any of the commonly used cutoff criteria methods for random forest modelling of the spatial distribution of environmental chemical hazards explicitly take into account the relative costs of misclassifying the hazard, indeed, they would seem to implicitly either ignore cost implications or assume that there is little material difference in the cost consequences of false-positive and false-negative classifications. As such, they are not designed to optimize utility taking into account combined mitigation, testing, and health impact costs. This begs the questions, does this really matter? And if it does matter, then how much does it matter? In particular, under what circumstances, is it materially important to consider the criteria used to obtain optimal cutoff values?

To test and illustrate the importance of using objective cost-optimized cutoff criteria, we present our analysis of a case study related to the machine learning modelling of the 2-D spatial distribution of groundwater arsenic in India. Our analysis is built upon our previous modelling (Wu et al. 2021b). We demonstrate that the use of such cost-optimized criteria not only results in the selection of numerically different cutoff values but also enables considerable reduction in overall potential mitigation/testing/health costs. Wider implications for the machine learning modelling of environmental chemicals of public health significance are discussed.

Methodology

Machine Learning Model

A binary target variable, groundwater arsenic, for India, with two possible values—"high" or "low" was determined using random foresting modelling, utilizing 145,813 geolocated (longitude/latitude) secondary groundwater arsenic concentration measurements from India and its neighbouring countries (Bangladesh, Pakistan, and Nepal) and 31 potential environmental predictors. The WHO provisional guide value for arsenic in drinking water, viz. 10 µg/L was adopted as the concentration used to classify concentrations as being either "high" or "low" in arsenic. The random forest model has been utilized to create a map at 1-km^2 (pixel) resolution of the predicted probability of groundwater arsenic exceeding 10 µg/L (Fig. 1a). The process of the random forest model generation and description of the dependent and independent variables is outlined here and provided in detail in our previous studies (Wu et al. 2021b, 2020; Podgorski et al. 2020).

Comparison of Non-cost-optimized Probability Cutoffs

A number of non-cost-optimized methods, as listed in Table 1, for determining cutoffs were separately used with the India groundwater arsenic random forest model to calculate the method-optimized cutoffs, together with their corresponding sensitivity, specificity, positive prediction values, and negative prediction values following the methods otherwise detailed in Wu et al. (2021b).

Calculation of Misclassification and Overall Costs

The misclassification costs, $Cost_{FP+FN}$, of a model can be expressed (Thiele and Hirschfeld 2020) as the sum of the costs arising from false-positive and false-negative classifications according to Eq. (1):

$$Cost_{FP+FN} = N_{FPpixels} \times \overline{CR_{pixel,FP}} + N_{FNpixels} \times \overline{CR_{pixel,FN}}, \quad (1)$$

where

$N_{FPpixels}$ and $N_{FNpixels}$ are the number of pixels, misclassified as a false positives and false negatives, respectively;

$\overline{CR_{pixel,FP}}$ and $\overline{CR_{pixel,FN}}$ are the weighted arithmetic mean per-pixel cost arising from misclassification of a pixel as a false positive or as a false negative, respectively.

We define here also, the overall costs, $Cost_{FP+FN+TP}$, of a model as the sum of costs arising from all false- and true-positive and negative model classifications and assuming that costs arising from a true-negative classification can be assumed to be zero, according to

$$\begin{aligned} Cost_{FP+FN+TP} = {} & N_{FPpixels} \times \overline{CR_{pixel,FP}} \\ & + N_{FNpixels} \times \overline{CR_{pixel,FN}} \\ & + N_{TPpixels} \times \overline{CR_{pixel,TP}}, \end{aligned} \quad (2)$$

where

$N_{FPpixels}$, $N_{FNpixels}$, and $N_{TPpixels}$ are the number of pixels classified as FP, FN, and TP, respectively;

$\overline{CR_{pixel,FP}}$, $\overline{CR_{pixel,FN}}$,, and $\overline{CR_{pixel,TP}}$ are the weighted arithmetic mean per-pixel cost arising from misclassification of a pixel as a false positive or as a false negative or classification as a true positive, respectively.

In practice, $\overline{CR_{pixel,FP}}$, $\overline{CR_{pixel,FN}}$,, and $\overline{CR_{pixel,TP}}$ are inconvenient to calculate, so for the case study of India groundwater arsenic, we calculated $Cost_{FP+FN}$ and $Cost_{FP+FN+TP}$ using the following methodology which renders the same results as Eqs. (1) and (2), respectively. We note that India is administratively divided into several

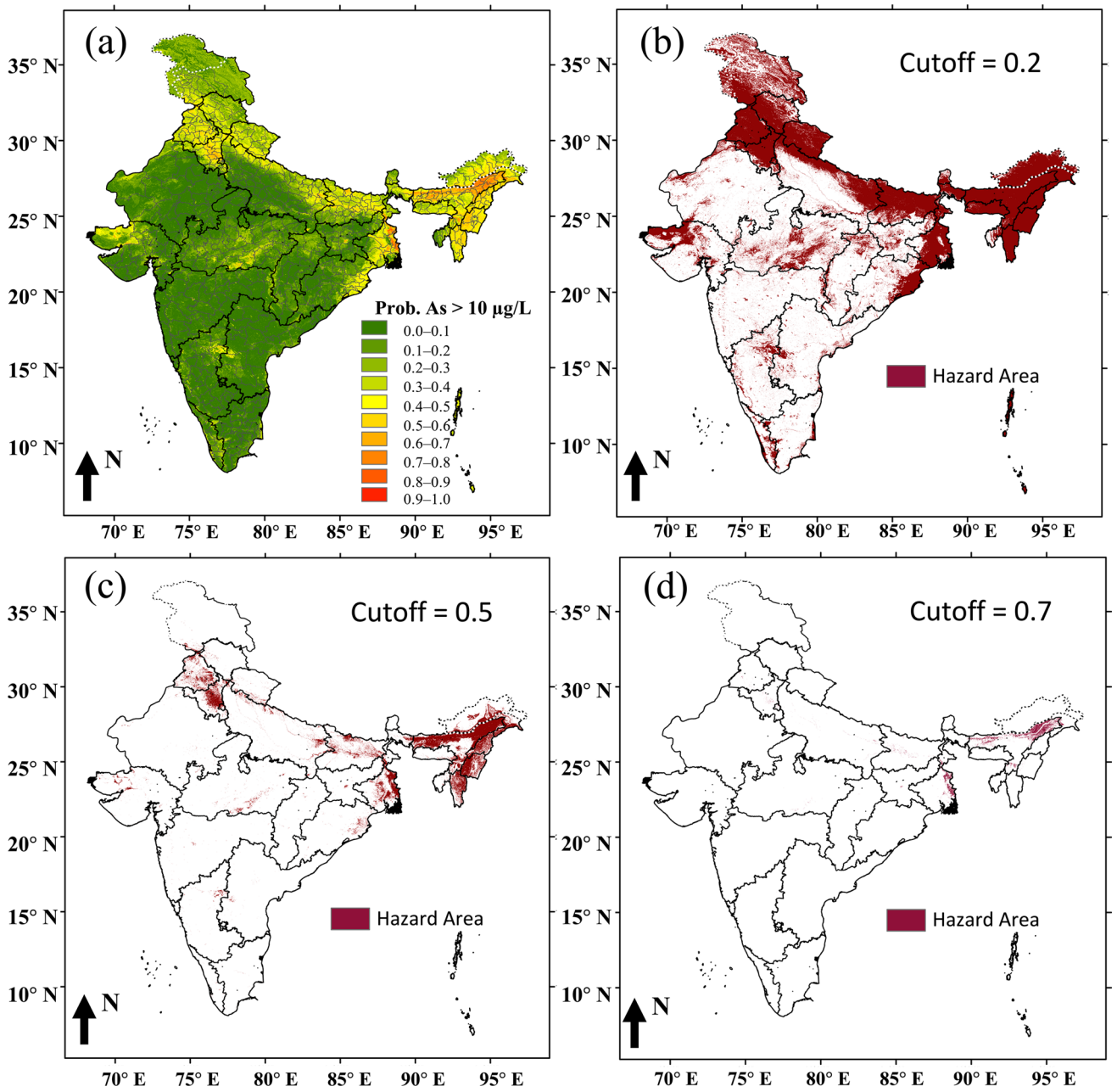

Fig. 1 Distribution of arsenic in groundwater in India as determined by random forest modelling. **a** Probability map of groundwater arsenic exceeding 10 µg/L (from Wu et al. 2021b; reproduced here under the terms of a Creative Commons CC-BY Licence). **b**–**d** Map of "high" groundwater arsenic hazard arising from using probability cutoffs of **b** 0.2, **c** 0.5, or **d** 0.7

