npj | clean water

www.nature.com/npjcleanwater

ARTICLE OPEN

Spatial association of surface water quality and human cancer in China

Zixing Wang[1,4], Wentao Gu[1,4], Xiaobo Guo[1,4], Fang Xue[1], Jing Zhao[1], Wei Han[1], Hairong Li[2,3], Wangyue Chen[1], Yaoda Hu[1], Cuihong Yang[1], Luwen Zhang[1], Peng Wu[1], Yali Chen[1], Yujie Zhao[1], Jin Du[1] and Jingmei Jiang[1]✉

Little is known about the association between surface water quality and cancer incidence, especially in China. Drinking water quality has been linked to the incidence of several cancers in individual-level studies. However, few studies have attempted to examine multiple pollutants and multiple cancers at population level. This study used water monitoring and population-level cancer data from across China to examine spatial associations between water pollutants and types of cancer. We found a "dose–response" relationship between the number of pollutants present at high levels and cancer incidence. These results provide evidence of a nationwide spatial association between water quality and cancer in China. The precise relationship varies with cancers and pollutants. However, the overall consistency of the "dose–response" relationship suggests that surface water quality is an important factor in cancer incidence. Our findings highlight new issues such as the changing effects when different pollutants co-exist and an increasing number of new cancer cases partially attributable to poor water quality. Our work also points to some ways to deal with these challenges.

npj Clean Water (2023)6:53 ; https://doi.org/10.1038/s41545-023-00267-5

INTRODUCTION

Surface water is an essential natural resource[1]. Deterioration of water quality creates enormous challenges to water availability, sustainable development and human well-being[2]. In assessing the health risk of poor surface water quality, carcinogenic effects are a key aspect. Relevant human evidence is considered essential for establishing surface water quality standards, that is, the minimum legislative requirements for safeguarding public health[3,4]. However, the association between surface water quality and cancer remains inadequately understood, especially in China, a country that accounts for nearly a quarter of global cancer cases[5,6]. This poses major limitations to the development of a suitable Chinese surface water quality standard, and is also deemed a driving factor in the discrepancy between standards adopted by different countries and authorities[7].

Drinking water sources and disinfection by-products have been linked to the risk of several cancers (especially oesophageal[8], gastric[9], colorectal and renal[10]) in cohort or case-control studies using individual questionnaire data. However, these designs have rarely been applied to analysis of surface water pollutants because it is challenging, if not impossible, to determine individual exposure level given complex exposure routes, including ingestion, dermal absorption and inhalation[3,11]. Another drawback with individual-level analyses is the slowness of evidence generation. Even beyond the scope of water research, only 16 agents have been added to the Group I carcinogen list during the past 10 years (and only two since 2019)[12,13]. Water monitoring and population-level cancer data have enabled ecological studies to make more efficient and comprehensive inspections of the association[14]. For instance, Yang et al. confirmed a connection between frequency of water pollution and population mortality from digestive cancers in the Huaihe River Basin in 2014. This was influential in promoting ongoing countermeasures in this specific area[15,16]. At the national level, China has now established one of the world's largest networks on water quality monitoring[17] and cancer registries[18], both covering all 31 province-level areas of the mainland. These efforts provide a unique opportunity to generate new information to policy-makers, scientists and the public. However, water bodies vary in their geographic background, pollutant level and combination[19]. The coexistence of multiple pollutants in surface water has not been examined, and the complex spatial patterns pose a methodological challenge for comprehensive appraisal of the attributable cancer burden.

In this study, we put forward a framework that integrates nationwide data from two large industries on surface water and cancer, and a design that decodes spatial associations between different pollutants and cancers. The study aimed to generate evidence on the systematic impact of surface water quality on cancer, determine the most influential water pollutants in different river basins, and forecast (and therefore possibly curb) a water quality-related cancer burden in the foreseeable future.

RESULTS AND DISCUSSION

Water quality assessment within an integrated framework

The quality of surface water in China is regularly assessed through the National Surface Water Environmental Quality Monitoring Network. During 2001–2021, this network experienced a large expansion in the number of water monitoring sections (from 454 to 3632; Supplementary Fig. 1). It now provides full coverage of all major rivers and administrative regions at prefecture level[17]. Figure 1a shows the locations of water monitoring sections. These

[1]Department of Epidemiology and Biostatistics, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College, No. 5 Dongdansantiao, Dongcheng District, 100005 Beijing, China. [2]Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, No.11 (A) Datun Road, 100101 Beijing, China. [3]College of Resources and Environment, University of Chinese Academy of Sciences, No.19 (A) Yuquan Road, Shijingshan District, 100049 Beijing, China. [4]These authors contributed equally: Zixing Wang, Wentao Gu, Xiaobo Guo.
✉email: jingmeijiang@ibms.pumc.edu.cn

