Position paper: Common mistakes and solutions for a better use of correlation- and regression-based approaches in environmental sciences

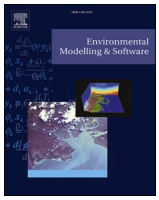
Damien Tedoldi, Boram Kim, Santiago Sandoval, Nicolas Forquet, Bruno Tassin

Environmental Modelling and Software 192 (2025) 106526

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



Position paper: Common mistakes and solutions for a better use of correlation- and regression-based approaches in environmental sciences

Damien Tedoldi [a,*], Boram Kim [a], Santiago Sandoval [b], Nicolas Forquet [c], Bruno Tassin [b]

[a] INSA Lyon, DEEP, UR7429, Villeurbanne, 69621, France
[b] LEESU, ENPC, Institut Polytechnique de Paris, Univ Paris Est Créteil, Champs-sur-Marne, France
[c] INRAE, UR REVERSAAL, Villeurbanne, 69625, France

ARTICLE INFO

Keywords:
Bivariate analysis
Data-driven modelling
Empirical modelling
Good practices
Linear regression
Statistical testing

ABSTRACT

While empirical modelling remains a popular practice in environmental sciences, an alarming number of misuses of correlation- and regression-based techniques are encountered in recent research, although these techniques are described in courses and textbooks. This position paper reviews the most common issues, and provides theoretical background for understanding the interests and limitations of these methods, based on their underlying assumptions. We call for a reconsideration of misleading practices, including: the application of linear regression to data points that do not display a linear pattern, the failure to pinpoint influential points, the inappropriate extrapolation of empirical relationships, the overrated search for "statistical significance", the pooling of data belonging to different populations, and, most importantly, calculations without data visualization. We urge reviewers to be vigilant on these aspects. We also recall the existence of alternative approaches to overcome the highlighted shortcomings, and thus contribute to a more accurate interpretation of the results.

1. Introduction

> "*Regarde de tous tes yeux, regarde !*" ["Look with all your eyes, look!"]
> — Jules Verne, Michel Strogoff, 1876.

Many environmental processes are prone to significant variability, multiple couplings, feedbacks and stochastic forcings, resulting in high complexity that often jeopardizes the predictive capacity of deterministic, physically-based models (Wainwright and Mulligan, 2004; Young et al., 1996). Additionally, environmental measurement, in the broadest sense – monitoring, sampling, analysis –, often entails important challenges: significant costs, time constraints and inherent non-controlled conditions of *in situ* measurements may result in scarce and/or low-quality data with few replicates, despite recent developments in high-frequency monitoring strategies (Razguliaev et al., 2024; Smits et al., 2025). This may help understand the widespread use of empirical approaches in environmental studies, in order to relate two (or more) quantities that happen to co-fluctuate. In general, correlation- and regression-based techniques aim to go beyond, and possibly consolidate, the (typically graphical) evidence of a certain association between variables, with different objectives: from the quantification of the degree of association (*e.g.* through a correlation coefficient) to the development of a predictive model. A series of methods – which, like any statistical model, are underpinned by their own fundamental assumptions – have been developed for the foregoing purposes, and are now widespread and implemented in most existing tools for data analysis.

The genesis of this paper stems from a concerning observation exposed in the initial section: in peer-reviewed articles published over the last two decades in environmental sciences, there appear to be increasing cases of misuse of such techniques, which disregard their mathematical properties and underlying assumptions. In so doing, the main risk is to draw potentially erroneous conclusions – *e.g.*, to seemingly identify a "significant" association between two independent variables – from the application of an inadequate approach as a methodological backbone. The problem is not new, nor is its recognition, and pleas for an enlightened use of regression can be found in works such as Deming (1943), Tomassone et al. (1983), Helsel and Hirsch (2002) and Berthouex and Brown (2002). What is novel, however, is the unprecedented scale of the issue – exacerbated by the recent massification of data combined with the development of easy-to-run software for data analysis – ironically associated to the growing idea that correlation- and regression-based techniques are "basic statistics". As summarized by Mikkonen et al. (2019), "this simplicity is deceptive as the line-fitting

* Corresponding author.
E-mail address: damien.tedoldi@insa-lyon.fr (D. Tedoldi).

https://doi.org/10.1016/j.envsoft.2025.106526
Received 15 April 2025; Received in revised form 11 May 2025; Accepted 16 May 2025
Available online 19 May 2025

procedure is actually quite a complex problem".

Obviously, the intention of this paper is not to stigmatize specific studies by pointing out their shortcomings. Our main objective is rather to raise awareness among environmental scientists – whether they are contributing as article authors or acting as reviewers – about the inherent limitations of commonly applied correlation- and regression-based techniques, to outline key caveats, and to discuss possible ways to overcome the encountered limitations. Following an initial part that sets out the scope and scale of the problem, the article is structured around a series of fundamental assertions, illustrated by several examples and detailed by a few theoretical considerations. The purpose of these theoretical elements is not to compile yet another volume on statistics, but to offer the reader the analytical means to precisely grasp the problematic nature of the practices examined herein as well as the relevance of the proposed improvements.

2. Problem statement

To outline the addressed issue, we have first compiled eight graphical examples of statistical shortfalls published between 2017 and 2023 in well-established journals in environmental sciences (Fig. 1): these graphs were redrawn to ensure anonymity, but we keep their references available for skeptical readers.

The problems embodied by these graphs may be categorized as follows.

1. Application of a linear regression and/or calculation of a Pearson correlation coefficient for data points that visibly show a nonlinear pattern (Fig. 1A);
2. Presence of an influential point which alone is responsible for the modeled "trend" as well as its "statistical significance" (Fig. 1B and C) – even if the rest of the data points clearly indicate an absence of trend (Fig. 1B);
3. Violation of the normality assumption for correlation testing (Fig. 1B, C and G);
4. Violation of fundamental assumptions on regression residuals: homoscedasticity (Fig. 1D) and normality (Fig. 1E);
5. Pooling of data points that belong to different populations (Fig. 1F, and presumably Fig. 1G);
6. Extrapolation of a linear model outside of the range of available data (Fig. 1F and G) – even reaching non-physical values (Fig. 1F);
7. Interpolation of isolated clusters of data points, without any evidence of the intermediate trend (Fig. 1C and H).

Most importantly, they were treated as "positive" results (as reflected by the presence of a p-value on many of the graphs or in their caption), *i.e.* the authors concluded that there is a relevant association between the two variables under study, and/or that a linear trend is appropriate to describe their relationship. The fact that these glaring shortcomings did not elicit sufficient objection on the part of co-authors, editors, and reviewers, to prevent their publication, is already a relatively disturbing observation.

To demonstrate that the problem extends far beyond this set of eight studies, while avoiding confirmation bias, a systematic literature search was carried out. The topic of microplastics was selected as a currently very active scientific field, and this keyword was searched on Web of Science in association with the keywords "correlation" or "influencing factor" (the search was done between September and October 2024). Each article, in order of appearance according to the "relevance" sorting criterion, from a journal with an impact factor greater than 2.5, was scanned to check that the content or supplementary material did present a correlation or regression analysis. The search was stopped at 100 articles meeting these criteria: the earliest and median publication dates were 2017 and 2023, respectively; the median impact factor (in 2024) of the journals from which they were extracted was 7.6. These 100 articles were then meticulously examined to check whether the statistical methods are appropriate, or whether they exhibit one or several of the above-mentioned flaws.

The number of publications without any identified problem relating to correlations and regressions amounts to 25 out of 100. Conversely, 55 articles present *at least* one patent shortcoming, the most frequent being: linear regression/calculation of Pearson correlation coefficient on data points with no linear pattern (23), "statistical significance" due to the presence of an influential point (20), inference in linear regression with non-Gaussian or heteroscedastic residuals (13), and extrapolation or interpolation issues (13). In the remaining 20 articles, the details provided do not allow a definite assessment of the appropriateness of the methodology. Typical examples of this situation include the following: the nature of the coefficients presented in correlation matrices – Pearson, Spearman, or any other – is not specified; no normality test is mentioned prior to testing the significance of Pearson correlation coefficient, and the data are not available; or the sample size on which a correlation test was performed is not specified, while the data include repeated measurements on the same sites. A more subtle shortcoming, not quantified here, pertains to publications that merely comment on whether or not a correlation coefficient is "significant", without any consideration for the value of the coefficient itself.

We should certainly be alerted by the fact that more than half, and potentially up to 3/4, of the "most relevant articles" on microplastics and their links to other environmental factors, have based their conclusions on flawed statistical methods. In particular, this shows that the existence and availability of statistics manuals is not a sufficient bulwark against these kinds of misleading practices. Hence, with this position paper, we hope to contribute to a more informed use of empirical modelling among environmental scientists and, in turn, to help reduce the prevalence of statistical errors in published research.

In the subsequent developments, explanations will be illustrated using examples generated with random sequence simulators, designed to reflect actual configurations encountered in published literature. The reader will note that these examples are representative of the patterns shown in Fig. 1 and have not been artificially exaggerated. Synthetic data have the pedagogical advantage of showing that recreating datasets such as those depicted in Fig. 1 involves the violation of at least one of the statistical assumptions underpinning usual correlation analysis and/or linear regression. Another advantage is reproducibility: for readers willing to replicate the figures and calculations presented below, R scripts (R Core Team, 2021) are provided in the supplementary material accompanying this article.

3. Influential points

3.1. Quadratic and variance-based criteria are sensitive to influential points

Consider the following two examples, illustrated on Fig. 2.

Example A

X and Y are independent random variables. Let us assume that under usual environmental conditions, each of them follows a normal distribution, with a mean value of $\mu_X = 5$ and $\mu_Y = 10$, and a standard deviation of $\sigma_X = 2$ and $\sigma_Y = 3$, respectively. Evidently, the nature and characteristics of these distributions are unknown to the experimenter.

