nature portfolio

# Peer Review File

## Spatial Consistency of Co-exposure to Air and Surface Water Pollution and Cancer in China.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in environmental epidemiology, water contamination, and health disparities

This is an interesting analysis of air pollutants, surface water contaminants, and cancer incidence across 30 provinces in China. It is an ambitious project. The research question is of great importance and I commend the authors for their hard work in this important area. The analysis as presented, however, does not meet the standards of the field, and therefore the interpretations are not appropriate. One suggestion is that the authors could break this analysis up into several papers so that adequate consideration, time, and modification can be spent on each component of the analysis (eg developing the SESEC as one paper; epidemiologic models with cancer incidence as a separate paper, although the timing of water data need to be resolved).

I have several major concerns:

This is a complicated, multi-part analysis. The authors present a non-standard approach for identifying spatial units with “high concentrations” of surface water and air pollutants. Yet, the methods section is extremely short and does not provide enough information such that a reader could reasonably replicate the analysis. A few non-exhaustive examples include:

Unclear how surface water monitoring data were collected, spatial resolution, missingness, etc. How are you handling non-detects? How many water contaminants had 75th percentile values below LODs? Are 75th percentile concentrations for contaminants relevant for health? Etc.

Definition of neighborhood spatial matrix is not clear.

Entire question about causative agents and shared mechanisms is missing here; biological plausibility needs to be explained; otherwise these indicators of poor environmental quality are likely tracking with other causative agents, including social determinants of health. Which of these exposures are established or potential carcinogens (eg Zn? Sulphate? Chemical oxygen demand? these are not relevant at all yet are included.)?

There are several concerns with the interpretation and framing of the PAF analysis. The authors are using outcome data from 2016, but exposure data from 2015 (air) and 2021 (water).

Are the authors confident they can conduct a PAR analysis on ecological, spatial-unit level data, with water pollutant concentrations measured after the cancer incidence data? What of the biologically relevant time frame for cancer development, which differs greatly across outcomes? The interpretations are inappropriate given these considerations.

Concerns about categorizing spatial units into Grades 1-IV:

The interpretation and utility of Moran's I in generating the final "Grades" seems inappropriate. I do not think the interpretation of the local Moran's I or its use in this paper fits the research question of interest. It is also not clear how the authors combined the 'threshold concentration assigned to each pollutant' into the Moran's I calculation to identify clusters. Does this mean the authors categorized each pollutant on a binary scale (exceeded threshold vs did not exceed threshold)? Something else?

The local Moran's I is also in reference to the mean of the variable and it is not clear whether the mean of the variable is health relevant; how does this choice align with the choice to select water threshold values at the 75th percentile?. Given the research objective of this paper, the best utility of Moran's I is to identify whether spatial clustering is happening so that the authors can account for spatial dependence in the epidemiologic models ("cancer-specific key pollutants"). It does not make sense to categorize areas as "high pollutant areas" based on Moran's I clusters of HH, HL, and HN; this does not necessarily identify areas where concentrations are high in a health-relevant manner. The HN/HL clusters just represent something very different than areas with high concentrations yet are categorized together. Why not just identify spatial units where concentrations exceed the thresholds identified in Table 1 (although, even this would be inappropriate given the water thresholds are % based and not health-relevant)? Moreover, the choice to define Groups I-IV (0-1, 2, 3-6 H for air and >=7 for water) seems arbitrary. Whether findings are robust to alternative grouping schemes is not discussed.

Given these considerations, interpretations in paragraph starting on line 156 such as "The common graded scale showed that 90 (41.1%) basic analytic units had high-level pollution in both air and surface water (Grade IV)" seem inappropriate.

Major concerns about cancer-specific epidemiologic analyses:

The epidemiologic analyses are inadequate and do not meet the standards of the field, and therefore interpretations are inappropriate. The authors have not adjusted for any areal-level confounders in these models, including social determinants of health variables that likely confound the association of interest. Given the substantial spatial dependence identified earlier in the paper, the authors also should account for spatial dependence in this modeling approach.

Reviewer #2 (Remarks to the Author): Expert in cancer epidemiology and risk, genetic and environmental factors

Overall, the title is brief and clear. The abstract provides a concise overview of the study. The research question is quite important – to investigate the combined impact of environmental pollutants in air and surface water on cancer incidence.

However, running through the article, we have some concerns and confusion about the study design and findings. Explanations or potential solutions from the authors might be required to make further progress.

1. In this study, the data for air and surface water pollutants were retrieved from 2015 and 2021 respectively, and for cancer incidence from 2016. The pollutant concentrations in air and surface water can change temporally, which can fluctuate not only seasonably but also annually. The levels of pollutants may alter greatly when regional or national policies are implemented to reduce emissions or when the industrialization process begins in a certain area. Thus, we have reservations about the suggested spatial consistency of pollutant measurements in air and surface water when sourced from years with a five-year interval. To allay this concern, the authors might need to provide evidence about the stability of the contaminant condition over the years. Additionally, the effects of environmental pollutants on human health might not manifest immediately, especially changes in cancer incidence in a population that has long latency periods. How do the authors perceive the causal relationship derived from the spatial consistency between cancer incidence and co-pollution grade?

2. Many factors could contribute to the variation in cancer incidence across regions. For example, the age structure of a population, lifestyle, and access to available medical resources could potentially influence the incidence of cancer. Not adjusting for these confounding factors may lead to biased results. Are there any confounding factors that the authors have looked at?

3. Based on the findings presented in Figure 2, there seems to be a protective effect of pollutants on the incidence of liver cancer. However, existing literature suggested the association between environmental pollution exposure (such as $NO_2$ and $PM_{2.5}$ in air, and toxic chemicals in surface water) and an increased risk of liver cancer development. Furthermore, a 'dose-response' relationship is observed in this study - higher co-pollution grades correlate with lower liver cancer incidence. Can the authors provide any insights or explanations for this unexpected association observed?

Beyond these, the clarity of the following aspects may need further elaboration or specificity from the authors.

4. The definition of pollutant cluster patterns is described in the figure legends and methods section, but I am uncertain about the number of neighbouring analytic units to determine the clustering feature of one specific unit.

5. Can the authors explain what the red line in Figure 3 represents?

6. Regarding the calculation of population population-attributable fraction, the cancer attribution was assessed for each grade. Can the authors give more explanation on the calculation? What is the reference group to be compared with? What does it mean by the minimum risk exposure distribution mentioned in the method section?

