

**EDITOR:**

Editor Decision:

Publish subject to minor revisions (further review by editor) (24 May 2018) by Stacey Archfield

Comments to the Author:

The manuscript has received two reviews. Both reviewers had a positive reaction to the manuscript and felt that the manuscript was appropriate for publication in HESS, subject to what I would characterize as minor revision based on the excellent and constructive review comments.

The authors have responded to each of the comments and I would now instruct the authors to make their proposed changes to their manuscript. It was not clear from the author responses how Comments #3 and 4 by Reviewer 1 will be addressed in the manuscript so I ask that the authors indicate this in their final responses to the reviewer comments.

I look forward to final acceptance of this manuscript once the review comments have been addressed.

Thank you for considering HESS for your work,  
Stacey

## **AUTHORS:**

Dear Stacey,

thank you very much for the information about the positive assessment of our manuscript.

We made the proposed changes to the manuscript and added the missing information on the changes in the manuscript regarding comments # 3 and # 4 of referee 1 in our final response to the referees. We hope that those proposed changes likewise covered the raised issues. During revision, we realized one detail concerning the results at site S\_121 which might irritate the reader. We included this in our response to comment # 4 b of referee 1 to further clarify how the local water quality relationships at a site or of single samples relate to the overall picture constituted by the components.

Please find our detailed final responses to all comments of the referees and the manuscript with the marked changes. For our final responses we marked the unaltered responses from the 8<sup>th</sup> of May 2018 with blue font color. New or modified responses are marked with red font color. Please note that we did not explicitly state the formal changes (spelling, remove of spaces, updating of links) we applied in the revised manuscript in our final responses. The series of apparently double figures in the marked up version of the manuscript are caused by the new treatment of missing values according to comment # 7 of referee 1.

We want to thank once again both referees for their positive feedback, their constructive comments and their time.

The comments helped us a lot to improve the quality and clarity of the manuscript and made us think about some aspects of our study in more detail.

Best regards

Christian Lehr

(on behalf of the authors)

## Anonymous Referee #1

Received and published: 16 March 2018

**REFeree:** This manuscript presents a new exploratory framework for detection of dominant changes in multivariate water quality data sets with irregular sampling in space and time. The paper is well written and I think it is a valuable contribution to the hydrological community. I recommend its publication after the following comments are addressed.

General comments:

1. On the novelty of the proposed framework: I think this manuscript can foster future research ideas and efforts that are aimed toward detecting dominant changes in watershed using multivariate data at multiple sites. I think this type of coherent and systematic investigation of watershed data is limited in the literature, since previous studies have tended to focus on either only a few sites or a few constituents.

**AUTHORS:** We thank the referee very much for these very positive statements!

**REFeree:** 2. On the abstract: I found it quite lengthy (469 words), which prevents readers from quickly grasping the key messages. Also, it is not customary to have more than one paragraph in the abstract.

**AUTHORS:** We shortened it and reformatted it to one paragraph. The new abstract reads:

“Time series of groundwater and stream water quality often exhibit substantial temporal and spatial variability which can rarely be traced back to single anthropogenic or natural drivers. Typical existing monitoring data sets, e.g. from environmental agencies, are usually characterized by relatively low sampling frequency and irregular sampling in space and / or time. This complicates the differentiation between anthropogenic influence and natural variability as well as the detection of changes in water quality which indicate changes of single drivers. We suggest the new term ‘dominant changes’ for changes in multivariate water quality data which concern 1) multiple variables, 2) multiple sites and 3) long-term patterns and present an exploratory framework for the detection of such ‘dominant changes’ in multivariate water quality data sets with irregular sampling in space and time. Firstly, we used a non-linear dimension reduction technique to derive components which provide a sparse description of the dominant spatiotemporal dynamics in the multivariate water quality data set. They were used to derive hypotheses on the dominant drivers influencing water quality. Secondly, different sampling sites were compared with respect to median component values. Thirdly, time series of the components at single sites were analysed for seasonal patterns and linear and non-linear trends. The approach uses spatial and temporal heterogeneities as a source of information rather than considering them as noise, and considers non-linearities explicitly. It is especially recommended for the exploratory assessment of existing

long term low frequency multivariate water quality monitoring data. We tested the approach with a joint stream water and groundwater data set quality consisting of 1572 samples, each comprising sixteen variables, sampled with a spatially and temporally irregular sampling scheme at 29 sites in the Uckermark region in northeast Germany from 1998 to 2009. Four components were derived and interpreted as 1) agriculturally induced enhancement of the natural background level of solute concentration, 2) redox sequence from reducing conditions in deep groundwater to post oxic conditions in shallow groundwater and oxic conditions in stream water, 3) mixing ratio of deep and shallow groundwater to the streamflow and 4) sporadic events of slurry application in the agricultural practice. Dominant changes were observed for the first two components. The changing intensity of the 1<sup>st</sup> component was interpreted as response to the temporal variability of the thickness of the unsaturated zone. A steady increase of the 2<sup>nd</sup> component at most stream water sites pointed towards progressing depletion of the denitrification capacity of the deep aquifer.”

**REFEREE:** 3. On the coverage of the monitoring data: The paper addresses the ‘time’ aspect of the collected water quality data but lacks a thorough discussion on the ‘discharge’ and ‘season’ aspects of those data. Were all constituents at these sites sampled roughly similarly across season? Were they sampled roughly similarly during normal-flow and stormflow conditions? Such information is important and can be simply shown with boxplots (e.g., with “month” and “discharge percentiles” as x-axes respectively.) If samples at these sites were not taken roughly similarly across season or discharge, how would that affect the validity of the proposed exploratory framework and the interpretation of the results? The authors should comment on that.

**AUTHORS:** The monitoring did not explicitly distinguish between normal-flow and stormflow conditions. It rather aimed to fulfill the approximately monthly sampling frequency in the streams. Each sample contained all 16 constituents except for the missing values (Table S3). The grab samples were taken on the days marked in the left panel of Figure 2. Thus, while there are definitely irregularities among the series and within series over the course of time, the sites were sampled roughly similarly across season. The most important systematic deviation from this rule were the Peege sites and the most upstream sites of the Quillow, which often desiccate in summer (p. 36, l. 780-782).

In general, the interpretation of the components should consider the temporal structure of the data set. E.g. systematic deviations as the ones describe above should be considered. Thus, we included it in our interpretation of the 1<sup>st</sup> component (p. 36, l. 778 et seq.).

If the monitoring would in general not have been performed roughly similarly across seasons, e.g. if one or more seasons would in general be missing, the estimation of the seasonality would not be applicable. If the monitoring would be such, that there would be different seasons sampled in different years, this would have to be considered in the estimation of the trend.

We agree that considering discharge data would be valuable. Unfortunately, the monitoring did not include discharge measurements. The monitoring aimed to cover the spatial and temporal variability ~~variety~~ of water quality along the Quillow stream its tributaries and the adjacent streams. Discharge data was only available at sites Q\_93 and S\_118. Thus we did not include it in the presented analysis.

To clarify this we changed the last sentence of the third paragraph of section 2.2 in the revised manuscript to:

“In total, sampling intervals between two consecutive samples varied between nine and 714 days (Figure 2). The sites were sampled roughly similarly across seasons (left panel Figure 2). The most important systematic deviation from this rule were the Peege sites and the most upstream sites of the Quillow (left panel Figure 2 and Figure 1), which often fall dry in summer (Merz and Steidl, 2015).”

We changed the third sentence of the fourth paragraph of section 5.3 to:

“This leads often to drought in the uppermost stream reaches (left panel Figure 2; Merz and Steidl, 2015).”

We changed the first sentence in the first paragraph of section 2.2 to:

“The monitoring aimed to cover the spatial variability and temporal of water quality along the Quillow stream, its tributaries and the adjacent streams. The main focus of the monitoring was the Quillow catchment.”

We added a new sixth sentence to the first paragraph of section 2.2:

“Discharge data was only available at sites Q\_93 and S\_118 (Figure 1). Thus we did not include it in the presented analysis.”

And we added a new last paragraph in section 5.4:

“In general, the interpretation of the components should consider the temporal structure of the data set. For example in this study the drying out of the streams at the Peege sites and the most upstream sites of the Quillow in summer was the most important systematic deviation from an otherwise roughly similar sampling across seasons (left panel Figure 2). This information was included in the interpretation of the 1<sup>st</sup> component (section 5.3). If the monitoring would in general not have been performed roughly similarly across seasons, e.g. if one or more seasons would in general be missing, the estimation of the seasonality would not be applicable. If the monitoring would be such that there would be different seasons sampled in different years, this would have to be considered in the estimation of the trend.”

**REFEREE:** 4. On the general applicability of the framework: Several points shall be discussed by the authors regarding the applicability of the framework, which can guide its application to monitoring network elsewhere.

a) Is the framework intended to solute data only? Sediment and total phosphorus are typically monitored by many programs. Do the authors recommend the inclusion of such constituents in the proposed framework?

**AUTHORS:** Technically it is possible to include other data than solutes. However, the multivariate components derived by the dimension reduction approach are at the basis of our interpretation. Thus including other types of data might in some cases complicate the interpretation.

In general we would not mix variables with different scales of measures (e.g. nominal variables and ratio scaled variables).

For the interpretation we recommend to keep in mind, that all included variables are z-scaled prior to the dimension reduction. Thus all of them are equally weighted. For example if we would include only one sediment variable to our set of 16 water quality solutes, we expect that it would not change too much of the derived components.

To address this issue, we extended the fourth paragraph in section 5.5 to the following two paragraphs:

“Technically it is possible to combine other data than solutes (e.g. sediment data, biological indicators, etc.) together with the solutes in one joined data set for the derivation of the components. However, the multivariate components derived by the dimension reduction approach are the basis of the subsequent interpretation of the results. It has to be considered as well that all included variables are equally weighted due to the z-scaling prior to the dimension reduction. Thus, including other types of data might in some cases complicate the interpretation. In general, we recommend not to mix variables with different scales of measures (e.g. nominal variables and ratio scaled variables) in the data base for the derivation of the components.

Instead, data which was not used in the derivation of the components can be used as additional information for their interpretation. For example in this study, we used in addition to the spatiotemporal features of the components other variables like groundwater level series,  $\text{Fe}_2^+$  and  $\text{HCO}_3^-$  concentration from the groundwater samples, the spatial distribution of land use, and expert knowledge on the study area for the derivation of the hypotheses. A thorough...”

**REFEREE:** b) What is the threshold for a constituent (or a site) to be included in the analysis? Specifically, how many samples are required for a constituent-site pair to be eligible? I am puzzled by the few stations in Figure 2 that have only 1-8 samples. I wonder whether these site-constituent pairs should be disregarded.

**AUTHORS:** It depends on the focus of the study which samples might be considered neglectable. In our case the reasoning was to provide an exploratory approach which enables to get an overview on as much of the available data as possible without too many decisions beforehand which samples / sites to disregard (see also first sentence of comment 4c of referee 1). We intentionally included all samples

available, as long as not more than two of the 16 monitored variables were missing (p. 11, l. 221-223). If the data is organized as in our application, that means that the solutes serve as variables and the samples as observations, than the dimension reduction approach is “blind” to the information which sample belongs to which site. This information is maintained as index of the samples / observations. It is used for example to calculate for each site the median of the component values (section 4.2) and to assess at each site the representation of interpoint distances from the original data space in the low-dimensional projection (Table S4).

Because the selection of data points at a site is only a subset of the global data set for which the dimension reduction was performed, the performances regarding the representation of interpoint distances differ between the individual sites (Table S4) as well as compared to the overall performance for the global data set (Table 2). At some sites it can even happen that adding more components does not for every component improve the representation of interpoint distances in the low-dimensional projection. This occurred in this study only at site S\_121 where the representation of interpoint distances with four components ( $R^2 = 0.68$ ) was slightly worse than with three components ( $R^2 = 0.66$ ) (Table S4). This indicated an anomaly at this specific site compared to all other sites with respect to the 4<sup>th</sup> component, respectively the solutes which mainly determine the 4<sup>th</sup> component. We traced this phenomenon back to one single sample from the 25<sup>th</sup> of May 2004 which comprised relatively high DOC values and at the same time relatively low values of  $K^+$ , which is opposing the relationships indicative for the 4<sup>th</sup> component (Figure 3). The deterioration of the representation of the interpoint distances after adding the 4<sup>th</sup> component at this site vanished in an Isomap analysis which was performed without this sample. We were not able to find an explanation for this exceptional sample. However, it underlined that by applying a dimension reduction method every single sample is put into perspective of the global features of the data set as depicted by the components.

This interplay between the local perspective on the water quality relationships in the subsets of the sites or in individual samples on the one hand and of the global perspective of the whole data set on the other hand is a key feature of the presented analysis. The derived components constitute a frame in which all samples are integrated independent of the number of sample per site. Thus, in our application we get the information of how those sites with very little samples group or behave in relation to the others. Even a few samples might indicate e.g. that the respective site behaves similar to other sites with respect to some components and very different with respect to other components. This information would be lost if those samples would be excluded beforehand.

To clarify this, we added the following sentences after the fourth sentence in the second paragraph of section 3.3.2:

“For the local assessment of representation of interpoint distances at the individual sites, only the data points from the respective sites were used. Because the selection of data points at a site is only a subset of the global data set for which the dimension

reduction was performed, the performances regarding the representation of interpoint distances differ between the individual sites as well as compared to the overall performance for the global data set. At some sites it can even happen that adding more components does not for every component improve the representation of interpoint distances in the low-dimensional projection.”

We replaced the last two sentences in the last paragraph of section 5.2 with:

“However, the performance of the representation of the interpoint distances after adding the 4<sup>th</sup> component differed substantially between the different sites (Table S4). In case of site S\_121 the representation of interpoint distances with four components ( $R^2 = 0.68$ ) was even slightly worse than with three components ( $R^2 = 0.66$ ) (Table S4). This indicated an anomaly at this specific site compared to all other sites with respect to the 4<sup>th</sup> component, respectively the solutes which mainly determine the 4<sup>th</sup> component. We traced this phenomenon back to one single sample from the 25<sup>th</sup> of May 2004 which comprised relatively high DOC values and at the same time relatively low values of  $K^+$ , which is opposing the relationships indicative for the 4<sup>th</sup> component (Figure 3). The deterioration of the representation of the interpoint distances after adding the 4<sup>th</sup> component at this site vanished in an Isomap analysis which was performed without this sample. We were not able to find an explanation for this exceptional sample. However, it underlined that by applying a dimension reduction method every single sample is put into perspective of the global features of the data set as depicted by the components. Overall, the 4<sup>th</sup> component underlines the necessity to develop and use methods in environmental data analysis which enable to consider non-linear processes as well as singular and site-specific events.”

We moved the last paragraph of section 5.4 as new first paragraph. And we rewrote the former first paragraph as new second paragraph continuing as the new beginning of the third paragraph. The latter rewritten second and third paragraph reads:

“It is important to note that our approach does not require the same number of samples per site (Figure 2). The derived components constitute a frame in which all samples are integrated independent of the number of sample per site. Thus, in our application we get the information of how those sites with very little samples group or behave in relation to the others. Even a few samples might indicate for example that the respective site behaves similar to other sites with respect to some components and very different with respect to other components. The influence of single samples for the integration of the different sites into the global pattern of the water quality relationships summarized by the 4<sup>th</sup> component is an illustrative example for that (section 5.2). Thus, even occasional sampling at some sites helps assessing the strength of effects of the respective drivers at these sites and might support or contradict hypotheses on spatial variability and related long-term patterns of those influences. This information would be lost if those samples would be excluded beforehand.



In addition, the approach followed here does not require identical temporal sampling resolution at all sites or synchronous sampling dates. Thus, a strictly regular sampling design, which is hardly feasible, is no prerequisite. Correspondingly, data from different monitoring programs could be used for a joint analysis. **Sampling intervals ...**”

**REFEREE:** c) For such multi-site and multi-constituent exploration, all available data should be considered to enhance the robustness of modeling results. However, not all the data are consistently available across the sites. Then, how should one handle the tradeoff between the number of constituents and the number of sites? If we rank all constituents by the number of applicable sites, C1, C2, C3, C3, . . . , C16, then what is the relative gain of sequentially adding extra constituents (from C1 to C16) into the analysis framework? Can an explicit rule be developed to prevent adding new constituents to the framework?

**AUTHORS:** Again, this depends on the focus of the study. In our case we aimed to maintain the spatial coverage of the monitoring. If the main focus is to get an understanding of the multivariate water quality dynamics in detail, it might be worthwhile in the sketched trade-off scenario to disregard some sites and gain some constituents.

We have not thought about an explicit rule to prevent adding new constituents so far. But what we think could be considered is a correlation analysis of all variables beforehand to rule out the variables that correlate stronger than a pre-defined threshold. However, we recommend not to stick only to the threshold, but to visually examine the scatterplots of the respective variables to check for systematic deviations from the global relationship. There might be e.g. some sites or seasons in which the otherwise tight relationship gets weaker.

What we did is to exclude the variables with **less more** than 5% missing values (p. 10, l. 218-219) to keep the possible effect of any method of replacement rather low.

We included those considerations as two new paragraphs in section 5.5 after the new paragraph related to comment 8 of referee 1 and comment 7 of referee 2 and prior to the new paragraph related to comment 4a of referee 1:

“For data sets in which the number of measured variables differs between the sites there is a trade-off between number of considered variables vs. number of considered sites. Depending on the focus of the study different selections of the data set can be used. For example if the main focus of the study is to analyse the multivariate water quality dynamics in detail it might be worthwhile to disregard some sites to be able to include more variables. If the focus is to maintain the spatial coverage of the monitoring, like in this study, more sites might be of more value than additional variables. Depending on the available resources a third option would be to perform two analyses, one focusing on more variables, one on more sites, and comparing the results. If it is possible to link the considered components, like we did in the preceding paragraph, this proceeding can be used for spatial extrapolation of

the hypotheses derived from the version which included more variables. However, in our case the sketched trade-off was not dramatic. Thus, we excluded only the variables with more than 5% missing values (section 2.2) to keep the possible effect of any method of replacement rather low.

To prevent adding variables with little information gain it is recommendable to perform a correlation analysis beforehand and rule out highly correlated variables. For this purpose we recommend not to rely only on a numerical measure of correlation, but to visually examine the scatterplots of the respective variables to check for systematic deviations from the global relationship. There might be e.g. some sites or seasons in which the otherwise tight relationship gets weaker pointing to local or temporal phenomena.”

**REFEREE:** 5. On the irregularity nature of the monitoring data: The authors have provided adequate references in many parts of the manuscript. One exception is on the irregularity of water quality data (~ line 110 and also Section 5.4). One reference that you may find useful is provided below, which discusses at least two points that are discussed in this manuscript, including (a) irregularity nature of water quality data and how to model that property and (b) fractal scaling in water quality data which may affect trend significance (including the trend approaches used here).

Zhang, Q., Harman, C. J., and Kirchner, J. W. (2018), Evaluation of statistical methods for quantifying fractal scaling in water-quality time series with irregular sampling, *Hydrol. Earth Syst. Sci.*, 22, 1175-1192, <https://doi.org/10.5194/hess-22-1175-2018>.

**AUTHORS:** We included the suggested reference in the revised manuscript at the end of the fourth paragraph in the introduction:

“Thus, in environmental monitoring practice, data sets with gaps and periods with corrupted measurements are more the rule rather than the exception (c.f., e.g., Zhang et al., 2018 for river quality data).”

and in section 3.4.2 as new eighth sentence after the new seventh sentence related to comment # 11 of referee 1:

“Consequently, we did not consider the possible effects of the irregular sampling on the long-term memory (fractal scaling) of the water quality series either (Zhang et al., 2018).”

**REFEREE:** Specific comments:

6. On Figure 2:

a) This is a well-designed figure.

**AUTHORS:** Thank you very much!

**REFEREE:** b) Consider adding vertical reference lines in the right panel to indicate 1-day, 1-week, and 1-month intervals.

c) Add additional reference lines to separate groundwater from stream water – refer to your treatment in Figure 5.

d) Consider using color to distinguish between median and mean.

**AUTHORS:** We updated the figure according to your suggestions.

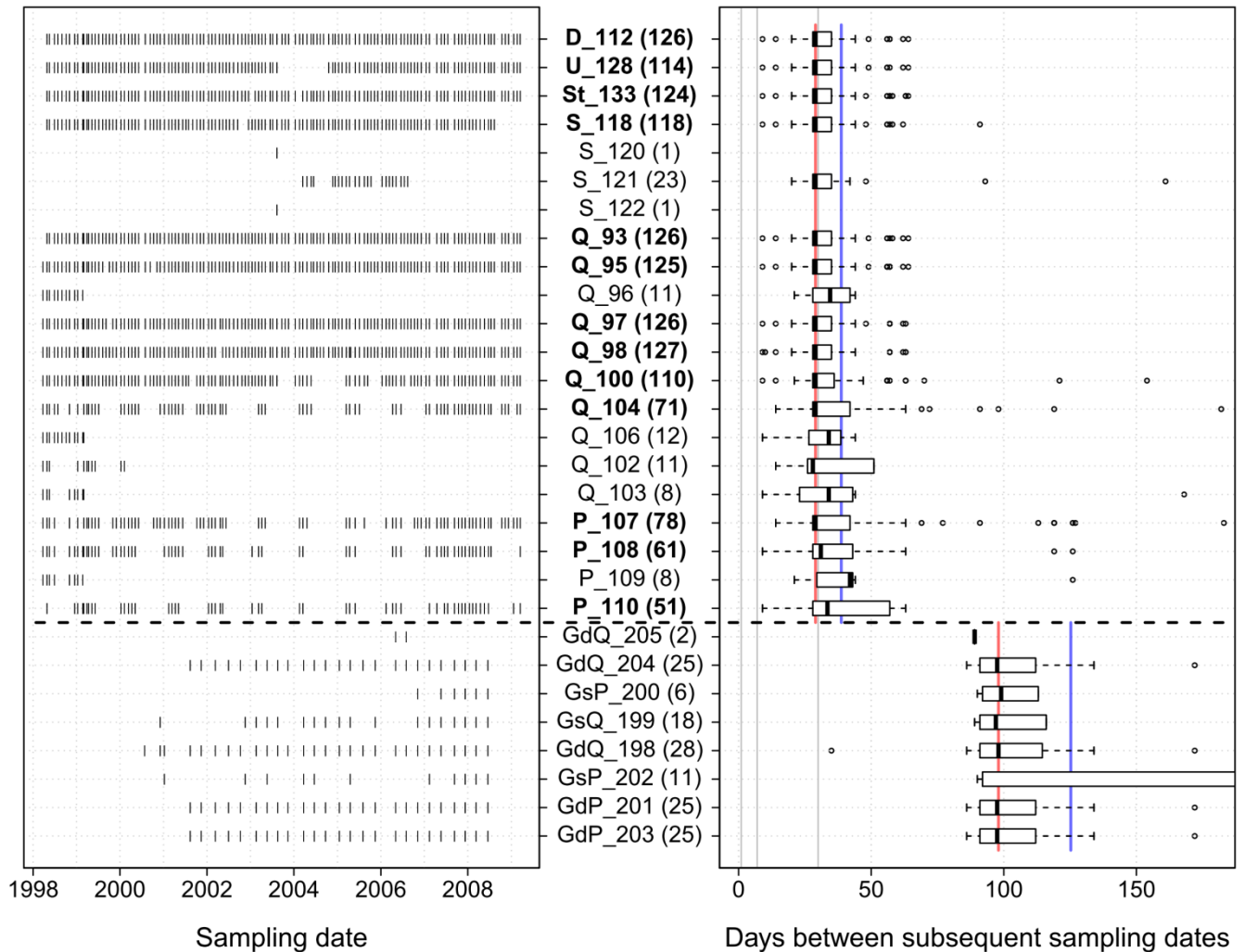


Figure 2 Left panel: Sampling dates at the sites for the whole monitoring period. Right panel: Boxplots of the variability of sampling intervals during the monitoring period. For better readability, the maximum of the x-axis is limited to 180 days. Median (red) and mean (blue) of sampling intervals are shown separately for the groundwater and stream water sites. Grey vertical lines mark the 1-day, 1-week and 1-month interval. Both panels: The dashed horizontal line separates groundwater sites (bottom) from stream water sites (top). Subscripts: P = Peege, Q = Quillow, S = Strom, St = Stierngraben, U = Ucker, D = Dauergraben, Gs = shallow groundwater, Gd = deep groundwater. The number of samples at each site is given in brackets. Names of the sites with more than 50 samples are printed bold.

**REFEREE:** e) Comment in the text on the apparent outlier in the site GdQ\_198 distribution.

**AUTHORS:** This was an exceptional sample taken during maintenance work. We included this information as fourth sentence in the third paragraph of section 2.2 in the revised manuscript:

“The one shorter sampling interval at site GdQ\_198 was an exceptional sample taken during maintenance work.”

**REFEREE:** f) Do the numbers in bracket represent the number of samples for one constituent or all constituents? Clarify.

**AUTHORS:** The numbers in bracket represent the numbers of samples. Each sample contained all constituents, except for the missing values (Table S3).

We added “Each sample contained measurements of all 16 variables.” prior to the sentence “Those water samples...” on page 11 line 221 in the revised manuscript.

**REFEREE:** g) Two of the sites have only one sample each. Justify why those sites should not be removed. In my opinion, those sites which only several samples should also be excluded unless their use can be justified.

**AUTHORS:** We aimed to demonstrate how the suggested exploratory approach can be used irrespective of those rather extreme differences between the numbers of samples per site to get an overview on as much of the available data as possible. While only of indicative value, it still can be interesting to see whether those single sample-sites plot / group different for the different components with respect to the other sites. Please see also our response to comment 4b) of referee 1.

**REFEREE:** 7. Line 248: I would suggest using median for the missing value replacement.

**AUTHORS:** In our case only a small percentage of samples were concerned (in the data set that was used for the dimension reduction at most for DOC: 3.44% and in the only for the comparison used groundwater samples at most  $\text{HCO}_3^-$ : 6.43% Table S3). We compared the two versions (missing value replacement with mean vs. missing value replacement with median). For the PCA, the scores of the first 10 components of the two versions yielded a  $R^2 > 0.99$ . For Isomap, the first 9 components yielded a  $R^2 > 0.99$  and the 10<sup>th</sup> component a  $R^2$  of 0.98. There were only minor differences in the site-specific cumulated  $R^2$  of the reproduction of the interpoint distances of the data in the projection by the first four components of Isomap at sites with  $n > 15$  (Table S4). Thus, for our case it did not really make a difference.

However, for other data sets this might be different. Thus, we agree that using median for the missing value replacement is in general the more robust approach.

Therefore, we updated the figures and results in the revised manuscript with the missing values replaced by the median and changed “mean” to “median” in the first sentence of section 3.1.

**REFEREE:** 8. Line 252: Provide references to justify the use of half detection limit for censored values. It is a typical practice but it has been pointed out that such treatment may cause issues to analysis – refer to the references below. This could be a problem for NO<sub>2</sub> and PO<sub>4</sub>, since the two species have significant proportions of censored values (Table S3).

Helsel, D.R., 2006. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, 65(11), pp.2434-2439.

Helsel, D. R. (2005). More than obvious: better methods for interpreting nondetect data. <https://pubs.acs.org/doi/pdf/10.1021/es053368a>.

**AUTHORS:** Thank you for this substantial comment and the provided references. As both referees raised this point, we will give a joint answer. Please see our response to comment 7 of referee 2.

**REFEREE:** 9. Line 262: How was the threshold of ‘50 samples’ chosen? It is still a small size.

**AUTHORS:** This threshold was a compromise between preferably long time series and the attempt to include preferably many of the series and sites in the analysis, to get an overview on the differences between the sites and catchments. The longest series in our data set comprised 127 samples. Thus, the data set as such is limited in this regard.

**REFEREE:** 10. Line 386 (Eq. 2): Check whether you want to use two equal signs in this equation.

**AUTHORS:** We rewrote the equation. Please see our response to comment 8 of referee 2.

**REFEREE:** 11. Line 421: The effect of autocorrelation on trend analysis is not only relevant to short-memory processes (e.g., AR(1) in Yue et al., 2002), but also long-memory processes (e.g., ARFIMA).

Cohn, T. A., and H. F. Lins (2005), Nature’s style: Naturally trendy, *Geophys. Res. Lett.*, 32, L23402, doi:10.1029/2005GL024476.

Zhang, Q., Harman, C. J., and Kirchner, J. W. (2018), Evaluation of statistical methods for quantifying fractal scaling in water-quality time series with irregular sampling, *Hydrol. Earth Syst. Sci.*, 22, 1175-1192, <https://doi.org/10.5194/hess-22-1175-2018>.

**AUTHORS:** We specified the addressed autocorrelation in the fifth sentence of section 3.4.2 as “short-term autocorrelation” and included the suggested references in the revised manuscript as new seventh sentence in section 3.4.2:

“Neither did we consider long-term memory and its effects on the statistical significance of the trends (Cohn and Lins 2005; Zhang et al., 2018).”

**REFEREE:** 12. Line 456: I think it should be 42% (per Table 2).

**AUTHORS:** 42% is correct. We corrected that in the revised manuscript.

**REFEREE:** 13. Line 459: In addition to temperature, PO<sub>4</sub> is also negatively correlated with PC 1.

**AUTHORS:** We decided to mention in the text for each component only the constituents which correlated strongest, because the interpretation was focused on those. The correlation with PO<sub>4</sub><sup>3-</sup> is negative, but almost zero. That is why we did not mention it. In the same manner, we proceeded for the other components.

**REFEREE:** 14. Line 463: This should be 18% (per Table 2).

**AUTHORS:** 18% is correct. We corrected that in the revised manuscript.

**REFEREE:** 15. Line 537: Check the label for  $n < 3$  in Figure 5, which should not be identical to  $n < 13$ .

**AUTHORS:** We changed the label for  $n < 3$  to “X” and reformatted the 4 plots in one column instead of a 2x2 matrix to enable larger labels for better readability.

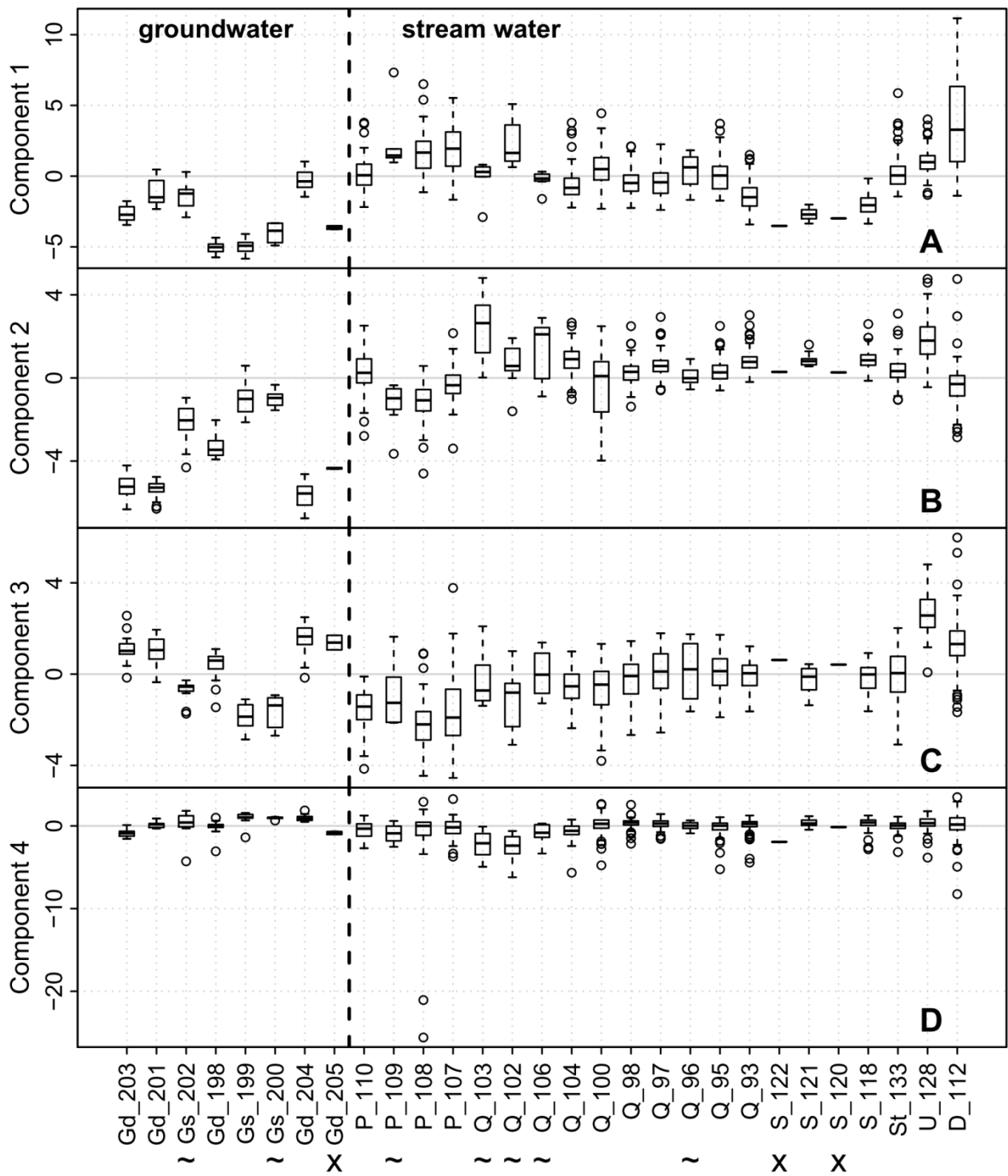


Figure 5 Boxplots of scores of component 1 to 4 at different sites. Sites with  $n < 13$  are marked with '~', those with  $n < 3$  with 'X'. Subscripts: P = Peege, Q = Quillow, S = Strom, St = Stierngraben, U = Ucker, D = Dauergraben, Gs = shallow groundwater, Gd = deep groundwater.

