Dear Dr. Middelburg,

Thank you for securing two helpful reviews of our work. Below we summarize the revisions we made in response to the reviewer suggestions. Reviewer suggestions are in normal text and our responses are in bold text to easily differentiate. We feel the reviews were helpful in improving the manuscript. We look forward to your further evaluation.

Thank you,
James Stegen on behalf of all co-authors

######

Reviewer 1:

In this manuscript Kew et al. discuss the use of intensity values for natural organic matter ions measured by ultrahigh resolution mass spectrometry. The authors review basic concepts, provide a summary of limitations, pitfalls, and then describe how the use/misuse of intensities could impact the employment of FTMS to obtain ecological metrics. The paper is concluded with a wonderful summary and recommendations to the community. The paper is not of a traditional format and is a hybrid between a review article, a critique, has some experimental data, as well as some computational modeling. It also has a lot of important citations provided to us, which is also very valuable. As a heavy FTMS user I enjoyed reading through and will be certainly coming back to in when I need to be reminded of how intensities can change if I were to change X, Y, or Z in my experiment.

Thank you for the encouraging remarks, we're glad to hear you see value in this work, especially as a heavy user of FTMS.

I really like figures 1,2, and 3. I agree with the authors, as they say in the response to comments doc, that most environmental users of FTMS do not have much formal education in mass spectrometry and they end up learning MS and other analytical methods "on the go". So while some of this theoretical background and experiments may be redundant with other papers/textbooks, often such resources are abundant with complex diagrams, lots of calculus, etc. - texts that are not super digestible by the broader biogeoscientist. This study is a perfect resource for our community and especially the uprising young scientists in non-chemistry departments (e.g., geology). The authors also nicely introduce the ecological metrics to non-ecologist readers (as a classically trained chemist I have no formal education in ecology, so it was nice to read a broader introduction of alpha/beta diversity indices and their uses).

Thank you for the further encouraging remarks, we're glad you find value in the introductory elements.

Given that this is a second round of review, this is already an excellent paper. I have one major comment and a few minor ones that can be easily remedied with a minor revision and overall aim to make the paper more receptable by the broader biogeoscientific community.

My biggest critique is that there is too much emphasis on the ecological indices. The PNNL group is the main group using these metrics with very few other groups using them once in a while - if we look at the global biogeosciences community, most people do not use mass spec data to calculate these diversity indices and do not use the data in an ecological framework/manner. People commonly plot van Krevelens, DBE vs C, calculate %CHO, %CHON, %CHOS, etc. types of formulas (%lignin, %lipid, etc.), other averages (average H/C, O/C, NOSC, AImod, etc.). Statistics like HCA, PCoA, PCA, NMDS are also very common, and so are now Spearman correlations with some other metrics (PARAFAC Fmax, CDOM metrics, etc.). I am not saying this is the right thing to do, I do think people should be doing more advanced things with MS data (ecological frameworking, neural networks, etc.), but this is where we are at and have to consider the state of the community right now. As this paper is targeting the broader community, it should contain as much relevant information as possible.

This comment helped us realize we needed to be more direct in the text that we include sample-level intensity-weighted trait/property/characteristic as one of our 'ecological metrics.' This is directly related to commonly used intensity-weighted averages (e.g., of H/C, etc.) that the reviewer points to above. The revised manuscript is more direct about including that as one of our three studied ecological metrics. In addition, we study Bray-Curtis dissimilarity, which is commonly used as input to NMDS and discuss the utility of Spearman rank-based correlation analyses. The revised manuscript is more direct about the relevance of the associated simulation-based analyses and outcomes to real-world FTMS NOM studies.

I like the ecological framing, but I do not think it would be so useful to the broader biogeoscientific community, because most of us do not use this type of ecological framing for our experiments at present and probably in the foreseeable future. For this reason, I think sections 4 and 5 are too long and should be combined into one section. The ecological section should come after a main section where the more common uses are discussed. All of the things I describe above involve the use of intensity values and the authors should discuss the relevant issues.