Reviewer #3 (Remarks to the Author): Expert in cancer epidemiology and risk, genetic and environmental factors

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author): Expert in environmental epidemiology, cancer risk, and pollution

The authors present a well-written study to categorize combined air and water pollutant exposures through a “common grading scale” spatially throughout China and investigate associations between those exposures categories (and the individual pollutants) with cancer incidence in these regions. This study addresses a need for approaches to further understand the relationships of complex, “real world” environmental exposures and cancer risk (and can be applied to other health outcomes as well). The authors present a novel approach to take environmental exposures measured at different scales and categorize into an overall grouping of sorts. There are a few questions that should be addressed, including with respect to the ecologic analyses investigating the associations with cancer incidence and estimating population attributable fractions.

Major points:

1. It is stated that “Threshold concentrations for these environmental quality items were set using the annual limit values recommended by the National Ambient Air Quality Standard for air pollutants”. Are the annual limits necessarily the most biologically relevant when it comes to cancer risk? Since there is a lot of literature on air and water pollutants and cancer risk it would be useful to see what metrics (annual measures, days exceeding threshold values (i.e. measure of peak exposures), etc) show the greatest associations with cancer risk to either justify this choice or provide a discussion point as to the appropriateness of the decisions made.

2. Related to point #1, how robust are the Grade Types to chances in the input pollution metrics (e.g. annual average limit vs # days exceeding a daily limit)?

3. The SHAP (machine learning approach) was used to identify pollutants with the largest impact on cancer incidence. Then those pollutants with the largest impact were used in regression models to get RRs for specific cancer risks, correct? Were the regression models adjusted for any factors? Given the high degree of correlation between some of the pollutants and the potential importance of regional characteristics, such as demographic characteristics of socio-economic factors at the regional level (median household income, median age, %urban/rural/population density, etc) it would be important to examine potential confounding factors with the observed associations.

4. Related to #3, please excuse me, as I’m not familiar with regional demographics in China, but regional socio-economic measures, median age, etc correlate with the co-pollution grade categories? These could be important correlations for consideration with health risk estimates.

5. In an ecologic analysis, if it is the case that the models are not adjusted for other factors that are known to be associated with cancer risk and may correlated with level of co-pollution exposure (and please clarify if that is not the case), is a population attributable fraction calculation the most appropriate measure, since the unadjusted RRs are used in the PAF model?

Minor points:

1. Consider revision to the title to inform the potential reader to the novel environmental exposure assessment: Spatial Consistency of co-exposures to Air and Surface Water pollution and Cancer in China – or something along those lines to better inform the reader of the content.

2. Line 63: “…but little is known about the impact on cancer…”, there is a fair amount of research on individual water pollutants and cancer risk – perhaps be more specific about the research gaps you are talking about (combined pollutants, spatial scale , whatever it is).

3. Lines 67-69: the fact that only 2 Group 1 carcinogens have been identified since 2019 is not necessarily a sign of slow progress. Your point here seem to be there are gaps in knowledge about carcinogenicity because of limited understanding of complex “real world” exposures – be more specific to that, if that is the point you intended.

4. The term “cluster” is used in the manuscript in the context of associations between cancer incidence and pollution exposures, which have spatial components. Maybe describe what is meant by cluster, since you don’t seem to be referring to the typical cancer cluster.

5. Extended figures 1 and 3: for 1, would some indication of the density of monitors in each region add to the readers understanding of potential monitor coverage? For 3, Is the map showing the catchment area of the cancer registry institutes? Would that catchment area be more informative than the location of just the institutes?

6. Line 205-206: “This confirms the combined effect of environmental conditions.”, No test was conducted to confirm this dose-response relationship, I would think “supports” or another term less definitive would be more appropriate.

7. Figure 5: The bar graphs are hard to read – a lot of data is being presented. Perhaps showing the overall environment PAF by cancer in the main graph and move individual pollutants to a supplemental material?

Dear editors and reviewers:

Thank you for providing us with the opportunity to revise our manuscript. We greatly appreciate the insightful comments and suggestions from the reviewers, as they have significantly contributed to improving the quality of our work, particularly in terms of the clarity of our methods and results.

We have attached the revised manuscript for your review. Here is a brief summary of the major revisions we made:

1. We have included a description of the data source and processing methods to provide more transparency and ensure reproducibility.
2. We have added an overview figure of our study design, highlighting the assumptions underlying our analytic approaches and presenting the workflow to improve the understanding of our methodology.
3. We have adjusted for social determinants in incidence risk ratio (RR) estimation, and subsequently updated the population attributable fractions (PAFs) accordingly.
4. We have removed water metal pollutants from our analysis due to the following reasons: a) Metal water pollutants in China have been subjected to stringent control measures since the implementation of surface water quality standards in 2002, resulting in very low concentrations. b) Only a small number of metal pollutants showed a significant relationship with cancer risk compared to the more prominent associations found between organic pollutants and cancer risk.
5. All our analysis results have been updated, and we have included several new supplementary figures to enhance the validity and clarity of our findings.

Please find below our detailed point-to-point response to the reviewers' comments.

**REVIEWER COMMENTS**

Reviewer #1 (Remarks to the Author): Expert in environmental epidemiology, water contamination, and health disparities

This is an interesting analysis of air pollutants, surface water contaminants, and cancer incidence across 30 provinces in China. It is an ambitious project. The research question is of great importance and I commend the authors for their hard work in this important area. The analysis as presented, however, does not meet the standards of the field, and therefore the interpretations are not appropriate. One suggestion is that the authors could break this analysis up into several papers so that adequate consideration, time, and modification can be spent on each component of the analysis (eg developing the SESEC as one paper; epidemiologic models with cancer incidence as a separate paper, although the timing of water data need to be resolved).

**Author reply**:

We greatly appreciate your valuable comments on our manuscript. This research is an important part of our environmental health project, which aims to investigate the complex relationship between real-world environments and human cancer risk. As you noted, this topic represents a significant challenge in terms of methodology. We have carefully considered your reviews and

others and have made critical revisions to the manuscript.
In the revised manuscript, we have included an overview figure (Figure 6) that summarizes the design and analysis processes. This figure clarifies each part of the work and their logical connections. Specifically, we used a spatial evaluation system that harmonizes nationwide data on air, surface water, and cancer incidence in China. We jointly applied spatial analysis, machine learning, and classical statistical models. Following a series of assumptions and justifications, we developed a graded scale of co-pollution that enabled us to transform a complex network with multiple pollutants and multiple types of cancers into an integrated evaluation of the overall relationship. This approach helped us identify, locate, and quantify the key pollutant components that contribute to this relationship. The results can be used to provide etiology clues and possible targets for intervention.
To address your concerns, we have performed additional analyses. Please refer to the point-by-point responses below for more details.

