A generalised approach for identifying influential data in hydrological modelling

D. Wright, M. Thyer, S. Westra, Benjamin Renard, D. Mcinerney

A generalised approach for identifying influential data in hydrological modelling

Authors: David P. Wright[1], Mark Thyer[1], Seth Westra[1], Benjamin Renard[2], David McInerney[1]

1. **School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, 5005, Australia**
2. **Irstea, UR Riverly, Lyon-Villeurbanne center, 69625 Villeurbanne, France**

Corresponding Author: David P. Wright, david.p.wright@adelaide.edu.au

Submission to Environmental Modelling & Software

Highlights:

1. Influential data points have a disproportionate impact on model predictions
2. A new generalised Cook's distance accurately identifies influential data points
3. More efficient (<1% computational cost) than standard case-deletion approaches
4. Applies to nonlinear regression and hydrological models with heteroscedastic errors
5. Can be used in a Bayesian framework with priors or data uncertainty

Abstract:

Influence diagnostics are used to identify data points that have a disproportionate impact on model parameters, performance and/or predictions, providing valuable information for use in model calibration. Regression-theory influence diagnostics identify influential data by combining the leverage and the standardised residuals, and are computationally more efficient than case-deletion approaches. This study evaluates the performance of a range of regression-theory influence diagnostics on ten case studies with a variety of model structures and inference scenarios including: nonlinear model response, heteroscedastic residual errors, data uncertainty and Bayesian priors. A new technique is developed, generalised Cook's distance, that is able to accurately identify the same influential data as standard case deletion approaches (Spearman rank correlation: 0.93-1.00) at a fraction of the computational cost (<1%). This is because generalised Cook's distance uses a generalised leverage formulation which outperforms linear and nonlinear leverage formulations because it has less restrictive assumptions. Generalised Cook's distance has the potential to enable influential data to be efficiently identified on a wide variety of hydrological and environmental modelling problems.

Keywords: *hydrologic model calibration, influence diagnostics, Cook's distance, generalised leverage*

1. Introduction

Hydrological model calibration is a critical component of model development as parameters generally cannot be determined directly from measurements but are instead inferred indirectly by calibrating the hydrological model to observed hydrological responses (e.g. daily streamflow) [*Beven*, 2011]. Studies increasingly have called for the use of "influence diagnostics" [e.g., *Foglia et al.*, 2009; *Foglia et al.*, 2007; *Hill et al.*, 2015; *Wright et al.*, 2015] to understand the extent to which model calibration outcomes are determined by a small number of data points that may be erroneous or unrepresentative of overall catchment behaviour. For example, *Wright et al.* [2015] showed that removing a single value of daily streamflow from a two-year calibration period could change the predicted streamflow by more than 25% in a semi-arid catchment. There are a range of influence diagnostics in the literature that have been used to identify which points are influential; the goal of this paper is to evaluate a generalised approach to identifying influential points that is both accurate and computationally efficient.

Influence diagnostics can be categorised into two different classes: "case-deletion" influence diagnostics and "regression-theory" influence diagnostics (see Figure 1). Case-deletion influence diagnostics measure the influence by censoring ("deleting") a data point ("case") from the set of calibration points, then re-calibrating the model. Once case-deletion has been performed, several approaches can be used to measure influence. The first approach is to evaluate Cook's distance [*Cook*, 1977], which is a commonly used measure of influence [*Cook*, 1977] and has been used in a large variety of regression problems [*Fox and Weisberg*, 2011]. The second approach is to quantify the difference between original and re-calibrated model parameters, model performance (such as objective function displacement) and/or model predictions of interest [*Wright et al.*, 2015]. Two further approaches to measure influence are DFFITS and DFBETA [see *Cook and Weisberg*, 1982]. These are not considered further in this study because DFFITS is conceptually identical to Cook's distance (see *Cook and Weisberg* [1982]), and DFBETA describes the impact of influential data on individual model parameter estimates only [*Fox and Weisberg*, 2011], whereas Cook's distance has the flexibility to look at the impact of influential points on parameters (including their interactions) and predictions.