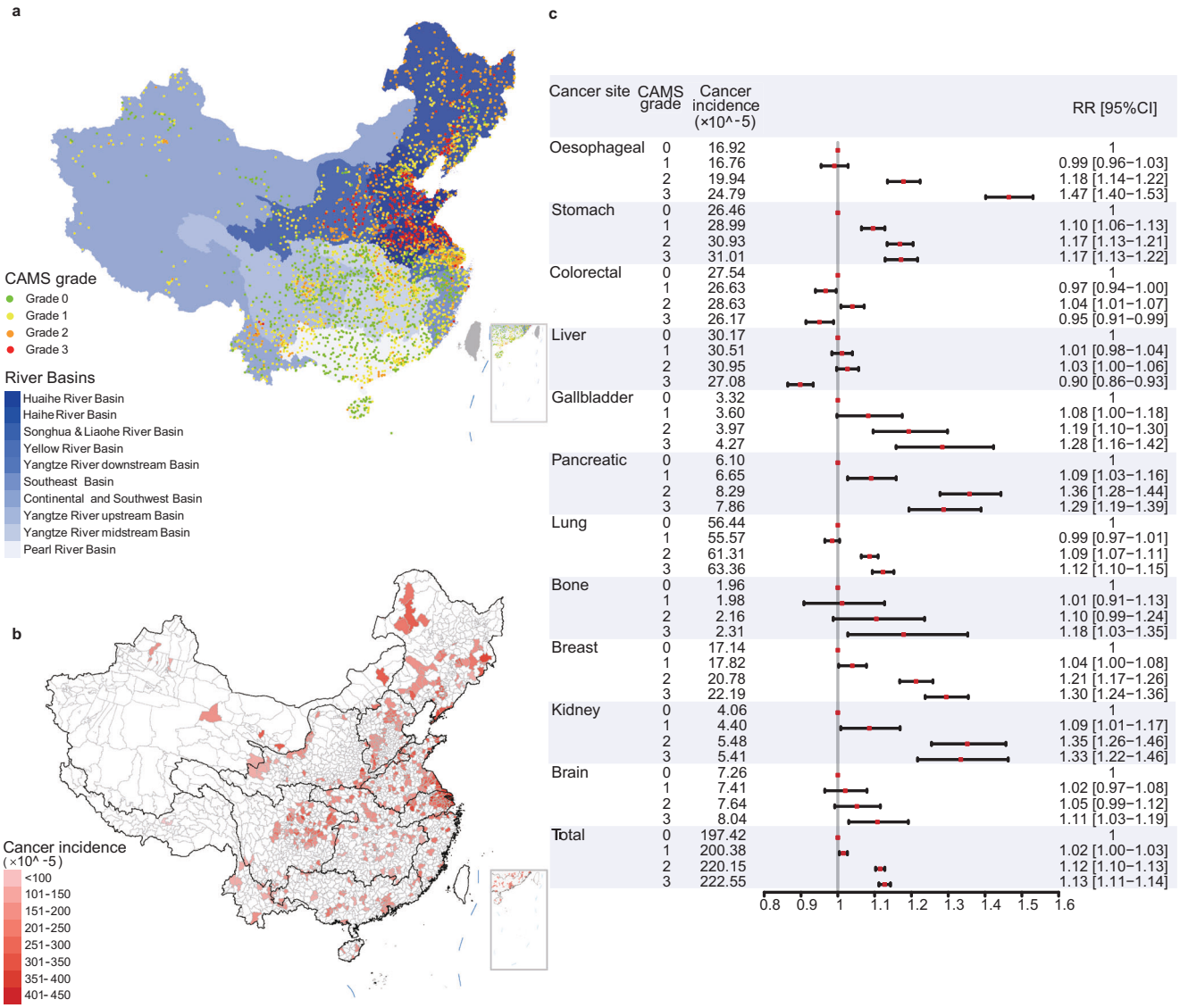

Fig. 1 Spatial association between surface water quality and cancer incidence. a Spatial distribution and pollution grading (as assessed by proposed "Cluster Analysis" of Multi-pollutants in Space design) of water monitoring sections. H denotes the number of pollution indicators in the high–high clusters or high–low outliers in the area: Grade 0 pollution (H = 0), Grade 1 pollution (H = 1–2), Grade 2 pollution (H = 3–5 H) and Grade 3 pollution (H ≥ 6). **b** Distribution of total incidence of 11 selected cancers in 486 cancer registry areas of China. The coloured boxes show cancer incidence in cases per 100,000 population. **c** Incidence and rate ratios (RRs) with 95% confidence intervals for each cancer in the area exposed to the different grades of surface water quality.

were matched to areas where cancer incidence data was available through the National Cancer Registry System (Fig. 1b). The matching design relied on the natural distribution of surface water and used each water quality monitoring section as the "centre" of a buffer zone with a 30 km radius. It retained 2331 out of 3632 water monitoring sections and 486 out of the 487 qualified cancer registries, covering a population of 380 million in all 31 provinces in mainland China. This was considered a representative sample of the country for both surface water (because of a non-artificial process; Supplementary Fig. 2) and its environmental influence on population cancer.

We considered all nine river basins in mainland China (Songhua & Liaohe, Haihe, Huaihe, Yellow, Continental, Southwest, Yangtze, Southeast and Pearl River Basins), and used this division to stratify the complex water network after a few modifications (see Fig. 1a). We sub-divided the Yangtze River Basin (the world's third longest river, the mainstream covering 11 provinces in China) into up-, mid- and down-stream basins to reflect discrepancies in both environment and population health[20]. We combined the Continental and Southwest River Basins for a similar reason (both located in western China and mainly include inland rivers). Annual summary data on the surface water quality in the monitored sections of these river basins in 2021 are given in Supplementary Table 1. The analysis included 21 basic pollution indicators because of their testability, comprehensiveness and high pollution share rate in China.

To make a comprehensive assessment, we applied the current Chinese standard, the Environmental Quality Standard for Surface Water (EQSSW version 2002)[21]. This sets a maximum of six levels for each pollution indicator, and the overall water quality is determined by the highest (i.e., worst) level. The water quality in 86.14% of the monitoring sections met the entry-level standard for sources of drinking water (Level III[21]). This was a marked improvement in surface water quality over time, compared with

previous reports that used the same criterion in China[17] (Supplementary Fig. 1). However, this quality standard, which has been in use for 20 years, was not designed for cancer risk assessment. The high number of monitoring sections meeting the standard across China raises questions about whether this criterion is useful in catalyzing further environment improvement or to benchmark between monitored sections.

To overcome these issues, we used the 75th percentile of each pollution indicator value as a threshold for defining inferior water quality level. This was stricter than EQSSW Level III for all the indicators except total nitrogen (Supplementary Table 1). Against these new standards (Fig. 1a), Western and Central China were largely free of inferior-quality surface water. However, the Eastern area had several areas with multiple indicators of inferior quality, particularly in the Huaihe, Haihe and Songhua & Liaohe Basins. Per-capita water availability in China is only a quarter of the world average[22], and it is particularly low in these areas (e.g., 37.5, 62.0 and 74.6 cubic metres per head in Beijing, Tianjin and Hebei in the Haihe River Basin) compared to the national average (2156.3 cubic metres per head in 2020)[23]. The varying distribution pattern reflects the validity of our proposed 75th percentile criteria and, more importantly, the difficulty of protecting and improving access to clean water.

Association between multiple pollutants and cancers

We ran a negative binomial regression analysis[24] for each pollution indicator, including any river basins in which over 20% of the monitored sections of water were of inferior-level quality. This value was used to avoid a "dilution" effect in analysing association and was selected because it approximates to the national average level using the 75th percentile threshold. Of the 21 pollution indicators, 11 (total nitrogen, petroleum, total phosphorus, permanganate index, chemical oxygen demand (COD), volatile phenol, fluoride (F^-), ammonia nitrogen (NH_3-N), arsenic, selenium, and zinc) were positively related to at least one cancer from the following sites (Supplementary Fig. 3): oesophagus (International Statistical Classification of Diseases 10th revision code: C15), stomach (C16), colorectum (C18–C21), liver (C22), gallbladder (C23–C24), pancreas (C25), lung (C33–C34), bone (C40–C41), breast (C50), kidney (C64–C66, C68) and brain (C70–C72, C32–C33, D42–D43). These 11 cancers were included in our analysis because of their significance in the Chinese population, assessed using incidence, mortality and 5-year survival rate[25,26].