As noted in our response above, the ecological framing goes beyond the diversity metrics and is quite relevant to a broad range of common analyses. We elaborate further on this in our next response, below. The reviewer also suggests shortening sections 4 and 5. We significantly shortened section 5, which describes the simulation model and associated results. We moved more than half the text to the supplementary materials and also moved two multi-panel figures to the supplementary materials. The text remaining in section 5 is written to be more accessible to a broader readership, and we leave the technical details to the supplementary material. We greatly prefer to keep section 4 as is because it provides a conceptual summary of how to best think about FTMS data when

doing NOM studies through space and/or time. It is written to be accessible to a broad audience and is an important part of the storyline; we ask for support to retain it.

Some examples I can easily come up with: Should we use intensity-weighed H/C, O/C, etc. metrics (sum of rel.intensity*metric), or just regularly calculated averages? Would it be better to report %CHO, CHON, etc. as % intensity (intensity of CHO formulas/all formulas) or % number of formulas (number of CHO formulas/number of total formulas)? When people look at a three-dimensional van Krevelen plot, and see that the most intense peaks are in the lipids region (for example), how confidently can we say "the sample is rich in lipids/most molecules are lipid-like". When people Spearman-correlate FTMS formula intensities with external metrics (e.g., proteinaceous component from PARAFAC), how much can we trust that the signifficantly correlating formulas correspond to proteins? Regarding statistics, should we combine FTMS results (%CHO, %CHON, etc., which are semi-quantitative) with truly quantitative data (DOC, SUVA, etc.) - everyone does it, even though all variables have to be of equal quantitativeness. Or should we not use FTMS intensities at all in such stats and just resort to presence-absense (see first paragraph of section 2.6.5 here: 10.1016/j.gca.2010.03.035). I can go on and on with coming up with similar questions, and I am sure that the authors can too - I don't necessaryly ask you to answer these exact questions in the revised version, but giving you ideas for talking points. I think this manuscript is the perfect place for addressing such questions and having this kind of discussion - this would be much more beneficial to the community.

These are all great questions and our work addresses many of them. We included more discussion on these examples as part of our recommendations, within section 6 (the last part of the paper). We kept this after the simulation model because the simulation model results are key to generating our recommendations. As noted above, we cut the simulation model section (section 5) by more than half to help with readability and broad accessibility.

After such section of describing the use of intensities in "common FTMS data workup approaches" you can follow up with the "using/misuing FTMS intensities in an ecological framerworks" section. I will also say that sections 4 and 5 are collectively too long, not so straightforward to read by someone who does not use ecological indicies on a regular basis (in my opinion, most of FTMS users), and so I recomment shorteing, not going that much into the weeds, and slightly streamlining with the thought of making that more friendly to the broader public.

See above; in short, we took the advice of shortening section 5 significantly.

A minor comment was that I was confused by the "within-peak" and "between-peak" terminologies. Regarding within-peak, I first thought of doing comparisons of the isomers that are under one m/z value. The between-peak sounds ackward and I though we would be comparing the noise level between two peaks.... I recommend changing these terms to something clearer. Suggestions: "same-peak comparisons" for within-peak and "peak-to-peak comparisons"/"different-peaks comparisons" for between-peak.

The language used here is difficult to navigate and we feel that no choice is clearly better than all others. For example, if we use 'different-peaks' instead of 'between-peaks' we'll end up with some sentences reading '...different-peak differences...', which feels potentially confusing. We're also careful to define our use of within-peak and between-peak at the start of Section 2:

"Here, we define 'within-peak' as comparing peak intensities of the same feature (i.e., m/z or molecular formula) across different sample spectra and 'between-peak' as comparing peak intensities across different features."

To help with any confusion regarding isomers, we added the following sentence to the same paragraph: "Both within-peak and between-peak comparisons are fundamentally based on the m/z observed within a mass spectrum and neither address comparisons across isomers."

Figure 2 also provides a visual definition of within-peak and between-peak. Given all these considerations, we propose retaining these terms.