I have several major concerns:
1. This is a complicated, multi-part analysis. The authors present a non-standard approach for identifying spatial units with "high concentrations" of surface water and air pollutants. Yet, the methods section is extremely short and does not provide enough information such that a reader could reasonably replicate the analysis. A few non-exhaustive examples include: Unclear how surface water monitoring data were collected, spatial resolution, missingness, etc. How are you handling non-detects? How many water contaminants had 75th percentile values below LODs? Are 75th percentile concentrations for contaminants relevant for health? Etc.
**Author reply**:
We have made revisions to the manuscript in order to provide detailed information for the replication of our analysis. Our responses are organized according to the numbered points you provided.
To enhance the methods section (Page 21), we have added the following details about the data source and processing. We included information regarding the establishment of surface water monitoring sections. These sections were selected strategically to represent the natural characteristics of river network density, runoff supply, and hydrological features. The guidelines outlined in the National Technical Specifications for Surface Water Environmental Monitoring were followed during this process. For measuring pollutant indicators, various methods such as spectrophotometry, colorimetry, and other standardized techniques recommended by the Environmental Quality Standards for Surface Water (EQSSW) were employed. It is important to note that some records fell below the limits of detection (LODs) of these methods. This occurred because alternative methods, which were equivalent or superior and complied with the applicability testing and International Organization for Standardization requirements, were utilized. Additionally, in some monitoring sections (ranging from 1.8% to 9.2% per month), data was missing due to conditions such as dryness, freezing, or other factors. To address this issue, we performed averaging over the analytic year for each specific section. In order to identify high-value status, we calculated the 75th percentile values for each pollutant. This practical criterion provides an intuitive approach and allows for sufficient variability in various pollutant indicators and their binary conditions when conducting high-value analysis. Furthermore, it offers a uniform method for determining thresholds of multiple pollutants, as we have observed their correlation with multiple types of cancer (npj Clean Water 2023;6:53). All of these calculated 75th percentile values exceeded the LODs established by the EQSSW-

recommended methods.

2. Definition of neighborhood spatial matrix is not clear.

**Author reply**:

We have incorporated the information regarding the use of weights in spatial analysis. We defined weights as the inverse of the distance between two analytic units and set the distance threshold as the minimum distance that guarantees each unit has at least one neighboring unit. Please refer to this definition in the revised manuscript on Page 22.

3. Entire question about causative agents and shared mechanisms is missing here; biological plausibility needs to be explained; otherwise these indicators of poor environmental quality are likely tracking with other causative agents, including social determinants of health. Which of these exposures are established or potential carcinogens (eg Zn? Sulphate? Chemical oxygen demand? these are not relevant at all yet are included.)?

**Author reply**:

As per your suggestion, we have included a discussion on possible carcinogenic mechanisms. For instance, we have highlighted the widespread use of permanganate index as a comprehensive indicator of nitrite and organic pollutants in surface water in China. Nitrite, when combined with amines, produces nitrosamine, which is a known carcinogen. Additionally, direct carcinogens and pre-carcinogens are organic substances that can cause changes in DNA. Please see Page 7 for further details.

Furthermore, we have added a discussion (Page 7) highlighting that pollutants do not exist in isolation, and therefore, do not act independently. Even pollutants that have not yet been acknowledged as carcinogenic can still impact the risk of cancer in populations. These impacts could arise from intricate interactions with known carcinogens. Several previous studies have provided support for this assumption (please see references no. 33 and 52), emphasizing the need for further investigation into the potential network mechanisms that connect multiple pollutants to the development of cancer.

We have also taken note of your suggestion regarding the screening of pollutants before performing negative binomial regression analysis. We have retained only those pollutants that positively contribute to the cancer risk (please refer to Page 23), and we have also adjusted for social determinants of health in effect estimation (Page 23). These adjustments could help enhance the validity of our observations. Please refer to Table 2 and Extended Data Figure 8 for further details.

4. There are several concerns with the interpretation and framing of the PAF analysis. The authors are using outcome data from 2016, but exposure data from 2015 (air) and 2021 (water). Are the authors confident they can conduct a PAR analysis on ecological, spatial-unit level data, with water pollutant concentrations measured after the cancer incidence data? What of the biologically relevant time frame for cancer development, which differs greatly across outcomes? The interpretations are inappropriate given these considerations.

**Author reply**:

We have added a discussion on several challenges of using typical epidemiology designs in effect estimation on Page 8. Firstly, it remains unknown how long environmental exposure leads to the development of different types of cancer. Secondly, accurate quantification of individual exposure levels to air and water is challenging. Thirdly, the complex interplay of genetic and environmental factors for cancer research makes it difficult to generalize existing pieces of epidemiology findings. To overcome these challenges, we examined the air-water-cancer association from another perspective, spatial analysis. One significant advantage of the study is its coverage of a vast majority of China's geographical area with a population of 377 million, providing statistical "parameters" that ensure temporal stability for both environmental data (annualized average concentrations) and cancer data (annual incidence). We have included evidence on the temporal stability of the environment and cancer data, which can be found in Extended Data Figures 9-10. The temporal stability is particularly pronounced when making comparisons between different regions, indicating that the environmental data used in our analysis are valid substitute indicators of the population exposure level. Please see added clarification points on Page 8.

Regarding PAR estimation, two factors were involved: RR and environmental exposure. Since we used national data, the environmental exposure represents the entire country well. For RR estimation, we considered the highest and lowest possible concentrations for each pollutant, using contentious variables to retain as much information as possible for exposures. To control for potential confounding, we have adjusted for social factors based on your suggestions. We believe that these supplements and updates will help provide a more reasonable interpretation of PAR.

Concerns about categorizing spatial units into Grades 1-IV:

The interpretation and utility of Moran's I in generating the final "Grades" seems inappropriate. I do not think the interpretation of the local Moran's I or its use in this paper fits the research question of interest. It is also not clear how the authors combined the 'threshold concentration assigned to each pollutant' into the Moran's I calculation to identify clusters. Does this mean the authors categorized each pollutant on a binary scale (exceeded threshold vs did not exceed threshold)? Something else?

