CHALMERS
UNIVERSITY OF TECHNOLOGY

Systematic visual analysis of groundwater hydrographs: potential benefits and challenges

Downloaded from: https://research.chalmers.se, 2024-04-20 06:07 UTC

Citation for the original published paper (version of record):
Barthel, R., Haaf, E., Nygren, M. et al (2022). Systematic visual analysis of groundwater hydrographs: potential benefits and challenges. Hydrogeology Journal, 30(2): 359-378. http://dx.doi.org/10.1007/s10040-021-02433-w

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Hydrogeology Journal
https://doi.org/10.1007/s10040-021-02433-w

PAPER

Systematic visual analysis of groundwater hydrographs: potential benefits and challenges

Roland Barthel[1] · **Ezra Haaf**[2] · **Michelle Nygren**[1] · **Markus Giese**[1]

Received: 9 July 2021 / Accepted: 28 November 2021
© The Author(s) 2021

Abstract
Visual analysis of time series in hydrology is frequently seen as a crucial step to becoming acquainted with the nature of the data, as well as detecting unexpected errors, biases, etc. Human eyes, in particular those of a trained expert, are well suited to recognize irregularities and distinct patterns. However, there are limits as to what the eye can resolve and process; moreover, visual analysis is by definition subjective and has low reproducibility. Visual inspection is frequently mentioned in publications, but rarely described in detail, even though it may have significantly affected decisions made in the process of performing the underlying study. This paper presents a visual analysis of groundwater hydrographs that has been performed in relation to attempts to classify groundwater time series as part of developing a new concept for prediction in data-scarce groundwater systems. Within this concept, determining the similarity of groundwater hydrographs is essential. As standard approaches for similarity analysis of groundwater hydrographs do not yet exist, different approaches were developed and tested. This provided the opportunity to carry out a comparison between visual analysis and formal, automated classification approaches. The presented visual classification was carried out on two sets of time series from central Europe and Fennoscandia. It is explained why and where visual classification can be beneficial but also where the limitations and challenges associated with the approach lie. It is concluded that systematic visual analysis of time series in hydrology, despite its subjectivity and low reproducibility, should receive much more attention.

Keywords Groundwater monitoring · Groundwater statistics · Visual inspection · Similarity · Time series

Introduction

Background and objectives

In hydrological literature, there are frequent references to how a visual inspection of time series was carried out as an initial step in data analysis. In fact, even if not mentioned explicitly, in most hydrological studies it can be assumed that plots of time series data were looked at in some way. In modelling studies, where time series of observed and simulated data are compared, the visual comparison of time series seems to play an important role, even if articles focus on the "hard" performance criteria used for objective evaluations (Crochemore et al. 2015; Seibert et al. 2016). While visual inspections and analyses seem to be carried out frequently, studies rarely mention exactly how such visual analyses were made, and how the results of such visual evaluations may have influenced the entire process and hence the results. Only few studies examine visual techniques further, and analyse their role and importance (Crochemore et al. 2015; Ehret and Zehe 2011; Seibert et al. 2016). This article focuses entirely on the visual analysis carried out in a study concerned with groundwater hydrograph similarity.

The visual analysis presented in this article is part of a larger research effort, which aims to adopt the concept of catchment classification and similarity, as used in surface hydrology to groundwater hydrology. This concept has been developed through a large community effort in the International Association of Hydrological Sciences (IAHS) decade 2003–2012, which was dedicated to PUB (predictions in ungauged basins, Hrachowitz et al. 2013). The underlying idea of the concept is that similar systems (catchments), in similar states, will respond similarly (responses here mean

✉ Roland Barthel
roland.barthel@gu.se

[1] Department of Earth Sciences, University of Gothenburg, 40530 Gothenburg, Sweden

[2] Department of Architecture and Civil Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden

Published online: 11 January 2022

discharge hydrographs) if exposed to similar input signals (see e.g., Wagener et al. 2007). One fundamental challenge of the approach is the detection of similarities between different system responses (expressed as time series of, for example, river discharge or groundwater levels). Methods developed and validated for comparing discharge hydrographs cannot simply be applied to groundwater hydrographs, as the latter exhibit some fundamentally different characteristics (Barthel 2014). Within the larger framework of the research, various different approaches for similarity detection and classification based on direct times series comparison (Haaf and Barthel 2018), as well as feature extraction with subsequent clustering (Heudorfer et al. 2019), were developed and compared. Haaf and Barthel (2018) used visual inspection of time series as a method in its own right, but did not provide a systematic analysis of the visual approaches used or the results obtained. Visual inspection, comparison and finally classification of time series were mainly a single step in data preprocessing, yet also led to important decisions in the development of the research as a whole.

Based on the research described in the preceding, this paper explores whether visual classification may be more than simply an initial step in data preparation. On the face of it, there are many arguments against this, including the contention that visual approaches are subjective, tedious and impractical to apply to large datasets, and that they are not reproducible and transferable to other datasets. However, there are indications that systematic visual inspection/classification is highly advantageous during the early stages of a study, which may be important in particular for groundwater systems and groundwater observations. Each groundwater hydrograph and its dynamics constitute the combined result of a multitude of factors: geology of the aquifer and the stratigraphy, unsaturated zone processes, land surface properties and land use, climate, human impact and its change over time (Giese et al. 2020). Moreover, the technical design of the observation well itself plays an important role. In groundwater hydrology, time series have not been explored as comprehensively as in surface hydrology (Barthel 2014); many features of groundwater hydrographs and their links to driving and controlling factors remain poorly understood. Further, quantitative information on the factors influencing groundwater dynamics is scarce—for example, very little is known about the unsaturated zone below the root zone in aquifers with a greater depth (e.g. >10 m) to groundwater (Barthel 2006; Harter and Hopmans 2004). Furthermore, groundwater time series are often short, contain gaps, or are influenced by human activity. Sparse data of poor and uncertain quality, a multitude of influencing factors, and many known and, possibly, even more unknown, unknowns are difficult for automated, algorithm-based approaches (data mining, multivariate statistics, etc.).

The overall objective of this paper is to evaluate the general value of systematic visual analysis and comparison of groundwater hydrographs. An attempt is made to determine whether visual analysis and classification can be more than just a preliminary, informal task in data analysis and whether it might even be mandatory for classification and similarity-based approaches in hydrogeology.

To avoid potential misunderstandings, let it be stated that this paper is not introducing a new method. Rather, it presents the discussion of a method that many hydrologists and hydrogeologists apply on an almost daily basis, but which is nonetheless hardly ever mentioned in scientific literature. Visual inspection of time series, to get an initial idea of their quality or what they might reveal, is something done by everyone. The impression from an initial visual inspection of time series data may often have considerable impact on choices made when applying formal scientific methodology, but they may even provide insights that "real" scientific methods not only may not provide but may also overlook. The examples used in the paper are not presented to provide proof of the validity of a new method, they are used to illustrate potential benefits and challenges. Readers who are interested in the concept that forms the background of this discussion, and the specific role of visual analysis therein, are referred to a recent article in *Hydrogeology Journal* by Barthel et al. (2021).

Moreover, it needs to be highlighted that visual analysis is subjective. Readers, expecting clear criteria and well defined, reproducible concepts will be disappointed. The subjectivity and lack of conclusiveness of the subject may also be reflected in the language used.

Visual inspection and classification as a tool in (groundwater) hydrology

It may seemingly contradict what was previously written, but there is, in fact, a large body of literature on visual classification, comparison of time series and visualization techniques for detecting similarities stemming from a huge range of different scientific disciplines. Many approaches do not compare plots of time series as such (as in this paper), but use special techniques for transforming time series into other objects, which are then visually compared. Overviews explaining purpose, visualization techniques and different fields of applications are provided by, for example, Gleicher et al. (2011) or Gogolou et al. (2019). Overall, the characteristics of the time series used and the goals of visual analysis in individual studies span a very wide range. However, the nature of the times series analysed and the goals of the analysis in this body of literature are often very different from the analysis of groundwater hydrographs presented

here; therefore, a systematic review of this field of study is not justified in the context of this article.

In (groundwater) hydrology, visual inspection and comparison of time series is quite frequently mentioned as part of a workflow or diagnostic tool. However, most of the related publications do not explicitly describe how the visual analysis was carried out and what role it played in the work flow (e.g., Guzha and Hardy 2009; Harrigan et al. 2014; Li et al. 2017). Of the few articles that not only mention the use of visual tools, but also address their role as a part of the methodology, the most relevant in the context of this article is Ehret and Zehe (2011). They described visual inspection as "a powerful tool for simultaneous, case-specific and multi-criteria (yet subjective) evaluation" and included it as an essential element in the development of an approach to determine hydrograph similarity. They even suggested that visual inspection and comparison of hydrographs may be even more important than objective metrics. They stated "Eye and brain are a powerful expert system for simultaneous, case specific multi-criteria evaluation which provides results in close accordance with the user's needs. Due to these obvious advantages, visual inspection is still standard procedure for calibration and validation in engineering practice." They also found, for example, that "peak time metrics are much easier verbalized and applied in visual inspection than formulated and coded, as it requires automated identification of individual events […]." Surprisingly, this is one of the very few indications in hydrological literature that visual inspection may, in some cases, be superior to automated identification; however, Ehret and Zehe (2011) also mentioned that the major drawback of visual inspection is that "it is subjective and hence irreproducible and it is not applicable on large data sets." Similar statements can be found in most articles on visual analysis.

Visual comparison of time series receives slightly more attention in hydrological studies when it is used for assessing the performance of hydrological models. One example of a study covering "visual performance measures"—the comparison of the time series of modelled output to the measured time series of the same variable (essentially a similarity analysis)—was comprehensively described by Ewen (2011). Crochemore et al. (2015) asked a group of 150 hydrologists to evaluate a set of 20 hydrographs of simulated and observed data visually, in order to find links between the numerical criteria of the model's performance and expert judgment. They found that expert judgement was highly variable from one expert to another but they ultimately recommended that quantitative and qualitative evaluations be combined, because "Visual evaluation benefits from the knowledge of experts and from their skill and experience, and can be very helpful in providing finely tuned assessments of model accuracy." A similar study (summarized in Crochemore et al. (2015)) was presented by Chiew and McMahon (1993). Seibert et al. (2016) argued that, for the purpose of comparing simulated to observed hydrographs, "visual hydrograph inspection is still the most widely used technique in hydrology as it allows for the simultaneous consideration of various aspects such as the occurrence of hydrological rainfall–runoff events, the timing of peaks and troughs, the agreement in shape, and the comparison of individual rising or falling limbs within an event." They also stated that "Visual hydrograph inspection is hence a powerful yet demanding evaluation technique which is still rather difficult to mimic by automated methods."

To the authors' knowledge, no studies exist that explicitly address visual similarity analysis as a processing step. In general, similarity analysis of groundwater hydrographs is still rare with these few exceptions—Allen et al. (2010), Rinderer et al. (2017) and Rinderer et al. (2019). Allen et al. (2010) classified mountain valley groundwater systems using a number of aspects including the seasonal reversals of recharge and discharge from snowmelt and rivers. Rinderer et al. (2017) and (2019) used time series classification of groundwater hydrographs on a small headwater catchment in Switzerland to upscale and model groundwater dynamics from point to catchment scale.

Study area and data

The examples presented in this article are taken from two main study areas (Fig. 1)—one in central Europe (mainly Southern Germany plus Austria, France, Switzerland) and one in northern Europe (mainly Sweden, plus Finland); they will be referred to as the German and the Swedish datasets respectively. Here, only those aspects related to the study area and data that are immediately relevant to the aspects in focus are mentioned. Readers interested in more details regarding the data used and the various studies carried out can refer to Barthel et al. (2005, 2008, 2012, 2016, 2021), Gaiser et al. (2008), Giese et al. (2020), Haaf and Barthel (2018), Haaf et al. (2020), Heudorfer et al. (2019), Mauser and Prasch (2016), Nickel et al. (2005), Nygren et al. (2020), Römer et al. (2016) as well as the PhD dissertations of Haaf (2020) and Heudorfer (2019).