Joint measurements of X and Y under standard conditions lead to samples x_0 and y_0 with a sample size $n = 15$. Then an additional observation shows a large deviation from the common response of X and Y, both of which are significantly increased: $x_1 = 20$ and $y_1 = 30$. A discerning observer would readily recognize that this new point does not belong to the same underlying population. Nevertheless, as in many examples mentioned in Section 2, we consider a situation where the data are treated as such in order to explore potential correlation/trend, using the Pearson correlation coefficient and linear regression.

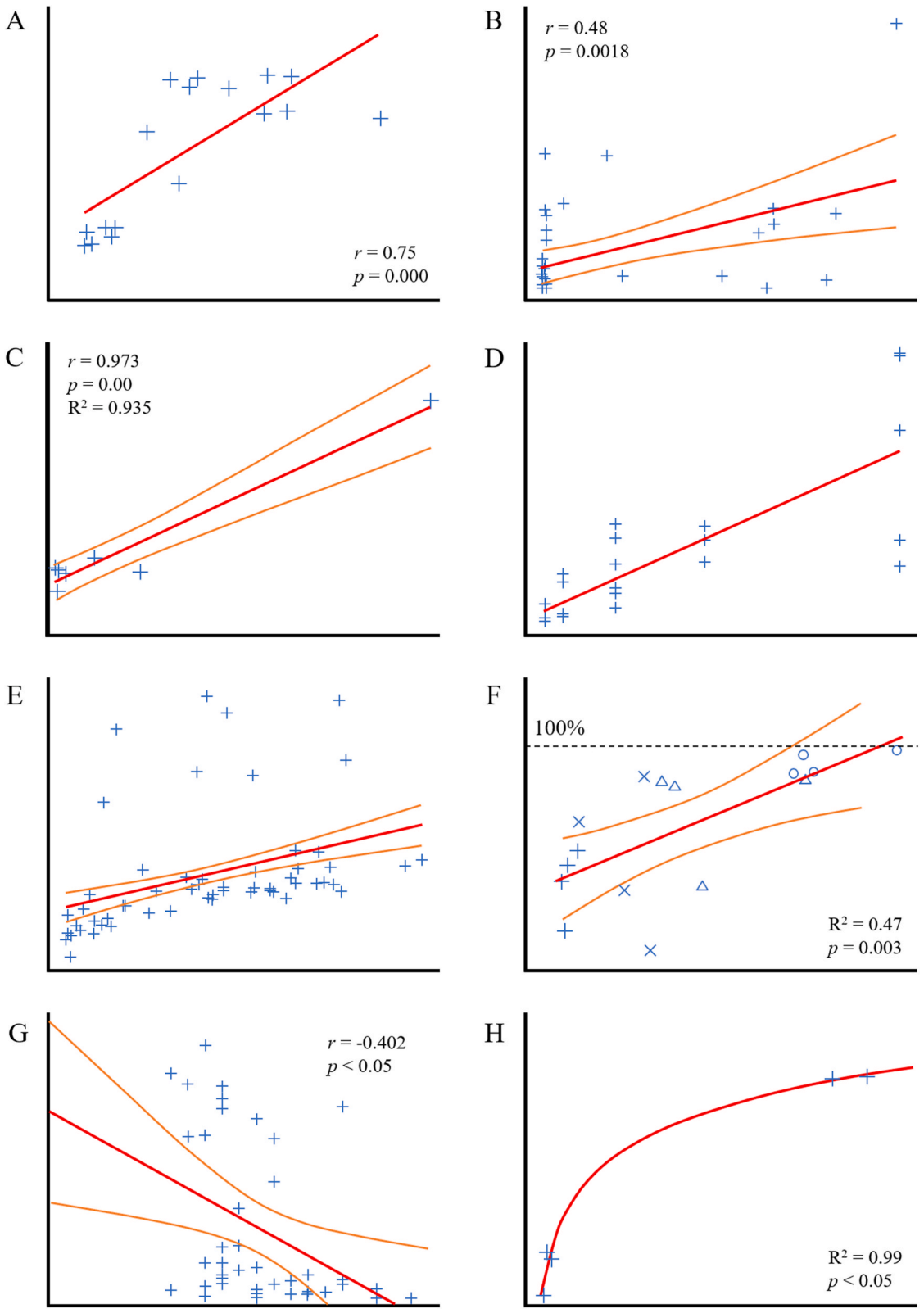


Fig. 1. Examples of misleading use of correlation- or regression-based techniques taken from scientific literature (A: Journal of Hazardous Materials; B: Journal of Exposure Science & Environmental Epidemiology; C: Environmental Pollution; D: Nature Communications Earth & Environment; E: Science of the Total Environment; F: Frontiers in Environmental Science; G: Environmental Science and Pollution Research; H: Environmental Science & Technology). To ensure anonymity, these figures have been redrawn and the legends and labels of the axes have not been included. The symbols correspond to the data points, the red line/curve to the least-squares regression model, and the orange curves, where present, to the 95 % confidence intervals on the regression line. On figure F, the different symbols correspond to different populations. When included on the original figures or in their captions, the notations r, p, and R^2 refer respectively to the Pearson correlation coefficient, the p-value associated with the test of significance of this correlation, and the coefficient of determination of the linear regression model.

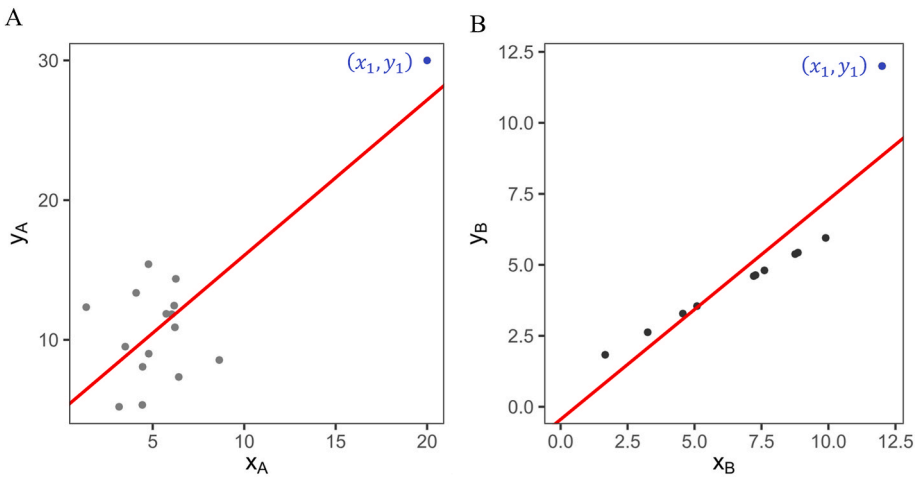


Fig. 2. Graphical illustration of Example A (left) and Example B (right). In Example A, while by construction there is no association between x_0 and y_0 (Pearson's $r = 0.05$ for the data points above), the addition of one point to the data series results in a much higher, and seemingly "significant", Pearson correlation coefficient ($r = 0.79$), as well as a positive slope of the regression line. The problem is different in Example B, where the two variables do have a linear association within a certain range, but the additional measurement results in a regression line that matches neither the linear portion of the data points nor the entire data series.

Example B

Consider two variables X and Y. In the usual range of variation of X, the two variables are linked by a deterministic linear relationship, whose coefficients are unknown to the experimenter. In this example, X is usually in $[0, 10]$, in which it follows a uniform distribution and $Y = 0.5X + 1$.

In order to characterize the relationship between X and Y, 10 joint measurements (x_0 and y_0) have been carried out in standard conditions, and an additional measurement has been taken outside this range, leading to $x_1 = 12$ and $y_1 = 12$. Linear regression is then applied to the whole dataset.

In both situations, the point (x_1, y_1) is called an *influential point*, since it has disproportionate effects on the correlation/regression results. This influence directly derives from the mathematical properties of usual correlation- and regression-based techniques. Generally, the indicator that is estimated by default when calling a "correlation" function on a statistical software is the Pearson correlation coefficient, r_{XY}. Readers will likely be familiar with the formula used to estimate this coefficient from n joint observations of X and Y:

$$r = \frac{\sum_{i=1}^{n} \left(x_{obs,i} - \bar{x}\right) \cdot \left(y_{obs,i} - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_{obs,i} - \bar{x}\right)^2 \cdot \sum_{i=1}^{n} \left(y_{obs,i} - \bar{y}\right)^2}} \tag{1}$$

where $x_{obs,i}$ (resp. $y_{obs,i}$) is the i^{th} observed value of the variable X (resp. Y), and $\bar{x}$ and $\bar{y}$ are their respective mean values. Less commonly acknowledged is apparently the sensitivity of this measure to values that deviate significantly from the center of the distribution. By construction, r tends towards $\frac{(x_1 - \bar{x}) \cdot (y_1 - \bar{y})}{|x_1 - \bar{x}| \cdot |y_1 - \bar{y}|} = \pm 1$ as the influential point (x_1, y_1) moves further away from the rest of the data points in both directions (like in Example A).

A similar effect occurs when computing the a and b coefficients of a regression line as *least-squares* estimates, which minimize the following criterion:

$$\sum_{i=1}^{n} \left(y_{obs,i} - (a x_{obs,i} + b)\right)^2 \tag{2}$$

Squaring the residuals inherently magnifies large differences, and therefore tends to discard models that deviate substantially from one or a few influential point(s). Hence, in Example B, the regression line displayed on Fig. 2B is considered a "better fit" than the line $Y = 0.5X + 1$ (*i.e.* the truly linear trend observed in the interval $[0, 10]$), which is penalized by the squared difference $(y_1 - (0.5x_1 + 1))^2$.