Spatial autocorrelation was found for each of the cancer-related pollution indicators, evaluated using the global Moran's method (Supplementary Table 2; all $p < 0.05$). We identified high–high (HH) and low–low (LL) clusters (i.e., neighbouring water sections with similar superior/inferior water quality level), and high–low (HL, high amongst low) and low–high (LH, low amongst high) outliers of the monitored water sections, using local Moran's *I*, to show the distribution pattern of each indicator (Supplementary Fig. 4). Overall, more HH clusters and HL outliers were observed in North China than the South, but this varied for specific pollution indicators. The pollution indicators therefore form a complex pairwise correlation structure (Supplementary Fig. 5).

We were interested in whether multiple pollutants coexisted in the same area and their joint effect on cancers. We therefore proposed a design, "Cluster Analysis of Multi-pollutants in Space (CAMS)". We graded each monitored water section by the number of pollution indicators in HH clusters or HL outliers. For example, if there were two pollution indicators in the HH clusters and one in the HL outliers, then the H value for this section would be 3. This gave a CAMS grading: Grade 0 ($H = 0$; 26.9% of the water sections), Grade 1 ($H = 1–2$; 38.4%), Grade 2 ($H = 3–5$; 24.0%), Grade 3 ($H \geq 6$; 10.7%). Compared with existing water quality classification rules (e.g., the EQSSW worst-level approach), the CAMS grading has advantages because it takes into account the coexistence of multiple pollutants in a specific location, and the impact of each area's relationships with neighbouring areas. This reflects the impact of both point and non-point source water pollution[27].

We found an approximate "dose–response" relationship between the CAMS grading and population incidence data for the selected cancers (Fig. 1c). Grades 1, 2 and 3 were associated with a relative increase in population incidence of 10%, 17% and 17% in stomach cancer, 9%, 36% and 29% in pancreatic cancer, and 9%, 35%, and 33% in kidney cancer compared with Grade 0 (all $p < 0.05$); no statistical significance when comparing the two highest levels. This suggests that these cancers are very sensitive to water quality, especially when there is more than one pollutant present. There was no significant increase in population incidence in oesophageal, breast, gallbladder or lung cancer when comparing Grade 1 to Grade 0. However, a joint effect became apparent at higher grades. For instance, there was a relative increase in incidence in oesophageal cancer of 18% (Grade 2) and 47% (Grade 3). A significant increase (18% and 11%) was seen in bone and brain cancer incidence in Grade 3 areas compared to Grade 0. This may only have been visible in Grade 3 areas because of the low incidence of these cancers. When combined, the overall incidence of the 11 selected cancers increased from 197.42 per 100,000 in Grade 0 to 222.55 per 100,000 in Grade 3 (a relative increase of 12.7%).

These results provide evidence of a nationwide spatial association between water quality and cancer in China. This study simultaneously demonstrates such an association for multiple cancers. The multiple-test results of the pairwise association found between each pollution indicator and each cancer site should be treated with caution. However, the consistency in the "dose–response" relationship for the majority of the cancer sites reinforces our confidence that the contribution of surface water quality to cancer is widespread.

It may be helpful to consider some insights into the CAMS grading to help explain its observed relationship with cancer incidence. For nearly all the specific pollutants, the quality was poorer in water sections of higher CAMS grades (Supplementary Fig. 6a–d): take pollutant F^- as an example, the rates of inferior quality (i.e., exceeding the 75th percentile standard) were 1.0% (Grade 0), 12.5% (Grade 1), 52.8% (Grade 2) and 90.8% (Grade 3). Besides that, there were widespread correlations between the pollutants, and these became more complex (positively or negatively correlated) and more apparent (higher absolute values of correlation coefficient) in Grades 2 or 3 (Supplementary Fig. 6e–h). These findings suggest that the CAMS grading could be used as a comprehensive measure to summarize both single roles of pollutants and the complex effects of their interactions. Another spatial analysis in the United States showed water-borne chemicals not currently recognized as carcinogens may contribute to the population cancer risk[28]. When interpreted with our results, this suggests that specific pollutants may not necessarily be carcinogenic on their own. Rather than assessing the health risk of individual pollutants in surface water, we may need to look more comprehensively.

Key cancer-related pollutants in specific river basins

There was a dramatic variation in cancer incidence between cancers and across river basins (Fig. 2a). To assess the environmental impact (water quality and other socioeconomic factors), we labelled the top three river basins with the highest average incidence per cancer, and applied a machine learning technique (SHapley Additive exPlanations (SHAP) values generated from an extreme gradient boosting algorithm) to provide deeper insights into river basin-specific contributions of the pollution indicators (Fig. 2b).

River basins that generally had poorer-quality water (Huaihe, Haihe, Songhua & Liaohe, and downstream Yangtze River Basins) ranked high in incidence of multiple cancers. The Huaihe River

