Process identification by principal component analysis of river water-quality data

W. Petersen a,*, L. Bertino b, U. Callies a, E. Zorita a

a GKSS Research Centre, Institute of Hydrophysics, Max-Planck-Str., D-21502 Geesthacht, Germany
b Ecole de Mines de Paris, Centre de Géostatistique, 35, rue Saint-Honoré, F-77305 Fontainebleau Cedex, France

Abstract

Time series of nutrient concentrations and related water quality parameters taken at several locations along the River Elbe were subjected to multivariate statistical analysis. The main question underlying this study is concerned with whether known interactions between water quality variables can be recovered as statistically significant covariance patterns. For this purpose, the standard technique of principal component analysis (PCA) was applied. Raw data and deviations from an estimated seasonal cycle were analysed. In both cases, two leading patterns of covariance was obtained, one discharge-dependent and the other related to biological activities. Linear regression modelling based on discharge and temperature was used to approximately eliminate the impact of meteorological forcing; this led to a large reduction of the seasonal component. The remaining partial variance of water-quality variables could be shown to be dominated by biological activities for which temperature is of secondary importance. Amplitudes of the pattern related to biological processes are much less correlated between different stations than those of the pattern induced by spatially homogenous discharge. The analysed covariance patterns agree well with general knowledge about basic dynamical processes in the river. Therefore, multivariate statistical analysis offers an objective method to estimate the observed strengths of the given processes that involve simultaneous changes of several water-quality parameters. Such an assessment is a prerequisite when observations are to be compared with corresponding results from process-oriented numerical models in order to increase the knowledge about the nutrient system. A related application would be to use it to identify the number of degrees of freedom needed to appropriately describe the nutrient system’s variability. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Exploratory statistical analysis; Elbe River; Nutrients; Oxygen; Primary production

1. Introduction

During its course from eastern Germany to the German Bight the Elbe River transports a high nutrient load into the North Sea. The yearly loads are in the range of 100,000–200,000 t/a N and 4000–5000 t/a P (ARGE Elbe, 1995a). The main sources of nutrients are urban, industrial or agricultural inputs. The calculated emissions are always higher than the measured loads in the river (Behrendt, 1994) reflecting a non-conservative behaviour and losses during transport in the river.
On one hand, the nutrient concentrations control
the biological activity (such as algal and bacterial
growth), resulting in changes in the oxygen bud-
get. On the other hand, biological activity con-
trols concentrations of nutrients (e.g. NH₄ or
NO₃) and also physical properties (dissolved, par-
ticle bound) that influence the nutrient transport
behaviour, including deposition. These interac-
tions have to be known in order to model water
quality and to predict nutrient concentrations and
loads.

Nutrient concentrations have been measured in
the Elbe for more than 20 years. At that time, the
Elbe nutrient data have been used to quantify the
transport of nutrients to the North Sea (ARGE
Elbe, 1995b). The changes of nutrient species and
their influence on the oxygen budget in the tidal
part of the Elbe since the reunification of Ger-
many in 1990 have been investigated by Berge-
mann et al. (1996) and Petersen et al. (1999), and
the fact that oxygen concentration is mainly con-
trolled by primary production has been success-
fully modelled in a simple model by Schroeder
(1997). However, so far no empirical approach
has been used in the analysis of this data set.

Long-term monitoring programs of water qual-
ity produce large sets of data which are often
difficult to interpret (Dixon and Chiswell, 1996).
Most data analyses focus on primarily qualitative
discussions of single effects or variables. The use-
fulness of reducing the complexity of water-qual-
ity data by exploratory statistical analysis has
been emphasised by several authors (e.g. Brown et
al., 1996). Recently, Vega et al. (1998) applied
statistical methods to separate the impact of pol-
lution on the quality of river water from that of
seasonal effects.

The empirical approach used in this paper was
to summarise and condense the information con-
tained in the measured data by the use of ex-
ploratory statistics, i.e. without modelling and
without any a prior knowledge about the nutrient
interactions. Principal component analysis (PCA)
was applied to multiconstituent chemical mea-
surements and physical parameters using a num-
er of different approaches. The aims of this
study were (1) to empirically identify the main
processes during the transport of nutrients and (2)
to provide a reduced description of the system
using a few significant and interpretable PCA
patterns which reflect the most relevant processes
and can be used in simplified diagnostic models.

2. Data

Data from manual measurements taken at 14
stations on the Elbe River in Germany, from
Wittenberg (Elbe 214 km) to the North Sea (Elbe
725 km), was used. The distance between any two
neighbouring stations ranges from 4 to 100 km.
Measurements were taken every 14 days by the
water authorities (Arbeitsgemeinschaft für die
Reinhaltung der Elbe (ARGE Elbe)).

A description of the data set can be found in
ARGE Elbe (1997). After the German reunifica-
tion in 1990, the mean nutrient concentrations
changed until 1993 when they seemed to have
reached a stable activity state. Since then, they
can be considered to have been stationary from 1
year to the next. Therefore, only 5 years (1993–
1997) of observations were selected for this study.
Some outlying measurements were removed from
the data, and measurements beyond the test sensi-
tivity limit were arbitrarily taken to be equal to
half the sensitivity. The investigated parameters
are shown in Table 1.

Only dissolved (from filtered water samples)
nutrients were considered as these are mainly
involved in biological cycles. A saturation index
(in percent) was used instead of oxygen concen-
trations to avoid the physical dependence of oxy-
gen concentrations on water temperature. For
Fig. 1. Schematic view of the stations in the Elbe River investigated and a table of abbreviations used.

