WP5.2:
Combination of MAR and adjusted conventional treatment processes for an Integrated Water Resources Management

Deliverable 5.2.13
Application of a data-driven approach for well field modelling

Summary

Work package WP 5.2 “Combination of Managed Aquifer Recharge (MAR) and adjusted conventional treatment processes for an Integrated Water Resources Management“ within the European Project TECHNEAU (“Technology enabled universal access to safe water”) investigates bank filtration (BF) + post-treatment as a MAR technique to provide sustainable and safe drinking water supply.
One of the tasks within the project is the testing of a data-driven approach for the identification (pattern recognition) and quantification of the key processes that drive the groundwater (GW) dynamics in observation wells (OW) near well fields of a BF waterworks. For this BUSSE (2010) used a multivariate statistical method (principal component analysis - PCA) with daily GW level time series of 41 OWs and was able to identify four processes that explained 95% of the total variance in the data set. On the one hand GW recharge (58.9%) and its temporal delay (3.3%) explain 62% of the GW level fluctuations within the study period. On the other hand any discernible impact of waterworks abstractions is limited to one of the three well fields with the highest production rate (29.8% of explained variance). In addition the infiltration of a marshy ditch into the GW accounts for another 2.9% of the GW level fluctuations.
Regarding the ability to identify driving forces for GW level fluctuations the main advantage for using PCA compared to process-driven GW flow modelling is that the driving forces for GW level fluctuations can be identified and quantified without requiring exact knowledge about the structural properties of the subsurface (e.g. aquifer transmissivities) and its input parameters (e.g. GW recharge, production rates). Note that the latter do not enter the PCA directly but are used for spatiotemporal interpretation of the results, which also requires some expertise. In addition, it is recommended to perform a sensitivity analysis of the PCA results in a next step, so that it can be tested whether the processes identified above are robust in case of changing input parameters such as:

- Reduced spatiotemporal resolution
- Study period with different boundary conditions (e.g. pumping regime)

The contents of this report were presented to the involved experts from the Berliner Wasserbetriebe (BWB). In agreement with their recommendations it was decided to focus further research within follow-up projects on the (i) sensitivity analysis of the PCA results and (ii) to apply nonlinear approaches for identification and quantification of processes that drive GW quality dynamics within the study area.

Contact
Michael Rustler, Dipl.-Geoök., KompetenzZentrum Wasser Berlin gGmbH
Email: michael.rustler@kompetenz-wasser.de
Phone: +49 (0) 30 536 53 825

TKI Categorisation

Classification									
Supply Chain		**Process Chain**		**Process Chain (cont'd)**		**Water Quality**		**Water Quantity (cont'd)**	
Source		**Raw water storage**		**Sludge treatment**		**Legislation/regulation**		- Leakage	
- Catchment	X	- Supply reservoir		- Settlement		- Raw water (source)		- Recycle	X
- Groundwater	X	- Bankside storage	X	- Thickening		- Treated water			
- Surface water	X	**Pretreatment**		- Dewatering		**Chemical**			
- Spring water		- Screening		- Disposal		- Organic compounds			
- Storm water		- Microstraining		**Chemical dosing**		- Inorganic compounds			
- Brackish/seawater		**Primary treatment**		- pH adjustment		- Disinfection by-products			
- Wastewater		- Sedimentation		- Coagulant		- Corrosion			
Raw water storage		- Rapid filtration		- Polyelectrolyte		- Scaling			
- Supply reservoir		- Slow sand filtration		- Disinfectant		- Chlorine decay			
- Bankside storage	X	- Bank filtration	X	- Lead/plumbosolvency		**Microbiological**			
Water treatment		- Dune infiltration		**Control/instrumentation**		- Viruses		**Consumers / Risk**	
- Pretreatment	X	**Secondary treatment**		- Flow		- Parasites			
- Primary treatment	X	- Coagulation/flocculation		- Pressure		- Bacteria		**Trust**	
- Secondary treatment		- Sedimentation		- pH		- Fungi		- In water safety/quality	X
- Sludge treatment		- Filtration		- Chlorine		**Aesthetic**		- In security of supply	X
Treated water storage		- Dissolved air flotation(DAF)		- Dosing		- Hardness / alkalinity		- In suppliers	X
- Service reservoir		- Ion exchange		- Telemetry		- pH		- In regulations and regulators	
Distribution		- Membrane treatment		**Analysis**		- Turbidity		**Willingness-to-pay/acceptance**	
- Pumps		- Adsorption		- Chemical		- Colour		- For safety	X
- Supply pipe / main		- Disinfection		- Microbiological		- Taste		- For improved taste/odour	X
Tap (Customer)		- Dechlorination		- Physical	X	- Odour		- For infrastructure	X
- Supply (service) pipe		**Treated water storage**						- For security of supply	X

Internal plumbing		- Service reservoir				Water Quantity		**Risk Communication**	
- Internal storage		**Distribution**						- Communication strategies	
		- Disinfection				**Source**		- Potential pitfalls	
		- Lead/plumbosolvency				- Source management	X	- Proven techniques	X
		- Manganese control				- Alternative source(s)	X		
		- Biofilm control				**Management**			
		Tap (Customer)				- Water balance	X		
		- Point-of-entry (POE)				- Demand/supply trend(s)			
		- Point-of-use (POU)				- Demand reduction			

TKI Categorisation (continued)

Contains		Constraints		Meta data					
Report	X	Low cost	x	*Michael Rustler, Gesche Grützmacher*					
Database	X	Simple technology	x	*KompetenzZentrum Wasser Berlin*					
Spreadsheet		No/low skill requirement	x	*Michael Rustler*					
Model	X	No/low energy requirement	x	*michael.rustler@kompetenz-wasser.de*					
Research	X	No/low chemical requirement	x						
Literature review		No/low sludge production	x						
Trend analysis		Rural location	x						
Case study / demonstration	X	Developing world location	x						
Financial / organisational									
Methodology	X								
Legislation / regulation									

Colophon

Title
Application of a data-driven approach for well field modelling

Author(s)
Michael Rustler (KWB), Julia Busse (KWB)

Quality Assurance
Gesche Grützmacher (KWB),
Gunnar Lischeid (ZALF)

Deliverable number
D 5.2.13

This report is: PU for public

Contents

List of Abbreviations

BWB	Berlin Water Company
GW	groundwater
MSA	measure of sampling adequacy
OW	observation well
PCA	principal component analysis
PC	principal component
SenGUV	Berlin Senate Department for Health, the Environment and Consumer Protection
SW	surface water
WW	waterworks
WSA Berlin	Water and Shipping Authority Berlin

1 Introduction

Background

Within the European project TECHNEAU (www.techneau.org) the Berlin Centre of Competence for Water (KWB) is investigating bank filtration (BF) and adjusted post-treatment as a managed aquifer recharge (MAR) technique to provide sustainable and safe drinking water supply. One of the tasks is the testing of data-driven models for identification (pattern recognition) of the key processes that drive the groundwater dynamics in observation wells near well fields of a BF waterworks. This report is a synthesis of the diploma thesis of BUSSE (2010) and summarises its main outcomes.

Objective

Operation of extended fields of production wells in urban areas encounters many challenges. The operator has to consider, e.g., restrictions with respect to maximum groundwater drawdown in adjacent areas, interactions with surface water bodies, or contamination risks from a variety of nearby sources. Often detailed groundwater models are used for optimising water resources management in these areas. However, application of these models is hampered by the usually encountered enormous subsurface heterogeneity, especially in Pleistocene sediments that prevail in Northern Germany.

In this study an alternative approach is followed. Groundwater level dynamics in various observation wells located in vicinity of the production wells is affected by the pumping regime in different well groups. These effects vary in space and time. For example, groundwater heads in wells close to single production wells will react more extensively and more rapid to changing production compared to more distant wells. A systematic analysis of these dynamics could provide a base for setting up a statistically sound empirical model that could then be used for optimising well field operation.

Correspondingly, the objective of the diploma thesis of BUSSE (2010) was to analyse available observation well hydrographs near a Berlin waterworks by using a multivariate statistical method (principal component analysis) to identify the key drivers of groundwater level dynamics through temporal and spatial pattern recognition. This method has already been successfully used in different case studies to identify the prevailing processes that impact the temporal and spatial groundwater dynamics (e.g. LONGUEVERGNE et al. 2007, LEWANDOWSKI et al. 2009, LISCHEID et al. 2010). Consequently it should be possible to quantify the relative impact of each driver (e.g. groundwater recharge, surface water infiltration, groundwater abstractions) on the groundwater level fluctuations in observation wells close to the waterworks.

In a next step (not covered within this work) this knowledge about the important processes can be used to develop an empirical data-based model (e.g. regression), which is able to derive the influence of different operational management schemes (i.e. well operations) on the groundwater level dynamics in observation wells close to the well fields of the waterworks. Consequently it can be used to optimise well field operation to different objectives (e.g. minimum drawdown).

2 Study area

The study area covers the area of the waterworks Wuhlheide and its proximate vicinity, which is located north of the River Spree in the eastern part of Berlin (Germany). The total subsurface catchment area of the waterworks is 114 km² (ZIPPEL & HANNAPEL 2008). However, within this work the study area is limited to a smaller part (15 km², see Figure 1) since a dense GW monitoring network with a high temporal resolution (mainly: daily) of GW level measurements is only available near the well fields. This is due to the fact that multiple brown field sites surround the waterworks for which the Berlin Senate has established a large monitoring network to survey the ongoing remediation measures (BLACH-RADAU 2008).