Jointly, the datasets comprise groundwater level time series from ~5,500 groundwater observation wells (~4,000 from central Europe). Additionally, a large variety of geographic, geological, hydrological and climate data were collected and used. Data were obtained from different agencies in several countries and are thus very heterogeneous in terms of time series length and resolution, metadata, geological and technical descriptions, etc. With respect to the raw data available for the studies summarized here, and in the context

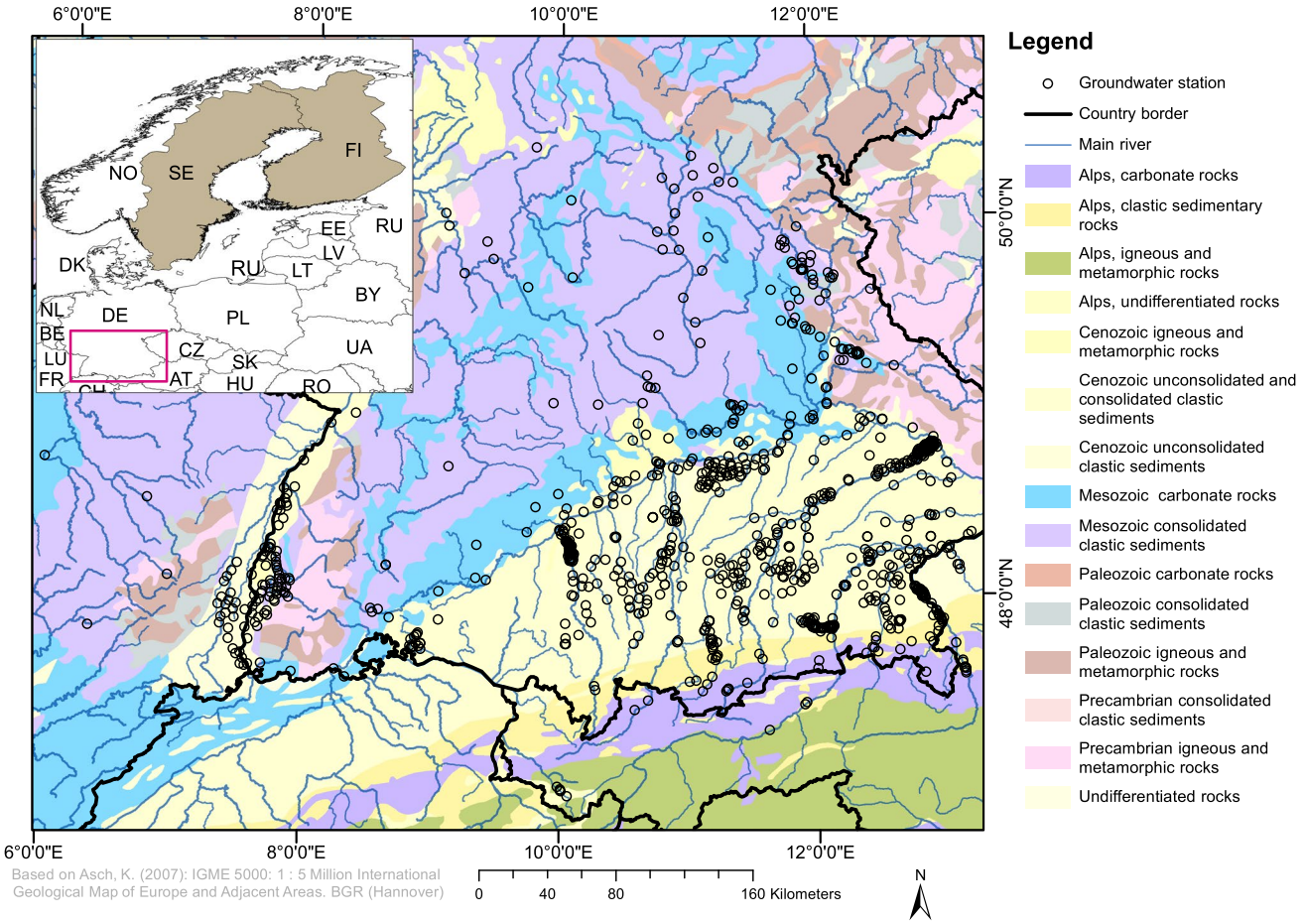


Fig. 1 Detailed map with observation well locations of the study area in southern Germany based on the International Geological Map of Europe IGME5000 (Asch 2007). Points are the locations of 751 out of ~5,000 observation wells in Central Europe used in the study described by Heudorfer et al. (2019). Note that Quaternary alluvial sediments, from which the majority of the observations were made, are often not explicitly distinguishable at this scale due to their small spatial extents. Quaternary alluvial sediments are typically located in narrow stretches alongside rivers. From Heudorfer et al. (2019). For the Fennoscandian study area (Sweden and Finland, highlighted in grey on the small map), observation well locations are not shown

of this article, it is important to note that most of the observation wells were:

- Relatively shallow, with only a few deeper than 200 m
- Often clustered in river valleys
- Often located relatively close to human settlements.

Additionally, the time series used:

- Differed widely in terms of length and regularity of measurement intervals, and total length of observation period
- Very often contained gaps, outliers, and (sometimes quite peculiar) irregularities (see, for example, plots d and h in Fig. 2).

A wide variety of examples of time series is shown in the results section; note that all plots are available in the electronic supplementary material (ESM) (Barthel et al. 2020).

To give the reader a flavour of the range of different appearances of time series within the dataset, Fig. 2 shows selected plots where plots a and b are very similar, while plot c is completely different (i.e. dissimilar) to both a and b. Plot f mostly resembles plot c, yet shows, in a rather subtle way, some overprinting of features found in plots a and b. This is even more pronounced in plot e, which could be described as a mix of the prominent characteristics of a, b and c. Plots d and h show examples of what is meant when a hydrograph is called "irregular" – in example d, the pattern for the first 4 years is apparently rather different (dissimilar) from the following years. Plot h shows a variety of "unusual" characteristics. First, there is a section in the middle that has a different (smoother) appearance than the sections at the beginning and at the end. Second, there are some deep downward peaks, the most prominent around 1995. It is important to keep this in mind as irregularities in

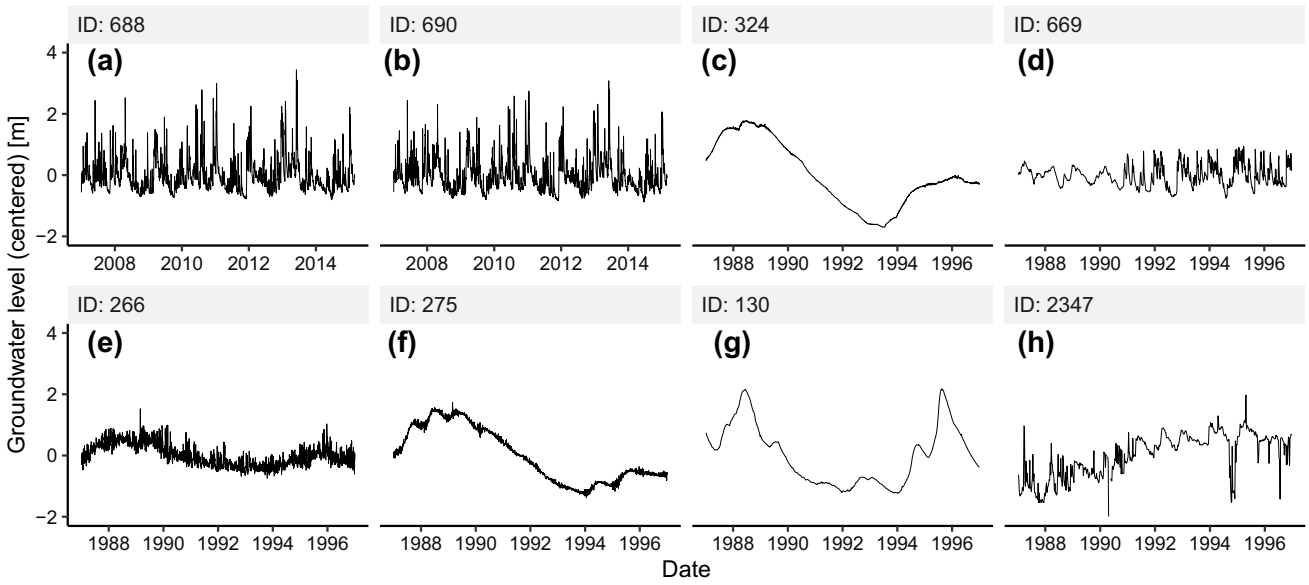


Fig. 2 **a**–**h** Groundwater hydrograph examples. Plots of daily measurements, spanning 10 years. Means are subtracted, but the original magnitude of fluctuations has been maintained

groundwater hydrographs are very frequent, occur in many peculiar and unexpected ways and form one main reason to carry out visual analysis.

It is highly recommended for the reader to take a look at the entire data set provided online in Barthel et al. (2020), as the within-group similarities and between group similarities of the time series are much more striking when looking at a large number of plots than at the few examples presented here in the main document.

Methods

Plots of 1,096 time series from southern Germany (841) and Sweden (255) were visually inspected and grouped according to similar appearance. To evaluate the performance of visual classification, the results were compared those from algorithm-based classification techniques applied to the same dataset. To explore whether the results of visual classification can be explained by the hydrogeological conditions where the measurements were taken, the time series classification results were compared with the results of a hydrogeological classification.

Visual comparison and classification of time series

Preprocessing and plotting

The following methodological considerations are primarily concerned with similarity analysis and classification, but will be applicable to many other uses of visual analysis as well. To achieve a meaningful visual comparison, plots do need to show differences between time series as clearly as possible, but otherwise they need to be identical (scaling, layout). This may sound obvious but is easier said than done in practice. Many choices exist, yet many of the best ones may not be feasible (for example, the plots required may be too large). The choice of plot has a significant impact on the visual appearance, and thus the analysis results. Choices include, for example, the length of the plotted time interval, temporal resolution, graphical resolution, standardization, and how to deal with gaps (either not plotting or filling with straight lines) among others. Even line thickness, colours and symbols have great impact. The requirement for identical layout requires harsh compromises. As groundwater head data are expressed either in meters above sea level or depth below surface, the subtraction of the mean (or shifting graphs along the y-axis) seems mandatory. Groundwater level fluctuations have very different magnitudes (in this dataset, from a few centimetres up to 15 m within a decade). Z-scores are an option to achieve comparable y-axis scaling, with the disadvantage that differences in magnitude, a significant hydrogeological feature, are lost. The solution taken here was to work with two sets of plots—one with just the mean subtracted, the other showing z-scores. Different plot lengths were also tested: 1, 3, 10 and 30 years. As time series start and end at different dates and have different lengths, plot periods become short if one wants each plot to have the same start and end dates. Longer plot periods with different start and end dates (the chosen option) have the disadvantage that particular events (e.g. a significant drought or flood creating characteristic peaks) may be visible in one

plot but not the other. Also, within a short period such as 1 year, signals with long wavelengths are lost, while long periods (30 years) mean that high frequency signals (visually) disappear if plots are not made extremely large. One option tested to overcome this issue was to use EEMD—ensemble empirical model decomposition; Wu and Huang (2009)—plots that show different frequency components of time series in parallel (Fig. 3e).

While different styles of plots emphasize different characteristics of time series, using many different plot styles in parallel has proven unfeasible and confusing, and a consistent classification of time series, taking several different features into account, cannot be achieved. A significant disadvantage of visual techniques is that the human brain, while very good at detecting differences in a small number of objects of the same kind, can only process a rather limited quantity of different information at a time.

