This manuscript addresses an important and timely issue: mercury bioaccumulation models are often too simplified to reproduce observed responses to mercury emission reduction strategies and tend to predict overly linear outcomes compared to empirical observations. In my view, this represents a meaningful contribution to modelling efforts supporting the Minamata Convention. However, the current version lacks a mechanistic discussion explaining why the assembled models behave in this linear manner. The authors has done a great job compiling the existing model results, and with additional depth in the discussion, the paper could not only identify structural weaknesses in current models but also highlight lessons for improving future model development. Additionally, there are several unclear sections, especially in the methods, that could be further clarified. Therefore, I recommend acceptance with major revisions.

Specific comments

- Line 26: The sentence's mentioned Hg is identified by the WHO as a top 10 chemical, but doest specify what for. I understand that this is the top 10 chemicals of most concerns, but that should be clarified or removed. I would recommend removing this as the first sub-sentence makes the first sentence very heavy.
- Line 33: Pre 1450 is not pre antrhopgenic but preindustrial.
- Line 44 states once deposited but not were. This makes the sediment in the next sentence sound like sediment in general but based on the rest of the sentence it sounds like the author specifically refers to sediment in water. Hg would not mehtylate for example in Saharan sediment.
- Line 47: Biomagnification is a sub process of bioaccumulation, so it doesn't bioaccumulate and biomangifies. It bioaccumulates through bioconcentration (bioaccumulation through direct uptake) and biomagnification (bioaccumulation through trophic interactions)
- Line 123: you stated the analysis was performed using Python but is inisufficient to reproduce it. The package or method used should be specified and ideally both the versions of both the packages and Python that was used for the analysis. Like you do in line 130 for QGIS I would consider it much easier to reproduce.

- The formatting of Tables 1–4 does not work in the manuscript's current layout. When printed, the table headers appear on one page while the corresponding values appear on the next, making the tables difficult to read. For example, in Table 1 the column header "reference" is split across lines, and several header rows span up to seven lines. In addition, the citation text within the tables often occupies four to five lines because the table columns are too narrow. I recommend either moving these tables to the Supplement, reformatting them in landscape orientation, or reducing the amount of information included so that the tables become readable.

- Line 204: The conclusion that changes in deposition are smaller than changes in emissions is expected. A significant portion of deposition originates from re-emissions and legacy Hg pools, so reductions in fresh anthropogenic emissions do *not* translate 1:1 into deposition changes.
- Lines 234 - 266: This section includes an interesting discussion, but is too long and geographically unfocused (China, Southern Hemisphere) compared to the rest of the paper which focusses on the US. Recommend shortening and focusing on the core point: observed nonlinearity vs. modeled linearity.
- Line 304: Growth dilution does not eliminate MeHg, it only reduces concentration.
- Line 328: Interpretation of R^2 is incorrect. Since the axes represent *percent change in deposition* vs. *percent change in fish MeHg*, R^2 = 0.63 means 63% of the variance in the *change* of fish MeHg bioaccumulation is explained by the *change* in Hg deposition. Not 63% of the MeHg variance in fish. This distinction is important because each point in the plot represents a difference between two scenario simulations, meaning the underlying variance in fish MeHg itself is not present in the dataset.
- Section 3.2: Model a) and b) are marked with: while model d) and c) are marked with –. This should be uniform.
- Figure 3 and line 200: It should be stated explicitly how the percentage differences were calculated, including the exact equations used. More importantly, the manuscript must clearly clarify between which scenarios the differences were computed between. I initially assumed that each point represented a comparison within the same model under different emission scenarios, but this cannot explain the cluster of points where Hg emissions show no change while deposition increases or decreases. This suggests that the comparisons may instead be made against a baseline model or a model average and some of the differences in model output are caused by other drivers than Hg emissions, but I did not clearly understand this from the text. If so, the procedure and its implications should be clearly described, as the choice of reference scenario can strongly influence the resulting patterns.
- Figure 4: In the scenarios where Hg deposition is expected to increase, there are two points labeled *L-HTL b+3* that share the same colour but show very different changes in fish MeHg bioaccumulation. The manuscript should clarify why two identically labeled scenarios produce such different outcomes. In addition, the colours are described as representing deposition-reduction categories, but the x-axis already represents the change in deposition. This creates redundancy and confusion. Finally, several points on the far right of the plot share the same change in Hg deposition but are assigned to different reduction categories (e.g.,

"middle" vs. "high"). It is unclear how the same deposition change can belong to multiple deposition-reduction classes, and the criteria for assigning these categories should be explained.

