Response to comments by Referee #2

We thank Anonymous Referee #2 for her/his for providing useful and constructive comments. We will carefully revise the manuscript and address all the points raised by the Referee. Particularly, we intend to replace the linear regression model that was criticized by referee #2 by a qualitative analysis.

The paper presents an interesting case study of the effect of vegetated treatment system to reduce contaminant inputs to surface waters. The topic is important, the field and lab work involved a lot of effort and the analysis can help to improve the understanding of contaminant loss and mitigation with vegetated treatment system. Therefore, it is worth publishing after additional clarification and correction. Overall, I have similar issues with the data analysis in the current state as reviewer 1. I support his remarks and think that they are very well stated. The main points to address as described below.

- The uncertainty in the target variables peak-concentration reduction rate and mass removal rate are very high and largely ignored in the analysis. Some reasons for the uncertainties are:
  - The contaminant concentrations are highly variable and the dynamics are difficult to capture with the applied measurement resolution. Important peaks can be missed and last data points do not always reflect baseflow concentrations. This missing information can lead to large errors in the peak-concentration reduction rates. Simple linear interpolation of concentration for mass calculation (L 157) can led to even higher bias for the mass removal rates. A flow proportional measurement could have been a better option. Was flow proportion measurements available?

Reply: We agree with referee #2 that flow proportional sampling is the preferred option if accurate mass balances are desired. As already stated in our response to the first referee we aimed at including both aspects (peak and mass reduction) of pesticide mitigation in our approach. As such we chose flow-triggered sampling at fixed times with dense sampling at the beginning and increasing intervals towards later times. As in any sampling strategy (limitation of resources) there is a tradeoff between number of samples per event and number of events to be sampled. We evaluate our approach as a meaningful compromise. However, we further agree that uncertainty is of this procedure is difficult to evaluate. We will include this problem in the discussion section.

  - The timing of the last application of the investigated substances is completely neglected. However, more pronounced peaks are to be expected shortly after the application. Standardizing the concentrations or considering relative reduction rate does not completely solve this issue.

Reply: We agree that detailed information on application rates would be beneficial for interpretation of the pesticide signals emerging from the catchment. We will discuss this aspect in more detail in a revised manuscript (also see reply to referee #1). We generally think that the timing of application is one of several factors that affect relative reduction rates via chemograph shape which is the key point of our study.

- I think a much better understanding of the uncertainty of the contamination measurement could be gained from a detailed analysis of the discharge behavior during the investigated events. This data is available in a much higher resolution (L86: Stream flow was measured every minute . . .). Unfortunately, they are neither shown in detail nor really used in the analysis. It would be interesting to see the sample points/concentrations during an event together with the discharge measurement on a higher resolution, since the dynamics of the contaminants are driven mainly by the hydrology. I think this additional information would give an inside about how well the concentration dynamics have been captured. Moreover, information about the application patterns would improve the interpretation as well.

Reply: We will provide a figure showing all events and compounds together with discharge dynamics in the revised manuscript and discuss this aspect.

- The regression analysis is done rather poorly and the procedure is neither well explained nor well presented.
  - It is not shown that the condition for a simple multiple linear regression are fulfilled, since the results are not validated or at least this information is not shown. I would at least expect a classical residual analysis in the supporting information.

Reply: We thank the reviewer for this point. We actually did perform a residual analysis in which we did not find indication for violation of the assumptions of multiple linear regression. The results should have been communicated in the original manuscript. However, as we intend to drop the regression model as a consequence to further objections (s. below) we consider this point obsolete.

  - Automatically remove outliers based on a doubtful model without further analysis is not a proper way to go. For example, if outliers are a real problem, robust regression could be a solution (e.g. library robustbase).

Reply: In the original manuscript we removed two outliers in the RC and one in the RM model based on Cook's distance and standard deviation with the intention to reduce the influence of high leverage points. We agree that we should have dealt more sensitively with this issue and will revise the handling of outliers, i.e. not remove data points in a qualitative analysis.

– I don't think the requirement of independence of the data is fulfilled in this context. Data points of the same discharge event for the different components are not expected to be independent. Maybe a mixed model (with the discharge event as random effect) could help (e.g. library lme4). However, I doubt that more than a nice qualitative analysis of contaminate dynamics in a catchment with a vegetated treatment system will be possible with this setup.

Reply: We thank referee #2 for this interesting objection to the model used in our study and realize that we should have more carefully considered the structure of our data. Although other studies likewise neglected this aspect (Bundschuh et al., 2016; Stehle et al., 2011), we agree that a mixed model would be a better option. The intention of the regression analysis was to facilitate a comparison of model parameters identified as influential with the results of other studies. Now that we realize that we realize that the results of such other studies should also be treated with caution, the statistical analysis becomes somewhat obsolete. We therefore decide to go with the recommendation by referee #2 and limit ourselves to a qualitative analysis in which we compare potentially influential variables to reduction rates of peak concentration and compound mass. Boxplots will be used to illustrate variability in events for data available on the level of compounds and vice versa.

- It is somehow obvious that dispersion has a stronger effect on substances with a more pronounced peak (as explained by reviewer 1).

Reply: We agree with referee #2 that it is obvious that dispersion affects well-defined peaks stronger than flattened signals. However, as already discussed in our response to reviewer #1, it is not adequately stressed in existing literature on wetland contaminant mitigation.