In many cases, such as those shown on Fig. 1B-C and Fig. 2A-B, preliminary visual assessment should be quite sufficient to identify influential points. This may be complemented by usual outlier detection techniques to reinforce the conviction that these points have likely been generated by different mechanisms, and therefore shall not be modeled with the same equation as the rest of the population. For more ambiguous settings, a helpful and probably underutilized indicator to pinpoint (potentially) influential points is the "leverage" that derives from the projection matrix. The latter quantifies the influence of each observation $y_{obs,j}$ on the predicted value $y_{sim,i}$, and leverage scores correspond to the diagonal elements:

$$\text{Leverage}_i = \frac{\partial y_{sim,i}}{\partial y_{obs,i}} \in [0, 1] \tag{3}$$

A value close to 1 for the i^{th} leverage score means that a variation in $y_{obs,i}$ will result in an identically large variation in $y_{sim,i}$, or graphically, that the regression line "sticks" to the i^{th} data point. A point with a high leverage score is not necessarily an influential point, as it may just be distant from the rest of the data points but aligned with them. However, it should raise the modeler's vigilance regarding its potential effect on the regression model, particularly if it is subject to some uncertainty. On R software, leverage scores are calculated using the *hatvalues* function.

An even more direct way of identifying influential points is provided by "leave-one-out" procedures, which consist of (i) omitting one observation among the n data points, (ii) applying a linear regression to the subset of remaining $n - 1$ points, and (iii) examining the change in the fitted parameters and/or in the predicted value for the omitted point. These steps are repeated n times (excluding all data points one at a time), thus highlighting – if any – the points whose absence would result in a significantly different regression model.

By way of illustration, these two methods have been applied to Example A (see supplementary material). The leverage score for all points drawn from the two independent Gaussian distributions, *i.e.*, x_0 and y_0, remains below 0.15 (median: 0.07), while it is as high as 0.85 for the point (x_1, y_1). Similarly, the absence of one point from the first series of data would only marginally alter the slope of the regression model (from 1.11 to a minimum of 1.07 and a maximum of 1.24), which would however drop to <0.1 if the influential point were left out.

3.2. The criteria used for evaluating the distance between observed and simulated values may have a major influence on the fitted relationship

Prior to any statistical consideration, regression can first be regarded as a calibration procedure, in which the model parameters are fitted to reach a minimum distance between observed and simulated values. That being said, it should be kept in mind that there are different ways to define a "distance" between two sets of values. As presented in the foregoing paragraph, usual applications of linear and nonlinear regressions are based on the least-squares method (Eq. (2)), in which the distance criterion is defined as the sum of squares of the residuals, *i.e.* the usual Euclidean or ℓ^2 norm:

$$Dist_{\ell^2}(y_{obs}, y_{sim}) = \sum_{i=1}^{n} \left(y_{obs,i} - y_{sim,i}\right)^2 \quad (4)$$

Conversely, a criterion based on the ℓ^1 norm considers absolute values of the residuals and thus mitigates the contribution of large differences:

$$Dist_{\ell^1}(y_{obs}, y_{sim}) = \sum_{i=1}^{n} \left|y_{obs,i} - y_{sim,i}\right| \quad (5)$$

Whatever distance criterion is chosen, it is essential to recognize that, when considering the *vertical* distances to the regression line/curve ($y_{obs,i} - y_{sim,i}$), all error is assumed to reside in Y, while X is assumed to be known without (or with negligible) uncertainty (Cantrell, 2008). This is, for example, appropriate when considering the calibration curve of any measurement device, *i.e.*, the output delivered by the device when it is submitted to known x values – ideally certified standards (Benisch et al., 2021). However, in the presence of uncertainties affecting both the X and Y variables, it has been shown that the slope of the regression line is underestimated (Francq and Govaerts, 2014; Mikkonen et al., 2019).

It can be easily conceived that the minimum of these two distance functions is not necessarily reached for the same set of model parameters. To demonstrate how different the results may be, let us go back to Example B. Fig. 3 compares the "optimal" linear and quadratic models derived from the minimization of the $Dist_{\ell^1}$ and $Dist_{\ell^2}$ criteria. The influential point for the least-squares estimates of the regression coefficients has, conversely, no effects on the coefficients fitted with the ℓ^1 norm: the latter do coincide with the "true" coefficients of the relationship between X and Y in the interval $[0, 10]$, even for a quadratic regression model, for which the best-fit coefficient of x^2 is almost zero.

This brings us to a fundamental statement: the "optimal" parameters of a model fitted to a data series are not intrinsic, they are a direct consequence of the distance criterion used, which should not be regarded as a "default" or "universal" option. The usual choice of the ℓ^2 norm has several theoretical advantages, provided that statistical assumptions on the regression residuals can be guaranteed (Section 4.2). For instance, if X and Y are indeed linearly related in the whole population, the least-squares estimates of the regression coefficients are unbiased. More generally, the calculations leading to, *e.g.*, confidence and prediction intervals of the regression, are not compatible with estimates derived from the ℓ^1 norm. However, it is certainly preferable to limit the ambitions of the regression approach while still achieving a reasonably accurate adjustment, rather than applying a flawed method, notably due to influential points – but more widely to a whole set of assumptions that will be shown to be actually quite restrictive, and often violated.

4. Regression as a modelling approach

4.1. Regression models aim at making predictions, not at drawing a line on a graph

Thus far, the described objective has been to achieve the closest fit to the observed data points by adjusting the parameters of a given analytical regression function. Things fundamentally change when regressions are regarded as a modelling approach. Then the aim is not just to *approximate* the available data, but to claim that the fitted relationship would still be valid for a new measurement – in other words, to *predict* the value of Y corresponding to any x value. As should be the case for all models, this objective goes hand in hand with an assessment of the magnitude of uncertainty. This broadened purpose leads to the introduction of a probabilistic framework, in which (i) Y is considered as a random variable, (ii) the whole regression model characterizes the conditional probability distribution $Y|\{X = x\}$, (iii) the regression function corresponds to the deterministic component of the model, and (iv) the residuals are treated as realizations of a random variable ϵ, which encompasses all sources of variability in Y that are not captured by the latter. Under the common approach of least-squares minimization, the regression function is an estimate of the conditional expectation $\mathbb{E}[Y|X = x]$, and the residuals are assumed to be Gaussian with specific properties (see Section 4.2).

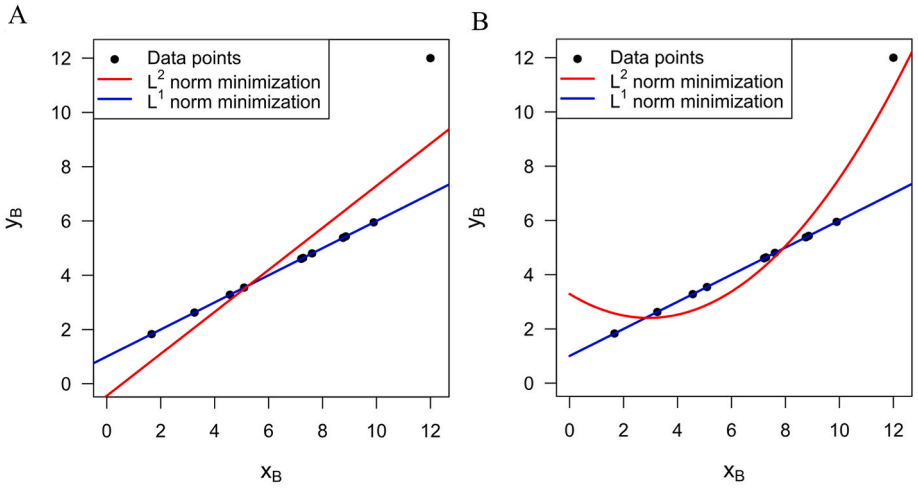


Fig. 3. Data points from Example B (10 measurements in a range where X and Y are linked by a linear relationship, and an additional point outside), along with the "optimal" linear (left) and quadratic (right) models fitted to the data by minimizing the $Dist_{\ell^1}$ (blue) and $Dist_{\ell^2}$ (red) criteria. The models adjusted with the $Dist_{\ell^2}$ criterion are shifted towards the influential point (the parabola naturally passes closer to it than the regression line due to an additional degree of freedom), but do not reproduce the linear trend between X and Y in the interval $[0, 10]$. Conversely, the "best fit" models obtained from the minimization of the $Dist_{\ell^1}$ criterion do match this linear trend and are not affected by the additional point.

Obviously, whether the focus is on the parameter estimates or on the model $Y|\{X = x\}$ depends on the study objectives. The problem arises when this goal is not explicitly stated, because these two matters do not imply the same level of complexity. Making the objectives explicit enables the elaboration of an appropriate strategy for conducting statistical analyses, while being aware of the formal framework and the assumptions to be checked.

Among the caveats that follow from the above, it should not be necessary to draw a "flat" line to demonstrate the lack of association between the two variables. More specifically, a horizontal regression line contains more implicit information than the statement of a non-significant correlation. It implies that the variability in Y cannot be explained by the variations in X, so that the "best" prediction one could give for an unmeasured y is the mean of the observed data. This, in particular, does not hold in cases such as skewed distributions for Y, where the mean value is poorly representative of the "center" of the distribution. More fundamentally, it misses the principal objective of the regression, *i.e.*, prediction, as Y is entirely unpredictable from X. In many cases, a visual assessment of the scatterplot should be sufficient to be convinced of the lack of association between X and Y. However, for readers in need of "mathematical proof", the calculation of an appropriate correlation coefficient (as detailed in Section 6.3) would be more suitable than a linear regression analysis – which, once again, should not be considered as an equivalent statistical tool.

4.2. Inference in regression relies on theoretical properties of the residuals (that should be checked)

When applying a linear regression method in scientific publications, it has become quite common to display, in addition to the data points and the regression line, two hyperbola branches representing the confidence interval associated with the regression line (Fig. 1). Some graphs also show the prediction interval that should contain, with a given confidence level, the value of an unmeasured y predicted from any x value. Although these plots may be "standard" outputs from various statistical software packages, it is essential to remember that the mathematical derivation of these curves – as well as other inference results from the linear regression model – hinges upon several properties of the residuals (Dodge, 2008):

- All residuals are independent realizations of the same random variable ϵ. The corollary of this assertion is that the residuals do not describe any pattern as a function of x (the contrary would indicate a nonlinear relationship between X and Y).
- The distribution of the residuals is Gaussian with a null mean and a constant variance σ^2 (which belongs to the model parameters).