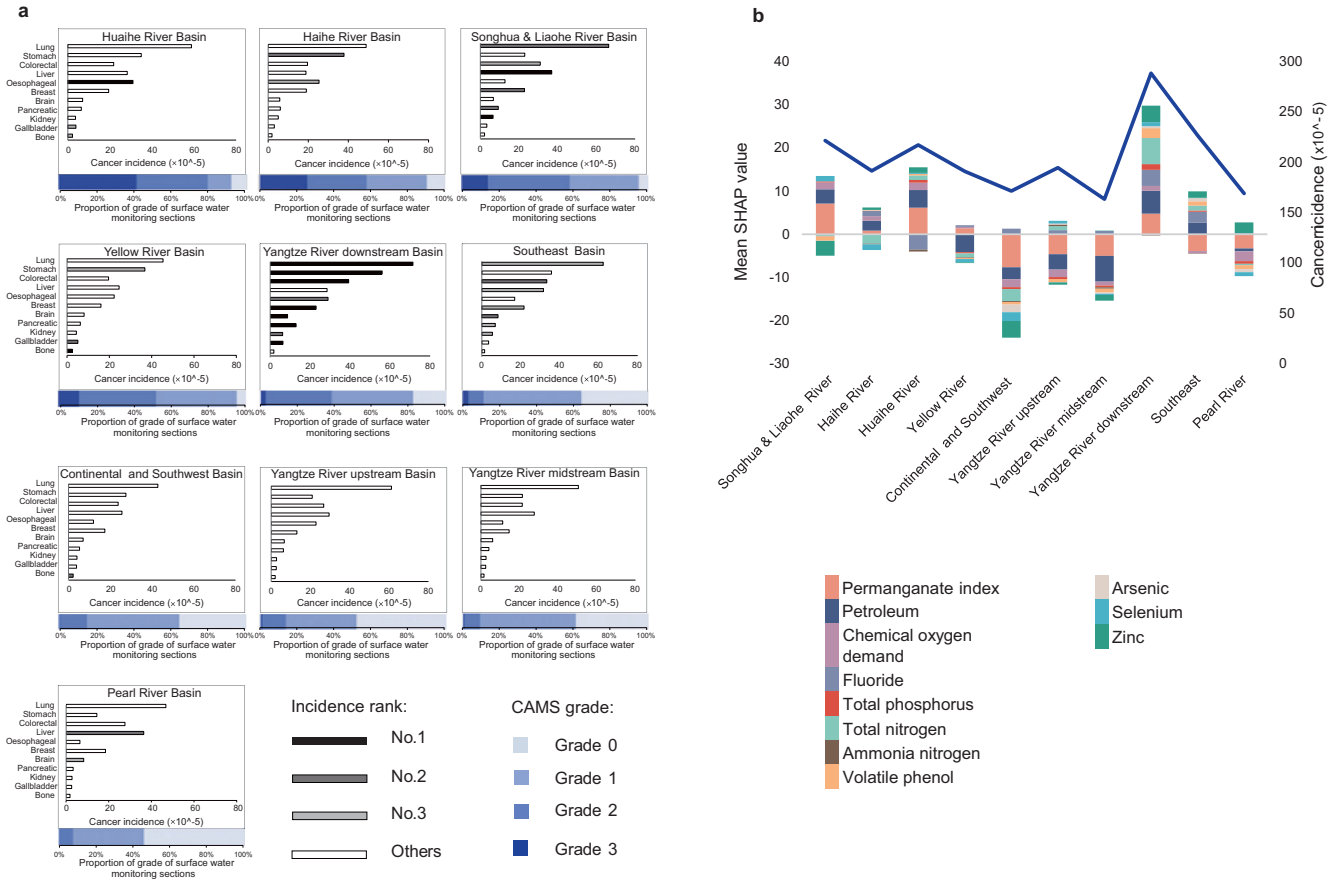

Fig. 2 Key cancer-related pollutants in river basin-specific contexts. a Ranked incidence of 11 selected cancers in 10 river basins with different grades of surface water quality. CAMS, proposed "Cluster Analysis" of Multi-pollutants in Space design. **b** Key pollution indicators affecting total cancer incidence in each river basin, assessed by SHapley Additive exPlanations (SHAP) values generated from an extreme gradient boosting algorithm.

Basin in particular is renowned for its high levels of pollution[19,29,30]. It ranked highest for oesophageal cancer and third for both gallbladder and bone cancers, even though the Chinese government has invested heavily in environmental restoration and cancer prevention in this area in recent years[31,32]. This finding emphasizes the difficulty in completely eliminating the impact of water pollution, especially in a short time.

Permanganate index, petroleum and chemical oxygen demand were important pollution indicators that could explain the high level of cancer incidence in the Huaihe River Basin. We obtained similar findings from the Haihe and Songhua & Liaohe River Basins. The Haihe River Basin is next to the Huaihe River Basin both geographically and in its water quality, and ranked second for stomach cancer and third for oesophageal cancer. In this basin in particular, F^- was related to cancer incidence. A previous study reported that the high level of F^- in this basin was related to the long-term over-exploitation of groundwater for agricultural irrigation in the area, resulting in a large amount of F^- in clay soils entering the groundwater and surface water[33]. Recent systematic reviews have not found a link between water F^- and cancer[34,35], but further work is needed to explain our observations in the Haihe River Basin. Another interesting finding was that selenium was related to cancer incidence in the Songhua & Liaohe River Basin. Previous observational studies have suggested that selenium has a protective effect against cancer[36], but recent randomized controlled trials found no beneficial effect of selenium supplements on cancer risk[37]. The relationship between selenium and cancer is therefore unclear. Consistent with our study, a long-term cohort study (28-year follow-up) among Italian citizens observed an increased risk of cancer with higher selenium concentration in the water supply system[38]. We therefore need to consider the inorganic form of selenium generally found in water because its biological properties may be markedly different from those of other chemical forms[36]. Laboratory studies have shown that the toxicity of inorganic (tetravalent) selenium greatly exceeds that of organic selenium[39].

Notably, the Songhua & Liaohe River Basin had better overall water quality than the Huaihe and Haihe River Basins but had the highest incidence of liver and kidney cancer, second-highest levels of pancreatic, lung and breast cancer, and third-highest colorectal cancer levels. Air pollution has also been reported to be associated with cancer[40], probably resulting from open burning of straw and other causes[41]. All these provide an explanation for this complex finding. Surface water pollution is likely one, but not the only, critical environmental driver of increased cancer.

One especially thought-provoking finding was from the Yangtze River Basin, which supports more than 40% of China's population and nearly half of its economy. Along the west-to-east course of the Yangtze River, the per-capita gross domestic product increases (e.g., 57,532 CNY up-stream, 65,481 CNY mid-stream, and 104,031 CNY down-stream in year 2020)[23], but the water quality decreases (possibly due to cumulative pollution from the branches, Supplementary Fig. 7) and cancer incidence increases. This contributes to very different pictures of the mode of impact of water quality on cancer across these three sub-basins, even though their water systems are closely connected. The up- and