**REFEREE:** 16. Line 675: This conclusion should be supported by some references.

**AUTHORS:** We included references and changed the last sentence of the last paragraph of section 5.1 to:

“The catchments of the ~~analysed~~ ~~analyzed~~ streams are only sparsely populated and mainly characterized by intensive agriculture (Table 1). In agricultural landscapes slurry is a typical source in which those nutrients occur in high concentration (Hooda et al., 2000). We are not aware of any other high-concentration sources of this combination of nutrients in the region. The little number of scores with very low scores implied that there were merely single events occurring at some of the sites only. This fits to the finding that the timing of slurry application is crucial for the amount of nutrient loss to the streams (Hooda et al., 2000; Cherobim et al., 2017). Thus, we interpreted the negative peaks of the 4<sup>th</sup> component as sporadic events of slurry application, being either unintentionally directly applied to the stream during the spreading of the slurry or being leached via surface runoff and tile drain discharge after application.”

References:

Cherobim, V. F., Huang, C.-H. and Favaretto, N.: Tillage system and time post-liquid dairy manure: Effects on runoff, sediment and nutrients losses, *Agricultural Water Management*, 184, 96–103, doi:10.1016/j.agwat.2017.01.004, 2017.

Hooda, P. S., Edwards, A. C., Anderson, H. A. and Miller, A.: A review of water quality concerns in livestock farming areas, *Science of The Total Environment*, 250(1), 143–167, doi:10.1016/S0048-9697(00)00373-9, 2000.



## **Anonymous Referee #2**

Received and published: 22 March 2018

**REFeree:** The manuscript proposes an exploratory framework for detection of dominant changes in multivariate water-quality data sets with irregular sampling in space and time. As stated in the introduction, many analysis methods assume regular temporal spacing, but many monitoring networks evolve over time resulting in irregularly spaced samples. The concept is good, some more effort needs to be put into the writing and analysis.

**AUTHORS:** Thank you for the positive statement!

**REFeree:** 1. The abstract is rather lengthy.

**AUTHORS:** We shortened it. Please see our response to comment 2 of referee 1

**REFeree:** 2. The introduction contains vague statements and extraneous adverbs. The first sentence of the article is "Numerous high frequency studies unravelled the high temporal variability of stream water quality." This is well known, as shown by the many references. It seems like the first sentence of the article should start with a stronger sentence about the problem at hand.

**AUTHORS:** We added the following sentence as first sentence of the introduction in the revised manuscript:

"Detecting of changes in water quality and the responsible drivers are of fundamental interest for water management purposes as well as for scientific analyses."

**REFeree:** The second paragraph of the introduction has the phrase "numerous different drivers at different scales." This is vague. Give an example, or qualify the drivers, such as climatic and land-use drivers.

**AUTHORS:** We rewrote the sentence to:

"Instead, a variety of biogeochemical processes (e.g., Stumm and Morgan, 1996; Neal, 2004; Beudert et al., 2015), climatic (e.g., Neal, 2004) and hydrological (e.g., Molenat et al., 2008) variability and anthropogenic influences, for example agricultural (e.g., Basu et al., 2010; Basu et al., 2011; Aubert et al., 2013) or forestal (e.g., Neal, 2004) land use, land use change (e.g., Scanlon et al., 2007; Raymond et al., 2008) or urbanization (e.g., Kroeze et al., 2013), interact at different scales impeding identification of clear cause-effect relationships."

**REFeree:** The second sentence of the third paragraph is either missing something or "determining" should be "determine."

**AUTHORS:** We rewrote the sentence to:

“Usually only a few dominant processes determine the main dynamics of stream flow, groundwater head or water quality (Grayson and Blöschl, 2000; Sivakumar, 2004; Lischeid et al., 2016).”

and rewrote the sentence on p. 41 l. 929-931 to:

“This is consistent with the prior assumption that there are a few dominant drivers which determine the main stream water and groundwater quality dynamics in the region.”

**REFEREE:** 3. In the description of the study area the mean annual precipitation and mean annual temperature are given for the federal state Brandenburg for 1961–1990. This does not overlap with the study period of 1990–2009 at all. With the common use and availability of climatic data, it would not take much effort to report precipitation and temperature for the study period. It is not clear what period the water balance variability values represent.

**AUTHORS:** We replaced the addressed lines in the revised manuscript with:

“At the ZALF weather station Dedelow, which is situated approximately 500m ~~NE~~northeast of Q\_97 (Figure 1), a mean annual precipitation of 550 mm and a mean annual temperature of 8.9° C was observed for the hydrological years within the study period (1997-11 to 2009-10). The mean annual climatic water balance for this period, calculated from daily precipitation and potential evapotranspiration, was found to be -103 mm, exhibiting high interannual variability with -148 mm in the summer half year and +45 mm in the winter half year.”

**REFEREE:** 4. The topography and soils sections are well written and informative.

**AUTHORS:** Thank you very much for this positive feedback!

**REFEREE:** 5. We know the data are collected irregularly, but are they collected to be representative of seasons and flow conditions, i.e., are there high-flow samples?

**AUTHORS:** Thank you for this comment. Referee 1 raised this point as well. Please see our response to comment 3 of referee 1.

**REFEREE:** 6. Figure 2 shows some sites with very little data, yet it seems like they were included. It is not clear how these help inform the method. It seems like there should be some minimum number of samples per year most of the years from 1998 - 2009 in order for a site to be included in the study. Some parts of the proposed framework were done for sites with more than 50 observations. It seems like the entire analysis should be done only with those sites. It is not clear how these low-sample sites fit with the rest of the sites.

**AUTHORS:** Thank you for this comment. Referee 1 raised this point as well. Please see our response to the comments 4.b) and 6.g) of referee 1.

**REFEREE:** 7. It has been very well documented that substituting a fraction of the reporting limit is an inappropriate method for dealing with censored data. See:

Gilliom, R.J., and Helsel, D.R., 1986, Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques: *Water Resources Reserach*, 22, 135–146.

Singh, A., and Nocerino, J., 2002, Robust estimation of mean and variance using environmental data sets with below detection limit observations: *Chemometrics and Intelligent Laboratory Systems*, 60, 69–86.

Helsel, D. R., 2005, More than obvious - Better method better methods for interpreting nondetect data: *Environ. Sci. Technol.*, 39(20), 419A–423A, DOI: 10.1021/es053368a

Helsel, D.R., 2005, *Nondetects and Data Analysis: Wiley-Interscience*, 250 p.

Helsel, D.R., 2006, Fabricating data - How substituting values for nondetects can ruin results, and what can be done about it: *Chemosphere*, 65(11), 2434–2439.

Helsel, D.R., 2012, *Statistics for Censored Environmental Data Using Minitab and R: John Wiley & Sons*, 324 p.

Admittedly, the percent of censored values is small, but substitution should really not be used anymore in water-quality analyses. I'm not sure if Isometric Feature Mapping can utilize censored values. However, the authors could estimate the mean and standard deviation of the constituents with censored values using regression on order statistics or maximum likelihood methods (see Helsel, 2012) before standardizing the variables. The Akritas-Thiel-Sen median line can be used for the trend analysis.

**AUTHORS:** Thank you for this substantial comment and the provided references. As both referees raised this point, we will give a joint answer.

First of all, we agree that the question of how to deal with censored values is crucial and has to be handled with care.

The censored values in our study are the values below the detection limit of the respective variable, thus the measurements which are considered to be too imprecise to be reported as a single number (according to Helsel, 2012). Still they yield important information, in particular the ratio of values below the detection limit in comparison to values above the detection limit (cf. Helsel, 2012, page 12). This information is provided for all variables in table S3. We agree that censored values are not a big issue for our data set, except for the variables  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$  (and  $\text{Fe}^{2+}$  for the additional groundwater data, which was not used to calculate the components).

In our case, the purpose of the replacement of values below the detection limit is not to estimate distributional parameters such as mean or standard deviation or to perform statistical tests (like in most applications of the provided references). The

purpose is merely to provide values for all 16 variables in a sample so that the dimension reduction method can be applied.

The standardizing of the variables before applying the dimension reduction method is to achieve equal weighting of the variables. Therefore, the estimation of mean and standard deviation for this purpose has to be based on all values of a variable – whatever values are used for replacement of the censored values.

We are not aware of an isometric feature mapping variant, which is able to explicitly deal with censored values.

Helsel (2012) suggests to perform dimension reduction methods on the rank scaled variables or on a rank based distance matrix if censored values occur. To our understanding, we have to deal here with the trade-off between derivation of more “correct” components (the rank based case) and the loss of information that occurs, in case the ratio scaled variables are transformed to ranks (namely the information on the relative distances of the data points to each other, for example how distant the value of rank  $x$  is to the value of rank  $x-1$  in comparison to rank  $x-2$ , etc.). For the exploratory purpose of our study, we prefer to maintain this information in the light of the fact that only 2 out of 16 variables are substantially affected.

Although in our case the calculation of the components included the substituted values, the components themselves do not contain censored values any more. Thus, the subsequent time series analysis of the component scores does not have to be designed especially for the treatment of censored values (e.g. Akritas-Thiel-Sen median).

Concerning the correlation of variables and components, we used the residual plots and the spearman rank correlation of residuals and components (Section 3.3.3, p. 17, l. 377-388). We admit that a problem arises with the calculation of the multiple linear regression and therefore the residuals are affected as well. Again, we have to deal with a trade-off between potential information loss regarding the 14 out of 16 variables compared to the more correct treatment of 2 out of 16 variables. Spearman rank is one of the methods recommended by Helsel (2012) for the calculation of correlations with variables with only one reporting limit (in our case the detection limit). However - as in the case of the components - the residuals themselves are calculated with the censored values, but they do not contain censored values as such. For example for  $\text{NO}_2^-$ , the values that were substituted with half of the detection limit would get all the same rank, while the residuals of the linear model of  $\text{NO}_2^-$  with three of the components do not contain same-ranked values any longer.

Those two decisions (rank-based dimension reduction method yes / no and use of multiple linear regression and the residuals yes / no) can be questioned. Here, we provided arguments, why we did so. Following our argumentation and proceeding, the subsequent time series analysis of component scores as well as the correlation analysis between residuals and components should be not problematic.

In addition, of the two affected variables only  $\text{PO}_4^{3-}$  is substantial for the interpretation of a component, namely component 4. In this specific case, the range of values of the 4<sup>th</sup> component “was spanned mainly by single large values of  $\text{NH}_4^+$ ,  $\text{PO}_4^{3-}$  and  $\text{K}^+$  that cannot be explained with the preceding three components (Figure S4). This highlights the importance of particular events for the 4th component.” (p.21, l. 483-486). This fits to the distribution of  $\text{PO}_4^{3-}$  values which exhibits a substantial part of values below the detection limit and some outstandingly large values.

We checked for the influence of the substitution of the two affected variables on the components by performing another PCA and Isomap based on a data set in which  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$  were excluded.

The correlation of the PCA scores of the interpreted components 1 to 4 of version 1 (with  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ ) vs. version 2 (without  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ ) yielded a  $R^2$  of cp1: 0.99, cp2: 0.99, cp3: 0.99, cp4: 0.71.

The correlation of the Isomap scores of the interpreted components 1 to 4 of version 1 (with  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ ) vs. version 2 (without  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ ) yielded a  $R^2$  of cp1: 0.99, cp2: 0.98, cp3: 0.97, cp4: 0.64.

The same correlations were found for a third version in which  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$  were excluded and all missing values were replaced with the respective median, instead of the mean as suggested by Referee 1 in Comment 7.

The comparison of the two versions with respect to the Spearman rank correlations of Isomap scores of the first four components and the residuals (please see Figure 3 in the manuscript for the respective values of version 1) yielded a  $R^2$  of cp1: 0.98, cp2: 0.99, cp3: 0.99, cp4: 0.88.

Thus the first three components are virtually identical. The fourth component is affected, because  $\text{PO}_4^{3-}$  is one of the important variables determining this component. Still, the similarity of the correlations of Isomap scores and component 4 of both versions suggest that even for this component the variables  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ , and therefore the substitution of values below the detection limit with half of the detection limit, did not substantially affect the derived components.

To summarize:

We agree that the treatment of censored values is an issue that has to be considered carefully, in our case especially for  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ . We decided for our data set and the amount of affected values / variables to go not for a rank based dimension reduction method, due to the loss of information. Therefore, we needed to provide numerical values for the values below the detection limit. We decided to choose half the detection limit as a simple marker. The calculation of the components, the multiple linear regression and the residuals is affected by the substitution. We showed that for our case the substitution did not substantially affect the interpretation of the results.

We included the following paragraph after the third paragraph in section 5.5. “Exploratory framework” in the revised manuscript:

“The treatment of censored values can substantially affect the derived components and the subsequent interpretation of the results and has to be considered carefully (Helsel, 2012 and references therein). For the application of Isomap, it is required to provide numerical values for the values below the detection limit. For simplicity, we here used half the detection limit as a maker for values below the detection limit. We checked for the effect of this substitution by comparing the Isomap results of the presented analysis with another Isomap analysis in which we excluded the two most affected variables  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$  (Figure S4). The correlation of the Isomap scores of the interpreted components 1 to 4 of version 1 (with  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ ) vs. version 2 (without  $\text{NO}_2^-$  and  $\text{PO}_4^{3-}$ ) yielded a  $R^2$  of cp1: 0.99, cp2: 0.98, cp3: 0.97, cp4: 0.64. The comparison of the two versions with respect to the Spearman rank correlations of Isomap scores of the first four components and the residuals (please see Figure 3 for the respective values of version 1) yielded a  $R^2$  of cp1: 0.98, cp2: 0.99, cp3: 0.99, cp4: 0.88. Thus the first three components are virtually identical. The 4<sup>th</sup> component is affected, because  $\text{PO}_4^{3-}$  is one of the important variables for this component (Figure 3). Still, the similarity of the correlations of Isomap scores and the 4<sup>th</sup> component of both versions suggests that the characteristics of the 4<sup>th</sup> component were not merely introduced by the substitution of the values below the detection limit for  $\text{PO}_4^{3-}$ . Thus, overall, the substitution did not substantially affect the interpretation of the considered components. For data sets which are more heavily affected by censored values other dimension reduction methods such as the rank based approaches suggested by Helsel (2012) should be preferred.”

References:

Helsel, D. R.: Statistics for Censored Environmental Data Using Minitab and R, 2nd ed., John Wiley & Sons., 2012.

**REFEREE:** 8. Check equation (2) in line 385. Should there be a plus sign between B0 and the summation symbol? Describe the components of the equation that were not already described in equation (1).

**AUTHORS:** We rewrote the addressed paragraph to:

“Correlation between scores of a selected component  $cp_x$  and values of single variables might be blurred due to the effects of other components on the same variable. We excluded those effects by analysing the relationships between scores of the selected component  $cp_x$  and the residuals of the multiple linear regression mlr of the single variable  $v_i$  at hand and the remaining other considered components  $CP \setminus x$  (residuals):

$$\text{cor}(cp_x, \text{residuals}[\text{mlr}(v_i, CP \setminus x)]) , \quad (1)$$

where  $CP \setminus x$  is the set of  $m$  considered components, without the selected component  $cp_x$ ,  $\beta_0$  and  $\beta_j$  the intercept and coefficients of the regression

$$mlr(v_i, CP \setminus x) = \beta_0 + \sum_{j \in \{CP \setminus x\}} \beta_j cp_j + residuals \quad (2)''$$

**REFEREE:** 9. In the interpretation of components, the authors describe using multiple linear regression, which is a parametric method that assumes a model linear in the parameters, but then make an argument for a non-parametric measure of correlation applied to the multiple linear regression results. This seems contradictory.

**AUTHORS:** The residuals of the multiple linear regression were used to exclude the influence of the respective other three components in the assessment of correlation between single variables and components (p. 17, l. 377-387). Thus, the aim was to facilitate the assessment of the specific contribution of a single component out of the four considered components, especially in the visual examination of the residual-plots (p.17, l. 387).

To summarize the relationships between residuals and components we used Spearman rank correlation (p.17, l. 388+389). Most of the global relationships in this study were linear (Figure S1-S4). This is usually not known beforehand. Using Spearman rank correlation enabled to consider non-linear relationships between residuals and components as well, as long as they are monotonic.

However, the main benefit in this study was that Spearman rank correlation is less sensitive to extreme values compared to Pearson correlation. This concerned especially the assessment of the relationships of the residuals of  $SO_4^{2-}$  and  $Cl^-$  with the 2<sup>nd</sup> component and the residuals of  $PO_4^{3-}$  and  $NH_4^+$  with the 4<sup>th</sup> component (Figure S2 and S4), which were way stronger expressed with Pearson correlation due to a few single extreme values.

In addition, if the step with the multiple linear regression is omitted, thus if the correlations between variables and components are assessed based on the measured variables and not the residuals, than the use of Spearman rank correlation yields the additional benefit that it can deal with censored values (because there is in our case only one detection limit per variable → cf. Helsel, 2012, p. 218).

To clarify this issue we will replace the last sentence in section 3.3.3 with:

“To summarize the relationships between components and residuals we used Spearman rank correlation, which enables to consider non-linear relationships as well, as long as they are monotonic. Besides, it is less sensitive to extreme values than Pearson correlation.”

and the 2<sup>nd</sup> sentence in the 3<sup>rd</sup> paragraph of section 5.5 with:

“Again, whether the relationships are linear, as it was for most of the global relationships in this study (Figure S1-S4), is usually not known beforehand. Summarizing the relationships between residuals and components with Spearman

rank correlation enables to consider non-linear relationships between residuals and components as well, as long as they are monotonic. However, the main benefit in this study was that Spearman rank correlation is less sensitive to extreme values compared to Pearson correlation. This concerned especially the assessment of the relationships of the residuals of  $\text{SO}_4^{2-}$  and  $\text{Cl}^-$  with the 2<sup>nd</sup> component and the residuals of  $\text{PO}_4^{3-}$  and  $\text{NH}_4^+$  with the 4<sup>th</sup> component (Figure S2 and S4), which were way stronger expressed with Pearson correlation due to a few single extreme values.”

**REFEREE:** 10. Consider presenting the methods and the results in the same order for parallel construction.

**AUTHORS:** Thanks for this comment. We tried different ways to structure the manuscript during the writing process before we ended up with the current structure. The reasoning was to firstly introduce separately all the tools in the methods section before we secondly present the results from the perspective of the different aspects of the dominant changes in the data set.

We still think that it is a reasonable compromise for the purpose of this study. The structures of the methods and results sections are not parallel as you mentioned. Instead, we explicitly introduced the structure of the results and discussion section in the section “3.2 Exploratory framework”. The purpose of this section is to wrap up all the methods in one consistent picture and illustrate the workflow.

**REFEREE:** 11. In the discussion, the conclusions on page 32 about the 1st component were not well supported. There were a lot of statements like "we assume a general effect," some process "might" happen, some processes "tend to enhance." The discussion of the 2nd component was better supported with information about the sediments in the area. Some of the material in the first paragraph of section 5.2 should be moved up to better support the conclusions about the 1st component.

**AUTHORS:** Interpretations of the components were developed in a systematic way, considering the aspects of the correlations of variables and components (section 5.1), the spatial patterns (section 5.2) and the temporal patterns (section 5.3) of the components. Any interpretation is not only based on section 5.1 but after putting the different pieces of information in section 5.1, 5.2 and 5.3 together (p. 17, l. 374-376). We would like to stick to this structure for the sake of clarity. As guidance for the reader, we present the hypotheses for the components already in section 5.1. Correspondingly, we formulated the hypothesis for the 1<sup>st</sup> component in section 5.1 in a careful manner, to express that the aspect of correlation among the solutes alone is merely one aspect which needs further support. This is realized in sections 5.2 and 5.3 in which we add the spatial and temporal patterns to the picture to strengthen our hypothesis.

To more explicitly state the background of our hypothesis for the 1<sup>st</sup> component in this early stage of the argumentation, we added a new introductory sentence for the 3<sup>rd</sup> paragraph in section 5.1 in the revised manuscript.



“The whole study region is characterized by relatively intense agriculture (Table 1).”

**REFEREE:** The discussion of the 4th component on page 33 seemed speculative. Has this been modelled or shown elsewhere?

**AUTHORS:** Thank you for this comment. Referee 1 raised this point as well. Please see our response to the comment 16 of referee 1.

**REFEREE:** 12. Page 37 states nicely some important implications of the observed water quality.

**AUTHORS:** Thank you very much for this positive feedback!

**REFEREE:** 13. Page 40, line 895, change "is" to "are."

**AUTHORS:** The “is” refers to “The assessment of ..... is less sensitive...”

**REFEREE:** 14. Page 40, line 901, "Complementary" does not seem like an appropriate word for this sentence.

**AUTHORS:**

OLD RESPONSE from 8 May 2018:

~~We rewrote the sentence to:~~

~~“In addition to the spatiotemporal features of the components we used other variables like groundwater level series,  $\text{Fe}^{2+}$  and  $\text{HCO}_3^-$  concentration from the groundwater samples, the spatial distribution of land use, and expert knowledge on the study area for the derivation of the hypotheses.”~~

NEW RESPONSE:

Considering the comment 4a) of Referee 1 we rewrote the addressed sentence. Please see our response there.

**REFEREE:** 15. Some of the results, discussion, and conclusions mention both PCA and Isomap, but some of the numbers, figures, results must come from one of them specifically. That should be made more clear.

**AUTHORS:** PCA is used here merely as a benchmark for the Isomap results (p. 15, l. 316+317) and to introduce the concept / functioning of dimension reduction methods to the reader, as we expected it to be more familiar to the hydrological community. To our knowledge it is the most established and most used dimension reduction method in hydrology. Another reason why we included it in the study is because some readers might want to apply the framework based on PCA alone.

Thus, all presented and discussed results are from Isomap except from the “benchmark” comparison with PCA (Table 2).

We clarified this in the revised manuscript. We moved the last sentence of the first paragraph of section 5.5 to the beginning of 5.1 and added another sentence:

“Non-linear Isomap performed in this study only slightly better with respect to the representation of interpoint distances than PCA (Table 2), suggesting that mainly linear relationships were of importance for the overall dynamics in the data set. As there were only minor differences, we will present in the following the results of Isomap only.”

The ~~second~~ **first** sentence in the second paragraph of section 5.5 reads now:

“Whether the relationships in the data set are mainly linear ones, as in this study, or whether there are considerably non-linear relationships as well, is usually not known in advance.”

**REFEREE:** 16. Check that numbers in the text agree with the numbers in the figures and tables.

**AUTHORS:** We carefully checked the manuscript. Unfortunately we missed the two numbers referee 1 pointed out (comment 12 and 14 of referee 1).

**REFEREE:** 17. In suggesting this approach, how do you know the results are sufficient?

**AUTHORS:** In our understanding, the sufficiency of the results depends on the purpose of the study.

Our purpose was to provide a framework for the exploratory analysis of dominant changes in the spatial and temporal features of multivariate water quality data sets. We think that we were able to demonstrate its applicability with the presented study.

**REFEREE:** Are there some measures of quality that can be incorporated into this?

A very basic measure of quality is to measure the amount of variance in the data set, which is assigned to the first components. For example a more or less evenly distributed variance among the first components indicates that there are no dominant structures in the data set the used method is sensitive for. This result in itself can be rather interesting. Apart from that it would be in this case most probably not possible to link the components to drivers which help to better understand the monitored system.

A next step can be to compare the results of different dimension reduction methods, as we did here with principal component analysis and isometric feature mapping (Table 2). If applicable, the results of the dimension reduction method can be evaluated with different performance measures (e.g. the PCA performance can be evaluated with the “classical” approach via the sizes of the eigenvalues that are assigned to the components, or the correlation of the distance matrices of the analysed data in the original data space and the projection, as it was done in this study).

Concerning the interpretation of the components, we want to emphasize once more that the suggested approach is an exploratory one. Testing the derived hypothesis - for example by correlating the results with additional data - is a next step. Another option would be to test the hypotheses with virtual or “real-life” experiments (p. 40, l. 907-909).

Depending on the structure of the data set (e.g. its spatial and temporal resolution, number of samples per site, etc.) one option could be to perform the suggested approach with different subsets of the data set and compare the derived spatial and temporal patterns for example for different regions or time periods. The same approach can be used to check the results for their dependence on specific selections of the data set, which can serve as an estimation of the representativeness of the results for the overall region and time period.

## **Detecting dominant changes in irregularly sampled multivariate water quality data sets**

**Christian Lehr<sup>1,2</sup>, Ralf Dannowski<sup>1</sup>, Thomas Kalettka<sup>1</sup>, Christoph Merz<sup>1,3</sup>, Boris Schröder<sup>4,5</sup>, Jörg Steidl<sup>1</sup> and Gunnar Lischeid<sup>1,2</sup>**

[1]{Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany}

[2]{University of Potsdam, Institute for Earth and Environmental Sciences, Potsdam, Germany}

[3]{Institute of Geological Sciences, Workgroup Hydrogeology, Freie Universität Berlin, Germany}

[4]{Landscape Ecology and Environmental Systems Analysis, Institute of Geoecology, Technische Universität Braunschweig, Langer Kamp 19c, 38106 Braunschweig, Germany}

[5] {Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), Altensteinstraße 6, 14195 Berlin, Germany}

15 Correspondence to: C. Lehr (lehr@zalf.de)

## Abstract

Time series of ~~groundwater and stream catchment~~ water quality often exhibit substantial temporal and spatial variability which can rarely be traced back to single ~~causal factors. Numerous~~ anthropogenic ~~or and~~ natural drivers ~~influence groundwater and stream water quality, especially in regions with high land use intensity. In addition,~~ typical existing monitoring data sets, e.g. from environmental agencies, are usually characterized by relatively low sampling frequency and irregular sampling in space and / or time. This complicates the differentiation between anthropogenic influence and natural variability as well as the detection of changes in water quality which indicate changes of single drivers. ~~Detecting such changes is of fundamental interest for water management purposes as well as for scientific analyses.~~

We suggest the new term 'dominant changes' for changes in multivariate water quality data ~~that which~~ concern 1) multiple variables, 2) multiple sites and 3) long-term patterns ~~1) more than a single variable, 2) more than one single site and 3) more than short term fluctuations or single events~~ and present an exploratory framework for the detection of such 'dominant changes' in multivariate water quality data sets with irregular sampling in space and time. Firstly, we used a non-linear dimension reduction technique to derive ~~multivariate water quality components. The~~ components which provide a sparse description of the dominant spatiotemporal dynamics in the multivariate water quality data set. ~~In addition, they can be~~ They were used to derive hypotheses on the dominant drivers influencing water quality. Secondly, different sampling sites were compared with respect to median component values. Thirdly, time series of the components at single sites were analysed for seasonal patterns and linear and non-linear trends. The approach uses spatial and temporal heterogeneities as a source of information rather than considering them as noise, and considers non-linearities explicitly. Spatial and temporal heterogeneities are efficiently used as a source of information rather than being considered as noise. Besides, non-linearities are considered explicitly. The approach It is especially recommended for the exploratory assessment of existing long term low frequency multivariate water quality monitoring data.

We tested the approach with a ~~large data set of~~ joint stream water and groundwater

data set quality consisting of 1572 samples, each comprising sixteen hydrochemical variables, sampled with a spatially and temporally irregular sampling scheme at 29 sites in the Uckermark region in northeast Germany from 1998 to 2009. Four components were derived and interpreted as 1) ~~the~~ agriculturally induced enhancement of the natural background level of solute concentration, 2) ~~the~~ redox sequence from reducing conditions in deep groundwater to post oxic conditions in shallow groundwater and oxic conditions in stream water, 3) ~~the~~ mixing ratio of deep and shallow groundwater to the streamflow and 4) sporadic events of slurry application in the agricultural practice. Dominant changes were observed for the first two components. The changing intensity of the 1<sup>st</sup> component ~~during the course of the observation period~~ was interpreted as response to the temporal variability of the thickness of the unsaturated zone. A steady increase of the 2<sup>nd</sup> component ~~throughout the monitoring period~~ at most stream water sites pointed towards progressing depletion of the denitrification capacity of the deep aquifer.

## 1 Introduction

Numerous high frequency sampling studies unravelled the high temporal variability  
65 of stream water quality (e.g., Kirchner et al., 2004; Cassidy and Jordan, 2011;  
Halliday et al., 2012; Neal et al., 2012; Wade et al., 2012; Aubert et al., 2013;  
Kirchner and Neal, 2013; Tunaley et al. 2016; Rode et al., 2016; Blaen et al., 2017).  
Therefore, monitoring water quantity and quality on the timescale of the hydrological  
response of the catchment is a key requirement for understanding water quality  
70 dynamics and its driving processes in detail (Kirchner et al., 2004; Neal et al., 2012;  
Halliday et al., 2012). While the development of sensor technology, data loggers and  
transmission technology hopefully will help to significantly increase the number of  
high-frequency monitoring programmes in the future, most of the existing monitoring  
programmes so far applied a rather low sampling frequency. Nonetheless, there is  
75 common agreement that for short periods with high-frequency data, longer periods of  
low-frequency monitoring provide invaluable context (Burt et al., 2011; Neal et al.,  
2012; Halliday et al., 2012; Bieroza et al., 2014). This is especially true for existing  
long term records which are required as reference to distinguish between natural  
short term and long term variability of the observed variables and the assessment of  
80 the effects of anthropogenic influence on water quality such as changes in land use  
in the catchment (Burt et al., 2008; Howden et al., 2011).

The intriguing temporal and spatial variability in water quality monitoring data sets  
can in most cases hardly be related to single causal factors. Instead, a variety of  
biogeochemical processes (e.g., Stumm and Morgan, 1996; Neal, 2004; Beudert et  
85 al., 2015), climatic (e.g., Neal, 2004) and hydrological (e.g., Molenat et al., 2008)  
variability and anthropogenic influences, for example agricultural (e.g., Basu et al.,  
2010; Basu et al., 2011; Aubert et al., 2013) or forestal (e.g., Neal, 2004) land use,  
land use change (e.g., Scanlon et al., 2007; Raymond et al., 2008) or urbanization  
(e.g., Kroeze et al., 2013), interact at different scales impeding identification of clear  
90 cause-effect relationships ~~(e.g., Stumm and Morgan, 1996; Neal, 2004; Scanlon et~~  
~~al., 2007; Raymond et al., 2008; Basu et al., 2010; Basu et al., 2011; Aubert et al.,~~  
~~2013; Kroeze et al., 2013; Beudert et al., 2015)~~. Usually a single solute is affected by  
numerous different drivers at different scales (cf., e.g., Molenat et al., 2008; Lischeid

et al., 2010; Schuetz et al., 2016 for  $\text{NO}_3^-$ ). Inversely, a single driver usually has an  
95 impact on various solutes (Massmann et al., 2004; Lischeid and Bittersohl, 2008).  
This suggests that trend analyses of single variables might easily be misleading with  
respect to the identification of driving factors. For this purpose techniques which are  
able to account for the interaction of multiple drivers and observed variables are  
preferable.