hundred districts, each composed of a number of 1 km^2 pixels for which we modelled groundwater arsenic status. Rounding errors arising from the imperfect alignment of some 1 km^2 pixels with district boundaries were determined to be of only minor importance in the context of this study. We further note that costs (related to well testing or well water remediation) arising from false-positive and true-positive classification will largely be incurred on a per well basis, whereas those (related to detrimental health outcomes) arising from a false-negative classification will largely be incurred on a per groundwater arsenic-exposed person basis.

Misclassification costs, $Cost_{FP+FN}$, were calculated according to

$$Cost_{FP+FN} = \sum \left(N_{FP,\ district} \times \frac{P_{district} \times Pro_{GW,district} \times Wells_{district}}{N_{pixel,district}} \right) \times CR_{wells,FP} + \sum \left(N_{FN,district} \times \frac{P_{district} \times Pro_{GW,district}}{N_{pixel,district}} \right) \times CR_{people,FN}, \quad (3)$$

where

$N_{FP,district}$ and $N_{FN,district}$ are the number of FP and FN classified pixels, respectively, based on the random forest model of groundwater arsenic exceeding 10 $\mu g/L$;

$P_{district}$ is the population of the subscripted district;

$Pro_{GW,district}$ is proportion of population in the subscripted district drinking groundwater via tubewells/boreholes and hand pumps;

$N_{pixel,district}$ is the total number of pixels in the subscripted district;

$Wells_{district}$ is the prevalence of wells utilized for drinking in subscripted district; and

$CR_{wells,FP}$ and $CR_{people,FN}$ are the relative unit costs for each well and each groundwater arsenic-exposed person.

Values of $P_{district}$, $Pro_{GW,district}$, and $N_{pixel,district}$ were obtained or derived from [India Census, 2011], whilst $Wells_{district}$ was approximated by assuming a ratio of 1 drinking water well per 20 people based on whole country estimates of population and wells [Government of India 2011a, b; CGWB 2022].

Overall costs, $Cost_{FP+FN+TP}$, were calculated according to

$$Cost_{FP+FN+TP} = \sum \left(N_{FP,district} \times \frac{P_{district} \times Pro_{GW,district} \times Wells_{district}}{N_{pixel,district}} \right) \times CR_{wells,FP} + \sum \left(N_{FN,\ district} \times \frac{P_{district} \times Pro_{GW,\ district}}{N_{pixel,\ district}} \right) \times CR_{people,\ FN} + \left(N_{TP,\ district} \times \frac{P_{district} \times Pro_{GW,\ district} \times Wells_{district}}{N_{pixel,\ district}} \right) \times CR_{wells,\ TP} \quad (4)$$

where

$N_{TP,district}$ is the number of TP classified pixels, respectively, in the subscripted district based on the random forest model of groundwater arsenic exceeding 10 $\mu g/L$ in India;

And, $CR_{wells,TP}$ is the relative unit cost arising for TP classified pixels on a per groundwater arsenic-exposed person basis.

Comparison of Costs Arising Using Cost-Optimized Cutoffs Compared to a Default Cutoff

Both misclassification and overall cost proportion difference comparisons between the use of cost-optimized cutoff values and a default cutoff of 0.5 were calculated for various ratios of CR_{FP}:CR_{FN} (for misclassification costs and discrete cost ratio value selected, see Table S1) and ratios of CR_{FP}:CR_{FN}:CR_{TP} (for discrete cost ratio values selected for overall costs calculations, see Table S2) according to Eqs. (5) and (6). These comparisons were made both on an all India basis and also for the individual states of Assam, Gujarat, Uttar Pradesh, and West Bengal states, which collectively exhibit a wide range of prevalence of high groundwater arsenic.

$$CPD_1 = \frac{Cost_{0.5} - Cost_{cost-optimizedcutoff}}{Cost_{0.5}}, \quad (5)$$

$$CPD_2 = \frac{Cost_{0.5} - Cost_{cost-optimizedcutoff}}{Cost_{cost-optimizedcutoff}}, \quad (6)$$

where

CPD_1 and CPD_2 are the cost proportion differences using Eqs. (5) and (6), respectively;

$Cost_{0.5}$ is the cost (misclassification or overall) arising from the use of a default cutoff of 0.5 (cf. Wu et al. 2021b); and $Cost_{cost-optimizedcutoff}$ is the cost (misclassification or overall) arising from the cost-optimized cutoff value.

Illustrative Example of Selecting Cost-Optimal Cutoff for Groundwater Arsenic in India

In order to estimate the potential cost savings arising from using a cost-optimized cutoff model as opposed to a default cutoff value, we further used estimated unit costs for well testing (FP), well remediation (TP), and detrimental groundwater arsenic-attributable health outcomes (FN) detailed in Table 2. For illustrative purposes, the costs of a true-positive (TP) classification were taken to be costs of remediation for each "high" arsenic well, the numbers of which were calculated as above; the costs of a false-positive (FP) classification were taken to be the costs of diagnostic testing for each "high" arsenic classified well, the costs of a false-negative (FN) classification were to be costs of health lives lost as result of unmitigated exposure to "high" arsenic groundwater and determined on a per well-user basis. The figures adopted here are broadly based on Wu et al. (2021b) supplemented by discussions with technology (remediation and chemical analysis) providers in India. The unit costs are intended to be illustrative not definitive, and, in any event, the actual unit costs may vary substantially from place to place, well-user to well-user and from time to time.