In Fig. 3a–d, plots clearly show the impact of standardization on plot appearance: the pairs (a and c; and b and d) show the same hydrograph, one with only the mean subtracted, the other as z-scores. Standardization has the advantage that all plots can be more easily fitted into a chosen plot area with equal size and axis scaling; however, it changes the appearance, drastically in some cases, whereby some characteristics are understated, while others are exaggerated. Again, it depends on the overall objectives of a study as to what is most appropriate. As always, "appropriate" is a relative term: it depends on the research question or hypothesis to be tested. In a study on the impact of long-term climatic changes on groundwater, high frequency signals may be less important, while those might be extremely interesting in a study on shallow groundwater dependent eco-systems. Figure 3e shows an EEMD plot of the same time series shown in Fig. 3b,d. Splitting the time series into different frequency components can be very helpful but, overall, this style of representation was not found to be very helpful in the classification stage of a classification; the additional information provided cannot be "processed" meaningfully. It may prove helpful in a stage where one goes from simple visual characterization to conceptual interpretation in relation to hydrogeological conditions.

In addition to the technical challenges associated with plotting, other challenges arise from the nature and quality of groundwater hydrographs. As described in section "Study area and data", many hydrographs are irregular, have different length and measurement intervals, gaps, etc. It is thus not always possible to visually compare hydrographs with identical technical descriptions—i.e. identical length, start date or measurement intervals—if one wants to include a large enough number of time series in the comparison. Compromises have to be made. In this context it is worth mentioning that of the more than 5,000 time series available in the projects, only about 1,300 were selected for the visual classification exercise. The rest were removed prior to visual inspection using automatic procedures. Reasons to remove time series from the data set were for example: very short time series, time series with start and end dates far outside the range of the majority of time series, time series with very long gaps, many gaps or very irregular measurement

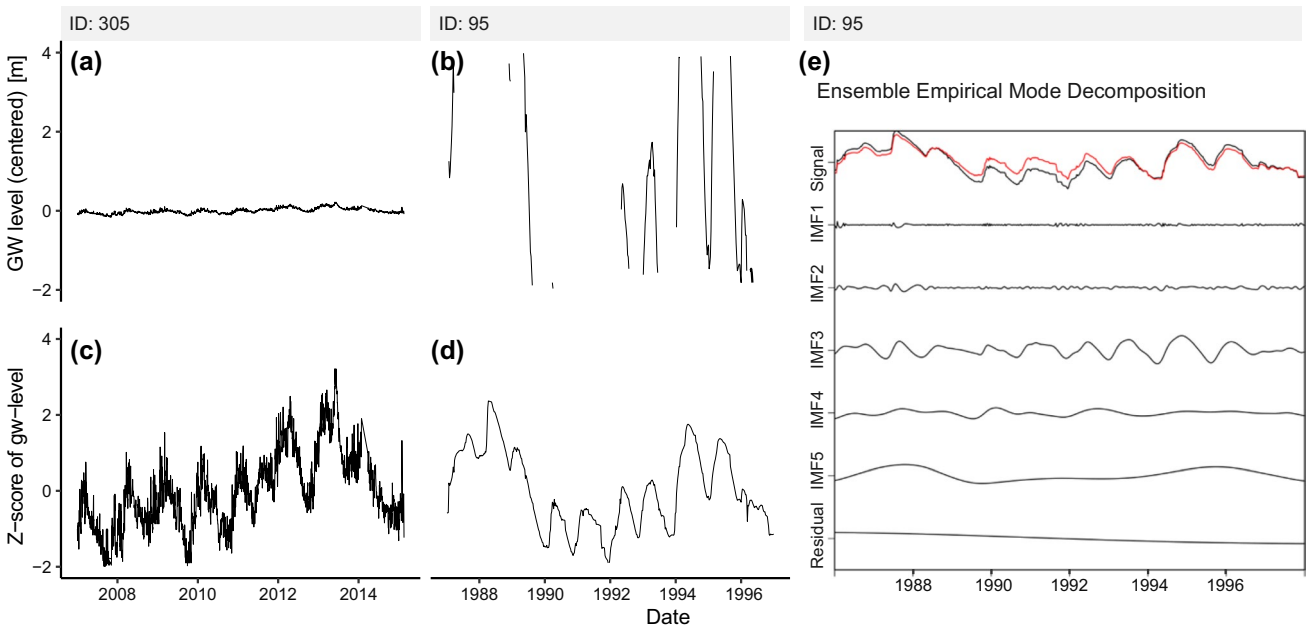


Fig. 3 Plots clearly showing some of the many different options to plot times series data and associated challenges: **a**–**b** only mean subtracted, **c**–**d** z-scores of the same time series (**a** and **b**), **e** ensemble empirical mode decomposition of the time series shown previously (**b** and **d**)

intervals. The ~100 time series classified as "irregular" in the present study, are those which passed the automatic quality control, but were found irregular on closer visual inspection.

The final preprocessing and plotting options chosen were:

- Plots of 10-year-long time series of daily data (with different start dates); shorter (3 years) and longer (30 years) time series were used in some unclear cases to support the decision.
- Original data with mean subtracted, on y-axis ranging 6 m (adjusted to the min and max values of the data); z-score data and EEMD were used in parallel, but only to help make classification decisions in very unclear cases.
- Gaps were filled using linear interpolation, but a set of plots with gaps left blank was also used (which option is better depends mostly on the length of the gaps).

Time series classification according to similarity

One of the most interesting methodological questions in relation to classification of time series is "What makes two time series appear similar or dissimilar?" Looking at the examples shown in Fig. 2a,b, most people will likely agree that the plots are quite similar, while the plot in Fig. 2c is clearly dissimilar from plots of Fig. 2a,b. In this case, where plots show strong similarity, to arrive at this conclusion is straightforward, a simple glance is sufficient. It becomes much more difficult when patterns are mixed or irregularities occur (see Fig. 2 and associated text). Figure 2e, for example, has features that resemble Fig. 2c and also Fig. 2a. How to deal with this is a great challenge and, ultimately, the point where subjectivity dominates the process. Without elaborating further on this, there are different options to visualize time series classification according to similarity. One can:

1. Look for characteristic patterns in a time series, make an inventory of patterns and then group the time series according to the combination of patterns found. Figure 4 explains what is meant by patterns in this sense.
2. Look at the overall visual appearance only, without making attempts to identify and characterize patterns.

The first attempts at visual classification in this study were carried out according to option 1. They were started

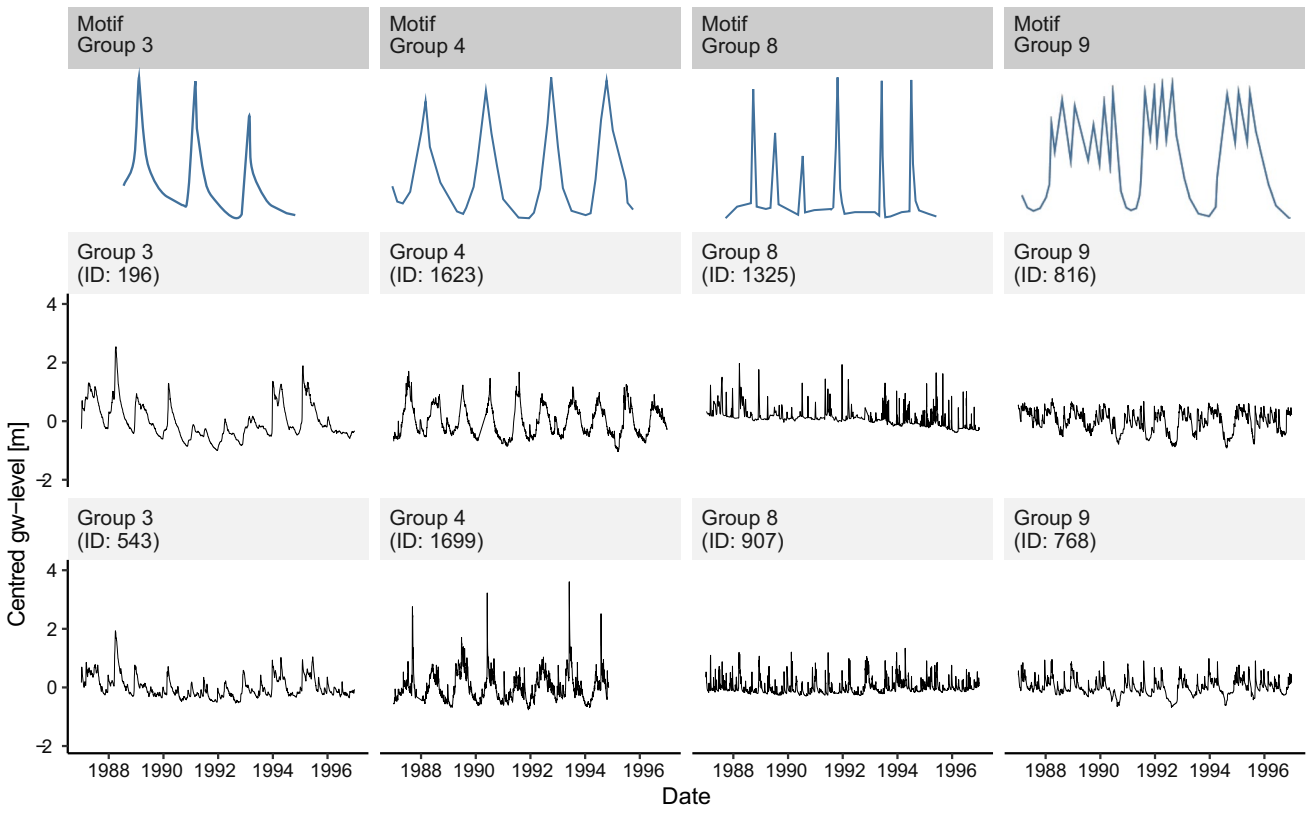


Fig. 4 Selected examples of "eye-catching" patterns in groundwater hydrographs. The upper row shows simplified sketches of four patterns, the middle row shows a relatively pure representation of the respective pattern above, the lower row shows the pattern apparently superimposed on other patterns

from a set of unique patterns (peak symmetry, upper/lower bounds, flashiness, etc.), after which it was attempted to identify the time series where those patterns occurred. Figure 4 shows four selected examples of "eye-catching" patterns. Some attempts were made to create an inventory of patterns based on verbal descriptions ("symmetric", "flashy") and sketches. Attempts were also made to explain potential hydrogeological causes. While visual pattern identification may be a valid and even less subjective approach to visual classification, paving the way for semiautomated or automated classification—see section 'Comparison with index-based classification' and Heudorfer et al. (2019)—it was finally decided not to follow the approach as it proved to be extremely tedious. The main challenge is that most time series do not show patterns in pure form, but show rather the superposition of many patterns. In Fig. 4, the uppermost row shows simplified versions of several patterns which frequently occur in the dataset. The middle row of Fig. 4 shows examples of time series that show the same patterns in relatively pure form. The examples in the lower row of Fig. 4 show superpositions of the respective pattern with other patterns. The eye seems well able to distinguish between basic patterns in such a mixture; the challenge remains to decide what pattern, or mixture of patterns, is characteristic for a specific group, subgroup or type. In relation to the attempts to explain patterns using expert hydrogeological knowledge, it should be noted that this may introduce a strong bias and wishful thinking ("this is a shallow unconfined aquifer and thus must go into this group", even if it looks different from time series from similar locations) and should probably best be avoided. It should also be noted that visual classification essentially struggles with the same issues as formal, algorithm-based classification—in this example, it is the struggle between structural similarity and shape-based similarity as explained in work such as Lin and Li (2009).

