Dear Dr. Singer,

Please find enclosed the revised version of manuscript “Methane, carbon dioxide, and nitrous oxide emissions from two clear-water and two turbid-water urban ponds in Brussels (Belgium)”

As recommended by the reviewer we analyzed the data using GLMM to include random effects when appropriate. For some tests, the data-sets were too small to use GLMM (the model did not converge) so we used other methods (Pearson or ANOVA). Please note that we found the same patterns in the data analysis as before, so the new statistical analysis did not change the overall conclusions of the paper.

As recommended, we have also separated the Results and Discussion sections.

As recommended, we simplified and shortened the text, and we have reduced the number of figures.

We have narrowed down the scope of the paper to the GHG emissions and we removed the analysis on methane oxidation and on methanogenesis pathways.

We are grateful for the editorial guidance and the reviewers’ suggestions and we sincerely hope that the present version of manuscript meets the requirements for acceptance in Biogeosciences.

Best regards,

Thomas Bauduin, Nathalie Gypens, Alberto Borges

# Associate Editor

I am sorry for the long time it took to process another round of reviews for your manuscript. Long manuscripts may also need a long time to review.

Reply: We are very grateful for the second round of reviews and your annotated manuscript.

I have received very mixed feedback from the two reviewers who have seen your manuscript also in its first version. Reviewer #1 considers that his/her concerns were well addressed and recommends acceptance of the manuscript upon having taken care of a few minor points. Reviewer #2 acknowledges the work you have put into addressing the raised concerns and into improving the manuscript, in particular with regard to improved flow in the abstract and an improved introduction that now includes hypotheses. This reviewer also still recognizes the value of your dataset on pond GHG dynamics. However, the final verdict of this reviewer was to reject the manuscript because of two reasons: (i) still existing serious deficits with regard to statistical analysis, and (ii) general readability difficulties and complexity of the study presentation given the very long combined results and discussion section. I gave your revised manuscript a careful read once more and have to agree with reviewer #2 regarding the statistical issues, yet I believe these issues can be fixed. I also still believe - as pointed out earlier - that your manuscript is really long and its presentation overly complex, thereby risking only limited attractivity to readers. I don´t think that this issue is critical enough to justify rejection of the paper, however. Once more, I side with reviewer #2´s recommendation to separate results and discussion to create a manuscript that leads the reader through the many results and large number of response variables in a less confusing way.

Reply: We have carefully considered the editorial recommendations. In order to accommodate the request to reduce the size (text length and number of figures) and to re-structure the manuscript into separate Results and Discussion sections, we have removed the data MOX and methanogenesis pathways based on the stable isotope data. The resulting manuscript is solely focused on the GHG emissions, more condensed, reduced in the number of figures, and structured in separate results and discussion. We have also reviewed the statistics (see below and in the response to reviewer 2).

A few more words on statistical issues:

1) Besides giving clear recommendations of how to take care of the temporal nature of your dataset, Reviewer #2 considers your use of the Pearson correlation coefficient inadequate given ignorance of the repeated measures nature of your data. The clear concern is that the current approach inflates significance of some explanatory variables, which is not best practice for temporally repeated measurements (even if you find significance). Please also respect the difference between correlation and regression. For example, in line 370 you argue for slopes of correlations to be significantly different, which is meaningless while also it remains unclear how this was tested. Another example is Fig. 4, where a functional relationship between %CH4 as a driver of bubble flux is implied - I see no justification for a regression in this case, maybe it would be better to plot temperature against %CH4.

Reply: We have used GLMM when the data-set was large enough. For some tests, the data-sets were too small to use GLMM (the model did not converge) so we used other methods (Pearson or ANOVA). Please note that we used ANOVA for repeated measures, so the test does incorporate to some extent the nature of temporally repeated measurements.

