

# Trend Assessment for Groundwater Pollutants: A Brief Review and Some Remarks

Francesc Oliva, Esteban Vegas, Sergi Civit, Teresa Garrido, Josep Fraile, and Antoni Munné

**Abstract** Groundwater is a valuable natural resource that needs to be assessed and protected. The European Union (EU) adopted new water legislation that includes the Water Framework Directive (WFD) and the Groundwater Daughter Directive (GWD). Both require the identification of sustained increasing pollution *trends* and their reversal. This is the second pillar of the WFD: such trends have to be identified for any pollutants that result in groundwater being characterized as at risk of not meeting the environmental objectives. Measuring these trends is necessary to determine and understand whether changes in land use, fertilizer application, pollution history, or climate change are affecting groundwater quality. However, in many cases, groundwater data series may not meet minimum requirements for classical statistical procedures employed in trend assessment: among other obstacles, data may be sparse, with missing or extreme values, censored data, seasonal effects, and autocorrelation. The aim of this chapter is to present and review several statistical methodologies that have been proposed and applied in recent years to deal with groundwater trend assessment, discussing the relative advantages and disadvantages of each one.

**Keywords** Catalonia, Daughter Directive, Groundwater, Monitoring, Quality trend assessment, Water Framework Directive

---

F. Oliva (✉), E. Vegas, and S. Civit

Department of Statistics - Faculty of Biology, University of Barcelona, Diagonal 643, 08028 Barcelona, Spain

e-mail: [foliva@ub.edu](mailto:foliva@ub.edu)

T. Garrido, J. Fraile, and A. Munné

Catalan Water Agency, Provença, 204-208, 08036 Barcelona, Spain

## Contents

1	Introduction .....	26
2	An Overview of Methodologies .....	28
2.1	Individual Monitoring Points (Stations) .....	28
2.2	Overall Assessment of GWBs .....	39
3	Assayed Procedures: A Case Study .....	42
3.1	Methods Assayed for Individual Monitoring Points .....	43
3.2	Methods Assayed for Overall GWB Trend Assessment .....	45
4	Results and Discussion .....	46
4.1	Individual Monitoring Points .....	46
4.2	Overall GWB Trend Assessment .....	55
5	Conclusions and Future Trends .....	57
	References .....	57

## 1 Introduction

Groundwater is a valuable natural resource that accounts for over 97% of all the freshwater available on Earth, excluding glaciers and ice caps (extracted from Groundwater Protection in Europe, [1]). It needs to be monitored and protected from chemical and organic pollutants, not only because groundwater is used as drinking water but also because it is an important resource for industry and agriculture and has recreational uses and environmental value [2]. Due to the fact that groundwater moves through the subsurface slowly, the impact of human activities may last for decades, and pollution events that occurred in previous years will probably continue to threaten us for several generations.

The European Union (EU) adopted new water legislation that includes the Water Framework Directive (WFD) (2000/60/EC) [3] and the Groundwater Daughter Directive (GWD) (2006/118/EC) [4]. The WFD is a regulatory framework for the protection of all natural waters which prescribes *environmental objectives* to be achieved by the end of 2015 and contemplates extensions to 2021 or 2027. The GWD for the protection of groundwater against pollution defines the specific environmental objectives of the WFD. It requires that threshold values be established (by the end of 2008) for pollutants related to the pressures identified as putting bodies of groundwater at risk (GWD, Article 3). These threshold values are quality standards, and so they are to be used to assess the chemical status of groundwater. The GWD also introduces measures to prevent or limit the introduction of pollutants into groundwater (GWD, Article 6).

Both the WFD (Article 17) and the GWD also require that *trends* must be identified for pollutants that characterize groundwater as being at risk of not meeting the environmental objectives of the WFD (WFD, Annex V 2.4.4; GWD, Article 5). According to the GWD, those trends must be statistically and environmentally significant (GWD, Article 2). The environmental significance relates to the potential future impact of an identified increasing trend in pollution. Such trends should be reversed when they reach 75% of the EU groundwater quality standard

values or specific threshold values (GWD, Article 5). In many countries, the concentrations of groundwater pollutants in aquifers are already approaching or have even exceeded statutory limits for drinking water. For this and other reasons, long-term monitoring of the concentrations of groundwater pollutant by both water utilities and regulatory agencies has become widespread. The achievement of a *good status* and the reversal of *significant and sustained* upward trends in the concentration of pollutants, including nitrates, are environmental objectives set out in Article 4 of the WFD. Trend reversal is to be achieved through the implementation of the Program of Measures (WFD, Annex VI) which aims to progressively reduce pollution and prevent further deterioration of bodies of groundwater (GWD, Article 5). Therefore, the statistical assessment of *trend reversal* is another matter of concern.

The properties that determine the quality of groundwater can vary over different time scales (daily, seasonally, or annually) and depend on the characteristics of the aquifer. In fact, the concentrations of a solute in samples at a single station depend on numerous factors including land use history, groundwater flow, and local groundwater pumping regimes, as well as seasonal and climate effects. Furthermore, the monitoring itself may involve elements that could make it difficult to accurately assess trends, for example, the sampling frequency, amount of missing data, length of monitoring period, and presence of uncontrolled covariate variables.

In many cases, the characteristics of groundwater data series may not meet the minimum requirements for classical statistical procedures employed to analyze trends. Among other obstacles, the data may be sparse and gappy and include censored data, extreme values or *outliers*, seasonal effects, and autocorrelation [5–7]. For EU members, in accordance with the WFD and GWD, two technical reports [8, 9] contain a series of general recommendations for groundwater trend assessment, including monitoring requirements, how to treat censored values, the length of data series, and the statistical methods to be employed. Besides those two excellent guides, a large number of papers have been published in many different journals over the last few decades. However, in our opinion, none of those works manage to definitively determine the procedures to be applied, because a procedure that may be adequate for a certain data series may be cumbersome or inappropriate for another series.

Another concern is related to individual monitoring points and data aggregation. Every groundwater body (GWB) is usually monitored at a certain number of sampling sites (stations). As required by the GWD (Article 5) and WFD (Article 13), member states must summarize the way in which the trend assessment at individual monitoring points within a GWB or group of GWBs has contributed to identifying that those bodies are subject to a significant and sustained upward trend or are experiencing a reversal of such a trend. Furthermore, the WFD (WFD, Annex V Section 2.3) states that in assessing the status of a GWB, the results of individual monitoring points within it are to be aggregated for the body as a whole. So apart from assessing the trend at every station, we have to assess the overall trend in the GWB.

In Catalonia, a groundwater quality monitoring network has been set up in accordance with the requirements of the WFD (WFD, Article 8 and Annex V). Control networks have been designed taking into account the hydrogeological model and the pressures that have been identified on each GWB. Currently, there are some 942 stations distributed across a total of 53 GWBs. Chemical quality is analyzed annually, and quantitative analysis is performed monthly. The average spatial density of the stations in a given GWB is 0.58 stations per 10 km<sup>2</sup>. This network for monitoring groundwater quality dates from 1994 in some strategic aquifers, and the first monitoring program in accordance with the WFD requirements began in 2007.

This chapter mainly focuses on presenting an overview of some of the main statistical techniques that have been proposed in recent years for the statistical testing of a groundwater trend. The methods presented here are rather simple and can easily be implemented using different commercial or open-source statistical software packages. In order to illustrate the procedures and highlight the main results, some of the GWBs and monitoring stations in Catalanian have been selected and the methods applied (see Sect. 3). The pollutants analyzed are nitrates and chlorides. All of the techniques are applied using packages and functions implemented in R [10].

## 2 An Overview of Methodologies

### 2.1 *Individual Monitoring Points (Stations)*

Groundwater monitoring networks provide observation of random variables (concentrations of pollutants) at each sampling site over time. Within the assessment of significant upward trends and trend reversal, we have to consider ([9], Sect. 6.2) the following issues: (a) a correct statistical method for assessing trends at each monitoring point, (b) how to deal with values below the limit of quantification (LOQ), (c) the appropriate length of time series, (d) how to establish baseline levels of substances, (e) what an acceptable level of confidence is in trend assessment, (f) how to establish a starting point for trend reversal, and (g) how to statistically demonstrate that a trend has been reversed.

In accordance with the GWD (Annex IV), to deal with individual concentrations below the LOQ, they are replaced by LOQ/2. Regarding the length of time series for yearly data, Grath et al. [8] recommend at least 8 measurements for the detection of an upward trend and 15 measurements in order to establish whether there is a significant breakpoint (trend reversal). In fact, that research clearly advises against long-term time series if the statistical method does not take into account a possible breakpoint. Nevertheless, we will see that a breakpoint, even within a short-term series, can seriously affect the performance of statistical methods that are recommended for trend assessment.

Reviewing the literature, the problem of detecting and estimating trends in hydrology data has a long history [5, 6, 11–27], and some of the publications report comparative analysis of different techniques. For example, Esterby [18] reviewed some parametric and nonparametric trend detection methods by applying them to water quality time series. Meanwhile, Hess et al. [21] present an overview of six linear methods used to analyze environmental time series. A comprehensive evaluation of 28 statistical methods for checking trend, homogeneity, seasonality, periodicity, and persistence in hydrologic time series was recently published [28]. So far, there is no general consensus regarding which method performs best in a given unknown situation, and few extensive comparisons of the proposed methods have been published. In this scenario, quite a lot of published work advises visual inspections by the user (among others, [29]) in order to decide which method is suitable, whether assumptions are valid, or to finally assess trends. This, however, is quite unrealistic; how many plots would the user have to examine considering all the stations and pollutants? Perhaps it would be thousands, so a robust automatic method (software) for assessing trends is absolutely vital. In this context, TTAinterfaceTrendAnalysis was developed in R package [30] to perform nonparametric trend analysis (Kendall test family) through an interactive GUI. Notwithstanding, other techniques to assess temporal trends in an automatic manner are absolutely crucial.

In this section, we review some of the more common statistical procedures used for trend assessment in groundwater quality data. The methods can be classified into two main general approaches: parametric methods (distribution dependent) and nonparametric methods (distribution free). However, the decision as to which procedures are most useful (to reveal change when it is present and not identify any change when there is no trend) depends on the characteristics of the data: the distribution (normal, skewed, symmetric, heavy-tailed, extreme values); the presence of outliers (exaggerated extreme values, perhaps due to measurement error or a one-off serious contamination event); and its seasonality, whether there are missing values (a few isolated values or large gaps), there is censored data (e.g., due to the LOQ), or there is some serial correlation and if it does contain a monotonic trend or an abrupt change [26]. If the assumptions made in order to apply a statistical test do not hold for the data, then the estimate of the significance level could be incorrect.

Finally, when we carry out a statistical test, it is necessary to define the null hypothesis  $H_0$  (in our case, the hypothesis of no trend) and the alternative hypothesis  $H_1$  (trend). Two types of errors can occur when we perform a test. A *type I error* occurs when the researcher rejects the null hypothesis when it is in fact true (*false positive*). The probability of committing a type I error is called the significance level, and it is frequently denoted as  $\alpha$  (the most common choice is  $\alpha = 5\%$ ). A *type II error* is present when the researcher fails to reject the null hypothesis which is in fact false (*false negative*). The probability of committing a type II error is usually denoted as  $\beta$ , and the complementary probability,  $1 - \beta$ , is known as the *power of the test*. Type I and type II errors are not complementary (i.e., their sum is not one), but they do stand in a relationship: if we decrease the chance of

committing a type I error (choosing a lower  $\alpha$  value), then  $\beta$  increases. The decision rule for rejecting the null hypothesis is in accordance with the  $p$ -value. The  $p$ -value is the probability of observing a test statistic equal to or more extreme than our experimental value, assuming that the null hypothesis is true. If the  $p$ -value is less than the significance level, we reject the null hypothesis.

### 2.1.1 Regression Models

Linear models are the most widely used framework from a parametric perspective. The classical approach to assessing a trend is based on fitting a linear regression (LR) or quadratic regression (QR) model [8]. As is well known, the simple LR model is:

$$x_i = \beta_0 + \beta_1 t_i + e_i \quad i = 1, \dots, n \quad (1)$$

where  $x_i$  is the value for the  $i$ th observation,  $t_i$  is the corresponding value for the independent variable (time),  $\beta_0$  is the intercept,  $\beta_1$  is the slope, and  $e_i$  are the residuals (assumed to be independent and identically distributed). The regression coefficients are estimated using the method of *ordinary least squares* (OLS). A  $t$ -test may be used to test that the true slope is not different from zero ( $H_0 : \beta_1 = 0$ ), which implies that there is no correlation between the pollutant and time ( $H_0 : \rho = 0$ ). So we can conclude that a linear trend exists if the  $p$ -value of this statistic is less than  $\alpha$  (the significance level). Nevertheless, a new assumption is made when we apply the test: the residuals,  $e_i$ , are normally distributed. Moreover, OLS are highly sensitive to outliers. In spite of this, as reported in Grath et al. [8], LR was the most used statistical method for trend assessment in the EU in 2001.