**Author reply**:

We categorized the pollutants based on threshold concentrations, and using the binary variables (high value or not), we then calculated the modified local Moran's I statistic for each pollutant. This approach allowed for independent evaluations of environmental conditions in specific analytic units and provided a broader understanding of clustering spatial patterns. Please refer to related discussion on Page 5. To emphasize the role of local Moran's I, we have included a new figure (Figure 1) that highlights its significance and added interpretations to the results. Additionally, we have made revisions to the methods section to address any confusion (Pages 21-22).

6. The local Moran's I is also in reference to the mean of the variable and it is not clear whether the mean of the variable is health relevant; how does this choice align with the choice to select water threshold values at the 75th percentile? Given the research objective of this paper, the best utility of Moran's I is to identify whether spatial clustering is happening so that the authors can account for spatial dependence in the epidemiologic models ("cancer-specific key pollutants"). It does not make sense to categorize areas as "high pollutant areas" based on Moran's I clusters of HH, HL, and HN; this does not necessarily identify areas where concentrations are high in a health-relevant manner. The HN/HL clusters just represent something very different than areas with high concentrations yet are categorized together. Why not just identify spatial units where concentrations exceed the thresholds identified in Table 1 (although, even this would be inappropriate given the water

thresholds are % based and not health-relevant)?

**Author reply**:

We utilized binary variables to denote high pollutant values. The local Moran's I statistic now references the chosen threshold (i.e., the 75th percentile for water pollutants and limit values for air pollutants) instead of the mean. For more details, please refer to the revised methods section on Pages 21-22.

The employment of Moran's method served a dual purpose. Firstly, as you mentioned, we employed it to verify the presence of spatial dependence for each pollutant. This is a prerequisite for our subsequent examination of the real-world environment, encompassing correlations within pollutants, across different pollutants, and even across environmental media. Secondly, the modified Moran's *I* takes into account both the binary pollution status (i.e., whether the unit exhibits high-value pollution on its own) and the status of neighboring units. This is crucial due to the constant movement of air and surface water, as well as the boundless nature of pollution. In Figure 1, we showcase the clustering patterns for each pollutant, which could be highly valuable for crafting regionally-tailored environmental policies. Please see the interpretation on Page 5.

Moreover, the choice to define Groups I-IV (0-1, 2, 3-6 H for air and >=7 for water) seems arbitrary. Whether findings are robust to alternative grouping schemes is not discussed.

**Author reply**:

Our selection of the grouping scheme was based on three considerations: order consistency, sufficient interval, and appropriate size for each group. Following your recommendation, we also investigated an alternative grouping scheme that meet these conditions, and found that the results were robust to this change. Please refer to Extended Data Figure 6 and added discussion on Page 22.

7. Given these considerations, interpretations in paragraph starting on line 156 such as "The common graded scale showed that 90 (41.1%) basic analytic units had high-level pollution in both air and surface water (Grade IV)" seem inappropriate.

**Author reply**:

We have revised the wording of the sentence as follows: "Based on our empirical grouping on the common graded scale, 78 basic analytic units (35.6%) had high-level pollution in both air and surface water, which corresponds to Grade IV on the scale" (Page 39).

8. Major concerns about cancer-specific epidemiologic analyses:

The epidemiologic analyses are inadequate and do not meet the standards of the field, and therefore interpretations are inappropriate. The authors have not adjusted for any areal-level confounders in these models, including social determinants of health variables that likely confound the association of interest. Given the substantial spatial dependence identified earlier in the paper, the authors also should account for spatial dependence in this modeling approach.

**Author reply**:

Based on your suggestion, we have made efforts to gather information on confounding factors at the area level. The revised manuscript includes the following covariates: per capita gross domestic product, the fraction of the population aged 65 years and older, and the urbanization rate. We presented results in the revised manuscript both with and without adjusting for these social

covariates. Please see Table 2 and Extended Data Figure 8 for more details.
Spatial regression models like geographically weighted regression (GWR) assume that the effect parameters vary across different locations. While this can enhance the precision of local estimates, it may hinder overall interpretation and increase sensitivity to errors. Therefore, instead of GWR, we employed negative binomial regression, a commonly used method in cancer research, to account for potential over-dispersion in cancer incidence data. Negative binomial regression assumes a consistent effect, aligning with the focus of our analysis. The results indicate small standard errors in effect estimation across three analytical scenarios: a) regression of co-pollution grades on cancer incidence (see Figure 3), b) regression of a specific pollutant on cancer incidence, adjusting for other pollutants, and c) regression of a specific pollutant on cancer incidence, adjusting for other pollutants as well as social factors (see Table 2 and Extended Data Figure 8). These findings suggest minimal influence of inter-unit correlation on effect estimation, thereby validating our results.

**Reviewer #2 (Remarks to the Author): Expert in cancer epidemiology and risk, genetic and environmental factors**

Overall, the title is brief and clear. The abstract provides a concise overview of the study. The research question is quite important – to investigate the combined impact of environmental pollutants in air and surface water on cancer incidence.
However, running through the article, we have some concerns and confusion about the study design and findings. Explanations or potential solutions from the authors might be required to make further progress.
1. In this study, the data for air and surface water pollutants were retrieved from 2015 and 2021 respectively, and for cancer incidence from 2016. The pollutant concentrations in air and surface water can change temporally, which can fluctuate not only seasonably but also annually. The levels of pollutants may alter greatly when regional or national policies are implemented to reduce emissions or when the industrialization process begins in a certain area. Thus, we have reservations about the suggested spatial consistency of pollutant measurements in air and surface water when sourced from years with a five-year interval. To allay this concern, the authors might need to provide evidence about the stability of the contaminant condition over the years. Additionally, the effects of environmental pollutants on human health might not manifest immediately, especially changes in cancer incidence in a population that has long latency periods. How do the authors perceive the causal relationship derived from the spatial consistency between cancer incidence and co-pollution grade?
**Author reply**:
Thank you for your valuable comments. The suggestions were very helpful.
In terms of temporality, there is still limited knowledge regarding the duration of environmental exposure that leads to the development of different types of cancer, particularly in relation to water and air environments. These environments are unique as individuals are exposed to them from birth and throughout their lives. To address this, we utilized annual average concentrations to reflect potential long-term impacts of pollution on cancer. For further clarification, please refer to the methods details on Page 21. According to your suggestion, we have generated maps displaying PM2.5 levels in China from 2000-2015. These maps help demonstrate the consistent patterns of

pollution across the country. It is worth noting that the overall pattern has remained relatively stable over the years, as depicted in Extended Data Figure 9. The findings indicate that local interventions aimed at reducing pollution may have limited impact on the overall pattern across the country. Additionally, we analyzed cancer incidence data over the years and found that the overall patterns also remained stable across the country (illustrated in Extended Data Figure 10), which further strengthens the evidence of a spatial association between pollution and cancers.