- Figure 5: I believe this figure should be removed or substantially revised. Currently, the figure is confusing and did not become clear until reading the text *after* it. That text makes the important point that models often predict a limited, linear response while bioaccumulation in nature can follow very different patterns. The text used excellent references from relevant geographical locations, but in my opinion the figure itself does not help illustrate that conclusion. It was also initially confusing that models are shown by a bar, but observational data is 2 points and 1 bar with error bars. Only after reading the text and comparign the years did I understand what meant what, from the figure itself this was not clear, and after reading the text the figure did not improve my understanding.
- The title of Section 3.3, *"Lake characteristics vs MeHg in fish, 2050,"* could be improved for clarity and readability. I recommend rephrasing it to more clearly reflect the content, for example to "Influence of Lake Characteristics on Projected 2050 Fish MeHg Levels" .

Structural comments

- I recommend shortening the conclusion and adding a separate Discussion section, or a section purely focussed on lesson for future lake Hg modeling. While the manuscript correctly notes that the models used underrepresent natural complexity, it does not provide a mechanistic explanation of why this occurs or in-depth discussion of why this is, and what it means. For example, the Hendricks and SERAFM models simulate bioaccumulation using BAF-based relationships, which are inherently linear; therefore, unless I misunderstand, any non-linearity between Hg deposition and fish MeHg must arise from Hg speciation processes rather than from the bioaccumulation step itself as the BAF is a linear relationship between marine MeHg and fish MeHg, as per definition. In models you should always make sure not to rediscover your own paramaterization, and observign linear behaviour in a BAF based models should be unsurpring. More broadly, all ecosystem models included in this analysis are box models, which cannot represent several hydrodynamic processes known to influence both MeHg bioaccumulation and ecosystems interactions.
  I think this raises 2 key questions that should be discussed. The first is the question if the MeHg models are actually wrong. If a lake model demonstrates that a 2% reduction in deposition in the lake would result in a 2% reduction in fish MeHg and we assume for a moment the model is completely correct. Then a

change in trophic structure in the lake can still cause the observed difference in fish MeHg to be different than predicted. But this does not mean the MeHg bioaccumulatoin model is per se wrong. It means that MeHg bioaccumulation in lakes can be improved using coupled mechanistic models, rather than box models.

A key question that follows is how these models can be improved. The dataset already contains useful information that could guide such a discussion. For instance, the PCA shows that water depth is negatively correlated with the percent change in fish MeHg, while watershed area is positively correlated. This is intuitive: the ratio of atmospheric deposition area to total water volume likely influences a lake's sensitivity to changes in deposition. Exploring these relationships more fully, with a direct focus on how to contribute to the observed non linearirty and what this means for models would add depth to the analysis and help translate the findings into concrete recommendations for improving future mercury bioaccumulation models. This is somewhat discussed in section 3.3, but here the focus is mostly on statistical analysis of what drives MeHg concentrations in fish in lakes, and I think a clearer direct coupling to what lessons can be learned from the collected data can greatly improve this manuscript.

I understand this may not be possible and I would support the publishing without this. But it is noteworthy that the link between Hg emission and deposition in most models is based on complex 3D atmospheric models that are in general well understood and validated. Whereas the lake component of this analysis links Hg deposition to MeHg in fish using much simpler models, even though this pathway is poorly understood and extremely complicated. If the data is available splitting the lake component into Hg deposition to marine MeHg concentrations and marine MeHg concentrations to MeHg in fish the study could help to identify which components are nonlinear in nature and linear in the models, and thus what components of the models need to be improved.

- A final point regards over-citation throughout the manuscript. It is ofcourse good to cite previous litarature, but I strongly recommend making a clear statement and supporting it with one or two relevant references, ensuring that it is evident which reference supports which claim. The most notable example is the citation block around line 73, which spans four lines and includes roughly 20 papers, yet the rationale for including each citation and their specific contributions is unclear. This block refers broadly to previous modelling work, but the selection appears inconsistent: for instance, studies such as Zhang et al. (2020) are included even though they focus on marine 3-D modelling, while other relevant marine studies (e.g., Rosati et al. 2023; Bieser et al. 2023) are omitted. The

problem is that if a reader wants to verify or research further in a citation block of 20 citations it is hard to know what the most relevant citations are.
I suggest rewriting this section to clearly describe the types of models that exist, what they capture, and what is missing, while keeping citations closely tied to the specific statements they support. I would suggest this section should focus on lake and freshwater models, with marine models potentially introduced in the discussion, where they can illustrate processes or mechanistic complexity missing from current lake models. Similar over-citation occurs in other parts of the manuscript (often blocks of 5–10 references), which reduces readability and makes it difficult to evaluate the specific statements.

Overall, I enjoyed reading the manuscript and I think the core argument, that MeHg bioaccumulation models overpredict linearity compared to nature is an important message. However, I do believe the manuscript would be greatly improved if it translated these observations into a meaningful discussion about the quality of the models and what should be improved upon.