In other words, if these properties are not met, the mathematical calculations on which inference is based are unsuitable, and outcomes such as confidence and prediction intervals are consequently meaningless. It is therefore fundamental to check the validity of these assumptions before proceeding any further.

For that purpose, despite the general obsession for "statistical proof", it is worth mentioning that considerable insight can be gained from visual assessments. This idea is epitomized by the famous example of "Anscombe's quartet", which presents four datasets with identical basic statistical properties that nevertheless display very different patterns on a scatterplot (Anscombe, 1973). Plotting the residuals against the explanatory variable (or against the predicted variable) may reveal either a deterministic trend (so that the linear model is not adapted to describe the data) or a change in point dispersion with increasing x or y (so that the residuals do not have a constant variance). Usual visualizations of the statistical distribution of the residuals, such as a histogram, boxplot, or Q-Q plot against the quantiles of the standard normal distribution, provide valuable information regarding the validity of the normality assumption, beginning with the symmetry of the distribution.

These assessments may be complemented by statistical testing. Among other possibilities, the presence of autocorrelation in the residuals may be detected through the Durbin-Watson test; heteroscedasticity (*i.e.*, non-constant variance of the residuals) may be detected using the Breusch-Pagan and the White tests (Breusch and Pagan, 1979; White, 1980). On R software, these tests can be found in the 'lmtest', 'car' or 'skedastic' packages (Farrar, 2024; Zeileis and Hothorn, 2002). On the other hand, checking the normality of residuals relies on possibly more common procedures such as the Shapiro-Wilk, Lilliefors or Anderson-Darling tests, among others. These procedures are apparently not considered a routine that should follow the application of linear regression – as reflected by the many publications in environmental science failing to report such verifications after presenting a linear regression.

4.3. Modelling the mean value is not the only way to perform regression

As explained above, the most common approach to linear regression, based on ordinary least-squares, addresses how the *mean* of Y evolves as X takes on different values. Although the mean is a measure of position that is well adapted for Gaussian (or at least symmetric) distributions, it can be much less meaningful if the conditional probability distribution $Y|\{X = x\}$ is significantly skewed, and provide a distorted view of the variable under study. Nevertheless, the mean is not the only statistical property that can be approached in a regression model. *Quantile* regressions have been designed to appraise the evolution of any quantile of the distribution of $Y|\{X = x\}$ as a function of x. There are as many regression functions as quantiles to be estimated, possibly with different slopes/trends: this gives the regression model more subtlety than the relatively strict hypothesis of homoscedasticity – a corollary of which is that all quantiles should increase of decrease in parallel with the mean.

The fitting procedure of a quantile regression involves the minimization of a specific criterion. To understand how the latter is constructed, let us first point out that for a single variable Y with n measurements $\left(y_{obs,1}, \ldots, y_{obs,n}\right)$, the quantile of order $\tau \in [0,1]$ can be estimated as the value q that minimizes the following sum (D'Haultfoeuille and Givord, 2014):

$$\sum_{i=1}^{n} \varphi_\tau\left(y_{obs,i} - q\right) \tag{6a}$$

where

$$\varphi_\tau(u) = \begin{cases} (1-\tau)\,|u| & u < 0 \\ \tau\, u & u \geq 0 \end{cases} \tag{6b}$$

In a quantile regression, the only difference is that the quantile of order τ is assumed to be a certain function of the explanatory variable instead of being a constant value. Therefore, the "distance" between y_{obs} and y_{sim} is given by:

$$Dist_{\text{quantile}_\tau}(y_{obs}, y_{sim}) = \sum_{i=1}^{n} \varphi_\tau\left(y_{obs,i} - y_{sim,i}\right) \tag{7}$$

Contrary to Eqs. (4) and (5), this is not strictly a mathematical distance, as y_{obs} and y_{sim} do not have symmetric roles in general. The only and noteworthy exception is for $\tau = 0.5$ (*i.e.* when estimating the median value), where the function $\varphi_{0.5}$ takes the simpler form $\varphi_{0.5}(u) = 0.5\,|u|$. To within a factor of 0.5, $Dist_{\text{quantile}_{0.5}}$ is equal to the above-mentioned $Dist_{\ell^1}$ (Eq. (5)): this explains, in a different way, the lack of impact of influential points on the coefficients obtained by minimizing the ℓ^1 norm (Fig. 3), the median being known to be much less sensitive than the mean to extreme values.

In practice, quantile regression can be performed on R software with the *rq* and *nlrq* functions (for linear and nonlinear models, respectively) implemented within the 'quantreg' package (Koenker et al., 2017).

5. The hazards of extrapolation

5.1. Extrapolation of empirical relationships should be made with caution

Regression-based approaches have become so standard in environmental science that perhaps not everyone questions the soundness of the statement: "I have observed *several* points aligned so *all* points should be aligned". To reiterate: as the fundamental assumption of empirical modelling, using a regression model to predict the *y* value corresponding to any *x* value implies that the fitted relationship between the two variables remains valid beyond the set of observations. More precisely, the processes whose effect on *X* and *Y* is observed, but not mechanistically described, are expected to operate in the same way in unmeasured situations.

Here again, caution is called for, remembering that the domain of validity of an empirical model is *a priori* circumscribed by the collected data points and experimental conditions. While interpolation of suitably scattered points can be considered reasonable to allow for prediction in similar contexts, there is no guarantee that the adjusted model will hold up for extrapolation – which encompasses both predictions outside the range of observed data and under different environmental conditions. Bartley et al. (2019) present a noteworthy example where log-transformed concentrations of chlorophyll *a* and total phosphorus in U.S. lakes are linked by a linear relationship within a certain range of values, but this relationship becomes totally unsuitable below and above this range. Justifying the universality of an empirical model is a research task in itself, which, if successful, can lead to far-reaching conclusions; but which, if omitted, may result in totally inaccurate predictions. Lang et al. (2010) illustrate misleading hydrological diagnosis due to rating curve errors, most of which were found to be associated with extrapolation issues. In the words of Hahn (1977): "extrapolation cannot be supported on statistical grounds alone; it must be justified by physical considerations". More generally, at a time when we are witnessing a rapid expansion of data mining-based modelling approaches, acknowledging the issues associated with extrapolation in the case of linear regression – which can be seen, after all, as one of the simplest data mining schemes – will certainly provide critical insight into these practices and the validity of such modelling results.

5.2. First illustration: avoiding the prediction of non-physical values

There is one situation in which incautious extrapolation conducts to scientific nonsense, namely when model predictions reach non-physical values. Many contexts come to mind in which the quantity under study is physically bounded: for instance, a water depth, a concentration, a density, *etc*., cannot be negative, a yield or a removal rate cannot be greater than 100 % … Yet, if this quantity is treated as the response variable of a regression model, it may happen that the arrangement of the data points leads to a fitted model outside this "physical range" (see Example C/Fig. 4). There is perhaps no major novelty in asserting that a *statistical* model is not *physically-based*, and consequently does not by itself integrate the specific constraints related to the response variable; however, the automatic use of statistical tools can result in this fact being overlooked. The prediction of non-physical values should obviously be regarded as a mathematical artifact, and these parts of the model should *at least* not be displayed on a graph. More fundamentally, such situations should raise the researcher's vigilance about the validity of the regression model, and the use they intend to make with it.

Example C is a case of patent inadequacy of a linear model to describe the whole experimental trend between the two variables under consideration. It is certainly useful to be reminded that various tools are available for fitting piecewise functions to this kind of data points. On R software, the 'segmented' package may be used for that purpose: it determines the characteristics of the different line sections and the abscissa of the slope break(s), along with standard error/confidence interval associated with each estimate.

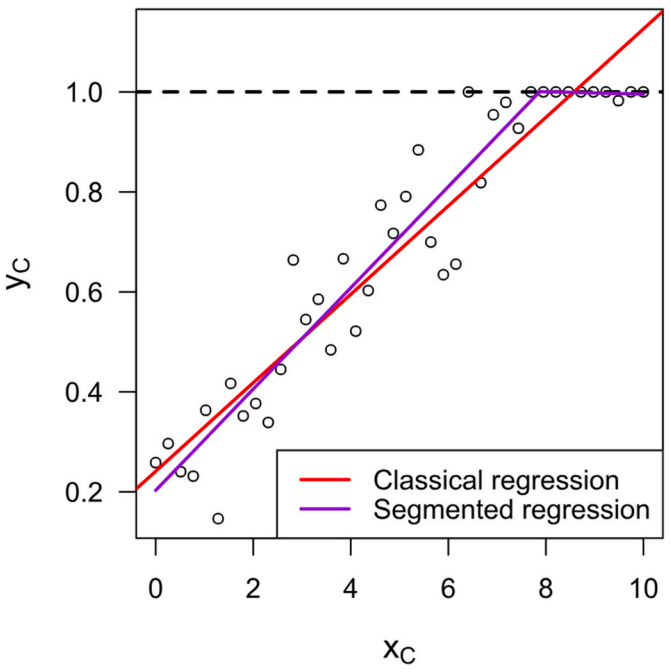


Fig. 4. Graphical illustration of Example C, where X and Y are linked by a linear relationship (affected by random noise) up to $Y = 1$, which is the upper bound for Y. Due to the arrangement of the data points, the linear model fitted from the application of a linear regression to the whole data set predicts "non-physical" values (*i.e.*, Y values greater than 1), contrary to the model derived from a piecewise regression (using the 'segmented' package in R – see supplementary material).

Example C

In a treatment device, the removal rate of a pollutant (Y, expressed as a fraction of the inflow concentration) is influenced by the amount of reagent used (X, in mg/L), as well as by other stochastic factors such as the water composition, that may either enhance or inhibit the treatment process. Let us assume that (i) Y has a deterministic linear dependence upon X up to $Y = 1$, whose coefficients are unknown to the experimenter; and (ii) the effects of the other mechanisms can be modeled by a Gaussian random variable (ϵ) with a mean value of $\mu_\epsilon = 0$ and a standard deviation of $\sigma_\epsilon = 0.1$. In this example, the theoretical expression of Y (which cannot be greater than 1) is given by: $Y = \min(0.2 + 0.1X + \epsilon\,; 1)$

40 campaigns have been carried out, during which different quantities of reagent x between 0 and 10 mg/L have been added to the treatment process, while assessing the removal rate of the target pollutant. Measurements of Y are exact. Linear regression is then applied to the data to characterize the effect of X on Y.