The case-deletion influence diagnostics are classified as "exact" because they make no assumptions regarding the type of regression model (linear/nonlinear) or the complexity of the residual error model (Gaussian, heteroscedastic, autocorrelated etc. - see *McInerney et al.* [2017]). This makes them particularly attractive for hydrological applications, where the hydrological models are generally nonlinear and assumptions related to the behaviour of the residuals, such as Gaussianity and homoscedasticity, are typically not supported by the data. The drawback with case-deletion based influence diagnostics is the high computational demand associated with re-estimating the parameters for every data point in the observed data (e.g. for a decade of daily data case-deletion requires ~3650 model re-calibrations). This renders influence analysis using case-deletion potentially infeasible for anything but the simplest hydrological models. A secondary issue with the case-deletion class is that anomalous results may arise when calibrating to complex response surfaces with multiple local optima [*Duan et al.*, 1992; *Kavetski et al.*, 2006], as each re-calibration may lead to parameter sets in different local optima. This may cause the case-deletion calibrated parameter sets to be different from each other, even if the data points have low influence on the actual model calibration. To address this issue the modeller may choose to increase the robustness of the optimisation; however, these efforts will compound the computational demands of the case-deletion re-calibrations.

In regression applications Cook's distance can alternatively be calculated using "regression-theory" influence diagnostics (see Figure 1). Regression-theory influence diagnostics have a significantly reduced computational demand as they do not require case-deletion re-calibration and instead rely on assumptions about the type of regression model (linear/nonlinear) and residual error model (Gaussian, homoscedastic etc.). The reduced computational demand is achieved by combining the following two components for each observed data point: (1) the leverage, which describes the rate of change of the predicted model output with respect to the corresponding observed output and can be used to assess the potential importance of individual observations [*Wei et al.*, 1998], and (2) the standardised residuals, which correspond to the raw residuals divided by the fitted standard deviation. By combining these two components to calculate Cook's distance, regression-theory influence diagnostics do not require additional re-calibrations and are therefore a more efficient alternative to the computationally demanding case-deletion influence diagnostics. There exist multiple alternative formulations of leverage, differing in the assumptions made about the fitted model and the probabilistic model of the residual errors. In circumstances where these assumptions are not violated regression-theory Cook's distance is equivalent to case-deletion Cook's distance.

Linear leverage is arguably the most widely used approach to approximate Cook's distance in regression problems [*Fox and Weisberg*, 2011], and is derived from standard linear regression theory and therefore inherits the assumptions of a linear model response (with respect to the model parameters) and Gaussian, homoscedastic and independent residual errors [*Cook and Weisberg*, 1982].When linear leverage is used in regression-theory Cook's distance (hereafter referred to as "linear Cook's distance") it also inherits these assumptions. This implies that linear Cook's distance may not be suitable for identifying the influential points in a hydrological modelling context as the hydrological model calibration violates the assumptions of linear regression, as a result of: 1) nonlinear model response [e.g. see discussion in *Kavetski and Kuczera*, 2007], and 2) heteroscedastic and non-Gaussian residual errors [e.g. see *Schoups and Vrugt*, 2010].

To address these limitations and expand the applicability of regression-theory influence diagnostics to more complex situations, *St. Laurent and Cook* [1992] proposed nonlinear leverage. Calculating Cook's distance by applying nonlinear leverage (hereafter referred to as "nonlinear Cook's distance") can take into account nonlinear model response, and is suitable for nonlinear models with Gaussian residuals. *Wright et al.* [2015] applied both linear and nonlinear Cook's distance in a hydrological modelling context and found that nonlinear Cook's distance provided higher performance than linear Cook's distance, in terms of a higher correlation with the influential points identified using case-deletion influence diagnostics. The limitation of *Wright et al.* [2015] is that the hydrological models were calibrated using a standard least squares objective function, which is known to perform poorly in a hydrological modelling context when the residual errors are non-Gaussian and/or heteroscedastic [see *McInerney et al.*, 2017].