Trends which are nonlinear (say quadratic, exponential, or with an abrupt change) will be poorly described by a linear slope coefficient. We can easily extend the linear model to QR by adding a new term to Eq. 1:

$$x_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + e_i \quad i = 1, \dots, n \quad (2)$$

If the coefficient  $\beta_2$  (quadratic term) is significant, we reject the null hypothesis that LR is acceptable. A quadratic fit implies a concave or convex function (second-order polynomial), and the *inflection point* is at time  $t_{IP} = -\beta_1/(2\beta_2)$ . Nevertheless, the quadratic model has its own limitations: (a) it cannot fit other nonlinear functions; (b) it is not easy to assess the trend before and after the inflection point (recall that a second-order polynomial never remains flat, which means we cannot say “no trend”); (c) the parameter estimates are greatly influenced by outliers; and (d) hypothesis tests are sensitive to departures from normality.

Therefore, it could be useful to apply a robust procedure. The *robust LR* (RLR) model can be stated as:

$$x_i = \beta_0 + \beta_1 t_i + \sigma e_i \quad i = 1, \dots, n \quad (3)$$

where  $\sigma > 0$  is a scale parameter. There are a number of ways to perform robust regression [31–36]. To deal with outliers in the  $x$ -direction, the most commonly used methods are based on *Huber-type estimates* [31, 37], which form a class of  $M$ -estimates. An  $M$ -estimate of  $\beta_1$  is a solution of:

$$\sum_{i=1}^n w_i e_i x_i = 0 \quad (4)$$

where  $e_i = x_i - (\beta_0 + \beta_1 t_i)$  are the residuals and  $w_i = W(e_i/\sigma)$  with  $W(u)$  a suitable weight function. Equation 4 is the equation of a weighted OLS estimate and must be obtained by an algorithm called *iteratively reweighted least squares* (IRLS). In the first iteration, each point is assigned equal weight, and the model coefficients are estimated using OLS. For subsequent iterations, the weights are recomputed so that points farther from model predictions in the previous iteration are given lower weights. The Huber  $M$ -estimator (the default in many software packages) is defined as  $W(u) = \psi_k(u)/u$ , with  $\psi_k(u) = \max(-k, \min(k, u))$ , where  $k$  is constant that the user has to specify. The estimate of  $\sigma$  is usually  $\hat{\sigma} = \text{MAD}$  (*median absolute deviation*). This approach is fairly robust against departures from normality and outliers in the  $x$ -direction. The model presented in Eq. 3 can easily be generalized to perform robust nonlinear regression.

Another approach is *loess* or *lowess*. Cleveland [38–40] proposed the *loess* algorithm as a flexible and robust method to fit a regression function which is suitable when there are outliers and we do not know the parametric model. The name is derived from *locally weighted scatterplot smoothing*, because a weighted least squares method is used to fit linear or quadratic functions at every local predictor point, based on its neighborhood. The *smoothing parameter* controls the fraction of the data contained in each local neighborhood, and data points are weighted by a decreasing function of their distance from the center of the neighborhood. Recommended in Grath et al. [2] as a method that is much more flexible with regard to the shape of a trend, an ANOVA test based on the *loess* smoother is described [2, 41] which allows us to examine both monotonic and non-monotonic trends. In spite of that, we want to mention several pitfalls of the system. (a) The election of the smoothing parameter is not easy, but it is absolutely essential, because it crucially influences the final result. (b) To perform a *loess* fit with short data series is a risky task which can lead to overfitting or underfitting of the data and may boost data autocorrelation. (c) In our experience, the method is not very outlier resistant, so a *loess* fit can become distorted due to the presence of outliers. (d) The ANOVA test proposed considers the *loess* fit as the true function (the error sum of squares is in fact the sum of squared *loess* residuals), so the outlined test is not reliable if we fail to fit the right function. Regarding the selection of a smoothing parameter, Hurvich et al. [42] propose an interesting and automatic determination based on AICc (the corrected Akaike information criterion) [43, 44],

as applied in Wen and Chen [45]. To summarize, for short data series, we prefer the robust regression approach outlined before *loess*; it is less risky and easier to perform, and we do not have to select a smoothing parameter.

Detecting step trends (an abrupt change in the mean level at a specific point in time) in a process is also an important topic. Various statistical methods are available for identifying and locating steps in a time series [46–48]. Piecewise LR is used to detect significant changes in a trend (*breakpoints*), which means there are two different linear relationships in the data with a sudden, sharp change in direction. In this case,  $x_i$  is modeled by splitting the linear predictor into two pieces:

$$x_i = \begin{cases} \beta_0 + \beta_1 t_i + e_i & t_i \leq \gamma \\ \beta_0 + \beta_1 t_i + \beta_2(t_i - \gamma) + e_i & t_i > \gamma \end{cases} \quad (5)$$

where  $\gamma$  is the breakpoint [49]. The slopes of the two lines are  $\beta_1$  and  $\beta_1 + \beta_2$ , so  $\beta_2$  can be interpreted as the difference between the slopes. The model given in Eq. 5, also named *two-section LR* (2SLR), forces continuity at the breakpoint [50]. It is straightforward to see that if  $\beta_2 = 0$ , we are in fact fitting simple LR. As is evident, the 2SLR model can be a useful approach for detecting trend reversals. Unfortunately, it is not robust when facing outliers.

### 2.1.2 Nonparametric Methods

Up to now, the methods we have summarized are based on linear models. Another possible approach is to apply a nonparametric model. Nonparametric methods tend to have been favored in the analysis of large datasets from national monitoring programs [18], since these methods involve fewer assumptions and they are less sensitive to outliers. There has been widespread use of Spearman's *rho* and Mann-Kendall (MK) *tau* statistics (especially the latter) to test for the presence of monotonic trends [5, 6, 51–62] and others.

The MK test is a well-known nonparametric rank-based method. Mann [63] originally used the test, and Kendall [64–66] derived the test statistic distribution. It is a distribution-free method (it does not require, e.g., normally distributed data), and it can be useful for detecting trends in time series when the data exhibit a monotonic function of time. MK test is robust against the influence of extreme values, suitable to be used with skewed variables [67], and appropriate for data that do not display seasonal variation (or for seasonally corrected data) and have negligible autocorrelation. It has been widely used and recommended by many researchers (see, e.g., [5, 60, 68]) to detect trends in the field of hydrology and in similar applications. The *tau b* correlation coefficient [65] is the natural estimator of the strength of the trend in the case of using the MK test and is easily derived.

In accordance with MK test, the null hypothesis  $H_0$  is that the data is a sample of  $n$  independent and identically distributed (iid) random variables [5, 69]. The test statistic, Kendall's  $S$  [65], is calculated as follows:



$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(x_i - x_j) \quad (6)$$

where  $\text{sign}(x_i - x_j)$  is equal to 1, 0, or  $-1$  according to the sign of the difference  $x_i - x_j$ . Under the null hypothesis,  $E(S) = 0$ , and

$$\text{var}(S) = \frac{n(n-1)(2n+5) - \sum_{j=1}^g k_j(k_j-1)(2k_j+5)}{18} = \sigma^2 \quad (7)$$

where  $g$  is the number of tied groups and  $k_j$  is the number of observations in the  $j$ th group (if there are no ties, delete the summation in the numerator of Eq. 7). For  $n > 10$ , the test statistic

$$Z_S = \begin{cases} (S-1)/\sigma & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ (S+1)/\sigma & \text{if } S < 0 \end{cases} \quad (8)$$

where  $\sigma = \sqrt{\text{var}(S)}$  follows a standard normal distribution [65]. The significance levels ( $p$ -values) can be obtained as follows:

$$p = 2(1 - \Phi(|Z_S|)) \quad (9)$$

where  $\Phi()$  is the cumulative distribution function (CDF) of a standard normal distribution.

The MK test detects monotonic trends, so it cannot assess a trend reversal. A sequential MK test, originally proposed in Sneyers [70], is used in Partal and Kahya [71] and Shifteh Some'e et al. [72]. Unfortunately, it is not a test in a statistical sense, but a heuristic visual procedure to detect a trend reversal. We consider this approach to be quite subjective, and we do not recommend it.

Hirsch et al. [5] developed an extension of Kendall's test called the Seasonal Kendall (SK) test that accounts for seasonality in the data. In order to estimate the magnitude of the trend, they developed the SK slope estimator, which is an extension of the estimator proposed by Theil [73] and Sen [74]. A modified version of the SK test that allows for both seasonal data and serial dependence has also been developed [6]. The literature contains many methods for dealing with seasonality (i.e., monthly or quarterly trends); however, the issue is beyond the scope of this review, which focuses on yearly data.

Spearman's  $\rho$  correlation is an alternative to Kendall's  $\tau$ , despite being less commonly used in trend assessment. The results of the two statistics and tests are very similar; however, the faster convergence to the normal distribution of the statistic based on Kendall's  $\tau$  makes it slightly preferable in the case of a small dataset [75].

If a significant trend is found, the magnitude (rate of change per unit time) can be estimated using the TS slope [73, 74]. This approach involves computing slopes for all pairs of data points  $(x_i - x_j)/(t_i - t_j)$  and then using the median of these slopes as an estimate of the overall slope:

$$\beta_{TS} = \text{Med}\left(\frac{x_i - x_j}{t_i - t_j}\right) \quad \forall i > j, \quad i, j = 1, \dots, n \quad (10)$$

Thus, it is fairly sensitive to the presence of extreme values but can handle a moderate number of values below the detection limit and missing values. The trend slope,  $\beta_{TS}$ , is a measure of monotonic change and represents the median rate of change for the selected period, assisting the user in comparing the magnitudes of trends for several stations. However, it must be said that it is a linear trend estimator.

Another option, which as far as we know has not been used in the area of monitoring GWBs, is RoCoCo (*robust rank correlation coefficients*), proposed by Bodenhofer [76, 77], who presents a set of measures to test for monotonic associations between two observables. These robust rank correlation measures are based on fuzzy orderings, and the tests developed seem to outperform the classical variants because they are more robust for small samples when facing noise. The MK test is ideally suited for detecting monotonic relationships, but if you have numerical data such as pollutant concentrations, they may contain noise. In that case, even small random perturbations of true values may obscure a monotonic association. The robust gamma correlation coefficient seems to overcome this problem. The basic idea behind this correlation is to replace the strict orderings in the definitions of concordant and discordant pairs by continuous functions that measure the degree to which one value is greater than another (fuzzy ordering). Given a dataset consisting of  $n$  pairs of observations,  $(t_i, x_i)_{i=1}^n$ , the scoring functions  $R_T$  and  $R_X$  are used to compute the overall score of concordant pairs,  $C$ , and the overall score of discordant pairs,  $D$ :

$$\begin{aligned} \tilde{C} &= \sum_{i=1}^n \sum_{j \neq i} \bar{T}(R_T(t_i, t_j), R_X(x_i, x_j)) \\ \tilde{D} &= \sum_{i=1}^n \sum_{j \neq i} \bar{T}(R_T(t_i, t_j), R_X(x_j, x_i)) \end{aligned}$$

The function  $\bar{T}$  is a triangular norm used for aggregating the relationships between  $t$  and  $x$  components. The final robust gamma rank correlation coefficient is then computed as:

$$\tilde{\gamma} = \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}} \quad (11)$$

in perfect analogy with Goodman's and Kruskal's gamma.

To test the significance of the robust gamma correlation coefficient, Bodenhofer and Klawonn [76] propose permutation testing. In order to calculate the correlations and to perform the tests, we use the R package RoCoCo [78].

### 2.1.3 Time Series Correlation

One of the underlying assumptions common to all the models and tests presented so far is that the data are serially uncorrelated. When there is serial time correlation, the performance of the tests at assessing the significance of the trend could be affected (type I error) and may give misleading results. The effects are well known [79, 80]: in the case of positive autocorrelation, the tests have a tendency to be risky (increase in false positives), while if the autocorrelation is negative, the tests perform in a conservative way (increase in false negatives). Environmental variables, such as hydrological data, frequently exhibit some form of positive autocorrelation [80, 81], and several authors have studied and quantified the type I error in a large amount of situations (both with analytical and simulation studies) and shown that they are often greater than the significance level adopted ([80, 82, 83], among others). If we are applying a parametric model, the time series statistics framework can be used to overcome the problem of autocorrelation. However, it is not easy to obtain reliable results when dealing with a small sample dataset, and probably for this reason, such a framework has not often been applied or recommended for annual data.

The most simple case is the presence of *lag-1 autocorrelation* (the reading  $x_i$  recorded at time  $t_i$  is correlated with the previous value  $x_{i-1}$ ), that is to say, a first-order autoregressive process AR(1). For annual data, it may be a suitable approach, because the serial correlation with a lag bigger than 1 year will in general be low. From now on in this work, we will assume only a possible significant lag-1 autocorrelation.

Several proposals have been made to avoid false trend detections resulting from autocorrelation. Generally speaking, such efforts can be classified into two different approaches. The first modifies the statistical test to account for the presence of the serial correlation. Further information regarding this first approach, including the advantages of its use, can be found in Yue and Wang [84] and Khaliq et al. [85]. The second approach transforms the original data so that it meets the assumption of no temporal dependence [86]. This second approach is often adopted via the procedure called *pre-whitening* (PW) [79, 87, 88]. In all cases, a key step in the process is the estimation of the serial dependence.