The results presented in this paper were reached exclusively using overall visual impression. It should be mentioned that even when one strives to carry out a characterization of the overall visual impression only, one automatically starts to look for patterns, even if those are not systematically described and named.

The classification process was carried out stepwise over several iterations. The main approach used was to display miniature plots on two computer screens in parallel with 20–80 miniature plots on each screen. The entire dataset with unsorted plots was shown on one screen, while a new folder with sorted plots was shown on the other. Plots that appeared similar were selected from the "unsorted" folder displayed on screen 1 and dragged and dropped to the sorted folder on screen 2. First, a very coarse sorting was carried out, leading to a small number (here, 9) of rather large groups. In the next step, subgroups were formed within these large groups, using the same drag and drop procedure as before. Subsequently, subgroups were split into sub-subgroups, called "types, using the same approach. In a final step, each type was compared with the other types and plots were redistributed to achieve maximum similarity in each subgroup. Cross-checks with other plot styles (see previous section) were carried out, and parts of the process were iterated several times. The "type-to-type" comparison sometimes led to splits, joins and removal of types, subgroups and groups.

The process resulted in a hierarchical scheme consisting of groups, subgroups and types. Groups were named accordingly using a a 3 digit scheme (X,Y,Z), where X identifies the group, Y the subgroup and Z the type.

A strictly hierarchical, exclusive classification (meaning each hydrograph belongs to exactly one type) leads to ambiguous results. Many hydrographs could belong to one type, but equally well to another. Visual inspection as used in this study does not provide the means to come to clear decisions due to the lack of objective criteria. For that reason, the authors have experimented with "fuzzy classifications" where a hydrograph can belong to more than one type. The methodology used to achieve the fuzzy classification and the respective results are presented in the electronic supplementary material (ESM).

Comparison of visual classification results with other data

The results of the visual classification were compared to the result of index-based classification (Heudorfer et al. 2019) and the results of direct comparison of time series based on different distance measures and subsequent clustering (Haaf and Barthel 2018). Moreover, visual classification results were compared to a set of hydrogeological descriptors and the results of hydrogeological classification (Giese et al. 2020). The comparison of visual classification with the results of the direct comparison is described in some detail in Haaf and Barthel (2018) and thus are not shown here.

Comparison with index-based classification

The index-based classification scheme is described in detail in Heudorfer et al. (2019). The fundamental idea is as follows—for each time series, a set of indices is calculated, each index expressing one characteristic feature (a "pattern") of groundwater dynamics. Heudorfer et al. (2019) developed a typology of groundwater dynamics with three major categories of features, namely structure, distribution and shape, each having several subcategories. Structure, for example, has the subcategories seasonal magnitude, seasonal timing, interannual variation and flashiness. Within each subcategory, one or more indices are defined—for example, the *Richard-Baker index*, in the subcategory *flashiness* is

the "sum of absolute values of day-to-day changes in head divided by the sum of scaled daily head" (Baker et al. 2004). The index *Average seasonal fluctuation*, belonging to the category "Seasonality magnitude", is the "Mean annual difference between the averaged 3 highest monthly groundwater heads per year and the averaged 3 lowest monthly groundwater heads per year" (Martens et al. 2013). Indices were calculated for 10-year time series of weekly and 5-year times series of daily data.

Heudorfer et al. (2019) considered a total of 62 indices, collected from different studies in hydrology and related fields of science. Of those 62, nine indices representing the previously mentioned subcategories were selected for comparison with the visual classification results (Table 1).

To compare the visual classification results with the index-based approach, mean and standard deviation for the selected indices from time series within each type, subgroup and group found through visual classification were calculated. A semiquantitative comparison was carried out with the aim of determining whether similar time series have similar index values.

Comparison with hydrogeological data

Within the scope of the wider research framework presented in this article, one main objective of time series classification is to be able to establish and use dependencies between time series characteristics (dynamic behaviour) and hydrogeological conditions. This has the ultimate aim of being able to use the found dependencies to make predictions. Therefore, the question as to whether a found classification scheme has any relation to hydrogeological conditions is a crucial one to be answered. It is assumed that the index-based classification introduced in the previous section shows some degree of relation to hydrogeological conditions, as many of the indices are based on theoretical and empirical considerations of the dynamic behavior of groundwater resources. The visual classification based on "visual appearance" only, however, cannot rely on such assumptions; therefore, the authors found it beneficial to take a closer look at this question. To do so, how selected numerical descriptors from a dataset containing geological and borehole information related to the found classification scheme were analysed. For the purpose of this article, this analysis was only carried out qualitatively. An in-depth quantitative analysis of the same question, based on indices, is described in Haaf (2020) and Giese et al. (2020). Those articles also explain how the descriptors used were developed and why and how selections for the purpose of those analyses were made.

Results

Visual comparison and classification of time series

The classification was carried out independently for both datasets (Germany/Sweden) and for a combination of the two datasets. The Swedish dataset was also independently grouped by two different observers. In the results section, only details of the results for the German dataset are present as the results for the Swedish dataset do not fundamentally differ. Results of the classification of Swedish data and the combined classification (German and Swedish data in one data set) can be found in the (ESM).

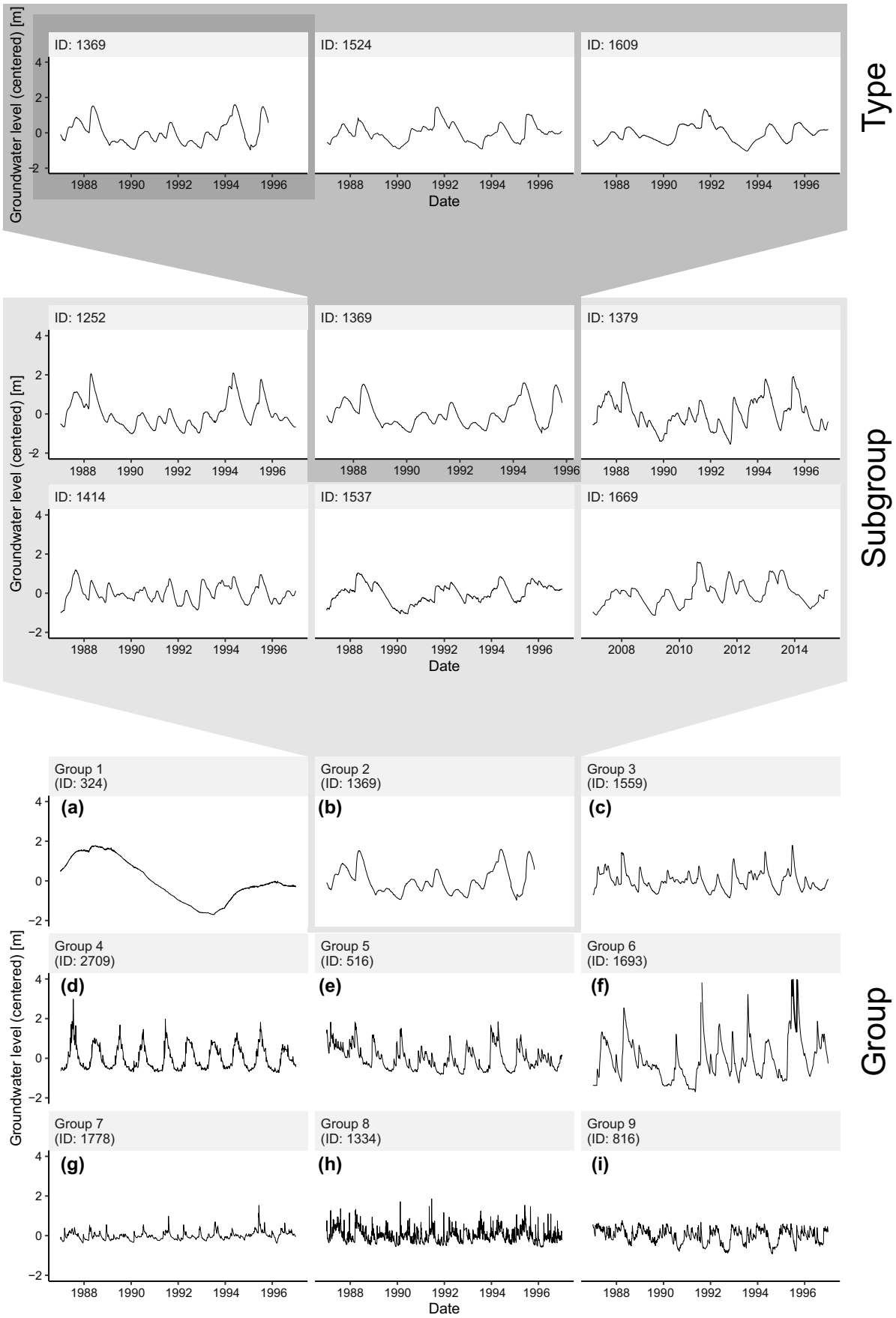
In the context of this study, both exclusive classification and fuzzy classification were tested. Exclusive classification means that each hydrograph is assigned to one type only, while in fuzzy classification each hydrograph can be assigned to many types.

Exclusive classification

For the German dataset of 841 time series, the procedure described in section 'Visual comparison and classification of time series' resulted in 82 visually distinguishable types within 28 different subgroups and nine groups. Up to 31 individual time series were grouped into one type, while a few types had only two or three members. Additionally,

Table 1 Indices selected for the comparison with visual classification results. See Heudorfer et al. (2019) for detailed explanations and references for all 62 indices

Aspect	Component	Index name	Index abbreviation
Structure	Seasonality magnitude	Coefficient of variation (CV) of mean minimum monthly head	cv.mon.min
	Seasonality timing	CV of date of annual minimum head	vardoy.min
	Flashiness	Richards pathlength	pathlength
	Interannual variation	Low/high pulse count	pulse.count.h
Shape	Slope	Recovery constant	recov.const
	Scale	Peak base time	peakbase.avg
Distribution	Modality	Bimodality coefficient	bimod
	Boundedness	Median with 0–1 scale	median
	Density	Mean of annual maximum	avg.ann.max


Type
ID: 1369
ID: 1524
ID: 1609
Groundwater level (centered) [m]
Date
Subgroup
ID: 1252
ID: 1369
ID: 1379
ID: 1414
ID: 1537
ID: 1669
Group
Group 1 (ID: 324)
Group 2 (ID: 1369)
Group 3 (ID: 1559)
Group 4 (ID: 2709)
Group 5 (ID: 516)
Group 6 (ID: 1693)
Group 7 (ID: 1778)
Group 8 (ID: 1334)
Group 9 (ID: 816)
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

◂**Fig. 5** Lower panel with nine plots: Examples of plots of time series from each of the nine groups defined by visual classification. Panel in the middle with six plots: examples of plots of time series from subgroups that were defined within group 2. Upper panel with three plots: examples of plots of time series from one type (type 2.2.1), which was defined within subgroup 2.2

there were about 29 time series that did not fit into any of the found types; thus, these were classified as "not fitting anywhere". Seventy-one hydrographs were put into a group called "irregular". Note that all time series with obviously peculiar behaviour, outliers, gaps etc. were removed from the dataset before starting the process of visual classification described here. The irregular time series detected here were, therefore, mainly those that were not spotted in the preparation stage. Plots of all the time series and classifications made are available in the ESM (Barthel et al. 2020).

Figure 5 shows the diversity of the time series used in this study, an overview of the used hierarchal classification scheme and a flavour of the achieved classification. The lower panel of Fig. 5 shows examples of time series from each of the nine groups created. The panel in the middle of Fig. 5 shows six representative examples of plots from one subgroup each, each belonging to group 2. The upper panel shows three representative examples from one type that was defined within the subgroup 2.2 of group 2.