- Although the clustering is done correctly, the connection with the discharge events is not well elaborated. Moreover, they are other clustering algorithms, which might be more robust (e.g. k-medoids. hierarchical Clustering). In L 207 it is written:"With the exception of cluster B which rather represented similar events (event 1 and event 4 in Fig. 2), overall clustering was controlled by similar behavior of contaminant groups." What was special by the event 1 and 4? Are these really exceptions? The contaminant groups seem to be important, however, I think the discharge dynamic and the application timing are important as well. Maybe it would also been interesting to cluster the discharge events. This data are also available in a higher resolution.

Reply: We agree with referee #2 that similar chemographs may emerge for many reasons, including compound-related (sorption affinity, degradability, application rate and timing), catchment-related (transport pathways, application areas) or event-specific (amount and dynamics of rainfall and subsequent discharge, incl. multiple peaks and spatial heterogeneity in rainfall) factors. As clustering is solely based on measured concentrations, the resulting clusters

are independent of whether we are aware of all relevant processes or not. We found that cluster A, C, and D mainly differed according to compound-related or catchment-related factors, separation of which is a bit challenging due to spatially separated application areas of fungicides and herbicides in the studied catchment (see also our comments to reviewer #1 above). In contrast, cluster B mainly reflected two different events. Examination of the discharge dynamics during the two events in cluster B did not reveal any obvious peculiarities. What made the two events in cluster B special was the absence of a time lag between peaks of fungicides and herbicides (and their TPs) that was usually observed and reflected in clusters A and D. The quicker response of herbicides in cluster B may be due to a recent application, however, exact application times and rates are unknown. We will include these observation in our updated manuscript.

We thank referee #2 for the suggesting to also consider alternative clustering algorithms. We agree that cluster centroids in k-medoids are more robust against outliers than cluster centers in k-means and revised our analysis correspondingly. We found similar clusters as when using k-means. In fact, partitioning between fungicides and other compounds was slightly better.

We thank referee # 2 for the suggestion to also cluster discharge events and agree that this may be helpful to check whether similar discharge events produced similar chemographs. However, we think that the number of discharge events with pesticide data is too low (n=10) for such an analysis. We will therefore rather include discharge conditions into comparison of cluster properties results of which will be included into the qualitative analysis of pesticide mitigation in the wetland as described above. We will make the rationale of the cluster analysis clearer and discuss the interpretation of the clusters in more detail in a revised manuscript.

Detailed comments

L100: Fig. 2: I guess the discharge shown in Fig. 2 is from G1. This should be included in the description.

Reply: This information will be added to the figure capture.

L 108: Overall, herbicides have been also shown to be very persistence. For examples, atrazine has been detected after 10 years without application. (e.g. https://doi.org/10.1021/acs.est.7b02529)

Reply: This was not meant to be a general statement. These lines (also s. next comment) just sum up, how the selected compounds are classified according to the Pesticide Properties Data Base (Lewis et al., 2016). We will make this more clear in a revised manuscript.

L 110: Azole pesticides are also persistent as indicated by many studies (e.g. https://doi.org/10.1016/j.envint.2020.105708)

Reply: (s. above)

L 131: What is the accuracy and precision of the method? Has the analytical method validated?

Reply: We will provide additional information about the method in a revised manuscript.

L 145: Why is the cluster analysis important for the calculation of the dispersion sensitivity index? The index could also be calculated without clustering.

Reply: We acknowledge that this phrasing was imprecise. The clustering revealed that there were major differences in chemograph shapes, particularly regarding the tailings of the breakthroughs. Thus, the clusters provided the idea for the index which is then considered a way to integrate the latter observation into further analysis.

L 191: From which mean? Do you mean 2 standard deviation from the prediction?

Reply: We acknowledge that our procedure for identification was debatable. As mentioned above we will change our analysis so that outliers do not have to be removed.

L 205: It doesn't make sense to talk about a peak in Cluster C ("T peak = 6h"). Not even the mean has a peak there.

Reply: This is true and will be corrected in a revised manuscript.

L 212: The surface runoff from the elevated vineyard has also to flow through the lower terrain slope to reach the river, expect that there are other shortcuts (streets, drains). See also reviewer 1).

Reply: We thank referee #2 for this comment as it points out that we have to consider the role of the catchment in more detail. We will revise our manuscript accordingly.

L 315: I do not understand the explanatory power for the different variable. Are they calculated by a univariate analyses? At least for me, the R-Output would be much easier to interpret.

Reply: The explanatory power of the resulted from decomposition of total R2 according to the method of Grömping (2006) (L. 193) which uses the mean over all possible orders of parameter addition to multivariate models. However, this will no longer be relevant as the regression model is discarded.

References

Bundschuh, M., Elsaesser, D., Stang, C., and Schulz, R.: Mitigation of fungicide pollution in detention ponds and vegetated ditches within a vine-growing area in Germany, Ecol. Eng., 89, 121–130, doi:10.1016/j.ecoleng.2015.12.015, 2016.

Grömping, U.: Relative importance for linear regression in R: The package relaimpo, Journal of Statistical Software, 17, 2006.

Lewis, K. A., Tzilivakis, J., Warner, D. J., and Green, A.: An international database for pesticide risk assessments and management, Hum. Ecol. Risk Assess., 22, 1050–1064, doi:10.1080/10807039.2015.1133242, 2016.

Stehle, S., Elsaesser, D., Gregoire, C., Imfeld, G., Niehaus, E., Passeport, E., Payraudeau, S., Schäfer, R. B., Tournebize, J., and Schulz, R.: Pesticide risk mitigation by vegetated treatment systems: a meta-analysis, Journal of environmental quality, 40, 1068–1080, doi:10.2134/jeq2010.0510, 2011.