This manuscript desperately needs to include some text (maybe one paragraph?) about normalization. Usually we take the raw intensity values of assigned peaks and divide the by the sum of all intensity values of assigned peaks, but is this the best way? Should we consider the sum of ALL spectral peaks (including isotopologues, blank peaks, etc.) or just normalize to the sum of intensities of the assigned formulas? Some ideas for talking points. I am familiar with this wonderful paper that provides some guidance: 10.1002/rcm.9068

We appreciate this suggestion and in response we added a short paragraph within a new sub-section (3.3). We believe that post-hoc normalization strategies (such as detailed in the referenced paper) are helpful for some applications, but cannot mitigate the underlying physical processes that cause peak intensities to be weakly related to true abundance. We feel that deeper discussion on normalization is beyond scope of this manuscript primarily because normalization isn't a clear solution to the challenges we raise. Regardless, we agree it's helpful to call this out directly in the manuscript, as done in the new text (provided immediately below).

The following text was added:

"3.3 Data Normalization Strategies

In the previous section, we use the peak intensities for each analyte without any normalization, only scaling to the base peak or between spectra to make comparison easier. However, more sophisticated or comprehensive normalization strategies may be useful when trying to make quantitative inferences of the data. Considerations may include whether to use the total intensity within a spectrum (including noise, isotopologues, and unannotated features), or to use just the peak intensity apportioned to annotated features. Additionally, non-linear or more sophisticated functions may have benefits. Such post-hoc statistical approaches have utility for some applications but do not resolve the fundamental, underlying physical origins of the weak connection between peak intensities and true concentrations. We refer readers to the work of

Thompson et al. (2021) for more insights into the theory and application of normalization of FTMS for complex mixtures..”

Lastly, there are a few different terms that we come across: intensity, magnitude, peak height. They are in my mind the same, but are there any nomenclature issues that need to be discussed in this regard? Should we stick to one uniform term for more comparable literature or keep using all of them interchangeably? Are there different connotations/flavors to these terms? Some ideas for talking points if you choose to include a terminology discussion (I do recommend! This paper would be the perfect spot for it).

We agree that consistent terminology is helpful within and across publications. We did a search across our manuscript to ensure we are using ‘intensity’ and not other related terms. We made a couple small adjustments to keep consistent terminology. We note that ‘magnitude’ may be ambiguous given different signal processing types (e.g. absorption or magnitude mode FT processing). ‘Peak height’ may be acceptable, and indeed can be helpful to avoid ambiguity if intensity is an area or height derived metric. However, intensity is commonly used, colloquially, and is the output column name from Vendor provided FTICR peak picking tools (e.g. Bruker DataAnalysis indicates column ‘I’ for peak intensity). Of course, we recommend that researchers define the terminology used in their papers where ambiguity or uncertainty may exist.

We added the following text to the start of Section 2 to clarify our suggested approach:

“Further, we suggest consistent use of the term ‘intensity’ in FTMS NOM studies to describe how much signal is observed for a given peak, as opposed to ‘height’, ‘magnitude’, or other alternatives. While terminology is not our central focus, it is useful to pursue consistency across studies.”

#######

Reviewer 2:

General comments:

This study mainly showed that the peak intensity of FTICRMS is not a proxy for abundance and concentration. However, this is somehow common sense and should be already known to most researchers. The experiments performed by the authors nicely proved that peak intensity should be used with caution, especially for quantification. Again it is already well known.

We agree that such information should be well known and basic knowledge to experienced mass spectrometrists, however many users of FTMS data are not formally trained in mass spectrometry. This manuscript serves as an accessible education on these pitfalls. Further, there are many papers that use peak intensities in ways that

implicitly assume the peak intensity is a proxy for abundance. More importantly, a key message from our paper is that despite the fact that peak intensities are not a direct proxy of abundance, the peak intensity data carries enough valid information to be useful. This is an important point that is not obvious and is not, to our knowledge, in the literature. It is also of practical value to the entire research community using FTMS to study NOM.

The author claimed that there are plenty of misuse studies of peak intensity, if this is true, I would suggest the author organize a table of literature to illustrate how serious this issue is.