We have changed the figure of %CH_4 as suggested but we kept as a supplemental the relation between %CH_4 and bubble flux. We agree that the %CH_4 is not a driver of the bubble flux but there is a justification to plot the data because it shows that %CH_4 increases with bubble flux. So, the increase of CH_4 ebullition with temperature results from an increase in both bubble flux and %CH_4. It is conceivable that the %CH_4 of bubbles would remain unchanged and that the ebullition of CH_4 was solely a function of bubble flux. The data show that ebullition of CH_4 increases with temperature as a result of an increase of both bubble flux and CH_4 content.

2) The hypotheses brought forward in lines 117-125 argue for effects of alternative stable states on GHG concentrations, emissions, partitioning into fractions of various GHGs. There is no hypothesis about seasons or any other effect. However, in the statistics section (line 272-279), we can find nothing about how the factor "alternative state" is tested, yet we read "four seasons were serving as independent factor". Also, pond ID seems to have been used as a fixed AND as a random factor, which is confusing and hardly correct. This serious disagreement between hypotheses and methods is then also evident in the R&D section, where at multiple places "significant" effects are stated, yet it remains unclear to which test you refer (examples in lines 370 (already mentioned above) or 506-510).

Statistical analyses were revised:

- We tried to compare the ponds according to ‘alternative state’; however, the data-set is only made up of 2 turbid-water ponds and 2 clear-water ponds and this number was insufficient to perform this comparison with GLMM using “alternative state” as fixed variable, and “pond” and “data” as random effects, as you predicted in the final paragraph of your decision letter (below). We therefore compared the 4 ponds with each other (pairwise) and then discussed the differences, considering that 2 were turbid-water and 2 were clear-water, as indicated by Chl-*a* values and macrophyte presence and abundance.

- We compare patterns of GHG versus drivers in the whole data-set (merging the 4 ponds) using GLMM (with sampling date and “pond” as a random factor to take into account repeated measurements over time).
- We investigate seasonal differences by comparing the ponds with each other in each season using repeated measures ANOVA (as GLMM did not converge due to insufficient data).

We have revised the text lines 117-125 (new lines 107-110 in the manuscript without track changes): “We test whether the differences between the four ponds are explained by the two alternative states in terms of (i) CO_2, CH_4, and N_2O dissolved concentration and diffusive emissions; (ii) ebullitive CH_4 emissions; (iii) relative contribution of CO_2, CH_4, and N_2O to the total GHG emissions in CO_2-eq.“). The comparisons described in line 370 have been removed and the comparisons in lines 506-510 (new lines 343-350) are now described in the materials and methods and follow the same methodology as the other comparisons described in the text. We have also revised the entire manuscript to ensure that all the comparisons described are explained in the material and methods, and that the results are fully included in the supplementary materials.

When reading your revised version I also stumbled upon another, likely minor question: You describe usage of an IRGA for a 30 ml headspace sample. I have worked with the Li-840 IRGA myself and wonder whether such a limited volume of gas actually creates a stable plateau reading.

Reply: We compared the measurements of pCO_2 with a Li-840 using a 30 ml headspace from discrete measurements with simultaneous pCO_2 measurements with a flow-through equilibrator system for a wide range of pCO_2 values in diverse environments (rivers and lakes). The agreement is excellent. Please refer to Figure 2 in Abril et al. (https://doi.org/10.5194/bg-12-67-2015) and Figure S22 in Borges et al. (https://doi.org/10.1126/sciadv.abi8716).

I have a few minor comments which I will make available to you as an annotated pdf. In some cases, your text is pretty hard to understand and rewriting suggested.

Reply: We have implemented all of your suggestions on the annotated pdf. Please note that there are no data points for the Netherlands and Québec plot because these are published relationships and are only shown data points for our own measurements. We prefer to keep the units of fluxes in mmol m^{-2} d^{-1} throughout the manuscript (even for the annual averages). This is to avoid mixing different units in the text and figures (which could be confusing for readers). Also the conversion of units from d^{-1} to yr^{-1} is straightforward.