To overcome the limitations of the assumptions of linear and nonlinear leverage, generalised leverage was developed by *Wei et al.* [1998]. Generalised leverage makes no assumptions of linear model response, and can be applied to a broad range of objective functions, including those with heteroscedastic and/or non-Gaussian residual error assumptions. It has been used in numerous regression applications [e.g. *Leiva et al.*, 2014; *Lemonte and Bazán*, 2015; *Osorio*, 2016; *Rocha and Simas*, 2011]; however, it has not been applied in the context of hydrological or other environmental modelling applications. Furthermore, generalised leverage is typically used as a standalone diagnostic and has not previously been applied as an input to calculate Cook's distance (hereafter referred to as "generalised Cook's distance") to identify influential points. This research gap presents an opportunity to determine if generalised Cook's distance can be used as an efficient approach to approximate case-deletion Cook's distance in a computationally frugal manner.

Given the substantial computational advantages of regression-theory influence diagnostics over case-deletion influence diagnostics, they show significant promise for application in the field of hydrological and other environmental modelling applications. However, before regression-theory influence diagnostics can be applied, the validity of the assumptions of the formulations of leverage will first need to be experimentally tested in the context of hydrological case-studies. An important issue to be investigated is the hypothesis that generalised leverage can be used to approximate case-deletion Cook's distance as it has not previously been combined with standardised residuals to measure the proposed generalised Cook's distance. This study will assess the performance of the different approaches within the class of regression-theory influence diagnostics (i.e. linear Cook's distance, nonlinear Cook's distance, and generalised Cook's distance) to reproduce case-deletion Cook's distance. The specific objectives of this study are to evaluate the ability of regression-theory influence diagnostics to identify influential points under the following modelling scenarios:

1. Linear and nonlinear regression models with either homoscedastic or heteroscedastic residual error;
2. A daily hydrological model including nonlinear model response and storage with heteroscedastic residual error; and
3. A stage-discharge rating curve model with Bayesian objective functions that include heteroscedastic residual error, data uncertainty and prior information.

For all three objectives, the regression-theory Cook's distance obtained using the linear, nonlinear and generalised leverage formulations will be compared to the case-deletion Cook's distance, in order to evaluate the extent to which the specific leverage formulation affects the performance of regression-theory influence diagnostics. The remainder of this paper is structured as follows. In Section 2 we describe the methodology, in Section 3 we introduce the three case studies selected to address the study objectives, and in Section 4 we apply the influence diagnostics to these case studies. In Section 5 we discuss the advantages and disadvantages of case-deletion and regression-theory influence diagnostics, the suitability of applying generalised Cook's distance to a broader class of hydrological and environmental models, and the future need to understand the key drivers of influential data.

2. Methodology

Influence diagnostics identify data points that exert a disproportionate impact on calibrated parameters, performance and/or predictions. In this study we consider the following classes of Cook's distance influence diagnostics:

1. Case-deletion based Cook's distance, which measures the influence of a single point by comparing model parameters, performance and/or predictions from calibration with and without that data point; and
2. Regression-theory influence diagnostics, which measure influence by combining the standardised residual and the leverage of each data point. We analyse and compare three approaches to determining the leverage, which produce three estimates of Cook's distance:
    i. Linear Cook's distance, which uses linear leverage,
    ii. Nonlinear Cook's distance, which uses nonlinear leverage, and
    iii. Generalised Cook's distance, which uses generalised leverage.

In this section we first introduce the general modelling framework, and then define the influence diagnostics, leverage, and the objective functions used in this study. We finish by describing the metrics that we will use to evaluate the performance of the regression-theory influence diagnostics.