100 On the other hand, despite their complexity, catchments are highly constrained  
systems. Usually only a few dominant processes ~~are dominant and~~ determining the  
main dynamics of stream flow, groundwater head or water quality (Grayson and  
Blöschl, 2000; Sivakumar, 2004; Lischeid et al., 2016). Using joint information from  
different solutes is an established way to derive hypotheses on processes or other  
105 causal factors that are dominant in the monitored data. For this purpose, dimension  
reduction techniques, especially the linear principal component analysis (PCA), have  
been used in analyses of multivariate water quality data for long, mostly as  
exploratory tool for descriptive process identification (e.g. Usunoff and Guzmán-  
Guzmán, 1989; Haag and Westrich, 2002; Cloutier et al., 2008) or for determining  
110 mixing ratios (e.g., Hooper et al., 1990; Capell et al., 2011). If the analysed data  
consist of time series of one or several variables observed at different sites, then the  
temporal features of the results of the dimension reduction can be analysed in a  
spatially explicit way, e.g. with respect to seasonal patterns or long term  
developments at the monitored sites (Lischeid and Bittersohl, 2008; Lischeid et al.,  
115 2010).

However, many of the methods commonly used for analysing temporal  
developments in monitoring data sets require regularly sampled data. In practice the  
spatiotemporal design of sampling campaigns and monitoring networks often evolves  
during the sampling period in an irregular way. In order to obtain a regularly sampled  
120 data set, additional information with a different sampling design, e.g. from pilot  
studies or single sampling campaigns, might not be utilized in the analysis at all.  
Further irregularities in the spatiotemporal structure of environmental monitoring data  
sets arise typically during the monitoring itself from a variety of reasons such as  
failure of sensors or data loggers, measurement errors, loss of samples, periods of



125 | ice or drought, etc. Thus, in environmental monitoring practice, data sets with gaps  
and periods with corrupted measurements are more the rule rather than the  
exception [\(c.f., e.g., Zhang et al., 2018 for river quality data\)](#).

Lischeid et al. (2010) suggested a combination of exploratory data analysis  
methods to detect and analyse dominant processes and their temporal development  
130 | in multivariate water quality data sets that is capable of dealing with irregular time  
series. We built on that and extended it towards the detection of ‘dominant changes’  
in time series of multivariate water quality data that are monitored at different sites,  
i.e. at different parts of a catchment or in different catchments within a region. In  
analogy to the dominant process concept (Grayson and Blöschl, 2000; Sivakumar,  
135 | 2004), we use the term ‘dominant changes’ in a broad and descriptive sense referring  
to systemic changes that clearly exceed the ‘usual’ range of heterogeneities in the  
temporal, spatial or inter-variable structure of the observed water quality data. We  
considered changes as dominant that concerned 1) main components of the  
multivariate water quality data set rather than single water quality variables  
140 | (multivariate components); 2) behaviour at various sites rather than at single sites  
(multiple sites); and 3) long-term behaviour rather than short-term fluctuations or  
single events (long-term patterns).

To identify the dominant changes, we combined exploratory data analysis methods  
for non-linear dimension reduction, spectral analysis, linear and non-linear trend  
145 | estimation and monotonic trend test in one exploratory framework. The suggested  
approach was tested with a multivariate water quality data set that has been sampled  
with a spatially and temporally irregular sampling scheme in northeast Germany from  
1998 to 2009. In the following, we present and discuss the results of our case study  
according to the three aspects of ‘dominant changes’: 1) multivariate components, 2)  
150 | multiple sites and 3) long-term patterns. We continue with a discussion of 4) effects  
of the irregular sampling and 5) methodological aspects of the exploratory framework.

## 2 Data

### 2.1 Study area

155 The study area is the upper part of the basin of the Ucker river located in the  
northeast of Germany, about 90 km north of Berlin, which drains to the Baltic Sea  
another 50 km further north. It is part of the Leibniz Centre for Agricultural Landscape  
Research (ZALF) long-term monitoring region AgroScapeLab Quillow, the LTER-D  
(Long Term Ecological Research Network, Germany) and the TERENO (Terrestrial  
160 Environmental Observatories, <http://teodoor.icg.kfa-juelich.de>) Northeastern German  
Lowland Observatory. Water samples have been taken in the adjacent catchments of  
Dauergraben (78.9 km<sup>2</sup>), Stierngraben (104.8 km<sup>2</sup>), and Quillow (399.4 km<sup>2</sup>) with its  
subcatchments Strom (235.8 km<sup>2</sup>) and Peege (25.6 km<sup>2</sup>) (Figure 1). At the ZALF  
weather station Dedelow, which is situated approximately 500m northeast of Q 97  
165 (Figure 1), a mean annual precipitation of 550 mm and a mean annual temperature of  
8.9° C was observed for the hydrological years within the study period (1997-11 to  
2009-10). The mean annual climatic water balance for this period, calculated from  
daily precipitation and potential evapotranspiration, was found to be -103 mm,  
exhibiting high interannual variability with -148 mm in the summer half year and +45  
170 mm in the winter half year.~~For the part of the Ucker catchment which is situated  
within the federal state Brandenburg a mean annual precipitation of 584.5 mm and a  
mean annual temperature of 8.3°C was found for the 1961-1990 period and a mean  
annual climatic water balance of -40.4 mm was estimated with the ARC/EGMO model  
(Lahmer et al., 2000). The mean climatic water balance exhibited high interannual  
175 variability with -181.4 mm in the summer half year and +141 mm in the winter half  
year.~~

The topography of the region developed basically during the Pomerian stage and  
the Mecklenburgian stage of the Weichselian ice age, i.e. 15,200 to 14,100 years  
before present. Altitude varies from 20 m in the lowlands of the Ucker river to more  
180 than 100 m above sea level in the southwestern part of the study area. During the  
Pleistocene, repeated advances and recessions of the ice sheet deposited highly  
heterogeneous unconsolidated sediments of about 150 m to 200 m thickness. The  
base consists of a thick Oligocene clay layer which separates the upper freshwater

groundwater system from saline groundwater underneath. Based on borehole  
185 surveys, up to seven aquifers divided by layers of till have been identified within the  
unconsolidated Quaternary sediments. In some parts of the region patches of  
halophilious plants are found in the lowlands indicating local upwelling of saline  
groundwater from the underlying Tertiary aquifer through windows of the Oligocene  
clay layer.

190 Loamy and sandy loamy soils prevail that developed from the till substrate. Most of  
the region is intensively used as cropland, although the fraction of arable land differs  
between the catchments (Table 1). Forests comprise only a minor fraction of the area  
(Table 1). Land cover did not change within the study period from 1998 to 2009. The  
195 riparian zone of the catchments is mostly used as grassland, underlain by peat and  
organic and sandy fluvial deposits. The hummocky landscape includes about 1300  
closed drainage basins and small ponds with an area of the water surface < 1 ha  
(Kalettka and Rudat, 2006; Lischeid et al., 2016). Many of the larger depressions  
have been connected by ditches to facilitate drainage. Partly, these ditches have later  
been replaced by underground pipes for land reclamation. In addition, agricultural  
200 soils are extensively drained by subsurface tile drainage systems. From the 13<sup>th</sup>  
century till the end of the 19<sup>th</sup> century, the energy of the natural water courses was  
also occasionally used to power mills. Today, those mills are not active any longer  
and have been replaced in most cases by weirs for water management or ramps. For  
more details on the study site, please see Merz and Steidl (2015).

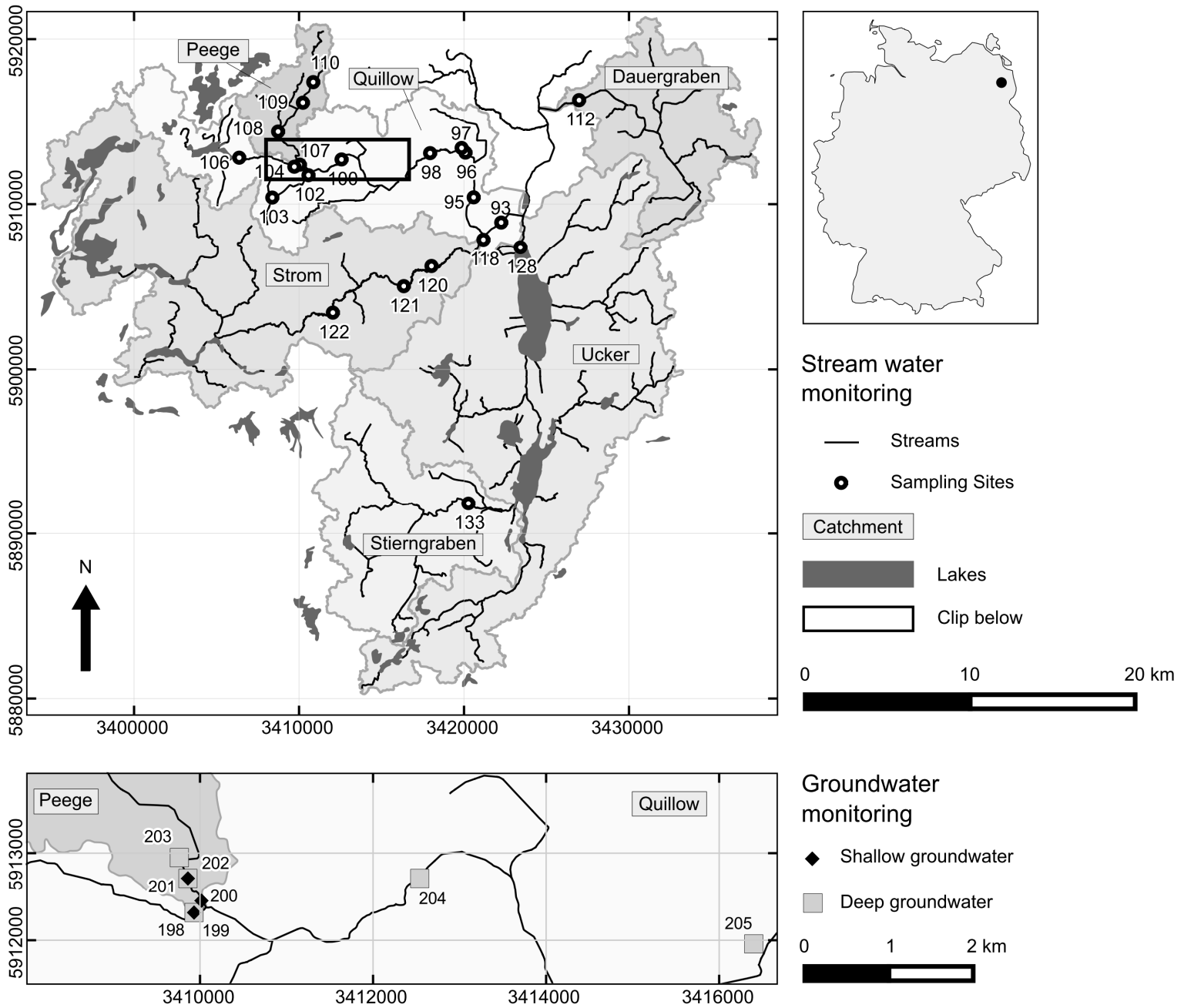


Figure 1 Map of the study area. Coordinates of UTM-zone 33N are given in m. Upper panel: Stream water monitoring sites and the location of the study area (Upper Ucker river catchment) within Germany. Lower panel: Section with the included groundwater monitoring sites. For better readability only the number of the ID of the monitoring sites is shown.

Table 1 Share of land use classes in the different catchments (percent of land cover) based on CORINE Land Cover data (2000).

	Settlements / Industry	Arable land	Grass-land	Lakes	Others	Wet-land	Wood-land
Dauergraben	1.7	92.1	4.1	1	0.3	-	0.8
Ucker	4.6	62.3	5.6	7.7	2.2	2.4	15.2
Stierngraben	1.4	61.2	15.8	1.2	0.9	-	19.5
Strom	2.2	54	7	6.9	1.2	-	28.7
Quillow	2.3	77	9.3	1.3	1.4	-	8.7
Peege	0	78.3	5.5	-	-	-	16.2

215

## 2.2 Sampling and analysis

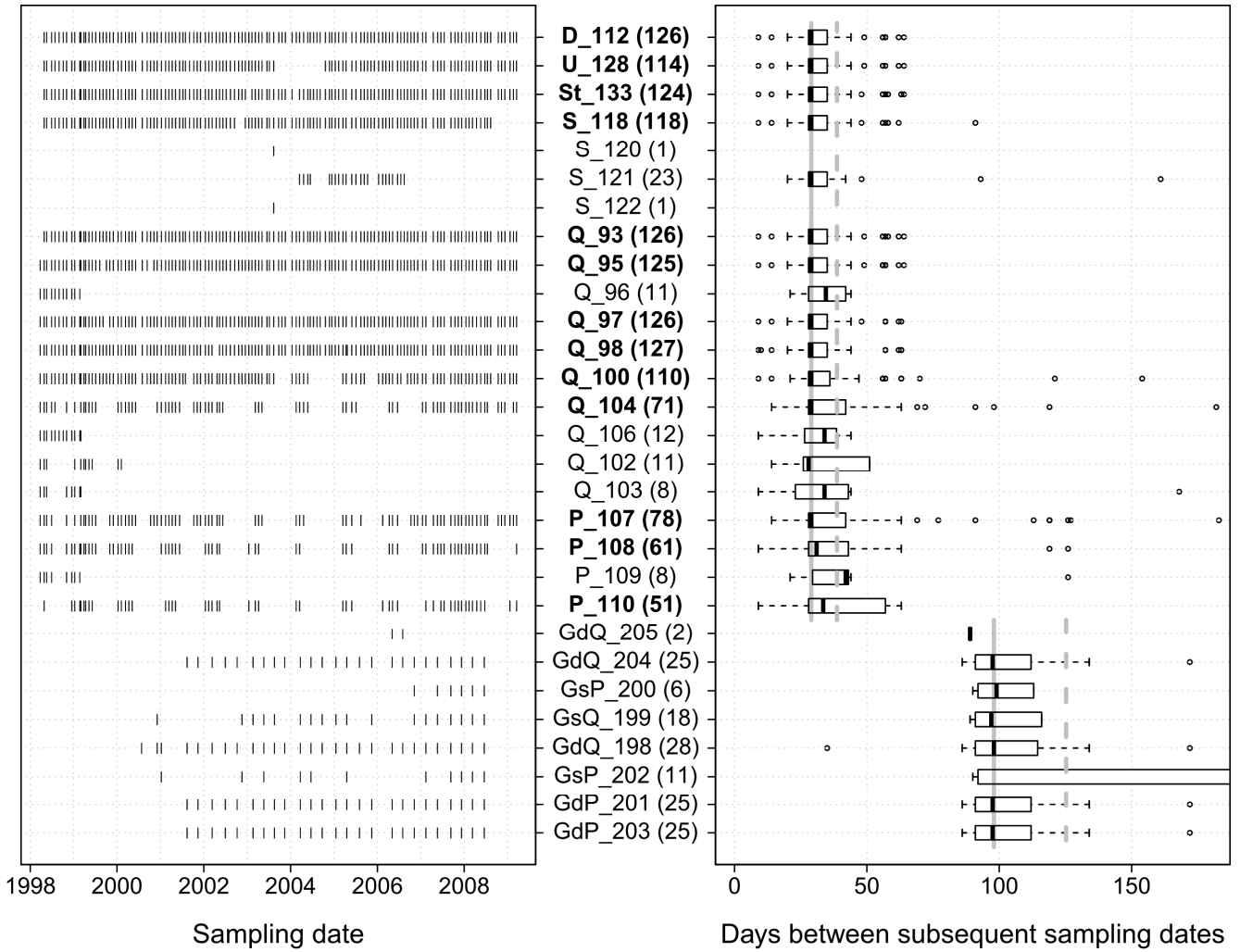
The monitoring aimed to cover the spatial and temporal variability of water quality along the Quillow stream, its tributaries and the adjacent streams. The main focus of the monitoring was the Quillow catchment. Here, eight sampling sites were located along the main stream, and another four at each of the two tributaries Peege and Strom (Figure 1 and Table S1). At the streams Dauergraben and Stierngraben and at the Ucker river, stream water quality was monitored at one site respectively. Stream water sampling started in 1998 and was performed until 2009. Discharge data was only available at sites Q\_93 and S\_118 (Figure 1). Thus we did not include it in the presented analysis. Groundwater quality was monitored in the Quillow catchment only, close to the middle reaches of the stream and close to the mouth of the Peege tributary, from 2000 to 2008 (Lower panel Figure 1). At this site, an up to 15 m thick horizontal till layer separates a shallow and very heterogeneous unconfined aquifer from a mainly confined deep aquifer. The separating till layer crops out further downstream (Merz and Steidl, 2015). Both aquifers were monitored (Table S2). The deep aquifer is known to be confined except at well Gd\_204. Groundwater level in the deep aquifer was measured daily with automatic data loggers at wells Gd\_198, Gd\_201, Gd\_203 and Gd\_204 (Merz and Steidl, 2014a).

Groundwater quality (Merz and Steidl, 2014b) and stream water quality (Kalettka and Steidl, 2014) monitoring in the Quillow catchment covers a wide range of water quality parameters. For the multivariate analysis in this study, we considered from the

joint groundwater and stream water quality data set only the 16 variables with less than 5% missing values, i.e.  $\text{NH}_4^+$ ,  $\text{NO}_3^-$ ,  $\text{NO}_2^-$ ,  $\text{PO}_4^{3-}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Mg}^{2+}$ ,  $\text{Ca}^{2+}$ ,  $\text{Cl}^-$ ,  $\text{O}_2$ , pH, water temperature, redox potential (Eh), electric conductivity (EC),  $\text{SO}_4^{2-}$ , and DOC (Table S3). Each sample contained measurements of all 16 variables. Those water samples for which more than two of the 16 monitored variables were missing were excluded from the analysis, resulting in a set of 1572 samples. In total, 0.69% of the values in the dataset were missing. In addition, we considered  $\text{HCO}_3^-$  and  $\text{Fe}^{2+}$  concentration from the groundwater monitoring (Table S3).

The number of temporal replicates varied between one and 127 per site (Figure 2). In general, streams were sampled at approximately monthly intervals, and groundwater samples were taken every three months. Median [mean] sampling intervals were 29 [38.7] days for stream water and 98 [125.3] days for groundwater. The one shorter sampling interval at site GdQ 198 was an exceptional sample taken during maintenance work. In total, sampling intervals between two consecutive samples varied between nine and 714 days (Figure 2). The sites were sampled roughly similarly across seasons (left panel Figure 2). The most important systematic deviation from this rule were the Peege sites and the most upstream sites of the Quillow (left panel Figure 2 and Figure 1), which often fall dry in summer (Merz and Steidl, 2015).

Further details on the data and measurement methods are provided by Merz and Steidl (2015). The selection of water quality data used in this article and the groundwater level data have been published under CC-BY 4.0 and can be accessed at doi: 10.4228/ZALF.2017.340~~http://open-research-data.ext.zalf.de/ResearchData/2017\_340.html~~ and doi: 10.4228/ZALF.2000.272 respectively.



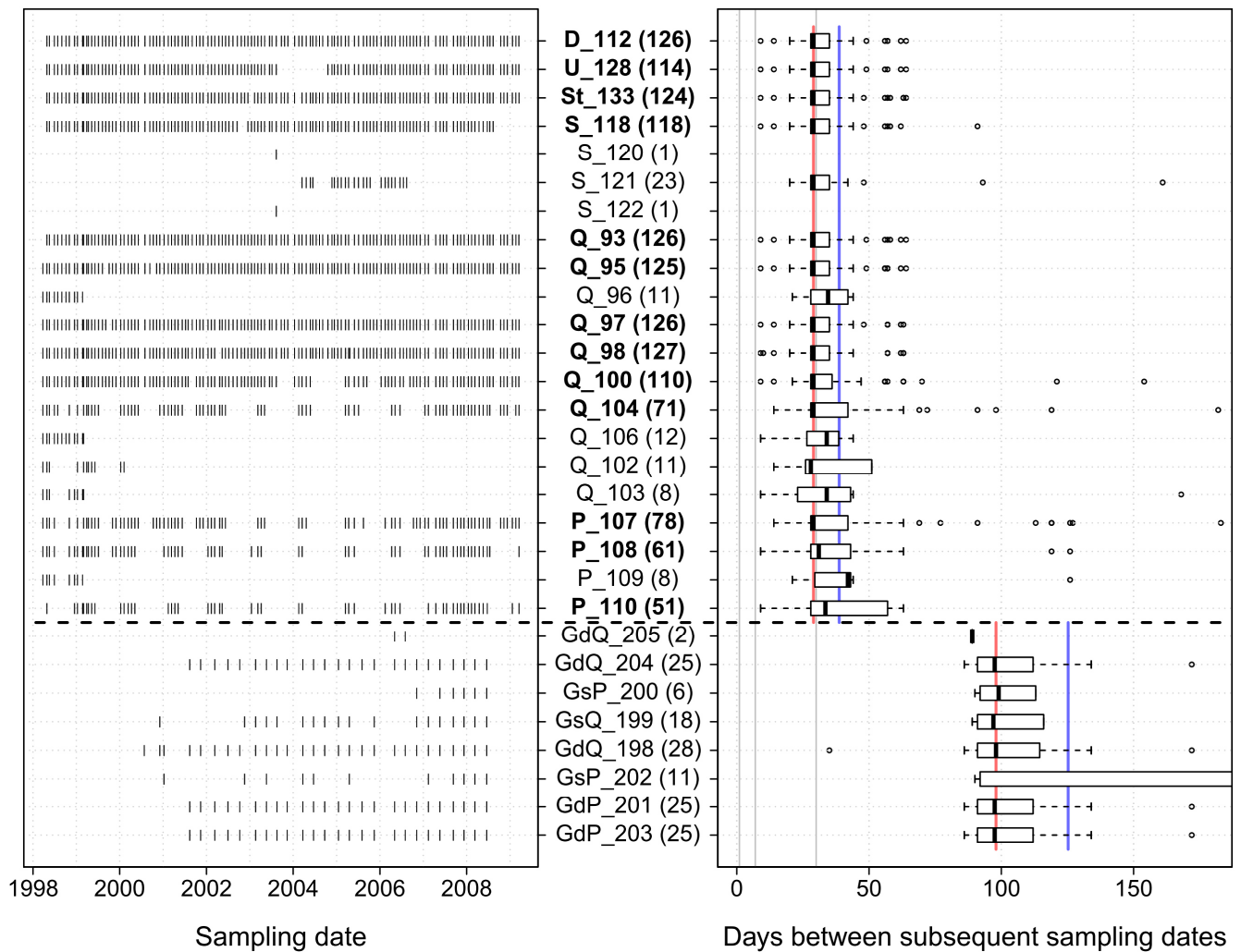


Figure 2 Left panel: Sampling dates at the sites for the whole monitoring period.  
 265 Right panel: Boxplots of the variability of sampling intervals during the monitoring period. For better readability, the maximum of the x-axis is limited to 180 days. Median (redgrey solid line) and mean (bluegrey dashed line) of sampling intervals are shown separately for the groundwater and stream water sites. Grey vertical lines mark the 1-day, 1-week and 1-month interval. Both panels: The dashed horizontal line separates groundwater sites (bottom) from stream water sites (top). Subscripts:  
 270 P = Peege, Q = Quillow, S = Strom, St = Stierngraben, U = Ucker, D = Dauergraben, Gs = shallow groundwater, Gd = deep groundwater. The number of samples at each site is given in brackets. Names of the sites with more than 50 samples are printed bold.

275



### 3 Methods

#### 3.1 Data preprocessing

Missing values were replaced by the median of the respective variable. This concerned at most DOC (3.44% of the values) and  $\text{NO}_2^-$  (2.54%), whereas the percentage of missing values was less than 2% for each of the other 14 variables (Table S3). Values below detection limit were replaced by 0.5 times that limit. To achieve equally weighted variables the values were z-normalized to zero mean and unit standard deviation for each variable separately.

#### 3.2 Exploratory framework

To identify the dominant changes, we firstly used the non-linear dimension reduction technique Isometric Feature Mapping to derive the main multivariate water quality components. To account for the interaction of groundwater and stream water, both groundwater and stream water samples have been analysed together in one joint analysis. Secondly, we studied differences between the sites with respect to median component values. Thirdly, we analysed the time series of the components at sites with more than 50 samples. Seasonal patterns were analysed with the Lomb-Scargle approach (Lomb, 1976; Scargle, 1982; Scargle, 1989) and – if significant – were subtracted from the series prior to trend analyses. Please note that the term ‘seasonal’ refers to the annual cycle throughout the article. Linear trends were estimated with the Theil-Sen estimator and tested for significance with the Mann-Kendall Test. Non-linear trends were depicted with the locally weighted regression (LOESS) approach (Cleveland, 1979; Cleveland and Devlin, 1988). We then related resulting low-frequency patterns to the long-term groundwater head dynamics, likewise determined as LOESS smooth of the de-seasonalised series. Time series analysis at different sites allowed to check whether long-term patterns were consistent, pointing to more general effects in the study area.

As the methods do not require regularly sampled data in space or time, we considered every sample as additional information of the spatiotemporal variability of

305 the observed water quality in the study area rather than noise. Consequently,  
irrespective of irregularities of sampling intervals at a site or differences in sampling  
intervals and numbers of samples between the different sites, we included as many  
samples in the analysis as possible to increase the informative value and support the  
representativeness of the study in space and time. This might lead to a bias in the  
310 determination of the components, as well as in the estimation of the trends of the  
components and their significance, if deviations from a regular sampling scheme  
follow a systematic pattern. To check for that, we tested the distribution of sampling  
intervals at all sites with  $N > 50$  (Table S1) for normality with the Shapiro-Wilk-test  
and the temporal development of the lengths of the sampling intervals for the whole  
315 observation period for monotonic trends with the Mann-Kendall-test. For all tested  
sites a Gaussian distribution of sampling intervals as well as a monotonic trend of the  
length of sampling intervals during the observation period was rejected.

### 3.3 Dimension reduction

320 Dimension reduction methods aim to represent a data set with a given number of  
dimensions (here the number of measured hydrochemical variables) in a new data  
space with substantially less dimensions. This is achieved by projecting the data in a  
new ordination system which makes a more efficient use of the intrinsic structures of  
the data set than the original one. The axes of the new ordination system are usually  
325 called 'components' or 'dimensions'. In the following, we will use the term  
'components'. For the values of a component we will use the term 'scores'. The  
reduction of the dataset's dimensionality is achieved by considering only some of the  
new components for further analysis. The selection process is a trade-off between  
reduction of the dimensionality and minimizing the loss of potentially informative  
330 structures. Typically only the first few components are selected as they depict the  
main structures in the data set.

In the projection, different methods focus on different aspects of the data. For  
example, PCA aims for maximizing variance on the first components, classical  
multidimensional scaling (CMDS) at preserving the interpoint distances of the input

335 data in the projection, and self-organizing maps (SOM) at preserving the  
neighbourhood relations (topology) of the input data in the projection (Lee, 2007). In  
the last years, a variety of non-linear dimension reduction methods has been  
developed (Van der Maaten et al., 2009). Although being sensitive to noisy data,  
Isometric Feature Mapping (Isomap; Tenenbaum et al., 2000) was one of the best  
340 performing approaches when applied to real-world-data (Geng et al., 2005). It has  
been successfully applied in environmental research disciplines, e.g. biodiversity  
studies (Mahecha et al., 2007), soil sciences (Schilli et al., 2010), climatology  
(Gámez et al., 2004), and biogeochemistry (Weyer et al., 2014).

### 345 **3.3.1 Principal component analysis**

In our study, the well-established linear principal component analysis (PCA) served  
as benchmark for the non-linear Isometric Feature Mapping. PCA is one of the most  
widespread dimension reduction methods going back to research of Pearson (1901)  
and Hotelling (1933). For a brief introduction to PCA, please see, e.g., Jolliffe and  
350 Cadima (2016), for a comprehensive one Jolliffe (2002). PCA aims to successively  
maximize the variance of the data set on the new calculated components. The scores  
of the components are calculated as weighted linear combinations of the original  
variables. The weights (loadings) of the linear combination define the axes of the  
data space in which the data is projected. The loadings are the eigenvectors derived  
355 from an eigenvalue decomposition of the covariance matrix of the analysed variables.  
If the analysed variables are z-normalized, as was done here, their covariance matrix  
is equivalent to the (Pearson) correlation matrix. The components are ordered with  
decreasing size of their eigenvalues. The share of variance that is assigned to a  
component is proportional to the size of its eigenvalue in relation to the sum of all  
360 eigenvalues. Thus, the ratio of total variance that is captured by the considered  
components gives a measure of performance of the PCA. PCA was performed in R  
(R Core Team, 2017) with the function 'princomp' of the default package 'stats'.

### 365 3.3.2 Isometric Feature Mapping

Isometric feature mapping (Isomap) is a non-linear extension of CMDS. It aims to approximate the global non-linearity in a dataset by local linear fittings (Geng et al., 2005). This is done by mapping approximated geodesic interpoint-distances to an Euclidean distance matrix via a neighbourhood graph  $G$  (Tenenbaum et al., 2000).  
370 The geodesic distance between two points is the distance along the surface of a (non-linear) manifold, in contrast to the straight-line Euclidean distance (Geng et al., 2005). The neighbourhood graph  $G$  consists of segments that connect every data point to its  $k$  nearest neighbours directly via Euclidean distances. For all non-connected points the shortest path along the neighbourhood graph  $G$  is computed as  
375 the smallest sum of connected segments via the Dijkstra-algorithm (Dijkstra, 1959). This approximation of the geodesic distances allows the adaptation of  $G$  to the global non-linear structures in a data set. The only free parameter  $k$  has to be optimized by checking the performance of several runs. The more linear the data, the higher will the optimum  $k$  be. If  $k$  equals the possible number of connections of one data point to  
380 all other data points, the approximations of the geodesic distances are equal to the Euclidean distances and the Isomap results are congruent to those of CMDS and linear PCA (Gámez et al., 2004). Finally the neighbourhood graph  $G$  is embedded in the Euclidean space.

In contrast to PCA, assessing performance based on the eigenvalues of the  
385 components is not applicable for Isomap. Performance of the dimension reduction of the Isomap approach was assessed and compared to performance of the PCA by the squared Pearson correlation coefficient ( $R^2$ ) of the interpoint distances in the high-dimensional data space and in the low-dimensional projection spanned by selected components (Lischeid and Bittersohl 2008; Lischeid et al., 2010). A perfect fit would  
390 yield a value of 1 and a value of 0 reflects no correlation between the distance matrices of the original data and of the projection. Please note, that with this measure the contribution of single components to the overall performance does not necessarily decrease monotonically with increasing order of the components, as it is the case for the eigenvalue-based performance measure of PCA. For the local  
395 assessment of representation of interpoint distances at the individual sites, only the

data points from the respective sites were used. Because the selection of data points at a site is only a subset of the global data set for which the dimension reduction was performed, the performances regarding the representation of interpoint distances differ between the individual sites as well as compared to the overall performance for the global data set. At some sites it can even happen that adding more components does not for every component improve the representation of interpoint distances in the low-dimensional projection. Isomap and the determination of the distance matrices were performed with the R-package 'vegan' (Oksanen et al., 2009).

### 405 3.3.3 Interpretation of components

The analysis focused on those components that explained a major fraction of the total interpoint distances. The considered components were regarded to reflect dominant drivers influencing water quality. Here, the term 'driver' was used for biogeochemical and hydrological processes as well as for anthropogenic influences affecting water quality. Correspondingly we formulated a hypothesis for each considered component. The interpretation of the components is based on analysing (i) the correlations between measured variables and component scores as well as (ii) spatial and temporal patterns of the scores.

Correlation between scores of a selected component  $cp_x$  and values of single variables might be blurred due to the effects of other components on the same variable. We excluded those effects by analysing the relationships between scores of the selected component  $cp_x$  and the residuals of the multiple linear regression mlr of the single variable  $v_i$  at hand and the remaining other considered components  $CP \setminus x$  (residuals):

$$420 \quad cor(cp_x, residuals[mlr(v_i, CP \setminus x)]) , \quad (1)$$

where  $CP \setminus x$  is the set of  $m$  considered components, without the selected component  $cp_x$ ,  $\beta_0$  and  $\beta_j$  the intercept and coefficients of the regression

$$mlr(v_i, CP \setminus x) = v_i = \beta_0 + \sum_{j=1}^m \beta_j cp_j + residuals \quad (2)$$

$$mlr(v_i, CP \setminus x) = \beta_0 + \sum_{j \in \{CP \setminus x\}} \beta_j cp_j + residuals \quad (2)$$

425 To assess the relationships between components and residuals we used bivariate  
scatterplots. To summarize the relationships between components and residuals we  
used Spearman rank correlation, which enables to consider non-linear relationships  
as well, as long as they are monotonic. Besides, it is less sensitive to extreme values  
than Pearson correlation.~~As a measure for monotonic but not necessarily linear  
430 correlation we used Spearman rank correlation.~~

### 3.4 Time series analysis

At sites with more than 50 samples, time series of component scores were  
analysed for seasonal patterns, linear trends and non-linear trends. The sites were  
435 compared with respect to the identified long-term patterns to detect general patterns  
in the study area. The significance level for trend and frequencies in this study was  
set to  $p \leq 0.05$ . At each site, the fractions of variance of a time series that were  
assigned to its seasonal pattern, linear trend or non-linear trend were determined as  
the  $R^2$  of the respective pattern with the component series. In case of significant  
440 seasonal patterns, the estimations of the trends were based on the de-seasonalised  
series. Accordingly, the fractions of variance assigned to the trends were determined  
as the  $R^2$  of the trend pattern with the de-seasonalised series. The decomposition of  
the time series in a seasonal component and a non-linear trend derived with LOESS  
was inspired by the STL-approach of Cleveland et al. (1990).