While it is important to acknowledge that spatial consistency alone does not establish a causal relationship, we believe validity of our findings would not be significantly impacted by the latency period issue.

2. Many factors could contribute to the variation in cancer incidence across regions. For example, the age structure of a population, lifestyle, and access to available medical resources could potentially influence the incidence of cancer. Not adjusting for these confounding factors may lead to biased results. Are there any confounding factors that the authors have looked at?

**Author reply**:

Based on your recommendation, we have gathered relevant national data, which includes per capita gross domestic product, the proportion of the population aged 65 years and older, and the rate of urbanization. After carefully adjusting for these factors, the results remained consistent and reliable. For further details, please refer to Pages 6-7, Table 2 and Extended Data Figure 8.

3. Based on the findings presented in Figure 2, there seems to be a protective effect of pollutants on the incidence of liver cancer. However, existing literature suggested the association between environmental pollution exposure (such as $NO_2$ and $PM_{2.5}$ in air, and toxic chemicals in surface water) and an increased risk of liver cancer development. Furthermore, a 'dose-response' relationship is observed in this study - higher co-pollution grades correlate with lower liver cancer incidence. Can the authors provide any insights or explanations for this unexpected association observed?

**Author reply**:

Thank you for bringing up this important question. In new Figure 3, which is Figure 2 in previous submission, all pollutants were considered on a common graded scale of co-pollution, which reflects different categories of air and surface water environmental conditions rather than the specific effects of individual pollutants. However, when we specifically focused on the pollutants and adjusted for social factors, we did not observe any relationship with liver cancer. These findings suggest that social factors play a significant role in the development of liver cancer. In fact, the prevalence of hepatitis B and C largely determines the profile of the liver cancer epidemic in China, which is high in boundary regions where the environmental quality is favorable. This observation indicates that the impact of the environment alone is insufficient to explain the burden of liver cancer. For a more detailed discussion on these points, please refer to Pages 6-7.

4. The definition of pollutant cluster patterns is described in the figure legends and methods section, but I am uncertain about the number of neighbouring analytic units to determine the clustering feature of one specific unit.

**Author reply**:

We have provided additional details regarding this point in the methods section. Specifically, we incorporated weights in our spatial analysis, which were defined as the inverse of the distance between two analytic units. To calculate this weight, we used the minimum distance that ensures each unit has at least one neighboring unit as the distance threshold. For further information, please refer to Page 22.

5. Can the authors explain what the red line in Figure 3 represents?

**Author reply**:

We have included an explanation of the red line in the revised figure legends. Please refer to Page 18. The red line represents the top 10 pollutants based on their SHAP values for the specific cancer type being analyzed.

6. Regarding the calculation of population population-attributable fraction, the cancer attribution was assessed for each grade. Can the authors give more explanation on the calculation? What is the reference group to be compared with? What does it mean by the minimum risk exposure distribution mentioned in the method section?

**Author reply**:

We have made revisions to the methods section to provide further clarification on these points. Please refer to Page 23.

To be specific here, we utilized the population attributable fraction (PAF) to assess the cancer attribution in each grade. This was done by using the burden attributable to continuous exposure method to calculate the PAF for each grade separately. The PAF was calculated for each cancer type, with the contribution of the $i$th pollutant in the $j$th ($j$=I, II, III, IV) grade determined using the following formula:

$$\mathrm{PAF}_{ij} = \frac{\int_{\min}^{\max} RR_i(x)P_{ij}(x)dx - \int_{\min}^{\max} RR_i(x)P_i^*(x)dx}{\int_{\min}^{\max} RR_i(x)P_{ij}(x)dx}$$

where $x$ represents the pollutant concentration, min and max are the minimum and maximum concentration values of each pollutant. $RR_i(x)$ is the incidence rate ratio at concentration $x$, which is obtained by negative binomial regression while adjusting for pollutants and social factors. The upper limit of $RR_i(x)$ is restricted to the value of the rate ratio at the threshold concentration to obtain the minimum estimate of PAF. $P_{ij}(x)$ is the concentration distribution of the $i$th pollutant in the $j$th grade. $P_i^*(x)$ is the theoretical minimum risk exposure distribution. It assumes that all units are exposed to concentrations below the threshold. By comparing the actual and assumed (referenced) conditions, we can estimate the burden.

Reviewer #3 (Remarks to the Author): Expert in cancer epidemiology and risk, genetic and environmental factors

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

**Author reply**:

Thank you for the reviews. They are of great help for improving the quality of our work.

**Reviewer #4** (Remarks to the Author): Expert in environmental epidemiology, cancer risk, and pollution

The authors present a well-written study to categorize combined air and water pollutant exposures through a "common grading scale" spatially throughout China and investigate associations between those exposures categories (and the individual pollutants) with cancer incidence in these regions. This study addresses a need for approaches to further understand the relationships of complex, "real world" environmental exposures and cancer risk (and can be applied to other health outcomes as well). The authors present a novel approach to take environmental exposures measured at different scales and categorize into an overall grouping of sorts. There are a few questions that should be addressed, including with respect to the ecologic analyses investigating the associations with cancer incidence and estimating population attributable fractions.

**Author reply**:

Thank you for providing valuable feedback on our manuscript. We have carefully considered your questions and made revisions accordingly. Please refer to the revised manuscript and our responses below.

Major points:

1. It is stated that "Threshold concentrations for these environmental quality items were set using the annual limit values recommended by the National Ambient Air Quality Standard for air pollutants". Are the annual limits necessarily the most biologically relevant when it comes to cancer risk? Since there is a lot of literature on air and water pollutants and cancer risk it would be useful to see what metrics (annual measures, days exceeding threshold values (i.e. measure of peak exposures), etc) show the greatest associations with cancer risk to either justify this choice or provide a discussion point as to the appropriateness of the decisions made.

**Author reply**:

In our investigation of the relationship between the "real-world" environment and cancer, we examined multiple pollutants and multiple types of cancer. Therefore, the analysis is primarily exploratory in nature. The decision to use annual average concentrations was based on several considerations: a) Annual average levels tend to smooth out temporal fluctuations, which are common in air and surface water pollutants. b) They represent long-term exposure levels, and cancers are often linked to long-term effects. c) Our previous study has shown that the annual average levels of surface water pollutants are associated with increased risk for multiple types of cancer (npj Clean Water 2023;6:53). We have incorporated these reasons into the manuscript as per your suggestion. Please refer to Page 21.