Within the manuscript, we do not make this claim. Within a previous response to a reviewer, we remarked that we do observe some manuscripts now using peak intensities wherein about a decade ago such use was less common. It is not our goal to highlight specific papers and critique them, but to raise awareness of how to most robustly use peak intensities.

The simulation model is great but I highly doubt its applicability considering so many speculation factors were introduced, including random error, etc. Besides, this model was to evaluate the matrix effect on peak intensity, and seems that it can only be used as an educational tool for people to understand the bias of FTRCRMS in the quantitative study. How people can use it for their own samples?

As with all useful models, simplifications are necessary. The simplifications do not undermine the value of the model as it is, and they provide guidance for future development of the model (e.g., inclusion of non-random error).

In terms of how people can use it for their own samples, we provide guidance on this within Section 6. While we expect more sophisticated versions of the model will be developed in future work, the starting point of using it for real sample sets is setting the simulations to use the same number of peaks observed in each real sample or each pair of real samples. The sample-level peak intensity distribution could also be constrained to be similar between the simulation model and real samples.

For the model part, I suggest the author simplify the sentences as it is hard for me to understand them and, if necessary, list the mathematical equations used for calculation. Add text to clearly and concisely describe what this simulation model is, how it can be applied to environmental samples, and how people can make use of it. This is very important because the model is the only highlight I can see from the manuscript even though it is very hard for me to understand all. I suggest a major revision for this manuscript and the author should pay attention to the concise of writing since the authors aim at people who are new in this field.

We cut the simulation model section (section 5) by more than half to help with readability and broad accessibility. As part of revising section 5, we replaced most of the simulation

model description with a shorter and more broadly accessible summary of what the model does.

For specific comments see below please:

The introduction part mainly described the diversity studies (yes or no question), then all of a sudden in the last paragraph, raised the concern of “quantification” (how much question), What is your point here? Could the diversity part be removed and more focused on how people used the peak intensity data incorrectly to support the concern of the authors?

We edited the introduction to include more emphasis on the mean trait (or property) values. Those were in the previous version, but not emphasized. Our intention across the introduction to present the ecological metrics as continuous variables, and did not intend to present them as answering a yes or no question. We carefully read over the introduction to ensure the text does not convey a binary yes/no perspective.

The study lacks a “Method” part, including information on compounds, standards, instruments, etc. Please provide. Sections 3.1 and 3.2 should be improved, use a scatter plot for Fig.4 instead so that you can simplify the text and make it easy for readers.

The Methods are detailed in the Supplementary Information file, and include information on chemicals, sample preparation, mass spectrometry measurements, mass spectrometry data analysis and the simulation model. We have opted to keep this bulk of text within the SI as it would unnecessarily lengthen the main text of the manuscript. We added a sentence to Section 3 to indicate the location of the Methods text.
“The experimental methods used are described in detail in the Supplementary Information.”

It seems that the homogeneity of the sample sets is not taken into consideration. Real-world within peak comparison might be more complicated if the studied set contains very heterogeneous samples.

The complexity of our empirical data increases from pure compounds in clean solvent (an ideal case), through to mixtures of chemicals in complex matrices (e.g. with organic matter added or artificial river water SPE eluent). While our sample set does not approach the heterogeneity of a real-world sample set, we show that even in the simple, idealized cases, peak intensities are not directly quantitative. Thus, in real-world studies with increased heterogeneity, this issue is only exacerbated.

We added a sentence to the penultimate paragraph of section 3 to this effect;
“A ‘real-world’ sample set would have even greater diversity and heterogeneity than presented here, and thus the issues with use of peak intensities for quantitative interpretation would only be exacerbated.”

L75: There was only one paper cited, but later sentence uses “these studies”, add more papers please.

Two additional citations have been added.

L113: “Using consistent sample concentrations” is impossible

We respectfully disagree here - in a typical NOM workflow it is readily possible to measure the TOC of the extract or to measure the mass of the dried extract. Several groups normalize the concentration (TOC) of their extracts prior to SPE, or prior to mass spectrometry measurement. Of course, frequently this is also not done, and variability in concentrations does exist.