445

#### 3.4.1 Lomb-Scargle method

Standard Fourier analysis requires equidistant time series which was not given in our  
study. Therefore the estimation of seasonal patterns in the time series was done with  
the Lomb-Scargle method, which is an extension of Fourier-Analysis to the uneven-  
450 spaced case genuinely invented in astrophysics (Lomb, 1976; Scargle, 1982). The  
application of the Lomb-Scargle method in this study follows to a large extent the  
workflow suggested by Glynn et al. (2006) as well as Hocke and Kämpfer (2009).

Details are given in the Appendix A. The implementation used in this manuscript can  
be accessed as R-script at [doi:  
10.4228/ZALF.2017.340](https://doi.org/10.4228/ZALF.2017.340)[http://open-research-data.ext.zalf.de/ResearchData/2017\\_340.html](http://open-research-data.ext.zalf.de/ResearchData/2017_340.html).

### 3.4.2 Theil-Sen estimator and Mann-Kendall test

The linear trend was estimated with the non-parametric Theil-Sen estimator which  
is the median of all interpoint slopes in a time series (Theil, 1950; Sen, 1968). The  
Mann-Kendall test (Mann, 1945; Kendall, 1990) was used to test for significant  
monotonic trends. Identified trends are not necessarily linear. Being based on rank  
correlation, data do not have to obey any specific distribution. Please note that we  
did not account for the effect of overestimation of the significance of trends with the  
Mann-Kendall test due to short-term autocorrelation (Yue et al., 2002). That would  
have required an assessment of the lag-1 autocorrelation which was hampered by  
the irregular sampling. Neither did we consider long-term memory and its effects on  
the statistical significance of the trends (Cohn and Lins 2005; Zhang et al., 2018).  
Consequently, we did not consider the possible effects of the irregular sampling on  
the long-term memory (fractal scaling) of the water quality series either (Zhang et al.,  
2018). Due to the limited number of samples per year and non-equidistant sampling,  
the seasonal Mann-Kendall test was not applicable (Figure 2). Instead, significant  
seasonal patterns according to the Lomb-Scargle approach were subtracted prior  
trend analysis. The Mann-Kendall test was performed with the R-package 'Kendall'  
(McLeod, 2011).

### 3.4.3 Locally weighted regression (LOESS)

We assessed non-linear trends and low-frequency patterns with locally weighted  
regression (LOESS; Cleveland, 1979; Cleveland and Devlin, 1988), where the  
smoothing is done by local fitting of a second order polynomial to each point  $x$  in the  
data set using weighted least squares. The weights for each value to be fitted are

scaled to the range from 0 to 1 by the distance  $d(x)$  between  $x$  and its  $q^{\text{th}}$  closest point. The ratio of  $q$  to the number  $n$  of all data points, i.e. the span of the local regression smoother, defines the degree of smoothing. We used the default  
485 smoothing span which is a proportion of  $q/n = 0.75$  of  $x$ 's nearest neighbours. Data points further away than the  $q^{\text{th}}$  data point do not contribute to the regression. Within the range of the span, the weights  $w_i$  of the neighbouring points  $x_i$  in the least square fit decrease with increasing distance of  $x_i$  to  $x$  symmetrically around  $x$  according to the tricubic weighting function  $w_i(x) = (1 - [ (|x_i - x|) / d(x) ]^3)^3$ . Again, significant  
490 seasonal patterns according to the Lomb-Scargle approach were subtracted prior trend analysis. For details about choosing different LOESS-parametrisations, please see Cleveland (1979) as well as Cleveland and Devlin (1988). Local extrema of the LOESS smooth were identified with the R-package 'EMD' (Kim and Oh, 2009; 2014.).

## 495 4 Results

### 4.1 Multivariate components

We achieved the best performance of the Isomap dimension reduction with  $k = 1300$  (Table 2). In the following, results are presented for the first four Isomap components representing 88% of the interpoint distances of the total data set. For  
500 single sites (with more than 15 samples), between 29 and 97 % of the respective interpoint distances were represented ([Table S4Table 2](#)).

The 1<sup>st</sup> component depicted [424%](#) of the interpoint distances of the total data set. Plotting residuals of the variables versus the 1<sup>st</sup> component showed strong positive correlations for  $\text{NO}_3^-$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Mg}^{2+}$ ,  $\text{Ca}^{2+}$ ,  $\text{Cl}^-$ , EC,  $\text{SO}_4^{2-}$ , DOC and slightly less, but  
505 still positive, correlations for  $\text{O}_2$  and Eh. Temperature was the only variable correlating negatively with the 1<sup>st</sup> component (Figure 3). Visualization of the component scores versus residuals of solute concentration revealed predominantly linear relationships (Figure S1).

The 2<sup>nd</sup> component reflected [189%](#) of the interpoint distances in the data. It  
510 exhibited clear positive correlation with  $\text{O}_2$  concentration, pH and Eh, and weaker



correlation with  $\text{Na}^+$ ,  $\text{K}^+$  and DOC. It was inversely correlated with  $\text{Ca}^{2+}$ , EC and  $\text{SO}_4^{2-}$  (Figure 3 and Figure S2). In the groundwater samples,  $\text{HCO}_3^-$  and  $\text{Fe}^{2+}$  had been determined as well. Both solutes were negatively correlated with this component (Figure 4 upper panel).  $\text{NO}_3^-$  concentration in the deep groundwater samples was very low (with 27% of the samples below detection limit) and did not show any clear correlation with the 2<sup>nd</sup> component. Low component scores in the groundwater came along with high  $\text{Ca}^{2+}$  and  $\text{HCO}_3^-$  concentration.

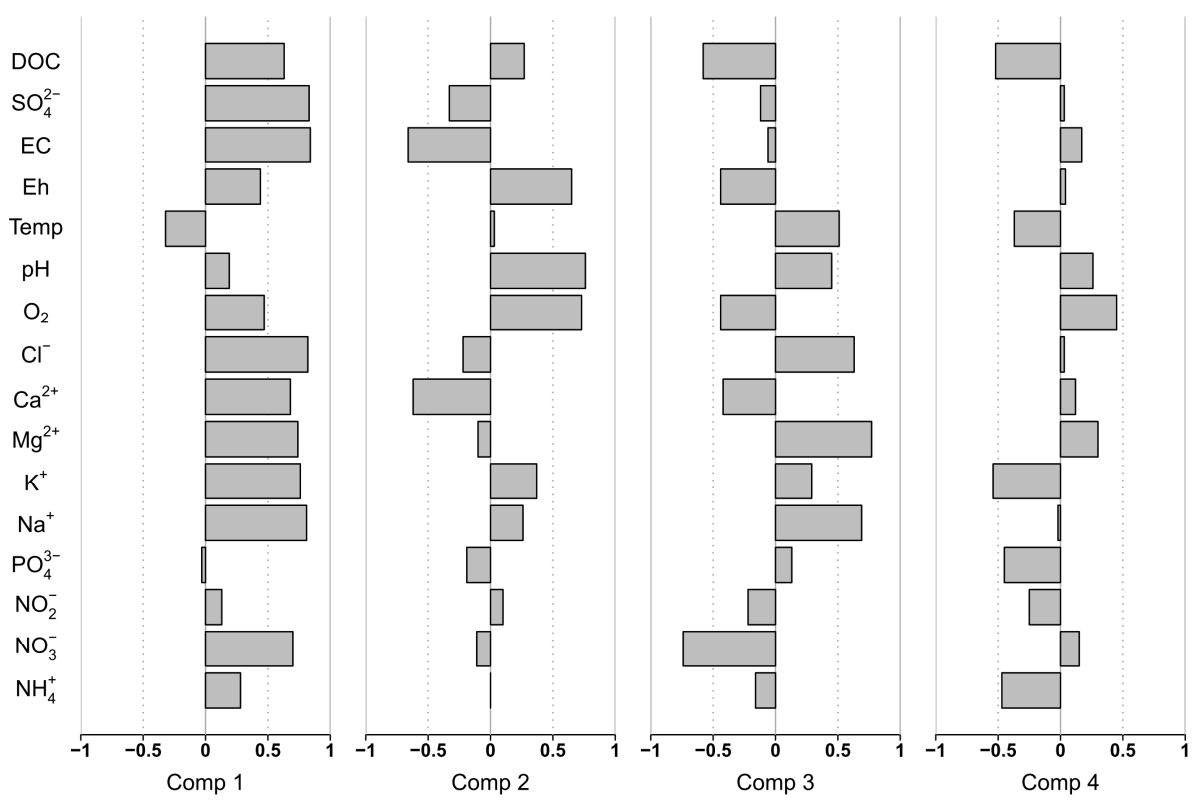
The relationship of scores of component one and two in the groundwater is shown in the lower panel of Figure 4. Except for the two shallow wells close to the Peege stream (Gs\_200, Gs\_202; cf. Figure 1) scores of the 1<sup>st</sup> and 2<sup>nd</sup> component are negatively related (Figure 4 lower panel).

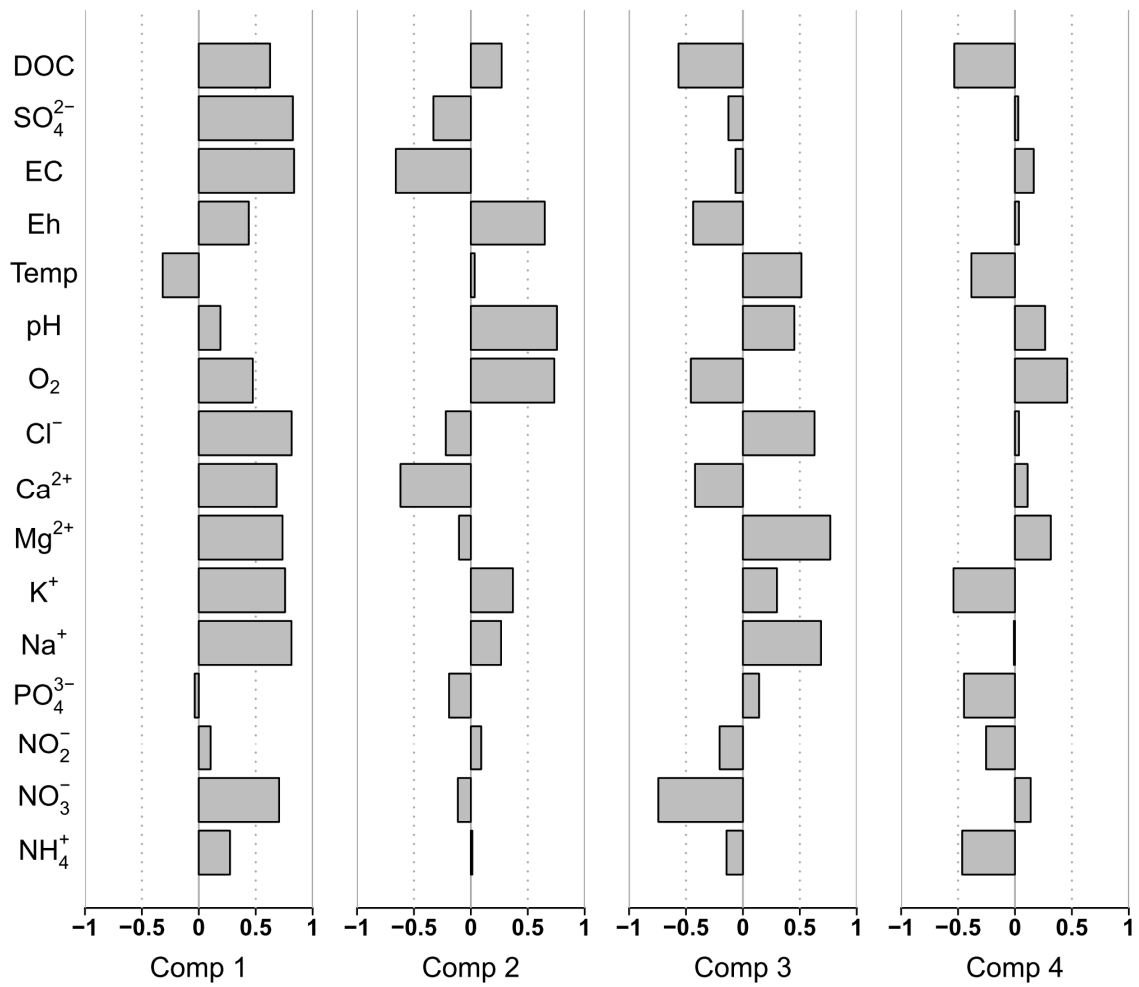
The 3<sup>rd</sup> component represented 6% of the interpoint distances in the data set. The residuals exhibited positive correlation for  $\text{Na}^+$ ,  $\text{Mg}^{2+}$ ,  $\text{Cl}^-$ , pH and temperature. Negative correlations were found for  $\text{NO}_3^-$ ,  $\text{Ca}^{2+}$ ,  $\text{O}_2$ , Eh, and DOC (Figure 3 and Figure S3-).

Another 22% of the interpoint distances in the data were assigned to the 4<sup>th</sup> component. Residuals of the component scores showed negative correlation for  $\text{NH}_4^+$ ,  $\text{PO}_4^{3-}$ ,  $\text{K}^+$ , temperature, and DOC and positive correlation for  $\text{O}_2$  (Figure 3 and Figure S4). The range of component values was spanned mainly by single large values of  $\text{NH}_4^+$ ,  $\text{PO}_4^{3-}$ , and  $\text{K}^+$  that cannot be explained with the preceding three components (Figure S4). This highlights the importance of particular events for the 4<sup>th</sup> component.

Table 2 Cumulated  $R^2$  of the reproduction of the interpoint distances of the data in the projection by the first ten components of the best Isomap run and linear PCA.

Component	1	2	3	4	5	6	7	8	9	10
Isomap	0.42	0.6	0.66	0.88	0.94	0.96	0.97	0.98	0.98	0.99
PCA	0.39	0.57	0.65	0.88	0.94	0.95	0.97	0.98	0.99	0.99





540 | Figure 3 Spearman-rank-correlation of a component and the residuals of the multiple linear regression of the measured variable and the remaining three other components.

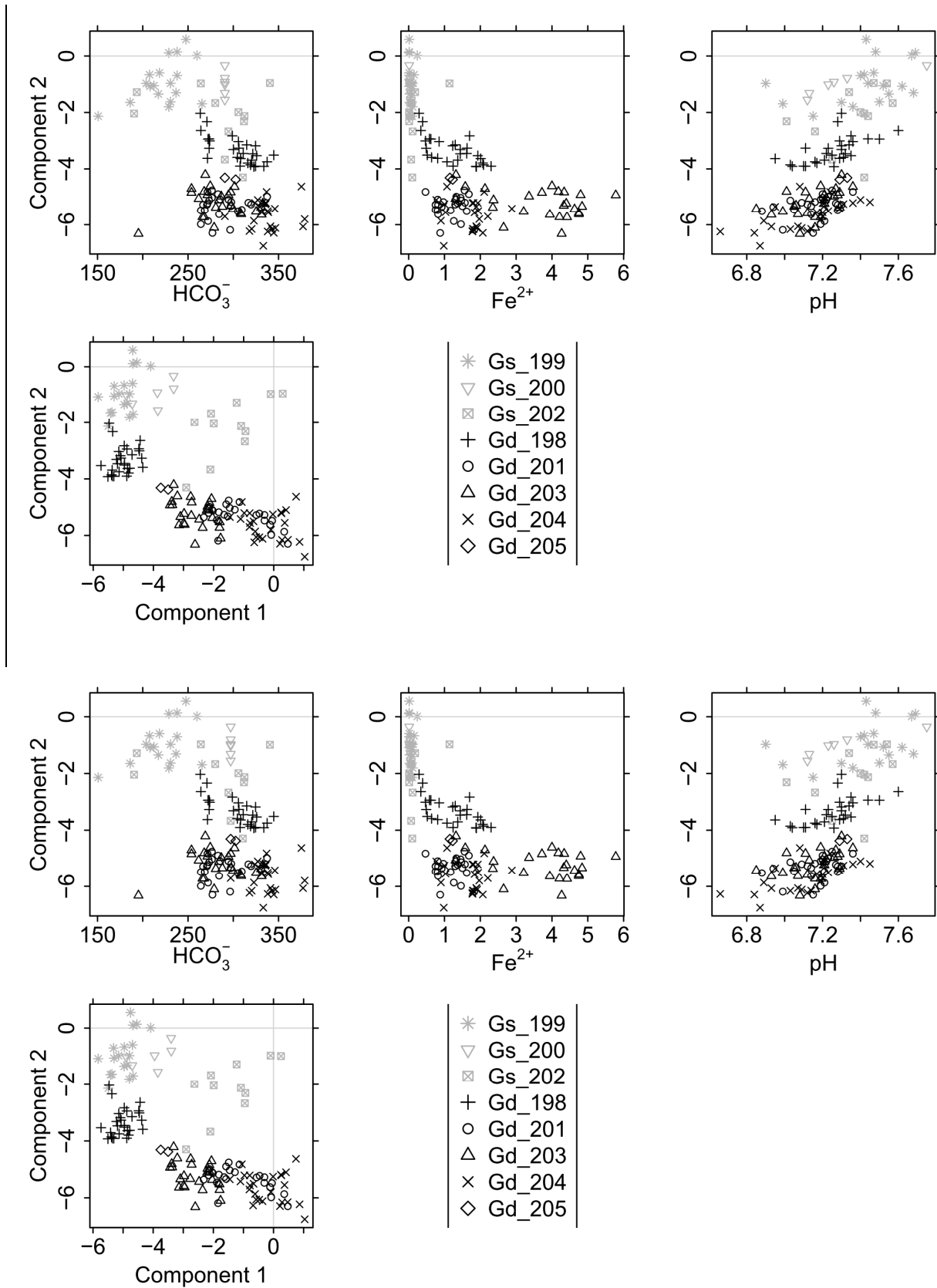


Figure 4 Upper panel: Selection of variables vs. scores of component 2 for the

545 groundwater samples. Concentration in  $\text{mgL}^{-1}$ . Lower panel: Scores of component 1  
vs. component 2 at the groundwater sites.

## 4.2 Multiple sites

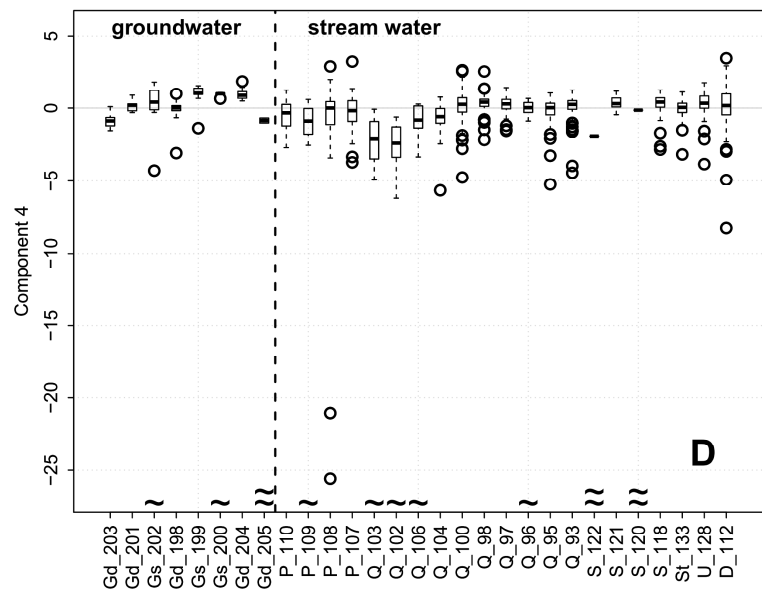
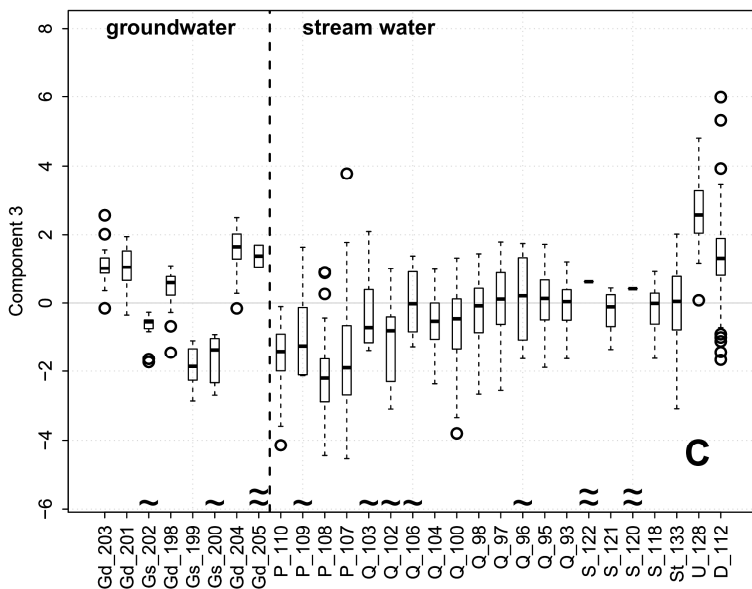
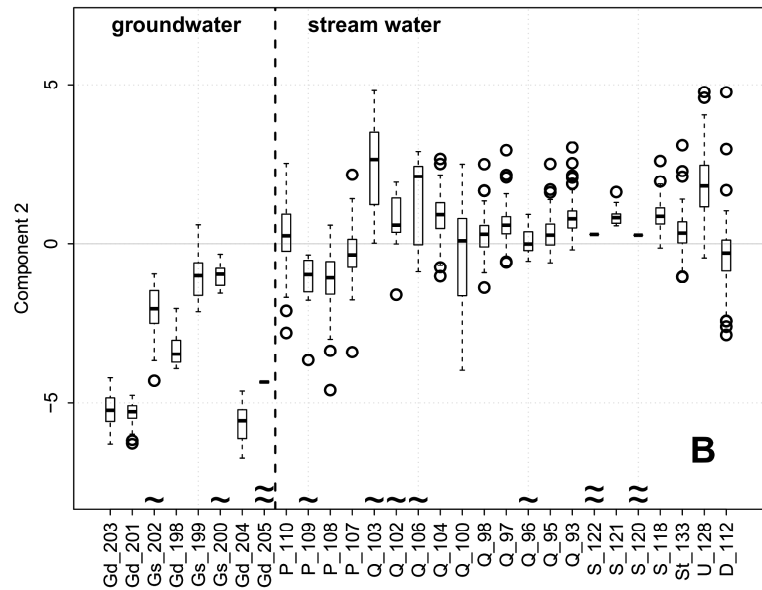
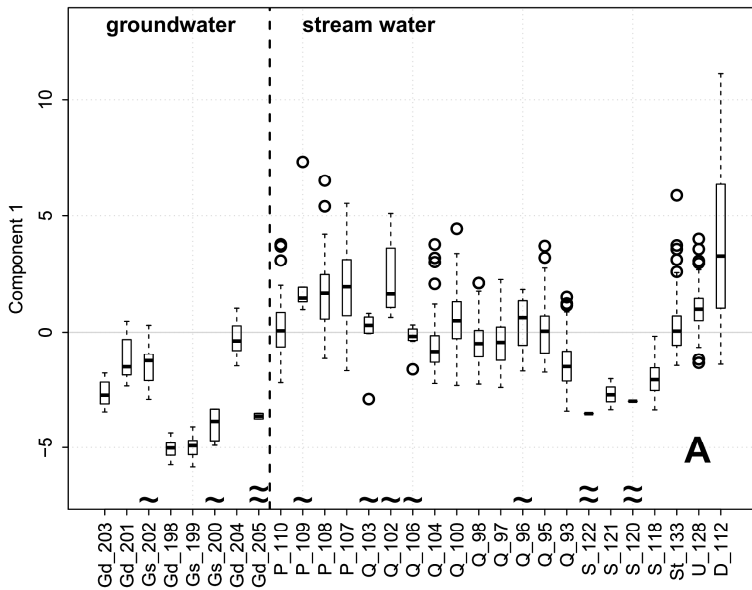
550 Median values of scores of the 1<sup>st</sup> component clearly differed between streams  
(Figure 5 A). At the Strom sites, the median score values were considerably lower  
than those from the other stream water sampling sites. The median values of scores  
of the sites at the Quillow and Stierngraben showed intermediate values followed by  
the Ucker site, the Peege sites and finally the Dauergraben with the highest median  
score value. Groundwater samples in general exhibited consistently low scores of the  
555 1<sup>st</sup> component, but without clear differences between deep and shallow groundwater  
samples. Mixing of water from different streams was visible at site Q\_93 downstream  
the confluence of the Quillow (Q\_95) and of the Strom stream (S\_118), as well as at  
site Q\_100 downstream the confluence of Q\_104, Q\_102 and P\_107 (Figure 1 and  
Figure 5 A).

560 Stream water samples exhibited the highest scores of the 2<sup>nd</sup> component, whereas  
low scores were limited to deep groundwater samples, and shallow groundwater  
samples were in an intermediate position (Figure 5 B). Median values of the stream  
water sites were approximately on the same level except for the sites Q\_103, Q\_106  
and U\_128 which exhibited noticeably higher median values than the other stream  
565 water sites and the two Peege sites P\_109 and P\_108, which exhibited median  
values on the same level as the shallow groundwater sites Gs\_199 and G\_200. The  
scores in the deep groundwater clearly showed the largest absolute values,  
indicating the significance of deep groundwater for this component (Figure 5 B).

570 Scores of the 3<sup>rd</sup> component in the deep groundwater were consistently higher  
than in shallow groundwater, while the stream water samples covered the whole  
range of values (Figure 5 C). Lowest scores of the 3<sup>rd</sup> component were found at the  
Peege sites and in the shallow groundwater, highest scores at Ucker, Dauergraben  
and the deep groundwater. At the Quillow stream, scores tended to increase from the  
spring to the outlet. The effect of mixing of tributaries with different water qualities

575 was visible along the course of the Peege and Quillow streams downstream of the  
 respective confluences at the sites P\_108, Q\_95 and Q\_93 (Figure 1 and Figure 5  
 C).

The range of values of the 4<sup>th</sup> component was strongly biased towards negative  
 values, caused by single events at some sites which exhibited very low values  
 580 (Figure 5 D).



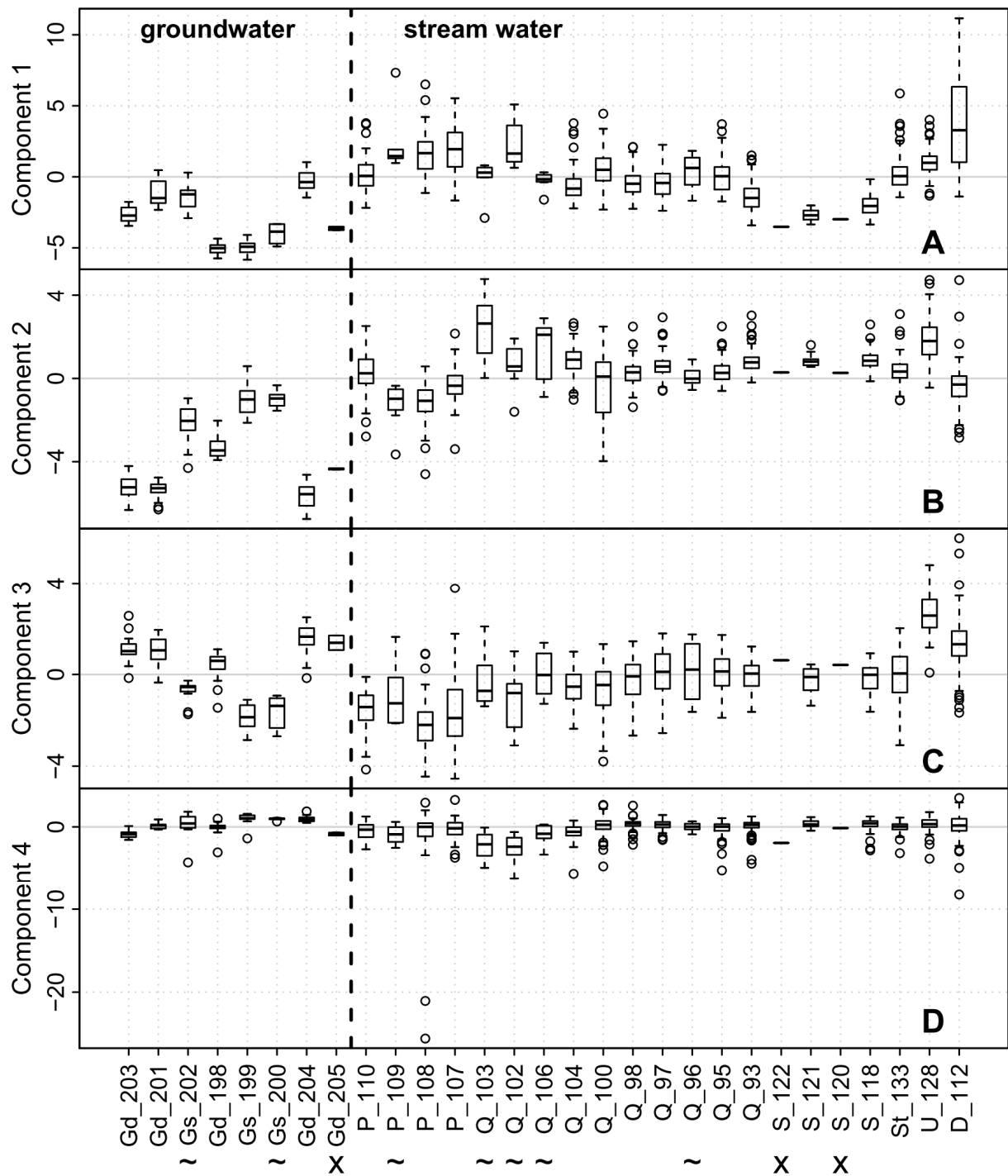


Figure 5 Boxplots of scores of component 1 to 4 at different sites. Sites with  $n < 13$  are marked with '~', those with  $n < 3$  with 'X'. Subscripts: P = Peege, Q = Quillow, S = Strom, St = Stierngraben, U = Ucker, D = Dauergraben, Gs = shallow groundwater, Gd = deep groundwater.

585

### 4.3 Long-term patterns

Time series of scores of the components were studied at sites with more than 50  
590 temporal replicates. This applied for 13 stream water sites (Table S1). All dominant  
frequencies (for details, please see Appendix A) interpreted as seasonal patterns had  
a period length in the range between 350 and 380 days. For de-seasonalisation  
these seasonal patterns were subtracted from the time series prior to analysis for  
linear and non-linear trends.