Fig. 2 depicts, for most of the variables a more or less marked annual cycle. High water discharges correspond with reduced conductivity (due to dilution effects) and higher concentrations of nitrate. The seasonal behaviour is reflected by high temperatures in the summer which correspond with increased pH and decreased phosphate and ammonia. At first sight, a major difference between the two stations is only recognisable for oxygen. At station Schnackenburg, oxygen increases during the summer period with high water temperatures, whereas at station Seemannshöft the oxygen saturation index sharply declines at this time. This characteristic difference between the estuarine and the riverine part of the Elbe becomes more clear when considering the mean seasonal cycle of selected parameters as calculated in chapter 4.2 which have been averaged over all stations in the respective river section (Fig. 3). The figure indicates different processes in the two river stretches during the months from April to October with respect to oxygen, pH and phosphate. Corresponding to an increase of oxygen in the riverine part the oxygen saturation index decreases in the estuarine part, indicating dominating oxygen production in the riverine part and more oxygen consumption in the estuarine part. A corresponding behaviour between the two river sections is also depicted for pH and phosphate. Only ammonia and nitrate (not shown in this figure) show a very similar behaviour in both sections. Therefore, in the following, all calculations were carried out separately for the two river stretches.
3. Methods

3.1. Principal component analysis (PCA)

Principal component analysis (Johnson and Wichern, 1992) is a technique widely used for reducing the dimensions of multivariate problems. The basic idea is to find a small number of uncorrelated linear combinations $pc_i(t)$ of the $N$ original variables $v_j(t)$-with $t$ denoting time — that is able to explain the essential part of the covariance structure in the data set of interest

$$pc_i(t) = \sum_{j=1}^{N} eof_j \frac{v_j(t) - \bar{v}_j}{\sigma_j}.$$  

(1)

Here, $\sigma_j$ is the standard deviation (S.D.) of variable $v_j$, which is used to remove from all variables their different physical dimensions. The vector components $\bar{v}_j$ denote the mean values of the observations. The new synthetic variables $pc_i$ are called principal components, and the vectors or patterns $eof_i$ with components $eof_j$ that map the original variables onto the principal components are called empirical orthogonal functions (EOFs, von Storch and Zwiers, 1999).

By scaling all observed variables with their standard deviations, the non-dimensional total variance of the observed data vector, being defined as the sum of variances arising from its $N$ individual components, becomes $N$. The first principal component, $pc_1$, is defined as that linear combination of the scaled vector components that exhibits the maximum variance $\lambda_1 = \text{var}(pc_1)$ among all possible combinations with coefficient vectors $eof^1$ of unit length. Similarly, the second principal component shows the maximum vari-

---

Fig. 2. Typical time series of the investigated parameters in the estuarine part (Seemannshöft, 629 km) and the riverine part (Schnackenburg, 475 km) of the Elbe River.
Fig. 3. Comparison of mean seasonal cycle of the oxygen saturation index, pH, phosphate and ammonia, being averaged over all stations in the riverine and estuarine part of the Elbe river, respectively (deviations from the average: 5 and 20%).

The EOF-patterns can conveniently be obtained as the eigen-vectors of the correlation matrix associated with the data (Johnson and Wichern, 1992). The variances \( \lambda_i = \text{var}(\text{pc}_i) \) of the principal components, i.e. of the projections of the original variables onto the patterns \( \text{eof}_i \), are obtained from the corresponding eigenvalues. The sum of these eigen-values equals the total variance \( N \) of the original scaled variables. Accordingly, the proportion of total variance explained by the \( i \)th principal component (or EOF) is

\[
\frac{\lambda_i}{\sum \lambda_i} = \frac{\lambda_i}{N} \tag{2}
\]

If variables in a data set vary partly coherently due to them being involved in common processes, a parsimonious description of their covariance structure may be achieved by retaining only the leading principal components with variances \( \lambda_i \) significantly larger than \( 1/N \).

3.2. Partial principal component analysis

Much of a nutrient’s variability can be explained by considering the external meteorological forcing which is reflected in variations of water temperature (WT) and discharge (DC). To interpret time series of nutrient observations, it is instructive to separate this externally induced signal from variations due to internal dynamical processes that depend on other factors.
Given some standardised variable $X$, the simplest estimates, $\hat{X}$, of its component governed by DC or WT or both of them are obtained by linear modelling,

$$\hat{X} = a_X \text{DC}$$

(3)

$$\hat{X} = b_X \text{WT}$$

(4)

$$\hat{X} = c_X \text{DC} + d_X \text{WT}$$

(5)

with the empirically specified regression coefficients $a_X$, $b_X$, $c_X$ and $d_X$. A consequence of the least squares regression approach is that the variance of $X$, $\text{var}(X)$, decomposes into the sum

$$\text{var}(X) = \text{var}(\hat{X}) + \text{var}(X - \hat{X})$$

(6)

due to the orthogonality between $\hat{X}$ and the residual $X - \hat{X}$ (Whittaker, 1990). The second term in Eq. (6) represents the partial variance of $X$ to be observed when WT and DC, and therefore also $\hat{X}$, are held fixed. If $X$ is a vector made up by several observed variables, the partial variance becomes a matrix. This partial covariance matrix can again be subjected to principal component analysis (in the following denoted as partial PCA) to find those dominant patterns of co-variation in the space of variables which cannot be explained by varying meteorological forcing.

In order to perform a PCA of the original data, the variables have to be standardised so that the covariance matrix $\text{var}(X)$ is a correlation matrix with unit diagonal. This can obviously not be true for the two partial matrices on the right-hand side of Eq. (6). In this paper, we decided to re-scale the partial variance matrix to have unit diagonal before subjecting it to PCA. This second scaling is not really necessary, since the physical dimensions of all variables in vector $X$ have already been removed at the very beginning of the statistical analysis. However, results with and without rescaling of residuals $X - \hat{X}$ do not differ significantly.

3.3. The bootstrap method

To assess the reliability of the EOF patterns analysed and discussed in this paper, the variability of individual EOF components was to be estimated. This is done in cases where experiments are repeated many times and the resulting sets of observations are all subjected to the same sort of analysis. In reality, only one single set of actual observations is available. However, computer intensive resampling techniques can be applied to simulate the process of repeated experiments by generating from the actual data set a series of synthetic data sets, each of which contains the same number of data points, as the original data set. Performing a PCA on each of the hypothetical data sets gives a series of simulated EOF patterns which may then be used to estimate the statistical accuracy of the analysed EOF pattern structure. The advantage of this computer-intensive analysis of uncertainty compared to analytical considerations is that no prior assumptions about the distribution of the data are needed.