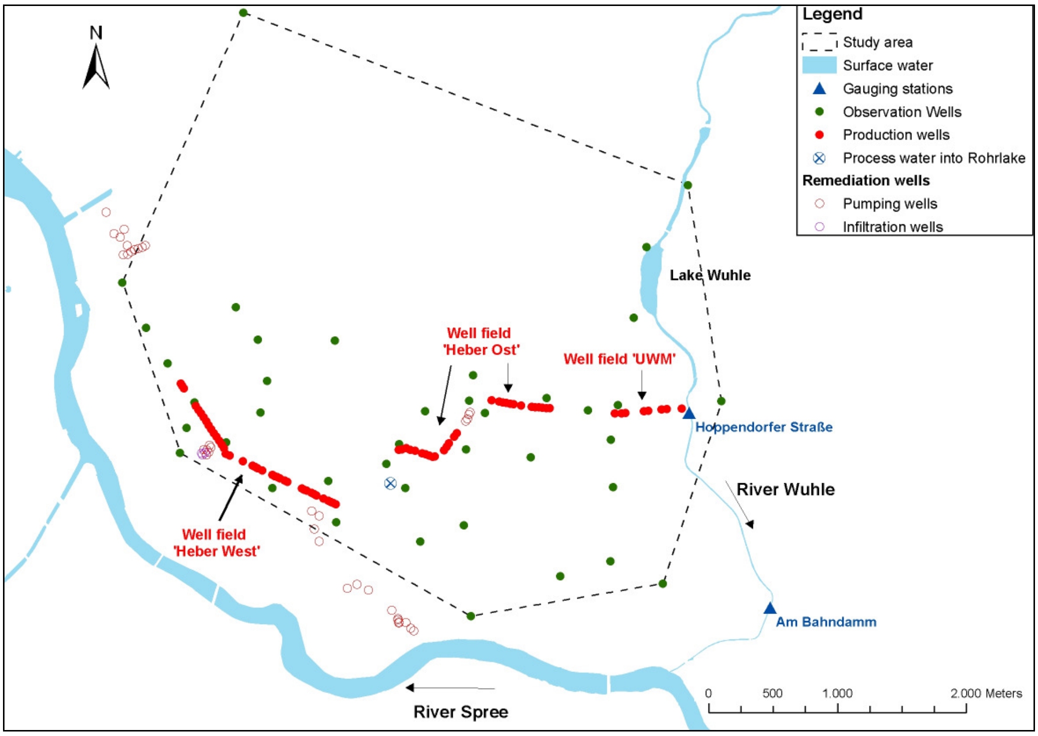

Figure 1 Study area (data source: BWB, SenGUV, TAUW GmbH)

Annual mean long term temperature within the study area varies between 8.5-10 °C (1961-1990, SENSTADT 2001). Annual mean precipitation of the BWB precipitation station at the WW Wuhlheide accounts to 544 mm/a (1995-2008), with slightly higher precipitation rates during the hydrological summer half year (April-Oct: 322 mm) and lower during the winter half year (Nov-April: 222 mm). GW recharge rates vary between 100 – 400 mm/a within the study area (SENSTADT 2007) depending on hydrometeorological boundary conditions and spatial properties (e.g. surface sealing, land use, soil type or vegetation). In general the GW flow direction is from north to south (SENSTADT 2009). However, due to the depression cone of the production well fields of the waterworks the natural GW flow direction is partly inversed close to the River Spree. As a result 26.2 % of the total pumped GW is infiltrated surface water (bank filtrate) originating from both River Spree and

Wuhle (ZIPPEL & HANNAPEL 2008). In total three well fields of the waterworks Wuhlheide abstract GW with an average production rate of 1014 m^3/h (2005-2009) from the alluvial deposits. Note that all production wells are screened in the 2nd aquifer (mean depth of the top of the filter screen below ground surface: 27m), while most of the observation wells are screened in the 1st aquifer (see Appendix A, Figure 12). This classification is based on hydrogeological maps that extrapolate small scale geologic information (i.e. borehole profiles). However, due to the heterogenic log-normal hydraulic conductivity distribution in the subsurface, this classification does not need to be true on a larger scale. In fact, a hydrograph analysis for two observation wells screened in 1st and 2nd aquifer, that are located quite distant from SW bodies and the waterworks well fields, yielded a comparable amplitude (see Appendix A, Figure 11 and Figure 12), which indicates a similar storage coefficient in both aquifers. Consequently, 1st and 2nd aquifer can be regarded as one hydraulically well connected unconfined aquifer, since storage coefficients would be several order of magnitudes lower in case of a confined aquifer (i.e. 10^{-3}- 10^{-5} compared to ~0.2).
Two of the three well fields ('Heber West', 'Heber Ost') use siphon wells with an estimated production rate of 15-25 m^3/h per well (DIESNER 2009). Multiple siphon wells are operated in tandem (so called: well groups), meaning that they cannot be turned on or off separately. In total there are 11 operable well groups (6 for the well field Heber Ost' and 5 for 'Heber West', respectively). The third well field is located in the eastern part of the study area near the River Wuhle and consists of 8 operable production wells that are equipped with submersible pumps allowing maximum pumping rates of 50 m^3/h per well (DIESNER 2009). In contrast to the siphon wells the later can be turned on or off separately, thus enabling a larger number of possible well operation schemes. Average GW abstraction shares for the three well fields account 62.5% ('Heber West'), 27% ('Heber Ost') and 10.4% ('UWM') of the total waterworks abstractions for the period 2005 to 2009. A small share of 4% of the abstracted GW (=0.011 m^3/s) is used as process water for filter washing, conveyed to a sewerage settling basin and subsequently the clear water is discharged into the marshy ditch 'Rohrlake' for re-infiltration.
In addition the GW hydraulics are also influenced by ongoing GW remediation measures (mainly pump-and-treat wells), which had an average production rate of 346 m^3/h in the year 2008 (BLACH-RADAU 2008). This is approximately 35 % compared to the average total production rate (2005-2009) used for drinking water supply.

3 Material and methods

The methodology for the application of the statistical tool (principal component analysis - PCA) to identify the key drivers and their spatial impact on groundwater level dynamics in the study area is illustrated in Figure 2. Each step of the flowchart is described in detail in the subsequent chapters. The open source software R (R DEVELOPMENT CORE TEAM 2009) is used for data selection, preparation and statistical analysis.

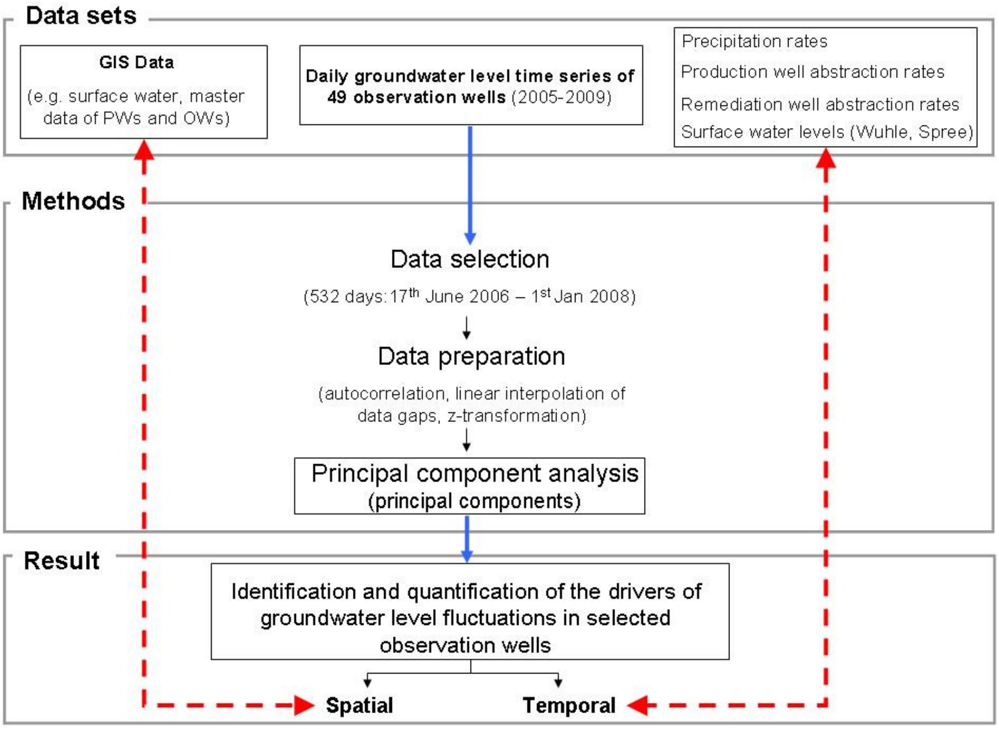

Figure 2 *Flow chart to identify and quantify the key processes that drive the groundwater level dynamics in selected observation wells*

Note that out of the many available data sets (see Chapter 3.1) the only input parameters which are used for the PCA are daily groundwater level time series from observation wells, since no information on groundwater levels in the production wells was available. All further information, either spatial (e.g. locations of surface water bodies) or temporal (e.g. production rates of well fields) are only used for the interpretation of the PCA results and thus do not impact the PCA results directly (see Chapter 4).