Results and Discussion

Machine Learning Model of Groundwater Arsenic Distribution in India

The random forest model for India of the probability of groundwater arsenic exceeding 10 µg/L is shown in Fig. 1a. A description and discussion of the characteristics of the distribution have been discussed previously in considerably

Table 2 Adopted illustrative potential unit costs arising from classification and misclassification of machine learning model of groundwater arsenic exceeding 10 µg/L

	Classified as high arsenic hazard	Classified as low arsenic hazard
Actual high arsenic hazard	Remediation targeted at one of more of hazard (source), exposure route or receptor indicator, costs likely to be highly variable depending upon local circumstances, 1,000,000 INR (*TP*; see text)	Model error may lead to avoidable arsenic-attributable detrimental health impacts, costs of which estimated at a notional 10,000,000 INR per well-user (*FN*; see text)
Actual low arsenic hazard	Model error may lead to cost of confirmatory arsenic testing estimated at 1000 INR per well (*FP*; see text)	No action indicated (*TN*; see text)

Table 3 Comparison of cutoff, sensitivity, specificity, positive prediction values (PPV), and negative prediction values (NPV) arising from applying non-cost-optimized probability cutoff criteria listed in Table 1 to the random forest model of distribution of groundwater arsenic exceeding 10 µg/L in India

Cutoff method	Cutoff	Sensitivity	Specificity	Positive prediction value (PPV)	Negative prediction value (NPV)
Default	0.500	0.944	0.957	0.940	0.960
SpEquSe	0.474	0.952	0.952	0.935	0.966
PPVequNPV	0.554	0.930	0.966	0.951	0.951
Youden	0.398	0.970	0.939	0.919	0.978
ROC01	0.452	0.958	0.949	0.930	0.970
MaxKappa	0.467	0.955	0.951	0.933	0.967
PreMatch	0.510	0.941	0.958	0.942	0.958
MaxNPVPPVProd	0.599	0.917	0.973	0.960	0.942

more detail (Wu et al. 2021b) and is not the focus of the current study. This probability distribution gives rise to very different overall areas classified as "high" (> 10 µg/L) groundwater arsenic hazard depending upon on the value of the probability cutoff value selected, viz. 0.2 (Fig. 1b), 0.5 (Fig. 1c), or 0.7 (Fig. 1d). These illustrated cutoff values encompass the range of cutoff values (0.4 to 0.6) arising from the use of commonly used non-cost-optimized cutoff criteria (Table 1) and for which the corresponding sensitivity, specificity, positive prediction values (PPV), and negative prediction values (NPV) are shown in Table 3. Clearly, the use of different cutoff criteria gives rise both to different cutoff values (Table 3) and to different extents of areas classified as "high" groundwater arsenic hazard (Fig. 1). It is further noted that, for higher cutoffs, specificity, and positive prediction value increase, whilst sensitivity and negative prediction value decrease. Further, for higher cutoffs, there is a decrease in the relative number of false positives but an increase in the relative number of false negatives.

Cost-Optimized Cutoffs as Function of Misclassification Costs

Calculated cost-optimized cutoffs for the whole of India groundwater arsenic distribution model as function of the relative costs of false-positive and false-negative classifications expressed as $CR_{wells,FP}$:$CR_{people,FN}$ are shown in Table S1.

As the cost ratio $CR_{wells,FP}$:$CR_{people,FN}$ increases from 1:1000 to 1000:1, the cost-optimized cutoffs become larger, increasing from just above zero, 0.00, to 0.68. The relationship between these cutoff values and $\log(CR_{wells,FP}:CR_{people,FN})$ is further illustrated in Fig. 2c. The equivalent relationships for the states of Gujarat (Fig. 2a), Uttar Pradesh (Fig. 2b), West Bengal (Fig. 2d), and Assam (Fig. 2e) states, where the prevalence of high arsenic in groundwaters are different, are also shown. In all cases, the relationships are monotonic increasing and sigmoidal in form, although over much of the range considered they can be roughly approximated by first-order linear fits with very high (> 0.94) R^2. The associations between cost-optimized cutoffs and $\log(CR_{wells,FP}:CR_{people,FN})$ can serve as a guide to choosing cost-optimized cutoff values when the cost ratio $CR_{wells,FP}:CR_{people,FN}$ is known.

For a given value of $\log(CR_{wells,FP}:CR_{people,FN})$, the calculated cost-optimized cutoff value is a strong function of the prevalence of high arsenic groundwaters in the area being considered. For example, for $\log(CR_{wells,FP}:CR_{people,FN}) = 0$, the calculated cost-optimized cutoff values increase monotonically with prevalence of high arsenic groundwaters as follows: Gujarat (cutoff value 0.27; prevalence 0.4%); Uttar Pradesh (0.28, 3%), India (0.30, 8%), West Bengal (0.30, 30%), and Assam (0.41, 67%).

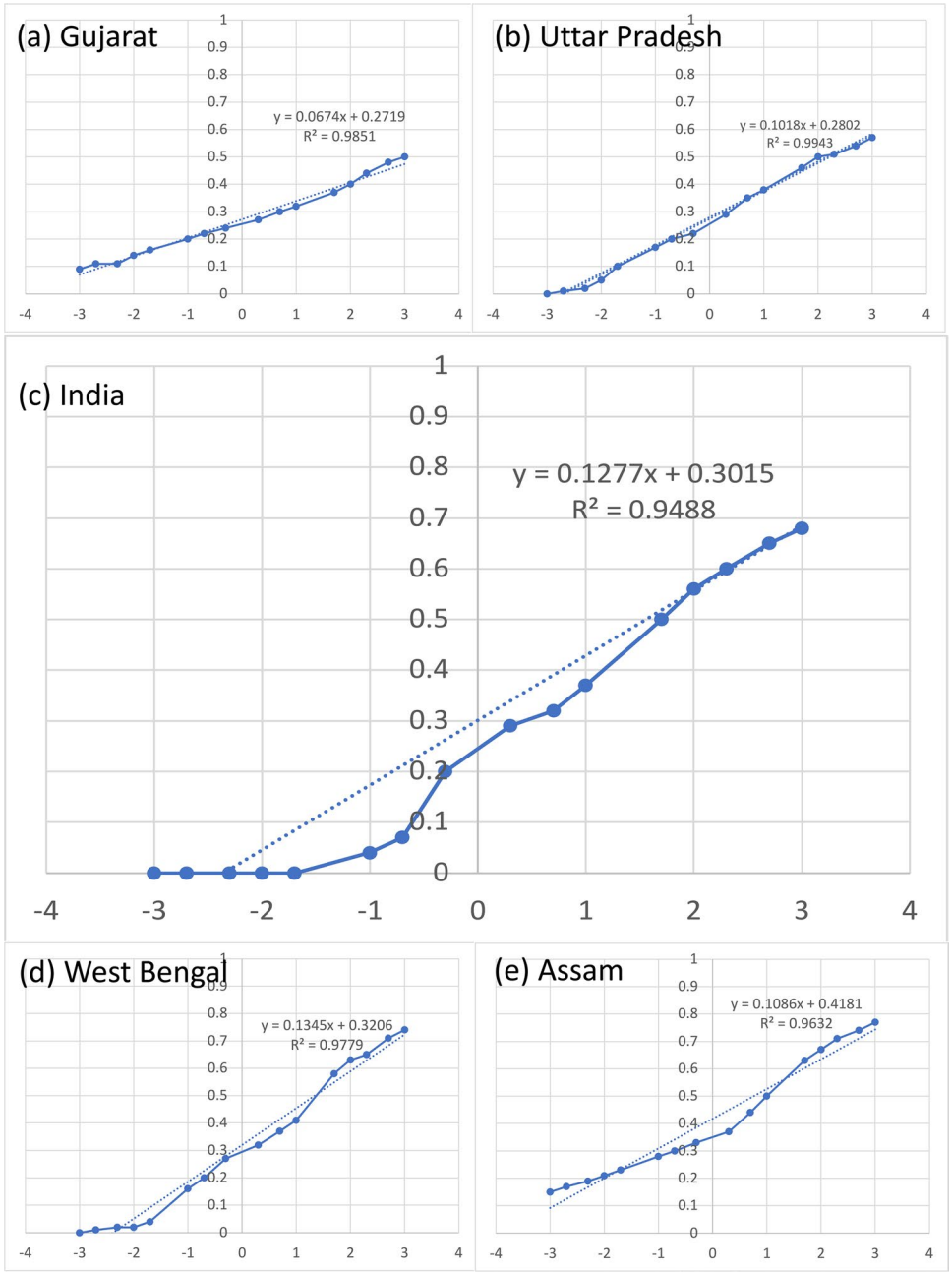

Fig. 2 Relationship between cost-optimized probability cutoffs and the relative costs of false-positive and false-negative classification expressed as log ($CR_{wells,FP}$:$CR_{people,FN}$) (see text for explanation) for **a** Gujarat, **b** Uttar Pradesh **c** India, **d** West Bengal, and **e** Assam, based on the machine learning model of the distribution of groundwater arsenic in India