2. Related to point #1, how robust are the Grade Types to chances in the input pollution metrics (e.g. annual average limit vs # days exceeding a daily limit)?

**Author reply**:

We were unable to investigate exposure metrics such as days exceeding threshold, which may be more precise than annual measures, due to limitations in the data on air and water pollution, which were only released on a monthly basis. As a result, we adopted a conservative approach by using

annual average levels to examine the relationship between pollutants and cancers. Please refer to our response to your comment #1 for further justification of this choice.
In response to your suggestion and feedback from other reviewers, we have conducted a sensitivity analysis regarding the grading of co-pollution status on cancer. Please refer to Extended Data Figure 6 and the related discussions on Page 6 for further details.

3. The SHAP (machine learning approach) was used to identify pollutants with the largest impact on cancer incidence. Then those pollutants with the largest impact were used in regression models to get RRs for specific cancer risks, correct? Were the regression models adjusted for any factors? Given the high degree of correlation between some of the pollutants and the potential importance of regional characteristics, such as demographic characteristics of socio-economic factors at the regional level (median household income, median age, %urban/rural/population density, etc) it would be important to examine potential confounding factors with the observed associations.
**Author reply**:
Regarding the first question, we utilized the SHAP approach to identify pollutants that had the greatest impact on cancer incidence. We retained the pollutants that positively contribute to the cancer risk and then employed a multivariable negative binomial regression to quantify the magnitude of their effects. We have included these method descriptions on Pages 22-23 for further clarification.
In response to your second question, we conducted two sets of multiple negative binomial regression analyses based on your suggestion. The first set of analyses adjusted for the selected pollutants, while the second set additionally adjusted for region-specific covariates, including per capita gross domestic product, fraction of population aged 65 years and older, and urbanization rate. The results demonstrated small standard errors in the estimation of effects, indicating that the correlation between some pollutants had minimal impact on the validity of our findings. Please find the added results in Table 2 and Extended Figure 8, along with a detailed description of the methods on Page 23.

4. Related to #3, please excuse me, as I'm not familiar with regional demographics in China, but regional socio-economic measures, median age, etc correlate with the co-pollution grade categories? These could be important correlations for consideration with health risk estimates.
**Author reply**:
We have taken your suggestion into account and included several regional social covariates in our updated analysis. We compared the estimated RRs with and without adjustment for these covariates. For more information and further clarification, please refer to our response to your comment #3.

5. In an ecologic analysis, if it is the case that the models are not adjusted for other factors that are known to be associated with cancer risk and may correlated with level of co-pollution exposure (and please clarify if that is not the case), is a population attributable fraction calculation the most appropriate measure, since the unadjusted RRs are used in the PAF model?
**Author reply**:
According to your suggestion, we have calculated the RRs by adjusting for both the pollutants and social factors. We have used these adjusted RRs to estimate the population attributable fractions (PAFs). The updated results can be found on Page 7, and we have also made revisions to Table 2

and Figure 5 to reflect these changes. Please refer to the updated results and figure for more details.

Minor points:

1. Consider revision to the title to inform the potential reader to the novel environmental exposure assessment: Spatial Consistency of co-exposures to Air and Surface Water pollution and Cancer in China – or something along those lines to better inform the reader of the content.

**Author reply**:

We have taken your advice into consideration and made the necessary revision to the title as you suggested.

2. Line 63: "…but little is known about the impact on cancer…", there is a fair amount of research on individual water pollutants and cancer risk – perhaps be more specific about the research gaps you are talking about (combined pollutants, spatial scale , whatever it is).

**Author reply**:

We have revised the wording of the sentence as follows: "While previous studies have investigated the impact of specific types of water pollution at a local level, there remains a limited understanding of the potential health effects of exposure to combined pollutants and their association with cancer incidence". Pleased see the revisions on Page 3.

3. Lines 67-69: the fact that only 2 Group 1 carcinogens have been identified since 2019 is not necessarily a sign of slow progress. Your point here seem to be there are gaps in knowledge about carcinogenicity because of limited understanding of complex "real world" exposures – be more specific to that, if that is the point you intended.

**Author reply**:

We have revised the wording of the sentences as "the existing evidence regarding environmental carcinogenicity has been developed in a fragmented manner, lacking a comprehensive evaluation system for understanding the holistic relationship between the real-world environment and cancer". Please see the revisions on Page 3.

4. The term "cluster" is used in the manuscript in the context of associations between cancer incidence and pollution exposures, which have spatial components. Maybe describe what is meant by cluster, since you don't seem to be referring to the typical cancer cluster.

**Author reply**:

To avoid any ambiguity, we have incorporated a clear definition of the spatial analysis term in question. Please refer to Page 5 for further details and clarification.

5. Extended figures 1 and 3: for 1, would some indication of the density of monitors in each region add to the readers understanding of potential monitor coverage? For 3, Is the map showing the catchment area of the cancer registry institutes? Would that catchment area be more informative than the location of just the institutes?

**Author reply**:

We have added description of the data source for the air quality data, with a web link that details the location of air quality monitors. Please also refer to Extended Figure 3, which shows the catchment areas of the cancer registry institutes.

6. Line 205-206: “This confirms the combined effect of environmental conditions.”, No test was conducted to confirm this dose-response relationship, I would think “supports” or another term less definitive would be more appropriate.

**Author reply**:

We have taken your suggestion into consideration and made the necessary modifications to the statement as suggested. Please refer to Page 6.

7. Figure 5: The bar graphs are hard to read – a lot of data is being presented. Perhaps showing the overall environment PAF by cancer in the main graph and move individual pollutants to a supplemental material?

**Author reply**:

We have recreated the figure in question, and we invite you to refer to the details provided in online Extended Data Table 4.

Thank you all for your time and effort in reviewing our work, and we hope that these modifications have addressed all of your concerns. Please let us know if there is anything else we can do to improve the manuscript. We look forward to receiving your positive feedback.

Authors

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

The authors did a lot of work based on previous feedback and made substantial improvements. The revisions including the new diagrams and discussion section added to the manuscript, along with the textual responses, have effectively addressed all of my previous concerns. I don't have any additional comments based on the latest version.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author):

The authors' thoughtful consideration of comments are revisions to the manuscript, particularly the additional information in the methods section, are greatly appreciated. As other reviewers have noted, there is a temporality issue with using exposure data from 2015 (air) and 2021 (surface water) and cancer incidence from 2016, since cancers may take a few to many years to develop. The authors have responded to this with additional discussion and PM2.5 mapping over time. While the PM2.5 maps show only a minor degree of change over time, PM does not directly correspond to the Grade types. Since relevant exposures may have happened 5, 10, 15+ years before development of many of the cancers, is it possible to map change in grade types over 10 years (or some sensible timeframe)? Since the association is between Grade type and cancer incidence, some confirmation on the stability of the Grade types spatially would enhance the confidence in the observed associations.