Figure 6 shows some of the challenges associated with visual similarity analysis, using examples of plots of time series from two types, one defined within group 8 (type 8.3.1), the other in group 9 (type 9.5.1). Types 8.3.1 and 9.5.1 are clearly dissimilar from each other, while the within-type similarity of type 9.5.1 appears to be higher than the one of type 8.3.1. Although this can partly be explained by the preprocessing and plotting options used, the chosen plots demonstrate one of the challenges of visual similarity analysis—the impression of similarity/dissimilarity is often a relative one, and it depends on the number of time series available. For type 9.5.1, very similar (i.e. similarity of all visible features) time series exist making them easy to group. For type 8.3.1, time series with very similar features exist—e.g. the apparent upper boundness, maxima flashier than minima, overall "noisy" appearance, no strong interannual variations—yet those features are somewhat different in each of the given examples. There is ultimately no way to express degree of similarity clearly or to describe the nature of similarity in visual analysis; it remains subjective and unstructured. To overcome this, further moves towards quantitative analyses have to be made.

Figure 7 (upper two rows) shows examples of time series plots that were classified as being "irregular" and thus excluded from classification. The lower row of Fig. 7 shows examples of time series which are regular, but classified as "not fitting anywhere"—there is no other time series that resembles each of them. As mentioned before, irregularities in time series are of significance, both as "disturbing features" that introduce errors and uncertainty, and as additional information on groundwater systems and system responses—whatever creates strange/irregular and exceptional features may inform us about the influence of properties and processes or changes of these. Much may be learned from analyzing those more closely. It is important to point out that smaller or larger irregularities occur in almost any groundwater hydrograph. Without being able to prove this, the authors claim that many irregularities can be identified much better through visual inspection than by formalized automated approaches. Please note, that the term "noise" in relation to irregularities is avoided here, as identifying signals as noise may imply that they are not the results of the system responses one is interested in. The irregular features, however, may very well be "real" signals, yet created by an unknown and unusual process.

The visual classification of the Swedish dataset was carried out independently (i.e. each observer used their own approach), and the results were compared. The details of this comparison do not add much to the overall conclusions other than confirming the subjectivity of the approach. On a "large scale", i.e. at the group level, results concur largely, while overlaps are fewer on the subgroup and type levels. In that sense, visual classification does not differ from automatic procedure, where the fine-grained classification results depend on, e.g. the chosen distance measures or clustering algorithms. Figure 8 shows examples from one subgroup created in the joint classification of Swedish and German data. Similarity between Swedish and German data is usually quite high at the subgroup level, but less strong at the type level. One main reason is that the Swedish dataset was created from bi-weekly measurements, usually leading to a smoother appearance. The second reason is that in many of the very shallow and thin aquifers in Sweden, groundwater levels frequently drop below the well bottom or reach the ground surface (wetlands). This creates an appearance of upper and lower bounds. In general, the diversity of types in Sweden is much lower, as there is less variety among hydrogeological settings—for example, hydrographs belonging in some of the subgroups of group 1, associated with deep confined aquifers, are missing in the Swedish dataset.

"Fuzzy" classification

The results presented so far represent an "exclusive" hierarchical classification—each time series belongs to one type only. However, results remain unsatisfactory—there are many cases where one time series fits in several types, depending on which of its visual features is weighted highest. This leads to an ambiguity of classifications, which can be explained by the mixture of patterns as demonstrated in

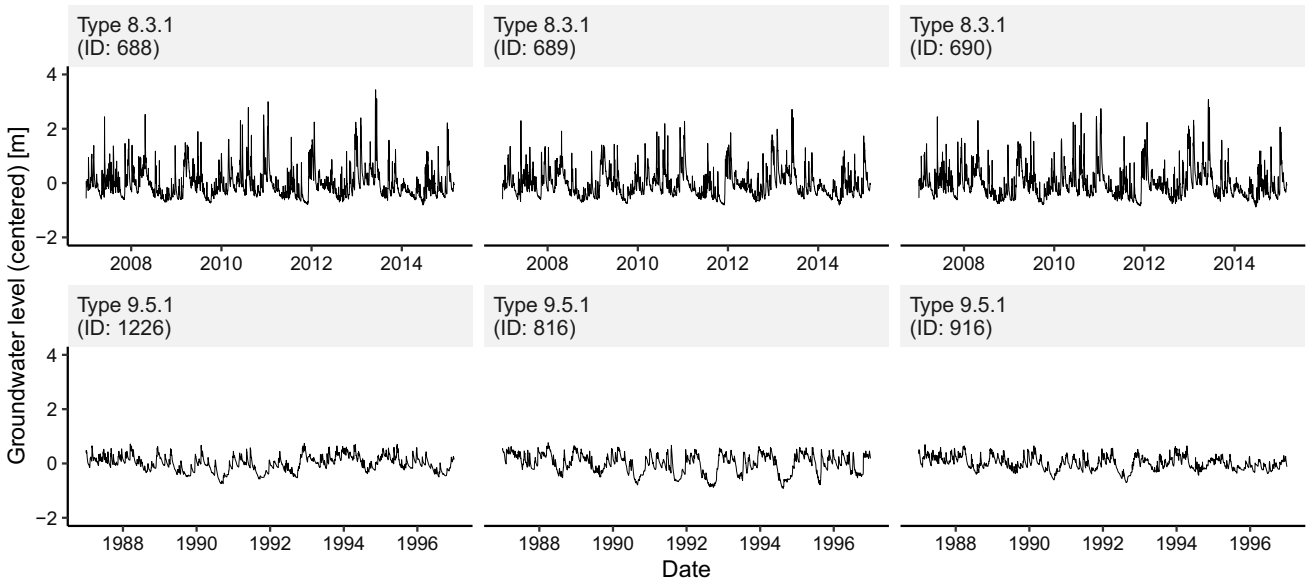


Fig. 6 Examples of plots of time series from two types

Fig. 4. In-between-type similarities across groups and subgroups exist. To explore this further, the between-similarities and within-similarities of types, subgroups and groups were determined visually using a similarity matrix (section 'Visual comparison and classification of time series'). As the results of this supplementary approach are not used to support the conclusions of this paper, they are, together with further explanations, shown in the ESM.

Summary of findings

Some of the findings of the visual classifications are summarized thus:

- As expected, the results of classification may differ, sometimes significantly, depending on preprocessing and plotting decisions (see Fig. 3, for example). This is particularly the case for time series with a high interannual and low intraannual fluctuation or overall range.
- Classification time series with a large diversity of superimposed differing patterns (see Fig. 4 for examples) is extremely difficult. They may be assigned to different types, subgroups or groups, depending on which pattern the observer gives the highest weight to.
- A strictly hierarchical and exclusive classification is not meaningful. This should be kept in mind when applying automated similarity analysis approaches, in particular those described by Haaf and Barthel (2018).
- Two different observers will produce different results. The more strategic decisions they can agree on (plotting, preprocessing, weighting of features and patterns), the more alike the results will be. However, large differences are still to be expected, as a comparison of visual classification of the Swedish dataset carried out by two observers revealed.
- Wishful thinking is a problem. The temptation to confirm preexisting conceptual ideas and hypotheses ("shallow-unconfined: must be a flashy type, etc.") is strong.
- The Swedish and German time series are very similar in general. Most types in the joint classification contain time series from both datasets, yet there are a number of types that seem to be specific to either of the regions. There are technical issues to be considered, however, as the Swedish time series are measured bi-weekly, giving them a smoother appearance.
- The majority of time series contain sections that look peculiar or irregular in one way or another, deviating more or less from their general pattern. These can be small spikes and shifts, longer-lasting periods of exceptionally high or low measurements, changes of fluctuation frequency, as well as exceptionally noisy or exceptionally smooth parts. This aspect should be very carefully considered when developing and using automated similarity analysis and should be the main motivation to carry out systematic visual analysis of (groundwater) time series data overall. The human eye can, to some degree, "ignore" sections that are different from the "usual" appearance of a time series. Automated procedures, if not applied using moving windows, will just mix them in.

Readers may be interested to learn how much time is needed to perform a meaningful visual inspection and classification of a dataset of around 1,000 hydrographs. Based

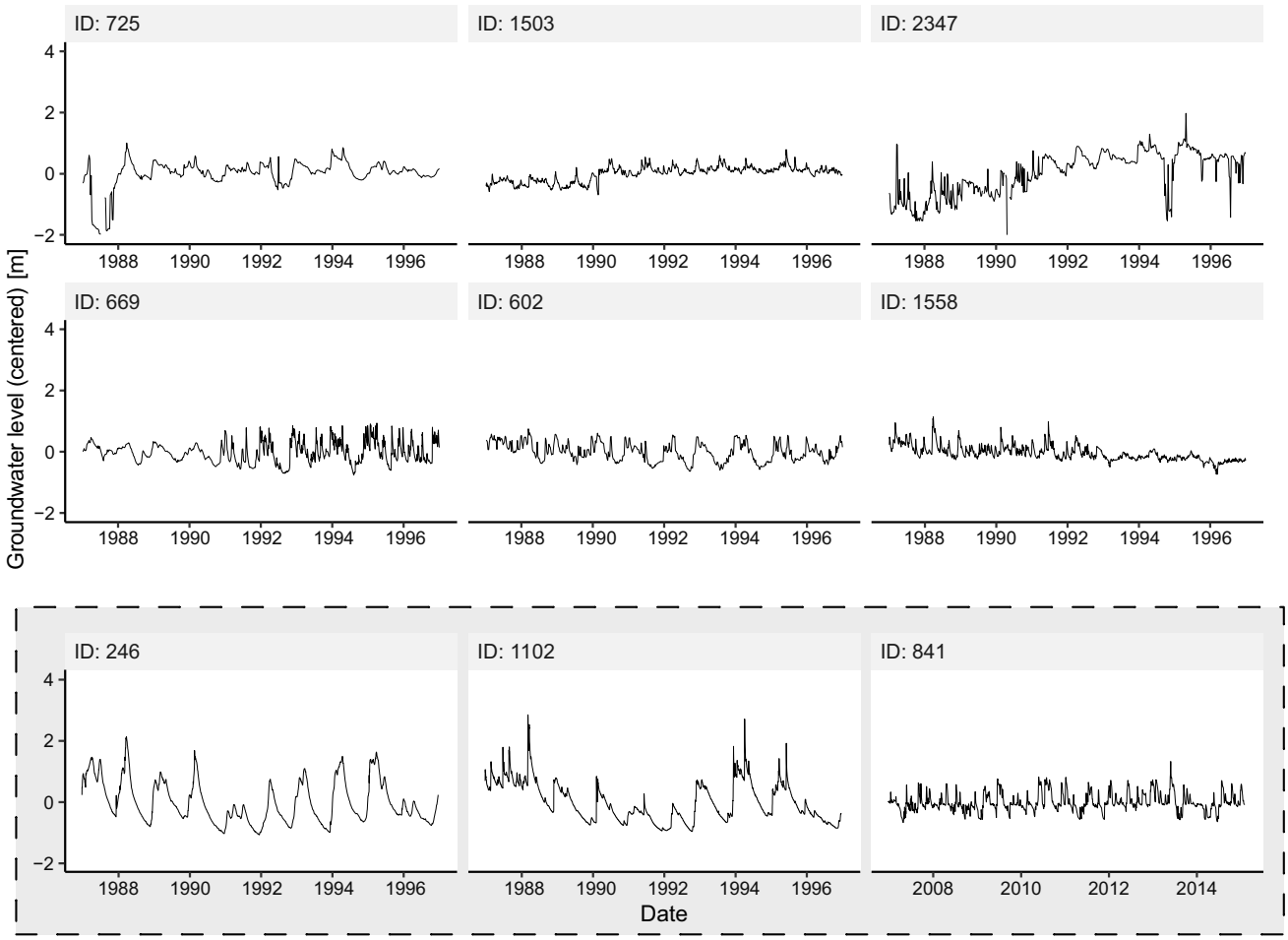


Fig. 7 Upper two rows: examples of plots of time series that were classified as being "irregular". Keep in mind that the main principle of the approach is to characterize hydrographs based on overall visual appearance only; therefore, similarity or irregularities are not explained. It is thus also left to the reader to identify what may be irregular here. Lower row: examples of time series which are regular (and seem normal), but classified as "not fitting anywhere" as they do not resemble any other time series

on the study presented here, this is difficult to answer; to carry out the study and to arrive at the final classification presented, it took several months. However, much of this time was not spent on the actual inspection and classification but on testing different approaches and ideas and the comparison with other approaches. Also, due to the lack of hard and objective criteria, it is hard to determine when one is done, i.e. when a satisfying result is achieved. Nevertheless, to provide some ideas, a rough estimate for a set of 1,000 time series, assuming that preprocessing and plotting decisions have already been made, would be that a reasonable classification at the group level can be made within several hours. From there, a refinement to subgroup level will require about 1 or 2 more days. Type level classification can be achieved in an additional week. A fuzzy classification, involving a type-to-type comparison will require another 2–3 days. For a set of 100 time series, everything, including fuzzy classification could be done in 1 day.