L145-L146: Add text to describe what samples, what compounds, and what chemical standards.

We have clarified that the details are in the supplementary information. The specific chemicals are also detailed in the context of the Figure.

L268-L269, list the publications that misuse peak intensities

We do not think it is necessary or helpful to explicitly call out papers for misusing peak intensities - that is not our intention or inference. Our manuscript describes why peak intensity usage should be done carefully and the caveats and considerations which must be taken into account.

L369: add literatures

The text was edited and references were added. It now reads: “There is significant value in using FTMS data to study NOM chemistry (Bahureksa et al., 2021; Cooper et al., 2022; Spencer et al., 2015; Stubbins et al., 2010), and it is vital that this be done based on rigorous use of the data.”

L275-L276, why do you choose 100 and 1000 peaks? Where are these peaks from? From what kind of samples?

These peak numbers represent a very sparse sample (100 peaks), and a more typical order of magnitude (1000) for FTMS NOM data acquired over the past decade. Newer and higher performing instruments can increase the number of detected features even more (10,000 or above), however our models did not significantly change when we evaluated more (10,000) peaks. We note that the law of large numbers, which may be a key factor in the model's observations, would suggest that results will only more closely reflect true properties as more peaks are added. Thus, the use of 100 and 1000 reflects more conservative scenarios.

Figures: Be consistent, either use figure and Fig, do not mix the usage

We edited text to make the usage consistent. We use (Fig. X) in parenthesis to indicate the referred to object, and 'figure(s)' in a sentence when the subject of the sentence is the figure.

Fig.4 Please change it to a scatter plot and give a correlation value (R2), this can help people understand the relationship between intensity and concentration.

We replaced the figure with a scatter plot. The Pearson r (not R2) correlation coefficients and p-values (calculated by the Python scipy stats module) have been calculated for each dataset and are included in a new supplementary table (Table S1). Figure 4 caption has been updated to reflect this.

In fact, in lines 162-163, the author also suggested a calibration curve, why not do it for this study?

As discussed in the text, comparisons of the same peak between samples are problematic because the same m/z (and same molecular formula) may be different isomers (or different relative amounts of the same isomers) between samples. As demonstrated in the empirical section, isomeric compounds can have starkly different ionization efficiency. Thus, a calibration curve would not resolve the case of NOM studies in which isomeric composition is unknown.

Our intention is to say: A calibration curve *could* be used in the case of a sample where you do know the specific chemical identity and can purchase a standard, but such efforts would require additional upfront separation (e.g., online liquid chromatography or ion mobility) to ensure the chemical you are calibrating is the analyte of interest. Such workflows would be extremely laborious, targeted, and sample-set specific.

What is relative intensity?

Relative intensity is a mass spectrometry term defined as the ratio of a signal of interest to the base peak (most intense peak in the spectrum). In Figure 4, we use a slightly modified definition to account for scaling across spectra, and this is detailed in the caption to Figure 4.

Fig. 8: How do you get the observed difference? How is the true difference calculated?

This is all done within the simulation model, and the text in Section 5 was edited to help make this clearer. The model generates a sample *in silico*, which is used as the true sample with zero error. The model then adds specific kinds of errors (as detailed in the manuscript) meant to represent the kind of error that occurs when running a real sample.

In this case we have both truth (i.e., no error) and observed (i.e., after error is introduced to the truth).

The R2 in Figure 8C is questionable: although the value appears to be high, it's obvious that the independent variables are heteroscedastic.

This figure and its companion (based on a sensitivity analysis) are now both in the Supplementary Material and are Figures S6 and S7. To highlight the issue with heteroscedasticity, we added the following text to the Supplementary Methods (Section 2.6): "However, we suggest caution when interpreting the R2 values associated with Figure S6C and S7C as the differences collapse when near zero, leading to heteroscedastic residuals that likely bias the R2." We also added the following text to the captions of Figures S6 and S7: "On panel C the R2 value should be interpreted with caution as the residuals are clearly heteroscedastic."