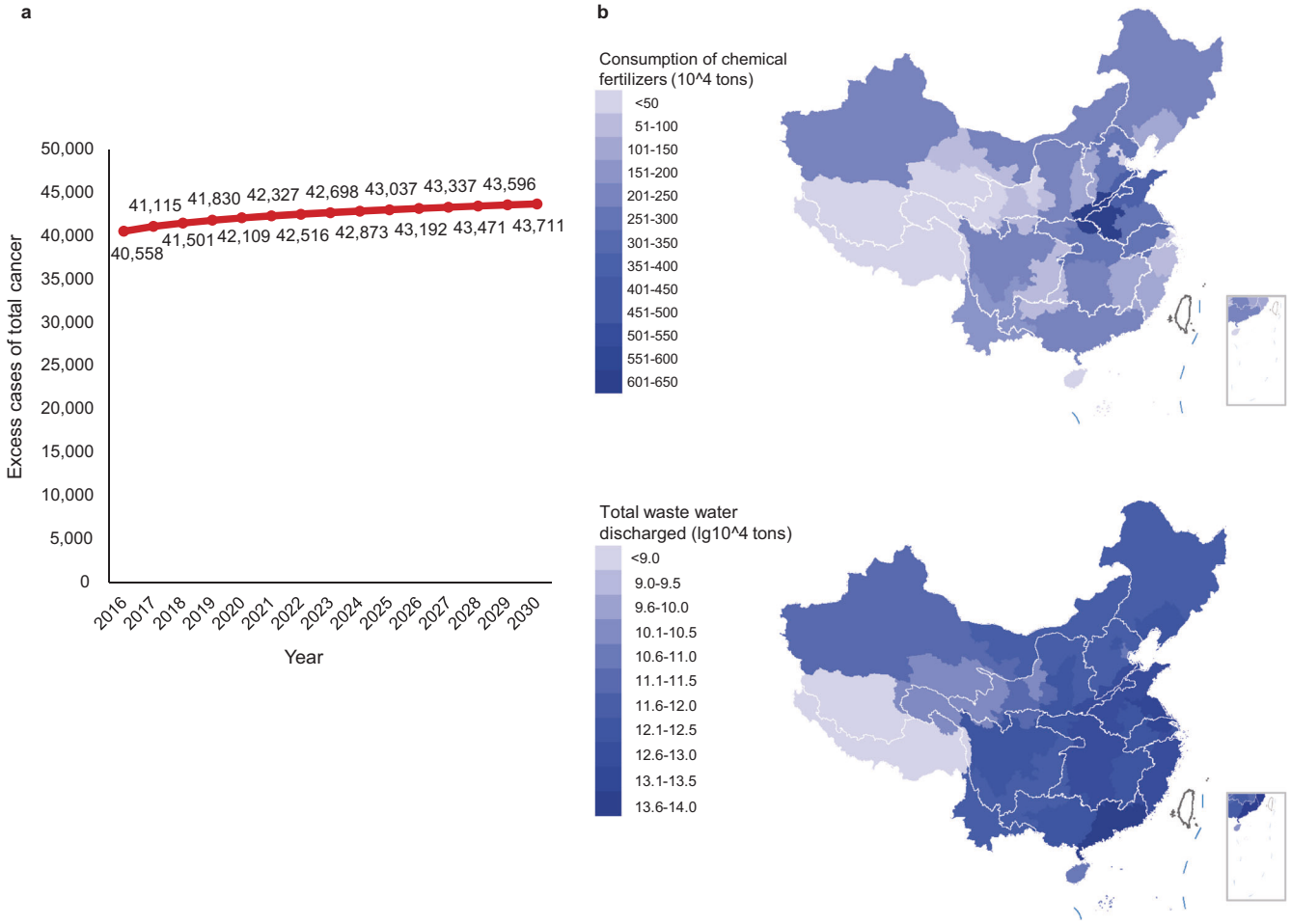

Fig. 3 Rising to current and future challenges in surface water-related cancer. a Projected excess cases of total cancer caused by the degree of surface water pollution in cancer registry areas of China, 2016–2030. **b** Distribution of the amount of fertilizers applied in different provinces of China. **c** Distribution of wastewater discharge in different provinces of China.

mid-stream basins had no cancer ranked among the country's top three for population incidence, and there was little or no association between cancer and water quality. However, the incidence of all cancers except for liver and bone cancers in the downstream basin was among the top three (resulting in the highest combined incidence, 288.50 per 100,000). There were multiple water pollution indicators that could explain this, including total nitrogen, petroleum, permanganate index, zinc, F^-, and volatile phenol (in descending order of influence). These patterns can be viewed as a mirror of China's east–west development discrepancy, and convincing proof of the impact of the environment. It strongly suggests a need to end the approach that has traded environmental quality for economic growth, and instead makes systematic efforts towards environmentally-friendly and sustainable development. In 2022, the Chinese government issued the Action Plan for Deepening the Yangtze River Protection and Restoration Campaign, clarifying the outstanding ecological and environmental problems that need to be addressed by 2025[42]. Our study provides a model to identify key targets that would allow effective, tailored countermeasures.

Rising to current and future challenges

Following our proposed 75th percentile thresholds and risk estimates corresponding to the CAMS grades, we estimated the population attributable fraction (PAF) and the number of water quality-related cancers (Supplementary Table 3). On average, 5.0% of all new cancers in current cancer registry areas can be attributed to poor quality surface water, with the most marked fractions in the Huaihe (9.3%), Songhua & Liaohe (7.3%) and Yellow River Basins (5.8%).

Our estimates provide a deep look at the impact of water quality on cancers, but they show only part of the environmental influence, which can be modified through proactive measures and government commitment. Another driving force in the cancer burden in China and many parts of the world, population aging, provides a dual challenge, and the impact is unlikely to cease in the foreseeable future[43,44]. The number of cancer cases worldwide is projected to double from the 2020 figures by 2070[45]. Assuming no further improvement in water quality across China, we estimate that the number of new cancer cases will rise by 7.8%, to 874,218 in the present cancer registry areas by 2030, and an increasing number each year will be attributable to poor water quality during 2022–2030 (Fig. 3a). We estimate that this will amount to a total of 388,431 new cases. This is hypothetical, but highlights the urgency of acting to stop a rising tide of potentially preventable cancer.

Finding and controlling the sources of pollution are likely to be the key solutions. Figure 3b, c shows the distributions of the amount of fertilizers applied and wastewater discharge in different provinces. This gives an insight into some of the sources of water pollution. The result suggests that there are multi-faceted, interrelated causes. First, to increase yields to supply the growing

population, farmlands in China (e.g., in the Huaihe River Basin) used to apply large amounts of fertilizers (highly soluble ammonia fertilizers in particular), which drained into waterways and increased total nitrogen pollution[46]. Second, products (including these chemical fertilizers) are manufactured in industrial areas (mainly in the Haihe, Songhua & Liaohe, and Yangtze downstream River Basins), where large numbers of chemical compounds are discharged into water as part of their production processes[47]. Third, domestic sewage discharge is also high in these areas because of the high population density[48]. The total amount of wastewater in China reached a peak in 2015, having grown by an average annual rate of 2.8%, and has only recently slowed down by an annual rate of −2.5%[23,49,50]. Good management of chemical fertilizers and wastewater outfalls is suggested as the best way to improve water quality in China.

According to a survey in 2022, water pollution has replaced air pollution as the primary public concern in China[51]. This reflects both the urgency and societal will for change. However, just as pollution problems do not develop in a day, the effects cannot be eliminated rapidly. Realizing ambitious goals in environmental protection and population health requires a generational time frame and sustained input into policy, research, and most importantly, creation of an environmentally-friendly culture. Another implication for developing countries is that pollution source governance in industry and agriculture may have an impact on economic development. Investment in green technology is vital in balancing development and environmental protection[52].

Breaking barriers in environmental health research

This paper proposes a paradigm in environmental health research, and therefore makes an important step in exploring the complex association between water quality and multiple cancers. It has integrated data from multiple sources in a unified spatial scale of river basins, a notable difference with the administrative system typically used in previous cancer research. One important barrier that hampered advances in both this field, and other environmental health fields, is the absence of a harmonized data framework. The environmental industry and the health industry have different focuses and operational structures (Supplementary Fig. 8), making the linkage between different data sources particularly challenging[53]. More cross-industry coordination is required to foster a more solid basis to build advances in data sciences and keep up with the increased concerns of the public about the environment.