3.1 Data sets

All data are stored in an MS ACCESS 2003 database. The available data sets (master data, time-dependant data) are provided by different institutions (BWB, SenGUV, WSA Berlin, TAUW GmbH). Master data of observation, production and remediation wells contain information about their geographic location and screening depths. While waterworks abstractions are restricted to the 2nd aquifer most of the observation wells are screened in the 1st aquifer (see Appendix A, Figure 12). Automatic GW level measurements of observation wells are available at high temporal resolution (mainly daily) for a period from 2005 to 2009 (see Appendix A, Figure 13). However, GW level measurements are not available for production wells and remediation wells. For these only abstraction and infiltration rates on different spatial and temporal scales are available. On the one hand abstraction and infiltration rates of remediation wells are measured for each well separately, but only on a low temporal resolution (bi-annual). One the other hand high temporal resolution (mainly: daily) is available in case of the production wells of WW Wuhlheide, but abstraction rates are not measured for each well separately. These are measured for the siphon well fields ('Heber West' and 'Heber Ost') and for the submersible pump wells field 'UWM' in total and subsequently down-scaled to each production well (well field 'UWM') or well group (siphon well fields 'Heber West'/'Heber Ost') by using a mathematical approach. Based on monthly well operation times and assumed production well/well group rates (estimated on quarterly well performance tests) the total production rate of both well fields is calculated for each well or well group (Kramer, BWB, personal communication, 19-Jun 2010). In addition the following temporal data is available at least for the period 2005-2009:

- Daily precipitation rates of the precipitation station Wuhlheide (provided by BWB).
- Daily surface water level measurements for the River Wuhle (gauging station: 'Am Bahndamm') and River Spree (gauging station: 'Mühlendammschleuse') provided by SenGUV and WSA Berlin, respectively
- Weekly surface water level measurements for the River Wuhle (gauging station: 'Hoppendorfer Straße') provided by SenGUV

3.2 Data selection and preparation

Preprocessing and analysis of the data was performed using the R software package. In a first step the longest time period for which daily groundwater level time series of the 49 observation wells with least data gaps are available is determined (see Appendix A, Figure 13). Consequently the time period for the principal component analysis is limited to 532 days (19th July 2006 – 1st January 2008). Only 32 of the 49 observation wells did not exhibit any data gaps within this period (Figure 2). Three observation wells with more than 15 % data daps are excluded from further analysis. For the remaining 14 observation wells data gap of length between 2 to 62 days (see Appendix A, Table 2) needed to be filled by linear interpolation.

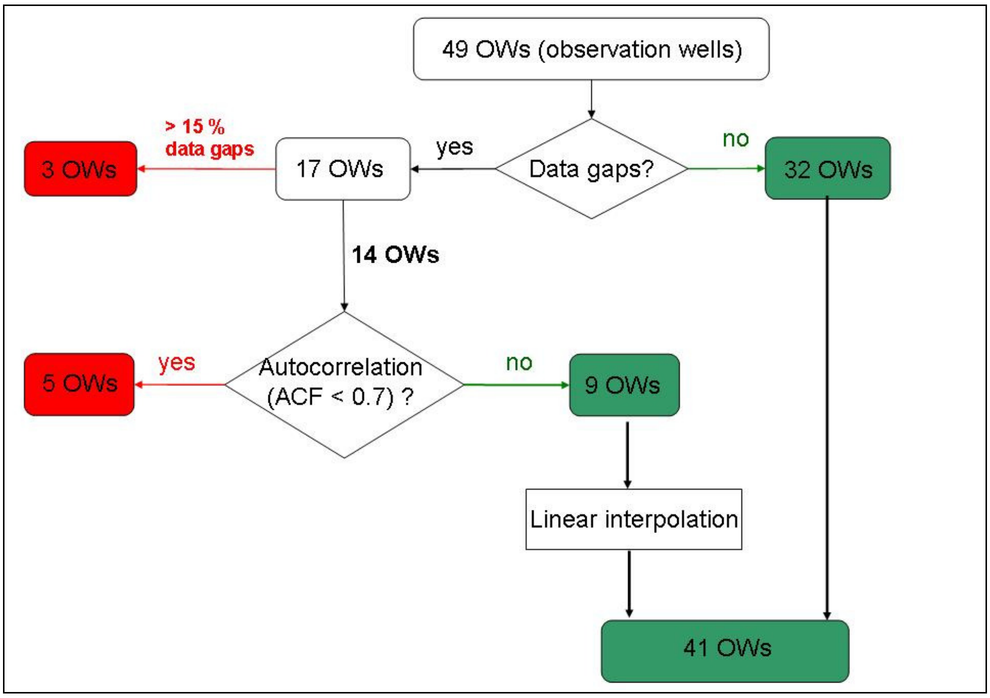

Figure 3 *Flow chart for data preparation of different observation wells (OW)*

In order to ensure that linear interpolation did not result in too large artefacts, the autocorrelation function of the time series was determined. To that end, the function 'acf ' was applied to the longest available time period (2005-2009) without data gaps to identify whether a linear interpolation of the data gaps is acceptable. Autocorrelation is the correlation of a signal xi (here: groundwater levels) with itself for a time lag k.For long time series with N groundwater levels the autocorrelation coefficient (r_k) is calculated according to Eq. 1 equation (MANIAK 2005):

$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x}) \cdot (x_{i+k} - \bar{x})}{(x_i - \bar{x})^2 \cdot (N-k)} \quad \forall k \quad in \quad 0 < k < N/4 \quad \text{(Eq. 1)}$$

with: $-1 < r_k < 1$, $\bar{x} = mean$

Interpolation is not performed for observation wells with an autocorrelation coefficient below 0.7 (=49% of explained variance) for the longest data gap period. Six observation wells, which show lower autocorrelation coefficients than 0.7, are excluded (see Appendix A, Figure 13). Consequently linear interpolation is performed in the software R by using the function ‘approx’ for the remaining nine observation wells which have an autocorrelation function above 0.7 for the longest continuous data gap period (see Appendix A, Table 2). In total data gaps filled by linear interpolation account for less than 0.4 % of all 41 observation wells.
In a last step z-transformation is used to normalise the groundwater level time series of each observation well to zero mean and unit variance, which assures that these will have a similar weighting in the subsequent principal component analysis.

3.3 Principal component analysis

The z-normalised daily groundwater level time series of 41 observation wells for a period of 532 days (19th July 2006 – 1st January 2008) are used as input parameters for the principal component analysis (PCA).
The goal of PCA is to explain as much information contained in the data as possible in as few components as possible. In statistics information content is expressed by the variability of a data set. The PCA searches for the direction in the multivariate space that contains the maximum amount of variability, which is the first principal component (PC1). The second principal component (PC2) contains the maximum amount of the remaining data variability under the constraint that is has to be orthogonal to the first PC. All subsequent PCs are constructed by the same principle (orthogonal to the preceding PC and accounting for the maximum amount of the remaining variability) and thus are uncorrelated to each other. Using z-normalised data, the PCA performs a decomposition of the correlation matrix into its 'eigenvectors' and 'eigenvalues'. For detailed information on the mathematical background of PCA the reader is referred to JOLLIFFE (2002).
The eigenvectors are the coefficients of the PCs spanning the new coordinate system. Note that since the input variables are time series of GW levels the PCs can be regarded as time series as well. The higher the loading (=correlation) of a single variable (here: GW level time series of one observation well) on a specific PC the more similar is its temporal behaviour compared to the PC dynamics. The amount of variability contained in each PC is expressed by the 'eigenvalues' which are equal to the sum of squared loadings (=variances) of all input variables (here: groundwater level time series of 41 observation wells) on the PC. On the contrary the communality of one variable is the fraction of variance explained by the considered PCs.

For using linear PCA the following preconditions should be satisfied:

- The Kaiser-Meyer-Olkin measure of sampling adequacy (MSA) is the best available method for assessing correlation matrices (BACKHAUS et al. 2008). Based on the anti-image-correlation matrix (GUTTMANN 1953) it calculates the amount of variability of a variable (here: observation well time series) that cannot be explained by regression with other variables. The MSA values can vary between zero and one. For performing PCA preferable MSA values should be above 0.8 (CURETON & D'AGOSTINO 1993).
- Data should exhibit a Gaussian distribution. However, that requirement is the most crucial for small data sets with a few tens of observations. In this study, the data set comprised 532 observations. For a data set of that size Gaussian distribution is close to irrelevant.

Note that linear PCA is the best available multivariate method for pattern recognition in case that all variables within the data set follow a linear dependency. However, if strong nonlinear relationships between the variables (here: bivariate correlations of OW time series) exist, nonlinear methods (e.g. nonlinear PCA) may outperform the linear PCA. To determine whether nonlinear relationships in the data set prevail scatter plots of GW level time series for all possible bivariate OW combinations are analysed. For a small amount of scatter plots quite close and approximately linear dependencies are identified. However, most of the scatter plots show loops, bifurcations and similar structures (data not shown). These structures arise due to the fact that the observation wells (OW) time series only show parallel behaviour within a limited time frame but are decoupled or even negatively correlated for other time periods. Note that whether this fact will limit the application of the PCA can only be identified 'a posteriori' in case that the time series of the most important PCs cannot be interpreted in a meaningful way. On the contrary the MSA, which is calculated in R with the function 'kmo.test' (KERNS 2007) for the z-normalised data matrix, yielded a value of 0.95 and thus is highly adequate for PCA. Consequently PCA is performed in the software R by using the function 'prcomp'.

4 Results and discussion

The interpretation of the PCA results is limited to the first four principal components with eigenvalues larger than one (Kaiser criterion), which already explain 95 % of the total variance in the data set (Table 1).