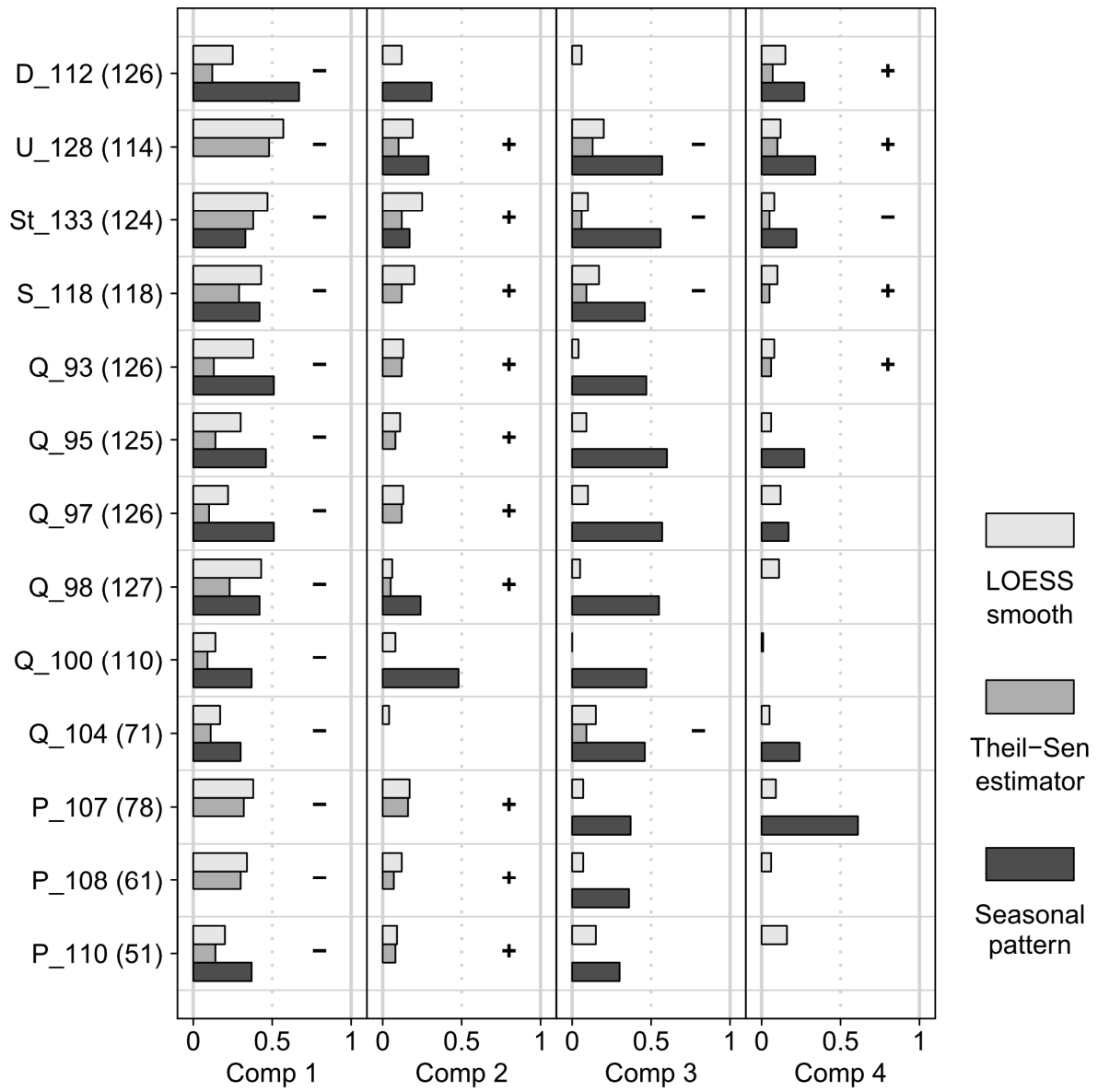
595 Most of the time series of the scores of the 1<sup>st</sup> component exhibited clear seasonal  
patterns with maximum scores during the winter season (Figure 6 and Figure 7).  
Between 30 and 67 % of the variance were assigned to the seasonal pattern. At all  
sites we found significant negative monotonic trends (Figure 6). The strongest  
decline was found at site D\_112, the weakest trend at site Q\_97 (not shown). The  
600 linear trend comprised between 9 and 48 % of the variance of the de-seasonalised  
time series (Figure 6). In contrast, the LOESS smooth depicted 14 to 57 % of the  
variance (Figure 6). It showed a decrease until December 2004 approximately and an  
increase thereafter (Figure 8). The de-seasonalised time series of groundwater  
heads showed a similar behaviour, with the minimum water level in June 2006  
605 (Figure 8). Timing of the minimum values of the scores of the 1<sup>st</sup> component varied  
between sites, spanning a range from 17<sup>th</sup> February 2004 to 17<sup>th</sup> of March 2009  
(Figure 8). As an example, Figure 7 gives the time series of scores of the 1<sup>st</sup>  
component at site Q\_93, the seasonal pattern extracted from the series and the de-  
seasonalised time series with the non-linear trend estimated with the LOESS  
610 smoother.

Unlike for the 1<sup>st</sup> component, only five of the thirteen considered time series of the  
2<sup>nd</sup> component exhibited a clear significant seasonal pattern, accounting for 17 to 48  
% of variance (Figure 6). The maxima of the seasonal patterns of the sites at Quillow  
and Ucker were in spring, at Stierngraben and Dauergraben in summer. In contrast,  
615 significant monotonic trends were found at most of the stream water sampling sites.  
All significant trends of the 2<sup>nd</sup> component were positive. The linear trend comprised  
between 5 and 16 % of the variance of the time series, while the LOESS smooth  
comprised between 4 and 25 %.



620 Values of the 3<sup>rd</sup> component showed a clear seasonal pattern with maxima in  
summer (Figure 6). Between 30 and 60 % of the variance were assigned to the  
seasonal signal. The only exception was site D\_112 where the seasonal pattern was  
distorted by strong maxima in the winters of 2003, 2004 and 2007. Only at four sites  
significant linear trends were found. All of them were negative, comprising between 6  
and 13 % of the variance. The LOESS smooth depicted between 0 and 21 % of the  
625 variance.

For the 4<sup>th</sup> component, significant seasonal patterns with maxima in summer were  
observed at 7 of the 13 analysed series, comprising between 17 and 61 % of the  
variance (Figure 6). Five sites showed a significant monotonic trend, comprising  
between 5 and 10 % of the variance. A negative trend was observed at site St\_133  
630 only. Four sites showed a positive trend. The LOESS smooth depicted between 1  
and 16 % of the variance.



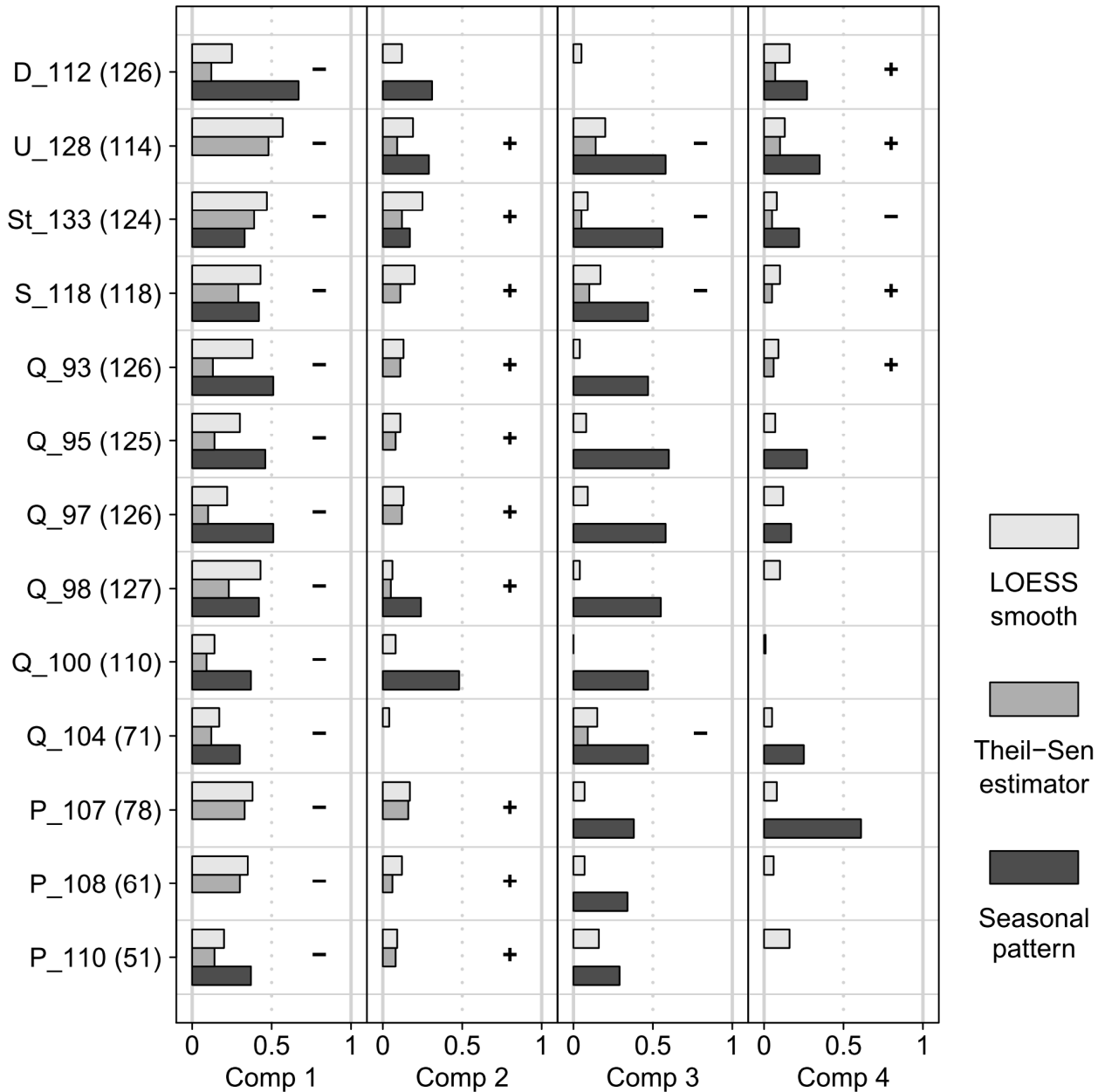
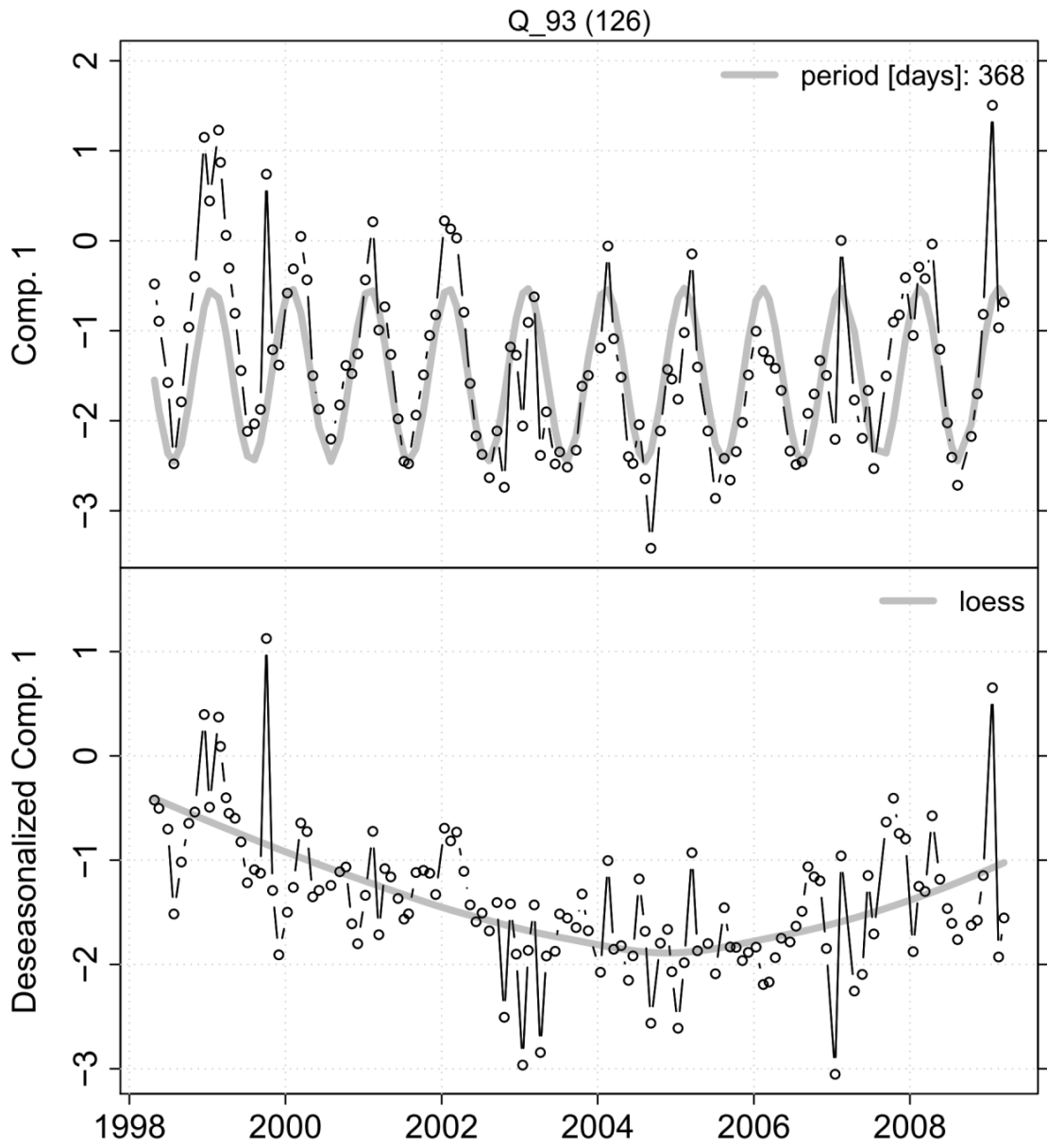


Figure 6 Fraction of variance of the time series of the Isomap component scores of sites with  $n > 50$  assigned to the seasonal pattern (dark grey) and the trend estimated by the linear Theil-Sen estimator (mid grey) as well as the non-linear LOESS smooth (light grey). Fraction of variance is derived as  $R^2$  of the scores of the respective component with the seasonal pattern or the estimated trend. Only significant seasonal patterns and linear trends are shown. The sign of the linear Theil-Sen estimator is given in the respective line. The number of samples at each site is given in brackets.



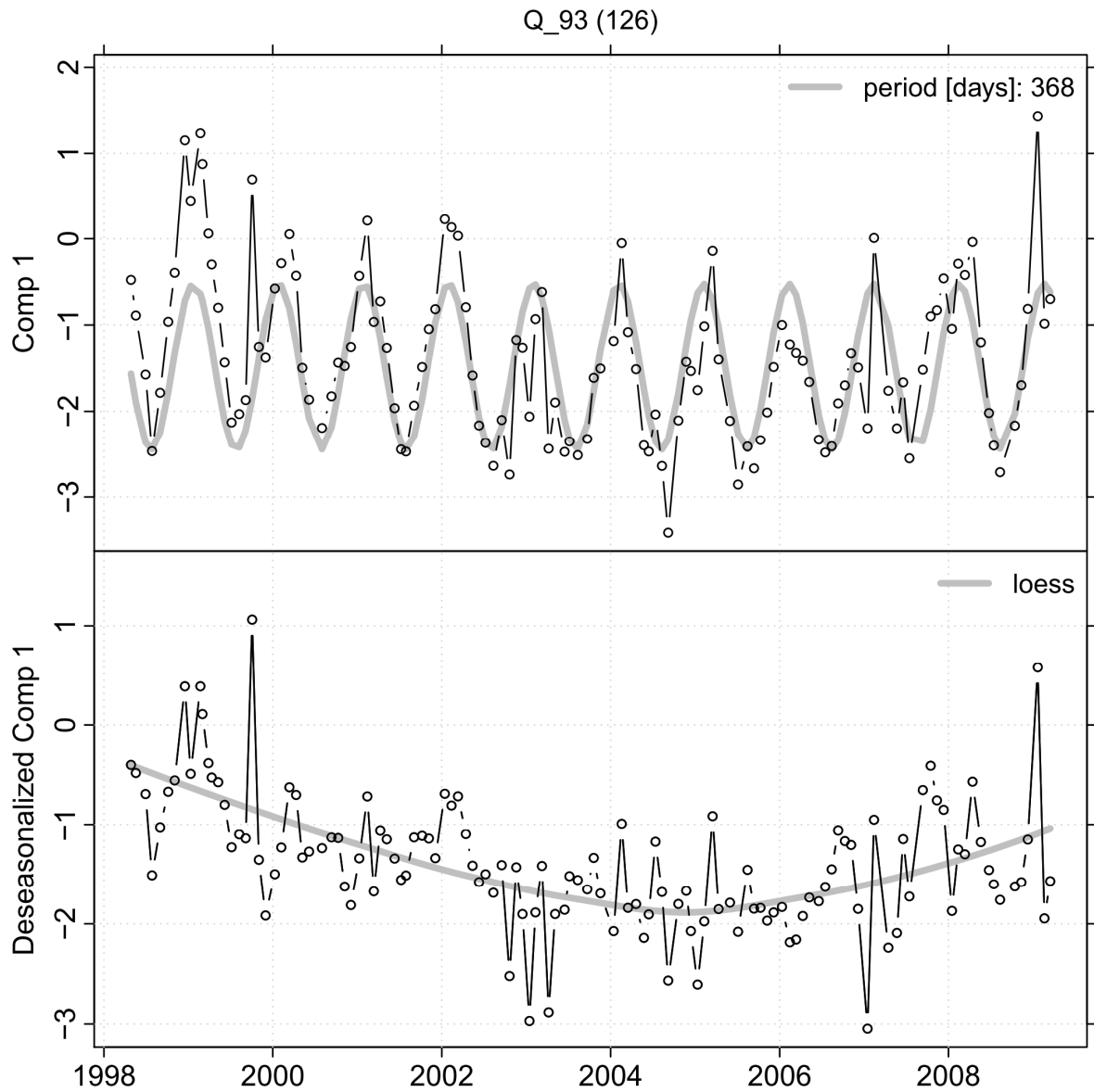
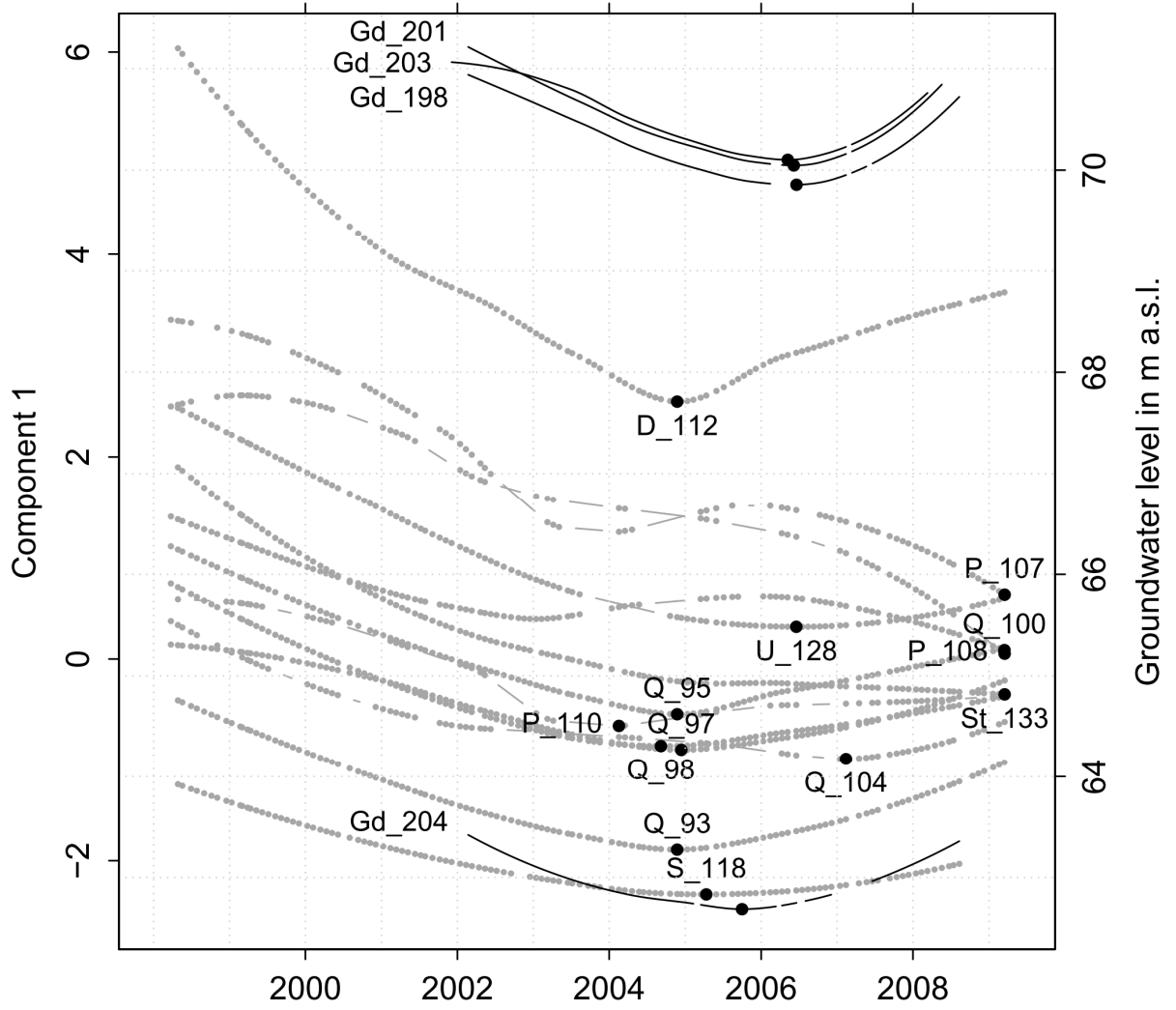
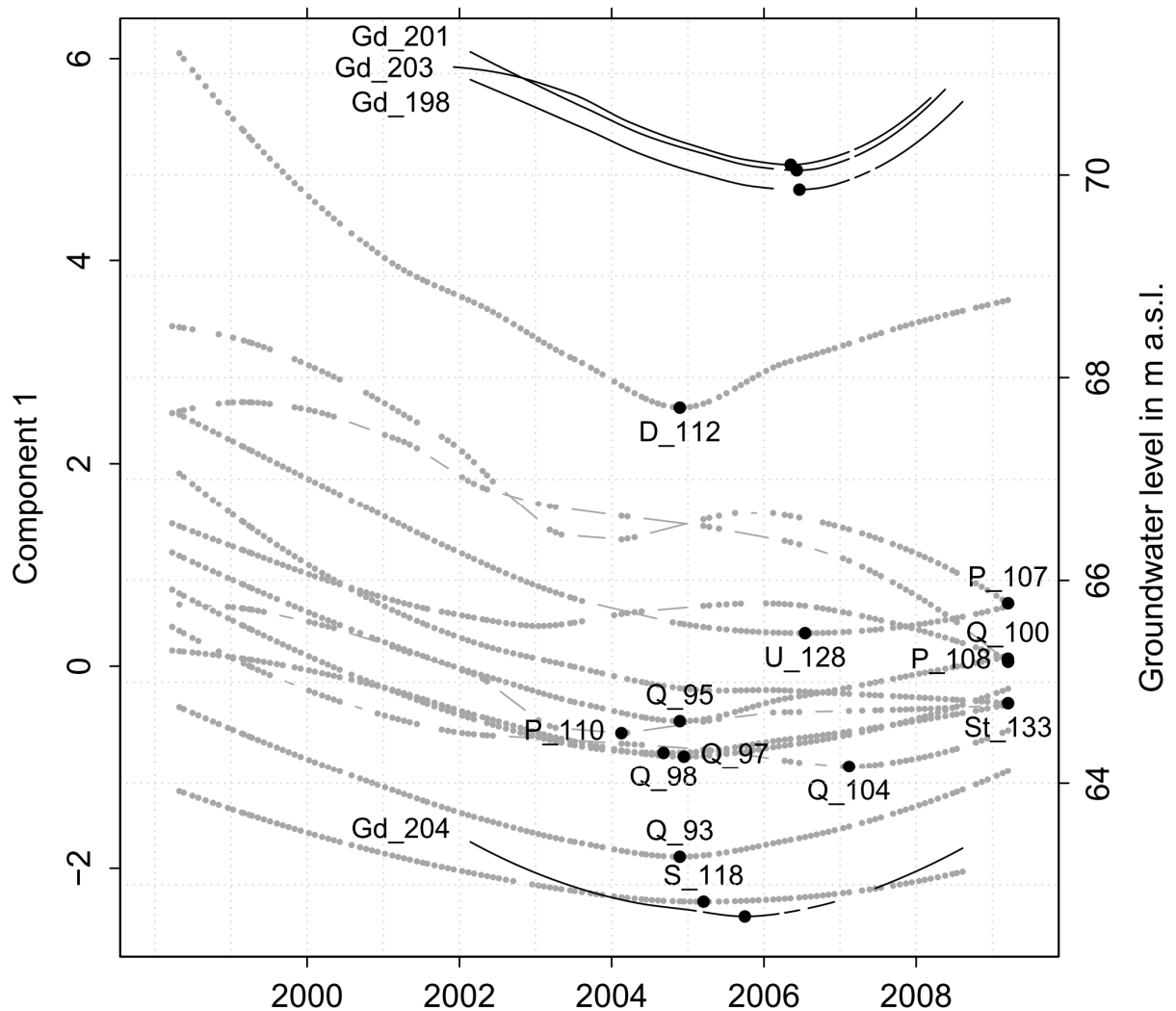


Figure 7 Upper panel: Time series of scores of the 1<sup>st</sup> component at site Q\_93 in  
 645 black and the seasonal pattern estimated with Lomb-Scargle in grey. Lower panel:  
 The de-seasonalised series in black and the non-linear trend estimated with LOESS  
 in grey. The number of samples is given in brackets.





650 Figure 8 Left y-axis: LOESS smooth of time series of the 1<sup>st</sup> component at sites with  $n > 50$  in grey. If a significant seasonal pattern was present, this was removed before smoothing. Right y-axis: LOESS smooth of the de-seasonalised groundwater level at four sites in black. The black dots mark the minima of the LOESS-smoothed series.

655

## 5 Discussion

### 5.1 Multivariate components

Non-linear Isomap performed in this study only slightly better with respect to the

representation of interpoint distances than PCA (Table 2), suggesting that mainly linear relationships were of importance for the overall dynamics in the data set. As there were only minor differences, we will present in the following the results of Isomap only.

For PCA and Isomap, the 1<sup>st</sup> component represents by definition the correlation structure that predominantly can be extracted from the set of variables as a whole. If all the loadings of the 1<sup>st</sup> component of a PCA have the same sign, it is a weighted average of all the analysed variables (Jolliffe, 2002; Jolliffe and Cadima, 2016). The stronger the analysed variables are linearly correlated, the more the 1<sup>st</sup> component approximates the arithmetic mean of all variables (for examples with hydrometric data see Lischeid, et al., 2010; Lehr et al., 2015). Furthermore, the 1<sup>st</sup> component serves as reference for all the subsequent components.

In this study each sample of the multivariate water quality data set is uniquely defined by a sampling site and a sampling date. Thus, the 1<sup>st</sup> component depicted a) for each sampling site the pattern that was most prominent in the time series of the variables correlating with the 1<sup>st</sup> component, and b) between the sampling sites the difference in concentration level of those variables. High values of the 1<sup>st</sup> component indicate synchronous appearance of relatively high Eh and EC together with relatively high concentration of NO<sub>3</sub><sup>-</sup>, Cl<sup>-</sup>, SO<sub>4</sub><sup>2-</sup>, Na<sup>+</sup>, K<sup>+</sup>, Mg<sup>2+</sup>, Ca<sup>2+</sup>, DOC, O<sub>2</sub> accompanied with relatively low temperature (Figure 3).

The whole study region is characterized by relatively intense agriculture (Table 1).

Thus, in addition to the natural background, we assume a general effect of the agricultural practice on the solute concentration level and the dynamics of the water quality series in the area. Enhanced concentration of NO<sub>3</sub><sup>-</sup>, Cl<sup>-</sup>, SO<sub>4</sub><sup>2-</sup> and Ca<sup>2+</sup> is typical for groundwater and stream water in regions with intense agriculture compared to forested areas (Broers and van der Grift, 2004; Fitzpatrick et al., 2007; Lischeid and Kalettka, 2012). Nitrogen and potassium are the main ingredients of mineral fertilizers. Cl<sup>-</sup> and SO<sub>4</sub><sup>2-</sup> are the dominating anions in potassium fertilizers. SO<sub>4</sub><sup>2-</sup> is a major ingredient of phosphorus fertilizers and ingredient in some nitrogen fertilizers. Calcite is present in some nitrogen fertilizers or is applied separately via liming. DOC might be leached from slurry application or via tile drains after



690 mechanical destruction of topsoil aggregates during tillage (Graeber et al., 2012). In addition, cations from the soil matrix might be leached by an enhanced anion concentration (mainly  $\text{NO}_3^-$ ) (Jessen et al., 2017). Overall the application of fertilizers and other agricultural practices like tillage tend to enhance the solute concentration of seepage water (Pierson-Wickmann et al., 2009). Thus, we interpreted the 1<sup>st</sup>  
695 component as the enhancement of the natural background level of solute concentration due to agricultural practices.

Compared to the 1<sup>st</sup> component, the relationships of the 2<sup>nd</sup> component with Eh, pH and  $\text{O}_2$  concentration were clearer expressed (Figure 3 and Figure S2). The range of the scores of the 2<sup>nd</sup> component was spanned by the lowest values in the  
700 deep groundwater and the highest values in the stream water (Figure 5 B) whereas shallow groundwater exhibited intermediate scores. This sequence corresponds to redox conditions expected in those water categories. Thus, we interpreted the 2<sup>nd</sup> component as a redox controlled component covering a sequence from reducing conditions in deep groundwater to post oxic conditions in shallow groundwater and  
705 oxic conditions in stream water.  $\text{O}_2$  and  $\text{NO}_3^-$  concentration in deep groundwater samples usually was below the detection limit which is a common feature in this region (Merz et al., 2009).  $\text{NO}_3^-$  in seepage and groundwater can be denitrified by microorganisms which use the oxidation of sulphides to sulphate as electron donor for denitrification (Massmann et al., 2003, Jørgensen et al., 2009). We ascribed the  
710 high  $\text{SO}_4^{2-}$  and  $\text{Fe}^{2+}$  concentration to oxidation of pyrite (Figure 4 upper panel and Figure S2). Pyrite and other sulphides are abundant in the Pleistocenic sediments of North Germany (e.g. Weymann et al., 2010). Consequently, the pH decreases, calcite gets dissolved and the  $\text{HCO}_3^-$  concentration increases. Part of the released  $\text{Ca}^{2+}$  replaces  $\text{Na}^+$  and  $\text{K}^+$  being sorbed to clay minerals.

715 We interpreted the clear separation in the 3<sup>rd</sup> component between relatively low scores for the shallow aquifer and relatively high scores for the deep aquifer as reflection of two opposing gradients (Figure 5 C). High concentration of  $\text{NO}_3^-$ ,  $\text{O}_2$  and DOC and relatively high Eh values being negatively related to the 3<sup>rd</sup> component (Figure 3) is indicative for groundwater close to the surface, whereas enhanced  
720 concentration of the positively related solutes  $\text{Na}^+$ ,  $\text{Mg}^{2+}$  and  $\text{Cl}^-$  is characteristic for

725 local upwelling of saline groundwater from the underlying Tertiary aquifers at greater depth (Hannemann and Schirrmeister 1998; Tesmer et al., 2007). The scores of the stream water samples, in turn, reflect the mixing ratio of groundwater from the two aquifers to the streamflow. We expect the baseflow maintained from the deep aquifer to be relatively enriched with geogenic solutes compared to the water that stems from the shallow aquifer or faster responding flow components like tile drain discharge and surface runoff. Water from the shallow aquifer is expected to be relatively enriched with surface born solutes compared to water that stems from the deep aquifer.

730 The range of values of the 4<sup>th</sup> component was dominated by single extremely low scores, reflecting samples with high concentration of  $\text{NH}_4^+$ ,  $\text{PO}_4^{3-}$ , and  $\text{K}^+$  (Figure S4).

The catchments of the analysed streams are only sparsely populated and mainly characterized by intensive agriculture (Table 1). In agricultural landscapes slurry is a typical source in which those nutrients occur in high concentration (Hooda et al., 2000). We are not aware of any other high-concentration sources of this combination of nutrients in the region. The little number of scores with very low scores implied that there were merely single events occurring at some of the sites only. This fits to the finding that the timing of slurry application is crucial for the amount of nutrient loss to the streams (Hooda et al., 2000; Cherobim et al., 2017). Thus, we interpreted the negative peaks of the 4<sup>th</sup> component as sporadic events of slurry application, being either unintentionally directly applied to the stream during the spreading of the slurry or being leached via surface runoff and tile drain discharge after application. We conclude that these negative peaks can be ascribed to slurry application, being either unintentionally directly applied to the stream or being leached via surface runoff and tile drain discharge after application.