Comparison of visual classification with other data

Comparison with index-based classification indices

To compare the results of visual classification to index-based classification, Fig. 9 shows the averaged within-subgroup values of nine selected indices. It can be seen that index value distributions show a distinct pattern for most groups and subgroups. Within groups, most indices usually have the same sign and magnitude, while one or two indices can differ significantly between subgroups of one group. There are, however, also exceptions—based on their index values, subgroups 1.4 and 1.5 are much closer to the subgroups in group 2 then they are to the other subgroups in group 1. It would definitely be interesting to look into this more, yet it is outside the scope of this article to discuss this further.

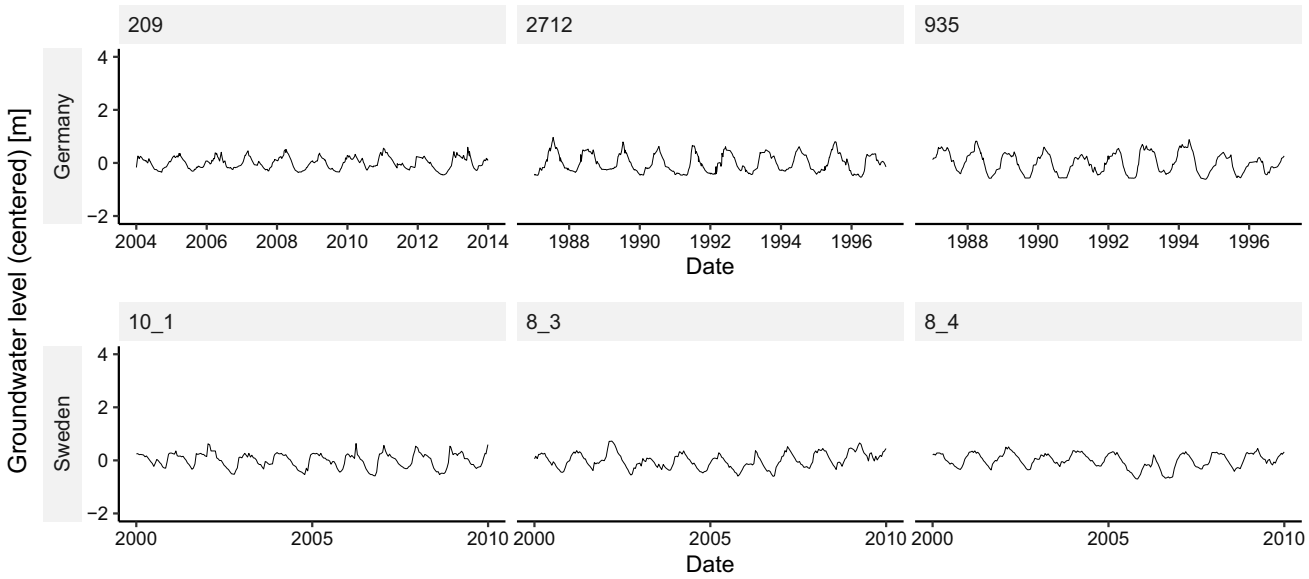


Fig. 8 Time series examples from type 4.1.3 created from a joint dataset of Germany and Swedish data. The upper row shows examples from Germany, the lower from Sweden. Please note that the German data were measured daily, the Swedish biweekly, leading to a smoother appearance

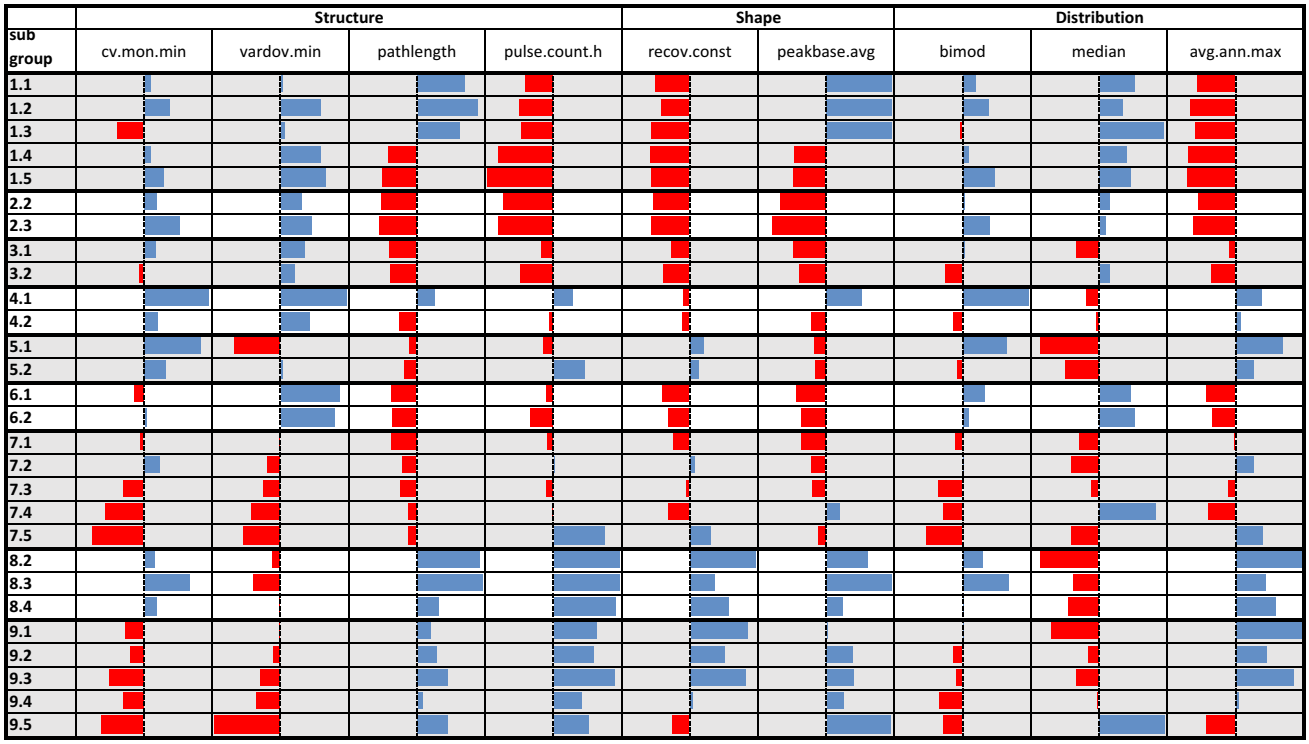


Fig. 9 Comparison of the subgroups defined through visual classification with the averaged within-subgroup values of nine selected indices (see Table 1). Index values are standardized (z-scores)

In many cases, index values can be easily explained by the visual appearance (and vice versa)—for example, the pulse.count.h index is low for the rather smooth time series that is dominated by long wavelengths in groups 1, 2 and 3, but has higher values for the rather flashy groups 8 and 9 (compare also with Fig. 5). The pulse.count.h index value for subgroup 9.5 is, on average, 0.67, while the same index value is 1.63 for subgroup 8.3, clearly reflected by the visual appearance of the given examples from these subgroups.

To demonstrate how sets of index values for individual time series relate to the actual visual impression of a time series, Fig. 10 shows the index values of four distinct time series taken from pairs of very close types along with selected plots of those types. The subgroups those types belong to were chosen as they demonstrate strong between-subgroup similarity (2.2 and 2.3, 8.2 and 8.3, respectively), or strong between-subgroup dissimilarity (subgroups from group 2 versus group 8). According to the index profiles, times series ID1550 and ID1369 seem to be very similar, while the visual appearance suggests some differences (thus two different subgroups). ID1825 and 1325, grouped into the same subgroup but different types therein, are also quite similar index-wise (apart from index pulse.count.h). Visually, the different appearance is mainly due to a smoother lower bound in 1325.

Comparison with hydrogeological data

Figure 11 shows the averaged, normalized (z-scores) values of six selected, commonly available observation well properties (descriptors) for all subgroups, determined through visual classification. All descriptors are related to thickness and depth, apart from elevation. For more information on the descriptors used in this study, please refer to Giese et al. (2020) and Haaf et al. (2020).

More comparisons between visual classification results and hydrogeological settings are shown in Barthel et al. (2021). Figure 11 indicates that there is a strong, yet not unambiguous relationship between hydrogeological properties and types of groundwater dynamics as determined with visual classification. A deeper analysis of the relationships shown in Fig. 11 would increase the understanding of groundwater system responses to change, but this also is outside the scope of this article. The authors would once again like to draw the reader's attention to the other studies published by the authors within the framework of this research, as well as the PhD dissertations presented by Haaf (2020) and Heudorfer (2019).

Discussion

The objective of this article was to discuss the benefits and challenges of visual inspection and classification. This was done using a visual classification scheme for a large set of groundwater hydrographs. It was evaluated as to how well this scheme matched the results of automated (formalized) classification, and whether the assigned classification had a meaningful relationship to hydrogeological conditions. It was found that, using a thorough, systematic classification based on visual appearance, a fine-grained hierarchal

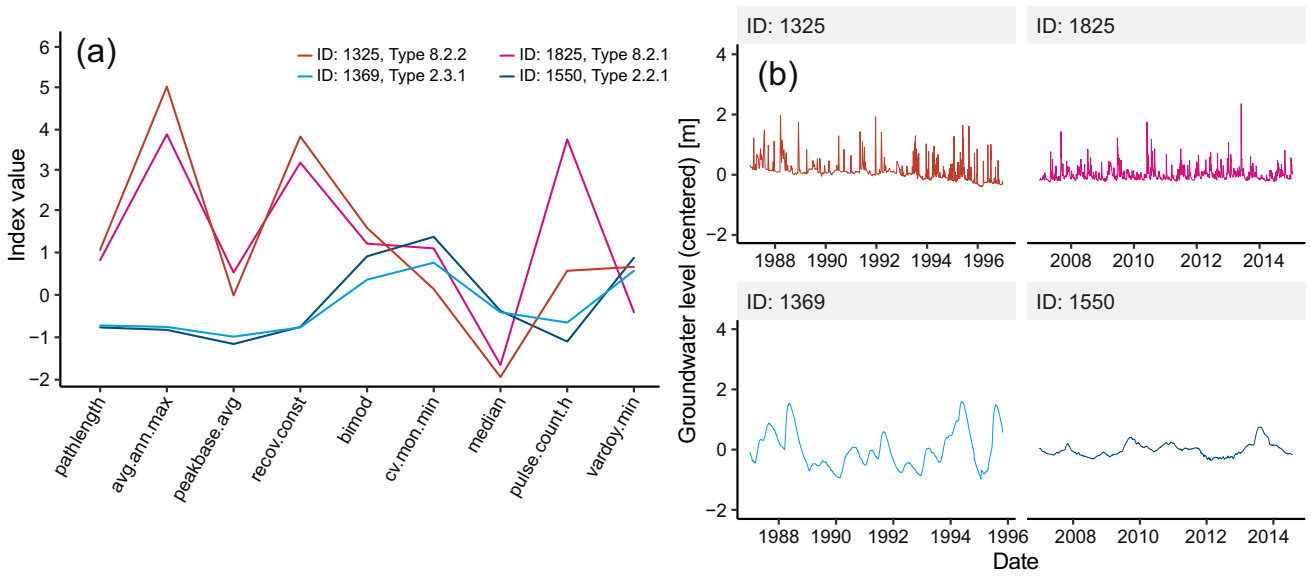


Fig. 10 **a** The averaged within-subgroup values for nine selected indices and four selected types. **b** Specimen hydrographs for the four selected types