Misclassification Costs

Misclassification costs ($Cost_{FP+FN}$) as function of probability cutoffs (101 values ranging from 0 to 1 with interval 0.01) are plotted for each discrete selected cost ratio in Fig. 3. Particularly if the ratio of the unit FP ($CR_{wells,FP}$) and FN ($CR_{people,FN}$) relative costs is very large or very small, (e.g. $CR_{wells,FP}$:$CR_{people,FN}$ of 1:1000, 1000:1, 1:500, 500:1, 1:200, and 200:1), then the misclassification costs tend to have very high values at cutoff values of 0 or 1, respectively, with the lowest misclassification costs occurring at cutoff values close to 1 or 0, respectively. Where the ratio of FP ($CR_{wells,FP}$) and FN ($CR_{people,FN}$) relative costs is between 2 and 100, however, the plotted curves are more obviously "U" shaped with the lowest misclassification costs arising from cutoff values in the range 0.2–0.6 that is closer to the widely used default value of 0.5.

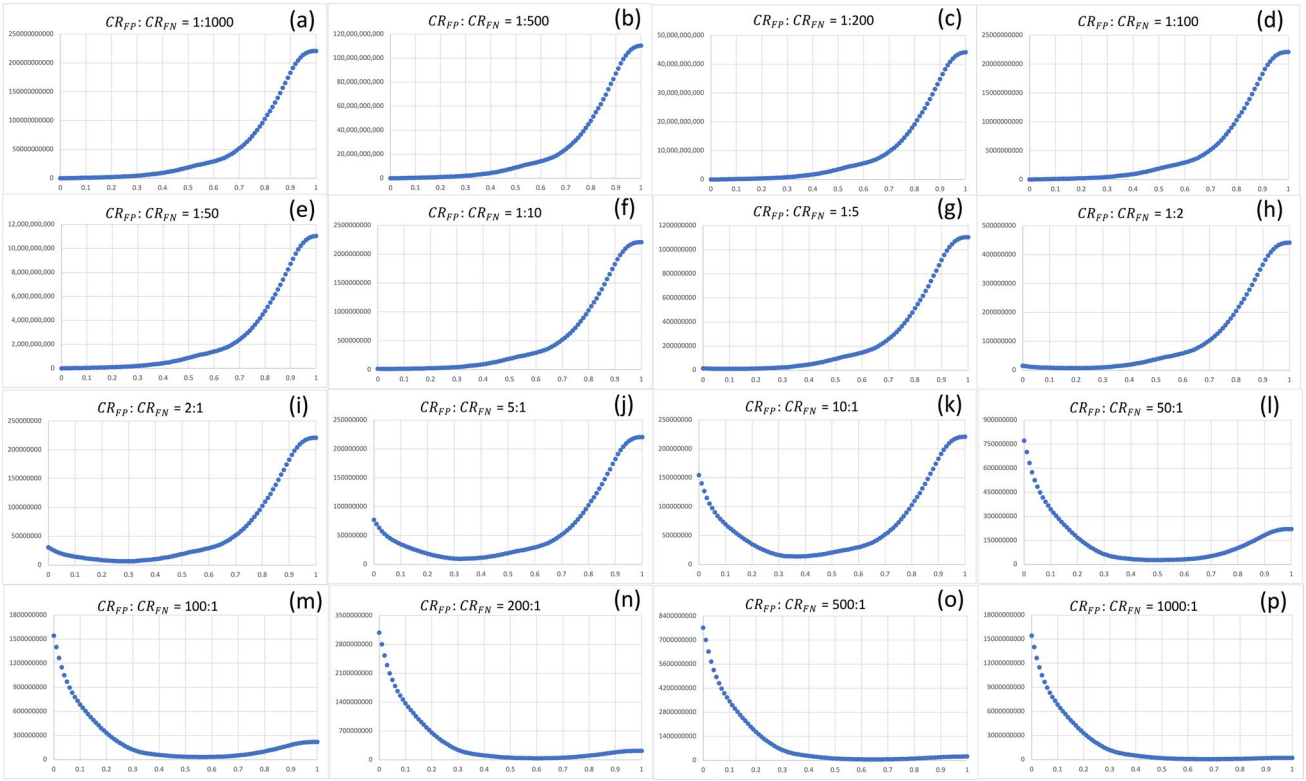

Fig. 3 Misclassification costs (Y-axis) as a function of cutoff value for discrete $CR_{wells,FP}$:$CR_{people,FN}$ ratios (X-axis) ranging from 1:1000 to 1000:1, viz. **a** 1:1000, **b** 1:500, **c** 1:200, d 1:100, e 1:50, **f** 1:10, **g** 1:5, **h** 1:2, **i** 2:1, **j** 5:1, **k** 10:1, **l** 50:1, **m** 100:1, **n** 200:1, **o** 500:1, **p** 1000:1. Calculated considering only costs arising from false-positive and false-negative misclassification costs

Overall Model-Dependent Costs

Cost-optimized cutoffs as a function of $CR_{wells,FP}$:$CR_{people,FN}$ and $CR_{wells,FP}$:$CR_{wells,TP}$ using the sets of cost ratio values tabulated in Table S2 are tabulated in Table S3 and illustrated in Fig. 4. These cost-optimized cutoffs varied from 0 to 1 and increased with both increasing $CR_{wells,FP}$:$CR_{people,FN}$ and decreasing $CR_{wells,FP}$:$CR_{wells,TP}$. Of these cutoffs, more than half of them exceeded the commonly used default cutoff value of 0.5. In all cases, the relationships are monotonic increasing and sigmoidal in form, although over much of the range considered they can be roughly approximated by first-order linear fits.

Comparison of Costs Arising Using Cost-Optimized Cutoffs Compared to a Default Cutoff

The misclassification costs (arising from FP and FN) from using cost-optimized cutoffs compared with these costs that would arise from the use of a default cutoff of 0.5 for each cost ratio are shown in Fig. 5. When these differences are expressed as a percentage of the costs arising from the use of a default cutoff of 0.5 (Fig. 5a–e), the relative costs vs cost ratios ($CR_{wells,FP}$:$CR_{people,FN}$) curves tend to be "V" shaped, with the minimum value arising when the cost ratio ($CR_{wells,FP}$:$CR_{people,FN}$) gives rise to the default cutoff of 0.5 also being the cost-optimized cutoff. Obviously, this only occurs for a particular value of ($CR_{wells,FP}$:$CR_{people,FN}$) and, in general, the cost-optimized cutoff will be different and gives rise to increasing substantial misclassification costs as the difference between the two cutoffs increases.