735  
740  
745

## 5.2 Multiple sites

750 The interpretation of the 1<sup>st</sup> component as agriculturally induced enhancement of the natural background level of most of the water quality variables is consistent with the spatial pattern of median component scores at the different sites. The highest scores were found in the Dauergraben stream and in the Peege stream (Figure 5 A).

Both catchments are characterized by intense agriculture, a relatively dense network of tile drains, and hardly any buffer strips along the streams leading to a rapid transmission of solute enriched waters from the fields to the streams. In contrast, the Strom stream exhibited the lowest scores among all streams. Compared to the other  
755 streams, the valley of the Strom stream is clearly deep cut. Therefore, the Strom stream is expected to receive along its whole length continuous and substantial groundwater inflow from the deep aquifer. In addition, the valley slopes are covered with forest and not in agricultural use, acting as a buffer strip for the agricultural impact. Furthermore, the fraction of arable land in the Strom catchment is smallest,  
760 and the fraction of woodland is largest compared to the other catchments (Table 1). Main parts of the Strom catchment are situated within a nature conservation area furthermore limiting the agricultural impact in its riparian zone.

Deep groundwater, shallow groundwater and the stream water were well separated by the 2<sup>nd</sup> component (Figure 5 B). Exceptions were the sites at the  
765 Peege, which are mainly supplied with water from tile drainage and the shallow aquifer and consequently yield median values similar to the shallow groundwater. The largest positive median values of the 2<sup>nd</sup> component, being higher than those of the other stream water sites, were observed at sites with less than 13 samples (Q\_103 and Q\_106) and at the site U\_128 which received at least partly waters from  
770 a different region than the other stream water sites (Figure 1 and Figure 5 B). Thus, for the purpose of this study, we restricted our analysis on the spatial variability of the redox component to the categories of deep groundwater, shallow groundwater and stream water.

However, we took a closer look at the non-linear structure that became apparent  
775 for the deep groundwater samples in some of the residual plots of the 2<sup>nd</sup> component (Figure S2). In addition, we related the groundwater values of the 2<sup>nd</sup> component to the 1<sup>st</sup> component and the  $\text{HCO}_3^-$  and  $\text{Fe}^{2+}$  concentration (Figure 4). The negative relationship between the 2<sup>nd</sup> component and the 1<sup>st</sup> component in the deep groundwater suggests that the agricultural load represented by the 1<sup>st</sup> component  
780 acts as a driver for the sulphide oxidation represented by the 2<sup>nd</sup> component. Among all deep groundwater wells, the deepest groundwater well Gd\_198 exhibited the

lowest scores of the 1<sup>st</sup> component (Figure 5 A) and the highest scores of the 2<sup>nd</sup> component (Figure 4 lower panel and Figure 5 B). This suggests that due to the relatively low agricultural load the oxidation of sulphides was the least pronounced among all deeper wells. Similar relationships between the extent of sulphate oxidation in the aquifer and agriculturally borne NO<sub>3</sub><sup>-</sup> input have been found in other studies (e.g., Zhang et al., 2009; Jessen et al., 2017 and references therein).

We expected the ratio of groundwater from the deep aquifer contributing to the streamflow to increase in general with increasing catchment size. The Peege stream is mainly fed by the shallow aquifer and yielded consequently median values of the 3<sup>rd</sup> component similar to the shallow groundwater sites (Figure 5 C). The streams of Quillow, Strom and Stierngraben, showed little higher median values, indicating the larger proportion of groundwater from the deep aquifer contributing to runoff compared to the Peege stream. The sites U\_128 and D\_112 showed the highest median values of the 3<sup>rd</sup> component among the stream water sites, being equal or even higher than those of the deep groundwater sites (Figure 5 C). Both sites have subsurface catchments that do not include the deep groundwater samplings sites in this study. Site D\_112 is on the eastern side of the river Ucker, while all groundwater sampling sites are on the western side of it (Figure 1). In addition, its higher median value of the 3<sup>rd</sup> component was partly due to several peaks during the winter time. Those coincide with high values of Cl<sup>-</sup>. These might indicate road salt application, but we did not investigate this further, as it considered only this single site. Site U\_128 is situated at the outlet of the lake Unteruckersee upstream of the confluence of the Quillow stream (Figure 1). There, we did not expect a contribution of the groundwater sampled in the Quillow catchment either.

All the stream water sampling sites with negative peaks of the 4<sup>th</sup> component are located near arable fields which are known to get fertilised by slurry (Figure 5 D). For example the two most affected sites Q\_102 and Q\_103 receive slurry input from a large pig farm close by (personal communication G. Verch). Overall, only a few slurry input events accounted for 22% of the representation of the interpoint distances of all the water quality samples of the water quality data set in the Isomap projection (Figure 5 D). However, the performance of the representation of the interpoint

815 distances after adding the 4<sup>th</sup> component differed substantially between the different  
sites (Table S4). In case of site S 121 the representation of interpoint distances with  
four components ( $R^2 = 0.68$ ) was even slightly worse than with three components ( $R^2$   
= 0.66) (Table S4). This indicated an anomaly at this specific site compared to all  
other sites with respect to the 4<sup>th</sup> component, respectively the solutes which mainly  
determine the 4<sup>th</sup> component. We traced this phenomenon back to one single sample  
from the 25<sup>th</sup> of May 2004 which comprised relatively high DOC values and at the  
820 same time relatively low values of  $K^+$ , which is opposing the relationships indicative  
for the 4<sup>th</sup> component (Figure 3). The deterioration of the representation of the  
interpoint distances after adding the 4<sup>th</sup> component at this site vanished in an Isomap  
analysis which was performed without this sample. We were not able to find an  
explanation for this exceptional sample. However, it underlined that by applying a  
825 dimension reduction method every single sample is put into perspective of the global  
features of the data set as depicted by the components. Overall, the 4<sup>th</sup> component  
underlines the necessity to develop and use methods in environmental data analysis  
which enable to consider non-linear processes as well as singular and site-specific  
events.~~However, the percentage of represented interpoint distances of all samples at~~  
830 ~~a specific site ranged from < 1% to 42% for sites with  $n > 18$  (Table S4). This~~  
~~underlines the necessity to develop and use such methods in environmental sciences~~  
~~which are able to consider non-linear processes and to deal with singular and site-~~  
~~specific events.~~

### 835 **5.3 Long-term patterns**

Dominant changes were observed for the first two components (Figure 6). We interpreted the non-linear long term trend of the 1<sup>st</sup> component at most stream water sites (Figure 8) as the response of stream water quality to the interannual variability of depth to groundwater. An increase in the thickness of the unsaturated zone leads  
840 in general to longer residence time of seepage water, increasing retardation and buffering of topsoil seepage water, which is reducing the solute concentration originating from the surface in the seepage water and consequently reducing the values of the 1<sup>st</sup> component.

Trends similarly shaped to the non-linear trend of the 1<sup>st</sup> component of stream  
845 water quality were observed for the water level in the deep groundwater. In general,  
the turning points of the deep groundwater head time series lag behind those of the  
scores of the 1<sup>st</sup> component of the stream water sites by approximately 1.5 years  
(Figure 8). The earlier date of the turning point at groundwater gauge Gd\_204 in  
850 October 2005 is most probably an artefact, caused by the effect of the large time  
gaps in 2006 and 2007 on the de-seasonalising at this site and has to be considered  
with care.

We suggest that the time lag between stream water chemistry and water level in  
the deep aquifer is due to different response times to changes in the moisture  
conditions of the unsaturated zone. Compared to the relatively fast response of the  
855 stream water quality, the groundwater level in the deep aquifer reacts slower. In  
general, the overall trend of groundwater recharge reflects a relatively slow response  
to changes in the regional water balance. The velocity of seepage in the sediments of  
the upstream region of the Quillow catchment is estimated to be roughly 1 m per  
year.

860 The seasonal patterns, i.e. the annual variability, in the time series of the scores of  
the 1<sup>st</sup> component in the streams were ascribed to transient hydraulic decoupling of  
the mostly affected topsoils from the streams in summer. Usually there is hardly any  
seepage during the dry summer months at all. This leads often to **desiccation**  
**of drought in** the uppermost stream reaches ([left panel Figure 2; Lischeid et al.,](#)  
865 [2017](#)[Merz and Steidl, 2015](#)). Thus, shallow groundwater and tile drain discharge,  
both sources with relatively high agricultural load, did not contribute to stream  
discharge during these periods and larger areas of the catchment got hydraulically  
decoupled from the stream network (Merz and Steidl, 2015). Similar effects of the  
seasonal variability of the hydrological connectivity of streams, groundwater and tile  
870 drainage in agricultural catchments on the concentration level of agriculturally born  
solutes in the stream water have been reported, e.g. for NO<sub>3</sub><sup>-</sup> in the Schaugraben  
study catchment in the North of Germany (Wriedt et al., 2007) and for NO<sub>3</sub><sup>-</sup> and Cl<sup>-</sup> in  
the Kervidy-Naizin catchment in western France (Molenat et al., 2008; Aubert et al.,  
2013).

875 The other dominant change of stream water chemistry observed in this study was  
the continuous increase of the 2<sup>nd</sup> component at most stream water sites (Figure 6).  
All of the sampling sites with very low values of the 2<sup>nd</sup> component were in the deep  
aquifer (Figure 5 B). The positive trends of the 2<sup>nd</sup> component at most stream water  
sites were ascribed to changes in the chemistry of the groundwater-borne baseflow.  
880 Considering the interpretation of the 2<sup>nd</sup> component, this translates into enhanced  
oxidation of geogenic sulphides in the deeper aquifer due to the continuous input of  
agriculturally born NO<sub>3</sub><sup>-</sup> and DOC and subsequent calcite dissolution. Geogenic  
sulphides, such as pyrite, serve as electron donors for denitrification. The  
consumption of the geogenic sulphides is irreversible and might lead to the depletion  
885 of the denitrification capacity in the deep aquifer in the long run (Merz et al., 2009;  
Zhang et al., 2009; Merz and Steidl, 2015). Consequently, buffering of NO<sub>3</sub><sup>-</sup> surplus  
from agricultural land use is expected to decrease and NO<sub>3</sub><sup>-</sup> concentration in the  
groundwater and the stream water is expected to increase. The hypothesised long-  
term development should be of concern for the water resources and environmental  
890 protection agencies with respect to future water quality and related international  
commitments, such as the Water framework (EU, 2000), the Groundwater (EU, 2006)  
and the Nitrate directive (EU, 1991) of the European Union. Substantial time lags  
have to be considered for the response of groundwater quality to measures that  
reduce leaching of NO<sub>3</sub><sup>-</sup> (e.g. Pierson-Wickmann et al., 2009; Meals et al., 2010). In  
895 the Quillow catchment, we expect travel times in the order of magnitude of decades  
for the seepage water to reach the deep aquifer.

We did not observe dominant changes for the other two water quality components  
during the course of the observation period. The main temporal feature of the 3<sup>rd</sup>  
component was a very distinct and steady seasonal pattern, as could be expected for  
900 the mixing ratio of groundwater from the deep aquifer. All stream water sites with  $n >$   
50, except for D\_112, showed a distinct seasonal pattern with maximum scores in the  
summer, which is consistent with the assumption that the fraction of deep  
groundwater in the streams is highest during this period (Figure 6). The seasonal  
pattern at site D\_112 was disturbed by the winter peaks we ascribed to road salt  
905 application (section 5.2).

Because of its strong dependence from single events (Figure 5 D), the results of the estimation of the seasonal patterns and the trends of the 4<sup>th</sup> component have to be considered with care. The maxima of the seasonal pattern in summer at some sites were interpreted as reduced nutrient inputs to the stream due to nutrient uptake of plants and maximum buffering capacity of the unsaturated zone in summer. There were no indications for any effects of those events that we ascribed to the direct effect of slurry application on samples taken on the subsequent sampling dates at the affected sites. This is presumably due to the width of the sampling interval (Figure 2).

In case of dependence of a component from single events, 'change' might be also related to clustering of those events during certain parts of the series, either for series at single sites or sets of series. Most of the 'extreme' events of the 4<sup>th</sup> component appeared during the first half of the observation period (not shown). However, because of the small number of clearly outstanding events, we did not investigate this further (Figure 5 D).

In this study, the presented analysis of changes in water quality was limited by the temporal resolution of the data. Aspects such as long-term memory effects, as indicated by fractal scaling of solute series (Kirchner et al., 2000) and the observed scale-crossing non-self-averaging behaviour of solute series (Kirchner and Neal, 2013) were not considered. However, we assume that the suggested use of multivariate components gives some robustness to the detected changes compared to the analysis of single solutes.

#### 5.4 Effects of the irregular sampling

There was an obvious spatial bias with a focus on the Quillow catchment itself, conditioned by the focus of the monitoring (section 2.2, Figure 1). Stream sampling sites were only partly independent from each other, as the same streams had been sampled along different reaches. This needs to be considered in the interpretation of the components. In our exploratory approach, differences between subsequent stream reaches helped to identify the effects of tributaries or groundwater that recharged between the respective sampling sites. In that way, the stream was used



as a measurement device for biogeochemical processes and water-borne solute transport in different parts of the catchment and the interlinkages of groundwater and stream water.

940 It is important to note that our approach does not require the same number of samples per site~~identical temporal sampling resolution at all sites~~ (Figure 2). The derived components constitute a frame in which all samples are integrated independent of the number of sample per site. Thus, in our application we get the information of how those sites with very little samples group or behave in relation to the others. Even a few samples might indicate for example that the respective site  
945 behaves similar to other sites with respect to some components and very different with respect to other components. The influence of single samples for the integration of the different sites into the global pattern of the water quality relationships summarized by the 4<sup>th</sup> component is an illustrative example for that (section 5.2). Thus, even o  
950 Occasional sampling at additional some sites helps assessing the strength of effects of the respective drivers at these sites and might support or contradict hypotheses on spatial variability and related long-term patterns of those influences. This information would be lost if those samples would be excluded beforehand.

955 In addition, the approach followed here does not require identical temporal sampling resolution at all sites or synchronous sampling dates. Thus, a strictly regular sampling design, which is hardly feasible, is no prerequisite. Correspondingly, data from different monitoring programs could be used for a joint analysis.

960 Sampling intervals at the sampling sites with  $N > 50$  were not normally distributed and biased towards deviations that are longer than the median (right panel Figure 2). Several series exhibited large time gaps. However, as sampling intervals did not change systematically throughout the monitoring period we assume that the effects on the results of the significance test with Mann-Kendall were negligible (section 3.2). In comparison, the trend estimations with Theil-Sen estimator and LOESS are more robust, as they incorporate the exact sampling dates explicitly in the calculations.  
965 Thus, we do not expect major effects on the sign of the Theil-Sen estimator or the general shape of the LOESS smooth at the given temporal resolution.

970 In general, the interpretation of the components should consider the temporal structure of the data set. For example in this study the drying out of the streams at the Peege sites and the most upstream sites of the Quillow in summer was the most important systematic deviation from an otherwise roughly similar sampling across seasons (left panel Figure 2). This information was included in the interpretation of the 1<sup>st</sup> component (section 5.3). If the monitoring would in general not have been performed roughly similarly across seasons, e.g. if one or more seasons would in general be missing, the estimation of the seasonality would not be applicable. If the  
975 monitoring would be such that there would be different seasons sampled in different years, this would have to be considered in the estimation of the trend.

980 ~~There was an obvious spatial bias with a focus on the Quillow catchment itself, conditioned by the focus of the monitoring (section 2.2, Figure 1). Stream sampling sites were only partly independent from each other, as the same streams had been sampled along different reaches. This needs to be considered in the interpretation of the components. In our exploratory approach, differences between subsequent stream reaches helped to identify the effects of tributaries or groundwater that recharged between the respective sampling sites. In that way, the stream was used as a measurement device for biogeochemical processes and water-borne solute~~  
985 ~~transport in different parts of the catchment and the interlinkages of groundwater and stream water.~~

## 5.5 Exploratory framework

990 The application of a dimension reduction approach was motivated by the assumption that drivers influencing water quality usually affect more than one solute, and that single solutes are affected by more than one driver. Like in preceding studies (e.g., Lischeid and Bittersohl, 2008; Lischeid et al., 2010), the representation of water quality data in low-dimensional space required only a few components to capture the 'main features' of the data set. ~~Non-linear Isometric Feature Mapping performed in this study only slightly better with respect to the representation of interpoint distances than PCA (Table 2), suggesting that mainly linear relationships~~

~~were of importance for the overall dynamics in the data set.~~

~~Whether the relationships in the data set are mainly linear ones, as in this study, or whether there are considerably non-linear relationships as well. This~~ is usually not

1000 known in advance. Thus, if the aim is to consider and check for possible non-linear relationships in the analysis we recommend using PCA as a linear benchmark for Isomap (demonstrated by Lischeid and Bittersohl, 2008). In a straightforward way this allows for 1) assessing whether the dominant correlation structures in the data set are mainly linear or non-linear, and 2) identifying those components, samples, sites  
1005 and periods deviating from the linear behaviour as captured by the PCA.

Based on the correlation of component scores and residuals, we formulated for each considered component a hypothesis on a dominant driver influencing water quality. ~~Again, whether the relationships are linear, as it was for most of the global relationships in this study (Figure S1-S4), is usually not known beforehand.~~

1010 ~~Summarizing the relationships between residuals and components with Spearman rank correlation enables to consider non-linear relationships between residuals and components as well, as long as they are monotonic. However, the main benefit in this study was that Spearman rank correlation is less sensitive to extreme values compared to Pearson correlation. This concerned especially the assessment of the relationships of the residuals of  $\text{SO}_4^{2-}$  and  $\text{Cl}^-$  with the 2<sup>nd</sup> component and the residuals of  $\text{PO}_4^{3-}$  and  $\text{NH}_4^+$  with the 4<sup>th</sup> component (Figure S2 and S4), which were way stronger expressed with Pearson correlation due to a few single extreme values.~~

1015 ~~The assessment of the relationships of scores and residuals with Spearman rank correlation considers non-linear monotonic correlations and is less sensitive to extreme values compared to Pearson correlation.~~

1020 The derived correlations differ from default loadings of PCA, which are defined as the coefficients of the linear combination of the analysed variables which is used to calculate the principal component scores. Those coefficients, scaled by the square root of the eigenvalue of the respective component, are equivalent to the Pearson correlation of PCA  
1025 component scores and analysed variables. It is important to note that the differences in the evaluation of the correlations of components and the measured variables might lead to different interpretations of the components.

1030 The treatment of censored values can substantially affect the derived components  
and the subsequent interpretation of the results and has to be considered carefully  
(Helsel, 2012 and references therein). For the application of Isomap, it is required to  
provide numerical values for the values below the detection limit. For simplicity, we  
here used half the detection limit as a maker for values below the detection limit. We  
1035 checked for the effect of this substitution by comparing the Isomap results of the  
presented analysis with another Isomap analysis in which we excluded the two most  
affected variables NO<sub>2</sub><sup>-</sup> and PO<sub>4</sub><sup>3-</sup> (Figure S4). The correlation of the Isomap scores  
of the interpreted components 1 to 4 of version 1 (with NO<sub>2</sub><sup>-</sup> and PO<sub>4</sub><sup>3-</sup>) vs. version 2  
(without NO<sub>2</sub><sup>-</sup> and PO<sub>4</sub><sup>3-</sup>) yielded a R<sup>2</sup> of cp1: 0.99, cp2: 0.98, cp3: 0.97, cp4: 0.64.  
The comparison of the two versions with respect to the Spearman rank correlations  
of Isomap scores of the first four components and the residuals (please see Figure 3  
1040 for the respective values of version 1) yielded a R<sup>2</sup> of cp1: 0.98, cp2: 0.99, cp3: 0.99,  
cp4: 0.88. Thus the first three components are virtually identical. The 4<sup>th</sup> component  
is affected, because PO<sub>4</sub><sup>3-</sup> is one of the important variables for this component  
(Figure 3). Still, the similarity of the correlations of Isomap scores and the 4<sup>th</sup>  
component of both versions suggests that the characteristics of the 4<sup>th</sup> component  
1045 were not merely introduced by the substitution of the values below the detection limit  
for PO<sub>4</sub><sup>3-</sup>. Thus, overall, the substitution did not substantially affect the interpretation  
of the considered components. For data sets which are more heavily affected by  
censored values other dimension reduction methods such as the rank based  
approaches suggested by Helsel (2012) should be preferred.

1050 For data sets in which the number of measured variables differs between the sites  
there is a trade-off between number of considered variables vs. number of  
considered sites. Depending on the focus of the study different selections of the data  
set can be used. For example if the main focus of the study is to analyse the  
multivariate water quality dynamics in detail it might be worthwhile to disregard some  
1055 sites to be able to include more variables. If the focus is to maintain the spatial  
coverage of the monitoring, like in this study, more sites might be of more value than  
additional variables. Depending on the available resources a third option would be to  
perform two analyses, one focusing on more variables, one on more sites, and  
comparing the results. If it is possible to link the considered components, like we did

1060 in the preceding paragraph, this proceeding can be used for spatial extrapolation of  
the hypotheses derived from the version which included more variables. However, in  
our case the sketched trade-off was not dramatic. Thus, we excluded only the  
variables with more than 5% missing values (section 2.2) to keep the possible effect  
of any method of replacement rather low.

1065 To prevent adding variables with little information gain it is recommendable to  
perform a correlation analysis beforehand and rule out highly correlated variables.  
For this purpose we recommend not to rely only on a numerical measure of  
correlation, but to visually examine the scatterplots of the respective variables to  
1070 check for systematic deviations from the global relationship. There might be e.g.  
some sites or seasons in which the otherwise tight relationship gets weaker pointing  
to local or temporal phenomena.

Technically it is possible to combine other data than solutes (e.g. sediment data,  
biological indicators, etc.) together with the solutes in one joined data set for the  
derivation of the components. However, the multivariate components derived by the  
1075 dimension reduction approach are the basis of the subsequent interpretation of the  
results. It has to be considered as well that all included variables are equally  
weighted due to the z-scaling prior to the dimension reduction. Thus, including other  
types of data might in some cases complicate the interpretation. In general, we  
recommend not to mix variables with different scales of measures (e.g. nominal  
1080 variables and ratio scaled variables) in the data base for the derivation of the  
components.

Instead, data which was not used in the derivation of the components can be used  
as additional information for their interpretation. For example in this study, we used in  
addition~~Complementary, we used for the derivation of the hypotheses~~ to the  
1085 spatiotemporal features of the components~~in combination with the spatial order of~~  
~~the sampling sites,~~ other variables like groundwater level series,  $\text{Fe}_2^+$  and  $\text{HCO}_3^-$   
concentration from the groundwater samples, the spatial distribution of land use, and  
expert knowledge on the study area for the derivation of the hypotheses. A thorough  
testing of the hypotheses, for example through hydrochemical modelling or numerical  
1090 experiments with virtual catchments was out of the scope of this study.

1095 However, an interpretation of the components as distinct drivers is no prerequisite for the further analysis of the components. In any case, the components constitute, and can simply be used as, a condensed representation of similar behaviour among the analysed variables according to the constraints of the used dimension reduction method.

1100 For PCA and Isomap each component describes subsequently the correlation structure that is most prominent in the remainder of what has not already been assigned to the higher-ranked components. This implies that each component has to be interpreted with respect to the higher ranked components. Also, the consideration of the respective other components in the interpretation of a component can be helpful to carve out its characteristics as it was done here with the residuals of the multiple linear regression of the respective three other components and the measured variables (e.g. Figure S1). Beyond that, it can be helpful to elucidate the interaction of the components as it was done here e.g. for scores of the 1<sup>st</sup> and 2<sup>nd</sup> component (Figure 4 lower panel).

1110 The sites differed substantially with respect to the median values of the four analysed multivariate components (Figure 5). However, these components comprised the largest fraction of the interpoint distances at any single site with more than 18 samples (Table S4). We conclude that our results identified major regional phenomena rather than site-specific peculiarities. This is consistent with the prior assumptions that there are a few dominant drivers which determine the main influencing stream water and groundwater quality dynamics in the region. were in fact the same at all sites. This gives some confidence to hypothesize that these drivers presumably play a major role even in adjacent catchments that have not been sampled so far.

1120 To detect and characterize the dominant changes in the multivariate water quality data we explored whether there were shifts in time in specific components, whether they were linear or non-linear in nature, and if trends did occur at many or only at single sites. For example for the scores of the 1<sup>st</sup> component, the Mann-Kendall approach identified monotonic trends at various stream water sampling sites (Figure 6). However, the linear trend estimation failed to detect the non-linear trend that was

observed at many series (Figure 8). This reflects the well-known sensitivity of global linear trend estimation to low-frequency patterns that are not entirely covered by the observation period (Koutsoyiannis, 2006; Milliman et al., 2008; Lins and Cohn, 2011).

1125 The LOESS smooths of the de-seasonalised series, on the other hand, did clearly reveal the similarity between the long term behaviour of groundwater level in the deep aquifer and series of the 1<sup>st</sup> component. In our exploratory approach, the LOESS smooth of the de-seasonalised series served as a descriptive tool for illustrating rather than for proving non-linear long-term patterns. No significance test  
1130 was applied. The outcome of the LOESS smoother highly depends on the parameterisation of the approach (i.e., the degree of smoothness) that would have to be justified prior testing of significance.

## 6 Conclusions

1135 We suggested and tested an exploratory approach for the detection of dominant changes in multivariate water quality data sets with irregular sampling in space and time. The combination of the selected methods aimed to provide a broadly applicable exploratory framework for typical existing monitoring data sets, e.g. from environmental agencies, which are often characterized by relatively low sampling  
1140 frequency and irregularities of the sampling in space and / or time. In the approach, we applied a dimension reduction method to derive multivariate water quality components and analysed their spatiotemporal features with respect to changes that concerned more than single sites, short-term fluctuations or single events.

The components can be used irrespective of an interpretation as drivers  
1145 influencing water quality. By definition, the components are a sparse description of the common dynamics among the water quality variables. Thus, similar behaviour in space and time among the water quality variables as well as systematic changes in the multivariate water quality data can be addressed in a purely descriptive manner. This can be used for example to test the often implicit assumption of constant  
1150 boundary conditions of scientific process and modelling studies. Furthermore, the components and their spatiotemporal features per se can serve as reference for

further studies, e.g. detailed process studies with higher temporal resolution, and the assessment of future developments of water quality in an area. In this study, the components were used to develop hypotheses on dominant drivers influencing water quality and to analyse the temporal and spatial variability of those influences.

It is emphasized that the presented approach is readily applicable with data from common monitoring programs without specific requirements concerning sampling frequency or regular distribution of sampling sites, sampling dates, and sampling intervals, except that there should be no systematic bias in the respective distribution. Even variables which have to be excluded from the derivation of the components, for example because of the amount of missing values or because they have been monitored only at subsets of the sampling sites, can be related to the components as additional information for their interpretation. For example in this study we used the concentration of  $\text{Fe}^{2+}$  and  $\text{HCO}_3^-$  in the groundwater as additional information for the interpretation of the 2<sup>nd</sup> component. Thus the approach allows an efficient use of existing monitoring data as well as the consideration of often neglected 'irregular' pieces of data from e.g. pilot studies or single sampling campaigns. Irregularities in the structure of a data set are not seen as fundamental hindrance, but as additional source of information. We see this as a major advantage for the analysis of comprehensive water quality monitoring programs, both from a scientific perspective and from a more applied point of view of e.g. water resources and environmental agencies. Therefore, we recommend the approach especially for the exploratory assessment of existing long term low frequency multivariate water quality monitoring data sets.

1175

### **Data availability**

A selection of R-scripts that covers the main steps of the exploratory framework is provided at [doi: 10.4228/ZALF.2017.340](https://doi.org/10.4228/ZALF.2017.340)~~http://open-research-data.ext.zalf.de/ResearchData/2017\_340.html~~ under CC-BY 4.0 licence. It comes together with the water quality data used in this manuscript and some examples of exploratory plots not included in this manuscript.

1180



## Acknowledgements

The long-term monitoring program that provided the data for this study would not  
1185 have been possible without the diligent work of many colleagues. We would like to  
thank Roswitha Schulz, Dorith Henning, Ralph Tauschke, Joachim Bartelt (†), Peter  
Bernd and Bernd Schwien for installation of sampling sites, including numerous  
groundwater wells, and for performing the sampling program in spite of sometimes  
harsh field conditions. In addition we acknowledge the painstaking work of Rita  
1190 Schwarz (†) and Melitta Engel in the chemical laboratory of the Institute of  
Landscape Hydrology as well as of the staff of the central chemical laboratory of the  
Leibniz Centre for Agricultural Landscape Research. We thank Gernot Verch of the  
research station Dedelow for the information on the historical development of  
agricultural land use in the study area. We thank the editor Stacey Archfield for the  
1195 smooth handling of the review process and the two anonymous referees for their  
contributions and constructive comments, especially the issue of censored values,  
which helped to improve the manuscript.

Christian Lehr received funding from the Leibniz Association (SAW-2012-IGB-  
4167) within the International Leibniz Research School: Aquatic boundaries and  
1200 linkages in a changing environment (Aqualink).

We used free software products under the GNU General Public Licence and thank  
the respective communities. Maps and the determination of the catchments' areas  
were carried out with Quantum GIS 2.14.1 (<http://www.qgis.org/index.php>) and  
statistical analyses and the graphs were performed using the R statistical software  
1205 environment, version 3.4.1 (R Core Team, 2017; <http://www.r-project.org>).

## References

- 1210 Aubert, A. H., Gascuel-Oudou, C., Gruau, G., Akkal, N., Faucheux, M., Fauvel, Y., Grimaldi, C., Hamon, Y., Jaffrézic, A., Lecoz-Boutnik, M., Molénat, J., Petitjean, P., Ruiz, L. and Merot, P.: Solute transport dynamics in small, shallow groundwater-dominated agricultural catchments: insights from a high-frequency, multisolute 10 yr-long monitoring study, *Hydrology and Earth System Sciences*, 17, 1379-1391, doi: 10.5194/hess-17-1379-2013, 2013.
- 1215 Basu, N. B., Destouni, G., Jawitz, J. W., Thompson, S. E., Loukinova, N. V., Darracq, A., Zanardo, S., Yaeger, M., Sivapalan, M., Rinaldo, A. and Rao, P. S. C.: Nutrient loads exported from managed catchments reveal emergent biogeochemical stationarity, *Geophysical Research Letters*, 37, doi: 10.1029/2010GL045168, 2010.
- 1220 Basu, N. B., Thompson, S. E. and Rao, P. S. C.: Hydrologic and biogeochemical functioning of intensively managed catchments: A synthesis of top-down analyses, *Water Resources Research*, 47, doi: 10.1029/2011WR010800, 2011.
- Beudert, B., Bäessler, C., Thorn, S., Noss, R., Schröder, B., Dieffenbach-Fries, H., Foullois, N. and Müller, J.: Bark Beetles Increase Biodiversity While Maintaining Drinking Water Quality, *Conservation Letters*, 8, 272-281, doi: 10.1111/conl.12153, 2015.
- 1225 Bieroza, M. Z., Heathwaite, A. L., Mullinger, N. J. and Keenan, P. O.: Understanding nutrient biogeochemistry in agricultural catchments: the challenge of appropriate monitoring frequencies, *Environmental Science: Processes & Impacts*, 16, 1676–1691, doi: 10.1039/C4EM00100A, 2014.
- 1230 Blaen, P. J., Khamis, K., Lloyd, C., Comer-Warner, S., Ciocca, F., Thomas, R. M., MacKenzie, A. R. and Krause, S.: High-frequency monitoring of catchment nutrient exports reveals highly variable storm event responses and dynamic source zone activation, *Journal of Geophysical Research: Biogeosciences*, 122, 2265–2281, doi: 10.1002/2017JG003904, 2017.
- Broers, H. P. and van der Grift, B.: Regional monitoring of temporal changes in

1235 groundwater quality, *Journal of Hydrology*, 296, 192-220, doi: 10.1016/j.jhydrol.2004.03.022, 2004.

Burt, T. P., Howden, N. J. K., Worrall, F. and McDonnell, J. J.: On the value of long-term, low-frequency water quality sampling: avoiding throwing the baby out with the bathwater, *Hydrological Processes*, 25, 828-830, doi: 10.1002/hyp.7961, 2011.

1240 Burt, T. P., Howden, N. J. K., Worrall, F. and Whelan, M. J.: Importance of long-term monitoring for detecting environmental change: lessons from a lowland river in south east England, *Biogeosciences*, 5, 1529-1535, doi: 10.5194/bg-5-1529-2008, 2008.