Table 1 Principal components (PC), eigenvalues, fraction of explained variance and cumulative fraction of explained variance (data from BUSSE 2010)

PC	Eigenvalue	Fraction of explained Variance [a] [%]	Cumulative fraction of Explained Variance [b] [%]
1	24.2	58.9	58.9
2	12.2	29.8	88.8
3	1.4	3.3	92.1
4	1.2	2.9	95.0
5	0.6	1.4	96.4
6	0.4	1.1	97.5
7	0.3	0.6	98.1
8	0.2	0.5	98.7
9 to 41	0.5	1.3	100.0

[a] eigenvalue divided by cumulative sum of eigenvalues
[b] cumulative sum of eigenvalues divided by sum of eigenvalues

The principal components and their loadings are based on time series of observed groundwater head. In a next step the spatial pattern of behaviour at different sites is investigated. The loadings of the observation well on each PC are interpolated with the nearest neighbour method and visualised in the geoinformation system ArcMap 9.1 (ESRI 2005). The algebraic sign of the loadings is incidentally allocated during PCA and thus can be either positive or negative. In addition the time series of the principal components are compared with time series of possible driving forces for temporal pattern recognition. Both steps are applied successively for the first four PCs, which are described in detail in the following subchapters.

4.1 First principal component

The first principal component explains 58.9% of the groundwater level fluctuations in the data set. Most of the observation wells show very high positive loadings (> 0.8) on this component (Figure 4) – apart from the observation wells located near to well field 'Heber West' (medium negative loadings) and Lake Wuhle (medium positive loadings). It is hypothesised that the first component represents the impact of groundwater recharge. Observation wells close to the well field 'Heber West' with medium negative loadings are more influenced by the higher pumping rates of this well field (15943 m^3/d compared to 8017 m^3/d total abstraction rate of well fields 'Heber Ost' and 'UWM' within the study period). A possible explanation for the lower positive loadings of three observation wells located near to Lake Wuhle is mainly due to low permeable soils in the alluvial floodplain and the additional impact of GW-SW interactions (e.g. reduced infiltration due to lake/stream bed clogging).

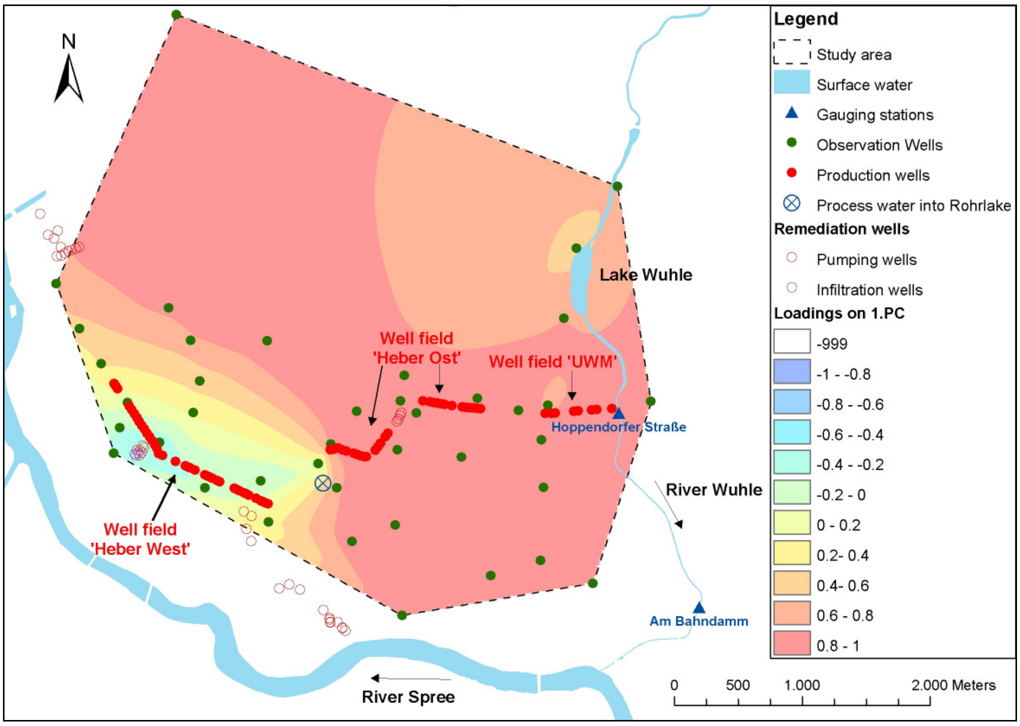

Figure 4 Spatial distribution of interpolated observation well loadings on the first principal component (modified from BUSSE 2010)

To confirm the hypothesis that the first PC describes the GW recharge process time series of the first PC values and cumulative daily precipitation rates are assessed in Figure 5. Scores of the first PC values increase after precipitation events. The sharp increase of the first PC scores between January and April 2007 is due to the cumulative precipitation of 186 mm (Jan.-March) on a nearly saturated soil at the end of the hydrological winter half year. The temporal delay of 17 days between last precipitation event (23rd March) and maximum score of the third PC (9th April) within this time frame can be regarded as mean delay of GW recharge inputs for all OWs due to the residence time within the unsaturated zone (depending on water retention curve, soil water content, unsaturated hydraulic conductivity function, depth to water table). Note that only increasing component values (i.e. groundwater levels) can be explained through precipitation events, while decreasing component values dependent on lateral GW discharge and additional evapotranspiration losses (especially during summer in case of low depths to the water table). Consequently the calculated Pearson correlation coefficient of 0.63 is quite high, which indicates that the initial hypothesis that the first PC describes the GW recharge process still holds true after temporal analysis.

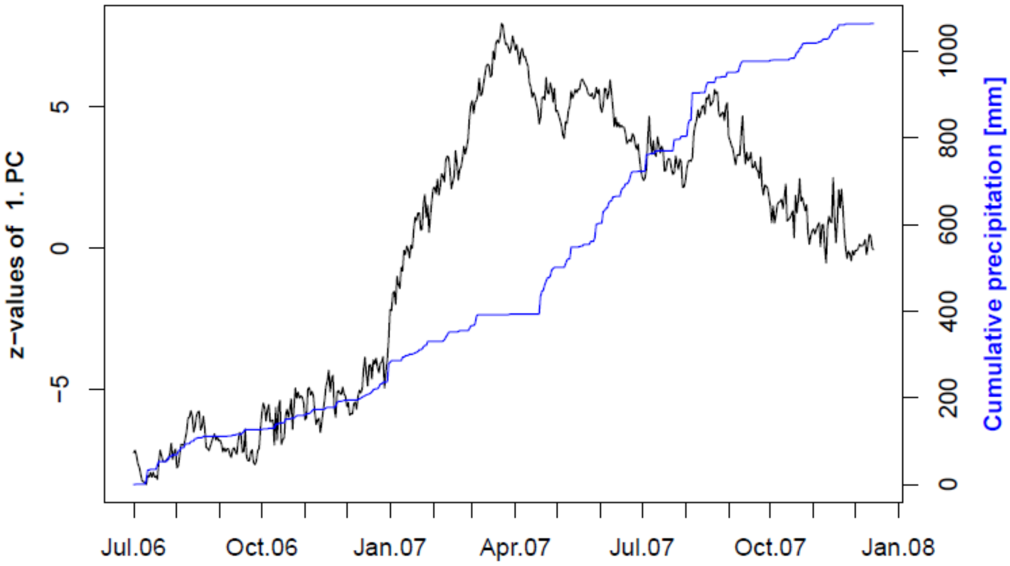

Figure 5 *Time series analysis of z-values of first principal component and cumulative daily mean precipitation measured at BWB precipitation station Wuhlheide (modified from BUSSE 2010)*

4.2 Second principal component

The loadings on the second principal component (29.8 % of explained variance) are highly negative for observation wells located near to the well field 'Heber West' (< -0.8), but increase in north east direction so that three observation wells located near to the Lake Wuhle show medium to high positive loadings (Figure 6). Since the negative loadings decrease with increasing distance from well field 'Heber West' it is assumed that the second component represents the impact of groundwater abstractions of this well field 'Heber West' on the groundwater levels of the surrounding observation wells.

For verifying this hypothesis the second PC time series and daily production rates of well field 'Heber West' are analysed (Figure 7). Cross-correlation between both time series yielded a maximum positive value of 0.69 for a time lag of 42 days. It can be regarded as quite high if one considers that the abstraction rates of this well field are not directly measured, but down-scaled from the combined abstractions of 'Heber West' and 'Heber Ost' by using a mathematical approach (see Chapter 3.1). The temporal delay of 42 days for the second PC values compared to the production rate of 'Heber West' results from the slow pressure transmission in the unconfined sandy aquifer around the well field (ASBRAND ET AL. 2004) and thus is quite plausible.

Consequently - after spatial and temporal pattern detection – it is concluded that the second PC represents the impact of abstraction rates of well field 'Heber West' on the water level fluctuations in observation wells nearby this well field.

