Review for bg-2023-184: From iron curtain to green belt: Shift from heterotrophic to autotrophic nitrogen retention in the Elbe River over 35 years of passive restoration

Overall:

The authors present estimates of DIN retention along a 100 km reach of the Elbe River across a period of dramatic reduction in industrial and wastewater processes. The authors use estimates of metabolism to partition DIN retention into autotrophic or heterotrophic compartments and craft an interesting and compelling narrative about how changes in organic pollution and nitrate result in shifting regimes of autotrophic or heterotrophic DIN uptake dominance. This work has strong potential, but I hope that the authors spend more time on developing a sturdier methodological framework before drawing their conclusions. This framework should be able to be replicated by a peer. I have listed several comments that should help them along this path. The figures are quite clear and informative—nice work!

The major comments I list will likely take considerable effort, but none are unreasonable given the work that has already been done. I caution the authors on their development of synthetic DO time series, subsequent metabolism calculations, and the ultimate uncertainty around those calculations. There are many "minor comments", some of which require more effort than others, but I regard them as not having strong influence on results or inference. Finally, I advocate for separation of Results and Discussion. I think it will help the authors synthesize their results more effectively, and will greatly help the reader in understanding the important take-homes from this work.

Major Comments:

2. Data and methods

This section would greatly benefit from a data collection or acquisition section that says where the data came from, the data resolution, the tools/instruments/methods used to collect/analyse it, the frequency of measurements, location of measurement, etc. The reader, even after reading the supplement, is not given this information. See for example:

L117-118: From where are these DIN data collected? Who collected them? What were the protocol? Protocols often change over such a time span, and these protocol may have differing uncertainties. It's worth mentioning and discussing.

In 2.2 Study Site, you can add also here the depth, width, other water quality parameters of interest (e.g., alkalinity, pH, phosphorus). You can further describe the changes in vegetation/trophic state over time, what is meant by the "vegetation period" that you refer to later.

2.4 Metabolism estimates and S4

A few notes on these sections and oxygen/metabolism analysis:

1) you say "light use efficiency (k600)" at the top of S4…I imagine this is just a typo. k600 is the gas exchange velocity.

2) you don't say if you used hierarchical modelling or not. How did you fit each day with the Bayesian model? A continuous time series, or in daily chunks? You should provide the equation and/or the code.
3) "On 16 % of all days, k600 was negative, and those days occurred when residence times, water temperatures were high and DOsat is > 100% (Fig. S8)"
I think this is possibly a function of the way you created your time series. The hour of peak DO strongly influences estimates of k, and you'll notice that your model is peaking after your observations. This is an important point as incorrect k strongly influences estimates of ER. Moreover, the timing of when you would expect DIN retention to be highest (high temperature, high residence time, and high GPP) are the days when your k is negative. So precisely during the period of the thing you are interested in, you have the worst estimates of metabolism. I caution the authors to be more careful here.
4) Adding to the above point, is there an inherent reason to use a sine function with a mean of 16 for phi? Perhaps a generalised additive model would be better? Probably not too important, but might help reduce your error.
5) It is well-known that there is an equifinality problem in simultaneously estimating GPP, ER, and K. One simple way to evaluate if your data have this problem is to look for collinearity in estimated ER and K. Strong linearity implies that ER and K are poorly constrained, and this should be reported. One of the ways previous authors have dealt with this issue is to hierarchically model K so that days with similar discharge have similar K. This might not be necessary for your data, but I suspect it is why you have so many negative K days.
6) An additional metric that should be considered when evaluating the fit of the Bayesian model with multiple chains is the Gelman-Rubin statistic, Rhat. That helps to diagnose your days with poor model convergence.

2.6 Data Preparation
L177–180: "To estimate the effects of using daily instead of hourly water temperature measurements, we calculated the mean diurnal temperature variability from 24 years of hourly water temperature in the Elbe, which is 1.1 deg C (+- 0.7). For typical *DO*, *T*, and *p* conditions at the 180 Elbe, this can lead to a deviation in DO_{sat} of a maximum of 5.4 % (see Fig. S5), which we neglect in the following analysis"
The mean over 24 years won't be informative for this analysis. I imagine, based on experience, that the temperature change in winter will be near 0, which will heavily bias your mean towards 0. Moreover, winter is when there is no DIN retention, so you're missing the effect of temperature variability during the period of most interest—summer. There can be much larger temperature changes in summer during low-flow (e.g., on the order of 5-6 degrees C), which can be up to 1 mg/L difference in DO_{sat}, or more like 10% change. I'm not saying you can reconcile this issue, but it may be worth considering this uncertainty in your calculations and uncertainty propagation.

2.7 Estimating the N demand of metabolic processes
L218–220: "As U_{obs} and U_{met} values have relatively high uncertainties during Regime 1 (Fig. 2; Fig. S9b), we chose a seasonality-based validation approach for the metabolic N demand model. For the U_{obs} and U_{met} time series, we compared the annual mean (μ),
220 the day of the peak (ϕ), and the seasonality index (*SI*)."
This seems a bit forced to present a good "fit". Why not just compare the model and the observations with standard RMSE, bias, etc. and then try to understand why "Regime 1" has

a poorer fit than the others? Perhaps the data is of lower quality or the simulated oxygen and subsequent metabolism data are incorrect?

3.1 DIN retention

There is a missed opportunity to discuss the changing peak in DIN retention from Julian day 106 to 182 (from mid-April to the beginning of July) across the study period. Considering that results and discussion are bundled in this work, why do you think this is? To me, the obvious explanation is that in Regime 3, DIN retention is controlled by GPP, which exhibits a seasonal peak around day 180–190, whereas previously retention was controlled by ER, which (typically) exhibits less seasonality. I'd advocate for separating Results and Discussion for this reason—it will allow the authors to be more synthetic in their writing, which the scope of their analysis seems to beg for.