Nevertheless, at this point, there is another obstacle to consider: the interaction between the trend and the autoregressive process [89–91]. If the data reflect a

deterministic trend, the estimate of the autocorrelation will become artificially inflated. For this reason, it has been proposed that the trend be extracted from the data (*detrending*) prior to the estimation of the autocorrelation [80]. Therefore, in order to evaluate the autocorrelation, two steps are involved: (1) detrending of the data (in the case of a significant trend) and (2) estimation of the autocorrelation within the detrended data.

In order to detrend (DT) the data, the following procedure has been proposed:

1. Calculate the Theil-Sen (TS) estimator with the original data,  $\beta_{TS}$ .
2. Assess whether the trend is significant using a  $(1 - \alpha) \times 100\%$  confidence interval of the slope. At least two different procedures have been described to build this confidence interval: a procedure based on order statistics [92, 93] and a bootstrapping procedure [94]. A controversial point is the election of the significance level (type I error);  $\alpha = 0.05$  ( $\alpha = 0.10$  if we choose a unilateral approach) is the usual value, but it could be too conservative in this step (low test power), so we recommend a value from 0.10 to 0.20 (0.15 and 0.20 are values often recommended in many stepwise variable selection procedures).
3. If zero is inside the confidence interval, we accept the hypothesis of no trend; and therefore the data remain unchanged:  $x_i^* = x_i$ . Otherwise, we remove the trend, transforming the data; thus,

$$x_i^* = x_i - \beta_{TS} t_i$$

At this point, we want to note an obvious fact that is poorly reported in the literature: despite TS being a robust nonparametric estimator, it simply evaluates a slope, and therefore, the detrending procedure described above only removes a linear trend. It is straightforward and easy to understand that if the data reflect a nonlinear trend, detrending with a linear function is inappropriate. Suppose we have a data series with a quadratic trend (Eq. 2); then the detrending procedure leads to:

$$x_i^* = x_i - \beta_{TS} t_i = \beta_0 + (\beta_1 - \beta_{TS}) t_i + \beta_2 t_i^2 + e_i \quad i = 1, \dots, n$$

which is another quadratic function of time (so the data reflect a new trend). Even for data with a linear trend (Eq. 1), it is not certain that detrending completely removes the trend:

$$x_i^* = x_i - \beta_{TS} t_i = \beta_0 + (\beta_1 - \beta_{TS}) t_i + e_i \quad i = 1, \dots, n$$

Can we be sure that  $\beta_1 - \beta_{TS} \approx 0$ ? It is not always, particularly with small datasets. To summarize, the detrending procedure may produce data that reflect a new trend. As far as we know, there have been no exhaustive studies of the consequences of the detrending procedure on the estimation of autocorrelation when it is applied to data with linear and nonlinear trends.

Once we establish that the data have no trend, we have to evaluate the autocorrelation. This is no easy matter, as can be seen in much of the published work

([95–98], among others). To estimate the lag-1 autocorrelation, the conventional estimator is widely used [99]:

$$r_1 = \frac{\sum_{i=2}^n (x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (12)$$

The statistic  $r_1$  is very similar to Pearson's correlation coefficient, but the numerator only has  $n - 1$  terms, and  $\bar{x}$  is used instead of calculating the average for  $x_1, \dots, x_n$  and  $x_1, \dots, x_{n-1}$ . Under the hypothesis  $\rho_1 = 0$ , the expected value of  $r_1$  is  $E(r_1) = -1/n$  [100]. On this basis, Huitema and McKean [101] proposed the modified estimator:

$$r_1^+ = r_1 + \frac{1}{n} \quad (13)$$

Nevertheless,  $r_1^+$  has a negative bias if  $\rho_1 > 0$  and  $n$  is small. In fact, the approximate bias of  $r_1$  is  $-(1 + 4\rho_1)/n$  [102]. To test  $H_0 : \rho_1 = 0$ , we use the approximation proposed by Moran [100] of  $\text{var}(r_1)$ :

$$\text{var}(r_1) = \frac{(n-2)^2}{n^2(n-1)} (1 - \rho_1^2)$$

which is an accurate estimation under the null hypothesis [103, 104]. So, under the null hypothesis  $\rho_1 = 0$ ,  $E_0(r_1^+) = 0$  is the expected value of  $r_1^+$ , and the variance is

$$\text{var}_0(r_1^+) = \frac{(n-2)^2}{n^2(n-1)} \quad (14)$$

It is now straightforward to use the statistic

$$Z_{r_1^+} = \frac{n\sqrt{n-1}}{(n-2)} r_1^+$$

which, under the null hypothesis, asymptotically follows a standard normal distribution. Therefore, it is now easy to calculate the  $p$ -value:

$$p = 2 \left( 1 - \Phi \left( \left| Z_{r_1^+} \right| \right) \right)$$

As we said above, the existence of serial correlation affects the MK test [67, 79]. Autocorrelation implies a change in the variance of the statistic  $S$ ,  $\text{var}(S)$ . Specifically, a lag-1 positive autocorrelation  $\rho_1 > 0$  increases the  $\text{var}(S)$ , which will

produce an increase in type I errors if we apply the usual MK test [84, 105]. In the case of a negative autocorrelation  $\rho_1 < 0$ , we will have exactly the opposite effect, but as we mention above, we expect a null or a positive time serial correlation with pollutants, and it is very rare to detect a negative autocorrelation. In order to solve the problem with the MK test, two strategies appear in the literature: data PW and modifying the MK test.

### Modified MK Test for Autocorrelated Data

One way to modify the MK test is by modifying  $\text{var}(S)$ . There are several proposals, and here we present the method described in Lettenmaier [106] and Hamed and Rao [107], reviewed and evaluated in Yue and Wang [84]. Once we have evaluated the lag-1 autocorrelation and rejected the null hypothesis  $\rho_1 = 0$  (if the null hypothesis is accepted, it is recommended to apply the usual MK test), we apply the modified MK (MKM) test. The modified variance of  $S$ ,  $\text{var}^*(S)$ , is calculated as follows:

$$\text{var}^*(S) = \frac{n}{n^*} \text{var}(S)$$

where  $n^*$  is the effective number of independent samples [108]. Matalas and Langbein [105] derived a formula for  $n^*$  in the case of a lag-1 autoregressive process:

$$n^* = \frac{n}{1 + 2 \frac{\rho_1^{n+1} - n\rho_1^2 + (n-1)\rho_1}{n(1-\rho_1)^2}} \quad (15)$$

To arrive at our estimate of  $\text{var}^*(S)$ ,  $\hat{\text{var}}^*(S)$ , we have to replace  $\rho_1$  with the estimator  $\hat{\rho}_1 = r_1^+$  in Eq. 15. Next, we apply (for  $n > 10$ ) the transformation specified in Eq. 9, replacing  $\sigma$  by  $\hat{\sigma}^* = \sqrt{\hat{\text{var}}^*(S)}$ .

Finally, we note that other modifications of the MK test have been published (e.g., [67, 107]).

### PW

The other approach to deal with autocorrelation is data PW [79] which means removing the serial correlation from the data prior to applying the test. This procedure reduces the occurrence of type I errors close to the adopted (nominal) significance level [86, 87]. Nevertheless, as we say above, the estimate of the autocorrelation will become artificially inflated if the data reflect a trend. For this reason, *trend-free PW* (TFPW) has been proposed [80]. TFPW is a four-step procedure: 1) DT the data; 2) estimate the autocorrelation; 3) remove the serial correlation; and 4) replace the trend in the data. According to Öñöz and Bayazit

[83], TFPW is frequently more powerful than the PW at detecting trends, and it is widely used in much work [80, 109–114].

We propose the following TFPW algorithm that deals with missing values:

1. Apply the detrending procedure and estimate the lag-1 autocorrelation  $\hat{\rho}_1 = r_1^+$ .
2. Is the autocorrelation significant? If  $H_0 : \rho_1 = 0$  is accepted (choose  $\alpha$  between 0.10 and 0.20), then do not transform the data, i.e., final data are original data:  $x_i^{***} = x_i$ .
3. If  $H_0 : \rho_1 = 0$  is rejected and  $\beta_{TS}$  (recall TS estimator) is not significant (the zero value is inside the confidence interval), remove the autocorrelation from the original data:  $x_i^{***} = x_i - \hat{\rho}_1 x_{i-1}$ .
4. If  $H_0 : \rho_1 = 0$  is rejected and  $\beta_{TS}$  is significant, then:
  - a. Remove the autocorrelation from the detrended data:

$$x_i^{**} = x_i^* - \hat{\rho}_1 x_{i-1}^*$$

where  $x_i^* = x_i - \beta_{TS} t_i$ . Here, another obstacle arises: what do we do if the data have missing values? Curiously, this is an important point that is absent from the work we have reviewed. We propose imputing them (e.g., by applying linear interpolation), only in order to be able to remove the autocorrelation; otherwise, every original missing value implies losing another final value in this step.

- b. Replace the trend in the data:  $x_i^{***} = x_i^{**} + \beta_{TS} t_i$ .

Once TFPW has been performed, you can apply the usual MK test and TS estimator to the transformed data,  $x_t^{***}$ . Finally, in Öñöz and Bayazit [83], a modified form of TFPW (referred to as MTFPW) that aims to reduce its probability of rejecting a true  $H_0$  was proposed. There is still some controversy in the literature regarding the most appropriate approach to correcting for serial correlation [115, 116]. For example, we think that many researchers forget that if  $\rho_1 = 0$  and we reject the null hypothesis (due to a type I error), PW introduces autocorrelation. In this unfortunate case, it is easy to calculate this serial correlation:

$$\text{corr}(x_i, x_{i-1}) = -\hat{\rho}_1 / (1 + \hat{\rho}_1^2)$$

where  $\hat{\rho}_1 = r_1^+$ . In summary, we may assume that further studies are still required to evaluate the performance of both PW and TFPW techniques.

## 2.2 Overall Assessment of GWBs

In this section, we deal with how to aggregate data from individual monitoring points (spatial aggregation) in order to assess a trend in a GWB as a whole. Grath

et al. [8] introduce the necessity to calculate trends for a GWB or group of GWBs. Since trends in GWBs as a whole cannot be observed directly, an approach has to be adopted to aggregating observations from individual monitoring points. Spatial correlation is a matter of concern here, because data series coming from various stations may be correlated. Although most studies ignore this fact in their interpretations of results, it must be kept in mind that the effect of cross-correlation in the data is to increase the expected number of trends [57].

Several approaches deal with overall trend assessment by grouping or blending the results from all the stations inside a GWB. Grath et al. [8] recommended carrying out the trend assessment with the mean values (average of all the stations for each period of time). An alternative is the average (or the median) of any statistic obtained from every single station within a GWB. In a case study of CIS Guidance No. 18 ([9], Sect. 10.8), two aggregation procedures are presented: (a) defining the median trend slope and (b) using age dating to aggregate time series (simply pooling data) along a standardized  $x$ -axis showing recharge time. In Douglas et al. [60], an interesting aggregation procedure based on the results of MK tests is described, and we will outline it below. However, an average or a median implies the assumption that the trend is homogeneous across all stations, which is difficult to justify. Thus, in Van Belle and Hughes [12], a preliminary test of the homogeneity of the trend direction was proposed. When the trend is heterogeneous across stations (e.g., an upward trend in one set of stations and a downward trend in another), any overall test of trend direction or slope estimator will be misleading.

### 2.2.1 S-Mean Method

The  $S$ -mean method [60] is based on the results of the MK test for every station. Specifically, we must compute:

$$\bar{S}_m = \frac{1}{m} \sum_{k=1}^m S_k \quad (16)$$

where  $S_k$  is the  $S$  statistic provided by the MK test for the  $k$ th station in a GWB with  $m$  stations. If the data series are cross-correlated, i.e., there is spatial correlation between stations, the variance of  $\bar{S}_m$  is

$$\text{var}(\bar{S}_m) = \frac{1}{m^2} \left[ \sum_{k=1}^m \text{var}(S_k) + 2 \sum_{k=1}^{m-1} \sum_{j=1}^{m-j} \text{cov}(S_k, S_{k+j}) \right]$$

Following Salas-La Cruz [117], the covariance between station  $k$  and station  $k + j$  is:



$$\text{cov}(S_k, S_{k+j}) = \sigma^2 \rho_{k,k+j}$$

where  $\rho_{k,k+j}$  is the cross-correlation coefficient between the two stations. Therefore, the variance of  $\bar{S}_m$  becomes:

$$\text{var}(\bar{S}_m) = \frac{1}{m^2} \left[ m\sigma^2 + 2 \sum_{k=1}^{m-1} \sum_{j=1}^{m-j} \sigma^2 \rho_{k,k+j} \right] = \frac{\sigma^2}{m} [1 + (m-1)\bar{\rho}_{xx}] \quad (17)$$

where:

$$\bar{\rho}_{xx} = \frac{2 \sum_{k=1}^{m-1} \sum_{j=1}^{m-j} \sigma^2 \rho_{k,k+j}}{m(m-1)}$$

is the average cross-correlation coefficient for the whole GWB. Considering that the overall GWB has no trend, it is obvious that  $E(\bar{S}_m) = 0$ . Therefore, asymptotically ( $n > 10$  could be enough for practical purposes), the statistic:

$$Z_{\bar{S}} = \frac{\bar{S}_m}{\sqrt{\text{var}(\bar{S}_m)}} \sim N(0, 1)$$

follows a standard normal distribution. Obtaining  $p$ -value is straightforward:

$$p = 2(1 - \Phi(|Z_{\bar{S}}|))$$

This procedure takes into account spatial correlation. In our opinion, it is much better than other approaches that simply mix or group data from the different stations. Nevertheless, we warn against the use of these approaches to blending the results of individual monitoring points: they implicitly assume a global trend for the whole GWB, and that is not always so. We outline below quite a different spatial aggregation method; it does not assume a common global trend.