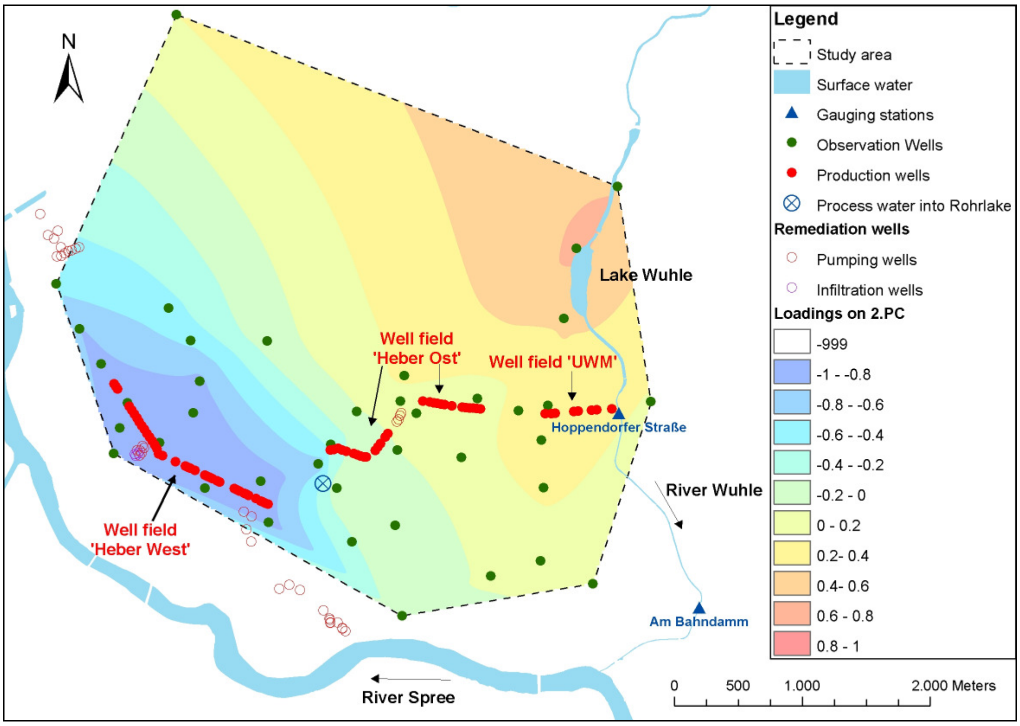

Figure 6 *Spatial distribution of interpolated observation well loadings on the second principal component (modified from (BUSSE 2010))*

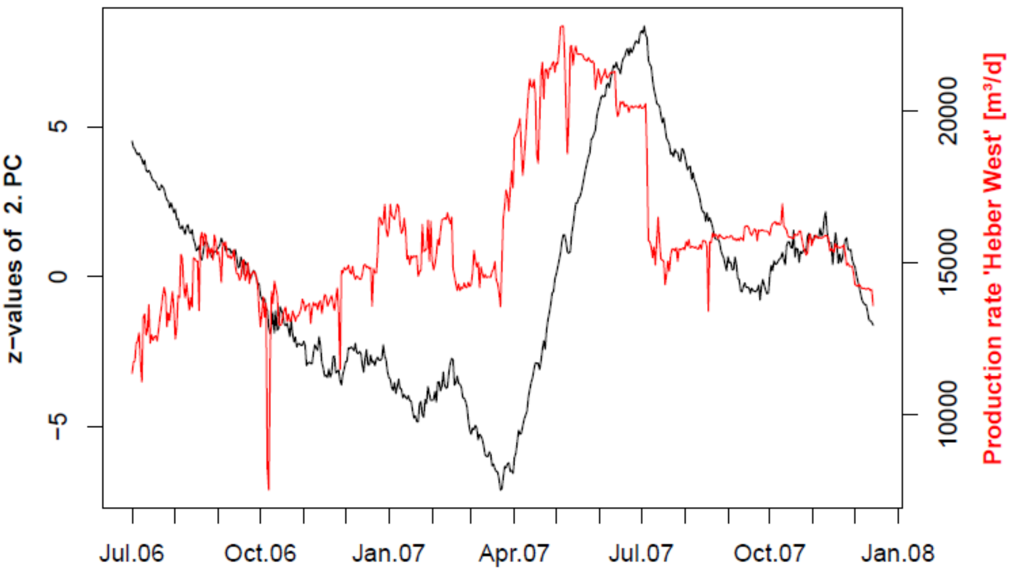

Figure 7 *Time series analysis of z-values of the second component and daily production rate of well field 'Heber West' (modified from BUSSE 2010)*

4.3 Third principal component

The third principal component explains 3.3% of the variance of the data set. Figure 8 shows that the strongest negative loadings on this component are restricted to observation wells located in the north eastern part along the River Wuhle (< -0.5) and close to the well field 'Heber West' (< -0.3). All other observation wells only have very weak loadings (-0.1 – 0.2). It is assumed that the third component represents the damping of the groundwater recharge inputs in the vadose zone, corresponding to the findings of LISCHEID et al. (2010).

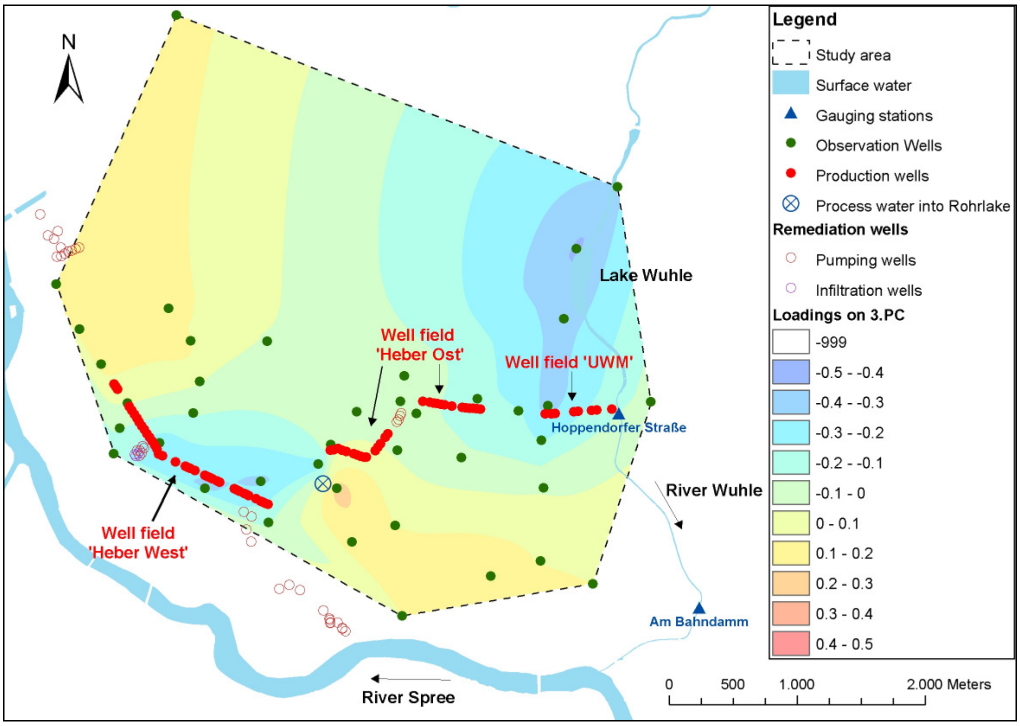

Figure 8 *Spatial distribution of interpolated observation well loadings on the third principal component (modified from BUSSE 2010)*

For verifying this hypothesis the 'mean depth to water table' for each observation well is used as a proxy and compared against the observation well loadings on the third component (not shown). In general, the lower the depth to the water table the higher the loadings on the third PC and vice versa. Consequently the more negative the loading on this component the stronger is the damping of the GW recharge signal. Taking into account the usually encountered enormous spatial heterogeneity of the hydraulic conductivity and of the water retention function in the vadose zone the Pearson correlation coefficient of -0.51 ($p < 0.05$) for all observation wells can be regarded as quite high.
In case that only OWs with high communalities (>0.66) on the first and third PC are considered (Figure 9), which represent GW recharge process (see Chapter 4.1) and its temporal damping, the correlation between OW loadings on third PC and the 'mean depth to the water table' is even -0.78. Thus it is concluded that the third PC describes the temporal damping of GW recharge inputs.

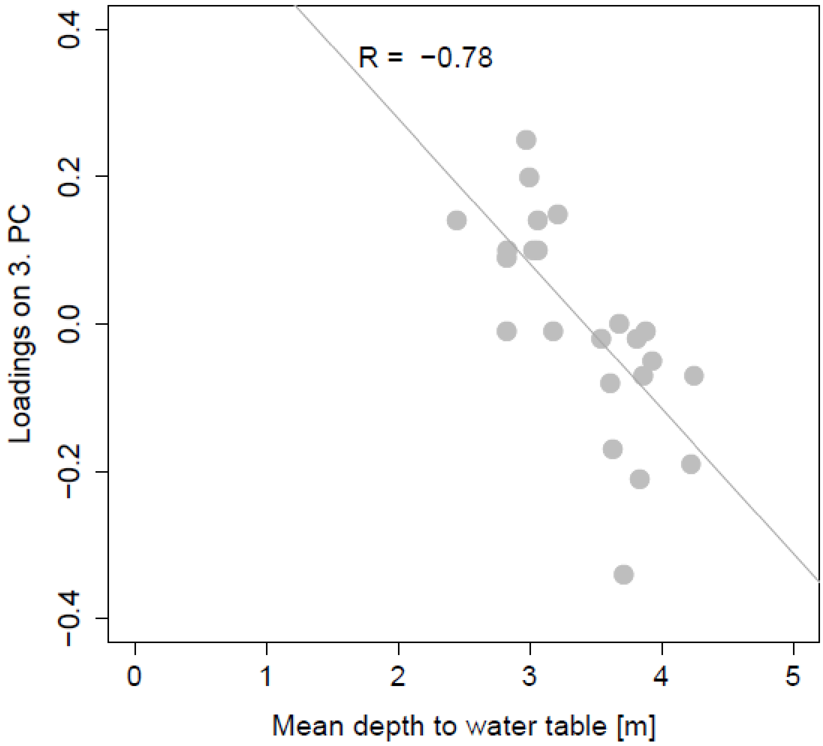

Figure 9 *Scatter plot of observation well loadings on the third principal component and mean depth to water table within the study period (modified from BUSSE 2010). Note that only 24 observation wells with communalities above 0.66 on first and third PC are shown.*

4.4 Fourth principal component

The fourth PC explains 2.9% of the total variance of the groundwater level fluctuations in the data set. Observation wells with the strongest negative loadings (-0.3 – -0.5) on this component are located near the discharge point of waterworks process water into the marshy ditch 'Rohrlake' (Figure 10). In addition the spatial extent of the area with the strongest negative loadings is nearly congruent with the course of the 'Rohrlake' (WIKIPEDIA 2008). Thus it is concluded that the fourth PC represents the infiltration of the Rohrlake into the underlying aquifer. However, it was not possible to verify this hypothesis by comparing the daily time series of the fourth component and monthly time series of process water discharge rates into the Rohrlake, due to the low temporal resolution of the latter.