2.1. General model framework

We define the general model response as:

$$\mathbf{y} = f\left(\boldsymbol{\alpha};\mathbf{X}\right)+\boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}=\left(y_1, y_2,..., y_n\right)$ is a vector of n observed responses, $f\left(.\right)$ is the model structure, $\boldsymbol{\alpha}=\left(\alpha_1,\alpha_2,...,\alpha_{m_\alpha}\right)$ is a vector of m_α model parameters, $\mathbf{X}$ is an $n \times k$ matrix of k observed inputs (e.g., precipitation, potential evapotranspiration (PET)), and $\boldsymbol{\varepsilon}$ is a vector of n residual errors. Residuals are further assumed to be realisations from a given probability distribution, with parameters $\boldsymbol{\beta}=\left(\beta_1,\beta_2,...,\beta_{m_\beta}\right)$ (e.g. a centred Gaussian distribution with unknown standard deviation). Thus, the entire set of m parameters to be calibrated are $\boldsymbol{\theta}=\left\{\boldsymbol{\alpha},\boldsymbol{\beta}\right\}$ which includes both the model parameters $\boldsymbol{\alpha}$ and the residual error model parameters $\boldsymbol{\beta}$.

2.1.1. Objective functions

In order to apply leverage to a broad class of objective functions used in hydrological modelling we consider the general form of the objective function, as suggested by *Wei et al.* [1998]:

$$\Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X})=\sum_{i=1}^{n}\rho_i\left(f_i\left(\boldsymbol{\alpha};\mathbf{X}\right),\boldsymbol{\beta};y_i\right) \tag{2}$$

where $\rho_i\left(.\right)$ is a function that describes the contribution of the i^{th} data point to the objective function, $f_i\left(\boldsymbol{\alpha};\mathbf{X}\right)$ is the i^{th} model prediction, $\Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X})$ and $f\left(\boldsymbol{\alpha};\mathbf{X}\right)$ are assumed to be twice differentiable with respect to $\boldsymbol{\theta}$ and $\mathbf{y}$. We will denote $\boldsymbol{\theta}$ as the model parameters that maximise Φ in equation (2), and $\mathbf{y}$ as the predicted response associated with $\boldsymbol{\theta}$, i.e. $\mathbf{y}=f\left(\hat{\boldsymbol{\alpha}};\mathbf{X}\right)$.

The generalised form in equation (2) can be adapted to a number of well-known objective functions in hydrological modelling as outlined in Section 2.4.

2.1.2. Standardised residuals

The standardised residuals, $\mathbf{v}$, which are required to estimate the regression-theory influence diagnostics introduced in Section 2.2.2, are obtained by dividing the raw residuals $\boldsymbol{\varepsilon}=\mathbf{y}-\mathbf{y}$ by their calibrated standard deviations, $\boldsymbol{\sigma}$:

$$\mathbf{v}=\frac{\boldsymbol{\varepsilon}}{\boldsymbol{\sigma}} \tag{3}$$

The vector $\boldsymbol{\sigma}$ is determined based on the assumed residual error model and the resultant objective function (see Section 2.4 for further details).

2.2. Influence diagnostics

This section provides a detailed description of the different influence diagnostics used in this study (see Figure 1 for an overview). Firstly, we present the case-deletion "class" of influence diagnostics and outline the approach used to calculate case-deletion Cook's distance. Secondly, we present the regression-theory "class" of influence diagnostics and outline the approaches used to calculate regression-theory Cook's distance using three formulations of leverage (i.e. linear, nonlinear and generalised leverage) to produce linear, nonlinear and generalised Cook's distance.

2.2.1. Case-deletion influence diagnostics

Case-deletion influence diagnostics describe the influence of masking a data point in model calibration and assessing the change to the model predictions, parameters and/or objective function value. Cook's distance can be measured exactly using case-deletion [see *Cook and Weisberg*, 1982]; note that in the statistical literature this case-deletion Cook's distance is sometimes referred to as "generalised Cook's distance" [e.g. *Das*, 2008]. Case-deletion based Cook's distance measures influence by comparing model predictions $\mathbf{y}$ based on using all of the calibration data and model predictions $\mathbf{y}^{(-i)}$ with the i^{th} point masked from the calibration data. For a given data point, case-deletion based Cook's distance is calculated by:

$$CD_i = \sum_{j=1}^{n} \frac{\left(\hat{y}_j - \hat{y}_j^{(-i)}\right)^2}{m \times \hat{\sigma}_j^2} \tag{4}$$

where σ_j is the calibrated standard deviation for the j^{th} data point, estimated from using all calibration data (i.e. $\mathbf{y}$).