Two minor points: 1. For table 2, listing the social factors that were included in model 2 (M2) in the legend would be useful, and 2. In figure 5a, the y-axis label is hard to read (resolution or size issue of something to be worked out in publication?).

Assessment of the response to Reviewer #1:

Overall, the authors made a good effort to respond to the comments from Review #1.

For comment #1, there is additional methods information provided, but a few things could still be clarified. For example, with regards to exposure data that is <LOD or missing the authors say "…we employed averaging techniques over the analytic year for each specific section." But how were values <LOD or missing data handled in the calculation of that average? Is missing data expected to impact relative rankings of the annual averages, for example, for water missing data during dry or freezing periods happen more in some areas than others and how does that impact the calculated average? Similar for air pollution monitoring, are there differential patterns by location of missing data that might impact the relative pollution ranks?

I'm also not sure what this sentence means: "…some records fell below the limits of detection (LODs) of these methods. This occurred because alternative methods, which were deemed equivalent or superior and complied with the applicability testing and International Organization for Standardization requirements, were permitted." The LOD is a quality of the analytic procedure, so having alternative methods may change the LOD. Maybe just explaining how <LOD values were handled is sufficient.

Responses to comments #2 and #3 seem fairly straightforward.

Comments #4 and #8, I think there are going to be questions/concerns with the cancer risk calculations. The exposure data is either closer to the cancer outcome data temporally than would be expected to have a causal effect (for air pollutants: 1-yr prior to cancer record is not enough lag period) or for the water data it is after the cancer outcome data. Therefore, providing some assurance as to the stability of these exposures (in a relative sense at least high exposure areas are historically high exposure areas, etc) is important. The authors do provide some support for this exposure stability, but Fig 9 only shows PM data and the scale changes on each map. The patterns of highs and low levels seems fairly stable, but the absolute values change – does this have impact on the models or risk estimates? And what about the water pollution patterns? They do incorporate some other social factors into their risk models, which is important and provide justification for their regression model choice.

Reviewer #5 (Remarks to the Author): Technical reviewer, expert in epidemiology and statistics, replaces Reviewer #1

This manuscript explores an important topic of how air and water pollution is related to cancer incidence in China. It uses comprehensive data to address this issue. However, I have major concerns about the statistical analysis in this paper. Please see my comments below.

1. The study design needs to be further clarified. For example, a clear sentence of how the 219 basic analytic units are selected is needed; what's the relationship between them and CRI or prefecture-level area? How to deal with data from multiple air/water monitoring sites and multiple CRIs within each basic analytic unit? How is values below LOD handled, which could impact the averaging process? Also, have the authors explored the characteristics of areas included in the analysis versus those that were not included? This may seriously bias the results because they were probably not included due to low socio-economic status and less accessibility, as can be seen from Extended Data Figure 1, 2 and 3. A similar map showing the selected 219 units is also needed to understand the spatial location of these units. Also, what method/software is used to calculate spatial autocorrelation?

2. The description and justification of using Moran's I to produce clusters is still confusing. What's the reason to first categorize each pollutant into high vs low? The local Moran's I is proposed for continuous variables, which fits the purpose just fine. Has the modified Moran's local I been investigated for its performance? If not, I would not trust it serves the same purpose as the original Moran's local I. Also the authors didn't give any detail on how to translate the local Moran's I, which is a continuous measure, to clusters, and it's unclear to me where the names of these clusters come from (clustered, outlier…). Once the clusters were assigned, the authors need to list the specific criteria for grouping, with details. It is also needed to clearly define the criteria of grades in main text/methods.

3. The description of SHAP is over-simplified. First, the reference of SHAP has not been published in a peer-reviewed journal. Second, random forest is already able to select most important predictors, what's the reason to add SHAP on top of it? Third, random forest has many tuning parameters. How are these specified or trained? The choice of 10 pollutants is arbitrary, simply being half of the total number is not a good enough reason. For the outcome model, since these spatial areas are not

independent from each other – those close by are likely more similar to each other, spatial correlation needs to be taken into account.

Dear editors and reviewers:

We appreciate your feedback and the valuable comments and suggestions on our revision version 1. We have attached the updated manuscript for your review. Please see our point-by-point responses below.

**REVIEWER COMMENTS**

Reviewer #2 (Remarks to the Author):

The authors did a lot of work based on previous feedback and made substantial improvements. The revisions including the new diagrams and discussion section added to the manuscript, along with the textual responses, have effectively addressed all of my previous concerns. I don't have any additional comments based on the latest version.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author):

The authors' thoughtful consideration of comments are revisions to the manuscript, particularly the additional information in the methods section, are greatly appreciated. As other reviewers have noted, there is a temporality issue with using exposure data from 2015 (air) and 2021 (surface water) and cancer incidence from 2016, since cancers may take a few to many years to develop. The authors have responded to this with additional discussion and PM2.5 mapping over time. While the PM2.5 maps show only a minor degree of change over time, PM does not directly correspond to the Grade types. Since relevant exposures may have happened 5, 10, 15+ years before development of many of the cancers, is it possible to map change in grade types over 10 years (or some sensible timeframe)? Since the association is between Grade type and cancer incidence, some confirmation on the stability of the Grade types spatially would enhance the confidence in the observed associations.

**Author reply**:

We deeply appreciate your support in reviewing our manuscript. Your comments are very helpful and constructive. Spatial stability of Grade types can indeed provide a strong support of our findings. However, we lacked historical data regarding this information. To enhance the explanation of this issue, we have added a critical discussion about this topic in the revised manuscript. Please refer to the last two paragraphs in the Discussion section for further details.

Two minor points: 1. For table 2, listing the social factors that were included in model 2 (M2) in the legend would be useful, and 2. In figure 5a, the y-axis label is hard to read (resolution or size issue of something to be worked out in publication?).

**Author reply**:

We have made the revisions as suggested. Please see the revised Table 2 and Figure 5.

Assessment of the response to Reviewer #1:

Overall, the authors made a good effort to respond to the comments from Review #1.