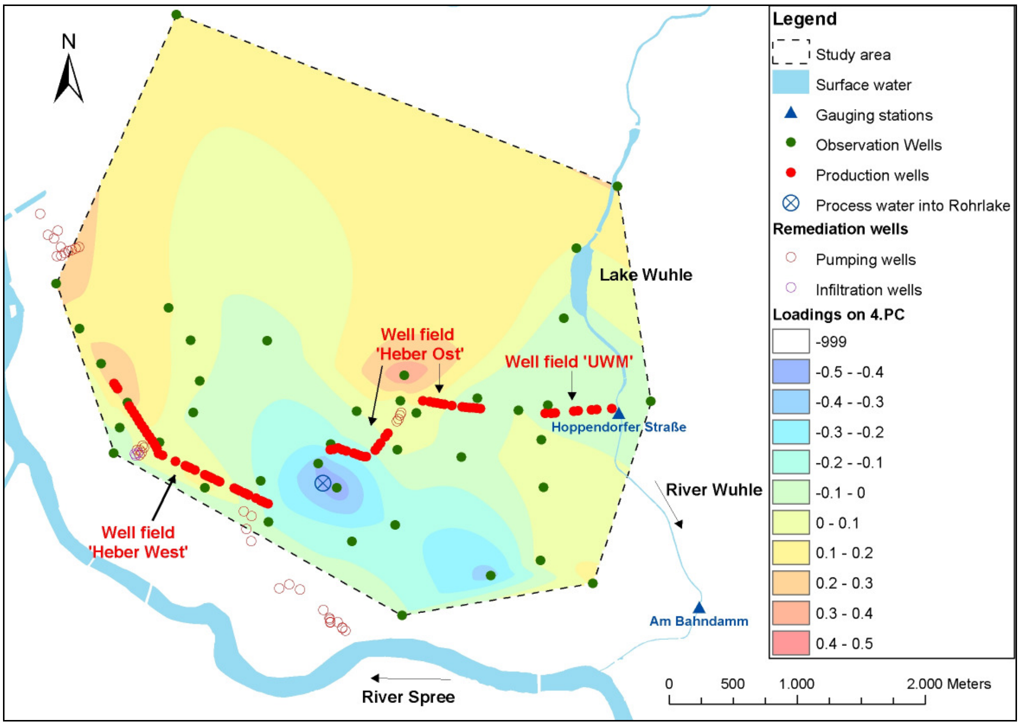

Figure 10 Spatial distribution of interpolated observation well loadings on the second principal component (modified from BUSSE 2010)

5 Conclusion and recommendation

The objective of the principal component analysis (PCA) was to determine the drivers of water level fluctuations in observation wells near a Berlin waterworks which may be influenced by multiple processes (GW-SW interaction, groundwater recharge, infiltration/abstraction wells used for groundwater remediation, production well fields for public water supply). The z-normalised daily groundwater level time series of 41 observation wells for the study period (19^{th} July – 1^{st} January 2008) easily satisfied the Kaiser-Meyer-Olkin measure of sampling adequacy with a value of 0.95. Thus these can be used as input parameters for the PCA.

PCA yielded that the first four PCs already explained 95 % of the total variance of the groundwater level fluctuations in the data set. In total 62.2% of the variance can be explained by groundwater recharge (first PC: 58.9%) and its temporal delay for different OWs (third PC: 3.3%), while any discernible impact of waterworks abstractions on groundwater level fluctuations is limited to the well field 'Heber West' (third PC: 29.8%). In addition the fourth PC (2.9% of explained variance) is identified as infiltration of the marshy ditch 'Rohrlake' into the groundwater. It is remarkable that this process is visible within the first four PCs, despite the fact that the discharge of process water into the 'Rohrlake' only accounts 14% compared to the combined abstractions of well fields 'Heber Ost' and 'UWM'. The most plausible explanation for this is that most observation wells are screened in the uppermost aquifer (see Appendix A, Figure 13), while both well fields abstract groundwater from the second aquifer but at a much lower rate (6800 m^3/d) compared to well field 'Heber West' (15940 m^3/d). Note that it was not possible to identify any impact of the remediation measures on the groundwater levels, since the production rates of the remediation wells were almost constant within the investigation period. However, the chosen PCA approach is based on analysis of temporal patterns and thus cannot detect any process that does not vary in time.

In a nutshell the main advantage for using PCA compared to process-driven groundwater flow modelling is that the driving forces for groundwater level fluctuations can be identified and quantified without requiring exact knowledge about the structural properties of the subsurface (e.g. aquifer transmissivities) and its input parameters (e.g. GW recharge, production rates). Note that the latter do not enter the PCA directly but are used for temporal and spatial interpretation of the results, which also requires some expertise. In case that a data set contains only variables with linear dependencies the PCA is the best available method for pattern recognition. However, its application is not limited to data sets with linear relationships, but it might be outperformed by nonlinear methods in case of strong nonlinear dependencies. Input data used for PCA showed some nonlinear relationships, but since the first four PCs can be interpreted in a meaningful way the application of the linear PCA is justified 'a posteriori'. In case that this precondition is not fulfilled a nonlinear PCA is required.

In addition it is not clear how the PCA results (time series of principal components, loadings of observation wells on PCs) will be influenced in case of changing input parameters such as:

- lower temporal resolution (e.g. weekly or monthly GW level times series of observation wells)
- higher spatial coverage (i.e. include more OWs for PCA at the expense of a lower temporal resolution)
- different boundary conditions (e.g. pumping regime)

Correspondingly, it is recommended to perform a sensitivity analysis in a next step. In case that the PCA results are not 'robust' to changing input parameters this would require an alternative interpretation of the dominant processes that drive the groundwater level dynamics in the observation wells.

In a next step the knowledge gained through PCA about the driving forces of the groundwater level fluctuations could be used to develop a statistical sound, empirical data-based model for well field modelling and operational optimisation (e.g. minimum drawdown), under the constraint that the boundary conditions do not change significantly over time (i.e. observed water levels during calibration period are representative even for more recent years). However, this approach cannot replace process-driven GW flow models (e.g. MODFLOW, FEFLOW) in case that the impacts of (i) changing boundary conditions (e.g. reduced bank clogging/GW recharge rate) or (ii) new well fields on the GW levels should be assessed in the framework of a scenario analysis.

The contents of this report were presented to the involved experts from the Berliner Wasserbetriebe (BWB). In agreement with their recommendations it was decided to focus further research within follow-up projects on the (i) sensitivity analysis of the PCA results and (ii) to apply nonlinear approaches for identification and quantification of processes that drive groundwater quality dynamics within the study area.

6 References

ASBRAND ET AL. (2004): Hydrogeologisches Strukturmodell für die Wasserwerke Wuhlheide und Kaulsdorf. Im Rahmen des wasserwirtschaftlichen Ressourcenmanagements der Berliner Wasser Betriebe Berlin, Fugro Consult GmbH

BACKHAUS, K.; ERICHSON, B.; PLINKE, W. & WEIBER, R. (2008): Multivariate Analysemethoden - Eine anwendungsorientierte Einführung. Heidelberg, Springer

BLACH-RADAU, C. (2008): Stand und Ziele der Grundwassersanierung im Anstrombereich des Wasserwerkes Wuhlheide Altlastensymposium Berlin, Senatsverwaltung für Stadtentwicklung.

BUSSE, J. (2010): Zeitreihenbasierte Analyse der Grundwasserdynamik in einem Brunnenfeld. Institut für Erd- und Umweltwissenschaften, Universität Potsdam

CURETON, E. E. & D'AGOSTINO, R. B. (1993): Factor analysis: An applied approach, Psychology Press

DIESNER, K. (2009): Hydrodynamische Bewertung geplanter eigenbewirtschafteter Brunnen im Wasserwerk Wuhlheide, Tauw GmbH: 32 p.

ESRI (2005): ArcMap 9.1

GUTTMANN, L. (1953): Image theory for the structure of quantitative variates Psychometrika 18(4): 277-296

JOLLIFFE, I. T. (2002): Principal component analysis. New York, Springer.