2.2.2. Regression-theory influence diagnostics

Regression-theory influence diagnostics avoid the computational burden of case-deletion re-calibration by making assumptions about the type of response model (linear/nonlinear) and residual error model (Gaussian, homoscedastic etc.). Regression-theory Cook's distance is calculated by combining the standardised residual of the i^{th} point (ν_i) with the leverage of i^{th} observation on the i^{th} prediction (L_{ii}) to give [*Cook and Weisberg*, 1982; *Fox and Weisberg*, 2011]:

$$CD_i = \frac{\nu_i^2}{m} \frac{L_{ii}}{(1-L_{ii})^2} \tag{5}$$

The approach used to determine the three different forms of Cook's distance (i.e. linear, nonlinear and generalised Cook's distance; Figure 1) is based on the corresponding forms of leverage (i.e. linear, nonlinear, and generalised leverage). In the next section, we provide a general definition of leverage followed by the three specific formations of leverage that are used to calculate regression-theory Cook's distance.

2.3. Leverage

Leverage generally can be defined as the rate of the change of the i^{th} predicted value, y_i, with respect to another j^{th} observed value, y_j [*Cook and Weisberg*, 1982; *Hoaglin and Welsch*, 1978; *St. Laurent and Cook*, 1992; *Wei et al.*, 1998]:

$$L_{ij} = \partial \hat{y}_i / \partial y_j \tag{6}$$

or in matrix notation:

$$\mathbf{L} = \frac{\partial \mathbf{y}}{\partial \mathbf{y}^T} \tag{7}$$

where $\mathbf{L}$ is an $n \times n$ matrix. The diagonal elements L_{ii} most directly reflect the impact of y_i on the model fit [*Cook and Weisberg*, 1982; *Hoaglin and Welsch*, 1978; *St. Laurent and Cook*, 1992], and are used for calculating regression-theory Cook's distance (Section 2.2.2).

2.3.1. Linear leverage

Linear leverage inherits the assumptions of standard linear regression theory; i.e. that the model response (with respect to the parameters) is linear and that residual errors are Gaussian, homoscedastic and independent. Under the assumptions of linear regression the general form of leverage in equation (7) can be expressed as $\mathbf{L}$ [*Fox and Weisberg*, 2011]:

$$\mathbf{L} = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}} \tag{8}$$

As linear leverage depends solely on the observed input $\mathbf{X}$, it can be calculated without model calibration using linear algebra. In a linear regression model with standard least squares (SLS) residual errors, regression-theory Cook's distance is equivalent to case-deletion Cook's distance [see *Cook*, 1977].

2.3.2. Nonlinear leverage

Nonlinear leverage does not assume a linear model response but retains the assumption that residual errors are Gaussian, homoscedastic and independent. Nonlinear leverage is dependent on the local sensitivity of the model predictions to small perturbations in model parameters [*St. Laurent and Cook*, 1992]. Nonlinear leverage is calculated after model calibration, and under the assumptions of nonlinear regression the general form of leverage in equation (7) can be expressed as $\mathbf{L}(\boldsymbol{\alpha})$ [*St. Laurent and Cook*, 1992; 1993; *Wei et al.*, 1998; *Wright et al.*, 2015]:

$$\mathbf{L}(\boldsymbol{\alpha}) = \frac{\partial f(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}}\left(\left(\frac{\partial f(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}}\right)^{T}\frac{\partial f(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}} - \sum_{i=1}^{n}\left(\left(y_i - y_i\right)\frac{\partial^2 f_i(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}^2}\right)\right)^{-1}\left(\frac{\partial f(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}}\right)^{T} \tag{9}$$

where $\frac{\partial f(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}}$ is the $n \times m_\alpha$ Jacobian matrix with i^{th} row $\frac{\partial f_i(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}}$, and $\frac{\partial^2 f_i(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\alpha}^2}$ is the $m_\alpha \times m_\alpha$ Hessian matrix associated with the i^{th} data point. Analytical derivatives are typically not available for hydrological models, and therefore we obtain estimates of the derivatives from central-difference numerical approximation [*Nocedal and Wright*, 2006]. When applied to a linear regression model with SLS residual errors, the nonlinear leverage simplifies to linear leverage, as shown in *Wei et al.* [1998].