5.3. Second illustration: accounting for censored data in a regression model

A slightly more subtle facet of this problem is encountered when the data are estimated with a certain limit of quantitation (LQ), which is the case, for example, for most contaminant concentrations in environmental matrices. Faced with such a situation, an unfortunately common practice remains to substitute LQ or LQ/2 for the left-censored values, before applying conventional statistics as if these were "genuine" measured data. Example D provides a typical illustration of this practice (Fig. 5), which calls for a number of comments. Rationally, it is far from obvious that a *data-driven model* might be able to predict Y values that

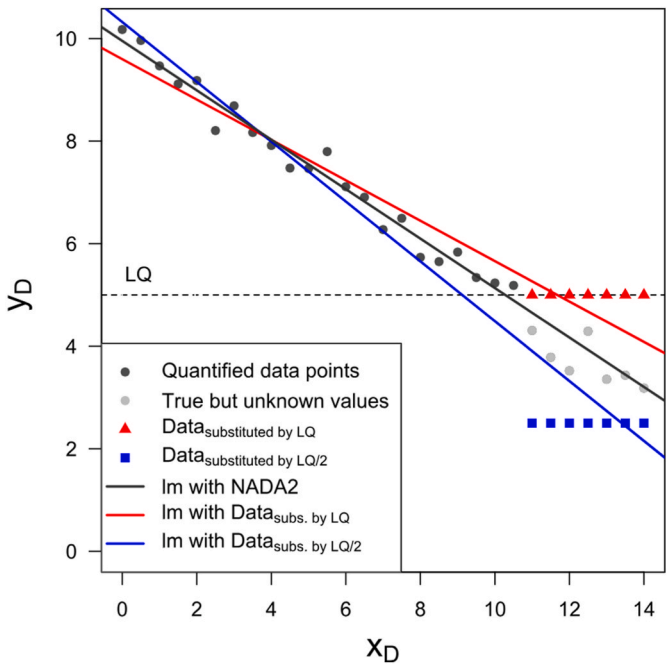


Fig. 5. Graphical illustration of Example D, where X and Y are linked by a linear relationship (affected by random noise), with Y measurements subject to a limit of quantitation LQ = 5. Grey points represent "true" but unknown Y values as they fell below the LQ, red (resp. blue) points represent the result of their substitution by LQ (resp. LQ/2). The linear models fitted after applying substitution methods to the left-censored data (red and blue lines) lead to a biased representation of the trend between X and Y, even in the range where Y is quantified. Conversely, the linear model fitted from maximum likelihood estimation (using the 'NADA2' package in R) provides a more adequate fit.

cannot be quantified experimentally. Such a purpose undoubtedly requires additional hypotheses to be made – and explicitly stated – within the statistical model underpinning the regression approach. Furthermore, the substitution of left-censored data may have an influence on the fitted regression coefficients, even, ironically, sacrificing a reliable reproduction of the linear trend which is actually observed where Y is quantified.

Adequately handling left-censored data has been the subject of comprehensive publications, and is well beyond the scope of this article. We shall simply mention the existence of the R packages 'NADA' and 'NADA2' (Julian and Helsel, 2024; Lee, 2020), and notably the *cencorreg* function, based on maximum likelihood estimation, the use of which is illustrated as an alternative (and preferable) regression method in Example D. For further details and recommendations, readers are invited to consult Helsel (2012) and the documentation associated with the above-mentioned R packages.

Example D

Consider two variables X and Y linked by the following relationship, whose coefficients are unknown to the experimenter:

$Y = 10 - 0.5X + \epsilon$ for $X \in [0,14]$

where ϵ follows a normal distribution with a mean value of $\mu_\epsilon = 0$ and a standard deviation of $\sigma_\epsilon = 0.3$.

Measurements of Y are taken for regularly spaced X values between 0 and 14, with a limit of quantitation LQ = 5. The coefficients of the relationship between X and Y are estimated after applying two relatively common approaches to the left-censored data, *viz.* LQ or LQ/2 substitution.

6. Misinterpretations of correlation tests

6.1. "Significant" correlations mean "significantly non-zero" correlations

In currently published research, testing "statistical significance" has become central to many demonstrations, including correlation analyses in which tables of p-values sometimes even replace data visualization. The purpose of this section is not to reopen the debate on the scientific value of this concept, which has been widely and helpfully discussed, *e.g.*, by Greenland et al. (2016), Amrhein et al. (2019), and Dushoff et al. (2019). However, it is certainly useful to recall some elements about the meaning of "significance" with regards to correlations, in order to understand that simply commenting on the "significant" nature of a correlation coefficient, without any further consideration, cannot be sufficient.

Although there are different ways of computing a correlation coefficient between two variables, the principle of a correlation test is globally the same. The null hypothesis $\mathscr{H}_0$ states that the "true" correlation coefficient (*i.e.* the correlation that exists in the whole population) between the two variables X and Y is equal to zero. A correlation test examines the compatibility between this hypothesis and the observed data, given the set of assumptions that underpin the statistical model. More precisely, the estimate of the correlation coefficient, calculated from the collected sample, is compared to its expected value under the null hypothesis, *i.e.* zero. If the other assumptions of the statistical model are warranted, then a sufficiently large difference, reflected by a sufficiently small p-value, may justify the rejection of $\mathscr{H}_0$, which means that there exists a non-zero correlation between X and Y.

Herein lies perhaps a source of misunderstanding in the interpretation of correlation tests: it should be recognized that this information alone is actually a relatively weak conclusion. *Some* degree of association between X and Y is indeed distinguishable from their other sources of fluctuations, which does not inform about the strength of this association, reflected by the value of the correlation coefficient itself (illustration on Fig. 6). Additionally, the results of a correlation test are

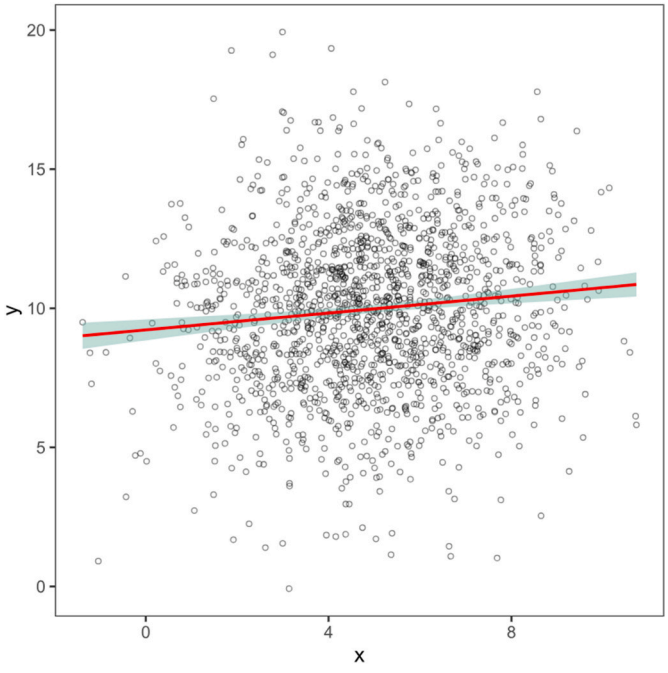


Fig. 6. Illustration of a data point arrangement (n = 1600) leading to an estimated Pearson correlation coefficient r = 0.10 between the two variables, with a 95 % confidence interval of 0.05–0.15. Testing the significance of this correlation leads to a p-value of 2.10^{-5}, well below the usual threshold of 0.05. Similarly, the regression line and corresponding 95 % confidence interval (shaded area) show a strictly increasing trend. However, although "significantly non-zero", it is not obvious that this correlation tells us much about the physical phenomena under investigation, or that the regression line provides a useful prediction of the response variable. This question should be answered by the researcher's knowledge, expertise and expectations, rather than by a statistical consideration.

overly sensitive to the sample size: for a significance level $\alpha = 5\%$ and a sample size $n = 10$, any value of the Pearson correlation coefficient greater than 0.63 will be considered "significant"; for $n = 100$ and $n = 1000$, this threshold drops to ~0.20 and ~0.06, respectively. It is therefore not so remarkable to find "significant" correlations among several hundreds of data points or more. Once again, rather than focusing on statistical significance, it is definitely more important to discuss the degree of association between the two variables, and the practical implications regarding the processes under study.

6.2. Testing the significance of the Pearson correlation coefficient is a parametric approach

Performing a correlation test on a Pearson coefficient requires assumptions to be made regarding the population distribution from which the sample is drawn. More specifically, the distribution of the test statistic is known if (i) observations are independent and identically distributed, (ii) X and Y follow a bivariate normal distribution, and (iii) X and Y are uncorrelated in the whole population ($\mathcal{H}_0$). The anomaly leading to a small p-value should come from this latter assumption and this one only, if one wants to demonstrate that the correlation between X and Y is "significant". The joint probability density function of a bivariate normal distribution is illustrated in the supplementary material: isovalue lines are ellipses centered on the mean point, whose minor axis tends towards zero as $r_{XY} \longrightarrow \pm 1$. In order for a Pearson correlation test to be meaningful, it is fundamental that the data points conform to this theoretical behavior, *i.e.* that they describe approximately an ellipse, otherwise a small p-value may be merely due to the different statistical distribution, rather than to the fact that $r_{XY} \neq 0$.