Given the (mixed) opinions by the two reviewers and my own evaluation I am asking you once more to revise your manuscript carefully. I consider this a request for major revision and plan to involve at least one additional reviewer as the reviewers who dealt with your manuscript so far are not willing to provide their expertise once more. I believe that both this fact as well as the long duration of the review process tells you about the challenges an overly long and hard-to-read manuscript creates for your readers. Please carefully revise and clarify usage of statistics. In this respect I see no need to "overdo" it, e.g. **testing an effect of "alternative stable state" may just be deemed impossible given the low sample size of 2 ponds for each factor level**, yet data may still be discussed in this light. However, clearly, your manuscript must be technically correct from a statistical standpoint, and if you bring forward hypotheses to be tested, then this must be reflected in appropriately chosen statistical methods and adequate results. I really urge you to consider shortening your manuscript wherever possible, even if no page limit is enforced by the journal. Splitting the R&D section into separate results and discussion sections is indeed a good advice that will help to make a long manuscript more digestible - you may argue otherwise.

Reply: We have implemented the suggestions of Reviewer#2 on the statistical analysis and have reduced the length of the manuscript and structured it in separate results and discussion.

#Reviewer 1

1) Scientific significance

Does the manuscript represent a substantial contribution to scientific progress within the scope of this journal (substantial new concepts, ideas, methods, or data)? Excellent

2) Scientific quality

Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)? Excellent

3) Presentation quality

Are the scientific results and conclusions presented in a clear, concise, and well structured way (number and quality of figures/tables, appropriate use of English language)? Excellent

For final publication, the manuscript should be accepted subject to minor revisions

Were a revised manuscript to be sent for another round of reviews: I would not be willing to review the revised manuscript.

Suggestions for revision or reasons for rejection (visible to the public if the article is accepted and published)

The authors have done an excellent job addressing the significant concerns I raised in my previous report. After addressing the few minor comments below, I recommend publication.

Line 21: Phytoplankton are also made up of OM. The two OM sources may be distinguished based on quality.

Reply: The reviewer is right. But we did observe higher ebullition in clear-water lakes than turbid-water lakes that should be logically related to macrophyte biomass.

Line 39: Higher or lower or comparable?

Reply: text now reads at lines 35-36 in the manuscript without track changes: “However, reported emissions of CH_4 from lakes (Rosentreter et al., 2021; Johnson et al., 2022) are equivalent or even higher compared to rivers (Stanley et al., 2016; Rocher-Ros et al., 2023)”

Line 763. States, “Such a hypothesis was consistent with an overall positive relation between %N2O and DIN in the urban ponds of the city of Brussels irrespective of presence or absence of macrophytes (Bauduin et al., 2024; this study).”However Line 380 states “More surprisingly, %N2O was not correlated with DIN (Table S3; Fig S3, S4, S5, S6) nor with individual forms of DIN (NH4+, NO2-, NO3-) in the four ponds individually or when all the data were pooled together for the individual forms of DIN (Table S3; Fig S7)”. It is true to only state that the N2O- DIN positive relationship in urban ponds was found in the previous study alone that tended to be more spatially based rather than including the “this study” citation as it is now.