### 2.2.2 Cross-Correlation: A Bootstrapping Procedure

Burn and Hag Elnur [61] propose an ingenious bootstrapping approach to determine the critical value for the percentage of stations expected to show a trend by chance. It can be summarized in the following steps:

1. Select a year at random from the entire period of time for which data are available (supposing that the sampled values are obtained yearly). Repeat this procedure for the required number of years, i.e., the number of years in the initial

dataset, so that the new dataset is in fact a resampling of the years, without altering the data from individual monitoring points.

2. Perform the MK test for every station (at a chosen significance level,  $\alpha$ ) and determine the percentage of stations with a significant trend.
3. Steps 1–2 are repeated  $k$  times ( $k = 600$  in Burn and Hag Elnur [61]), to obtain the empirical distribution of the percentage of stations that are significant at the  $\alpha$  level. Sort the  $k$  percentage values recorded and determine the percentile  $(1 - \alpha) \times 100\%$ ,  $p_{\text{crit}}$ .
4. If the actual percentage of stations showing a trend in the GWB analyzed is greater than  $p_{\text{crit}}$ , it will be considered significant.

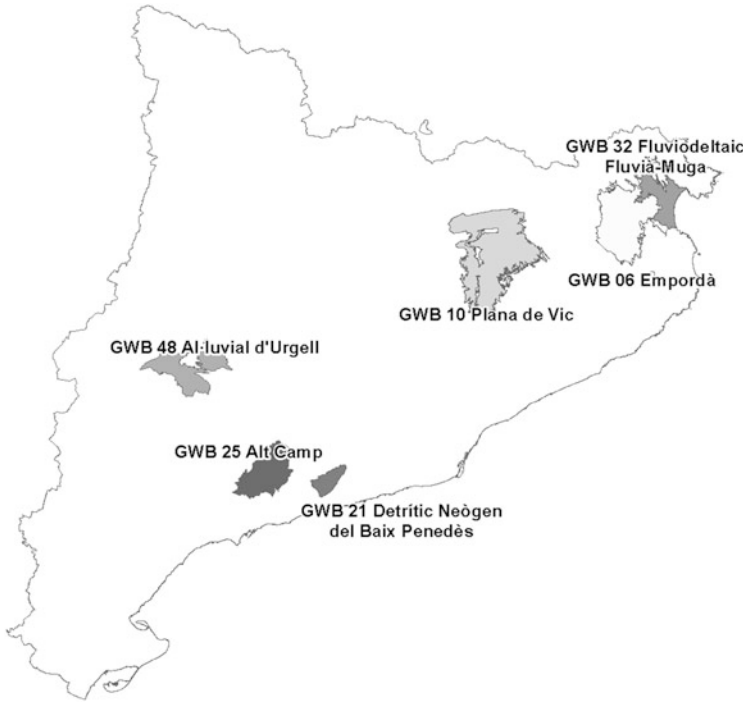
Note that any temporal structure (a trend) that exists in every single station will not be reproduced in the resampled datasets because of the bootstrapping process. However, the cross-correlations in the original data are preserved, allowing us to evaluate the effect of spatial correlation within the GWB. In Sect. 4, we will provide a modified procedure that outperforms the algorithm just outlined here.

### 3 Assayed Procedures: A Case Study

In order to illustrate the procedures and highlight the main results, some of the stations and GWBs in Catalonia were selected (Fig. 1), and some of the methods described above were applied. We first determined the trends for individual monitoring points, in accordance with nine specific methods (see below); and then we applied two distinct methods to show how data may be aggregated in order to determine the trend for a GWB as a whole. GWBs with a long monitoring period and many sampling sites were selected to be analyzed. Nitrate and chloride concentrations were analyzed in order to identify significant upward and downward trends.

Figure 2 shows the scatterplots of the twelve stations used to compare the nine methods. These stations were not selected at random, but specifically in order to show the performance of the methods in different scenarios. One possible classification of these monitoring points based on the data shown, according to a rough trend assessment, is the following:

- Linear or monotonic trends: stations N1, N2, N3 (the latter two with extreme values), and N7.
- Overdispersion: stations N4, N5, and N6. It is difficult to assess the trends due to overdispersion; perhaps the data could fit a quadratic or even a cubic curve.
- Trend reversal: evidently this is not certain, but quite possible, at N8, N9, C1, C2, and C3.

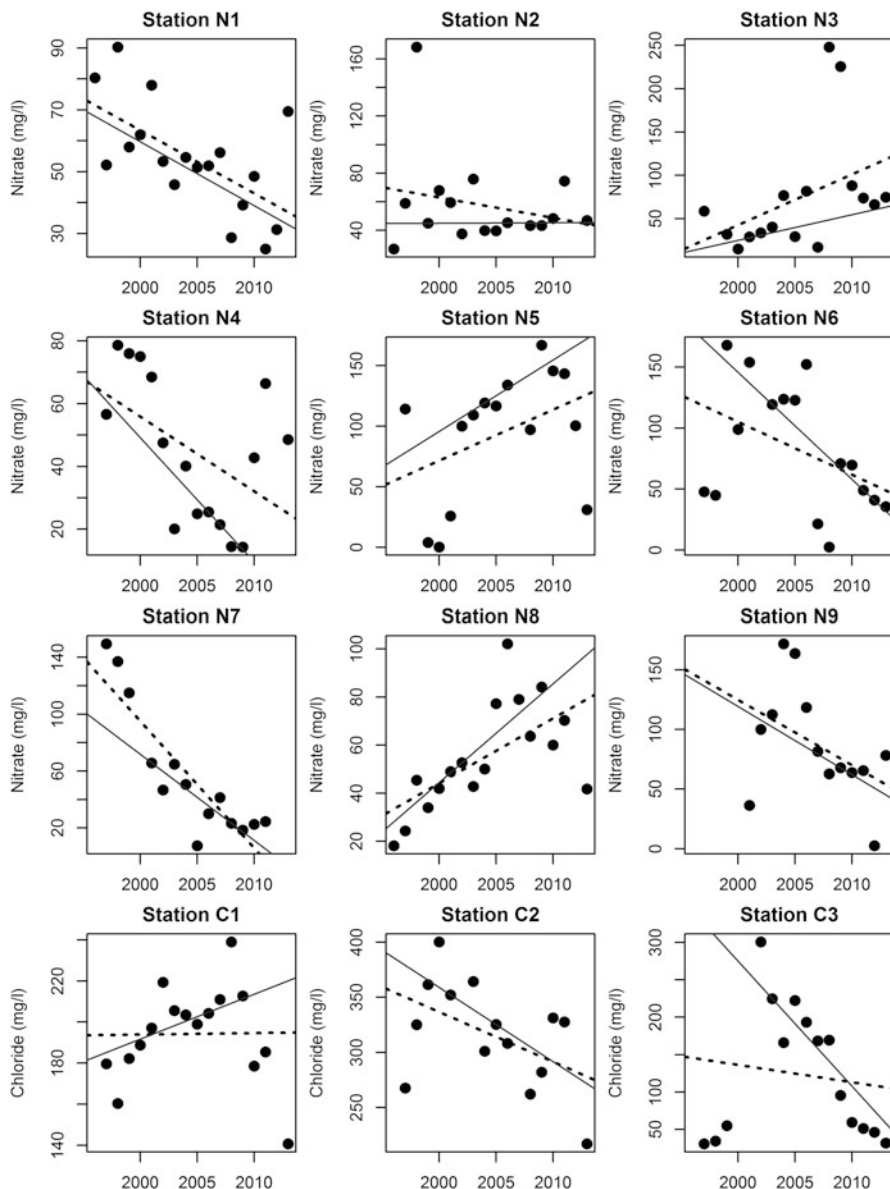


**Fig. 1** GWBs located in Catalonia and sampling sites currently used to monitor chemical status

### 3.1 Methods Assayed for Individual Monitoring Points

For each assayed method outlined below, we describe how it was performed (using package(s), function(s), or our own scripts in R), and we highlight the main results obtained in two categories: *goodness of fit* and *trend estimation*.

1. LR, using the *lm* function. Goodness of fit: Pearson's correlation coefficient ( $r$ ),  $R^2$ , and  $p$ -value. Trend estimation:  $\hat{\beta}_1$  (Eq. 1) and  $p$ -value for  $\hat{\beta}_1$ .
2. RLR, with the *rlm* function (MASS package, [35]). Setup: Huber-type function with parameter  $k = 1.345$  (default value),  $\hat{\sigma} = \text{MAD}$ . Goodness of fit:  $R^2$ . Trend estimation:  $\hat{\beta}_1$  (Eq. 3) and  $p$ -value for  $\hat{\beta}_1$ .
3. MK test and TS estimator (MK&TS), with the MK function (Kendall package, [118]) and *mblm* function (package mblm, [119]). Goodness of fit: *tau b* correlation coefficient and  $p$ -value. Trend estimation: TS estimator (Eq. 10).
4. RoCoCo and TS estimator (RCC&TS), using the RoCoCo package [78] to calculate the RoCoCo. Setup:  $R_T(t_i, t_j) = 1$  if  $t_j > t_i$  (zero otherwise),  $R_X(x_i, x_j) = \max\{0, \min((x_j - x_i)/r)\}$ , respectively named classical strict ordering and truncated linear scoring, and  $\bar{T}(x, y) = \min(x, y)$ . Goodness of



**Fig. 2** Scatterplots of pollutant concentrations (nitrates and chlorides) from 1996 to 2013 for the selected monitoring points, showing the LR fit (*dotted line*) and TS fit (*solid line*)

fit: robust gamma correlation coefficient and  $p$ -value. Trend estimation: TS estimator (Eq. 10).

5. MK modified test for autocorrelated data and TS estimator (MKM&TS). We programmed our own code for the calculus of the modified variance estimate,

$\text{var}^*(S)$ ; lag-1 autocorrelation,  $\hat{\rho}_1 = r_1^+$ ; and  $p$ -values. If serial data had missing values, we proposed the following heuristic correction to the estimation of lag-1 autocorrelation (Eqs. 13 and 14):

$$r_1^+ = \frac{(n+g)}{n} r_1 + \frac{1}{n}, \quad \text{var}_0(r_1^+) = \left(\frac{n+g}{n}\right)^2 \frac{(n-2)^2}{n^2(n-1)}$$

Goodness of fit: *tau b* correlation coefficient and  $p$ -value. Trend estimation: TS estimator (Eq. 10).

6. TFPW before applying the MK test and TS estimator (TFPW + MK&TS). We programmed our own code in order to perform the TFPW, as described in Sect. 2.1.3. A threshold value of  $\alpha = 0.2$  was used to test the significance of the TS estimator and lag-1 autocorrelation. Goodness of fit: *tau b* correlation coefficient and  $p$ -value. Trend estimation: TS estimator (Eq. 10).
7. TFPW before applying RoCoCo and TS estimator (TFPW + RCC&TS). This method combines TFPW with method 4. Goodness of fit: robust gamma correlation coefficient and  $p$ -value. Trend estimation: TS estimator (Eq. 10).
8. Two-section LR (2SLR). Breakpoints detected with the *piecewise* function (SiZer package, [120]). Significance test for the difference of slopes: Davies test [121], applied with the *davies.test* function (segmented package, [122]). Goodness of fit: none. Trend estimation:  $\hat{\beta}_1$  and  $\hat{\beta}_1 + \hat{\beta}_2$  (Eq. 5) and  $p$ -value for  $\hat{\beta}_2$  (Davies test).
9. QR, using the *lm* function. Goodness of fit:  $R^2$  and  $p$ -value. Trend estimation:  $\hat{\beta}_1$  and  $\hat{\beta}_2$  (Eq. 2) and  $p$ -value for  $\hat{\beta}_2$ .