More generally, the normality assumption, where variability is assumed to arise from random fluctuations symmetrically distributed around a constant central value, is often questionable when dealing with environmental variables – starting with the non-negative nature of most of the quantities involved. As a matter of fact, an observation shared by scientists from different research fields is that contaminant concentrations in various environmental matrices (air, water, soil, and even biological tissues) frequently follow a lognormal distribution (Andersson, 2021; Limpert et al., 2001). Finally, it must be borne in mind that normality tests do not *prove* that X follows a normal distribution, they just *fail to reject* it (which is often the case with small sample sizes), leading to accept the hypothesis as a "default" choice. Achieving some certainty as to the validity of the normality assumption is, here again, quite a demanding task.

6.3. There are non-parametric alternatives to the Pearson correlation coefficient

It is somehow paradoxical that the Pearson coefficient is the default option in most software for calculating the "correlation" between two variables, when it is actually the most restrictive; this is due to the above-mentioned reasons, but also, more commonly admitted, because it is only suitable for characterizing the strength of *linear* associations. When dealing with variables that either describe a visible albeit nonlinear trend, or simply do not comply with the normality assumption, alternative, non-parametric approaches are possible – and welcome.

The Spearman correlation coefficient is defined as the Pearson coefficient applied to the data series $rank(x_{obs})$ and $rank(y_{obs})$. The coefficient proposed by Kendall counts positively (resp. negatively) the number of pairs of data points positioned in ascending (resp. descending) order, and normalizes the sum by "n choose 2", *i.e.* $\frac{n(n-1)}{2}$. The purpose of the present discussion is not to compare these two coefficients – there is an abundant literature on the subject, see for instance Xu et al. (2013) – but to remind environmental scientists that these two approaches have a much wider range of validity than the Pearson coefficient: they are neither affected by monotonic but nonlinear behaviors, nor by extreme values, nor constrained by the normality assumption. The trade-off for this genericity is that conclusions will be less refined, insofar as Spearman and Kendall correlation coefficients do not reflect the linearity of the relationship, if any.

It is also worthwhile to recall that further alternatives exist to the aforementioned correlation coefficients for quantifying the degree of association between two variables. From Information Theory, the Average Mutual Information (AMI) can be named as a popular indicator in fields such as bioinformatics, communications, and electronics, but remains rarely employed in environmental studies. The AMI indicator is based on the probabilistic definition of independence, where two random variables X and Y are independent when the joint probability density function $p_{X,Y}(x,y)$ equals the product of the marginals $p_X(x) \cdot p_Y(y)$ for all x and y. Therefore, the degree of dependence between X and Y can be defined based on a notion of "distance" between $p_{X,Y}(x,y)$ and $p_X(x) \cdot p_Y(y)$. For this purpose, the Kullback-Leibler Divergence, D_{KL} (Kullback and Leibler, 1951), can be adopted, leading to define the AMI between X and Y as:

$$AMI(X,Y) = D_{KL}\left(p_{X,Y} \,\|\, p_X \cdot p_Y\right) = \iint p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x) \cdot p_Y(y)}\right) dxdy \quad (8)$$

$AMI(X,Y)$ is a non-negative and symmetric quantity, with units in bits when employing $\log_2$ in Eq. (8). Two possible interpretations of $AMI(X,Y)$ are: the amount of information shared between X and Y, or the amount of entropy (uncertainty) reduced in Y by knowing X (or vice-versa from mentioned symmetry). The main advantage of $AMI(X,Y)$ when compared to correlation coefficients probably remains its ability to detect nonlinear and non-monotonic associations between random variables (Kraskov et al., 2004). Furthermore, $AMI(X,Y) = 0$ is mathematically equivalent to the statistical independence of X and Y, which is not guaranteed by a zero correlation between X and Y (a classical counterexample is given by $Y = |X|$, with X in $[-1,1]$).

As $AMI(X,Y)$ is based on probability distributions, this indicator can be calculated following two approaches: continuous or discrete estimators. For this purpose, a series of algorithms and R packages are described in Sales and Romualdi (2011).

6.4. The statistical model underpinning correlation testing assumes independent observations

Finally, an often overlooked hypothesis would certainly deserve further consideration: the observations are assumed to be *independent* realizations of the random variables under study. Particular caution should be exercised when processing data measured consecutively at the same location (*e.g.*, the water level or flow rate in a river, the concentration of a certain pollutant in a water body or the atmosphere), which may be linked to each other by deterministic processes governing the temporal evolution of the variable: this is known as temporal autocorrelation (Hyndman and Athanasopoulos, 2021). When dealing with spatially distributed information, autocorrelation may also occur in space, if observations at neighboring locations tend to be similar – as expressed, for example, by the decreasing trend of the variogram as the lag distance tends towards 0. In both situations, the absence of autocorrelation between observations should be checked before considering their correlation with other variables.

The burden of autocorrelation can be illustrated with Example E and Fig. 7. In this situation, concluding that the association between X and Y is "significant" from a correlation test performed on repeated measurements would obviously be flawed: repeated measurements are relevant to mitigate random errors, and thus reduce the uncertainty on the mean value at each site; but they should not be included as such in a correlation analysis, because they are not independent from each other at the level of the whole population (*i.e.* all of the possible study sites).

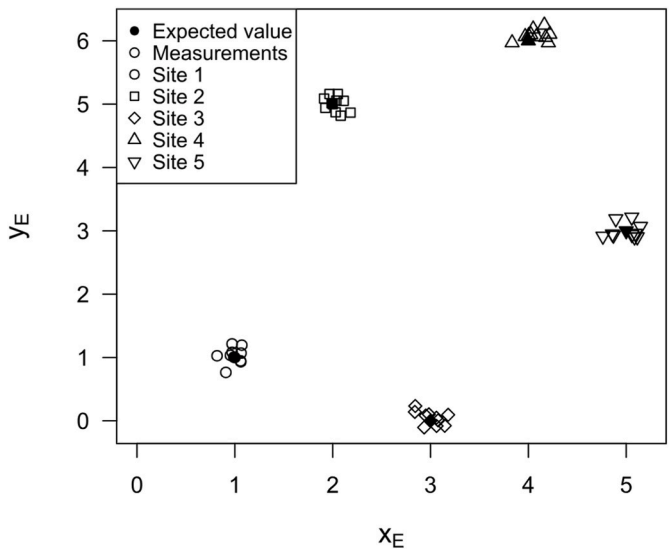


Fig. 7. Graphical illustration of Example E, where X and Y are independent random variables, measured 10 times at 5 study sites. With the five data points corresponding to the expected values of X and Y in the five sites (black symbols), the estimate of the Pearson correlation coefficient is 0.31, with a 95 % confidence interval of [–0.79, 0.94]. In other words, it is clear that the null hypothesis ($r_{XY} = 0$) cannot be rejected – as reflected by a p-value of 0.61. However, when including all repeated measurements (white symbols) in the data points on which a correlation test is performed, the estimate of r_{XY} remains approximately the same (the small difference is due to random sampling), but the sample size artificially increases from $n = 5$ to $n = 50$, leading to a seemingly "significant" correlation between X and Y (p-value = 0.02).

Example E

X and Y are independent random variables. Their expected values at five independent study sites are as follows: Site 1: (1,1); Site 2: (2,5); Site 3: (3,0); Site 4: (4,6); Site 5: (5,3).

10 repeated measurements of X and Y are carried out in each site. The latter are prone to some uncertainty, modeled by a normal distribution centered on the expected value, with a standard deviation of $\sigma_X = \sigma_Y = 0.1$. A correlation test is carried out on the entire set of 50 data points.

Despite being rarely considered in environmental studies, autocorrelation in time or space can have broader implications in data analysis than the potential violation of assumptions in statistical testing (*e.g.* for forecasting, model selection, identifying patterns and uncertainty propagation). Overall, different tools are available to assess data autocorrelation, including the Autocorrelation Function (ACF), the Partial Autocorrelation Function (PACF) and semivariograms. Autocorrelation can also be modeled by means of covariance models (Simpson et al., 2010).

7. Model evaluation

7.1. The sole use of correlation as a model evaluation metric remains a hazardous practice

Beyond regression techniques, the evaluation of any quantitative model consists of confronting simulated values to "reference" values, which most often originate from observations/measurements (but may also correspond to the outputs of a reference model). To evaluate the model's goodness of fit, it is rather common to consider the data points $y_{sim,i}$ against $y_{obs,i}$, independently from any temporal dynamics or spatial distribution. Various metrics can be found to compare these two series, most of which are presented in the comprehensive review by Hauduc et al. (2015). We would, however, like to warn against two potentially misleading practices, namely the sole use of the Pearson correlation coefficient (r) between y_{sim} and y_{obs}, and/or inappropriate use of the coefficient of determination (R^2).

First of all, it is fundamental to recall that a "perfect model fit" would correspond to all points lying on the first bisector (*i.e.*, $y_{sim,i} = y_{obs,i}\ \forall i \in [\![1,n]\!]$), which is actually a much stronger condition than all points being "aligned". In other words, to consider that the evaluated model is capable of replicating observed values, the existence of an association between y_{sim} and y_{obs} is indeed a necessary element – the opposite would mean that model predictions fluctuate independently from observations – but this argument alone cannot be considered as sufficient evidence to conclude that the model performance is "satisfactory". To simply illustrate this point, let us consider Example F (Fig. 8). This corresponds to a situation leading to a linear correlation $r = 1$ between y_{sim} and y_{obs}, yet where probably no one would consider the model predictions to be accurate, as they consistently underestimate the observed values. It should therefore be understood that exhibiting a "significant" correlation between simulated and observed values, without further consideration, is of limited interest for model evaluation. Naturally, analogue conclusions would be delivered about the model performance, if adopting the squared version of r, *i.e.* r^2.