1245 Capell, R., Tetzlaff, D., Malcolm, I., Hartley, A. and Soulsby, C.: Using hydrochemical tracers to conceptualise hydrological function in a larger scale catchment draining contrasting geologic provinces, *Journal of Hydrology*, 408, 164-177, doi: 10.1016/j.jhydrol.2011.07.034, 2011.

1250 Cassidy, R. and Jordan, P.: Limitations of instantaneous water quality sampling in surface-water catchments: Comparison with near-continuous phosphorus time-series data, *Journal of Hydrology*, 405, 182–193, doi: 10.1016/j.jhydrol.2011.05.020, 2011.

[Cherobim, V. F., Huang, C.-H. and Favaretto, N.: Tillage system and time post-liquid dairy manure: Effects on runoff, sediment and nutrients losses, \*Agricultural Water Management\*, 184, 96–103, doi: 10.1016/j.agwat.2017.01.004, 2017.](https://doi.org/10.1016/j.agwat.2017.01.004)

1255 Cleveland, R., Cleveland, W., McRae, J. and Terpenning, I.: STL: A seasonal-trend decomposition procedure based on loess, *Journal of Official Statistics*, 6, 3-73, 1990.

Cleveland, W. S.: Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 74, 829-836, doi: 10.1080/01621459.1979.10481038, 1979.

1260 Cleveland, W. S. and Devlin, S. J.: Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association*, 83, 596-610, doi: 10.1080/01621459.1988.10478639, 1988.

Cloutier, V., Lefebvre, R., Therrien, R. and Savard, M. M.: Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system, *Journal of Hydrology*, 353, 294-1265 313, doi: 10.1016/j.jhydrol.2008.02.015, 2008.

[Cohn, T. A. and Lins, H. F.: Nature's style: Naturally trendy, \*Geophysical Research Letters\*, 32, doi: 10.1029/2005GL024476, 2005.](#)

EU: Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources, *Official Journal of the European Communities*, 1-8, 1991. 1270

EU: Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy, *Official Journal of the European Communities*, 1-70, 2000.

EU: Directive 2006/118/EC of the European Parliament and of the Council of 12 December 2006 on the protection of groundwater against pollution and deterioration, *Official Journal of the European Union*, 19-31, 2006. 1275

Fitzpatrick, M., Long, D. and Pijanowski, B.: Exploring the effects of urban and agricultural land use on surface water chemistry, across a regional watershed, using multivariate statistics, *Applied Geochemistry*, 22, 1825-1840, doi: 1280 10.1016/j.apgeochem.2007.03.047, 2007.

Gámez, A. J., Zhou, C. S., Timmermann, A. and Kurths, J.: Nonlinear dimensionality reduction in climate data, *Nonlinear Processes in Geophysics*, 11, 393-398, doi: 10.5194/npg-11-393-2004, 2004.

Geng, X., Zhan, D.-C. and Zhou, Z.-H.: Supervised nonlinear dimensionality reduction for visualization and classification, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35, 1098-1107, doi: 1285 10.1109/TSMCB.2005.850151, 2005.

Glynn, E., Chen, J. and Mushegian, A.: Detecting periodic patterns in unevenly spaced gene expression time series using Lomb--Scargle periodograms,

1290 Bioinformatics, 22, 310-316, doi: 10.1093/bioinformatics/bti789, 2006.

Graeber, D., Gelbrecht, J., Pusch, M. T., Anlanger, C. and von Schiller, D.: Agriculture has changed the amount and composition of dissolved organic matter in Central European headwater streams, *Science of The Total Environment*, 438, 435-446, doi: 10.1016/j.scitotenv.2012.08.087, 2012.

1295 Grayson, R. and Blöschl, G. (Ed.): *Spatial patterns in catchment hydrology: observations and modelling*, Cambridge University Press Cambridge, UK, 2000.

Haag, I. and Westrich, B.: Processes governing river water quality identified by principal component analysis, *Hydrological Processes*, 16, 3113-3130, doi: 10.1002/hyp.1091, 2002.

1300 Halliday, S. J., Wade, A. J., Skeffington, R. A., Neal, C., Reynolds, B., Rowland, P., Neal, M. and Norris, D.: An analysis of long-term trends, seasonality and short-term dynamics in water quality data from Plynlimon, Wales, *Science of The Total Environment*, 434, 186-200, doi: 10.1016/j.scitotenv.2011.10.052, 2012.

1305 Hannemann, M. and Schirrmeister, W.: Paläohydrogeologische Grundlagen der Entwicklung der Süß-/Salzwassergrenze und der Salzwasseraustritte in Brandenburg, *Brandenburg Geowissenschaftliche Beiträge*, 5, 61-72, 1998.

[Helsel, D. R.: Statistics for Censored Environmental Data Using Minitab and R, 2<sup>nd</sup> ed., John Wiley & Sons, 2012.](#)

1310 Hocke, K.: Phase estimation with the Lomb-Scargle periodogram method, *Annales geophysicae*, 16, 356–358, 1998.

Hocke, K. and Kämpfer, N.: Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram, *Atmospheric Chemistry and Physics*, 9, 4197-4206, doi: 10.5194/acp-9-4197-2009, 2009.

1315 [Hooda, P. S., Edwards, A. C., Anderson, H. A. and Miller, A.: A review of water quality concerns in livestock farming areas, \*Science of The Total Environment\*, 250\(1\), 143–167, doi: 10.1016/S0048-9697\(00\)00373-9, 2000.](#)

Hooper, R. P., Christophersen, N. and Peters, N. E.: Modelling streamwater chemistry as a mixture of soilwater end-members — An application to the Panola Mountain catchment, Georgia, U.S.A., *Journal of Hydrology*, 116, 321-343, doi: 1320 10.1016/0022-1694(90)90131-G, 1990.

Horne, J. and Baliunas, S.: A prescription for period analysis of unevenly sampled time series, *The Astrophysical Journal*, 302, 757-763, 1986.

Hotelling, H.: Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24, 417-441, doi: 1325 10.1037/h0071325, 1933.

Howden, N. J. K., Burt, T. P., Worrall, F. and Whelan, M. J.: Monitoring fluvial water chemistry for trend detection: hydrological variability masks trends in datasets covering fewer than 12 years, *J. Environ. Monit.*, 13, 514-521, doi: 10.1039/C0EM00722F, 2011.

1330 Jessen, S., Postma, D., Thorling, L., Müller, S., Leskelä, J. and Engesgaard, P.: Decadal variations in groundwater quality: A legacy from nitrate leaching and denitrification by pyrite in a sandy aquifer, *Water Resources Research*, 53, 184-198, doi: 10.1002/2016WR018995, 2017.

Jolliffe, I.: *Principal Component Analysis*, 2nd ed., Springer, 2002.

1335 Jolliffe, I. T. and Cadima, J.: Principal component analysis: a review and recent developments, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 374, doi: 10.1098/rsta.2015.0202, 2016.

1340 Jørgensen, J. C., Jacobsen, O. S., Elberling, B. and Aamand, J.: Microbial Oxidation of Pyrite Coupled to Nitrate Reduction in Anoxic Groundwater Sediment, *Environmental Science & Technology*, 43, 4851-4857, doi: 10.1021/es803417s, 2009.

Kalettko, T. and Rudat, C.: Hydrogeomorphic types of glacially created kettle holes in North-East Germany, *Limnologica - Ecology and Management of Inland Waters*, 36, 54–64, doi: 10.1016/j.limno.2005.11.001, 2006.

- 1345 Kalettka, T. and Steidl, J.: Measurement of stream water chemical ingredients, Quillow catchment, Germany, doi: 10.4228/ZALF.1998.265, 2014.
- Kendall, M.: Rank Correlation Methods, Charles Griffin Book Series, 1990.
- Kim, D. and Oh, H.-S.: EMD: A Package for Empirical Mode Decomposition and Hilbert Spectrum, The R Journal, 1, 40-46, 2009.
- 1350 Kim, D. and Oh, H.-S.: EMD: Empirical Mode Decomposition and Hilbert Spectral Analysis, 2014.
- Kirchner, J., Feng, X. and Neal, C.: Fractal stream chemistry and its implications for contaminant transport in catchments, Nature, 403, 524-527, 2000.
- Kirchner, J., Feng, X., Neal, C. and Robson, A.: The fine structure of water-quality  
1355 dynamics: the (high-frequency) wave of the future, Hydrological Processes, 18, 1353-1359, doi: 10.1002/hyp.5537, 2004.
- Kirchner, J. W. and Neal, C.: Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection, Proceedings of the National Academy of Sciences, 110, 12213-12218, doi: 10.1073/pnas.1304328110,  
1360 2013.
- Koutsoyiannis, D.: Nonstationarity versus scaling in hydrology, Journal of Hydrology, 324, 239-254, doi: ~~doi:~~10.1016/j.jhydrol.2005.09.022, 2006.
- Kroeze, C., Hofstra, N., Ivens, W., Löhr, A., Strokal, M. and van Wijnen, J.: The links between global carbon, water and nutrient cycles in an urbanizing world — the  
1365 case of coastal eutrophication, Current Opinion in Environmental Sustainability, 5, 566-572, doi: <https://doi.org/10.1016/j.cosust.2013.11.004>, 2013.
- Lahmer, W., Dannowski, R., Steidl, J., Pfützner, B.: Flächendeckende Modellierung von Wasserhaushaltsgrößen für das Land Brandenburg, Studien und Tagungsberichte Schriftenreihe des Landesumweltamtes Brandenburg, 27, 2000.
- 1370 Lee, J. A. and Verleysen, M.: Nonlinear Dimensionality Reduction, Springer, 2007.

Lehr, C., Pöschke, F., Lewandowski, J. and Lischeid, G.: A novel method to evaluate the effect of a stream restoration on the spatial pattern of hydraulic connection of stream and groundwater, *Journal of Hydrology*, 527, 394-401, doi: 10.1016/j.jhydrol.2015.04.075, 2015.

1375 Lins, H. F. and Cohn, T. A.: Stationarity: Wanted Dead or Alive?, *JAWRA Journal of the American Water Resources Association*, 47, 475-480, doi: 10.1111/j.1752-1688.2011.00542.x, 2011.

Lischeid, G. and Bittersohl, J.: Tracing biogeochemical processes in stream water and groundwater using non-linear statistics, *Journal of Hydrology*, 357, 11-28, doi: 1380 10.1016/j.jhydrol.2008.03.013, 2008.

Lischeid, G.; Krám, P. and Weyer, C.: In: Müller, F.; Baessler, C.; Schubert, H. & Klotz, S. (Ed.), *Tracing Biogeochemical Processes in Small Catchments Using Non-linear Methods*, Springer Netherlands, 2010.

Lischeid, G., Kalettka, T., Merz, C. and Steidl, J.: Monitoring the phase space of ecosystems: Concept and examples from the Quillow catchment, Uckermark, *Ecological Indicators*, 65, 55-65, doi: 10.1016/j.ecolind.2015.10.067, 2016.

~~Lischeid, G., Kalettka, T., Holländer, M., Steidl, J., Merz, C., Dannowski, R., Hohenbrink, T., Lehr, C., Onandia, G., Reverey, F. and Pätzig, M.: Natural ponds in an agricultural landscape: External drivers, internal processes, and the role of the terrestrial-aquatic interface, *Limnologica – Ecology and Management of Inland Waters*, doi: 10.1016/j.limno.2017.01.003, 2017.~~

Lomb, N.: Least-squares frequency analysis of unequally spaced data, *Astrophysics and space science*, 39, 447-462, 1976.

Mahecha, M. D., Martínez, A., Lischeid, G. and Beck, E.: Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data, *Ecological Informatics*, 2, 138-49, doi: 1395 10.1016/j.ecoinf.2007.05.002, 2007.

Mann, H. B.: Nonparametric Tests Against Trend, *Econometrica*, 13, 245-259,



1945.

1400 Massmann, G., Pekdeger, A. and Merz, C.: Redox processes in the Oderbruch polder groundwater flow system in Germany, *Applied Geochemistry*, 19, 863-886, doi: 10.1016/j.apgeochem.2003.11.006, 2004.

Massmann, G., Tichomirowa, M., Merz, C. and Pekdeger, A.: Sulfide oxidation and sulfate reduction in a shallow groundwater system (Oderbruch Aquifer, Germany),  
1405 | *Journal of Hydrology*, 278, 231-243, doi: [10.1016/S0022-1694\(03\)00153-7](https://doi.org/10.1016/S0022-1694(03)00153-7), 2003.

McLeod, A.: Kendall: Kendall rank correlation and Mann-Kendall trend test, 2011.  
URL: <https://CRAN.R-project.org/package=Kendall> (last access: 08 October 2017).

Meals, D. W., Dressing, S. A. and Davenport, T. E.: Lag Time in Water Quality Response to Best Management Practices: A Review, *Journal of Environmental*  
1410 | *Quality*, 39, 85-96, 2010.

Merz, C. and Steidl, J.: Measurement of groundwater heads, Quillow catchment, Germany, doi: 10.4228/ZALF.2000.272, 2014a.

Merz, C. and Steidl, J.: Measurement of ground water chemical ingredients, Quillow catchment, Germany, doi: 10.4228/ZALF.2000.266, 2014b.

1415 | Merz, C. and Steidl, J.: Data on geochemical and hydraulic properties of a characteristic confined/unconfined aquifer system of the younger Pleistocene in northeast Germany, *Earth System Science Data*, 7, 109-116, doi: 10.5194/essd-7-109-2015, 2015.

Merz, C., Steidl, J. and Dannowski, R.: Parameterization and regionalization of  
1420 | redox based denitrification for GIS-embedded nitrate transport modeling in Pleistocene aquifer systems, *Environmental Geology*, 58, 1587, doi: 10.1007/s00254-008-1665-6, 2009.

Milliman, J. D., Farnsworth, K. L., Jones, P. D., Xu, K. H. and Smith, L. C.: Climatic and anthropogenic factors affecting river discharge to the global ocean, 1951–2000,  
1425 | *Global and Planetary Change*, 62, 187–194, doi: [10.1016/j.gloplacha.2008.03.001](https://doi.org/10.1016/j.gloplacha.2008.03.001),

2008.

Molénat, J., Gascuel-Oudou, C., Ruiz, L. and Gruau, G.: Role of water table dynamics on stream nitrate export and concentration in agricultural headwater catchment (France), *Journal of Hydrology*, 348, 363-378, doi:

1430 | <http://dx.doi.org/10.1016/j.jhydrol.2007.10.005>, 2008.

Neal, C.: The water quality functioning of the upper River Severn, Plynlimon, mid-Wales: issues of monitoring, process understanding and forestry, *Hydrology and Earth System Sciences*, 8, 521-532, doi: 10.5194/hess-8-521-2004, 2004.

1435 | Neal, C., Reynolds, B., Rowland, P., Norris, D., Kirchner, J. W., Neal, M., Sleep, D., Lawlor, A., Woods, C., Thacker, S., Guyatt, H., Vincent, C., Hockenhull, K., Wickham, H., Harman, S. and Armstrong, L.: High-frequency water quality time series in precipitation and streamflow: From fragmentary signals to scientific challenge, *Science of The Total Environment*, 434, 3-12, doi: 10.1016/j.scitotenv.2011.10.072, 2012.

1440 | Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. and Wagner, H. 2017. *vegan: Community Ecology Package*. URL: <https://CRAN.R-project.org/package=vegan> (last access: 08 October 2017).

1445 | Pearson, K. F.: LIII. On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, 2, 559-572, doi: 10.1080/14786440109462720, 1901.

Pierson-Wickmann, A.-C., Aquilina, L., Martin, C., Ruiz, L., Molénat, J., Jaffrézic, A. and Gascuel-Oudou, C.: High chemical weathering rates in first-order granitic catchments induced by agricultural stress, *Chemical Geology*, 265, 369-380, doi:

<http://dx.doi.org/10.1016/j.chemgeo.2009.04.014>, 2009.

1450 | Press, W., Flannery, B., Teukolsky, S., Vetterling, W. and others: *Numerical recipes*, Cambridge university press, 2007.

Raymond, P. A., Oh, N.-H., Turner, R. E. and Broussard, W.: Anthropogenically enhanced fluxes of water and carbon from the Mississippi River, *Nature*, 451, 449-

452, doi: 10.1038/nature06505, 2008.

1455 Scanlon, B. R., Jolly, I., Sophocleous, M. and Zhang, L.: Global impacts of conversions from natural to agricultural ecosystems on water resources: Quantity versus quality, *Water Resources Research*, 43, doi: 10.1029/2006WR005486, 2007.

Scargle, J.: Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data, *The Astrophysical Journal*, 263, 835-853, 1460 1982.

Scargle, J.: Studies in astronomical time series analysis. III-Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data, *The Astrophysical Journal*, 343, 874-887, 1989.

Schilli, C., Lischeid, G. and Rinklebe, J.: Which processes prevail?: Analyzing 1465 long-term soil solution monitoring data using nonlinear statistics, *Geoderma*, 158, 412-420, doi: 10.1016/j.geoderma.2010.06.014, 2010.

Schuetz, T., Gascuel-Oudou, C., Durand, P. and Weiler, M.: Nitrate sinks and sources as controls of spatio-temporal water quality dynamics in an agricultural headwater catchment, *Hydrology and Earth System Sciences*, 20, 843-857, doi: 1470 10.5194/hess-20-843-2016, 2016.

Sen, P. K.: Estimates of the Regression Coefficient Based on Kendall's Tau, *Journal of the American Statistical Association*, 63, 1379-1389, doi: 10.1080/01621459.1968.10480934, 1968.

1475 Sivakumar, B.: Dominant processes concept in hydrology: moving forward, *Hydrological Processes*, 18, 2349-2353, doi: ~~DOI~~:10.1002/hyp.5606, 2004.

Stumm, W. and Morgan, J.: *Aquatic chemistry*, Wiley, 1996.

R Core Team 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/> (last access: 08 October 2017).

- 1480 Rode, M., Wade, A. J., Cohen, M. J., Hensley, R. T., Bowes, M. J., Kirchner, J. W., Arhonditsis, G. B., Jordan, P., Kronvang, B., Halliday, S. J., Skeffington, R. A., Rozemeijer, J. C., Aubert, A. H., Rinke, K. and Jomaa, S.: Sensors in the Stream: The High-Frequency Wave of the Present, *Environmental Science & Technology*, 50, 10297–10307, doi: [10.1021/acs.est.6b02155](https://doi.org/10.1021/acs.est.6b02155), 2016.
- 1485 Tenenbaum, J., Silva, V. and Langford, J.: A global geometric framework for nonlinear dimensionality reduction, *Science*, 290, 2319-2323, 2000.
- Tesmer, M., Möller, P., Wieland, S., Jahnke, C., Voigt, H. and Pekdeger, A.: Deep reaching fluid flow in the North East German Basin: origin and processes of groundwater salinisation, *Hydrogeology Journal*, 15, 1291-1306, doi: 10.1007/s10040-007-0176-y, 2007.
- 1490 Theil, H.: A rank-invariant method of linear and polynomial regression analysis, 1, 2 and 3, 53, 386-392, 1950.
- Tunaley, C., Tetzlaff, D., Lessels, J. and Soulsby, C.: Linking high-frequency DOC dynamics to the age of connected water sources, *Water Resources Research*, 52, 5232–5247, doi: [10.1002/2015WR018419](https://doi.org/10.1002/2015WR018419), 2016.
- 1495 Usunoff, E. J. and Guzmán-Guzmán, A.: Multivariate Analysis in Hydrochemistry: An Example of the Use of Factor and Correspondence Analyses, *Ground Water*, 27, 27-34, doi: [10.1111/j.1745-6584.1989.tb00004.x](https://doi.org/10.1111/j.1745-6584.1989.tb00004.x), 1989.
- Van der Maaten, L., Postma, E. and van den Herik, H.: Dimensionality Reduction: A Comparative Review, *TiCC-TR 2009-005*, 2009.
- 1500 Wade, A. J., Palmer-Felgate, E. J., Halliday, S. J., Skeffington, R. A., Loewenthal, M., Jarvie, H. P., Bowes, M. J., Greenway, G. M., Haswell, S. J., Bell, I. M., Joly, E., Fallatah, A., Neal, C., Williams, R. J., Gozzard, E. and Newman, J. R.: Hydrochemical processes in lowland rivers: insights from in situ, high-resolution monitoring, *Hydrology and Earth System Sciences*, 16, 4323–4342, doi: [10.5194/hess-16-4323-2012](https://doi.org/10.5194/hess-16-4323-2012), 2012.
- Weyer, C., Peiffer, S., Schulze, K., Borken, W. and Lischeid, G.: Catchments as

heterogeneous and multi-species reactors: An integral approach for identifying biogeochemical hot-spots at the catchment scale-, Journal of Hydrology, 519, Part B, 1510 1560-1571, doi: 10.1016/j.jhydrol.2014.09.005, 2014.

Weymann, D., Geistlinger, H., Well, R., von der Heide, C. and Flessa, H.: Kinetics of N<sub>2</sub>O production and reduction in a nitrate-contaminated aquifer inferred from laboratory incubation experiments, Biogeosciences, 7, 1953-1972, doi: 10.5194/bg-7-1953-2010, 2010.

1515 Wriedt, G., Spindler, J., Neef, T., Meißner, R. and Rode, M.: Groundwater dynamics and channel activity as major controls of in-stream nitrate concentrations in a lowland catchment system?, Journal of Hydrology, 343, 154-168, doi: 10.1016/j.jhydrol.2007.06.010, 2007.

1520 Yue, S., Pilon, P., Phinney, B. and Cavadias, G.: The influence of autocorrelation on the ability to detect trend in hydrological series, Hydrological Processes, 16, 1807-1829, doi: ~~DOI:~~ 10.1002/hyp.1095, 2002.

1525 Zhang, Y.-C., Slomp, C. P., Broers, H. P., Passier, H. F. and Cappellen, P. V.: Denitrification coupled to pyrite oxidation and changes in groundwater quality in a shallow sandy aquifer, Geochimica et Cosmochimica Acta, 73, 6716-6726, doi: 10.1016/j.gca.2009.08.026, 2009.

[Zhang, Q., Harman, C. J., and Kirchner, J. W.: Evaluation of statistical methods for quantifying fractal scaling in water-quality time series with irregular sampling, Hydrology and Earth System Sciences, 22, 1175-1192, doi: 10.5194/hess-22-1175-2018, 2018.](#)

1530

## Appendix A

### Lomb-Scargle

A given discrete time series  $Y(t_i)$  with  $(i = 1, \dots, N)$  and centred around zero can be described as a superposition from sin- and cos-terms with amplitudes  $a$  and  $b$ , time  
1535  $t_i$ , angular frequency  $\omega = 2\pi f$  and a noise term  $n_i$ .

$$Y(t_i) = a \cos \omega t_i + b \sin \omega t_i + n_i \quad (1)$$

Lomb (1976) introduced an additional factor Tau to consider for deviations from the evenly spaced case.

$$\tau_j = \frac{1}{2\omega_j} \bullet \arctan 2 \left( \frac{\sum_i^N \sin 2\omega_j (t_i - t_{ave})}{\sum_i^N \cos 2\omega_j (t_i - t_{ave})} \right) \quad (2)$$

1540

The constant  $t_{ave} = \min(t) - \max(t)$  scales the term to the centre of the period covered by the series for every frequency  $j$ . If the starting point of the series is set to zero  $t_{ave}$  enables to correct for offsets between the spectral components and thus allows to correctly reconstruct the original series out of its spectral components (Hocke 1998;

1545

Hocke and Kämpfer, 2009).

With these two extensions of the time term, equation 1 can be rewritten as

$$Y(t_i) = A \cos(\omega(t_i - \tau - t_{ave}) + \phi) + n_i \quad (3)$$

with amplitude  $A = \sqrt{a^2 + b^2}$  and phase  $\phi = \arctan(b/a)$ .

The Lomb-Scargle periodogram  $P_N(\omega)$  (equation 4) normalized with the total variance  
1550 of the data  $\sigma^2$  equals the linear least square fit of the time series model in equations 1 and 3 for a certain frequency (Lomb, 1976; Press et al., 2007).

$$P_N(\omega) = \frac{1}{2\sigma^2} \left\{ \frac{\left( \sum_i^N Y(t_i) \cos[\omega_j(t_i - \tau - t_{ave})] \right)^2}{\sum_i^N \cos^2[\omega_j(t_i - \tau - t_{ave})]} + \frac{\left( \sum_i^N Y(t_i) \sin[\omega_j(t_i - \tau - t_{ave})] \right)^2}{\sum_i^N \sin^2[\omega_j(t_i - \tau - t_{ave})]} \right\} \quad (4)$$

The amplitudes  $a$  and  $b$  can be computed out of the square root of the corresponding sin- and cos-terms of the normalized Lomb-Scargle periodogram, which yields the  
 1555 normalized power spectral density at certain frequencies (Hocke and Kämpfer, 2009).

$$a = \sqrt{\frac{2}{N}} \frac{\sum_i^N Y(t_i) \cos[\omega_j(t_i - \tau - t_{ave})]}{\sqrt{\sum_i^N \cos^2[\omega_j(t_i - \tau - t_{ave})]}} \quad b = \sqrt{\frac{2}{N}} \frac{\sum_i^N Y(t_i) \sin[\omega_j(t_i - \tau - t_{ave})]}{\sqrt{\sum_i^N \sin^2[\omega_j(t_i - \tau - t_{ave})]}} \quad (5)$$

Different modified series can be reconstructed out of any set of spectral components. So the method might be used i.e. as band-pass-filter or filling of gaps in the series (Hocke and Kämpfer, 2009).

1560 The number of frequencies in which the series is decomposed is calculated with the empirical formula derived out of Monte Carlo simulations by Horne and Baliunas (1986) (Glynn et al., 2006; Press et al., 2007).

$$N_{indep} \approx -6.362 + 1.193N + 0.00098N^2 \quad (6)$$

The false-alarm probability or statistical significance level  $p$  of the  $P_N(\omega)$  value at a  
 1565 certain frequency is calculated with equation (Scargle, 1982; Glynn et al., 2006; Press et al., 2007).

$$p = 1 - (1 - e^{-z})^M \quad (7)$$

$M$  is the number of test frequencies which is here set to  $N_{indep}$  and  $z$  is the tested value of  $P_N(\omega)$  at a certain frequency. To diminish aliasing, which means reappearing  
 1570 of higher frequencies' power in the power of lower ones, the highest test frequency is set to the Nyquist-rate  $f_{max} = f_{Nyquist} = 1/(2\Delta t)$ . Because of the irregular sampling, the

sampling rate  $\Delta t$  is approximated here by the average sampling interval  $\Delta t = (t_N - t_1) / N$ . The lowest test frequency is the inverse of the sampling range  $f_{min} = 1 / (t_N - t_1)$  (Scargle, 1982; Press et al., 2007).

1575 Although  $N_{indep}$  should be the number of independent frequencies in the signal it is possible that frequencies lying close to each other 'share' the same underlying trigger. This leakage of power is promoted by the uneven sampling and oversampling of the frequency domain  $M > N$  (Scargle, 1989; Horne and Baliunas, 1986). In addition, the effect may be enhanced because of local high sampling density,  
1580 autocorrelation in the data or very strong momentum of the underlying trigger.

With regard to these circumstances, which apply especially for the groundwater level series in this study, only the 'dominant' frequencies were used to identify seasonal patterns. The term 'dominant' frequency is used here for the peaks in between groups of significant frequencies. If such groups build 'significance-plateaus' the  
1585 median of this plateau is taken as dominant frequency.



## Supplementary material

1590 Table S1 Stream water sampling sites. The abbreviation in the ID refers to the corresponding catchment. N: number of samples.

ID	N	Easting	Northing	Catchment
D_112	126	3426969	5916330	Dauergraben
U_128	114	3423416	5907370	Ucker
St_133	124	3420262	5891835	Stierngraben
S_118	118	3421173	5907839	Strom
S_120	1	3418025	5906225	Strom
S_121	23	3416348	5905013	Strom
S_122	1	3412048	5903419	Strom
Q_93	126	3422251	5908887	Quillow
Q_95	125	3420582	5910416	Quillow
Q_96	11	3420084	5913122	Quillow
Q_97	126	3419850	5913404	Quillow
Q_98	127	3417941	5913091	Quillow
Q_100	110	3412572	5912708	Quillow
Q_104	71	3409712	5912268	Quillow
Q_106	12	3406372	5912814	Quillow
Q_102	11	3410569	5911755	Quillow
Q_103	8	3408376	5910401	Quillow
P_107	78	3410047	5912392	Peege
P_108	61	3408727	5914397	Peege
P_109	8	3410232	5916180	Peege
P_110	51	3410858	5917416	Peege

1595 Table S2 Sampling sites for groundwater quality and groundwater level. The abbreviation in the ID refers to the corresponding catchment. The subscripts Gs = shallow groundwater and Gd = deep groundwater give additional information on the respective groundwater layer. All groundwater wells are inside the Quillow catchment. N: number of samples.

ID	N	Easting	Northing	Depth of screen m (a.s.l)	Depth of screen m below ground
Gd_205	2	3416412	5911941	40.55 - 38.55	15 - 17
Gd_204	25	3412546	5912702	49 - 47	16 - 18
Gs_200	6	3410020	5912439	74.10 - 73.10	4.0 - 5.0
Gs_199	18	3409934	5912302	72.20 - 71.20	3.0 - 4.0
Gd_198	28	3409934	5912302	51.27 - 53.27	22 - 24
Gs_202	11	3409863	5912702	74.14 - 73.14	4.0 - 5.0
Gd_201	25	3409863	5912702	65.79 - 63.79	12.5 - 14.5
Gd_203	25	3409764	5912942	63.46 - 61.46	16 - 18

1600

Table S3 Measurement details of the analysed variables. Before the data analysis  $\text{NH}_4^+$  was calculated from  $\text{NH}_4\text{-N}$ ,  $\text{PO}_4^{3-}$  from  $\text{o-PO}_4\text{-P}$ , and the concentration of  $\text{HCO}_3^-$  was converted from  $\text{mmolL}^{-1}$  to  $\text{mgL}^{-1}$ .