### 3.2 Methods Assayed for Overall GWB Trend Assessment

- (1) *S-mean* method, described in Sect. 3.2: Nevertheless, missing values and ties are not considered in Douglas et al. [60]; thus, we propose a modified variance estimation of the statistic  $\bar{S}_m$ . The estimation of  $\text{var}(S) = \sigma^2$  could be approximated by  $\hat{\sigma}^2 \approx m\hat{\sigma}_{\bar{S}}^2$ , where:

$$\hat{\sigma}_{\bar{S}}^2 = \frac{1}{m^2} \left[ \sum_{k=1}^m \text{var}(S_k) \right] \quad (18)$$

Therefore:

$$\text{var}(\bar{S}_m) = \hat{\sigma}_{\bar{S}}^2 [1 + (m-1)\bar{\rho}_{xx}] \quad (19)$$

- (2) *Permutational* approach for testing overall GWB trend assessment (PTA): based on the procedure described in Burn and Hag Elnur [61], we propose a

modified method that (a) is a permutation test, so it gives a significance assessment (in terms of  $p$ -value), and (b) evaluates not only significant overall trends but upward and downward trends in a separate manner. The algorithm can be implemented in the following way:

1. For a given GWB dataset, perform trend test for every station. Given a significance level  $\alpha$ , collect and count separately the number of significant upward and downward trends.
2. Randomly select a year from the entire period of time for which data are available (assuming a yearly sampling period). Repeat this random selection without replacement for the required number of years. The new dataset is in fact a permutation of years, without altering any data from individual monitoring points.
3. Perform the test for every station (significance level  $\alpha$ ) and determine the number of stations with significant upward and downward trends.
4. Repeat steps 2–3  $k$  times ( $k = 9,999$  permutations in our study). Sort the values of upward and downward significant trends separately, to obtain the empirical distributions of the number of stations with significant trends (upward and downward). Combine these with the actual values for the GWB obtained in step 1.
5. Finally, calculate the two percentiles,  $P_{\text{exp}}$ , of the actual values obtained in step 1, so complementary values  $p = 1 - P_{\text{exp}}$  (one for upward trends and the other for downward trends) perform like  $p$ -values.

Any of the methods summarized in Sect. 3.1 for individual monitoring points could have been applied, but for the purpose of comparing with  $S$ -mean, we performed the MK test.

## 4 Results and Discussion

### 4.1 Individual Monitoring Points

First of all, we briefly discuss the issue of autocorrelation. At nine of the twelve stations, significant positive autocorrelation was detected (Table 1). Of these nine stations, in five cases, we previously detrended (due to a significant TS slope). Therefore, because positive autocorrelation increases the likelihood of a false positive in the trend assessment, it seems absolutely essential to use a method that takes into account the autocorrelation (i.e., MKM) or to perform TFPW.

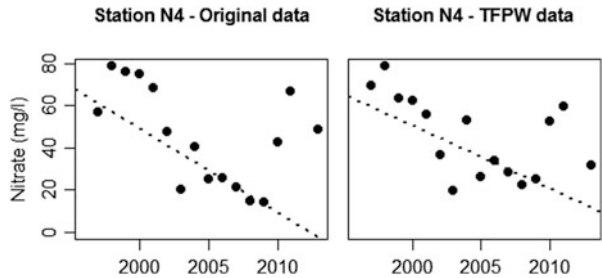
However, further analysis revealed serious flaws. Is there real autocorrelation of the data, or is it a spurious detection due to erroneous detrending? As you may recall, after linear detrending to data, if it is not appropriate (i.e., the data reflect a nonlinear trend), the transformed data will show a trend that will affect the estimate of autocorrelation. We cannot be sure exactly what happened (we are referring here to the overestimation of the autocorrelation in our case study), but it is clear that at

**Table 1** Lag-1 autocorrelation (estimate and *p*-value) and detrending applied (yes or no) according to TS confidence interval for all individual monitoring points (stations)

Groundwater body	Station code	Station label	Detrending	Lag-1 autocorrelation	
				$\hat{\rho}_1$	<i>p</i> -value
GWB48	25248-0005	N1	YES	−0.2644	0.2200
GWB21	43074-0058	N2	NO	−0.0382	0.8805
GWB10	08129-0011	N3	YES	0.3677	<b>0.1154</b>
GWB21	43140-0078	N4	YES	0.5348	<b>0.0220</b>
GWB49	25152-0006	N5	YES	0.5707	<b>0.0247</b>
GWB49	25217-0009	N6	YES	0.3191	<b>0.1719</b>
GWB25	43161-0159	N7	YES	0.6785	<b>0.0097</b>
GWB21	43020-0056	N8	YES	0.2858	0.2211
GWB33	17085-0012	N9	NO	0.7203	<b>0.0106</b>
GWB39	08301-0035	C1	NO	0.4850	<b>0.0378</b>
GWB49	25242-0008	C2	NO	0.4337	<b>0.1187</b>
GWB23	08231-0025	C3	NO	0.7896	<b>0.0010</b>

Labels N and C indicate nitrate and chloride pollutants, respectively

**Fig. 3** Comparison plots of raw data and TFPW transformed data for station N4, with TS fitted line



various stations with significant autocorrelation, the original trend seems not to be linear, so that the detrending will not have been effective (misrepresenting the data). The nonlinear tendency is quite evident, for example, in the case of N4, N7, C1, C2, and C3. Let us look in more detail at the case of N4: the original nonlinear trend (maybe a quadratic or cubic trend) means that the detrending actually increased the linear trend (Fig. 3), so that the two methods with TFPW detect a more significant trend than the respective non-TFPW methods.

Next, we discuss the results of the nine methods in the same order as that in which they appear in Table 2 and that they were described in Sect. 3.

LR is a good method when the data fit a linear trend reasonably well (Fig. 2). However, there are several pitfalls associated with it when treating groundwater data. It can be greatly affected by extreme values; for example, consider station N3 (Fig. 4): LR fails to detect an upward trend due to the inclusion of two extreme values in the time series which increase the residual values substantially. Also, the method has difficulties in cases of overdispersion (N5, N6) or a poorly defined nonlinear trend (N4). Obviously, it is not useful in cases with a trend reversal: it was

**Table 2** Goodness of fit, trend estimation, and trend assessment (upward ↑, no trend →, downward ↓) using the methods assayed (LR, RLR, MK&TS, MKM&TS, RCC&TS, TFPW + MK&TS, TFPW + RCC&TS, 2SLR, and QR) for the stations selected

Station	Method	Goodness of fit			Trend estimation			Assessment (↑ → ↓)
		<i>r</i>	$\sqrt{R^2}$	<i>p</i> -value	Estimates of parameters		<i>p</i> -value	
N1	LR	−0.6189	0.6189	<b>0.0062</b>	−2.035		<b>0.0062</b>	↓
	RLR	−	0.7043	−	−2.334		<b>0.0035</b>	↓
	MK&TS	−0.4641	−	<b>0.0080</b>	−2.136		−	↓
	RCC&TS	−0.4691	−	<b>0.0049</b>	−2.136		−	↓
	2SLR	−	−	−	2.753	40.457	<b>0.0373</b>	↓ 2012 ↑
	QR	−	0.6570	<b>0.0145</b>	−4.701	0.157	0.2755	LR
N2	LR	−0.2311	0.2311	0.3891	−1.428		0.3891	→
	RLR	−	0.3425	−	−0.315		0.7386	→
	MK&TS	0.0418	−	0.8569	0.146		−	→
	RCC&TS	0.0448	−	0.8185	0.146		−	→
	2SLR	−	−	−	−3.079	2.277	0.8809	LR
	QR	−	0.2529	0.6508	−3.690	0.136	0.7084	LR
N3	LR	0.4089	0.4089	0.1031	5.142		0.1031	→
	RLR	−	0.5321	−	3.109		<b>0.0492</b>	↑
	MK&TS	0.3235	−	<b>0.0765</b>	2.876		−	↑
	RCC&TS	0.3252	−	<b>0.0766</b>	2.876		−	↑
	MKM&TS	0.3235	−	0.2158	2.876		−	→
	TFPW + MK&TS	0.2833	−	0.1373	3.298		−	→
	TFPW + RCC&TS	0.2867	−	0.1343	3.298		−	→
	2SLR	−	−	−	−8.150	7.085	0.9447	LR
	QR	−	0.4104	0.2749	6.731	−0.092	0.8880	LR



N4	LR		-0.4080	0.4080	0.1040	-1.785	0.1040	→
	RLR		-	0.4081	-	-1.785	0.1040	→
	MK&TS		-0.3382	-	0.0638	-2.030	-	↓
	RCC&TS		-0.3502	-	0.0573	-2.030	-	↓
	MKM&TS		-0.3382	-	0.2837	-2.030	-	→
	TFPW + MK&TS		-0.4500	-	0.0170	-2.363	-	↓
	TFPW + RCC&TS		-0.4432	-	0.0167	-2.363	-	↓
	2SLR		-	-	-	-4.225	8.369	LR
	QR		-	0.5221	0.1078	-6.794	0.304	LR
	LR		0.3722	0.3722	0.1557	3.541	0.1557	→
N5	RLR		-	0.4394	-	3.774	0.1651	→
	MK&TS		0.3500	-	0.0649	4.089	-	↑
	RCC&TS		0.3590	-	0.0586	4.089	-	↑
	MKM&TS		0.3500	-	0.3033	4.089	-	→
	TFPW + MK&TS		0.2190	-	0.2763	2.581	-	→
	TFPW + RCC&TS		0.2093	-	0.2856	2.581	-	→
	2SLR		-	-	-	6.951	-69.301	LR
	QR		-	0.4240	0.2758	10.668	-0.411	LR
	LR		-0.3600	0.3600	0.1571	-3.307	0.1571	→
	RLR			0.3600	-	-3.307	0.1571	→
N6	MK&TS		-0.2794	-	0.1275	-4.398	-	→
	RCC&TS		-0.2882	-	0.1207	-4.398	-	→
	MKM&TS		-0.2794	-	0.2626	-4.398	-	→
	TFPW + MK&TS		-0.3167	-	0.0957	-4.7725	-	↓
	TFPW + RCC&TS		-0.3382	-	0.0764	-4.772	-	↓
	2SLR		-	-	-	19.779	-10.186	↑ 2001 ↓
	QR			0.6000	0.0452	13.059	-0.969	↑ 2003 ↓

(continued)

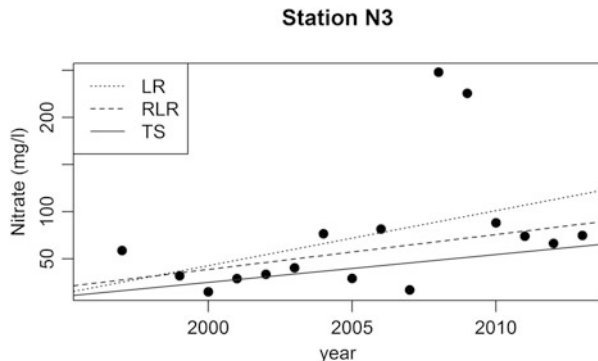
Table 2 (continued)

Station	Method	Goodness of fit			Trend estimation		Assessment
		$r$	$\sqrt{R^2}$	$p$ -value	Estimates of parameters	$p$ -value	
N7	LR	-0.9041		<b>0.0000</b>	-9.738	<b>0.0000</b>	(↑ → ↓)
	RLR	-	0.9120		-9.703	<b>0.0000</b>	↓
	MK&TS	-0.7524	-	<b>0.0001</b>	-9.650	-	↓
	RCC&TS	-0.7900	-	<b>0.0000</b>	-9.650	-	↓
	MKM&TS	-0.7524	-	<b>0.0602</b>	-9.650	-	↓
	TFPW + MK&TS	-0.7582	-	<b>0.0002</b>	-8.483	-	↓
	TFPW + RCC&TS	-0.8408	-	<b>0.0000</b>	-8.483	-	↓
	2SLR	-	-	-	-20.198	-3.678	↓ 2002 ↓
	QR	-	0.9758	<b>0.0000</b>	-23.944	0.954	↓ 2009 ↑
	LR	0.6271	0.6271	0.0071	2.697	<b>0.0071</b>	↑
N8	RLR	-	0.7411	-	3.065	<b>0.0008</b>	↑
	MK&TS	0.5147	-	<b>0.0045</b>	3.248	-	↑
	RCC&TS	0.5282	-	<b>0.0022</b>	3.248	-	↑
	2SLR	-	-	-	6.120	-5.134	↑ 2007 ↓
	QR	-	0.8125	<b>0.0005</b>	10.512	-0.475	↑ 2007 ↓
	LR	-0.3733	0.3733	0.2089	-4.536	0.2089	→
N9	RLR	-	0.4543	-	-4.488	0.2441	→
	MK&TS	-0.2527	-	0.2284	-4.037	-	→
	RCC&TS	-0.2599	-	0.2233	-4.037	-	→
	MKM&TS	-0.2527	-	0.5655	-4.037	-	→
	TFPW + MK&TS	-0.0769	-	0.7603	-2.831	-	→
	TFPW + RCC&TS	-0.1010	-	0.6543	-2.831	-	→
	2SLR	-	-	-	28.337	-17.15	↑ 2004 ↓
	QR	-	0.7886	<b>0.0077</b>	46.354	-2.544	↑ 2005 ↓