Interestingly, for the random forest model of groundwater arsenic exceeding 10 µg/L in this study, the cost ratio ($CR_{wells,FP}$:$CR_{people,FN}$) at which the default cutoff value would also be the cost-optimized cutoff value is clearly a strong function of the prevalence of high groundwater arsenic. The equality of costs arising from the use of the cost-optimized cutoff and that of the default cutoff of 0.5 for all of India and Gujarat, Uttar Pradesh, West Bengal, and Assam states was found to occur for cost ratios ($CR_{wells,FP}$:$CR_{people,FN}$) in the range of 10:1 to 1000:1, approximately as follows: Gujarat: 1000 (high groundwater arsenic prevalence 0.4%), Uttar Pradesh: 100 (3%), the whole of India: 50 (8%), West Bengal: a value between 10 and 50 (30%), and Assam: 10 (67%). Thus, the cost ratio values, where 0.5 is the cost-optimized cutoff with lowest misclassification cost, decrease with the increasing of prevalence of high

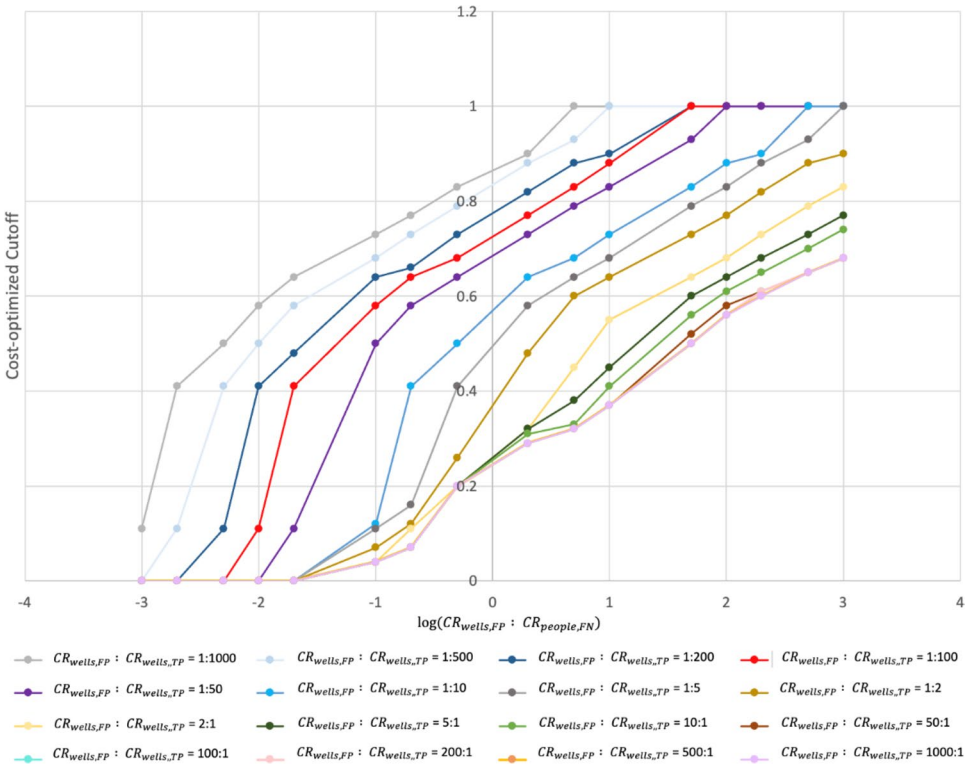

Fig. 4 Relationship between cost-optimized probability cutoffs and log ($CR_{wells,FP}$:$CR_{people,FN}$) at different discrete $CR_{people,FN}$: $CR_{wells,TP}$ ratios (see Table S2 for CR values selected for each point) for the random forest model of groundwater arsenic distribution in India. Note that where the ratio of costs arising from a false positive are very low compared to those for a false negative, the cost-optimized probability cutoff tends to 0 (which tends to classify all samples as "high" groundwater arsenic)

groundwater arsenic for the 5 different regions considered here.

Where the cost differences (CPD_2) are expressed relative to the costs arising from the use of cost-optimized cutoffs (Fig. 5f–j) similar relationships are observed, with the major difference being that (i) the magnitude of relative cost difference is much greater than when expressed relative to costs arising from the use of the default cutoff of 0.5 and (ii) the shape of the resultant curves are much more asymmetrical.

It is clear that, for this case study, in addition to misclassification and overall cost ratios, the actual prevalence of high groundwater arsenic concentrations also materially impacts the selection of cost-optimized cutoffs. When the high groundwater arsenic prevalence is low (e.g. Gujarat state), the proportion of FP misclassified pixels tends to be high. In contrast, when the groundwater arsenic prevalence is high (e.g. Assam state), the proportion of FN misclassified pixels tends to be low.

Further, it is evident that, for whole country models of groundwater arsenic, the inclusion of sub-regions (e.g. states) with highly different prevalence of high groundwater arsenic means that data from high groundwater arsenic prevalence states (e.g. Assam) will impact the model for low groundwater arsenic prevalence states (e.g. Gujarat) and vice versa. We speculate, therefore, that whole country modelling may not give the best cost-optimized models for smaller sub-regions and that global models may not give the best cost-optimized models for individual countries, particularly where there are wide differences in the prevalence of high arsenic groundwaters and where mechanisms leading to such high arsenic groundwaters may vary different from place to place (cf. Wu et al. 2021a). Interestingly, this is very analogous to the conclusions of Chen et al. (2020), albeit that their study was with respect to sub-populations of pregnant women with highly different likelihoods of bearing children with trisomy or open neural tube defects.

Illustrative Example of Selecting Cost-Optimal Cutoff for Groundwater Arsenic in India

Using the specific illustrative unit cost values ($CR_{wells,FP}$, $CR_{people,FN}$, and $CR_{wells,TP}$) listed in Table 2), a cost-optimized cutoff value between 0.00 and 0.01 was determined for the all India random forest model of groundwater arsenic (Fig. 6). The, perhaps surprising, closeness of this cutoff value to zero is due to the unit treatment and other costs for individual people at risk of suffering avoidable arsenic-attributable detrimental health impacts, $CR_{people,FN}$, being substantially higher than the unit costs of well remediation or of chemical analytical testing. An implication of this all India model is that the whole of India should be classified as an area of high groundwater arsenic to optimize model-related costs (health treatment, well remediation, chemical analysis); however, an all India model may not be the