A unique methodological challenge in assessing environment quality and its impact on health outcomes is the coexistence of multiple pollutants. Pollution often has a spatial autocorrelation, making multiple pollutants inter-correlated and giving different combination patterns in different regions. This phenomenon means that assessing health risks for individual pollutants is difficult because their interactions provide additional hazards that may not be explicable with current knowledge. To address these issues and a limit in the pairwise analysis of correlation in current methods, we proposed the CAMS design, which is comprehensive because it retains both the quality and spatial variation information. It is also versatile, transparent, easy-to-interpret and validated by our results. This design allowed us to restore the complicated real-world associations from previously separate findings, and generate knowledge about joint effects in a clear way. We hope this design may be useful for future studies and action to improve environmental and public health.

The EQSSW standards have been in use for 20 years in China, and many other countries have also not updated similar standards during the past two decades or more[7]. Using the current 75th percentile values, we found several pollution indicators were related to population cancer risk. We suggest that the EQSSW is not suitable for cancer risk assessment, and recommend that the 75th percentile thresholds should be updated annually to provide a dynamic goal for ongoing progress. This approach would also motivate local officers to benchmark against national levels. We used the national 75th percentile (i.e., the upper quartile) as a threshold because it allows for variation across river basins and is considered suitable to indicate high levels. Yet we do not suggest these values are necessarily the best cut-off values for safeguarding public health (for instance, the findings were similar if 80th percentile values were used; Supplementary Fig. 9). The negative findings for the other pollution indicators may be the result of not knowing the correct cut-off values to use. There are therefore many challenges in updating EQSSW and harmonizing standards across the world[7].

In this study, we used annual average values of the water quality monitoring data to smooth temporal fluctuations. Similarly, population incidences, as parameters in each cancer registry, are relatively robust over time. Population-level data (i.e., parameter) has such good properties that help to address the issue of time lag between exposure and cancer in individual-level studies, and preserve spatial variations (see Supplementary Table 1) that are important to explore law of the nature. Analysis of these data has been widely embraced in a number of study fields, but very little in medicine. This may be because of the traditional focus on individuals, instead of a larger perspective embedded in a broader ecosystem, where each individual is a 'cell' that is deeply influenced by the surroundings. A shift from human-centred to ecosystem- or nature-centred perspectives would help to break traditional boundaries in health-related studies, and ultimately, create a better living environment for all.

METHODS

Analytic workflow

Supplementary Fig. 10 gives an overview of the analytical framework used in this study.

Data source

Surface water quality data (2021) were derived from all (3632) monitored sections of the National Surface Water Environmental Quality Monitoring Network, Ministry of Ecology and Environment of China[54]. The surface water monitoring sections took into account the natural attributes of the basin such as basin area, river network density, runoff supply and hydrological characteristics.

Cancer incidence data were extracted from the Annual Report of the China Cancer Registry (2019), which was drawn from 487 registries in 31 provincial-level places in mainland China[55]. It covered a total population in China of 381.6 million people (193,632,323 males and 187,933,099 females), accounting for 27.60% (24.3% for urban areas and 32.0% for rural areas) of the national population at the end of 2016.

The data on the amount of applied fertilizers and wastewater discharge were extracted from the China Statistical Yearbooks and provincial statistical yearbooks (2019–2021)[23]. The data on forecast population size were obtained from the United Nations World Population Prospects (2019)[56]. Map data were obtained from the China Resource and Environment Science and Data Center.

Spatial analysis

Spatial mapping was used to show the spatial distribution and pattern of cancer incidence and surface water pollution. For spatial autocorrelation of each pollution indicator, we used the global Moran's *I* (see formula (1)) and local Moran's *I* (see formula (2)) to confirm the overall spatial autocorrelation and identify the

pattern of local spatial clusters[57–59].

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad (1)$$

$$I_i = \frac{x_i - \overline{x}}{S_i^2} \times \sum_{j=1, j\neq i}^{n} w_{ij}(x_j - \overline{x}), S_i^2 = \frac{\sum_{j=1, j\neq i}^{n}(x_j - \overline{x})^2}{n-1} \quad (2)$$

where n denotes the number of monitored surface water sections; x_i the value of the pollution indicator for the i^{th} section ($i = 1, 2, \cdots, n$), x_j the value of the pollution indicator for the j^{th} neighbouring section ($j = 1, 2, \cdots, n; i \neq j$); $\overline{x}$ the average value of each pollution indicator; and w_{ij} the neighbouring relations of sections i and j.

We assumed that as the distance from a river increases, a concomitant decrease occurs in the probability that a population will be influenced by water pollution in the river. We selected 30 km as the distance beyond which the health effect from surface water pollution is negligible. This was determined using a predefined buffer analysis informed by previous studies on interaction between riverine chemistry and land use[60] and on spatial distribution of persistent organic pollutant autocorrelation[61]. The buffer analysis process[57] involves generating buffers around existing monitored surface water sections and then identifying or selecting areas based on whether they fall inside or outside the boundary of the buffer. The cancer incidence and rate ratios (RR) in areas of different surface water quality grade were calculated based on buffer analysis, and 95% confidence intervals (CI) of the RRs were derived from negative binomial regression models.

Statistical analysis

Categorical variables are shown as frequencies and percentages, and continuous variables as medians (25th percentile, 75th percentile) if they did not satisfy a normal distribution. Pairwise correlation cross-pollution indicators were examined using Spearman's correlation analysis. Negative binomial regression[24] was used to test the effect of single pollution indicators on different cancers.

To make full use of the original values of the pollution indicators and accommodate the complex correlations between them, we conducted extreme gradient boosting (XGBoost), a scalable tree boosting and effective machine learning algorithm, to confirm our findings[62]. However, one of the disadvantages of this approach is the limited interpretability of the complex underlying feature interaction and non-linear structure. On the basis of game theory, SHAP values allowed us to estimate the mean impact on model output magnitude for each of the input features[63]. SHAP values were generated from the XGBoost algorithm (parameters tuned: learning rate = 0.04, max depth = 3, subsample ratio = 0.8), to allow us to identify the most important pollution indicators affecting multiple cancers and the primary pollution indicators affecting cancers in each river basin.

The excess cases (EC), i.e., expected cancer reductions, were computed via population attributable fractions (PAFs) (see formula (3))[64,65].

$$\mathrm{PAF} = \frac{\sum_{i=0}^{3} p_{e_i}\mathrm{RR}_i - 1}{\sum_{i=0}^{3} p_{e_i}\mathrm{RR}_i} \quad (3)$$

where p_{ei} denotes the prevalence of exposure i among the total population, that is, the proportion of population at Grade i, and RR_i the relative risk of cancer incidence at Grade i. The number of EC was given by $\mathrm{EC} = \mathrm{No.\ of\ incidence} \times \mathrm{PAF}$. The prevalence of graded exposure levels was evaluated by Thiessen polygon analysis[57]. Considering the trend of cancer incidence in 2010–2016[25,55], we used linear regression to project an annual total incidence of the selected cancers in 2017–2030.