C1	LR	0.0850	0.0850	0.0850	0.7462	0.381	0.7462	→
	RLR	–	0.2547	–	–	0.999	0.3835	→
	MK&TS	0.2500	–	–	0.1740	2.087	–	→
	RCC&TS	0.2501	–	–	0.1752	2.087	–	→
	MKM&TS	0.2500	–	–	0.4042	2.087	–	→
	TFPW + MK&TS	–0.0500	–	–	0.8219	–0.131	–	→
	TFPW + RCC&TS	–0.0348	–	–	0.8622	–0.131	–	→
	2SLR	–	–	–	–	4.3420	–16.363	↑ 2008 ↓
	QR	–	0.7718	0.0018	0.0018	12.514	–0.737	↑ 2004 ↓
	LR	–0.3530	0.3530	0.1974	–	–3.037	0.1974	→
C2	RLR	–	0.3500	–	–	–3.014	0.2321	→
	MK&TS	–0.1429	–	–	0.4884	–3.156	–	→
	RCC&TS	–0.1777	–	–	0.3657	–3.156	–	→
	MKM&TS	–0.1429	–	–	0.6482	–3.156	–	→
	TFPW + MK&TS	–0.2967	–	–	0.1546	–3.037	–	→
	TFPW + RCC&TS	–0.3032	–	–	0.1579	–3.037	–	→
	2SLR	–	–	–	–	30.090	–9.011	↑ 2000 ↓
	QR	–	0.6339	0.0458	0.0458	13.517	–1.006	↑ 2003 ↓
	LR	–0.0152	0.0152	0.9553	–	–0.243	0.9553	→
	RLR	–	0.0328	–	–	0.044	0.9918	→
C3	MK&TS	–0.1333	–	–	0.4995	–5.675	–	→
	RCC&TS	–0.1584	–	–	0.4306	–5.675	–	→
	MKM&TS	–0.1333	–	–	0.7862	–5.675	–	→
	TFPW + MK&TS	–0.1810	–	–	0.3731	–1.497	–	→
	TFPW + RCC&TS	–0.1891	–	–	0.3400	–1.497	–	→
	2SLR	–	–	–	–	45.157	–20.88	↑ 2002 ↓
	QR	–	0.8494	0.0002	0.0002	47.842	–2.863	↑ 2004 ↓

If a breakpoint (2SLR) or an inflection point (QR) is detected, we show the year in the assessment column

**Fig. 4** Scatterplot for station N3 showing three fitted lines, adjusted according to the LR, RLR, and TS methods assayed



not able to detect any trend (N9, C1, C2, C3), or it only detected one trend (an upward trend for station N8). In the case of N7, LR detects the downward trend perfectly, but we cannot identify the slope change. Finally, LR does not take autocorrelation into account, and as previously explained, this fact can affect the significance of the model. In this scenario, we could have applied TFPW to the data; but as we mention earlier and we will return to later, TFPW is a risky method. Another approach would be to apply any of the methodologies that have been published and implemented in a lot of software for the treatment of time series. In short, LR is highly sensitive to outliers and overdispersion; it only detects one linear trend (and therefore it cannot detect a trend reversal) and does not take autocorrelation into account. In general, we advise against the use of LR for trend detection.

RLR has proved to be resistant to extreme values. For the example station N3 (Fig. 4), LR failed to detect an upward trend due to the inclusion of two extreme values; however, RLR found the increasing trend. We suggest it as an alternative to LR, and we are a bit surprised that it has not been more widely recommended in the area of groundwater monitoring. It does still have limitations though: (1) it only works well for a linear trend; (2) it cannot detect trend reversal; and (3) it does not take autocorrelation into account.

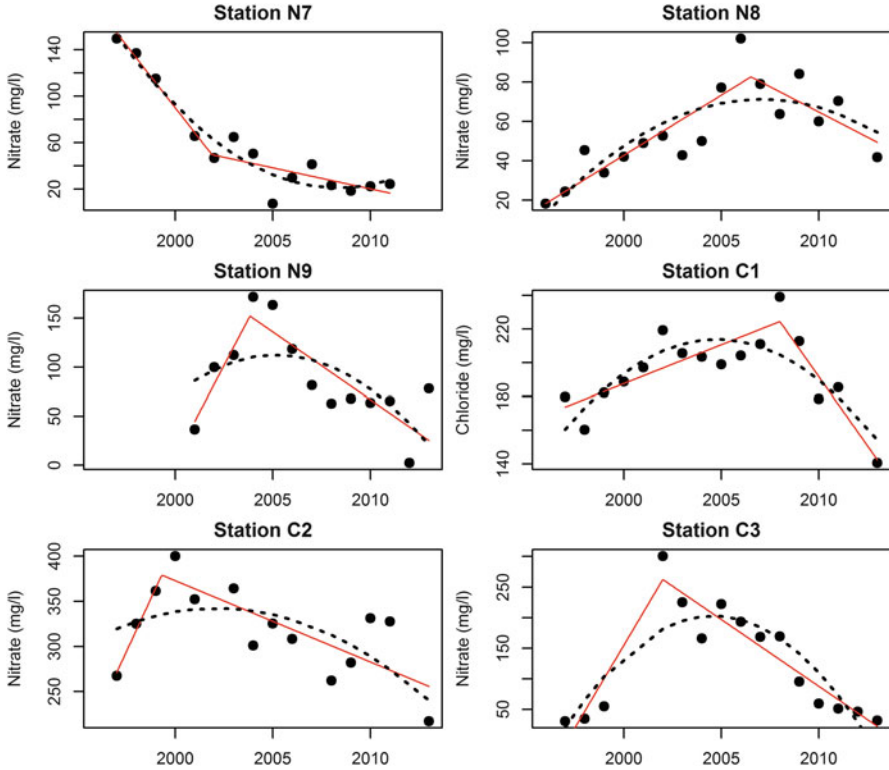
MK test and RoCoCo (with TS estimator of the slope): as evident from Table 2, these two methods usually give almost the same final results. RoCoCo, based on fuzzy scorings and more complex than Kendall's tau b, does not improve trend detection. More theoretical and simulation studies are necessary to determine in what situations RoCoCo may become a better method to detect a trend. In particular, perhaps another choice of the fuzzy scoring function could achieve better results. Compared to RLR, the only difference is that MK test and RoCoCo detected a downward trend in N4 and an upward trend in N5. An important point to remember is that MK test and RoCoCo detect monotonic trends but are not helpful in detecting trend reversals, as we can see from the results for stations N9, C1, C2, and C3. Finally, we would like to highlight again that trend estimation with TS estimator, despite being robust, is just linear (Fig. 2).

MKM test for autocorrelated data: as we state above, this has the advantage of not manipulating the original data, but simply changing the variance of the MK  $S$  statistic. In our case study, all nine autocorrelations detected were positive, so  $\text{var}^*(S) > \text{var}(S)$ , and the MKM test offers greater  $p$ -values than MK test. As we can see, MKM test becomes a much more conservative test than MK test and only detects a significant trend in station N7. It is difficult to quantify the effect, but it is evident that the incorrect estimate (even spurious) of autocorrelation makes MKM test very conservative. Although we cannot draw definitive conclusions from just a few examples of stations, we are afraid that the MKM test is exceedingly conservative for trend detection.

TFPW with MK test or RoCoCo: as we have discussed previously, it is not clear whether TFPW is a successful technique or leads to involuntary contamination of the original data. It is certainly the latter case when the data present a nonlinear trend. If we examine stations N4 and N6 carefully, we notice a peculiar feature. In both cases, TFPW clearly increases trend detection (in N6, this represents a shift from nonsignificance to significance). The risk due to both the data detrending and the estimate of the autocorrelation can be serious, and TFPW could produce misleading results; therefore, we advise against the blind application of TFPW. In any case, we recommend applying the test without TFPW first. If the hypothesis of no trend is accepted, then do not apply TFPW to the data. If a significant trend is detected, then apply TFPW and redo the test and finally choose the larger  $p$ -value (this conservative option is suitable if we accept that there is a null or a positive autocorrelation).

2SLR displays a surprisingly high (maybe too high in some cases) capacity to detect significant breakpoints (stations N1, N6, N7, N8, N9, C1, C2, and C3), sometimes even at the end of the time series (Fig. 5). We do not know the reliability of the Davies test (used to test for different slopes and so the significance of the breakpoint) in the case of small data series and autocorrelation. However, a point of change does not necessarily indicate trend reversal: it could simply be a change in the slope, as at station N7. Finally, detecting two significantly different slopes and with different signs does not necessarily imply trend reversal: one of the two sections could represent no trend. There are several R packages that estimate breakpoints (such as SiZer and segmented), and they also calculate the bootstrap confidence intervals of the slopes, but such intervals are completely meaningless for short time series, due to their considerable length. In summary, 2SLR seems to be a versatile method that is suitable for detecting trend (which implies abandoning linear or monotonic methods), but it must be studied more carefully. On the one hand, the Davies test seems to be a bit risky (a tendency toward false positives). On the other, if a breakpoint is detected, we need a good strategy for making a decision regarding the trend: a monotonic trend (slope change detected, but no trend change), no current trend, or a trend reversal. Finally, 2SLR does not take autocorrelation into account.

QR detects non-monotonic trends in the same stations as 2SLR does (Fig. 5), although in some cases the two differ in their estimates of the breakpoint (inflection point in the case of QR). Due to the concavity or convexity of a quadratic function, it sometimes seems to detect the breakpoint too earlier or too late. It does not allow



**Fig. 5** Plots showing 2SLR (solid line) and QR (dot line) fitted models, for stations N7, N8, N9, C1, C2, and C3

for assessing the trend before and after the inflection point in an easy and automatic manner: how do we assess whether the trend is actually no trend? Finally, it does not take into account autocorrelation. Despite QR exhibiting interesting and valuable behavior, we recommend applying 2SLR to detect breakpoints.

In summary, we regard it absolutely necessary to incorporate methods such as 2SLR in order to detect trend reversal. We believe excessive emphasis has been placed on the detection of monotonic trends, and the broad recommendation of the MK test is not justified. We also consider the application of a 3SLR method (three-section LR) for longtime series (at least 20 years). The question is how to decide whether one or two trend reversals are present; so we have to compare several linear models using the framework of GLM (general linear models) or other criteria such as AIC [43, 44]. In the scenario of a monotonic trend, once we have abandoned the multi-section model and trend reversal, we agree that the MK test and TS slope should be recommended. As an alternative, we suggest robust linear models, which could be a very useful alternative as they can tolerate a relatively large proportion of extreme values and overdispersion.

Autocorrelation is a sensitive and controversial issue. We consider it unresolved, since MKM test seems overly conservative and TFPW is risky. As we have seen, TFPW can lead to incorrect data correction (instead of PW, we may be inducing a spurious trend due to incorrect detrending). We recommend a priori decision criteria for the expert: if it is known that annual data autocorrelation is low, it is better not to consider its existence. If medium or high autocorrelation is possible, we need to use methods that are capable of adequate detrending or a simultaneous estimation of the trend and the autocorrelation. This is not an easy matter, and we also question the reliability (efficiency and robustness) of the estimates in both detrending and the calculation of autocorrelation when time series only contain a few data points.

## **4.2 Overall GWB Trend Assessment**

The two methods assayed produce very different results (Table 3). The method based on the *S*-mean statistic (mean of the *S* statistics from all the stations sampled from a GWB) was able to detect three significant global increasing trends (GWB21, GWB25, and GWB32), but no trend in the other three GWBs. However, the PTA method detected significant upward trends in four GWBs and two GWBs with significant upward and downward trends at the same time.

Obviously, the message sent to the management agents is quite different. See, for example, GWB48: the decision from the *S*-mean method is “no trend” (due to the conjunction of three stations showing upward trends and five showing downward trends), but the message from the PTA method is that we have both stations with significant upward trends and with significant downward trends. So we probably have two sections or areas in the GWB: one with an increasing trend and the other with a decreasing trend.

In our opinion, the *S*-mean method is only applicable if the overall GWB has the same trend; but evidently we do not know this fact a priori. So to summarize, we advise against using this approach to assess the overall GWB trend.

The PTA method appears to be a simple and effective method for global evaluation of a GWB that takes spatial correlation into account. It is necessary to consider ways to overcome the major weakness of PTA: it does not use the *p*-value of each station, that is to say, it only considers the number of stations with increasing and decreasing trends.

In short, we should abandon methods such as those that consider the mean values of all the stations. The *S*-mean takes into account spatial correlation, but it could offer a misleading assessment because an implicit assumption in this kind of technique is to consider GWB trends as homogeneous. We recommend an approach such as PTA to deal with overall GWB assessment.