Abbreviation	Parameter	Unit	Measuring accuracy / detection limit	missing values in %	n samples < detection limit in %
Stream water and groundwater					
pH	pH value		0.01	0	0
Eh	Redox potential	mV	1	0.57	0
EC	Electric conductivity	$\mu\text{Scm}^{-1}$	1	0	0
Temp	Water temperature	$^{\circ}\text{C}$	0.1	0	0
$\text{O}_2$	Oxygen	$\text{mgL}^{-1}$	0.1	1.91	0.25
$\text{NH}_4\text{-N}$	Ammonium nitrogen	$\text{mgL}^{-1}$	0.01	0.57	0.76
$\text{o-PO}_4\text{-P}$	Phosphorus of orthophosphate	$\text{mgL}^{-1}$	0.01	0	37.53
DOC	Dissolved organic carbon	$\text{mgL}^{-1}$	0.05	3.44	0
Anions					
$\text{Cl}^-$	Chloride	$\text{mgL}^{-1}$	0.03	0	0
$\text{NO}_2^-$	Nitrite	$\text{mgL}^{-1}$	0.03	2.54	65.52
$\text{NO}_3^-$	Nitrate	$\text{mgL}^{-1}$	0.03	0.38	2.93
$\text{SO}_4^{2-}$	Sulfate	$\text{mgL}^{-1}$	0.02	1.34	0
Cations					
$\text{Na}^+$	Sodium	$\text{mgL}^{-1}$	0.01	0	0
$\text{K}^+$	Potassium	$\text{mgL}^{-1}$	0.02	0	0
$\text{Mg}^{2+}$	Magnesium	$\text{mgL}^{-1}$	0.02	0	0
$\text{Ca}^{2+}$	Calcium	$\text{mgL}^{-1}$	0.03	0	0
Only groundwater					
$\text{Fe}^{2+}$	Iron(II)	$\text{mgL}^{-1}$	0.03	0	8.57
$\text{HCO}_3^-$	Hydrogen carbonate	$\text{mmolL}^{-1}$	0.01	6.43	0

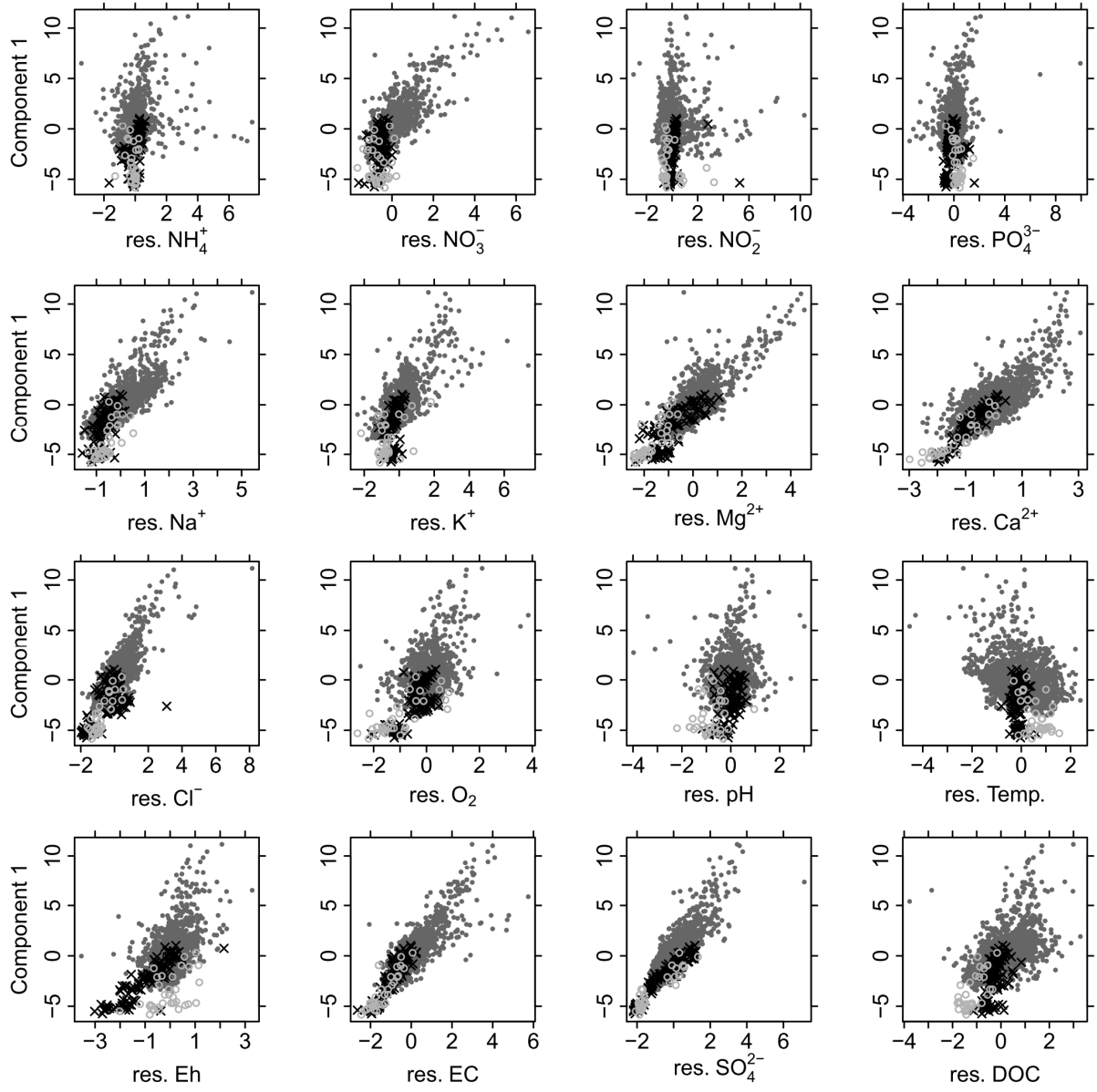
1605

Table S4 Site-specific cumulated R<sup>2</sup> of the reproduction of the interpoint distances of the data in the projection by the first four components of Isomap at sites with n > 15.

Subscripts: P = Peege, Q = Quillow, S = Strom, St = Stierngraben, U = Ucker, D =

1610 Dauergraben, Gs = shallow groundwater, Gd = deep groundwater.

ID	Gd_203	Gd_201	Gd_198	Gd_204	Gs_199	P_110	P_108	<del>P</del> <del>Q</del> <sub>7</sub> <sup>-10</sup>	Q_104	Q_100
N	25	25	28	25	18	51	61	78	71	110
Cp. 1	0.1	0. <del>62</del> <del>59</del>	0.01	0.1 <del>54</del>	0.01	0.25	0.3 <del>67</del>	0.1 <del>12</del>	0.21	0.27
Cp. 2	0.33	0.6 <del>96</del>	0.2 <del>54</del>	0.3 <del>31</del>	0.08	0.38	0.4 <del>68</del>	0.27	0.33	0.61
Cp. 3	0.49	0.79	0.6 <del>3</del>	0.5 <del>4</del>	0.2 <del>1</del>	0.45	0.5 <del>65</del>	0.5 <del>34</del>	0.41	0.64
Cp. 4	0. <del>549</del>	0.8	0.74	0.5 <del>56</del>	0.29	0. <del>596</del>	0.97	0.65	0.7 <del>45</del>	0.8
ID	Q_98	Q_97	Q_95	Q_93	S_121	S_118	St_133	U_128	D_112	
N	127	126	125	126	23	118	124	114	126	
Cp. 1	0.33	0.3 <del>65</del>	0.2	0.11	0.35	0.15	0.3	0.11	0.54	
Cp. 2	0.43	0.46	0.31	0.24	0.43	0.2 <del>54</del>	0.45	0.2 <del>78</del>	0.64	
Cp. 3	0.59	0.58	0.35	0.3	0.6 <del>87</del>	0.39	0.59	0.4 <del>1</del>	0.73	
Cp. 4	0.72	0.66	0.6 <del>78</del>	0.72	0.6 <del>65</del>	0. <del>769</del>	0.67	0.5 <del>23</del>	0.83	



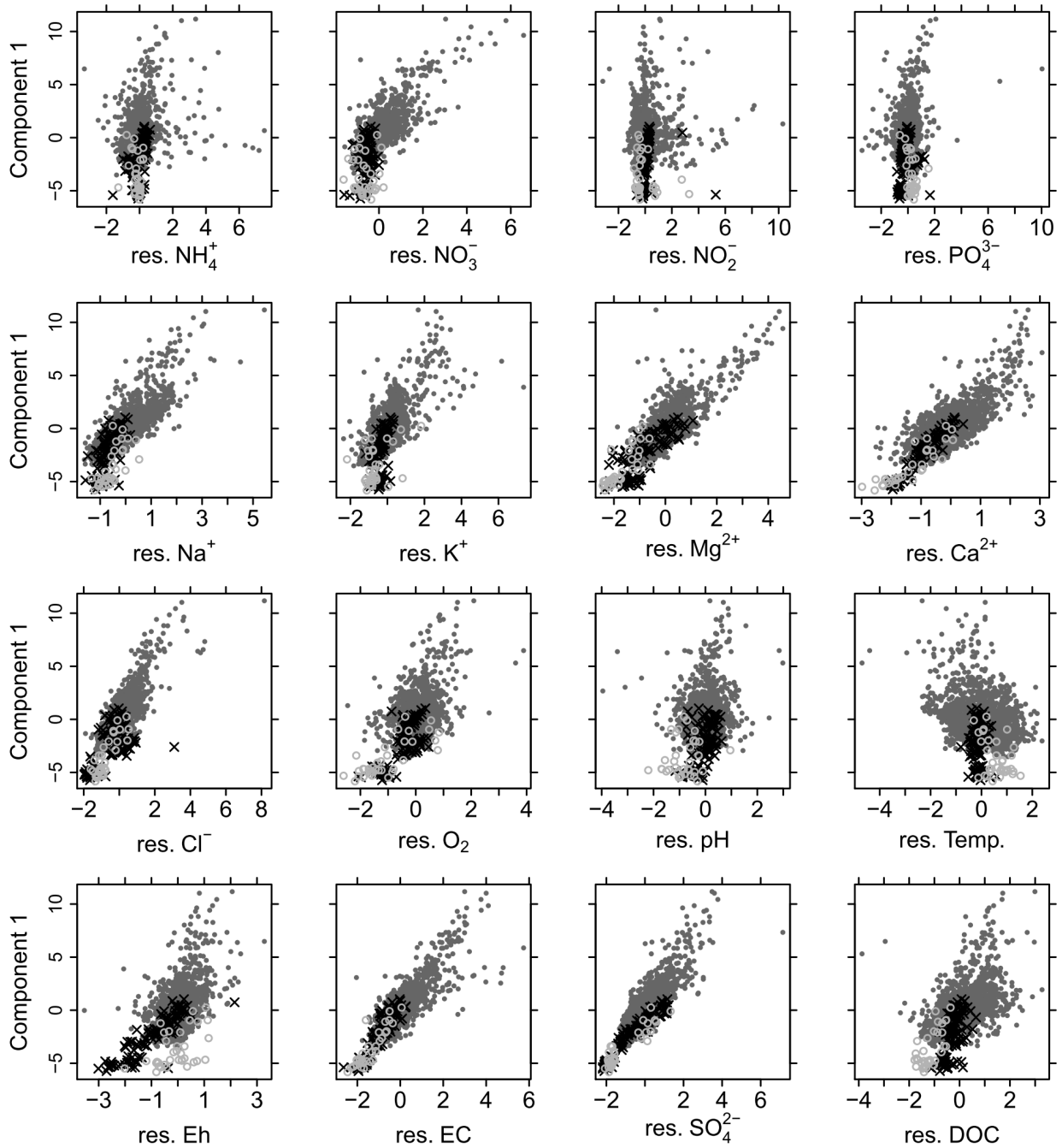
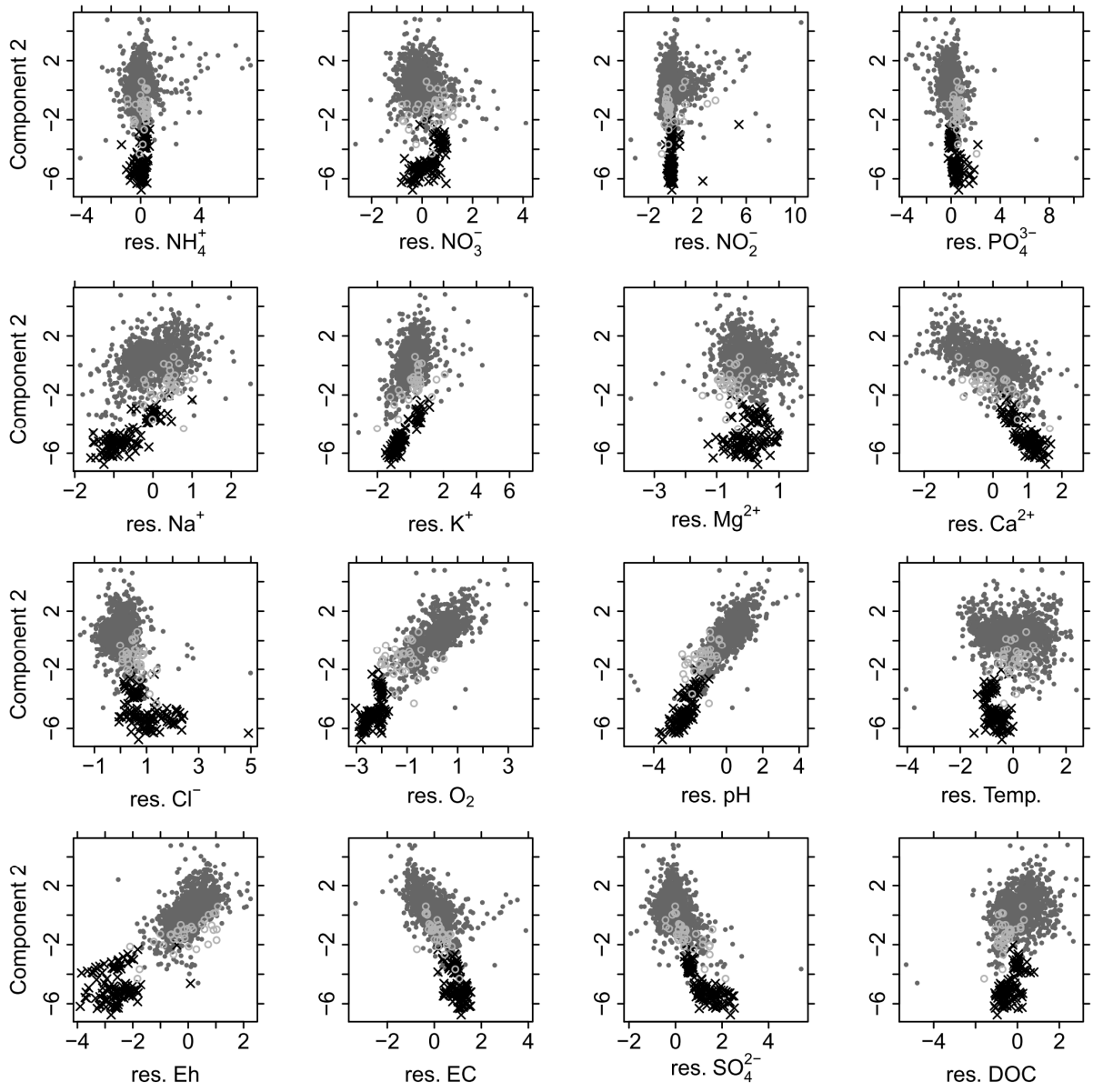


Figure S1 Residuals of the multiple linear regression of selected single variables and component 2-4 vs. scores of component 1. Grey filled dots: stream water. Light grey open circles: shallow groundwater. Black x-mark: deep groundwater.



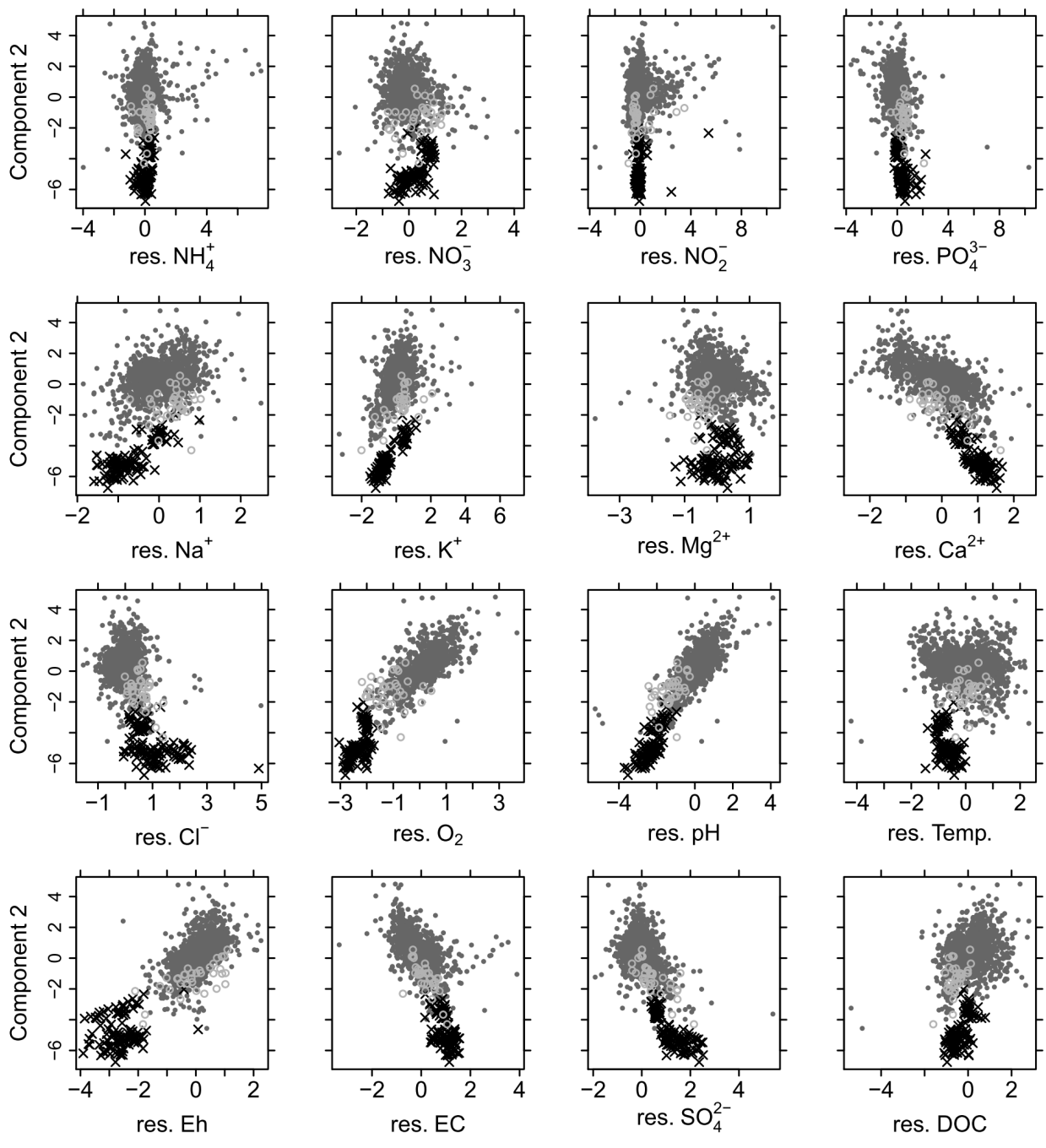
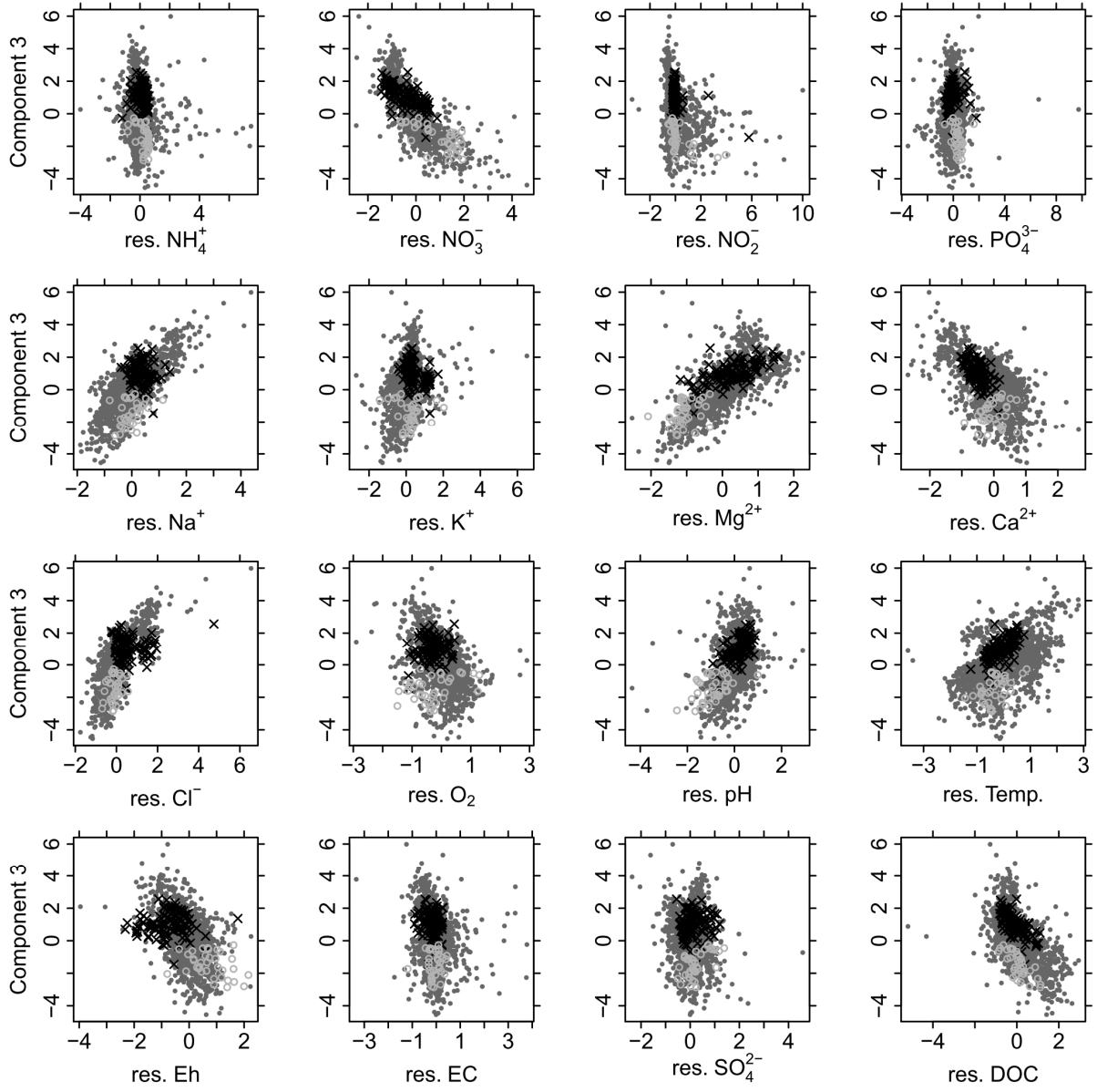


Figure S2 Residuals of the multiple linear regression of selected variables on component 1, 2 and 4 vs. scores of component 2. Grey filled dots: stream water. Light grey open circles: shallow groundwater. Black x-mark: deep groundwater.





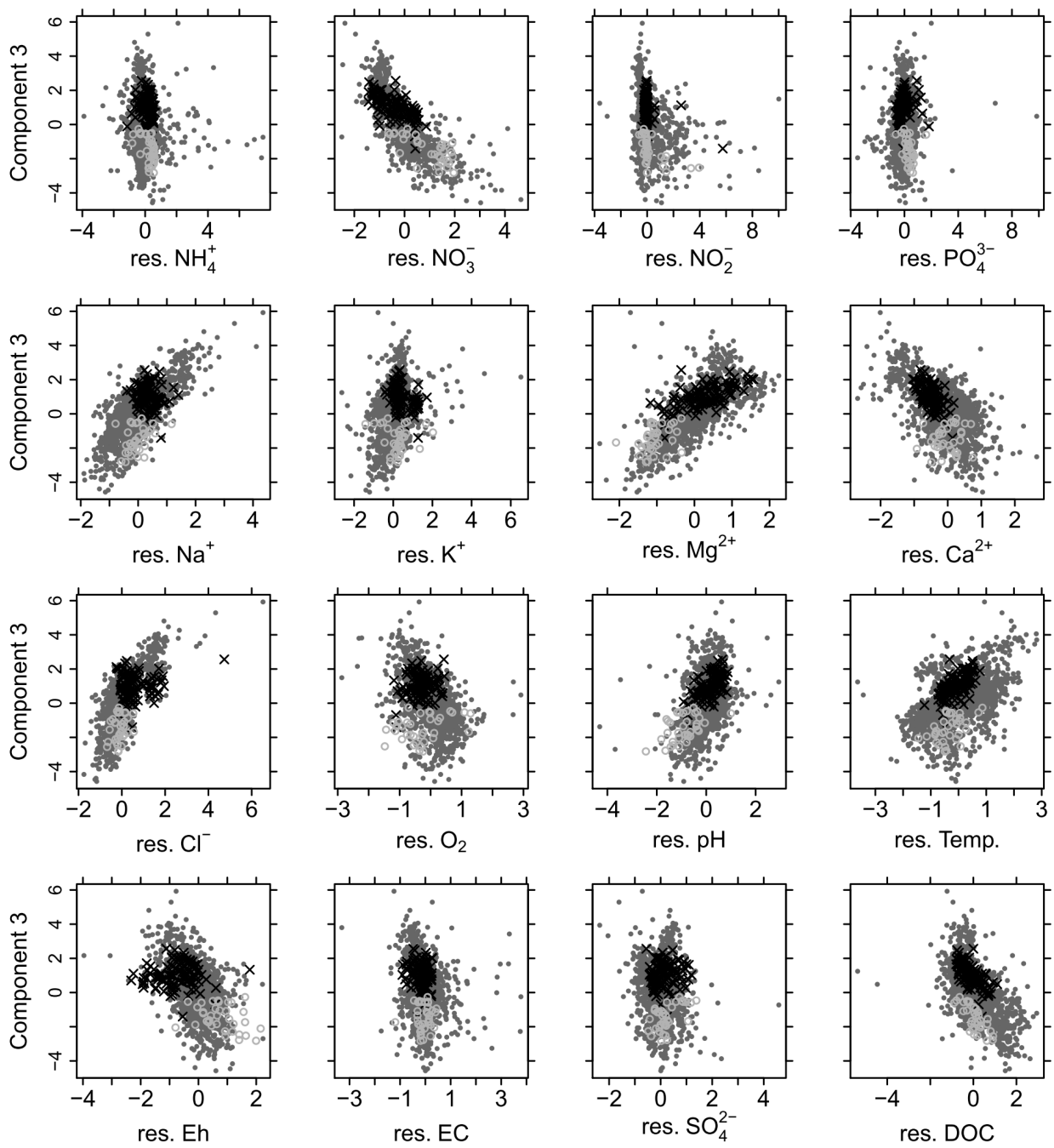
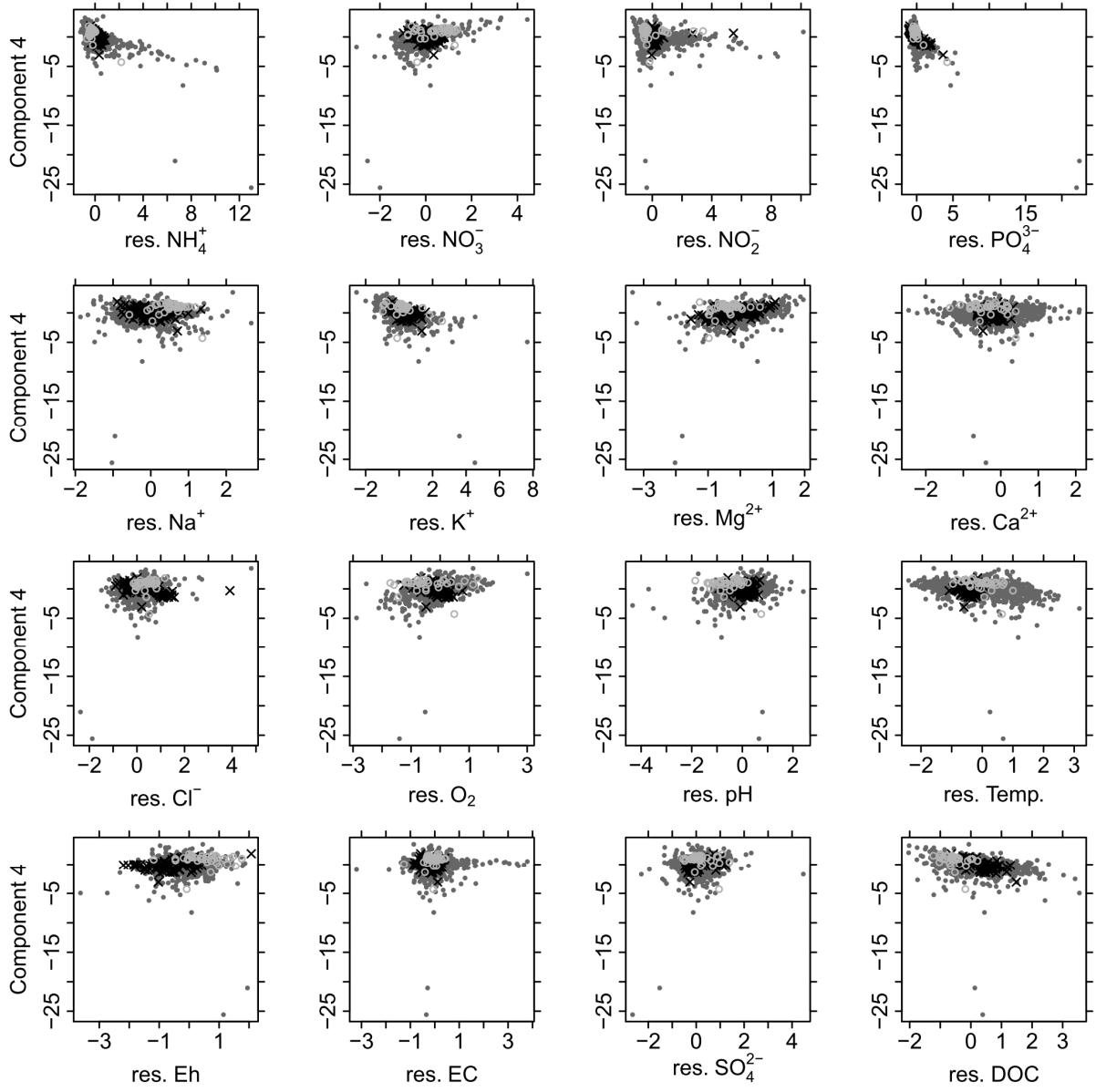


Figure S3 Selection of residuals of the multiple linear regression of single variables and component 1, 2 and 4 vs. scores of component 3. Grey filled dots: stream water. Light grey open circles: shallow groundwater. Black x-mark: deep groundwater.



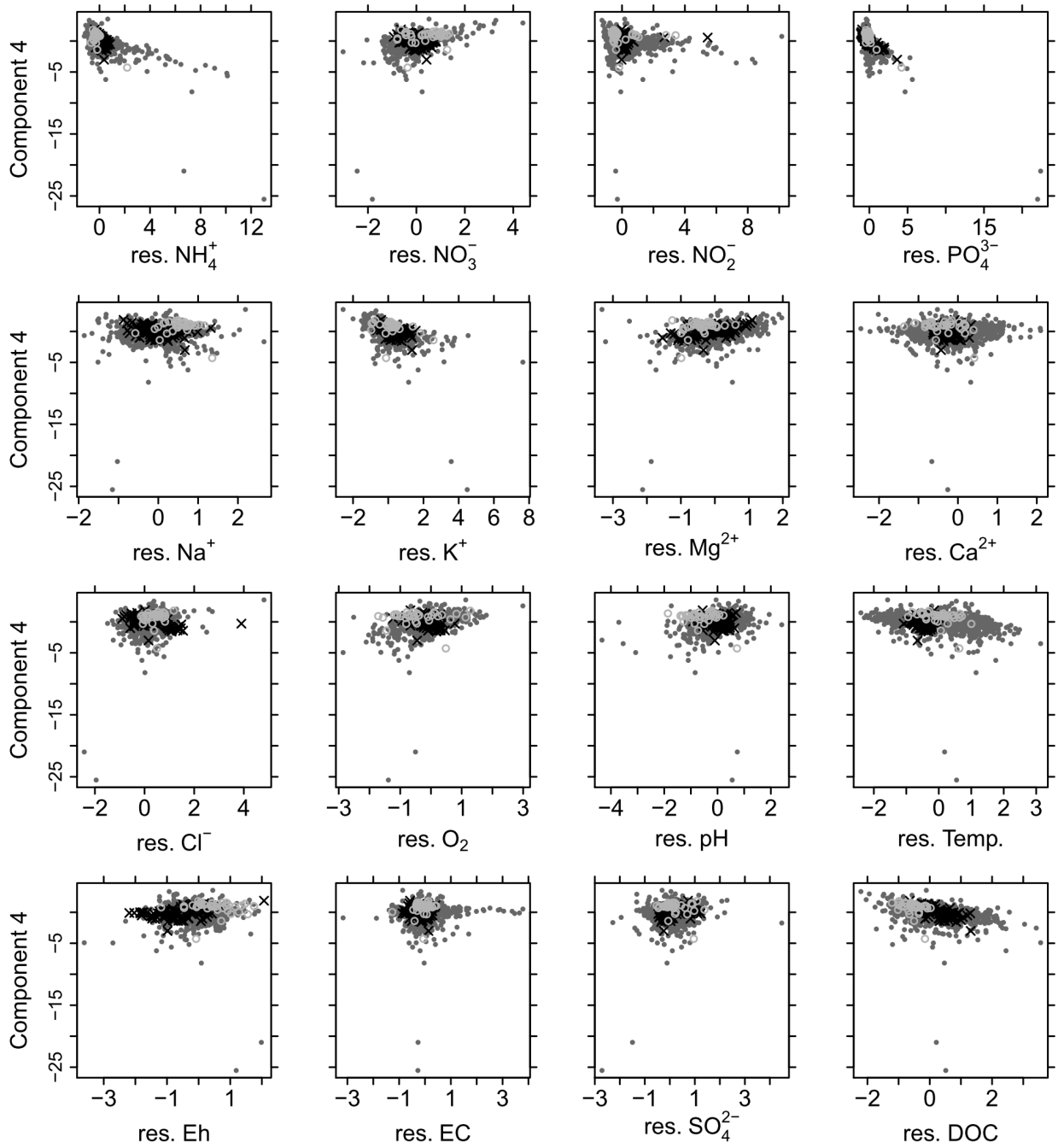


Figure S4 Selection of residuals of the multiple linear regression of single variables and component 1-3 vs. scores of component 4. Grey filled dots: stream water. Light grey open circles: shallow groundwater. Black x-mark: deep groundwater.