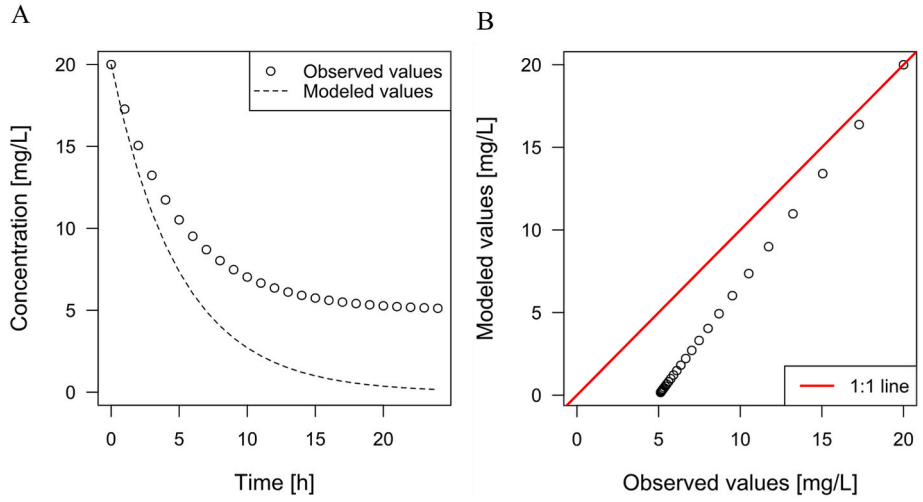


Fig. 8. Graphical illustration of Example F (modelling the first-order decay of pollutant concentrations in a water body disregarding the background concentration). (A) Temporal evolution of observed and modeled concentrations over a 24-h period. (B) Correlation between observed and modeled values ($r = 1$). Despite this perfect linear correlation between y_{sim} and y_{obs}, probably no one would consider the model predictions to be accurate, had they looked at the data.

Indeed, a widespread practice, even implemented in different software and simulation packages, remains to calculate a squared version of the Pearson correlation coefficient r as a performance indicator of a model. Squaring r in this modelling context is commonly thought to bring an "improved version" of the indicator r standalone by "eliminating" negative r values, "penalizing" poor performances and thereby gaining interpretation in the results. However, the r^2 indicator remains as weak as r for very general settings. A more serious conceptual issue is that this r^2 indicator is frequently cited as the coefficient of determination, or more commonly named R^2 (according to the standard notation). We wish to raise particular awareness about the validity of this assertion, by demonstrating why the coefficient of determination and the squared Pearson correlation coefficient remain two interchangeable quantities only in a very limited number of contexts. Common theoretical misconceptions and potential pitfalls when not interpreting these two concepts properly are discussed in the following section.

Example F

A pollutant is discharged into a water body, where its degradation follows the k-C* model (first-order decay with a background concentration):

$C(t) = C^* + (C_0 - C^*)e^{-kt}$

where C_0 and C^* are respectively the initial and background concentrations ($C_0 =$ 20 mg/L, $C^* = 5$ mg/L), and k is the degradation rate constant ($k = 0.2\ \mathrm{h}^{-1}$). Water is sampled every hour over a 24 h period (sampling times are referred to as t_i). Let us assume that water sampling and analysis lead to exact measurements, so that $y_{obs,i} = C(t_i)$.

Consider a situation where, as a result of a flawed analysis of the system, the pollutant fate is modeled omitting the background concentration. The value of the parameter k is known from preliminary tests. Then simulated values are given by:

$y_{sim,i} = C_0\, e^{-kt_i}$

Instead of visualizing the data, model evaluation is carried out through numerical indicators only, and notably the Pearson correlation coefficient between y_{sim} and y_{obs}.

7.2. In general, the coefficient of determination is not equivalent to the squared Pearson correlation coefficient

The coefficient of determination (hereafter referred to as CD to avoid any confusion) compares y_{sim} and y_{obs} through the following equation:

$$\mathrm{CD} = 1 - \frac{\sum_{i=1}^{n}\left(y_{obs,i} - y_{sim,i}\right)^2}{\sum_{i=1}^{n}\left(y_{obs,i} - \bar{y}\right)^2} \tag{9}$$

where $\bar{y}$ is the mean of the observed values. Two immediate interpretations follow directly from the definition: one the one hand, the CD compares the amount of residual variance *versus* the variance of the data itself; on the other hand, it describes "how good the model is performing – according to the $Dist_{r^2}$ criterion – when compared to simulating the output with a naive estimation: the mean of the observed data". An additional interpretation stems from the following variance decomposition, the validity of which is restricted to ordinary least-squares linear models:

$$\sum_{i=1}^{n}\left(y_{obs,i} - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_{sim,i} - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_{obs,i} - y_{sim,i}\right)^2 \tag{10}$$

It follows from the combination of Eqs. (9) and (10) that CD quantifies the proportion of variance explained by the linear model.

Whatever interpretation is adopted, it remains quite far from the original purpose of the Pearson correlation coefficient. However, despite their different objectives, it happens that *in the specific context of simple linear regression fitted with the least-squares method*, the following three quantities are identical: the CD, the squared Pearson correlation coefficient between y_{obs} and x_{obs} and, consequently, the squared Pearson correlation coefficient between y_{obs} and $y_{sim} = ax_{obs} + b$. This could be a reason for confusion between the CD and r^2 when addressing more general contexts outside of linear regression.

In the hydrological and environmental modelling literature, the CD is often called the Nash-Sutcliffe Efficiency (NSE) coefficient (Nash and Sutcliffe, 1970). In these scientific disciplines, it is well known that, contrary to r^2, the NSE coefficient can be lower than zero, meaning that the model performance is more deficient – once again, according to the $Dist_{r^2}$ criterion – than modelling the output with the mean of the observed data. Indeed, when the model or calibration approach is different from linear regression with least-squares estimates, or when a fitted model is evaluated against a dataset different from the training data, the variance decomposition according to Eq. (10) is no longer guaranteed. Consequently, the residual sum of squares, $\sum_{i=1}^{n}\left(y_{obs,i} - y_{sim,i}\right)^2$, can be greater than the total sum of squares, $\sum_{i=1}^{n}\left(y_{obs,i} - \bar{y}\right)^2$. For instance, in Example F, CD $= -0.04$. More broadly, it is a good modelling practice to separate the training and validation datasets, even for linear regression, so as to evaluate the model's goodness of fit on data with which it has not been trained.

In sum, when carrying out model evaluation, simulated values, y_{sim}, have to be *compared to* observed values, y_{obs}, rather than *correlated to* y_{obs} (as with r), nor *predicted from* y_{obs} (as with r^2). On the one hand, a correlation between y_{sim} and y_{obs} does not support, by itself, the accuracy of the model predictions, as it does not prove that the data points are distributed around the first bisector. On the other hand, the coefficient of determination of the model (*i.e.*, the series of calculations that produces y_{sim}) should not be confused with the coefficient of determination of the (essentially useless) linear regression between y_{sim} and y_{obs} – the only interest of which could be to check that the average trend between simulated and observed values remains close to the 1:1 line. Omitting the proper calculation and objectives of one or the other indicator can easily lead to erroneous conclusions (*e.g.*, under-judging modelling biases).

More generally, instead of focusing our attention on lists of indicators, it is certainly advisable to reconsider the primary objective of a model evaluation approach: to question whether the magnitude of the errors, $y_{obs,i} - y_{sim,i}$, is acceptable for our own model application. This will be addressed by combining much broader considerations than goodness of fit metrics, such as the accuracy sought by the modeler, but also the uncertainties inherent in the model itself and in the collection of observed data.

8. Conclusion: graphical verification is as important as statistical verification

Most environmental scientists are facing the inherent challenge of establishing some evidence about complex processes that are often appraised through scarce, partial and/or uncertain data. To this end, the methods inherited from statistical modelling open up a wide range of possibilities, overcoming a mechanistic description of phenomena when this proves virtually impossible to implement. The adoption of correlation- and regression-based approaches in environmental studies is therefore understandable, and in many cases perfectly justified. It is also likely that these approaches have been fostered by the democratization of statistical tools and software which have facilitated data processing and analysis.

However, the search for metrics, trends, "statistical significance" – sometimes encouraged by the publication process itself – or the mere reproduction of current practices seem to have led, in certain situations, to an automatic use of these tools that has gradually turned away from the data themselves. As outlined above, any statistical result stems from a series of mathematical calculations, themselves underpinned by different hypotheses, which can easily be overlooked when running today's easy-to-use software. The application of certain statistical methods considered "standard" or "basic" is often tantamount to inferring, sometimes without even realizing it, the distribution of the data. Unfortunately, nature does not often conform to the mental projections that would allow routine use of these approaches. In other words, it is a

complete illusion to imagine that a result or a graph is necessarily "correct" or "accurate" because it has been produced by a statistical tool; and it is even more illusory to consider that this result can have more value, or even replace, the expertise of the individual who generated the data and proposes an interpretation thereof.

It is therefore necessary to maintain constant vigilance with regard to what can and cannot be asserted about the variables and processes of interest. As experimental scientists, we need to be more aware of the possibilities of misrepresenting data (listed above in a non-exhaustive way) and not let our own biases/perspectives distort correlations between variables or even the data themselves. In our opinion, the first act of vigilance, and probably the most effective, is to devote time to data contemplation. Visualization techniques have been sometimes depreciated as not constituting "statistical proof", but we are convinced that they should be considered as one of the most important steps in data analysis.

In sum, although widely employed, correlation- and regression-based analyses are actually expertise-demanding approaches. Configurations allowing rigorous application of these methods to environmental data are perhaps not as numerous as the scientific literature would suggest. As a closing message to this discussion, we propose below a series of "good practices" which we hope may contribute to raising awareness among environmental scientists and improving the statistical treatment associated with bivariate data.

9. "Good practices" in implementing correlations and regressions

During data collection and preprocessing

- Design the experiment to maximize the chances of collecting independent, identically distributed observations. Keep track of apparently "non-standard" environmental conditions (*e.g.*, a flood event in a river) that may result in a different parent population for the observations.
- Keep track of left-censored data (if any), together with the corresponding limit of quantitation, and avoid substitution methods.
- Visualize the data and formulate preliminary assumptions regarding (i) the degree of association between variables and (ii) the structure of the data points.
- In particular, check for potential influential points and piecewise/nonlinear trends.
- Check the absence of temporal or spatial autocorrelation in the collected data, with additional normality tests if parametric statistics are to be applied to the dataset.