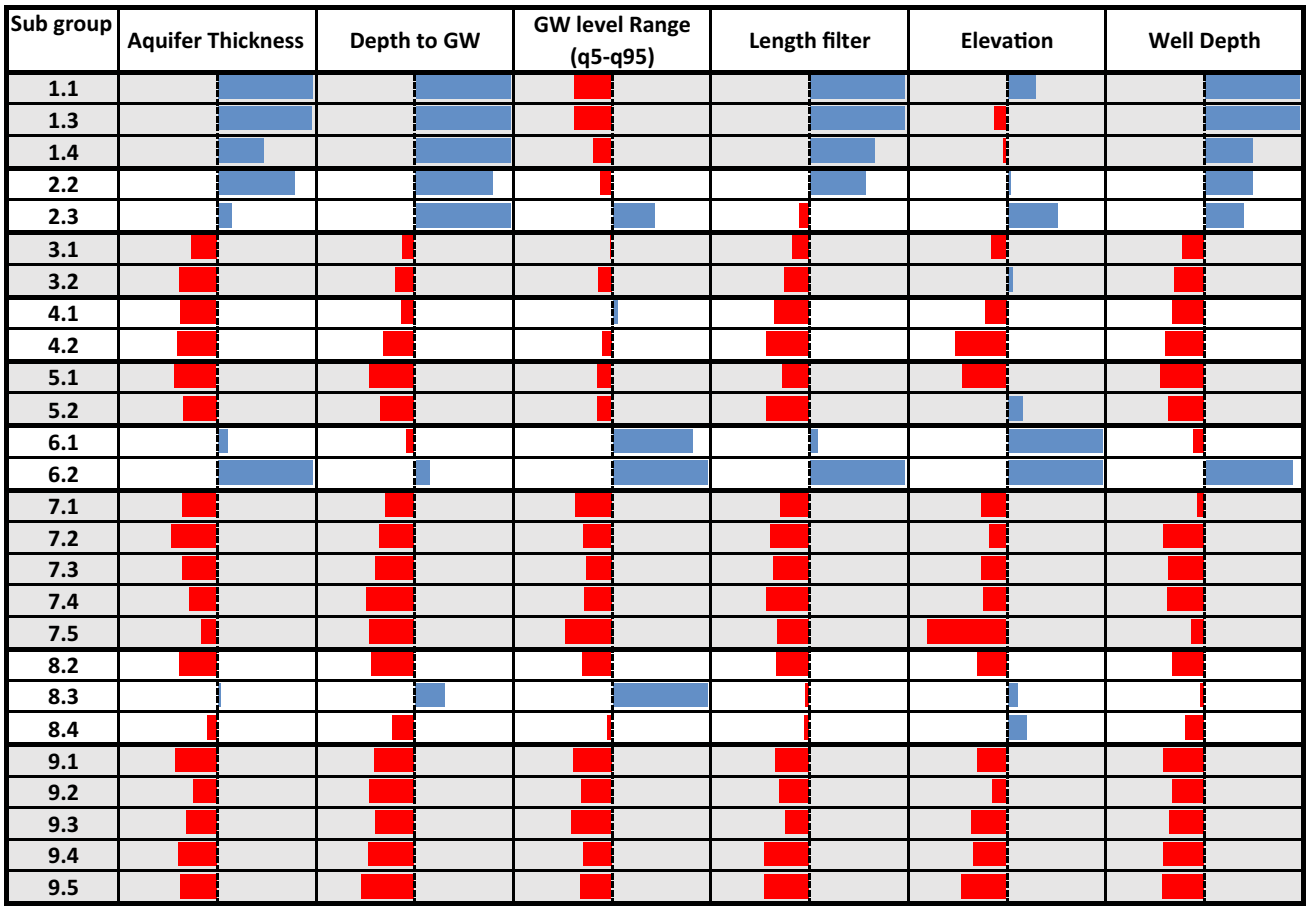


Fig. 11 Averaged, normalized (z-scores) values of six selected observation well parameters (descriptors) for all subgroups determined through visual classification. Note that for some subgroups, there were either no or not enough complete records of descriptors to do this analysis

classification of groundwater dynamics could be achieved that matched the results of automated, index-based classification quite well. The index-based time series classification scheme, through a variety of different indices and a clear topology, makes different dimensions of similarity transparent and accessible, and allows for a straightforward comparison to visual classification. The comparison between the two showed a good match—for most types of time series established through visual classification, the individual members showed quite similar values for a group of representative indices. The indices were also quite consistent with the hierarchal scheme of the groups, subgroups and types. Indices differed significantly for groups that were visually dissimilar and were close for groups that were similar. There were, however, also cases where visual and index-based classifications did not match very well, and time series where neither classification method delivered convincing results.

Not unexpectedly, it was found that the classifications based on visual appearance could be linked to hydrogeological conditions quite well. The relationships are not always straightforward—in some cases, similar time series are from different hydrogeological settings or a similar hydrogeological setting creates dissimilar time series. Yet, such a scheme provides plenty of possibilities for improving knowledge and understanding of groundwater systems. It can be used to carry out systematic cross-checks, and to find and improve explanations of groundwater systems behaviour, based on a condensed and structured dataset.

Overall, it can be concluded that visual classification is a valid approach and could be used to make predictions of the dynamic behaviour of distinct hydrogeological settings. However, the more interesting question, in the context of this article, is whether systematic visual analysis is necessary and worth such a big effort. Why not use the more systematic, formal and automated, and objective, procedures for time series characterization and classification from the beginning instead of the tedious and rather subjective visual approach?

This question is unfortunately difficult to answer systematically and quantitatively. One of the problems is that, despite demonstrating that visual and index-based classifications deliver similar results, it is still unknown which

is more useful. It is known from other studies (Haaf et al. 2020; Heudorfer et al. 2019; Giese et al. 2020) that index-based classification does not deliver definitive classification results. It was also shown that there is no absolute way to determine the performance of classification approaches (Haaf and Barthel 2018); therefore, it is difficult to decide, based on the comparison presented in this article, whether visual analysis performs much worse (making it unnecessary) or much better (making it mandatory) than index-based classification. However, even if it cannot be directly proven through the results presented in this article, it seems likely that visual analysis may significantly help to improve automated schemes, in particular the index-based classification. Nevertheless, there seems to be aspects of visually apparent dissimilarity that indices cannot capture (suggesting development of new indices or improvement of existing ones). Table 2 shows a comparison of the advantages, disadvantages and limitations found in the three different classification approaches applied in the wider framework of this research. The direct comparison approach by Haaf and Barthel (2018) is included.

It is the authors' opinion that the advantage of visual classification is the ability to handle time series data that are inhomogeneous in terms of measurement intervals, start and end date, gaps and irregularities—all features typical of groundwater hydrographs (Collenteur 2021; Peterson et al. 2017). Handling in this sense, not only means foremost detecting and characterizing irregularities, but also using data which otherwise is too poor for numerical evaluation. Of the raw data available for this study, only a relatively small fraction could be used with the formal automated procedures; and to distinguish suitable from nonsuitable time series, visual inspection had to be used. Another important advantage of visual classification is detecting new, yet unknown patterns in the data, and to detect situations where patterns are overprinted with others (see Fig. 4); some further examples are given in Barthel et al. (2021).

On the other hand, the disadvantages of visual classification are significant: it is tedious, subjective and not reproducible. Additionally, it lacks explanatory power and is highly reliant on choices made in preprocessing and plotting. Full-scale, systematic visual classification carried out on a large dataset of more than 1,000 time series does not seem to be appropriate, while the authors regard systematic visual analysis of data sets with less than 100 times series as almost mandatory, data sets of 100–1,000 time series can very well be considered for a systematic analysis, but it is recommended to start with subsets first.

Finally, the authors would like to highlight one discussion point related to time series classification in general—time series of groundwater levels are the result of many complicated and interrelated conditions and processes. The resulting response patterns are thus manifold across many different features, patterns and aspects. Degree of similarity can vary over time and/or be affected by unusual situations (drought, human interference). In visual classification,

Table 2 Comparison between different classification (similarity detection) approaches

Approach:	Visual classification	Index-based classification[a]	Direct comparison of time series[b]
Results intuitive, transparent	Partly	Yes	Partly
Allows for process-based explanations	Only using expert knowledge	Partly	No
Tolerance to irregularities and gaps in time series	High	Medium to low	Very low
Regular intervals between measurements required	Tolerant to a certain degree	Mostly yes	Yes
Same length of time series required	No	Partly	Yes
Time series must have same start and end date	No	No	Yes
Preprocessing effort required	Low	Medium	Medium
Influence of preprocessing on results	Very high	Low	Low
Influence of plot layout on results	High	N/A	N/A
Chances of spotting known unusual time series behaviour	High	Medium	Medium
Chances of spotting previously unknown unusual time series behaviour/patterns	High	Medium to low	Very low
Reproducibility	Very low	High	High
Max number of time series	~1,000	Almost unlimited	High
Time required	Very high	Low	Low
Transferability	Medium to low	High	High
Subjectivity	Very high	Low	Low

[a]Heudorfer et al. (2019)

[b]Haaf and Barthel (2018)

it depends on which patterns the viewer perceives as most important, and which patterns the chosen plotting option emphasizes most. It is impossible to know if those patterns also have a strong hydrogeological relevance. This problem is not unique for visual classification though—it applies to automated algorithm-based procedures as well. As long as the mechanisms and properties that lead to a certain dynamic behaviour of groundwater levels are not fully understood, it will neither be possible to determine the optimum classification scheme, nor to use similarity and classification in groundwater for robust predictions.

Conclusions

A systematic visual inspection and similarity analysis of a data set of more than 1,100 time series was performed. This was not done as a standalone approach but to support other approaches within a much larger framework of research (Barthel et al. 2021). The aim of this study on visual classification was thus not to prove the validity of a concept, or to quantitatively prove its superiority or inferiority in comparison to other approaches, but to evaluate how it can most beneficially be used as a support for other methods. The main conclusions from this study are:

- Visual classification (i.e. classification of time series based on perceived similarity), if carried out systematically and after thorough consideration of preprocessing and plotting options, is an excellent tool to identify patterns, irregularities and peculiar features that can be used in various ways to develop, cross-check and enhance other, automated, approaches.
- Visual inspection is particularly powerful where data are poor and heterogeneous, which is often the case for groundwater time series.
- Visual classification, despite the advantages listed in the preceding, cannot form a standard approach to time series classification because of its obvious disadvantages: subjectivity, tediousness, low reproducibility and reusability. It is only justified, and in particular in groundwater studies even mandatory until greater understanding has been reached, in the early stages of an analysis as a tool to improve automated classification procedures.
- Even in the early stages of a classification study, an effort as large as the one carried out for the purpose of this article (>1,000 time series) is hardly justified. It is recommended to create a classification for a smaller subset of the dataset (100–200 time series) and to carry out random cross-checks with the rest of the dataset. Irregular features and particularities can be detected with far less systematic approaches.
- For groundwater data, the focus should be on identifying and interpreting irregularities in time series, as those are plentiful and have interesting characteristics. With an inventory and characterization of the found irregularities, automated procedures to detect and understand irregularities can be better targeted and are less error-prone. The authors found that noise and outliers in groundwater time series are easily misinterpreted by automated procedures. Having a clear idea of characteristic patterns and the mode of their occurrences can help to improve this.
- In this study, time series classification was made based on "overall visual appearance only" (see section 'Time series classification according to similarity'). An approach using "visual pattern recognition" (see Fig. 4) may be even more difficult and tedious to accomplish, but may yield results which are more straightforward to use to improve automated similarity analysis.