The bootstrap method (Efron 1982; Efron and Tibshirani 1986) is a resampling method which generates the fake data sets in a particular way. Let, for instance, one data point comprise the observations of all the variables represented in a given EOF pattern and let the original data set contain $N$ data points of this type. To simulate a second data set of the same size, the bootstrap procedure randomly draws $N$ data points from the original data set, replacing the selected data point with an identical copy in the original data set after each draw. In other words, each data point can be selected repeatedly, and the simulated data set of size $N$ differs from the original data set by the fact that some random fraction of the data points (typically $\sim 37\%$; Press et al. 1992) is discarded while the other data points are duplicated several times. A large number of bootstrap samples can easily be generated in this way. All uncertainty estimates in the present study have been based on 1000 bootstrap samples. Treating these synthetic data as if they were real, the range between the 0.05 quantile and the 0.95 quantile is indicated for each EOF pattern. This range roughly corresponds to plus or minus twice the standard deviation. One has to be cautious in interpreting these quantiles as defining strict confidence intervals because the theoretical prerequisite of all the data points being independent is not satisfied by the present data set. Autocorrelations due to seasonal cycles or trends being present in
the data will generally cause an underestimation of uncertainty. Nevertheless, the error bars provide useful estimates of the accuracy of individual EOF components relative to the uncertainty of others.

4. Process identification by PCA

4.1. Riverine part: original data

First, EOF calculations were applied to the original data from the riverine stations from Wittenberg (214 km) to Boizenburg (559 km), upstream of the weir at Geesthacht. All observations of a particular variable (e.g. some nutrient) which were available from the different stations were pooled, i.e. measurements from different stations were treated as if they had all been made at the same location. Serial correlations between observations at different time were neglected.

4.1.1. PCA

The correlation matrix of the original data for the riverine stations is given in Table 2. Some correlations are very weak. The highest correlation found is that between ammonia and water temperature (correlation coefficient $=-0.7$). In some cases correlations between pairs of parameters do not provide the full amount of information that is available from plotted time series (Fig. 2).

After applying a PCA to this multi-variable data set, two leading EOFs can be distinguished that explain a significant portion (39 and 20%, respectively) of the total data set variance (Fig. 4).

4.1.1.1. First EOF

Within the part of variance described by this pattern (upper panel of Fig. 4), variations of temperature, pH and oxygen have an opposite sign in comparison with variations of ammonia and nitrate. The contribution of con-

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Correlation matrix when pooling all observations (original data set for the riverine stations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WT</td>
</tr>
<tr>
<td>WT</td>
<td>1.00</td>
</tr>
<tr>
<td>DC</td>
<td>1.00</td>
</tr>
<tr>
<td>PH</td>
<td>1.00</td>
</tr>
<tr>
<td>NH$_4$</td>
<td></td>
</tr>
<tr>
<td>PO$_4$</td>
<td></td>
</tr>
<tr>
<td>O$_2$s</td>
<td></td>
</tr>
<tr>
<td>NO$_3$</td>
<td></td>
</tr>
<tr>
<td>Cond</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4. Two leading EOF patterns (EOF 1 and EOF 2), the corresponding eigen-values (Eigen) and the explained variance (Var) from original data set in the riverine part of the Elbe River.
ductivity is negligible; phosphate and discharge are only weakly involved. The pattern can be interpreted in terms of biological activity as either primary production by algae or the subsequent microbial decomposition of the algae. During primary production of biomass (algae) by photosynthesis, the nutrients PO₄ and NH₄ as well as carbonate ions are assimilated, leading to a decrease in their concentrations. Removal of the acidifying carbonate ion increases the pH. At the same time, oxygen is released by the photosynthesis process. The inverse process is the process of biomineralization. As algae biomass is decomposed by micro-organisms, oxygen is consumed and nutrients and carbon dioxide are released; this results in lowered concentrations of oxygen and increased pH and nutrient concentrations. As the signs of the EOF loadings are arbitrary, the two processes cannot be distinguished by the EOF. We will call this first EOF pattern the 'biological mode'.

It seems reasonable to assume that reduced primary production is triggered by lower temperatures, probably as a result of reduced radiation intensity. This presumed causal relation seems to be confirmed by the temperature component in the biological mode. However, using the original data in the calculations of this mode, the effects caused by biological activity and the seasonal effect could not be clearly discriminated from one another. Both high temperatures and high biological activity mainly occur from spring to autumn.

4.1.1.2. Second EOF. The negative correlation between the nutrient concentrations and freshwater discharge (lower panel of Fig. 4) can be understood as dilution effect. Salt, governing the conductivity mainly comes from the Saale tributary, except for the stations in Estuary, for which the main source of salt is mixing with seawater. The main phosphate source is urban efflux. As these point sources are not related to the river discharge, the loads of salt and PO₄ are approximately constant; therefore, both the conductivity and the concentration of phosphate are negatively correlated with discharge due to dilution by higher discharge. See ARGE Elbe (1995b) for more details.

Nitrate concentrations do not exhibit a dilution effect. The main sources of nitrate are agricultural activities (ARGE Elbe, 1995b), and its increased drainage from fields after precipitation events produces a positive correlation with discharge.

Ammonia is known to be released in bovine breeding areas and waste plants, i.e. ammonia should more or less exhibit a dilution effect. However, it is also influenced by biological activity (assimilation by algae, conversion to NO₃ by nitrifying bacteria and release by the biomineralization of biomass). This biological activity can only to be taken into account from spring to autumn due to low water temperatures in the winter months.

According to these considerations, it seems justified to consider discharge as the common cause for the co-variation of concentrations and conductivity summarised in the pattern of the second EOF. Since the contribution of temperature as the second important external forcing parameter is negligible in the second EOF, we will call this pattern the 'discharge mode'.