We conducted the data integration and negative binomial regression in SAS 9.4 software (SAS Institute Inc., Cary, NC, USA). The XGBoost algorithm and SHAP plots used the R package XGBoost[66] and SHAPforxgboost[67]. All spatial analyses used the ArcMap module in ArcGIS 10.8 software (ESRI, Redland, CA).

DATA AVAILABILITY

Sources of raw public dataset used within the paper are summarized in the 'Methods' section. The datasets analysed during the current study are available from the corresponding author on reasonable request.

CODE AVAILABILITY

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 22 January 2023; Accepted: 21 June 2023;
Published online: 08 July 2023

REFERENCES

1. Rodell, M. et al. Emerging trends in global freshwater availability. *Nature* **557**, 651–659 (2018).
2. Leder, K., Sinclair, M. & McNeil, J. Water and the environment: a natural resource or a limited luxury? *Med. J. Aust.* **177**, 609–613 (2002).
3. US-EPA. *Methodology for Deriving Ambient Water Quality Criteria for the Protection of Human Health. EPA-822-B-00-004* (United States Environmental Protection Agency, 2000).
4. Ministry of Ecology and Environment of China. *Technical Guideline for Deriving Water Quality Criteria for the Protection of Human Health. HJ 837-2017* (Ministry of Ecology and Environment of China, 2017).
5. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
6. WHO IARC. *World Cancer Report* (WHO International Agency for Research on Cancer, 2020).
7. van Winckel, T. et al. Towards harmonization of water quality management: a comparison of chemical drinking water and surface water quality standards around the globe. *J. Environ. Manag.* **298**, 113447 (2021).
8. Sheikh, M. et al. Individual and combined effects of environmental risk factors for esophageal cancer based on results from the golestan cohort study. *Gastroenterology* **156**, 1416–1427 (2019).
9. Picetti, R. et al. Nitrate and nitrite contamination in drinking water and cancer risk: a systematic review with meta-analysis. *Environ. Res.* **210**, 112988 (2022).
10. Jones, R. R. et al. Ingested nitrate, disinfection by-products, and risk of colon and rectal cancers in the Iowa Women's Health Study cohort. *Environ. Int.* **126**, 242–251 (2019).
11. Tong, S., Li, H., Tudi, M., Yuan, X. & Yang, L. Comparison of characteristics, water quality and health risk assessment of trace elements in surface water and groundwater in China. *Ecotoxicol. Environ. Saf.* **219**, 112283 (2021).
12. WHO IARC. *Agents Classified by the IARC Monographs*, Vol. 1–132 (WHO International Agency for Research on Cancer, 2022).
13. Cogliano, V. J. et al. Preventable exposures associated with human cancers. *J. Natl Cancer Inst.* **103**, 1827–1839 (2011).
14. Kirch, W. *Encyclopedia of Public Health: Ecological Study* (Springer, 2008).
15. Yang, G. & Zhuang, D. *Atlas of the Huai River Basin Water Environment: Digestive Cancer Mortality* 1st edn (Springer, 2014).
16. Ren, H. et al. Association between changing mortality of digestive tract cancers and water pollution: a case study in the Huai River Basin, China. *Int. J. Environ. Res. Public Health* **12**, 214–226 (2014).
17. Ministry of Ecology and Environment of China. *Report on the State of the Ecology and Environment in China* (Ministry of Ecology and Environment of China, 2021).
18. Wei, W. et al. Cancer registration in China and its role in cancer prevention and control. *Lancet Oncol.* **21**, e342–e349 (2020).
19. Ma, T. et al. China's improving inland surface water quality since 2003. *Sci. Adv.* **6**, eaau3798 (2020).
20. River Water Conservancy Commission, Ministry of Water Resources of China. *Distribution Map of Important Control Sections of Water Resources in the Yangtze River Basin* (River Water Conservancy Commission, Ministry of Water Resources of China, 2022).
21. Ministry of Ecology and Environment of China. *Environmental Quality Standards for Surface Water. GB 3838–2002* (Ministry of Ecology and Environment of China, 2002).