For comment #1, there is additional methods information provided, but a few things could still be clarified. For example, with regards to exposure data that is <LOD or missing the authors say "…we employed averaging techniques over the analytic year for each specific section." But how were values <LOD or missing data handled in the calculation of that average? Is missing data expected to impact relative rankings of the annual averages, for example, for water missing data during dry or freezing periods happen more in some areas than others and how does that impact the calculated average? Similar for air pollution monitoring, are there differential patterns by location of missing data that might impact the relative pollution ranks? I'm also not sure what this sentence means: "…some records fell below the limits of detection (LODs) of these methods. This occurred because alternative methods, which were deemed equivalent or superior and complied with the applicability testing and International Organization for Standardization requirements, were permitted." The LOD is a quality of the analytic procedure, so having alternative methods may change the LOD. Maybe just explaining how <LOD values were handled is sufficient.

**Author reply**:

We agree with your comment and have added details and an explanation about how LODs and missing values were handled. Please refer to the first paragraph (Data source and processing) in the Methods section of the revised manuscript.

Responses to comments #2 and #3 seem fairly straightforward.

Comments #4 and #8, I think there are going to be questions/concerns with the cancer risk calculations. The exposure data is either closer to the cancer outcome data temporally than would be expected to have a causal effect (for air pollutants: 1-yr prior to cancer record is not enough lag period) or for the water data it is after the cancer outcome data. Therefore, providing some assurance as to the stability of these exposures (in a relative sense at least high exposure areas are historically high exposure areas, etc) is important. The authors do provide some support for this exposure stability, but Fig 9 only shows PM data and the scale changes on each map. The patterns of highs and low levels seems fairly stable, but the absolute values change – does this have impact on the models or risk estimates? And what about the water pollution patterns? They do incorporate some other social factors into their risk models, which is important and provide justification for their regression model choice.

**Author reply**:

Your comment is very important. To enhance the explanation of this issue, we have added a critical discussion about this topic in the revised manuscript. Please refer to the last two paragraphs in the

Discussion section for further details. In addition, we have recreated each map in Extended Data Fig. 7 and categorized the units according to their yearly quartiles.

Reviewer #5 (Remarks to the Author): Technical reviewer, expert in epidemiology and statistics, replaces Reviewer #1

This manuscript explores an important topic of how air and water pollution is related to cancer incidence in China. It uses comprehensive data to address this issue. However, I have major concerns about the statistical analysis in this paper. Please see my comments below.

1. The study design needs to be further clarified. For example, a clear sentence of how the 219 basic analytic units are selected is needed; what's the relationship between them and CRI or prefecture-level area? How to deal with data from multiple air/water monitoring sites and multiple CRIs within each basic analytic unit? How is values below LOD handled, which could impact the averaging process? Also, have the authors explored the characteristics of areas included in the analysis versus those that were not included? This may seriously bias the results because they were probably not included due to low socio-economic status and less accessibility, as can be seen from Extended Data Figure 1, 2 and 3. A similar map showing the selected 219 units is also needed to understand the spatial location of these units. Also, what method/software is used to calculate spatial autocorrelation?

**Author reply**:

Thank you for your professional advice. Your in-depth insight into the details has greatly contributed to the improvement of our manuscript. We have added details about the definition and location of the 219 basic analytic units, data preprocessing, and comparison between the included areas and those not included, among others. Please refer to the second paragraph in the "Spatial Evaluation System for Environment and Cancer" section, the second paragraph in the Discussion section, the first section (Data source and processing), and the last paragraph in the Methods section, Extended Data Table 5, and Extended Data Fig. 1.

2. The description and justification of using Moran's I to produce clusters is still confusing. What's the reason to first categorize each pollutant into high vs low? The local Moran's I is proposed for continuous variables, which fits the purpose just fine. Has the modified Moran's local I been investigated for its performance? If not, I would not trust it serves the same purpose as the original Moran's local I. Also the authors didn't give any detail on how to translate the local Moran's I, which is a continuous measure, to clusters, and it's unclear to me where the names of these clusters come from (clustered, outlier…). Once the clusters were assigned, the authors need to list the specific criteria for grouping, with details. It is also needed to clearly define the criteria of grades in main text/methods.

**Author reply**:

Your comment is very important. To dispel any confusion, we have added a separate paragraph to complement the whole process of determining pollution grade. We have also explained the details and considerations of the revised Moran's index. Please refer to the first three paragraphs in the sections "Distribution of co-pollution measured on a common graded scale" in the main manuscript, and "Modified local Moran's I index for identification of spatial clustering patterns" in the Methods.

3. The description of SHAP is over-simplified. First, the reference of SHAP has not been published in a peer-reviewed journal. Second, random forest is already able to select most important predictors, what's the reason to add SHAP on top of it? Third, random forest has many tuning parameters. How are these specified or trained? The choice of 10 pollutants is arbitrary, simply being half of the total number is not a good enough reason. For the outcome model, since these spatial areas are not independent from each other – those close by are likely more similar to each other, spatial correlation needs to be taken into account.

**Author reply**:

As you suggested, we have replaced the reference for SHAP and added more details about SHAP and random forest. We have also added explanation on the choice of 10 most important pollutants. Please refer to "A mixed modelling strategy for identification of cancer-specific key pollutants" in the Methods section. About the spatial correlation, spatial regression models like geographically weighted regression (GWR) assume that the effect parameters vary across different locations. Whereas this can enhance the precision of local estimates, it may hinder the overall interpretation and increase sensitivity to errors. Therefore, instead of GWR, we adopted negative binomial regression, a commonly used method in cancer research, to account for potential overdispersion in the cancer incidence data. Negative binomial regression assumes a consistent effect, aligning with the focus of our analysis. The results indicated small standard errors in the effect estimation. These findings suggest a minimal influence of inter-unit correlation on the effect estimation, thereby validating our results. Spatial correlation has been fully considered in the process of establishing pollution grade. We have added illustrative information about this consideration; please refer to the second paragraph in the sections "Spatial Evaluation System for Environment and Cancer" and "Distribution of co-pollution measured on a common graded scale".

REVIEWERS' COMMENTS

Reviewer #4 (Remarks to the Author):

The author's response to comments is again appreciated. While there are some data limitations regarding the timeframe of exposure data (2015-air and 2021-water) that have impact on the ability to draw associations between those exposures (classified as grade types) and cancer incidence, the additional comment on this limitation in the discussion is appreciated. As are further explanations on some of the statistical methods choices.

Reviewer #5 (Remarks to the Author):

The authors have made significant improvement on the manuscript since the last review. I appreciate the efforts and have no further comments.