Whether a certain EOF can be attributed either to a mainly biological or to a mainly physical (e.g. discharge) process can be assessed by considering the proportion of the original variables being associated to the pattern. The amount η that can be reproduced from the \( p_i \)th principal component \( p_i \) is given as:

\[
η_j(t) = \bar{η}_j + σ_j p_i(t) \text{eof}_j
\]  

(7)

In Fig. 5, predictions of discharge, water temperature and pH calculated according to Eq. (7) from the first and the second EOF, respectively, are compared to the corresponding original observations. Discharge and water temperature are external forcing parameters, while pH is a parameter characteristic of biological activity. When prediction of the parameters is made using only the first EOF (EOF₁), temperature and pH agree well with the measured data, while discharge is not predicted by the first EOF. When prediction is made using only the second EOF (representing the discharge mode), only the discharge agrees well with the original data, while water temperature and pH are not predicted by this EOF.
4.1.2. Separation of the impact of external forcing

So far, original data that have included the annual cycle was analysed. This approach might be questioned since slow variations on an annual time scale are primarily responses to systematic external forcing rather than random events. In Section 4.2, the PCA will be repeated, this time considering deviations of the actually observed values from estimated seasonal reference components. However, seasonal components estimated from time series covering just 5 years are rather uncertain and their usage may introduce a considerable amount of arbitrariness into the outcome of the statistical analysis. Therefore, the present section explores an approach to eliminate the impact of external forcing by regressing nutrient variables and related parameters on discharge and water temperature. The latter two parameters are supposed to be the dominant factors which make the aquatic system ‘feel’ the season. Probably the main weakness of this assumption resides in the neglect of radiation as a third external forcing factor, although variations of radiative fluxes can be expected to be partly represented by signals in the temperature records. The objective of this section is to analyse that proportion of the total variability in the data set which remains unpredictable when the actual values of discharge and temperature are known. For this purpose, the deviations of the normalised actual observations \( X \) from the corresponding estimates \( \hat{X} \), gained by evaluating Eqs. (3)–(5), will be subjected to PCA.

The coefficients \( a_X, b_X, c_X \) and \( d_X \) from Eqs. (3)–(5) that represent least-squares fits to the data are specified in Table 3. As all variables were standardised, the coefficients \( a_X \) and \( b_X \) appearing in regression equations with only one explanatory variable are equal to the correlations in Table 2 between discharge \( a_X \) or temperature \( b_X \) and the respective response variable, \( X \), of interest. For some response variables (for instance \( \text{NH}_4 \)), an estimate based on temperature is clearly superior to an estimate based on discharge; for others (for instance conductivity), discharge is the more effective predictor. When both predictors are combined, coefficient \( c_X \), for instance, may become zero, indicating that discharge provides no additional information beyond that already contained in temperature. Note that discharge and temperature are not completely independent of each other, variations of temperature can partly...
be estimated from variations of discharge and vice versa. As an example, consider $X = O_2s$, $a_X$ is not zero but $c_X$ vanishes; this can be understood when the prediction $DC \rightarrow O_2s$ is split into a two-step prediction $DC \rightarrow T \rightarrow O_2s$. In the latter, all of the information in $DC$ about $O_2s$ is channelled through temperature so that by knowing the water temperature, discharge becomes non-informative.

Fig. 6 depicts linear estimates based on discharge and temperature — obtained by evaluating Eq. (5) — together with the corresponding standardised observations of $NO_3$, $NH_4$, pH, and $O_2s$. The meteorologically induced time series (bold lines) are smooth and exhibit some similar behaviour to the seasonal components shown in Fig. 9. Regarding $NO_3$, the predictions reproduce a significant proportion of the observed variance. The underlying effects are, increasing nitrate concentrations with increasing discharge, caused by drainage from fields after precipitation events and fixation of nitrate in biomass in the warmer periods. Regarding ammonia, dependence on temperature, governing the biological activity, is dominant. Note, however, that the unusually high ammonia levels in winter 1996 and winter 1997 do not correlate with the external forcing. We have determined however, that the results of the PCA

<table>
<thead>
<tr>
<th>Explanatory variable</th>
<th>Coefficient</th>
<th>PH</th>
<th>NH₄</th>
<th>PO₄</th>
<th>O₂s</th>
<th>NO₃</th>
<th>Cond</th>
<th>DC</th>
<th>WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>$A_X$</td>
<td>0.19</td>
<td>0.01</td>
<td>-0.39</td>
<td>-0.14</td>
<td>0.38</td>
<td>-0.33</td>
<td>-</td>
<td>-0.24</td>
</tr>
<tr>
<td>WT</td>
<td>$B_X$</td>
<td>0.53</td>
<td>-0.70</td>
<td>-0.14</td>
<td>0.52</td>
<td>-0.56</td>
<td>0.07</td>
<td>-0.24</td>
<td>-</td>
</tr>
<tr>
<td>DC and WT</td>
<td>$C_X$</td>
<td>-0.07</td>
<td>-0.17</td>
<td>-0.45</td>
<td>-0.01</td>
<td>0.26</td>
<td>-0.33</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DC and WT</td>
<td>$D_X$</td>
<td>0.51</td>
<td>-0.75</td>
<td>-0.25</td>
<td>0.52</td>
<td>-0.49</td>
<td>-0.01</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3
Original data, riverine part: Regression coefficients when modelling water quality parameters as function of discharge and/or water temperature [Eq. (3) Eqs. (4) and (5)]

Fig. 6. Observations of nitrate, ammonia, pH, and oxygen at station Schnackenburg (thin lines) and the corresponding predictions obtained by regressing on water temperature and discharge (bold lines).
presented below are not very sensitive to these unusual periods with high residuals. For oxygen and pH, short-term deviations from the predictions are much more pronounced than for NO$_3$ and NH$_4$. The conclusion to be drawn is that biological activity (e.g. algae growth and/or microbial decomposition), which involve both oxygen and pH, is only weakly governed by meteorological forcing, at least by discharge and temperature alone.