Reply: Text now reads lines 283-299: “The %N_2O values ranged from 32 to 826% (Fig. 3). Undersaturation of N_2O with respect to atmospheric equilibrium was observed 66 times out of the 187 measurements. Low values of %N_2O were generally observed in spring and summer and high values of %N_2O were generally observed in fall and winter in the four ponds (Fig. 3). During spring, the %N_2O was lower in the Pêcheries pond (90±11%) than the Leybeek (138±30%, p=0.0043, Table S3) and the Tenreuken (138±41, p=0.0057, Table S3) ponds. During summer, the %N_2O was lower in the Pêcheries pond (78±17%) than the Leybeek (191±104%, p<0.0001, Table S3) and the Silex (126±49%, p=0.001, Table S3) pond, and lower in the Tenreuken pond (133±106%) than the Leybeek pond (p=0.0219, Table S3). During fall, %N_2O was lower in the Pêcheries pond (103±33%) than the Leybeek pond (190±70%, p=0.0174, Table S3). For the all sampling period, %N_2O was lower in the Pêcheries pond (94±28%) than the Leybeek (178±82 %, p<0.0001, Table S7), Tenreuken (140±77%, p<0.0001, Table S7) and Silex (144±113%, p<0.0001, Table S7) ponds, and was lower in the Tenreuken pond than the Leybeek pond (p=0.0038, Table S7). When data were pooled together, %N_2O was correlated negatively with water temperature and positively with DIN and NH_4^+ (Table S4). In individual ponds, %N_2O was negatively correlated with water temperature in the Leybeek, Pêcheries, and Tenreuken ponds (Table S5). %N_2O was positively correlated with NO_3^- in the Leybeek pond and with NH_4^+ in the Pêcheries and Tenreuken ponds (Table S8). %N_2O was positively correlated with Chl-*a* and TSM in the Tenreuken pond, and negatively with Chl-*a* in the Leybeek pond (Table S5), probably reflecting the negative correlation of Chl-*a* and TSM with water temperature in the Tenreuken pond and the positive correlation of Chl-*a* with water temperature in the Leybeek pond (Table S6).”

#Reviewer 2

1) Scientific significance

Does the manuscript represent a substantial contribution to scientific progress within the scope of this journal (substantial new concepts, ideas, methods, or data)? Good

2) Scientific quality

Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)? Fair

3) Presentation quality

Are the scientific results and conclusions presented in a clear, concise, and well structured way (number and quality of figures/tables, appropriate use of English language)? Poor

For final publication, the manuscript should be rejected

Were a revised manuscript to be sent for another round of reviews: I would not be willing to review the revised manuscript.

Suggestions for revision or reasons for rejection (visible to the public if the article is accepted and published)

Line comments:

Line 34: I suggest closing the abstract with a statement on how this work contributes to knowledge on carbon cycling from freshwater/ponds.

Reply: In our opinion, the change of annual emissions with precipitation in response to El Niño is the most important result of our work and we prefer to close the abstract with this. We have added new lines 19-22 of the abstract in the manuscript without track changes the following sentence that shows how this work might contribute to knowledge on carbon cycling from freshwater ponds: “These findings imply that it might be necessary to account for the presence of submerged macrophytes when scaling ebullitive CH_4 fluxes in ponds at larger scale (regional or global) (particularly if Chl-*a* is used as a descriptor), although possibly less critical for diffusive CH_4, CO_2, and N_2O fluxes.” and in new lines 23-25: “The temperature sensitivity of ebullitive CH_4 fluxes decreased with increasing water depth, implying that shallow sediments would respond more strongly to warming (*e.g.* heat waves)."

Line 61: replace “dynamic” with “dynamics”.

Reply: Done

Line 61: Can authors cite this? I don’t doubt that this is the case for some ecosystems but I’m not confident that this statement can be easily generalized across all freshwaters/ponds.

Reply: We do not see a problem here. Photosynthesis does affect CO_2 content in water and will be balanced by respiration (e.g. Sand-Jensen and Staehr 2007). We did not state that photosynthesis was the only process that controlled CO_2 in ponds or that phytoplankton/macrophytes occur in all freshwaters/ponds. However, in the context of the paper (urban ponds in a European city) it is unlikely that ponds are devoid of both phytoplankton and macrophytes. It is equally unlikely that photosynthesis from phytoplankton/macrophytes does not affect at all CO_2 dynamics of the studied ponds.

Line 91: I think it’s great that authors have added context about ebullition and methane oxidation in this paragraph and the next (starting on line 101). This suggestion may not be necessary, but if authors want to reduce the length of the introduction, I recommend combining and shortening these two paragraphs. I’d keep sentences on line 91-95 and remove the sentence on line 106-107 (this sounds like it belongs in methods) with some rewording to make the content flow better.

Reply: We have removed the section on methane oxidation and methanogenesis pathway from the Introduction that is shorter in the present version.