To summarize, systematic visual inspection and classification is a highly valuable tool, but it must not be overrated. In the context of classification and similarity in groundwater data, it will play a significant role until the optimal approaches for groundwater hydrograph similarity and classification are found, all the important features characterizing times series are known and understood, and all the disturbing features can be identified, eliminated and dealt with using automated procedures. As this ideal situation is still a long way off, visual analysis will, for the foreseeable future, have a role to play.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10040-021-02433-w.

Acknowledgements Data for the studies presented in this article were provided by the Bavarian State Office of the Environment (Bayerisches Landesamt für Umwelt, www.lfu.bayern.de), the State Office for Environment, Measurements and Nature Conservation in Baden-Württemberg, Germany (LUBW Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg, Grundwasserdatenbank), the Federal Ministry for Agriculture, Forestry, Environment and Water Management in Austria and the Geological Survey of Sweden.

Funding Open access funding provided by University of Gothenburg. This work was supported by the Swedish Research Council Formas under Grant number 2016-00513.

Data availability The majority of data was hereby provided under the condition that it is not distributed to third parties and that, when printed or presented in other ways, the exact location of the observations is not shown. The raw data used in this study can thus unfortunately not be made available. However, plots of all hydrographs used are contained in the ESM (Barthel et al. 2020).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Allen DM, Whitfield PH, Werner A (2010) Groundwater level responses in temperate mountainous terrain: regime classification, and linkages to climate and streamflow. Hydrol Processes 24:3392–3412. https://doi.org/10.1002/Hyp.7757

Asch K (2007) Europe's geology on-line: the IGME 5000. 5th European Congress on Regional Geoscientific Cartography and Information System. Rev Catalana Geograf XII(29). http://www.rcg.cat/articles.php?id=88. Accessed Dec 2021

Baker DB, Richards RP, Loftus TT, Kramer JW (2004) A new flashiness index: characteristics and applications to Midwestern rivers and streams. J Am Water Resour Assoc 40:503–522. https://doi.org/10.1111/j.1752-1688.2004.tb01046.x

Barthel R (2006) Common problematic aspects of coupling hydrological models with groundwater flow models on the river catchment scale. Adv Geosci 9:63–71

Barthel R (2014) HESS Opinions "Integration of groundwater and surface water research: an interdisciplinary problem?" Hydrol Earth Syst Sci 18:2615–2628. https://doi.org/10.5194/hess-18-2615-2014

Barthel R, Rojanschi V, Wolf J, Braun J (2005) Large-scale water resources management within the framework of GLOWA-Danube, part A: the groundwater model. Phys Chem Earth 30:372–382. https://doi.org/10.1016/j.pce.2005.06.003

Barthel R, Jagelke J, Götzinger J, Gaiser T, Printz A (2008) Aspects of choosing appropriate concepts for modelling groundwater resources in regional integrated water resources management: examples from the Neckar (Germany) and Ouémé catchment (Benin). Phys Chem Earth 33:92–114. https://doi.org/10.1016/j.pce.2007.04.013

Barthel R, Reichenau TG, Krimly T, Dabbert S, Schneider K, Mauser W (2012) Integrated modeling of global change impacts on agriculture and groundwater resources. Water Resour Manage 26:1929–1951. https://doi.org/10.1007/s11269-012-0001-9

Barthel R, Seidl R, Nickel D, Buttner H (2016) Global change impacts on the Upper Danube Catchment (Central Europe): a study of participatory modeling. Reg Environ Change 16:1595–1611. https://doi.org/10.1007/s10113-015-0895-x

Barthel R, Haaf E, Giese M, Nygren M (2020) Visual classification results. https://doi.org/10.6084/m9.figshare.13281395.v1

Barthel R, Haaf E, Giese M, Nygren M, Heudorfer B, Stahl K (2021) Similarity-based approaches in hydrogeology: proposal of a new concept for data-scarce groundwater resource characterization and prediction. Hydrogeol J. https://doi.org/10.1007/s10040-021-02358-4

Chiew FHS, McMahon TA (1993) Assessing the adequacy of catchment streamflow yield estimates. Soil Res 31:665–680. https://doi.org/10.1071/sr9930665

Collenteur RA (2021) How good is your model fit? Weighted goodness-of-fit metrics for irregular time series. Ground Water 59:474–478. https://doi.org/10.1111/gwat.13111

Crochemore L, Perrin C, Andréassian V, Ehret U, Seibert SP, Grimaldi S, Gupta H, Paturel JE (2015) Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrol Sci J 60:402–423. https://doi.org/10.1080/02626667.2014.903331

Ehret U, Zehe E (2011) Series distance: an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrol Earth Syst Sci 15:877–896. https://doi.org/10.5194/hess-15-877-2011

Ewen J (2011) Hydrograph matching method for measuring model performance. J Hydrol 408:178–187. https://doi.org/10.1016/j.jhydrol.2011.07.038

Gaiser T, Printz A, von Raumer HGS, Götzinger J, Dukhovny VA, Barthel R, Sorokin A, Tuchin A, Kiourtsidis C, Ganoulis I, Stahr K (2008) Development of a regional model for integrated management of water resources at the basin scale. Phys Chem Earth, Parts A/B/C 33:175–182. https://doi.org/10.1016/j.pce.2007.04.018

Giese M, Haaf E, Heudorfer B, Barthel R (2020) Comparative hydrogeology: reference analysis of groundwater dynamics from neighbouring observation wells. Hydrol Sci J. https://doi.org/10.1080/02626667.2020.1762888

Gleicher M, Albers D, Walker R, Jusufi I, Hansen CD, Roberts JC (2011) Visual comparison for information visualization. Inform Visual 10:289–309. https://doi.org/10.1177/1473871611416549

Gogolou A, Tsandilas T, Palpanas T, Bezerianos A (2019) Comparing similarity perception in time series visualizations. IEEE Trans Visual Comput Graphics 25:523–533. https://doi.org/10.1109/tvcg.2018.2865077

Guzha AC, Hardy TB (2009) Application of the distributed hydrological model, TOPNET, to the Big Darby Creek Watershed, Ohio, USA. Water Resour Manage 24:979–1003. https://doi.org/10.1007/s11269-009-9482-6

Haaf E (2020) Towards prediction in ungauged aquifers: methods for comparative regional analysis. PhD Thesis, University of Gothenburg, Germany

Haaf E, Barthel R (2018) An inter-comparison of similarity-based methods for organisation and classification of groundwater hydrographs. J Hydrol 559:222–237. https://doi.org/10.1016/j.jhydrol.2018.02.035

Haaf E, Giese M, Heudorfer B, Stahl K, Barthel R (2020) Physiographic and climatic controls on regional groundwater dynamics. Water Resour Res 56:WRCR24909. https://doi.org/10.1029/2019wr026545

Harrigan S, Murphy C, Hall J, Wilby RL, Sweeney J (2014) Attribution of detected changes in streamflow using multiple working hypotheses. Hydrol Earth Syst Sci 18:1935–1952. https://doi.org/10.5194/hess-18-1935-2014

Harter T, Hopmans JW (2004) Role of vadose-zone flow processes in regional-scale hydrology: review, opportunities and challenges. In: Papers for the Frontis Workshop on Unsaturated-Zone Modeling: Progress, Challenges and Applications, Wageningen, The Netherlands, 3-5 October 2004

Heudorfer B (2019) Groundwater dynamics during drought: an index-based analysis. PhD Thesis, University of Freiburg, Freiburg, Germany

Heudorfer B, Haaf E, Stahl K, Barthel R (2019) Index-based characterization and quantification of groundwater dynamics. Water Resour Res 55:5575–5592. https://doi.org/10.1029/2018wr024418

Hrachowitz M, Savenije HHG, Blöschl G, McDonnell JJ, Sivapalan M, Pomeroy JW, Arheimer B, Blume T, Clark MP, Ehret U, Fenicia F, Freer JE, Gelfan A, Gupta HV, Hughes DA, Hut RW, Montanari A, Pande S, Tetzlaff D, Troch PA, Uhlenbrook S, Wagener T, Winsemius HC, Woods RA, Zehe E, Cudennec C (2013) A decade of predictions in ungauged basins (PUB): a review. Hydrol Sci J 58:1198–1255. https://doi.org/10.1080/02626667.2013.803183

Li S, Gitau M, Bosch D, Engel BA, Zhang L, Du Y (2017) Development of a soil moisture-based distributed hydrologic model for determining hydrologically based critical source areas. Hydrol Processes 31:3543–3557. https://doi.org/10.1002/hyp.11276

Lin J, Li Y (2009) Finding structural similarity in time series data using bag-of-patterns representation. In: Scientific and Statistical Database Management SSDBM 2009, Lecture Notes in Computer Science. 21st International Conference, SSDBM 2009, New Orleans, LA, USA, June 2009

Martens K, Van Camp M, Van Damme D, Walraevens K (2013) Groundwater dynamics converted to a groundwater classification as a tool for nature development programs in the dunes. J Hydrol 499:236–246. https://doi.org/10.1016/j.jhydrol.2013.06.045

Mauser W, Prasch M (2016) Regional assessment of global change impacts: The Project GLOWA-Danube. Springer, Heidelberg, Germany

Nickel D, Barthel R, Braun J (2005) Large-scale water resources management within the framework of GLOWA-Danube: the water supply model. Phys Chem Earth 30:383–388. https://doi.org/10.1016/j.pce.2005.06.004

Nygren M, Giese M, Kløve B, Haaf E, Rossi PM, Barthel R (2020) Changes in seasonality of groundwater level fluctuations in a temperate-cold climate transition zone. J Hydrol X 8:100062. https://doi.org/10.1016/j.hydroa.2020.100062

Peterson TJ, Western AW, Cheng X (2017) The good, the bad and the outliers: automated detection of errors and outliers from groundwater hydrographs. Hydrogeol J 26:371–380. https://doi.org/10.1007/s10040-017-1660-7

Rinderer M, McGlynn BL, van Meerveld HJ (2017) Groundwater similarity across a watershed derived from time-warped and flow-corrected time series. Water Resour Res 53:3921–3940. https://doi.org/10.1002/2016wr019856

Rinderer M, Meerveld HJ, McGlynn BL (2019) From points to patterns: using groundwater time series clustering to investigate subsurface hydrological connectivity and runoff source area dynamics. Water Resour Res. https://doi.org/10.1029/2018wr023886

Römer T, van Heyden J, Barthel R (2016) Data on quantity and quality of groundwater. In: Mauser W, Prasch M (eds) Regional assessment of global change impacts: The Project GLOWA-Danube. Springer, Heidelberg, Germany, pp 177–184

Seibert SP, Ehret U, Zehe E (2016) Disentangling timing and amplitude errors in streamflow simulations. Hydrol Earth Syst Sci 20:3745–3763. https://doi.org/10.5194/hess-20-3745-2016

Wagener T, Sivapalan M, Troch P, Woods R (2007) Catchment classification and hydrologic similarity. Geogr Compass 1(4):901–931

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal 01:1–41. https://doi.org/10.1142/s1793536909000047

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.