Table 3 Overall GWB trend assessment

GWB	Stations (#)	Trend assessment (# and %)				GWB overall trend assessment			
		↑	→	↓	%	PTA (significance)	S-mean	Z	p
GWB06	8	3	4	1		↑ <b>0.0079</b>			
		37.5%	50.0%	12.5%			3.63	0.4513	0.6517
GWB10	22	8	12	2		<b>0.0011</b>			
		36.4%	54.5%	9.1%		0.2446	5.68	0.9178	0.3587
GWB21	10	3	6	1		<b>0.0111</b>			
		30.0%	60.0%	10.0%		0.3481	17.80	2.7299	<b>0.0063</b>
GWB25	25	5	16	4		<b>0.0165</b>			
		20.0%	64.0%	16.0%		<b>0.0456</b>	10.00	1.7285	<b>0.0839</b>
GWB32	13	4	8	1		<b>0.0058</b>			
		30.8%	61.5%	7.7%		0.4256	14.69	2.0778	<b>0.0377</b>
GWB48	15	3	7	5		<b>0.0579</b>			
		20.0%	46.7%	33.3%		<b>0.0043</b>	-11.20	-1.451	0.1469

Column 2 indicates the total number of stations within the GWB. Columns 3–5 show the number and percentage of stations with upward trends ↑, no trends →, and downward trends ↓, respectively. Overall GWB trend assessment is shown in the second half of the table, for the PTA and S-mean methods assayed; in bold numbers, we highlight significance trends



## 5 Conclusions and Future Trends

Regarding individual monitoring points, inflection points have to be identified in a proper and efficient manner. In our opinion, this is a key step in assessing groundwater trends. On the one hand, the detection of a breakpoint is necessary to demonstrate trend reversal; on the other, techniques that detect linear or monotonic trends could be considerably affected by trend reversal. It is not true that short data series solve the problem: a breakpoint in the middle of the series could affect the performance of linear or monotonic trend techniques, and it could produce misleading results. In addition, short series imply low test power. In short, we strongly advise applying methods with the capacity to identify breakpoints (such as piecewise LR).

Researchers also need to carefully consider the techniques for dealing with outliers or extreme values. In this concern, the MK test is a good method (if no breakpoint is present); but we think much more work needs to be done in exploring robust models, such as robust regression.

Another issue for future research is autocorrelation. Though much work has dealt with this matter, we consider that it is far from resolved. PW is a risky method and can even distort the data, so we recommend that the approach only be used with great care.

In relation to overall GWB trend assessment, a data aggregation method is needed. We want to emphasize the risks of methods that consider the trend as homogeneous within the GWB, because they could hide the behavior of sets of stations with opposed trends. Instead, we recommend approaches such as PTA.

Groundwater trend assessment is certainly an exciting area of research. To date, much knowledge has been gained regarding the performance of different methods, but not enough to conclude which approach is best in daily practice without analyzing every data time series individually. Nevertheless we must remember that due to the number of GWBs, stations, and pollutants, visual inspection and decisions for each data series are unrealistic. Therefore, an expert agent (software) to make decisions and apply the proper techniques in an automatic manner is absolutely essential. We are currently working on this issue.

## References

1. EC (2008) European Commission Groundwater Protection in Europe. The new groundwater directive-consolidating the EU regulatory framework, vol 35. doi: 10.2779/84304
2. Grath J, Ward R, Scheidleder A, Quevauviller P (2009) General introduction: objectives of groundwater assessment and monitoring. In: Quevauviller P, Fouillac A, Gralh J, Ward R (eds) Groundwater monitoring. Wiley, Chichester, pp 3–22
3. EU (2000) European Parliament and Council Directive 2000/60/EC of 23 October 2000 establishing a framework for Community action in the field of water policy (OJ L 327, 22/12/2000, p. 1) as amended by European Parliament and Council Decision 2455/2001/EC (OJ L 331, 15/12/2001, p. 1), European Commission, Brussels

4. EU (2006) European Parliament and Council of the European Union European Parliament and Council Directive of 12 December 2006 on the protection of groundwater against pollution and deterioration (2006/118/EC). Offi J Eur Commun L372/19
5. Hirsch RM, Slack JR, Smith RA (1982) Techniques of trend analysis for monthly water quality data. *Water Resour Res* 18:107. doi:[10.1029/WR018i001p00107](https://doi.org/10.1029/WR018i001p00107)
6. Hirsch RM, Slack JR (1984) A nonparametric trend test for seasonal data with serial dependence. *Water Resour Res* 20:727
7. Broers HP, Visser A, Chilton JP, Stuart ME (2009) Assessing and aggregating trends in groundwater quality. In: Quevauviller P, Fouillac A, Gralh J, Ward R (eds) *Groundwater monitoring*. Wiley, Chichester, pp 189–206
8. Grath J, Scheidleder A, Uhlig S et al (2001) The EU water framework directive: statistical aspects of the identification of groundwater pollution trends, and aggregation of monitoring results. Final Report. 63
9. EC (2009) European Commission Environment Common implementation strategy guidance for the water framework strategy (2000/60/2006). Guidance Document No 18. Guidance on groundwater status and trend assessment, vol 82
10. R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.r-project.org/>
11. Lettenmaier DP, Conquest LL, Hughes JP (1982) Routine streams and rivers water quality trend monitoring review. Charles W. Harris Hydraulics Laboratory, Technical Report No. 75. University of Washington
12. Van Belle G, Hughes JP (1984) Nonparametric tests for trend in water quality. *Water Resour Res* 20:127–136
13. Helsel DR (1987) Advantages of nonparametric procedures for analysis of water quality data. *Hydrol Sci J* 32:179–190. doi:[10.1080/02626668709491176](https://doi.org/10.1080/02626668709491176)
14. Lettenmaier DP (1988) Multivariate nonparametric tests for trend in water quality. *Water Resour Bull* 24:505–512. doi:[10.1111/j.1752-1688.1988.tb00900.x](https://doi.org/10.1111/j.1752-1688.1988.tb00900.x)
15. Hirsch RM, Alexander RB, Smith RA (1991) Selection of methods for the detection and estimation of trends in water quality. *Water Resour Res* 27:803–813
16. McLeod AI, Hipel KW, Bodo BA (1991) Trend analysis methodology for water quality time series. *Environmetrics* 2:169–200. doi:[10.1002/env.3770020205](https://doi.org/10.1002/env.3770020205)
17. Esterby SR (1993) Trend analysis methods for environmental data. *Environmetrics* 4:459–481
18. Esterby S (1996) Review of methods for the detection and estimation of trends with emphasis on water quality applications. *Hydrol Process* 10:127–149
19. Reckhow KH, Kepford K, Hicks WW (1993) Statistical methods for the analysis of lake water quality trends. Report 841-R-93-003
20. Loftis J (1996) Trends in groundwater quality. *Hydrol Process* 10:335–355
21. Hess A, Iyer H, Malm W (2001) Linear trend analysis: a comparison of methods. *Atmos Environ* 35:5211–5222. doi:[10.1016/S1352-2310\(01\)00342-9](https://doi.org/10.1016/S1352-2310(01)00342-9)
22. Helsel DR, Hirsch RM (2002) Trend analysis. In: *Statistical methods in water resources. Techniques of Water Resources Investigations, Book 4, chapter A3*. U.S. Geological Survey, pp 323–355
23. Kundzewicz ZW, Robson AJ (2004) Change detection in hydrological records—a review of the methodology / Revue méthodologique de la détection de changements dans les chroniques hydrologiques. *Hydrol Sci J* 49:7–19. doi:[10.1623/hysj.49.1.7.53993](https://doi.org/10.1623/hysj.49.1.7.53993)
24. Helsel DR, Frans LM (2006) Regional Kendal test for trend. *Environ Sci Technol* 40(13): 4066–4073
25. Chang H (2008) Spatial analysis of water quality trends in the Han River basin, South Korea. *Water Res* 42:3285–3304. doi:[10.1016/j.watres.2008.04.006](https://doi.org/10.1016/j.watres.2008.04.006)
26. Visser A, Dubus I, Broers HP et al (2009) Comparison of methods for the detection and extrapolation of trends in groundwater quality. *J Environ Monit* 11:2030–2043. doi:[10.1039/b905926a](https://doi.org/10.1039/b905926a)

27. NNPSMP (2011) Statistical analysis for monotonic trends, Tech Notes 6. Office of Water US Environmental Protection Agency
28. Machiwal D, Jha MK (2012) Trend and homogeneity in subsurface hydrologic variables: case study in a hard-rock aquifer of western India. In: Machiwal D, Jha MK (eds) Hydrologic time series analysis: theory and practice. Springer, Netherlands, pp 165–180. doi:[10.1007/978-94-007-1861-6\\_8](https://doi.org/10.1007/978-94-007-1861-6_8)
29. Wahlén K, Grimvall A (2010) Roadmap for assessing regional trends in groundwater quality. *Environ Monit Assess* 165:217–231. doi:[10.1007/s10661-009-0940-7](https://doi.org/10.1007/s10661-009-0940-7)
30. Devreker D, Lefebvre A (2014) TTAinterfaceTrendAnalysis: an R GUI for routine temporal trend analysis and diagnostics. *J Oceanogr Res Data* 7:1
31. Huber PJ (1981) Robust statistics. Wiley, New York. doi:[10.1002/0470010940](https://doi.org/10.1002/0470010940)
32. Hampel FR, Rousseeuw PJ, Ronchetti E, Stahel W (1986) Robust statistic: the approach based on influence functions. Wiley, New York
33. Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
34. Marazzi A (1993) Algorithms, routines, and S functions for robust statistics. CRC, New York
35. Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. doi:[10.2307/2685660](https://doi.org/10.2307/2685660)
36. Bellio R, Ventura L (2005) An introduction to robust estimation with R functions. *Proceedings of 1st International Work*, pp 1–57
37. Huber PJ (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *Ann Stat* 1: 799–821
38. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74:829–836
39. Cleveland WS, Devlin SJ, Grosse E (1988) Regression by local fitting. *J Econ* 37:87–114
40. Cleveland WS, Grosse E (1991) Computational methods for local regression. *Stat Comput* 1:47–62
41. Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc* 83:596–610. doi:[10.2307/2289282](https://doi.org/10.2307/2289282)
42. Hurvich CM, Simonoff JS, Tsai CL (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike Information Criterion. *J R Stat Soc Ser B* 60:271–293
43. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) Second international symposium on information theory, Tsahkadsor, Armenia, USSR, September 2–8, 1971, Akadémiai Kiadó, Budapest, pp 267–281
44. Akaike H (1974) A new look at the statistical model identification. *Autom Contr IEEE Trans* 19:716–723. doi:[10.1109/TAC.1974.1100705](https://doi.org/10.1109/TAC.1974.1100705)
45. Wen F, Chen X (2006) Evaluation of the impact of groundwater irrigation on streamflow in Nebraska. *J Hydrol* 327:603–617. doi:[10.1016/j.jhydrol.2005.12.016](https://doi.org/10.1016/j.jhydrol.2005.12.016)
46. Hirsch RM (1988) Statistical methods and sampling design for estimating step trends in surface water quality. *Water Resour Bull* 24:493–503. doi:[10.1111/j.1752-1688.1988.tb00899.x](https://doi.org/10.1111/j.1752-1688.1988.tb00899.x)
47. Hirsch RM, Gilroy EJ (1985) Detectability of step trends in the rate of atmospheric deposition of sulfate. *Water Resour Bull* 21:773–784. doi:[10.1111/j.1752-1688.1985.tb00171.x](https://doi.org/10.1111/j.1752-1688.1985.tb00171.x)
48. Kundzewicz ZW, Robson AJ (eds) (2000) Detecting trend and other changes in hydrological data. World Climate Programme—Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD no. 1013. World Meteorological Organization, Geneva, Switzerland
49. Seber GAF, Wild CJ (1989) Nonlinear regression. Wiley, New York. doi:[10.1002/0471725315](https://doi.org/10.1002/0471725315)
50. Toms JD, Lesperance ML (2003) Piecewise regression: a tool for identifying ecological thresholds. *Ecology* 84:2034–2041
51. Steele TD, Gilroy EJ, Hawkinson RO (1974) Techniques for the assessment of areal and temporal variations in streamflow quality. Open File Report, U.S. Geol. Surv., Washington, D.C.