To look more closely at the relation between temperature and biological activity let us employ a much simplified model of biological activity, being defined exclusively in terms of pH, O$_2$s and PO$_4$ changes. A PCA of the observations of these three variables yields a leading principal component that explains 61% of the total variance of the three variables. The corresponding EOF pattern (right panel in Fig. 7) portrays the characteristic effect of algae growth, namely increased pH linked with decreased phosphate and increased oxygen. The actual amplitude of this pattern, i.e. the actual value of the principal component, may be considered as a generalised observation...
combining three physical variables in a characteristic way. The fact that the observed amplitude is negative is an indication of relatively — with respect to the mean of all data points — low biomass concentrations, possibly resulting from intense mineralisation processes.

The left panel of Fig. 7 depicts a scatter plot of temperature versus the leading principal component (EOF-1). The fitted regression line reveals a clear increase of the mean principal component value (‘biomass’) as a function of temperature. Each data point is marked with respect to the season in which it was recorded. In winter, only a small amount of biomass exists, giving rise to negative principal-component values. In summer, the mean amount of biomass is positive; the actual values are however, widely scattered around the mean. It seems reasonable to take the size of this scatter as an indication of the strength of the competing biological processes affecting biomass. Although sufficiently high temperatures are a prerequisite for biological processes to be active, the actual values of biomass resulting from counteracting production and degradation processes depend on several other factors, such as light conditions (including self shading of algae), nutrient deficiencies, silicate limitations (for diatoms), populations of micro-organisms and zooplankton, infections by fungi, bacteria or viruses, and wash out by extreme discharge events. Consequently, it is not surprising that a significant proportion of the total variability of biomass remains unexplained by regressing on temperature alone.

The relative strengths of algal growth and algal degradation vary among the seasons. According to Fig. 7, there seems to be a tendency for the principal components observed in spring to be systematically higher than the principal components observed in autumn, even when spring temperatures are lower. In springtime, early strong algal blooms occur as radiation intensity gradually increases and the availability of nutrients is still high. The start of the counteracting process, biomass decomposition, is delayed with respect to algal growth, because micro-organisms and zooplankton need higher temperatures and a sufficient amount of substrate to develop. In contrast, at the end of summer and in the autumn, when water temperatures are still moderate but radiation intensity has already decreased, biomass mineralisation often prevails over algae growth, so that the principal component related to biomass is relatively low. This is another indication that the balance between biomass production and decomposition is not a function of temperature alone.

We will now analyse the correlation structure of the deviations \( X - \bar{X} \) of the normalised observations \( X \) from the predictions \( \hat{X} \) based on meteorological forcing. Fig. 8 shows the leading EOF pattern of these residual components when estimating the externally forced components \( \hat{X} \) as functions of discharge (Fig. 8a) or as functions of discharge and temperature (Fig. 8b). Fig. 8c differs from Fig. 8b by the fact that nitrate and conductivity are excluded from the analysis.

By subtracting the responses to variations of discharge from the observed data (using Eq. (1)), the leading EOF (Fig. 8a) does not significantly change with respect to the first EOF (biological mode) of the original data (Fig. 4). This indicates that the biological mode is more or less independent of discharge. The eigenvalue of 3.1 remains the same as it was for the original data, while the percentage of explained variance increases from 39 to 45\%, due to the reduction in total variance caused by holding discharge fixed. The eigenvalue of the second EOF (specified in the table of Fig. 8a) decreases from 1.6 to 1.3 with respect to the original data. This change is surprisingly moderate considering the fact that this EOF was identified as a discharge mode in the original data set.

Even when eliminating the responses to variations of both discharge and temperature (using Eq. (5)), the biological mode still remains the leading mode that explains most of the residual variance (Fig. 8b). The pattern remains stable, with the exception of the contribution of ammonia. Ammonia is assimilated when algae grow but is also consumed by micro-organisms that convert ammonia to nitrate by nitrification. Thus, significant ammonia concentrations arise only in winter, when biological activity is negligible (cf. the time series in Fig. 2). This strong correlation with temperature makes the ammonia component in the residual EOF pattern disappear. The eigenvalue corresponding to the first EOF significantly
decreases from 3.1 to 1.9, which indicates that in a data sample constrained by constant temperature the variability of biological activity is reduced. This pattern and the eigen-values remain unchanged even when nitrate and conductivity are excluded from calculation (Fig. 8c) showing that these two parameters do not influence the biological dynamic.

4.2. Riverine part: deviations from the seasonal normal

As stated in the previous subsection another way of subtracting the influence of systematic external forcing with a natural annual cycle is to consider deviations from a mean annual cycle. It is clear that, if two observed time series have similar annual cycles, the corresponding variables are not necessarily connected by physical, chemical or biological processes. Correlations between the original data can be spurious in the sense that the two variables of interest become independent of each other once the state of the mechanism binding the two variables together is known. For instance, the negative correlation between temperature and ammonia concentrations might vanish when the annual cycle in farming activities is accounted for. Therefore, in this section the PCA of the previous section will be repeated, this time considering deviations of the observations from an estimated mean annual cycle. This seasonal cycle will be defined on a monthly basis for each variable and for each station separately by averaging all of the measurements from each calendar month in the 5-year period. The observed deviations from this mean annual cycle will be denoted (seasonal) anomalies.

