Reply to the reviewer comments

Referee 4

We are grateful for this review and the helpful comments of referee 4, which will improve the manuscript. We are happy that also referee 4 considers the data of interest to the readers of Biogeosciences and clearly states that he/she supports publication in this journal. However, also referee 4 raised several critical points.

The main critical points are in accordance with the main criticism raised by the other referees: linguistic problems, manuscript length and speculations. As already commented we totally agree with these points. A revised manuscript will be corrected by a native speaker and speculative statements will be weakened according to the suggestions of all referees (see also reply to referee 1, 2, and 3). Furthermore, we agree with referee 4 that the most interesting part with scientifically new data is on the linkage between DOM and metabolism and we will focus a revised version on this topic as suggested.

Besides those general points, referee #4 raised several technical concerns regarding general methodical questions and statistical data analysis. Here we answer to the main critical points along the numbering (points 1 – 16) of the referee. We agree with all minor suggestions which are not listed here and will consider them in the revised manuscript.

1) "… clearly distinguish between a fluorescence pattern and the potentially (!) underlying chemical information…. "

> We will clarify the difference between DOM and CDOM in a revised manuscript in more detail and will more carefully distinguish between both in the description and interpretation of the data. See also our reply to referee 3.

2) Travel time "t": "One t [for propane injection] is definitely too short to reach equilibrium ("plateau") conditions along the entire reach. In fact, the peak time of a salt slug corresponds to the time point of maximum conductivity change (slope) of a metered (continuous addition), i.e., in the middle of the rising limb of the breakthrough curve of conductivity but potentially far from plateau conditions. The propane addition is a metered addition, it can therefore by no means be assumed that equilibrium conditions are reached after one t. …"

> We do not agree with this point. With the slug injection of the conservative tracer and the continuous propane injection we used a technique that is generally accepted for the detection of t and k (e.g. Reichert et al. 2009, Demars et al. 2011, and many other authors). We give a detailed description of the method in the reply to referee 3 and the slug injection generally followed standard procedure (Bales and Nardi 2007, U.S. Geological Survey Techniques and Methods 4-C2). However, if using the continuous salt addition, t is calculated from "the middle of the rising limb of the breakthrough curve of conductivity" which is the equivalent of the peak time of the conductivity if using the salt injection (cp. comment 2 referee 4, Hauer and Lamberti, Eds. (2007) Methods in Stream Ecology). This t can in both cases adequately be used to calculate

the DO changes between both stations. To our knowledge there is no literature available were researchers waited for 2 or even 4 t before sampling, as stated by the referee. If propane is sampled after one t at station one and all further samples are taken with the `wave´ the propane plateau is sampled and the same water body is sampled. Especially the last point is a strong benefit of our approach. To avoid further confusions, we will describe this approach in more detail.

3) "The authors compute both P/R and NEP to give information about the relative importance of GPP and CR24. However, in their argumentation P/R and NEP are almost used in a redundant way, not respecting the actual differences between the two in terms of their relative an absolute meaning (e.g, page 18268, first paragraph). "

This statement is probably based in misunderstanding as both (P/R and NEP) give different and individually relevant information. P/R is the ratio (GPP/CR24) whereas NEP gives the difference (GPP-CR24, seepage 18261-18262). The ratio P/R gives information about the trophic state (phototrophic/heterotroph) and NEP is a rate that indicates the net oxygen production of a stream and therefore the status of the metabolic balance (see page 18267). We checked the total manuscript to ensure that both parameters were used in the correct way.

"The same is true for the PARAFAC components, where once a ratio and then an absolute fluorescence is used, seemingly without any underlying reasoning other than achieving nice correlations."

As stated in the reply to referee 1 we tested all suggested ratio correlations and single component correlations. We will include all tested correlations in a revised manuscript (see detailed reply to referee 1).

"Furthermore, to examine the linkage between metabolism and DOM, it would make intuitive sense to combine ratios of P/R with ratios of fluorescent components, or to combine an absolute fluorescence of C2 with an absolute measure of metabolism such as GPP. But also here, authors seem to "pick" nice correlations, rather than follow a hypothesis-driven approach (see Fig. 9 for instance).