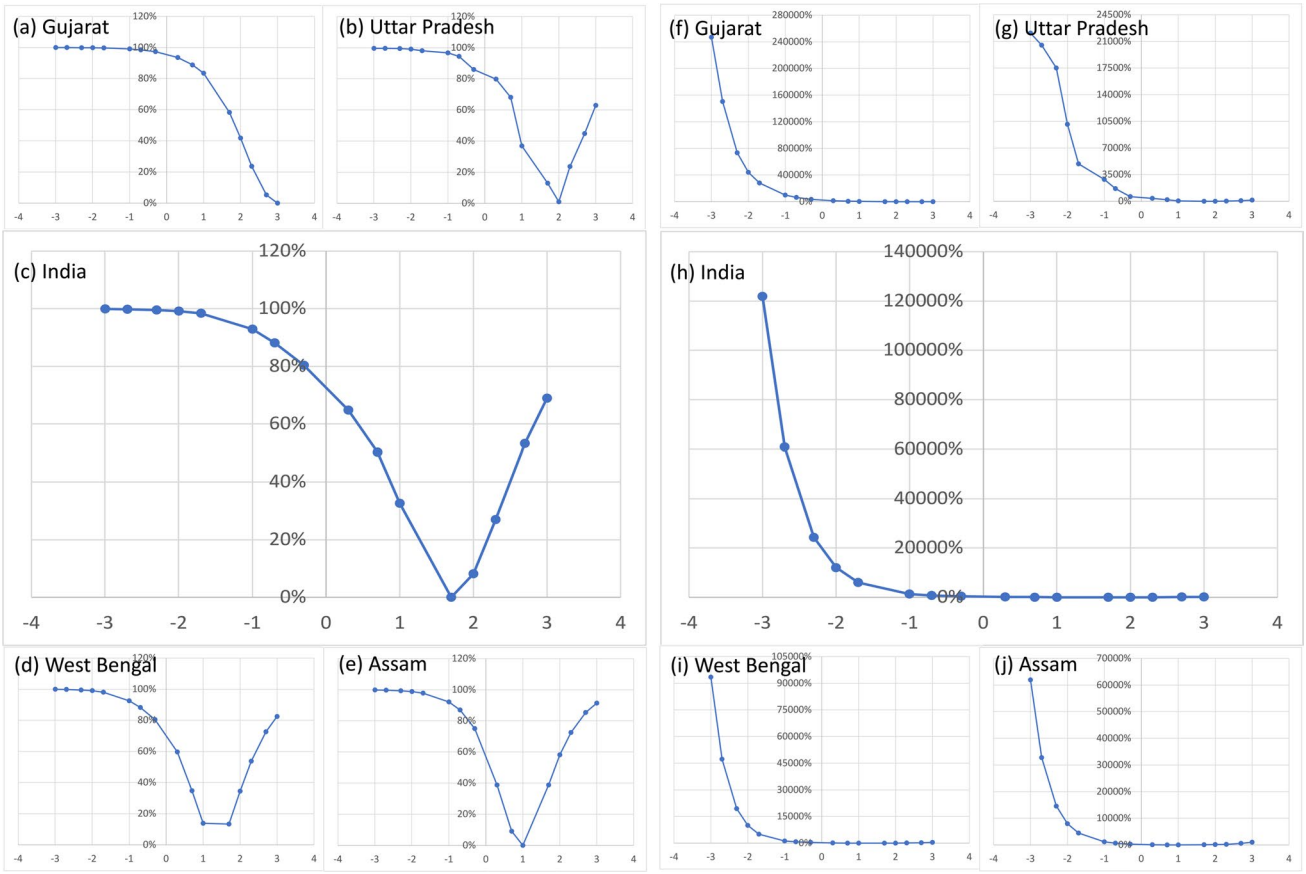

Fig. 5 Calculated potential model associated excess costs as a function of misclassification cost ratio ($CR_{wells,FP}$:$CR_{people,FN}$), expressed as percentage of costs arising from the use of (i) default cutoff value of 0.5 (a–e) (see Eq. (5)); (ii) cost-optimized cutoff value (f–j) (see Eq. (6)) for the states of Gujarat (**a**, **f**), Uttar Pradesh (**b**, **g**), West Bengal (**c**, **h**), and Assam (**d**, **i**) and for the whole of India (**e**, **j**)

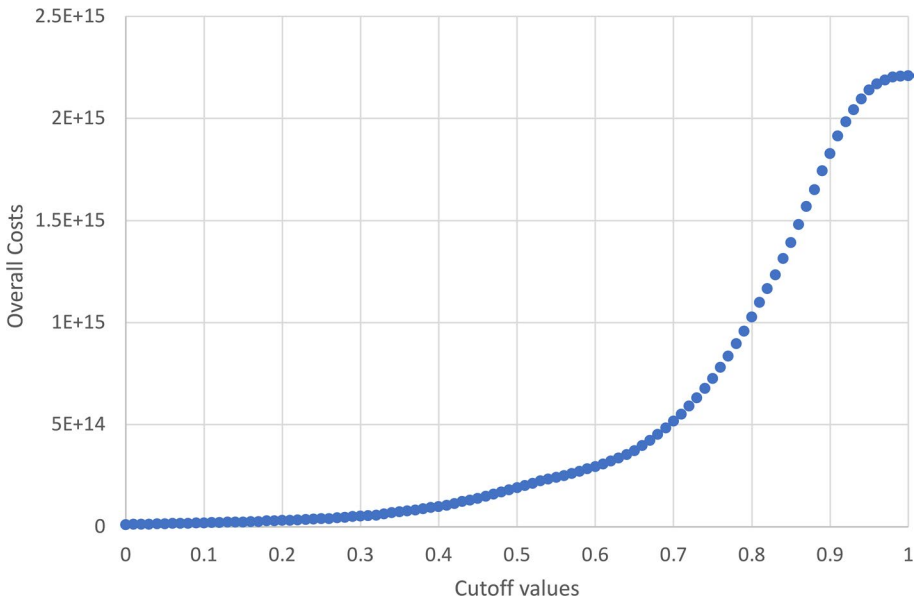

Fig. 6 Overall costs (FP, FN, and TP relative costs) a function of probability cutoff, based on the machine learning model of the distribution of groundwater arsenic in the whole of India. The unit cost values as defined in Table S2

optimal model upon which to inform policy for the reasons previously discussed. Notably, Maxim et al. (2014) warned of the excess costs arising from indiscriminate use of medical tests in populations with a very low prevalence of the conditions being tested for and this warning is also relevant to the large scale modelling of environmental chemical hazards, such as high arsenic groundwaters.

Conclusion

Probability maps of environmental chemical hazards generated by machine learning models are generally converted into hazard area maps by setting some probability cutoff. Choosing the probability cutoff is a crucial process to determining the modelled area of high hazard but existing methodologies for this are not designed to optimize costs related to health impacts, well remediation, and testing.

We demonstrate that for a case study of random forest-modelled groundwater arsenic distribution in India, the use of objective cost optimization criteria for selecting probability cutoff not only gives rise to probability cutoff different to those obtained from the most commonly published methods ((e.g. cutoff where sensitivity equals to specificity, cutoff where positive prediction values equal to negative prediction values, a default cutoff of 0.5) but also substantially reduces overall potential (health impacts, remediation, analytical) costs arising from the use of the model in informing practice.

The magnitude of the benefit of using cost-optimized probability cutoff criteria compared to commonly used default methods depends upon (i) the ratios of costs arising from false-positive, false-negative, and true-positive model classifications and (ii) the underlying prevalence of "high" (in this case study, higher than 10 µg/L arsenic) groundwater arsenic.

Where the distribution of high groundwater arsenic is highly heterogeneous, as it is in India, the greatest cost benefits may arise from the use of models of smaller, more granular areas at a more detailed scale (e.g. individual states; cf. Wu et al. 2020). State and basin scale modelling of groundwater arsenic using locally relevant cost bases for cost optimization of probability cutoffs is therefore indicated. We suggest that this would be a productive direction for studies not only of groundwater arsenic distribution in India but also of the distribution of other chemical hazards in various environmental media (waters, soils, crops) in India and other areas.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12403-023-00581-w.

Acknowledgements We thank and acknowledge the Newton Fund, the Natural Environmental Research Council of UK Research and Innovation (NERC, UK), and the Department of Science and Technology (DST, India) for their joint funding of the Indo-UK Water Quality project, FAR-GANGA (https://www.farganga.org/). We thank Biswajit Chakravorty, Abhijit Mukherjee, Joel Podgorski, Laura Richards, and Dipankar Saha for discussions. The opinions expressed in this paper however do not necessarily reflect those of any of the organizations or individuals whom we acknowledge here.

Author Contributions Conceptualization, DAP, methodology, software, modelling, validation, and formal analysis: RW, Co-writing—original draft preparation and reviewing and editing of the manuscript: RW and DAP, and funding acquisition—DAP. Both authors have read and agreed to the published version of the manuscript.

Funding This research was funded in part by the Newton Fund, the Natural Environmental Research Council (UK) (NE/R003386/1), and the Department of Science and Technology (India) (DST/TM/INDOUK/2K17/55(C) & 55(G)).

Data Availability Data presented in this study not otherwise available from the references and organizations indicated in the text may be available on request from the corresponding authors.