Fig. 9 depicts the estimated seasonal component (plotted as an average over all stations) together with the anomalies calculated at all riverine stations. Most of the seasonal cycles could be eliminated, although periodic patterns still remain.
Table 4  
Correlation matrix obtained from pooling the observed deviations from estimated local mean seasonal cycles at all riverine stations

<table>
<thead>
<tr>
<th></th>
<th>WT</th>
<th>DC</th>
<th>pH</th>
<th>NH₄</th>
<th>PO₄</th>
<th>O₂s</th>
<th>NO₃</th>
<th>Cond</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>1.00</td>
<td>0.10</td>
<td>0.29</td>
<td>-0.41</td>
<td>-0.23</td>
<td>0.23</td>
<td>-0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>DC</td>
<td>1.00</td>
<td>-0.30</td>
<td>-0.48</td>
<td>-0.20</td>
<td>-0.21</td>
<td>0.37</td>
<td>-0.66</td>
<td></td>
</tr>
<tr>
<td>PH</td>
<td>1.00</td>
<td>-0.14</td>
<td>-0.21</td>
<td>0.46</td>
<td>-0.17</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NH₄</td>
<td>1.00</td>
<td></td>
<td>0.35</td>
<td>-0.13</td>
<td>-0.15</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO₄</td>
<td></td>
<td>1.00</td>
<td>-0.29</td>
<td>0.05</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₂s</td>
<td></td>
<td></td>
<td>1.00</td>
<td>-0.22</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO₃</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>-0.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

Fig. 10. Leading EOF patterns (EOF-1 and EOF-2), the corresponding eigen-values (Eigen) and the proportion of total variance explained by the respective principal component (Var), calculated for the seasonal anomalies in the riverine part of the Elbe River (cf. Table 3).

for discharge and ammonia. The anomalies are not very different for different stations, indicating that the relevant processes are not local. However, differences among the stations seem to be a bit more pronounced for the parameters oxygen, pH and phosphate, which are modified by biological activity.

4.2.1. PCA

Table 4 provides the correlation matrix obtained when the deviations from the estimated seasonal cycle at all riverine stations are pooled. The EOF patterns of the two leading principal components, which explain 30.4% and 26.5% of the total variance, respectively, are depicted in Fig. 10. The two corresponding eigen-values, 2.43 and 2.12, respectively, are not well separated; therefore, the structures of the EOF patterns are relatively uncertain. However, the first EOF, with reliable positive contributions from discharge and nitrate in contrast to a significant negative contribution from conductivity, agrees roughly with the pattern of the discharge mode determined to be the second EOF of the original data (Fig. 4). A major difference in comparison with the discharge pattern obtained from the original data is that, with regard to anomalies, the negative contribution of phosphate is no longer significant.

The second EOF again depicts the typical pattern of a biological mode with significant parallel contributions of temperature, pH and oxygen having an opposite sign to ammonia and phosphate. This type of covariation can be attributed to the production and decomposition of biomass. Note that the second pattern in Fig. 10 agrees well with the leading EOF pattern of the original
data after subtracting the linear responses to variations of the discharge (Fig. 8a).

4.2.2. *Separation of the impact of external forcing*

A partial PCA similar to the one in Section 4.1.2 was performed in order to identify the remaining impact of external forcing on observations after the estimated seasonal cycle was subtracted. Again, the responses of observed variables to external forcing are assumed to be linear. Table 5 provides the coefficients obtained by fitting the corresponding regression models in Eqs. (3)–(5) to the observed anomalies.

**Fig. 11a** depicts the EOF pattern that corresponds to the leading principal component, which explains 33.2% of total variance in the data set after linear responses to variations of discharge—estimated from Eq. (3) — were subtracted. The pattern is not very different from the second EOF of the total anomalies (Fig. 10) and can clearly be attributed to biological activity. The amount of uncertainty in the pattern is considerably reduced compared to Fig. 10. Not surprisingly, regressing on discharge eliminated most of the discharge mode; thus the biological mode is now the only significant signal remaining in the data set. All other eigen-values are too close to each other to

Table 5
Seasonal anomaly data, riverine part: Regression coefficients when modelling water quality parameters as function of discharge and/or water temperature [Eq. (3) Eqs. (4) and (5)]

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>Response variable (X)</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pH</td>
<td>NH₄</td>
</tr>
<tr>
<td>DC</td>
<td>Aₓ</td>
<td>−0.30</td>
</tr>
<tr>
<td>WT</td>
<td>Cₓ</td>
<td>0.29</td>
</tr>
<tr>
<td>DC and WT</td>
<td>Cₓ</td>
<td>−0.33</td>
</tr>
<tr>
<td>WT</td>
<td>Dₓ</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Fig. 11. Leading EOF patterns of seasonal anomalies (riverine stations) when (a) regressed on discharge (DC); (b) regressed on discharge (DC) and water temperature (WT) and (c) regressed on discharge (DC) and water temperature (WT) excluding the variables nitrate (NO₃) and conductivity (Cond) from analysis.
allow for a meaningful interpretation of the corresponding individual patterns. The biological EOF pattern determined from the anomalies is not strongly affected by separating the impact of discharge; this agrees well with the observation that the biological mode identified from the total anomalies (Fig. 10) is very similar to the corresponding pattern obtained after removing the impact of discharge variations from the original observations (Fig. 8a).

The biological pattern remains unchanged even when separating the linear responses to discharge and water temperature from the anomalies (Fig. 11b), although including temperature as an additional external-forcing variable diminishes the eigen-value of the biological mode from 2.32 to 1.79. The most uncertain contributions in the pattern are those by nitrate and conductivity. As the amplitudes of these contributions are near zero, it seems reasonable to repeat the partial PCA with fixed discharge and temperature on a data set without the anomalies of nitrate and conductivity. As the amplitudes of these contributions are near zero, it seems reasonable to repeat the partial PCA with fixed discharge and temperature on a data set without the anomalies of nitrate and conductivity. It turns out (Fig. 11c) that the leading EOF containing the four remaining variables reproduces the corresponding sub-pattern of Fig. 11b perfectly. As the loading of nitrate and conductivity were small, the eigen-value did not change.

Comparing Figs. 8 and 11, the differences due to the analysis of data with and without a seasonal component can be summarised as follows. After subtracting the linear responses to variations of discharge, both approaches yield well-defined first EOFs that have similar structures and may be attributed to biological processes. Adding water temperature as a second predictor introduces some major structural differences which, however, are restricted to the most uncertain components: ammonia, nitrate and conductivity. It is found that in both approaches excluding these three variables from the data set subjected to PCA does not alter the subpattern constituted by the three variables pH, PO₄ and O₂S (not shown for NH₄). The characteristic pattern of parallel variations of pH and oxygen and variations with an opposite sign of phosphate appears to be the stable kernel of the biological mode, regardless of whether seasonal components and/or linear responses to external forcing are included in the data set or not.