We have to reject the wrong impression that we pick nice correlations rather than following a hypothesis-driven approach. P/R is an indicator of the trophic state of a stream. It is a clear consequence of our hypothesis (linkage between metabolism and DOM composition) to correlate it with FI, the indicator of the DOM source. It is also a consequence of our hypothesis to correlate C2 (indicating autochthonous and probably phototrophically produced material) with P/R rather than with GPP, because GPP is daily NEP corrected for the assumed daily respiration. This account for the fact that freshly produced material is often rapidly degraded within the microbial food web (indicated by CR). We will explain this in a revised manuscript.

4) "I do not agree that the inner filter effect can be neglected in this study. The methods associated with EEMs are pretty straightforward and follow widely agreed standards in the scientific community. It is not difficult to carry out an IFE correction and I don't see any reason why this should not be done here. A lot of PARAFAC-argumentation boils down to comparison of identified components with components

published in the literature. How can we expect this to be successful if methods for EEM correction differ among studies?"

> Mobed et al. (1996, Environmental Science & Technology) described absorbance correction as an essential tool "for accurate representation and comparison of the EEMs of the humic substances at high concentrations". The inner filter effect correction is generally done if high concentrations are measured. Thus, if EEM's are measured for samples where the absorption is above approximately 10 m^{-1} at wavelengths between 300 and 600 nm (Stedmon and Bro 2008 L&O, Stedmon and Markager 2005 L&O), quenching effect can be expected and a correction is indicated. There is no benefit for the EEMs if inner filter correction is done for samples with low absorption (e. g. Stedmon and Bro 2008 L&O, Parlanti et al. 2000 and 2002 Organic Geochemistry, Yamashita et al. 2008 L&O, Murphy et al. 2006 Environ. Sci. Technol.). Absorption between 300 and 600 nm measured in our samples was not that high, so that it was not necessary to carry out the IFE correction. We will explain this in more detail.

5) "… 254 is a used wavelength for computations but was not measured according to methods. Please explain or correct."

> We thank the referee for this hint. We used excitation at 255 nm, not 254 nm for the calculation of HIX (P 18263, Line 21-24). This mistake will be corrected in a revised manuscript.

6) "…Please pick Spearman or Pearson correlation but don't use both. A lot of the regression analyses should rather be correlation analyses, as there is no clear identification of independent and dependent variable."

> We will follow these suggestions and we will use only non-parametrical tests in a revised manuscript (see also reply to referee 3). We will also cancel all regressions if no clear causalities are known (for example the test of linear relationship, Fig. 7).

7) "Landcover and seaons: A difference between non-forest and forest streams can not be statistically tested. First, the non-forest streams differ substantially with regard to discharge at least, so should not be regarded replicates. The "landcover" effect may as well be a discharge effect."

> A detailed response to a similar statement is given in the reply to referee 1. It was not the aim to use the stream types as (mathematical) replicates such as needed for ANOVA. We replicated in both time and space and applied correlation analyses to test for relationships. It is thus correct, that differences between the two types cannot be statistically tested. Nevertheless, the land cover is known to heavily alter stream metabolism by different mechanisms (e.g. shading by canopy). Therefore, the classification is in our opinion vital to characterize the streams. However, we agree that hard conclusions on mechanisms in explaining apparent differences in the categories (e.g. land cover versus discharge effect) are not possible and such mechanisms must be discussed with caution. We will consider this point in the revision.

"Second, the real sample size for both forest and non-forest is only 2. Plots like the boxplots in fig 10 suggest a much more powerful statistical analysis, which however has to be considered as almost

completely built on pseudo- replicated data (e.g. measurements from two consecutive dates and multiple seasons from the same system). Unless seasons and consecutive dates are accounted for as within-subjects factors in some sort of ANOVA or similar model a valid analysis is not possible."

> This statement seems to be based on misunderstanding the meaning of box plots (Fig. 10), which are standard forms of a graphical representation of numerical data. They indicate differences between groups without making assumptions of the underlying statistical distribution. Box plots allow the identification of outliers and give information about the variability within groups in contrast to bar charts, for example. It is thus correct to use box plots for the representation of the temporal variability within our groups. We checked the manuscript to ensure that there is no over-interpretations of box plots.