KERNS, G. K. (2007): R function: KMO sampling adequacy. https://stat.ethz.ch/pipermail/r-help/2007-August/138049.html

LEWANDOWSKI, J.; LISCHEID, G. & NÜTZMANN, G. (2009): Drivers of water level fluctuations and hydrological exchange between groundwater and surface water at the lowland River Spree (Germany): Field study and statistical analyses Hydrol. Processes 23(15): 2117-2128

LISCHEID, G.; NATKHIN, M.; STEIDL, J.; DIETRICH, O.; DANNOWSKI, R. & MERZ, C. (2010): Assessing coupling between lakes and layered aquifers in a complex Pleistocene landscape based on water level dynamics Advances in Water Resources, Doi:10.1016/j.advwatres.2010.08.002

LONGUEVERGNE, L.; FLORSCH, N. & ELSASS, P. (2007): Extracting coherent regional information from local measurements with Karhunen-Loeve transform: Case study of an alluvial aquifer (Rhine valley, France and Germany) Water Resources Research 43(4)

MANIAK, U. (2005): Hydrologie und Wasserwirtschaft. Berlin, Heidelberg, Springer-Verlag.

R DEVELOPMENT CORE TEAM (2009): R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, http://www.R-project.org

SENSTADT (2001): 04.02 Langjähriges Mittel der Lufttemperatur 1961-1990 (Ausgabe 2001). Digitaler Umweltatlas, Senatsverwaltung für Stadtentwicklung, http://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/ia402.htm

SENSTADT (2007): 02.17 Grundwasserneubildung (Ausgabe 2007). Digitaler Umweltatlas, Senatsverwaltung für Stadtentwicklung, http://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/ia217.htm

SENSTADT (2009): 02.12 Grundwasserhöhen des Hauptgrundwasserleiters und des Panketalgrundwasserleiters (Ausgabe 2009). Digitaler Umweltatlas, Senatsverwaltung für Stadtentwicklung, http://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/ik212.htm

WIKIPEDIA (2008): Karte des Hohen Wallgrabens und der Rohrlake, Berlin. http://de.wikipedia.org/w/index.php?title=Datei:Hoher_Wallgraben.png&filetimestamp=20090104040357

ZIPPEL, M. & HANNAPEL, S. (2008): Ermittlung des Grundwasserdargebotes der Berliner Wasserwerke mittels regionaler numerischer Grundwasser-strömungsmodelle Grundwasser 13: 13

Appendix A

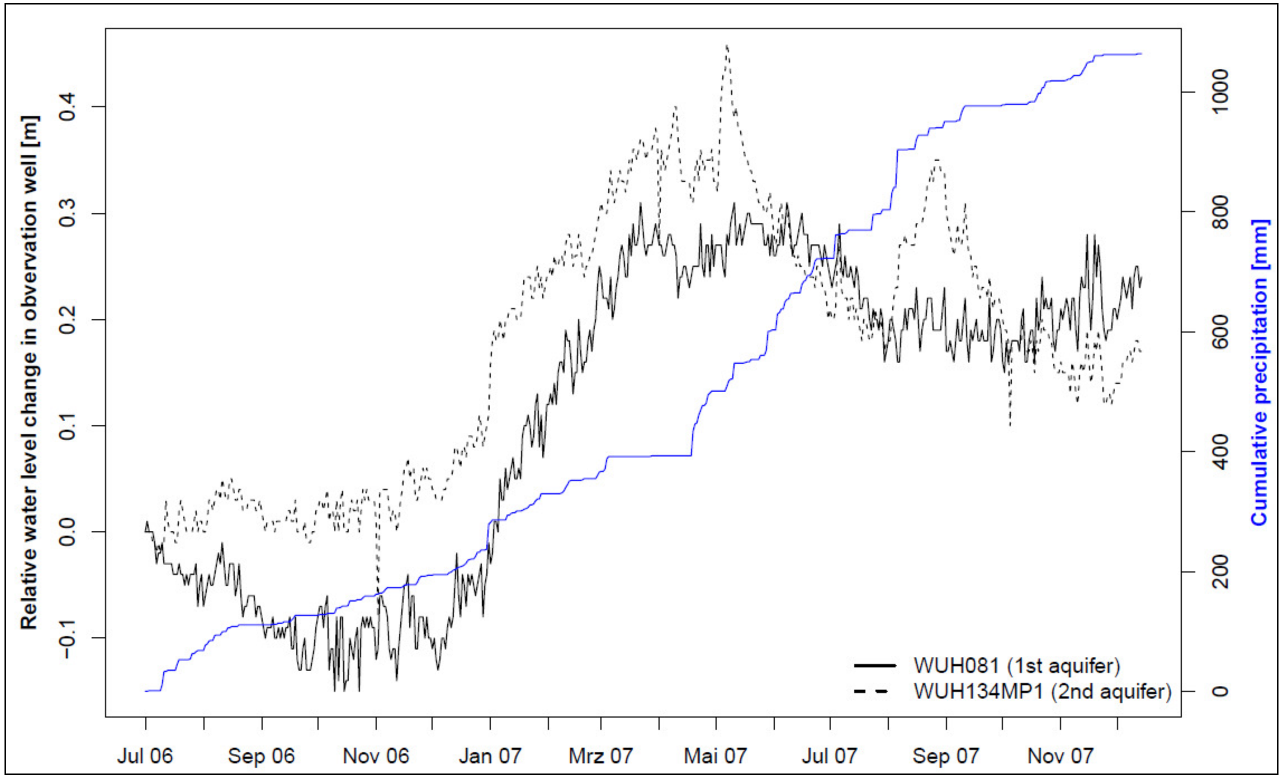

Figure 11 *Hydrograph analysis for two observation wells screened in different aquifers and located quite distant from the nearest waterworks well field (WUH081: > 3100 m, WUH134MP1: > 950m) and surface water bodies (see Appendix A, Figure 12). Note that the water levels in both observation wells are normalised to the inital water level at the beginning of the study period*

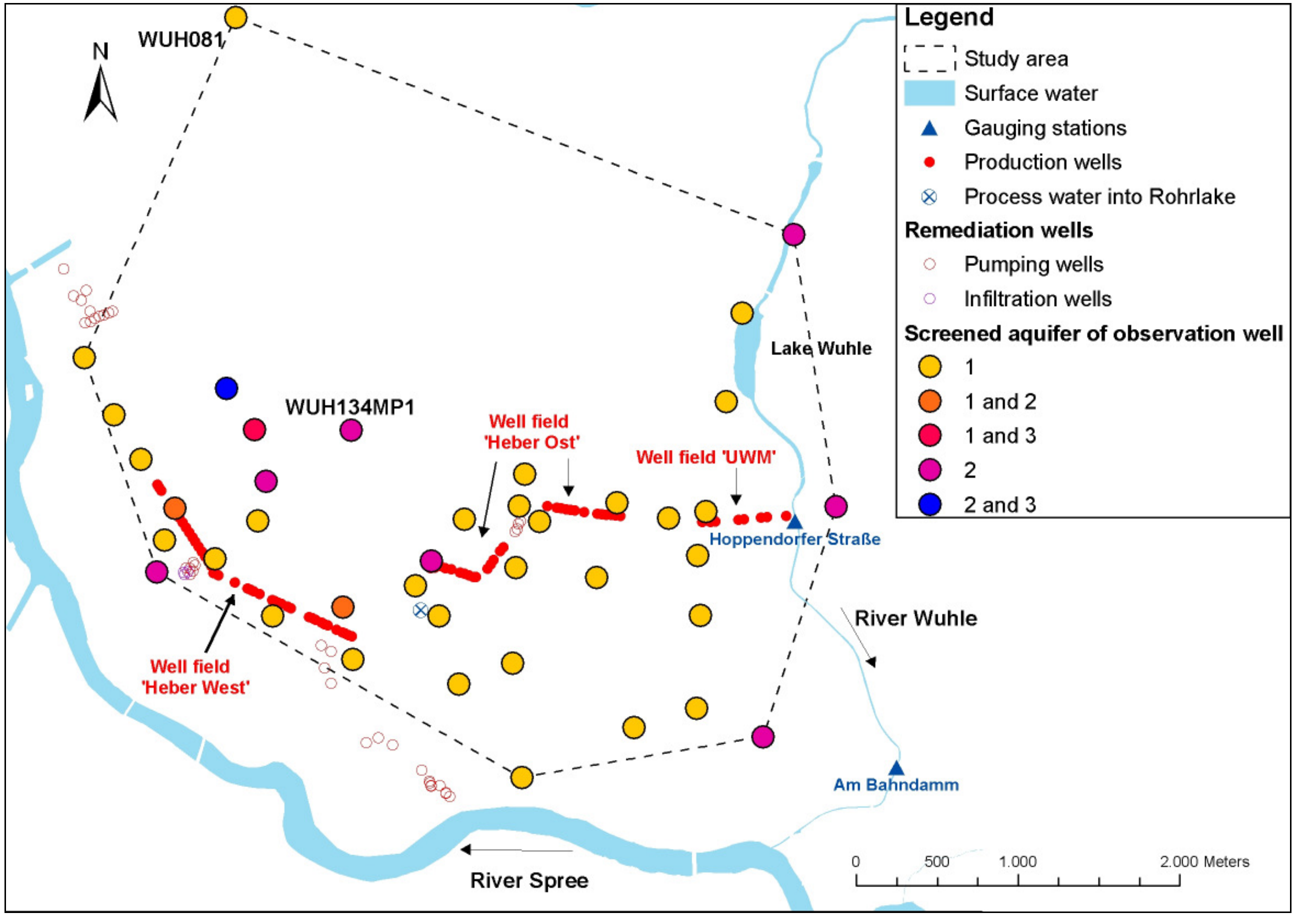

Figure 12 *Screened aquifer of observation wells. All waterworks production wells are screened in the 2nd aquifer (data source: BWB, SenGUV, TAUW) Labelled OWs (WUH081, WUH134MP1) are used for hydrograph analysis (see Appendix A, Figure 11)*