4.3. Estuarine part

In this section PCA is applied to data from the estuarine reach of the Elbe between Zollenspieker (599 km) and Brunsbüttel (693 km). Data from the station Cuxhaven (725 km) is not included as this station is located downstream of the mixing zone of marine and freshwater and is, therefore, already dominated by a strong marine signal.

4.3.1. PCA of original data set

Considering the original data, again two well-separated EOF patterns can be identified (Fig. 12), explaining 37 and 21%, respectively, of the total data set variance.
The first EOF has its strongest contribution from discharge, ammonia, phosphate and nitrate but also from temperature and oxygen. The pattern differs from the discharge mode in the riverine part of the Elbe (Fig. 4) because ammonia has an opposite sign to phosphate and because there is a strong contribution from temperature. This modification could be explained by spurious correlations between typical seasonal variations of discharge and water temperature. Typically, discharge is high during winter, when low temperature causes high ammonia concentration by suppressing biological processes.

The second EOF (lower panel Fig. 12b) is independent of discharge. Regarding its structure, which is dominated by pH, phosphate and oxygen, it can be considered to be the biological mode. However, in the estuary this mode is of secondary importance (explaining only 21.3% of total variance), while in the riverine part the biological mode is the leading mode, accounting for 39.2% of total variance in the data set. This indicates the minor importance of biological processes in the estuary. The very small contribution of water temperature to the pattern may be understood by considering that temperature, as a seasonal indicator, is already involved in the first mode. Note that the biological mode in the estuary (Fig. 12) resembles the corresponding pattern in the riverine part after the responses to meteorological forcing have been removed (Fig. 8).

4.3.2. Separation of the impact of external forcing

In this section, estimated linear responses to external forcing are eliminated before subjecting the estuarine data to PCA. External forcing is defined in terms of temperature and discharge. Results are summarised in Fig. 13.

As in the riverine part of the Elbe (Fig. 8), the biological mode (upper panel in Fig. 13) is found to be the leading mode, explaining most (31.0%) of the anomaly variance that cannot be attributed to external forcing. The structure of the pattern is not significantly different from the corresponding subpattern of the lower panel in Fig. 12.

The second EOF exhibits strong loadings of ammonia and, with an opposite sign, phosphate and conductivity. This mode has been called the ‘Tidal mode’, assuming that high conductivity reflects the presence of marine water (high salinity) during periods with the estuarine mixing zone being shifted upstream. Marine water is characterised by relatively low ammonia concentrations. A plausible explanation for the enhanced phosphate concentrations during periods of higher salinity in the mixing zone is the increased osmotic pressure causing cell walls to break with subsequent release of phosphate.

4.3.3. Removing the seasonal component

Instead of separating responses to external forcing, we analyse deviations from estimated mean seasonal cycles. The two leading EOFs obtained
in this way are depicted in Fig. 14. They may be compared with the EOFs in Fig. 10, which resulted for the riverine stations.

Again, typical patterns are obtained that can be considered to be a discharge mode (EOF 1) and a biological mode (EOF 2). As in the riverine part of the Elbe (Fig. 10), the discharge mode in the estuary explains the biggest proportion of variance of seasonal anomalies. The structure of the estuarine discharge mode shows that positive discharge anomalies usually correspond with positive concentration anomalies of nitrate, negative concentration anomalies of ammonia and phosphate and relatively low conductivities; this is very similar to the riverine discharge EOF in Fig. 10.

Comparing the estuarine biological mode with its riverine counterpart in Fig. 10, significant differences remain confined to the temperature and ammonia components. For the estuary, temperature disappeared from both leading EOFs after the mean seasonal cycles of observations were removed. One should, however, be aware of the high uncertainty in the analysed temperature component of the biological mode. The different behaviour of ammonia may be caused by the more dominating nitrification processes (oxidation of ammonia to nitrate) in the estuarine part. In this process oxygen will be consumed, thus oxygen and ammonia have the same sign.

Having identified the structure of a certain mode, the local strength of this mode may be analysed based on data from any station. An interesting question is whether the time evolution of these modes bear some relationship among all the stations or at least within a certain spatial separation range. This question can be addressed by analysing the correlation between the amplitudes of the modes at the different stations. The analysis will be done for seasonal anomalies rather than original observations to exclude correlations that are merely due to the overall station independent seasonal cycle.

According to Eq. (1), the time series of local amplitudes of any mode (principal components) can be calculated by projecting each locally observed anomaly data vector onto the respective EOF pattern of interest. To exemplify correlations between different stations, we take the amplitude at Schnackenburg as the basis time series, and calculate the simultaneous and the lagged correlation with the amplitudes at all other stations. It turns out that the lagged correlations fall off very rapidly to zero with increasing separation from Schnackenburg. Therefore only the simultaneous correlations are shown in Fig. 15.

It can be seen that the coherence range of the ‘discharge’ mode is quite long and reaches almost the whole river length. Only at the last stations near the North Sea the correlation with Schnackenburg is somewhat lower. This result means that, on the time scale of fortnightly measurements, the amplitude of the ‘discharge’ mode evolves almost in phase at all stations. This is consistent with the fact that discharge observations (average values over 14 days) are very homogeneous along the river.
On the other hand the spatial coherence of the ‘biological’ mode is much shorter, i.e. the time series at Schnackenburg corresponding to the ‘biological mode’ is only correlated with the neighbouring stations. This result indicates that the spatial range in which this mode evolves simultaneously is much more limited. This behaviour of the biological mode is therefore not consistent with an important role of water temperature, since the spatial coherence of temperature should be of the order of several hundred kilometres as the comparison of the temperature data of all stations shows. The results in Fig. 15, therefore, support the discussion presented in Section 4.1.2.

A possible interpretation for the short spatial range of coherence of the biological mode would be the importance of radiation and other local impacts (e.g. grazing by zooplankton), which can vary spatially much more randomly. Unfortunately, long time series of radiation data that could help corroborate this hypothesis are not available for all the stations in this study.