Figure 3

The summary time series of ER pre-1990 highlight a likely issue with this analysis. This is a very unusual pattern for ER, even in a more heterotrophic system. What is likely happening is that you are dealing with k600 and ER equifinality during this period, and thus ER estimates are hard to trust. I again caution the authors to be careful with their metabolism analysis. It may be simpler (and instructive) to apply a hydraulic equation (Raymond et al. 2012) to estimate k600 each day, and then only estimate GPP and ER using their Bayesian approach. How much does this change your results and inference?

Figure 4

I really like Figure 4a, but I do not get the rationale for 4b-d. Why use the annual mean, the seasonality index, and the day of peak DIN retention as your metrics for your model? This is never clearly explained, and it seems forced. The results from Figure S9 with reported RMSE and bias seem important to include here, as well (i.e., not in the supplement). The entirety of this analysis hinges on one free parameter, the growth efficiency of heterotrophs – a relatively un-measured/unknown quantity in large rivers. What happens if ER is systematically biased in earlier years (and it likely is) in this analysis? PQ also varies greatly depending on organisms, light, nitrate, and O2 conditions, all of which are changing over time. I understand the need to simplify this analysis (and I appreciate the power of back-of-the-envelope calculation), and that perhaps this is just a first step into more detailed work, but that needs to be made more obvious. More effort needs to go into the Methods to provide a rigorous framework for the analysis set forth here. That in addition to a more fair-handed view of the uncertainty in this result.

Minor Comments:

L26: No Diamond et al. 2022b in references.

L28: No Ehrhardt et al., 2019 in references.

L59: GDR not defined yet.

L66: Need to defined "heterotrophic" and "autotrophic-dominated" metabolic regimes. Metabolic regime can have many interpretations; which are you referring to?

L70-71: GPP and ER not yet defined.

L74: DO not yet defined.

L82-83: “GPP in the Elbe is mostly caused by phytoplankton (Hardenbicker et al., 2014)…” Can you be more quantitative here? Moreover, If that’s true, is it really “retention”? Wouldn’t this just be a change in form from DIN to particulate organic nitrogen? Which is not retained, but is transported downstream to be respired and turned back into DIN?

L83: “its” is for phytoplankton or GPP here?

L85: “10%” of…total retention? What does this number refer to?

L86: “nitrification” I think a short paragraph defining the author’s conception of “retention” is warranted. What role does nitrification play in retention?

L89: “DIN retention” Again, so far it sounds like autorophic DIN uptake, not retention.

L91: Is the hypothesis you are referring to the previous sentence? That sentence is not very easy to transform into a testable hypothesis as written. Please spend some more time to clarify your main question and hypothesis.

L122: “…as described in Wachholz et al. (2022)” That’s fine, but you could briefly say the gap-filling method. Do the two sites have nearly identical water chemistry? Was it a linear regression?

L123: “The discharge gage is located 50 km downstream of the sampling site used to estimate DIN load (station Geesthacht).” But it’s also 161km from the upstream site used to calculate load, right? That’s far.

L125: “…a previous mass balance study (Ritz and Fischer, 2019) assumed the errors to be ≤ 5% in the Elbe.” Can you be more specific here? The mass balance study assumed—or calculated?—the errors for this specific section of the Elbe to be 5% from the sampling gage? Hard to know what this means without digging into the other paper.

L139: “…was computed based on Gaussian error propagation (Section S2).” The 10% error assumption for C and Q is fine if you could show (with data or reference) that it’s conservative. I’m not sure it is. Wouldn’t the upstream and downstream Q have different uncertainty? You mention lack of “noteworthy” tributaries, but lateral gain in flow may be a few percent as well?

L144–145: “We calculate R_{obs}, RR_{obs}, and U_{obs} for both DIN and NH_4-N.” Why? Why not also NO3?

L149: “…when a discharge mass balance was considered.” I don’t understand this. Don’t you use the same Q for Lout and Lin? Please clarify.

L151–153: Please reformulate this sentence…I think it may have been two sentences originally.

Equation IV: “par” is not defined in the text.

L157: “…k600 is the gas exchange coefficient…” should be gas exchange “rate” or “velocity”, not coefficient.

L157: “…Schmidt number…” should be Schmidt number “for oxygen”.

L160: “Section S4”. See above comment in major comments.

L185: "*PQ* and *RQ* describe the ratio of O_2 produced/ consumed per CO_2 consumed/ produced." Small quip here: my understanding is that PQ has units O2/CO2, whereas RQ has units CO2/O2. I don't think it matters much in this sentence, but in Equations V and VI I think it does.

L198: "…however, it is well known that the measured ER not only caused by heterotrophic bacteria." Check phrasing.

L213: "…*curve_fit…"* More detail needed here. What method does this function use? An educated reader should be able to replicate this analysis.

L251: put units of "%" on "19 to 34"

Figure 3: Increase font size, please.

L290–291: "…high internal consistency." What does this mean?

L314: "…but it is unclear how that translates to rivers." Why? It's the same process, right?

L317–319: You've mentioned this several times and I would tend to agree…are there data to support this? E.g., BOD5?

L348–349: Can you expand on this logic? Why does low GE_{het} align with low NH_4? Wouldn't lower NH_4 necessitate organisms to be more efficient?

L369: "Considering our hypothesis…" What hypothesis? There doesn't seem to be one stated.

L408: "…of an autotrophic to heterotrophic regime shift…" But it's been heterotrophic the entire time, correct? Please be more clear on your use of these terms.