Declarations

Conflict of interest The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. We note that no jurisdictional claims are made or implied in any boundaries presented in figures in this article.

References

Amini M, Abbaspour KC, Berg M, Winkel L, Hug SJ, Hoehn E, Yang H, Johnson CA (2008) Statistical modeling of global geogenic arsenic contamination in groundwater. Environ Sci Technol 42(10):3669–3675. https://doi.org/10.1021/es702859e

Bhattacharya P, Polya DA, and Jovanović D (2017) *Best practice guide for the control of arsenic in drinking water*. International Water Association Publishing, ISBN13: 9781843393856

Bretzler A, Lalanne F, Nikiema J, Podgorski J, Pfenniger N, Berg M, Schirmer M (2017a) Groundwater arsenic contamination in Burkina Faso, West Africa: Predicting and verifying regions at risk. Science Total Environ 584:958–970

Bretzler A, Berg M, Winkel L, Amini M, Rodriguez-Lado L, Sovann C, Polya DA, Johnson A (2017b) Geostatistical modelling of arsenic hazard in groundwaters. In: Bhattacharya P, Polya DA, Jovanovic D (eds) Best practice guide for the control of arsenic in drinking water. IWA Publishing, London

Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F (2020) Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. J Med Syst 44(8):1–12. https://doi.org/10.1007/s10916-020-01597-4

Cao H, Xie X, Shi J, Wang Y (2022) Evaluation the validity of class balancing algorithms-based machine learning models for geogenic contaminated groundwaters prediction. J Hydrol. https://doi.org/10.1016/j.jhydrol.2022.127933

CGWB (Central Ground Water Broad). (2022). National Project on Aquifer Management (NAQUIM). http://cgwb.gov.in/AQM/NAQUIM.html Accessed 4 May 2022

Chakraborty M, Sarkar S, Mukherjee A, Shamsudduha M, Ahmed KM, Bhattacharya A, Mitra A (2020) Modeling regional-scale groundwater arsenic hazard in the transboundary Ganges River Delta, India and Bangladesh: Infusing physically-based model with machine learning. Sci Total Environ 748:141107. https://doi.org/10.1016/j.scitotenv.2020.141107

Chen Y, Wang X, Li L, Lu S, Zhang Z (2020) (2020) New cut-off values for screening of trisomy 21, 18 and open neural tube defects (ONTD) during the second trimester in pregnant women with advanced maternal age. BMC Pregn Childbirth 20:776. https://doi.org/10.1186/s12884-020-03464-z

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104

Connolly CT, Stahl MO, DeYoung BA, Bostick BC (2021) Surface flooding as a key driver of groundwater arsenic contamination in Southeast Asia. Environ Sci Technol 56(2):928–937. https://doi.org/10.1021/acs.est.1c05955

De Menezes MD, Bispo FHA, Faria WM, Goncalves MGM, Curi N, Guilherme LRG (2020) Modeling arsenic content in Brazilian soils: what is relevant? Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.136511

Dhamija S, Joshi H (2022) Prediction of groundwater arsenic hazard employing geostatistical modelling for the Ganga basin India. Water 14:2440. https://doi.org/10.3390/w14152440

ECDC (2021) Options for the use of rapid antigen detection tests for COVID-19 in the EU/EEA—first update. European Centre for Disease Prevention and Control. Technical Report 26 October 2021. https://www.ecdc.europa.eu/en/publications-data/options-use-rapid-antigen-tests-covid-19-eueea-first-update Accessed 20 Jan 2022

Erickson ML, Elliott SM, Brown CJ, Stackelberg PE, Ransom KM, Reddy JE, Cravotta CA III (2021) Machine-learning predictions of high arsenic and high manganese at drinking water depths of the glacial aquifer system, northern continental United States. Environ Sci Technol 55(9):5791–5805. https://doi.org/10.1021/acs.est.0c06740

Feinstein SH (1975) The accuracy of diver sound localization by pointing. Undersea Biomed Res 2(3):173–184 (**PMID: 15622737**)

Gail MH, Pfeiffer RM (2005) On criteria for evaluating models of absolute risk. Biostatistics 6(2):227–239. https://doi.org/10.1093/biostatistics/kxi005

Galen RS (1986) Use of predictive value theory in clinical immunology. Manual of clinical laboratory immunology, 3rd ed. American Society for Microbiology, Washington pp. 966–970

Government of India, (2011a) Census of India: population enumeration data. https://censusindia.gov.in/2011census/population_enumeration.html Accessed 10 Feb 2020

Government of India (2011b) Census of India: HH-6 households by main source of drinking water and location. https://www.censusindia.gov.in/2011census/Hlo-series/HH06.html Accessed 10 Feb 2020

Greiner M (1995) Two-graph receiver operating characteristic (TG-ROC): a Microsoft-EXCEL template for the selection of cut-off values in diagnostic tests. J Immunol Models 185(1):145–146. https://doi.org/10.1016/0022-1759(95)00078-O

Greiner M (1996) Two-graph receiver operating characteristic (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification costs. J Immunol Models 185(1):93–94. https://doi.org/10.1016/0022-1759(96)00013-0

Greiner M, Pfeiffer D, Smith R (2000) Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med 45(1–2):23–41. https://doi.org/10.1016/S0167-5877(00)00115-X

Grimes DA, Schulz KF (2002) Uses and abuses of screening tests. Lancet 359:881–884. https://doi.org/10.1016/S0140-6736(02)07948-5

Habibzadeh F, Habibzadeh P, Yadollahie M (2016) On determining the most appropriate test cut-off value: the case of tests with continuous results. Biochimia Medica 26(3):297–307. https://doi.org/10.11613/BM.2016.034

Hengl T, Mendes de Jesus J, Heuvelink GB, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Kempen B (2017) SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12(2):e0169748. https://doi.org/10.1371/journal.pone.0169748

Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley, New York, p 375

Jia X, Cao Y, O'Connor D, Zhu J, Tsang DC, Zou B, Hou D (2021) Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. Environ Pollut 270:116281. https://doi.org/10.1016/j.envpol.2020.116281

Kebonye NM, John K, Chakraborty S, Agyeman PC, Ahado SK, Eze PN, Nemecek K, Drabek O, Boruvka L (2021) Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via X-ray fluorescence spectroscopy. Geoderma. https://doi.org/10.1016/j.geoderma.2020.114792

Kelly MJ, Dunstan FD, Lloyd K, Fone DL (2008) Evaluating cutpoints for the MHI-5 and MCS using the GHQ-12: a comparison of five different methods. BMC Psychiatr 8(1):1–9. https://doi.org/10.1186/1471-244X-8-10

Knierim KJ, Kingsbury JA, Belitz K, Stackelberg PE, Minsley BJ, Rigby JR (2022) Mapped predictions of manganese and arsenic in an alluvial aquifer using boosted regression trees. Groundwater 60(3):362–376. https://doi.org/10.1111/gwat.13164

Kumar S, Pati J (2022) Assessment of groundwater arsenic contamination using machine learning in Varanasi, Uttar Pradesh India. J Water Health 20(5):829–848. https://doi.org/10.2166/wh.2022.015

Lado LR, Hengl T, Reuter HI (2008) Heavy metals in European soils: a geostatistical analysis of the FOREGS geochemical database. Geoderma 148(2):189–199. https://doi.org/10.1016/j.geoderma.2008.09.020

Lewis JD, Chuai S, Nessel L, Lichtenstein GR, Aberra FN, Ellenberg JH (2008) Use of the noninvasive components of the mayo score to assess clinical response in ulcerative colitis. Inflamm Bowel Dis 14(12):1660–1666. https://doi.org/10.1002/ibd.20520