5. Conclusions

Water-quality data taken over a 5-year period (1993–1997) at several locations along the Elbe river were subjected to exploratory statistical analysis. Multivariate statistical analysis offers the tools by which to estimate the observed strengths of relevant biogeochemical processes that involve simultaneous changes of several water-quality parameters. Such an assessment is a prerequisite when observations are to be compared with process-oriented numerical simulations to increase knowledge about the nutrient system.

Principal component analysis (PCA) is a standard technique used to obtain a condensed description of the interaction structure between a number of variables — defined in terms of covariances or correlations — by considering only a few linear combinations (principal components) of the original variables. The coefficients of these linear combinations make up vectors — in this paper the notation of empirical orthogonal functions (EOFs) was used — such that the value of a particular principal component is obtained by projecting the multivariate observations onto the respective EOF. Therefore, principal components are the amplitudes of the corresponding EOFs.

Observations of eight water-quality variables were analysed for the riverine and estuarine parts of the Elbe separately, pooling all observations at stations within each of these two river stretches. For both partial data sets, it was found that two

![Fig. 15](image_url)
principal components (or EOFs) are sufficient to describe nearly 60% of the observed total variance in the data. One of these principal components — the first in the riverine and the second in the estuarine part of the Elbe, can be attributed to biomass being modified by biological processes. The corresponding EOF exhibits parallel contributions (loadings) from pH and oxygen, which have an opposite sign to the contribution from phosphate. In the riverine part, the second principal component is clearly governed by discharge. According to the corresponding EOF, a positive discharge anomaly generally implies negative concentration anomalies caused by dilution. An exception to this are nitrate concentrations, which increase with discharge due to drainage from fields after strong precipitation. The estuarine counterpart of the riverine discharge mode has a modified structure, with a strong contribution from temperature and with ammonia anomalies that have an opposite sign to phosphate anomalies. We suspect that these differences arise from a spurious correlation between the seasonal cycles of discharge and water temperature, the latter parameters governing ammonia concentrations. It has been found (not discussed in the paper) that the dominant EOF patterns would not significantly be modified if prior to statistical analysis observed concentrations were transformed to a logarithmic scale to be consistent with the definition of the pH value in terms of H-ion concentrations.

Being aware of possibly spurious correlation's between observations due to similar annual cycles, the data were re-analysed after subtracting an estimated mean annual cycle from each variable. For the riverine part of the Elbe, the two characteristic covariance patterns previously identified remained discernible, although the EOFs changed in several details. Together, the two modes explain 57% of the total variance. However, the amplitudes of the two modes are of comparable magnitude, which makes it difficult to identify the exact structure of the two individual EOFs. Considering confidence intervals that were estimated using a bootstrap technique, the differences between EOFs obtained from original data and EOFs obtained from seasonal anomalies are judged to be not very significant. In the estuarine part of the Elbe, the analysis of anomalies reveals a clear discharge mode that corresponds well with that of the riverine counterpart. In the estuarine biological mode, the characteristic covariance pattern between pH, oxygen and phosphate remains unaffected by the removal of the mean seasonal cycle. Ammonia, however, appears with the opposite sign. This sensitivity to the kind of analysis could be due to the fact that only in winter are ammonia concentrations high enough to be detectable.

The question of whether a seasonal component should be separated from the original observations is closely related to the question of how much variability in the data set can be attributed to external meteorological forcing. In this study, external forcing was assumed to be represented by the discharge and the water temperature. Estimated responses of water quality variables to changes in these two parameters were separated from the data by employing linear regression models; the resulting residuals were subjected to PCA.

An important result is that in both the riverine and estuarine parts of the Elbe, biological activity appears to be the dominant source of variability that is uncorrelated with discharge and temperature. The corresponding EOFs are the only significant patterns of covariation that can be identified in this particular data set. At first sight, biological activity being independent of temperature may seem unrealistic. In fact, the mean amplitude of the biological mode was shown to increase with temperature. However, particularly in summer the scatter of this amplitude around its mean is very large. This scatter can be attributed to variable relative strengths of counteracting processes such as biomass production by photosynthesis and biomass degradation by micro-organisms and/or uptake by zooplankton. All these processes depend on various factors (e.g. light conditions, nutrient deficiencies, algal infections etc.) which are to a large extent independent of temperature. This is also confirmed by the low correlations between the amplitudes of the biological mode at different stations (neglecting the mean seasonal cycle), although water temperatures vary coherently at all stations.
A stable kernel of the biological EOF that appeared consistently in all kinds of analysis is constituted by the three variables pH, oxygen and phosphate. The amplitude of this composite pattern being affected by biomass production and decomposition might provide a more reliable and more informative assessment of a river’s actual trophic state than the observation of just one of the three original variables. That phosphate is the only nutrient in this pattern confirms the results of other investigations (e.g. Bergemann et al., 1996) that in the Elbe River primary production is mainly controlled by the availability of phosphate. Concentrations of nitrogen are always too high to be a limiting factor. After eliminating the estimated responses of water-quality parameters to external forcing, the internal variability of the biological mode might be monitored as an indicator of the status of eutrophication. Long-term trends might be useful to identify anthropogenic impacts on water quality. Whether internal variability of the biological mode could even be used to characterise different rivers or river sections is an open question that needs further investigation.

Acknowledgements

The authors thank M. Bergemann of ‘Arbeitsgemeinschaft zur Reinhaltung der Elbe’ (ARGE Elbe) for providing the data of the Elbe River and H. von Storch for helpful discussions and valuable comments.

References

ARGE Elbe, 1995a. 20 Jahre Arbeitsgemeinschaft für die Reinhaltung der Elbe (Eigenverlag), Hamburg, Germany
ARGE Elbe, 1995b. Nährstoffstudie der Elbe Teil 2 von Schmilka bis zur See. Arbeitsgemeinschaft für die Reinhaltung der Elbe (Eigenverlag), Hamburg, Germany