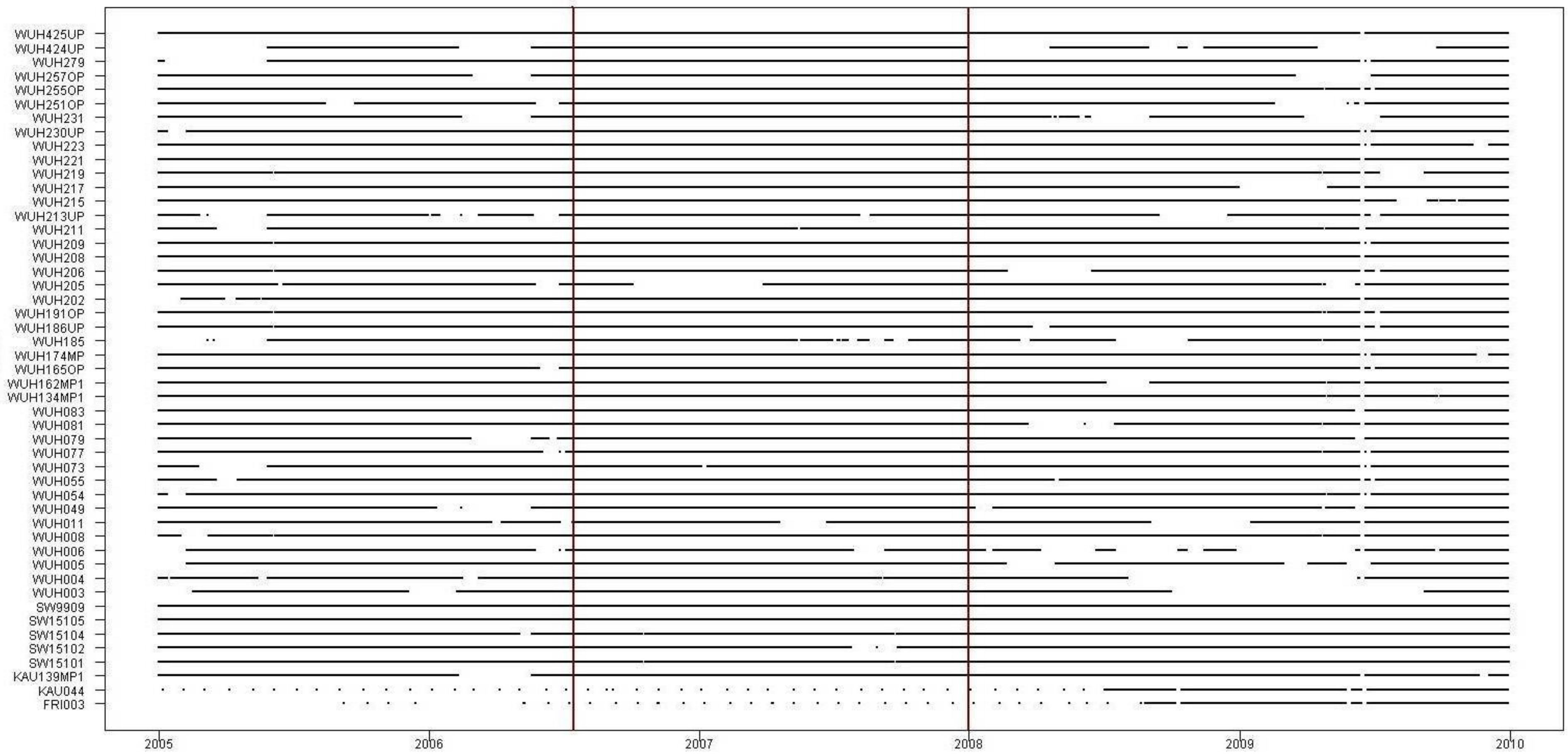

Figure 13 *Temporal coverage of the 49 observation wells with daily groundwater level measurements and visually selected study period (indicated by vertical lines); modified from BUSSE 2010*

Table 2 *Observation wells with percental data gaps within the study period and autocorrelation coefficients (r_k) for the longest continuous data gap period (k). Note that observation wells with data gaps > 15% (marked with *) and r_k values < 0.7 (marked with **) are excluded from PCA; modified from* B*USSE* *2010*

Observation well	Data gaps [%]	Longest data gap period k [days]	r_k
FRI003 *	96.43		
KAU044 *	96.43		
WUH205 *	32.89		
WUH011 **	11.84	62	0.058
WUH006 **	8.08	42	0.443
WUH213UP **	3.57	13	0.582
SW15102 **	11.65	33	0.618
WUH185 **	13.16	21	0.652
SW15101	2.82	3	0.905
SW15104	2.82	4	0.909
WUH073	1.88	7	0.918
WUH162MP1	0.38	2	0.96
WUH003	0.56	3	0.967
WUH279	0.56	3	0.978
WUH211	0.75	4	0.983
WUH008	0.56	3	0.986
WUH004	0.38	2	0.987

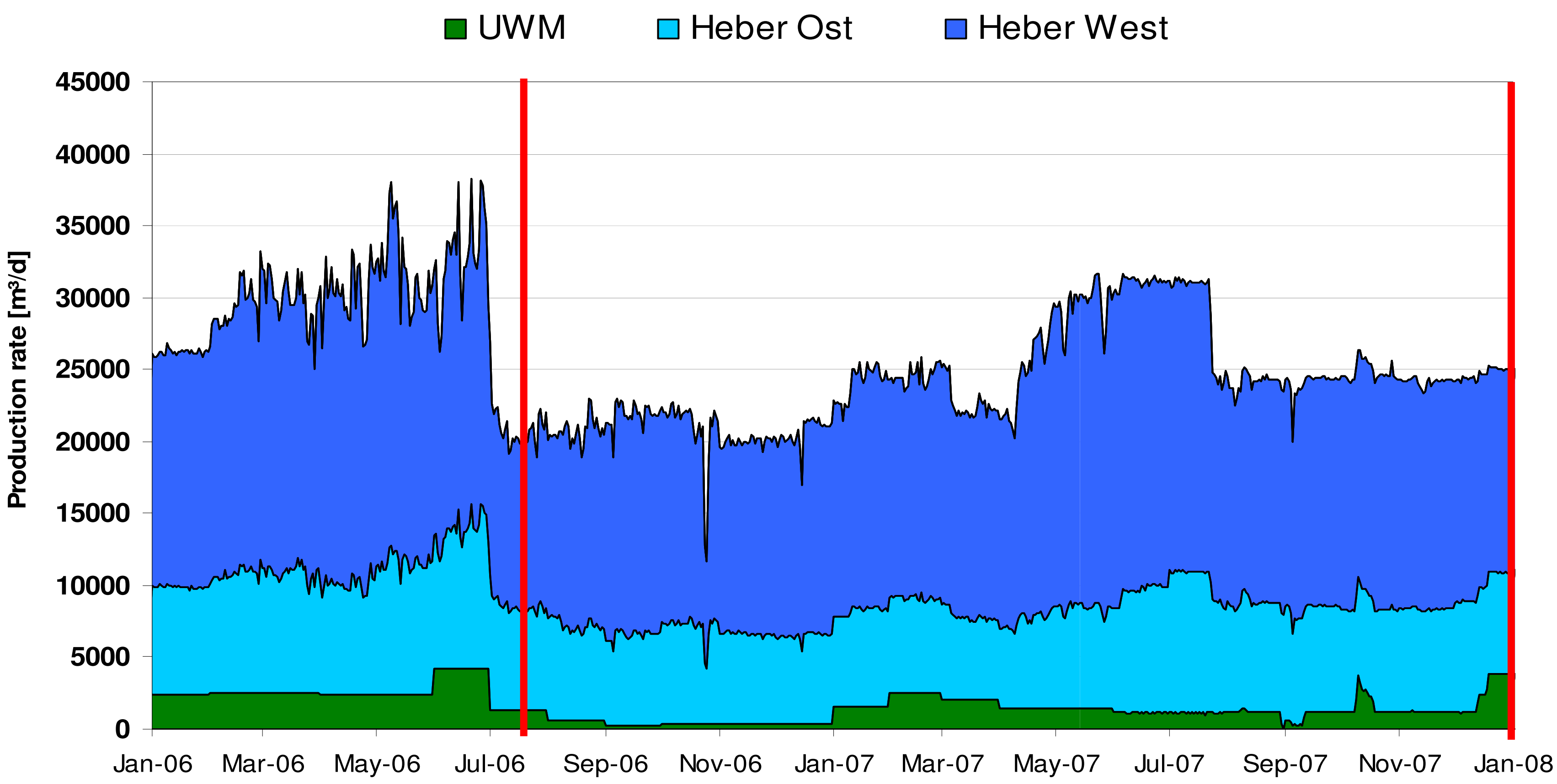

Figure 14 Production rates of the waterworks well fields (study period used for PCA is indicated by red lines), data source: BWB