Li H, Wu Y, Liu S, Xiao J, Zhao W, Chen J, Alexandrov G, Cao Y (2022) Decipher soil organic carbon dynamics and driving forces across China using machine learning. Global Change Biol 28(10):3394–3410. https://doi.org/10.1111/gcb.16154

Lombard MA, Bryan MS, Jones DK, Bulka C, Bradley PM, Backer LC et al (2021) Machine learning models of arsenic in private wells throughout the conterminous United States as a tool for exposure assessment in human health studies. Environ Sci Technol 55(8):5012–5023. https://doi.org/10.1021/acs.est.0c05239

López-Ratón M, Rodríguez-Álvarez MX (2019) Paxkage 'OptimalCutpoints'—computing optimal cutpoints in diagnostic tests. https://cran.r-project.org/web/packages/OptimalCutpoints/OptimalCutpoints.pdf Accessed 20 March 2020

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. J Appl Ecol 38(5):921–931. https://doi.org/10.1046/j.1365-2664.2001.00647.x

Maxim D, Niebo R, Utell MJ (2014) Screening tests: a review with examples. Inhal Toxicol 26(13):811–828. https://doi.org/10.3109/08958378.2014.955932

Metz CE (1978) Basic principles of ROC analysis. Semin Nucl Med 8(4):283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

Mikkonen HG, van de Graaff R, Mikkonen AT, Clarke BO, Dasika R, Wallis CJ, Reichman SM (2018) Environmental and anthropogenic influences on ambient background concentrations of fluoride in soil. Environ Pollut 242:1838–1849. https://doi.org/10.1016/j.envpol.2018.07.083

Millot R, Charlet L and Polya DA (2011) Un fléau mondial: la contamination de l'eau par l'arsenic. Pour Le Science*, 408*, October 2011, 76–82 [in French]

Mukherjee A, Sarkar S, Chakraborty M, Duttagupta S, Bhattacharya A, Saha D, Bhattachrya P, Mitra A, Gupta S (2021) Occurrence, predictors and hazards of elevated groundwater arsenic across India through field observations and regional-scale AI-based modeling. Sci Total Environ 759:143511. https://doi.org/10.1016/j.scitotenv.2020.143511

Ottong ZJ, Puspasari RL, Yoon D, Kim KW (2022) Predicting as contamination risk in red river delta using machine learning algorithms. SSRN. https://doi.org/10.2139/ssrn.3952430

Park Y, Ligaray M, Kim YM, Kim JH, Cho KH, Sthiannopkao S (2016) Development of enhanced groundwater arsenic prediction model using machine learning approaches in Southeast Asian countries. Desalin Water Treat 57(26):12227–12236. https://doi.org/10.1080/19443994.2015.1049411

Perović M, Šenk I, Tarjan L, Obradović V, Dimkić M (2021) Machine learning models for predicting the ammonium concentration in alluvial groundwaters. Environ Model Assess 26(2):187–203. https://doi.org/10.1007/s10666-020-09731-9

Phelps CE, Mushlin AI (1988) Focusing technology assessment using medical decision theory. Med Decis Making 8:279–289. https://doi.org/10.1177/0272989X8800800409

Podgorski J, Berg M (2020) Global threat of arsenic in groundwater. Science 368(6493):845–850. https://doi.org/10.1126/science.aba1510

Podgorski JE, Labhasetwar P, Saha D, Berg M (2018) Prediction modeling and mapping of groundwater fluoride contamination throughout India. Environ Sci Technol 52(17):9889–9898. https://doi.org/10.1021/acs.est.8b01679

Podgorski J, Wu R, Chakravorty B, Polya DA (2020) Groundwater arsenic distribution in India by machine learning geospatial modeling. Int J Environ Res Public Health 17(19):7119. https://doi.org/10.3390/ijerph17197119

Podgorski J, Araya D, Berg M (2022) Geogenic manganese and iron in groundwater of Southeast Asia and Bangladesh-machine learning spatial prediction modeling and comparison with arsenic. Sci Total Environ 833:155131. https://doi.org/10.1016/j.scitotenv.2022.155131

Polya DA, Sparrenbom C, Datta S, Guo HM (2019) Groundwater arsenic biogeochemistry—Key questions and use of tracers to understand arsenic-prone groundwater systems. Geosci Front 10:1635–1641. https://doi.org/10.1016/j.gsf.2019.05.004

Rodríguez-Lado L, Sun G, Zhang Q, Xue H, Zheng Q, Johnson CA (2013) Groundwater arsenic contamination throughout China. Science 341:866–868. https://doi.org/10.1126/science.1237484

Ruidas D, Pal SC, Islam ARMT, Saham A (2022) Hydrogeochemical evaluation of groundwater aquifers and associated health hazard risk mapping using ensemble data driven model in a water scares plateau region of eastern India. Expo Health. https://doi.org/10.1007/s12403-022-00480-6

Sajedi-Hosseini F, Malekian A, Choubin B, Rahmati O, Cipullo S, Coulon F, Pradhan B (2018) A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci Total Environ 644:954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054

Sharib J, Esserman L, Koay EJ, Maitra A, Shen Y, Kirkwood KS, Ozanne EM (2020) Cost-effectiveness of consensus guideline based management of pancreatic cysts: the sensitivity and specificity required for guidelines to be cost-effective. Surgery 168:601–609. https://doi.org/10.1016/j.surg.2020.04.052

Tan Z, Yang Q, Zheng Y (2020) Machine learning models of groundwater Arsenic spatial distribution in Bangladesh: influence of holocene sediment depositional history. Environ Sci Technol 54(15):9454–9463. https://doi.org/10.1021/acs.est.0c03617

Tesoriero AJ, Gronberg JA, Juckem PF, Miller MP, Austin BP (2017) Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resour Res 53(8):7316–7331. https://doi.org/10.1002/2016WR020197

Thiele C, Hirschfeld G. (2020). Cutpointr: improved estimation and validation of optimal cutpoints in R. *arXiv preprint* https://doi.org/10.48550/arXiv.2002.09209

UNICEF/WHO (2018) Arsenic primer. Guidance on the investigation & mitigation of arsenic contamination. https://www.unicef.org/media/91296/file/UNICEF-WHO-Arsenic-Primer.pdf Accessed 15 Aug 2022

Vermont J, Bosson JL, Francois P, Robert C, Rueff A, Demongeot J (1991) Strategies for graphical threshold determination. Comput Methods Program Biomed 35(2):141–150. https://doi.org/10.1016/0169-2607(91)90072-2

Winkel L, Berg M, Amini M, Hug SJ, Annette Johnson C (2008) Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. Nat Geosci 1(8):536–542. https://doi.org/10.1038/ngeo254

Wu R, Podgorski J, Berg M, Polya DA (2020) Geostatistical model of the spatial distribution of arsenic in groundwaters in Gujarat State India. Environ Geochem Health 43(7):2649–2664. https://doi.org/10.1007/s10653-020-00655-7

Wu R, Alvareda EM, Polya DA, Blanco G, Gamazo P (2021a) Distribution of groundwater arsenic in uruguay using hybrid machine learning and expert system approaches. Water 13(4):527. https://doi.org/10.3390/w13040527

Wu R, Xu L, Polya DA (2021b) Groundwater arsenic-attributable cardiovascular disease (CVD) mortality risks in India. Water 13(16):2232. https://doi.org/10.3390/w13162232

Youden WJ (1950) Index for rating diagnostic tests. Cancer 3(1):32–35. https://doi.org/10.1002/1097-0142(1950)3:1%3c32::AID-CNCR2820030106%3e3.0.CO;2-3