During data analysis

- Select appropriate methods for assessing the degree of association between variables based on the conclusions of the previous steps. Bear in mind the potential shortcomings of the Pearson correlation coefficient, which should not be considered as a "default" approach.
- Remember that evidencing a "significant" correlation is often a relatively weak conclusion in itself, which calls for further analysis.
- If linear or nonlinear regression is to be applied to the data, clearly state the objective: to approximate the data points with a given analytical function, or to use the regression model to make predictions. Set the next steps accordingly.
- Keep in mind the interest of quantile regression, as a potential alternative to ordinary least-squares regression (that estimates the mean of the response variable).
- Remember that nonlinearity is often the rule rather than the exception, and do not strive to fit a linear model to a set of data points that visibly do not conform to a linear behavior. More subtle alternatives are possible:
  - If the linear trend is only visible over a certain range of X and Y values, determine a linear model valid for this range.
  - If several distinct line sections are visible, apply a piecewise regression.
  - If the data points describe a nonlinear trend, apply a nonlinear regression.
  - It should not be necessary to draw a flat line (let alone a confidence interval) to prove that no correlation is observed between two variables.
- To be able to display and use inference results from a regression (*e.g.*, confidence and prediction intervals), check the residuals for normality, homoscedasticity and absence of autocorrelation.

During model use and dissemination

- Revert to visual assessment in order to question the accuracy of the adjusted model(s); do not depreciate common sense if they seem inappropriate or imprecise, whatever the values of the statistical indicators such as p-values or R^2.
- Do not use "significant" correlations as a sufficient justification for model validity.
- To avoid unnecessary confusion, we recommend using the *coefficient of determination* whenever prediction is the aim, and *correlation* when describing associations between variables. Do not calculate the former by squaring the Pearson correlation coefficient, but rather as 1 – residual sum of squares/total sum of squares.
- Always report the results of regression approaches together with a range of validity for the fitted relationship.
- Do not extrapolate empirical relationships unless a robust justification can be provided for doing so.

During review process

- The reviewers should look at the data.
- The reviewers should not systematically encourage the production of "standardized" statistical results as the only acceptable proof of what is being discussed.
- The statistics used must be consistent with the data on which the article is based.

CRediT authorship contribution statement

Damien Tedoldi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Formal analysis, Conceptualization. **Boram Kim:** Writing – review & editing, Visualization, Software, Formal analysis. **Santiago Sandoval:** Writing – review & editing, Validation, Formal analysis. **Nicolas Forquet:** Writing – review & editing, Validation, Software, Formal analysis. **Bruno Tassin:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are grateful to Julie Figueras for her meticulous proofreading of this paper and her helpful advice on its content.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2025.106526.

Data availability

Examples correspond to synthetic data (*i.e.*, generated with random sequence simulators). The R scripts allowing the replication of the calculations and figures are provided as supplementary material.

References

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9.

Andersson, A., 2021. Mechanisms for log normal concentration distributions in the environment. Sci. Rep. 11, 18075. https://doi.org/10.1038/s41598-021-97822-2.

Anscombe, F.J., 1973. Graphs in statistical analysis. Am. Statistician 27, 17–21. https://doi.org/10.1080/00031305.1973.10478966.

Bartley, M.L., Hanks, E.M., Schliep, E.M., Soranno, P.A., Wagner, T., 2019. Identifying and characterizing extrapolation in multivariate response data. PLoS One 14, e0225715. https://doi.org/10.1371/journal.pone.0225715.

Benisch, J., Helm, B., Bertrand-Krajewski, J.-L., Bloem, S., Cherqui, F., Eichelmann, U., Kroll, S., Poelsma, P., 2021. Operation and maintenance. In: Bertrand-Krajewski, J.-L., Clemens-Meyer, F., Lepot, M. (Eds.), Metrology in Urban Drainage and Stormwater Management: Plug and Pray. IWA Publishing, pp. 203–262. https://doi.org/10.2166/9781789060119_0203.

Berthouex, P.M., Brown, L.C., 2002. Statistics for Environmental Engineers, 2. Lewis Publishers, Boca Raton.

Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 47, 1287–1294. https://doi.org/10.2307/1911963.

Cantrell, C.A., 2008. Technical Note: review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems. Atmos. Chem. Phys. 8, 5477–5487. https://doi.org/10.5194/acp-8-5477-2008.

Deming, W.E., 1943. Statistical Adjustment of Data. Wiley, New York.

D'Haultfoeuille, X., Givord, P., 2014. La régression quantile en pratique. Econ. Stat. 471, 85–111. https://doi.org/10.3406/estat.2014.10484.

Dodge, Y., 2008. The concise encyclopedia of statistics. In: Springer Reference, 1st. ed. Springer, New York.

Dushoff, J., Kain, M.P., Bolker, B.M., 2019. I can see clearly now: Reinterpreting statistical significance. Methods Ecol. Evol. 10, 756–759. https://doi.org/10.1111/2041-210X.13159.

Farrar, T., 2024. Package 'skedastic': heteroskedasticity diagnostics for linear regression models. R Package Version 2.0.2. University of the Western Cape, Bellville, South Africa.

Francq, B.G., Govaerts, B.B., 2014. Measurement methods comparison with errors-in-variables regressions. From horizontal to vertical OLS regression, review and new perspectives. Chemometr. Intell. Lab. Syst. 134, 123–139. https://doi.org/10.1016/j.chemolab.2014.03.006.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol. 31, 337–350. https://doi.org/10.1007/s10654-016-0149-3.

Hahn, G.J., 1977. The hazards of extrapolation in regression analysis. J. Qual. Technol. 9, 159–165. https://doi.org/10.1080/00224065.1977.11980791.

Hauduc, H., Neumann, M.B., Muschalla, D., Gamerith, V., Gillot, S., Vanrolleghem, P.A., 2015. Efficiency criteria for environmental model quality assessment: a review and its application to wastewater treatment. Environ. Model. Software 68, 196–204. https://doi.org/10.1016/j.envsoft.2015.02.004.

Helsel, D.R., 2012. Statistics for censored environmental data using Minitab and R, Second. Statistics in Practice. Wiley, Hoboken, New Jersey.

Helsel, D.R., Hirsch, R.M., 2002. Statistical Methods in Water Resources - Techniques of Water Resources Investigations. U.S. Geological Survey, Reston, VA. Book 4, chapter A3.

Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: Principles and Practice, Third print edition. Otexts, Online Open-Access Textbooks, Melbourne, Australia.

Julian, P., Helsel, D.R., 2024. Package 'NADA2': data analysis for censored environmental data. R package version 1.1.6.

Koenker, R., Chernozhukov, V., He, X., Peng, L. (Eds.), 2017. Handbook of Quantile Regression, first ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781315120256.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. 69, 066138. https://doi.org/10.1103/PhysRevE.69.066138.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86. https://doi.org/10.1214/aoms/1177729694.

Lang, M., Pobanz, K., Renard, B., Renouf, E., Sauquet, E., 2010. Extrapolation of rating curves by hydraulic modelling, with application to flood frequency analysis. Hydrol. Sci. J. 55, 883–898. https://doi.org/10.1080/02626667.2010.504186.

Lee, L., 2020. Package 'NADA': nondetects and data analysis for environmental data. R package version 1, 6, 1.1.

Limpert, E., Stahel, W.A., Abbt, M., 2001. Log-normal distributions across the sciences: keys and clues. Bioscience 51, 341. https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.

Mikkonen, S., Pitkänen, M.R.A., Nieminen, T., Lipponen, A., Isokääntä, S., Arola, A., Lehtinen, K.E.J., 2019. Technical note: effects of uncertainties and number of data points on line fitting – a case study on new particle formation. Atmos. Chem. Phys. 19, 12531–12543. https://doi.org/10.5194/acp-19-12531-2019.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, Part I — a discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Razguliaev, N., Flanagan, K., Muthanna, T., Viklander, M., 2024. Urban stormwater quality: a review of methods for continuous field monitoring. Water Res. 249, 120929. https://doi.org/10.1016/j.watres.2023.120929.

Sales, G., Romualdi, C., 2011. *Parmigene* — a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics 27, 1876–1877. https://doi.org/10.1093/bioinformatics/btr274.

Simpson, S.L., Edwards, L.J., Muller, K.E., Sen, P.K., Styner, M.A., 2010. A linear exponent AR(1) family of correlation structures. Stat. Med. 29, 1825–1838. https://doi.org/10.1002/sim.3928.

Smits, A.P., Hall, E.K., Deemer, B.R., Scordo, F., Barbosa, C.C., Carlson, S.M., Cawley, K., Grossart, H., Kelly, P., Mammola, S., Pintar, M.R., Robbins, C.J., Ruhi, A., Saccò, M., 2025. Too much and not enough data: challenges and solutions for generating information in freshwater research and monitoring. Ecosphere 16, e70205. https://doi.org/10.1002/ecs2.70205.

Tomassone, R., Lesquoy, E., Millier, C., 1983. La régression: nouveaux regards sur une ancienne méthode statistique. Actualités Scientifiques et Agronomiques de l'Inra Institut National de la Recherche Agronomique. Masson, Paris.

Wainwright, J., Mulligan, M. (Eds.), 2004. Environmental Modelling: Finding Simplicity in Complexity. Wiley, Chichester, England; Hoboken, New Jersey.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838. https://doi.org/10.2307/1912934.

Xu, W., Hou, Y., Hung, Y.S., Zou, Y., 2013. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. Signal Process. 93, 261–276. https://doi.org/10.1016/j.sigpro.2012.08.005.

Young, P., Parkinson, S., Lees, M., 1996. Simplicity out of complexity in environmental modelling: occam's razor revisited. J. Appl. Stat. 23, 165–210. https://doi.org/10.1080/02664769624206.

Zeileis, A., Hothorn, T., 2002. Diagnostic checking in regression relationships. R. News 2/3, 7–10.