22. Liu, J. & Yang, W. Water management. Water sustainability for China and beyond. *Science* **337**, 649–650 (2012).
23. National Bureau of Statistics of China. *China Statistical Yearbook 2021* (China Statistics Press, 2021).
24. Hilbe, J. M. *Negative Binomial Regression* 2nd edn (Cambridge University Press, 2011).
25. Zheng, R. et al. Cancer incidence and mortality in China, 2016. *J. Natl Cancer Cent.* **2**, 1–9 (2022).
26. Zeng, H. et al. Cancer survival in China, 2003-2005: a population-based study. *Int. J. Cancer* **136**, 1921–1930 (2015).
27. Li, Y. in *Water Pollution Control* 25–26 (China WaterPower Press, 2018).
28. Hendryx, M., Conley, J., Fedorko, E., Luo, J. & Armistead, M. Permitted water pollution discharges and population cancer and non-cancer mortality: toxicity weights and upstream discharge effects in US rural-urban areas. *Int. J. Health Geogr.* **11**, 9 (2012).
29. Xu, J. et al. Assessing temporal variations of Ammonia Nitrogen concentrations and loads in the Huaihe River Basin in relation to policies on pollution source control. *Sci. Total Environ.* **642**, 1386–1395 (2018).
30. Zhai, X., Xia, J. & Zhang, Y. Water quality variation in the highly disturbed Huai River Basin, China from 1994 to 2005 by multi-statistical analyses. *Sci. Total Environ.* **496**, 594–606 (2014).
31. State Council of China. *Water Pollution Control Action Plan* (State Council of China, 2015).
32. State Council of China. *Water Pollution Prevention and Control Plan for Key River Basins (2011-2015)* (State Council of China, 2012).
33. JI, X., LI, B., Yang, K. & Sun, Z. Spatial and temporal distribution characteristics of fluoride in surface water of china. *Earth Environ.* **50**, 787–796 (2022).
34. European Commission Directorate-General for Health and Consumers, Scientific Committees. Critical review of any new evidence on the hazard profile, health effects, and human exposure to fluoride and the fluoridating agents of drinking water. http://ec.europa.eu/health/scientific_committees/environmental_risks/docs/scher_o_122.pdf (2011).
35. McDonagh, M. S. et al. Systematic review of water fluoridation. *BMJ* **321**, 855–859 (2000).
36. Vinceti, M., Crespi, C. M., Malagoli, C., Del Giovane, C. & Krogh, V. Friend or foe? The current epidemiologic evidence on selenium and human cancer risk. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* **31**, 305–341 (2013).
37. Vinceti, M. et al. Selenium for preventing cancer. *Cochrane Database Syst. Rev.* **1**, CD005195 (2018).
38. Vinceti, M. et al. Cancer incidence following long-term consumption of drinking water with high inorganic selenium content. *Sci. Total Environ.* **635**, 390–396 (2018).
39. Marschall, T. A., Bornhorst, J., Kuehnelt, D. & Schwerdtle, T. Differing cytotoxicity and bioavailability of selenite, methylselenocysteine, selenomethionine, selenosugar 1 and trimethylselenonium ion and their underlying metabolic transformations in human cells. *Mol. Nutr. Food Res.* **60**, 2622–2632 (2016).
40. Wang, Q., Wang, J., Zhou, J., Ban, J. & Li, T. Estimation of PM2.5-associated disease burden in China in 2020 and 2030 using population and air quality scenarios: a modelling study. *Lancet Planet. Health* **3**, e71–e80 (2019).
41. Cheng, Y. et al. Dramatic changes in Harbin aerosol during 2018–2020: the roles of open burning policy and secondary aerosol formation. *Atmos. Chem. Phys.* **21**, 15199–15211 (2021).
42. Ministry of Ecology and Environment of China. *Action Plan for Deepening the Yangtze River Protection and Restoration Campaign* (Ministry of Ecology and Environment of China, 2022).
43. de Magalhaes, J. P. How ageing processes influence cancer. *Nat. Rev. Cancer* **13**, 357–365 (2013).
44. Balducci, L. & Ershler, W. B. Cancer and ageing: a nexus at several levels. *Nat. Rev. Cancer* **5**, 655–662 (2005).
45. Soerjomataram, I. & Bray, F. Planning for tomorrow: global cancer incidence and the role of prevention 2020–2070. *Nat. Rev. Clin. Oncol.* **18**, 663–672 (2021).
46. Bodirsky, B. L. et al. Reactive nitrogen requirements to feed the world in 2050 and potential to mitigate nitrogen pollution. *Nat. Commun.* **5**, 3858 (2014).
47. Han, D., Huang, G., Liu, L., Zhai, M. & Gao, S. Multi-regional industrial wastewater metabolism analysis for the Yangtze River Economic Belt, China. *Environ. Pollut.* **284**, 117118 (2021).
48. Jin, L., Zhang, G. & Tian, H. Current state of sewage treatment in China. *Water Res.* **66**, 85–98 (2014).
49. National Bureau of Statistics of China. *China Statistical Yearbook 2015* (China Statistics Press, 2015).
50. National Bureau of Statistics of China. *China Statistical Yearbook 2011* (China Statistics Press, 2011).
51. Journal of Moderate Prosperity & State Information Center of China. China ecology development INDEX. https://kns.cnki.net/KXReader/Detail?invoice=iDAOXI12vo%2BQ9AG3650z0O8imKVjznBECDTtlxiozn1DYEGd%2B3ZlnQbYdDI9a%2BECoSjjDpUe6tzGK%2BnP8uJEpOCqVaC31nhAUsQDEUpCNuByq7wfQF%2FewvUef9Y4WOQfhUQB5IUnwFuGzbZdeAZ%2BcZ9hCGjHwtI7%2B6xUvLyqm1A%3D&DBCODE=CJFD&FileName=CHXK202216023&TABLEName=cjfdlast2022&nonce=40F79646A9594C4DAF2A55EFD08EF038&TIMESTAMP=1683082108301&uid=. (2022).
52. Show, P. L., Lau, P. L. & Foo, D. C. Y. Green technologies: innovations, challenges, and prospects. *Clean Technol. Environ. Pol.* **20**, 1939 (2018).
53. Ban, J. et al. Environmental Health Indicators for China: data resources for Chinese environmental public health tracking. *Environ. Health Perspect.* **127**, 44501 (2019).
54. China National Environmental Monitoring Centre, Ministry of Ecology and Environment of China. National surface water environmental quality monitoring network. http://www.cnemc.cn/ (2021).
55. National Office for Cancer Prevention and Control of China, National Cancer Center. *Annual report of the china cancer registry 2019* (China People's Medical Publishing House, 2019).
56. United Nations, Department of Economic and Social Affairs Population Division. *World Population Prospects 2019: Highlights. ST/ESA/SER.A/423* (United Nations, Department of Economic and Social Affairs Population Division, 2019).
57. Xiao, G. *Actual Practice of Spatial Statistics* (China Science Press, 2018).
58. Cressie, N. A. C. *Statistics for Spatial Data* (John Wiley & Sons, 1993).
59. Shi, Z. *Theory and Practice of Spatial Analysis* (China Science Press, 2020).
60. Herath, I. K., Wu, S. J., Ma, M. H., Jianli, W. & Chandrajith, R. Tracing controlling factors of riverine chemistry in a headwater tributary of the Yangtze River, China, inferred from geochemical and stable isotopic signatures. *Environ. Sci. Pollut. Res. Int.* **26**, 23899–23922 (2019).
61. Chen, W., Ni, J., YANG, H., Wei, R. & Yang, Y. Spatial heterogeneity and autocorrelation of polycyclic aromatic hydrocarbons in the sediment of Minjiang River in Fuzhou City. *Environ. Sci.* **33**, 1687–1692 (2012).
62. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery* (2016).
63. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (NIPS, 2017).
64. Khosravi, A., Nazemipour, M., Shinozaki, T. & Mansournia, M. A. Population attributable fraction in textbooks: time to revise. *Glob. Epidemiol.* **3**, 100062 (2021).
65. Mansournia, M. A. & Altman, D. G. Population attributable fraction. *BMJ* **360**, k757 (2018).
66. Chen, T. et al. xgboost: Extreme gradient boosting. R package version 1.6.0.1. https://CRAN.R-project.org/package=xgboost (2022).
67. Liu, Y. & Just, A. SHAPforxgboost: SHAP plots for 'XGBoost'. R package version 0.1.1. https://CRAN.R-project.org/package=SHAPforxgboost (2021).

ACKNOWLEDGEMENTS

This study was supported by the CAMS Innovation Fund for Medical Sciences (grant no: 2021–1-I2M-022). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

AUTHOR CONTRIBUTIONS

J.J. conceived and led the research. H.L. provided expert guidance regarding geography. W.G., X.G. and Z.W. performed the analyses and F.X. and J.Z. performed result verification. F.X, W.H. and W.G. led the software. Y.H. provided software guidance. X.G., L.Z., C.Y. and W.C. led literature research. H.L., W.C., Y.H. and C.Y. supported the data collection and preparation. P.W., Y.C., Y.Z. and J.D. contributed to data curation. Z.W., W.G. and X.G. wrote the original draft. All the authors were involved in rounds of critical revisions, read and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41545-023-00267-5.

Correspondence and requests for materials should be addressed to Jingmei Jiang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.