"…, I also note that the same season may actually mean very different dates for the different streams (sampling dates were up to 1.5 months apart for two systems in the same season, a considerable time distance especially for spring). I therefore do not think that seasons can nor should be compared, nor that this allows the use of "season" as a factor in any analysis (or at least this must be done with great caution)."

> We agree that we have to discuss results (e.g. the spring results on the between-stream differences) with caution due to different sampling dates. We also see that care is needed because single events can overlay seasonal patterns. Nevertheless, by knowing that relevant drivers undergo seasonal dynamics and after detecting variations in four seasons in all streams, we should address season as a factor in the discussion. We will ensure that statements seasonal comparisons are given with caution.

8) "Some of the statistical analyses (e.g., page 18266 last paragraph; page 18270 second paragraph) make the impression of a not very responsible combination of working with selected variables (and excluding others) and simultaneous exclusion of "outlier" cases (season-stream combinations). This gives some of the analysis a trial-and-error touch, which seems irresponsible and not very hypothesis-driven. If there are any outliers, I would prefer to see them in a graph still, maybe the correlation can still be computed without the outlier when indicated as such. "Outlier" data should not be considered "wrong" simply because it does not fit a model."

> Our statistical analyses are hypothesis driven, do not have a "trial-and error touch" and we did not treat outliner simply as "wrong"! In fact, we used a standardized and reproducible procedure following the tutorial of Stedmon and Bro (2008) to correct the used EEM's for outliers. At the end, only two EEM's, from 100 were rejected with this approach. With regard to metabolism values, we excluded the Hassel summer data when correlating metabolism parameters (GPP, P/R) with other values (TP, FI) because stream metabolism in the Hassel was clearly affected in summer by discharge events and by heavy modifications by cows. Because it is well known that such local events strongly alter the stream metabolism, it was our a priori decision to exclude the values from the analyses (which aimed to focus on the relationship of DOM-characteristics and stream metabolism and not on instantaneous effects of cattle excrement). We have presented these values in figure 5 and described the data handling in detail.

9) “If GPP is a function of light and TP, why did the authors not just consider a multivariate regression model? Maybe some of the outliers are not really outliers then…”

> In fact, we tested the effect of both light and TP in a multivariate regression model and found that this model is not valid because the data are not normally distributed.

10) “Some of the fluorescent components are interpreted as if having two fractions (e.g. page 18268, chapter 3.3.). I do not think that this is correct. Rather, one population of chemically similar molecules must be considered to produce manifold fluorescent signals. The same molecule may indeed produce two peaks in an EEM, as I believe. I am however, not an expert on this.”

> We totally agree that one molecule can produce different peaks. But when speaking about components we mean components modeled with PARAFAC analysis. PARAFAC components (C1, C2, C3) are a mixture of `chemical´ components (Table 2). For example, the PARAFAC component C3 is a composition of UVC humic-like substances, amino acids and/or proteins. Thus, PARAFAC components can have more than one fraction. We will clarify the difference between PARAFAC components and chemical components in a revised manuscript.

11) “Component ratios and “regressions” with components: I was totally confused by the report of these results on page 18268, where C1:C2 meant something different as C1/C2. Please clarify.

> We thank the referee for this comment. C:C is the same as C/C. We will use C:C consistently in the revised manuscript.

“Also, regression analysis seems really inadequate here, as independent and dependent variables cannot be clearly identified. Rather, correlation analysis should be chosen here. … Reading this paragraph I also think that it could be worthwhile to consider a ratio (C1+C3)/C2, this should give very similar information as beta:alpha.”

> We will follow this suggestion. We will focus on the significant correlations and delete the information on regression analysis (cp. reply to referee 3). We will also test the suggested correlation (C1+C3)/C2.

12) Table 1: can you add velocity to this table? It must be quite small and I am not sure if this is correct (e.g. about 1-3 cm/s for the first two entries in the table).

> We haven’t measured velocity. We will past calculated values in table 1.

15) “Fig. 5: According to my understanding there should be a maximum of 16 points in this graph. There are much more. Where are these coming from? …”

> This figure is based on single values. We measured stream metabolism, including irradiance on two days (n=2) for each stream (n=4) in different seasons (n=3) (P 18260, Fig. 4).

16) “All these graphs should show correlations rather than regressions. Modeled lines are therefore not adequate.”

We will follow this suggestion.