Line 263: I understand that the authors explored an LMM approach to look at environmental drivers of GHGs and methane processes but that they lost significance between variables they expected to correlate to GHGs (i.e., chl-a as a driver of pCO2) and so maintained their current approach. Still, I firmly believe the current statistical approach is inappropriate for the dataset as it does not account for pseudoreplication/repeated measures. When testing out the LMM approach, did authors incorporate sampling date into the formula? If using the glmmTMB package, this can be done by adding the date as a random effect (e.g., pCO2 ~ chla + nutrients + turbidity + (1|pond) + (1|date)), or better, by using the temporal autocorrelation function ar1(), which accounts for the similarity in samples collected close in time (e.g., pCO2 ~ chla + nutrients + turbidity + (1|pond) + ar1(date|pond); date and pond are in ar1() so that autoregression is applied to data based on the site). I strongly suggest GLMM or LMM because significance is going to be severely inflated due to repeated measures (Type I error), which inflates degrees of freedom. Any significance seen using the Pearson approach may be misleading, despite that authors expect them to occur.

Reply: The statistics section has been revised to use GLMMs where appropriate. We performed GLMMs to investigate relationships between data and comparisons between ponds. When the model did not converge, we kept Pearson for the relationships between data, and repeated measures ANOVA which, to a certain extent, considers the nature of the repeated measures over time. The following analyses were carried out, as described in the section on statistics new lines 212-226:

"For the data-sets covering the whole sampling period, for pCO_2, dissolved CH_4 concentration, $\%N_2O$, bubble flux, $\%CH_4$ in bubbles, and both ebullitive and diffusive CH_4 fluxes, generalized linear mixed models (GLMMs) were constructed that included water temperature, rainfall, $\%O_2$, Chl-*a*, TSM, DIN, SRP as fixed effects, and "pond" and "sampling date" as a random effect to account for repeated measurements via the *lme4* package (Bates et al., 2015) in R version 4.4.1 (R Core Team, 2021). When comparing data among the four ponds, "sampling date" was used as a random effect and post-hoc tests were performed using estimated marginal means (*emmeans* package) to assess pairwise differences between ponds.

For comparisons between the four seasons, GLMMs did not converge due to insufficient number of data points. Comparisons on log-transformed data were then made using repeated measures Analysis of variance (ANOVA) with Tukey's honestly significant difference (HSD) post-hoc tests.

The relationships between the annual means of CH_4, CO_2 and N_2O fluxes and the annual means of a subset of variables (Chl-*a*, macrophyte cover, surface area, depth) were tested with Pearson's linear or quadratic regressions. The modelled bubble fluxes in Silex pond were compared to measured values with Pearson's linear regression.

Statistical significance was set at $p < 0.05$ for all analyses. For comparisons presented on boxplots, different lower-case letters indicate a significant difference between groups."

Line 368: I suggest repeating what the correlations were (positive or negative) to remind readers.

Reply: Done

Figures 3, 7, 11: It might help to simplify the boxplots in these figures by making one box for turbid and one for clear ponds and overlay the points from those sites, with points as a unique shape or color for specific sites. That would leave 2 boxplots per season and would make statistical tests simpler since authors seek to look at differences between clear and turbid sites, not individual sites (gas variable ~ stable state type rather than ~ site). Individual site variation can still be seen in the temporal plots on the right side.

Reply: The data-set of only 2 turbid-water ponds and 2 clear-water ponds is insufficient to test variable ~ type. This was also noted by the Editor in the decision letter. The best we can do is test differences among the 4 ponds (pairwise) and then discuss these differences taking into account that two are clear and the other two are turbid. Consequently, the graphs were kept with the boxplots of the 4 ponds separate rather than grouping them 2 by 2.

Figures 2, 4, 5: These can be supplemental figures.

Reply: We have removed 4 figures from the original submission, so the total number of figures is now 9. We preferred to keep these figures as we think they support the core results and conclusions of our work.