52. Cailas MD, Cavadias G, Gehr R (1986) Application of a nonparametric approach for monitoring and detecting trends in water quality data of the St. Lawrence River. *Water Pollut Res J Can* 21(2):153–167
53. Berryman D, Bobée B, Haemmerli J (1988) Nonparametric tests for trend detection in water quality time series. *J Am Water Resour Assoc* 24:545–556
54. Taylor CH, Loftis JC (1989) Testing for trend in lake and groundwater quality time series. *J Am Water Resour Assoc* 25:715–726. doi:[10.1111/j.1752-1688.1989.tb05385.x](https://doi.org/10.1111/j.1752-1688.1989.tb05385.x)
55. Zetterqvist L (1991) Statistical estimation and interpretation of trends in water quality time series. *Water Resour Res* 27(7):1944–1973. doi:[10.1029/91WR00478](https://doi.org/10.1029/91WR00478)
56. Yu Y, Zou S, Whittemore D (1993) Non-parametric trend analysis of water quality data of rivers in Kansas. *J Hydrol* 150:61–80
57. Lettenmaier DP, Wood EF, Wallis JR (1994) Hydro-climatological trends in the continental United States, 1948–88. *J Clim* 7:586–607. doi:[10.1175/1520-0442\(1994\)007<0586:HCTITC>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<0586:HCTITC>2.0.CO;2)
58. Gan TY (1998) Hydroclimatic trends and possible climatic warming in the Canadian Prairies. *Water Resour Res* 34:3009–3015
59. Cun C, Vilagines R (1997) Time series analysis on chlorides, nitrates, ammonium and dissolved oxygen concentrations in the Seine river near Paris. *Sci Total Environ* 208:59–69
60. Douglas E, Vogel R, Kroll C (2000) Trends in floods and low flows in the United States: impact of spatial correlation. *J Hydrol* 240:90–105
61. Burn DH, Hag Elnur MA (2002) Detection of hydrologic trends and variability. *J Hydrol* 255:107–122. doi:[10.1016/S0022-1694\(01\)00514-5](https://doi.org/10.1016/S0022-1694(01)00514-5)
62. Broers HP, van der Grift B (2004) Regional monitoring of temporal changes in groundwater quality. *J Hydrol* 296:192–220. doi:[10.1016/j.jhydrol.2004.03.022](https://doi.org/10.1016/j.jhydrol.2004.03.022)
63. Mann HB (1945) Non-parametric test against trend. *Econometrica* 13:245–259
64. Kendall MG (1938) A new measure of rank correlation. *Biometrika* 30:81–93. doi:[10.2307/2332226](https://doi.org/10.2307/2332226)
65. Kendall MG (1962) Rank correlation methods, 3rd edn. Hafner, New York
66. Kendall M (1975) Multivariate methods. Charles Griffin, London
67. Hamed KH (2008) Trend detection in hydrologic data: the Mann–Kendall trend test under the scaling hypothesis. *J Hydrol* 349:350–363. doi:[10.1016/j.jhydrol.2007.11.009](https://doi.org/10.1016/j.jhydrol.2007.11.009)
68. Gan TY (1992) Finding trends in air temperature and precipitation for Canada and North-eastern United States. In: Kite GW, Harvey KD (eds) Using hydrometric data to detect and monitor climatic change. Proceedings of NHRI Workshop No. 8, National Hydrology Research Institute, Saskatoon, SK, pp 57–78
69. Kahya E, Kalaycı S (2004) Trend analysis of streamflow in Turkey. *J Hydrol* 289:128–144. doi:[10.1016/j.jhydrol.2003.11.006](https://doi.org/10.1016/j.jhydrol.2003.11.006)
70. Sneyers R (1990) On the statistical analysis of series of observations. World Meteorological Organization, Technical Note 143, Geneva, Switzerland
71. Partal T, Kahya E (2006) Trend analysis in Turkish precipitation data. *Hydrol Process* 20: 2011–2026. doi:[10.1002/hyp.5993](https://doi.org/10.1002/hyp.5993)
72. Shifteh Some'e B, Ezani A, Tabari H (2012) Spatiotemporal trends and change point of precipitation in Iran. *Atmos Res* 113:1–12. doi:[10.1016/j.atmosres.2012.04.016](https://doi.org/10.1016/j.atmosres.2012.04.016)
73. Theil H (1950) A rank-invariant method of linear and polynomial regression analysis. *NederlAkadWetensch Proc* 53:386–392 (Part I), 521–525 (Part II), 1397–1412 (Part III)
74. Sen PK (1968) Estimates of the regression coefficient based on Kendall's Tau. *J Am Stat Assoc* 63:1379–1389
75. Hamed KH (2014) The distribution of Spearman's rho trend statistic for persistent hydrologic data. *Hydrol Sci J*. doi:[10.1080/02626667.2014.968573](https://doi.org/10.1080/02626667.2014.968573)
76. Bodenhofer U, Klawonn F (2008) Robust rank correlation coefficients on the basis of fuzzy orderings: initial steps. An overview of rank correlation measures. *Mathwa Soft Comput* 15:5–20

77. Bodenhofer U, Krone M, Klawonn F (2013) Testing noisy numerical data for monotonic association. *Inf Sci* 245:21–37. doi:[10.1016/j.ins.2012.11.026](https://doi.org/10.1016/j.ins.2012.11.026)
78. Bodenhofer U, Krone M (2013) RoCoCo an R package implementing a robust rank correlation coefficient and a corresponding test, software manual. Institute of Bioinformatics, Johannes Kepler University Linz
79. Von Storch H (1995) Misuses of statistical analysis in climate research. In: von Storch H, Navarra A (eds) *Analysis of climate variability SE-2*. Springer, Berlin, pp 11–26
80. Yue S, Pilon P, Phinney B, Cavadias G (2002) The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrol Process* 16:1807–1829. doi:[10.1002/hyp.1095](https://doi.org/10.1002/hyp.1095)
81. Yue S, Wang CY (2002) The influence of serial correlation on the Mann–Whitney test for detecting a shift in median. *Adv Water Resour* 25:325–333. doi:[10.1016/S0309-1708\(01\)00049-5](https://doi.org/10.1016/S0309-1708(01)00049-5)
82. Yue S, Pilon P (2004) A comparison of the power of the t test, Mann–Kendall and bootstrap tests for trend detection. *Hydrol Sci J* 49(1):21–37
83. Önöz B, Bayazit M (2011) Block bootstrap for Mann–Kendall trend test of serially dependent data. *Hydrol Process* 26:1–19
84. Yue S, Wang CY (2004) The Mann–Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. *Water Resour Manag* 18:201–218. doi:[10.1023/B:WARM.0000043140.61082.60](https://doi.org/10.1023/B:WARM.0000043140.61082.60)
85. Khaliq MN, Ouarda TBMJ, Gachon P et al (2009) Identification of hydrological trends in the presence of serial and cross correlations: a review of selected methods and their application to annual flow regimes of Canadian rivers. *J Hydrol* 368:117–130. doi:[10.1016/j.jhydrol.2009.01.035](https://doi.org/10.1016/j.jhydrol.2009.01.035)
86. Hamed KH (2009) Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. *J Hydrol* 368:143–155. doi:[10.1016/j.jhydrol.2009.01.040](https://doi.org/10.1016/j.jhydrol.2009.01.040)
87. Yue S, Wang CY (2002) Applicability of prewhitening to eliminate the influence of serial correlation on the Mann–Kendall test. *Water Resour Res* 38:4–1–4–7. doi:[10.1029/2001WR000861](https://doi.org/10.1029/2001WR000861)
88. Miao C, Ni J (2010) Implement of filter to remove the autocorrelation’s influence on the Mann–Kendall test: a case in hydrological series. *Int J Food Agric* 8:1241–1246
89. Matalas NC, Sankarasubramanian A (2003) Effect of persistence on trend detection via regression. *Water Resour Res* 39. doi:[10.1029/2003WR002292](https://doi.org/10.1029/2003WR002292)
90. Yue S, Pilon P (2003) Interaction between deterministic trend and autoregressive process. *Water Resour Res* 39. doi:[10.1029/2001WR001210](https://doi.org/10.1029/2001WR001210)
91. Perron P, Yabu T (2009) Estimating deterministic trends with an integrated or stationary noise component. *J Econ* 151:56–69
92. Salmi T, Maatta A, Anttila P, Ruoho-Airola T, Amnell T (2002) Detecting trends of annual values of atmospheric pollutants by the Mann–Kendall test and Sen’s slope estimates. *Publ Air Qual* 31:1–35
93. Tabari H, Talaei PH (2011) Analysis of trends in temperature data in arid and semi-arid regions of Iran. *Glob Planet Change* 79:1–10. doi:[10.1016/j.gloplacha.2011.07.008](https://doi.org/10.1016/j.gloplacha.2011.07.008)
94. U.S. EPA (2009) US Environmental Protection Agency. Statistical analysis of groundwater monitoring data at RCRA facilities. Unified Guidance. Appendix C3. EPA 530/R-09-007
95. Arnau J, Bono R (2001) Autocorrelation and bias in short time series: an alternative estimator. *Qual Quant* 365–387
96. Arnau J, Bono R (2002) A program to calculate the empirical bias in autocorrelation estimators. *Psicothema* 14:669–672
97. Kan R, Wang X (2010) On the distribution of the sample autocorrelation coefficients. *J Econ* 154:101–121. doi:[10.1016/j.jeconom.2009.06.010](https://doi.org/10.1016/j.jeconom.2009.06.010)
98. Solanas A, Manolov R, Sierra V (2010) Lag-one autocorrelation in short series: estimation and hypotheses testing. *Int J Method Exp Psychol* 31(2):357–381
99. Fuller WA (1976) *Introduction to statistical time series*. Wiley, New York
100. Moran PAP (1948) The interpretation of statistical maps. *J R Stat Soc Ser B* 3:243–251

101. Huitema BE, McKean JW (1991) Autocorrelation estimation and inference with small samples. *Psychol Bull* 110:291–304. doi:[10.1037//0033-2909.110.2.291](https://doi.org/10.1037//0033-2909.110.2.291)
102. Kendall M, Ord JK (1990) *Time series*. Edward Arnold, London
103. Decarlo LT, Tryon WW (1993) Estimating and testing autocorrelation with small samples: a comparison of the c-statistic to a modified estimator. *Behav Res Ther* 31:781–788
104. Matyas TA, Greenwood KM (1991) Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behav Assess* 13:137–157
105. Matalas NC, Langbein WB (1962) Information content of the mean. *J Geophys Res* 67: 3441–3448. doi:[10.1029/JZ067i009p03441](https://doi.org/10.1029/JZ067i009p03441)
106. Lettenmaier DP (1976) Detection of trends in water quality data from records with dependent observations. *Water Resour Res* 12:1037–1046. doi:[10.1029/WR012i005p01037](https://doi.org/10.1029/WR012i005p01037)
107. Hamed KH, Rao AR (1998) A modified Mann-Kendall trend test for autocorrelated data. *J Hydrol* 204:182–196. doi:[10.1016/S0022-1694\(97\)00125-X](https://doi.org/10.1016/S0022-1694(97)00125-X)
108. Bayley G, Hammersley J (1946) The “effective” number of independent observations in an autocorrelated time series. *J R Stat Soc* 8:184–197
109. Yue S, Pilon P, Phinney B (2003) Canadian streamflow trend detection: impacts of serial and cross-correlation. *Hydrol Sci J* 48:51–63
110. Aziz OI, Burn DH (2006) Trends and variability in the hydrological regime of the Mackenzie River Basin. *J Hydrol* 319:282–294. doi:[10.1016/j.jhydrol.2005.06.039](https://doi.org/10.1016/j.jhydrol.2005.06.039)
111. Novotny EV, Stefan HG (2007) Stream flow in Minnesota: indicator of climate change. *J Hydrol* 334:319–333. doi:[10.1016/j.jhydrol.2006.10.011](https://doi.org/10.1016/j.jhydrol.2006.10.011)
112. Kumar S, Merwade V, Kam J, Thurner K (2009) Streamflow trends in Indiana: effects of long term persistence, precipitation and subsurface drains. *J Hydrol* 374:171–183. doi:[10.1016/j.jhydrol.2009.06.012](https://doi.org/10.1016/j.jhydrol.2009.06.012)
113. Oguntunde PG, Abiodun BJ, Lischeid G (2011) Rainfall trends in Nigeria, 1901–2000. *J Hydrol* 411:207–218. doi:[10.1016/j.jhydrol.2011.09.037](https://doi.org/10.1016/j.jhydrol.2011.09.037)
114. Tabari H, Abghari H, Hosseinzadeh Talaei P (2012) Temporal trends and spatial characteristics of drought and rainfall in arid and semiarid regions of Iran. *Hydrol Process* 26: 3351–3361. doi:[10.1002/hyp.8460](https://doi.org/10.1002/hyp.8460)
115. Bayazit M, Önöz B (2007) To prewhiten or not to prewhiten in trend analysis? *Hydrol Sci J* 52:611–624. doi:[10.1623/hysj.52.4.611](https://doi.org/10.1623/hysj.52.4.611)
116. Hamed KH (2008) Discussion of “To prewhiten or not to prewhiten in trend analysis?”. *Hydrol Sci J* 53:667–668. doi:[10.1623/hysj.53.3.667](https://doi.org/10.1623/hysj.53.3.667)
117. Salas-La Cruz JD (1972) Information content of the regional mean, vol 2. In: *Proceedings of the international symposium on uncertainties in hydrologic and water resource systems*. University of Arizona, Tucson, pp 646–660
118. McLeod AI (2011) Kendall: Kendall rank correlation and Mann-Kendall trend test. R package version 2.2. <http://CRAN.R-project.org/package=Kendall>
119. Komsta L (2013) mblm: Median-Based Linear Models. R package. version 0.12. <http://CRAN.R-project.org/package=mbm>
120. Sonderegger D (2012) SiZer: SiZer: significant zero crossings. R package version 0.1-4. <http://CRAN.R-project.org/package=SiZer>
121. Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74:33–43
122. Muggeo VMR (2008) Segmented: an R package to fit regression models with broken-line relationships. *R News* 8:20–25

Experiences from Ground, Coastal and Transitional  
Water Quality Monitoring  
The EU Water Framework Directive Implementation in  
the Catalan River Basin District (Part II)  
Munné, A.; Ginebreda, A.; Prat, N. (Eds.)  
2016, XX, 339 p., Hardcover  
ISBN: 978-3-319-23903-3