2.3.3. Generalised leverage

Generalised leverage makes no assumptions of linear model response, and can be applied to a general class of regression models and a broad range of objective functions, including those with heteroscedastic and/or non-Gaussian residual error assumptions. Generalised leverage is calculated after model calibration and takes into account the curvature of the objective function about the whole set of calibrated parameters $\boldsymbol{\theta}$. In this case the general form of leverage in equation (7) can be expressed as $\mathbf{L}(\boldsymbol{\theta})$ [*Wei et al.*, 1998]:

$$\mathbf{L}(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\theta}} \left(-\frac{\partial^2 \Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X})}{\partial \boldsymbol{\theta}^2} \right)^{-1} \frac{\partial^2 \Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X})}{\partial \boldsymbol{\theta} \partial \mathbf{y}^T} \tag{10}$$

where $\frac{\partial f(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\theta}}$ is the $n \times m$ Jacobian matrix with i^{th} row $\frac{\partial f_i(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\theta}}$ (note that $\frac{\partial f_i(\boldsymbol{\alpha};\mathbf{X})}{\partial \boldsymbol{\beta}} = 0$), $\frac{\partial^2 \Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X})}{\partial \boldsymbol{\theta}^2}$ is a $m \times m$ Hessian matrix and $\frac{\partial^2 \Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X})}{\partial \boldsymbol{\theta} \partial \mathbf{y}^T}$ is a $m \times n$ matrix. Generalised leverage can be applied to any objective function that takes the general form in equation (2). Generalised leverage simplifies to nonlinear leverage in the case of a nonlinear regression model and SLS residual errors, as shown in *Wei et al.* [1998].

2.4. Objective functions used in this study

This section introduces the range of different objective functions that will be used in the case studies to evaluate the performance of the differing implementations of regression-theory Cook's distance.

2.4.1. Standard least squares

Assuming independent and identically distributed (i.i.d.) Gaussian residual errors, the following log likelihood can be used as an objective function:

$$\Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \log(p_N(y_i - f_i(\boldsymbol{\alpha};\mathbf{X}) \mid 0,\sigma^2)) \tag{11}$$

where $p_N(x \mid \mu,\sigma^2)$ is the Gaussian probability density at x assuming constant mean μ and variance σ^2. As the standard deviation σ is unknown it will be estimated, and therefore we have $\beta = \{\sigma\}$. Note that the Nash-Sutcliffe efficiency [*Nash and Sutcliffe*, 1970] objective function that is commonly applied in hydrological calibration corresponds to the assumptions of constant-variance and Gaussian residual errors of the standard least squares (SLS) objective function. Note that (11) is a particular case of the general objective function in equation (2).

2.4.2. Weighted least squares

Residual errors in hydrological applications are generally heteroscedastic [see *Schoups and Vrugt*, 2010; *Sorooshian and Dracup*, 1980] and to account for this non-constant variance we apply a weighted least squares (WLS) objective function. Due to this heteroscedasticity in hydrological residual errors it is common to replace the constant standard deviation σ in equation (11) with a standard deviation $\boldsymbol{\sigma}$ that varies in time, so that the non-constant variance acts as a "weight" for each residual [e.g. *McInerney et al.*, 2017; *Thyer et al.*, 2009]. A common covariate for modelling heteroscedasticity in streamflow errors is the predicted streamflow itself [e.g. *Schoups and Vrugt*, 2010; *Thyer et al.*, 2009]. Following *Evin et al.* [2014] we consider the standard deviation of residuals to be a linear function of simulated streamflow, such that:

$$\boldsymbol{\sigma} = \beta_1 \mathbf{y} + \beta_2 \tag{12}$$

The objective function becomes:

$$\Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \log(p_N(y_i - f_i(\boldsymbol{\alpha};\mathbf{X}) \mid 0, \sigma_i^2)) \qquad (13)$$

As the parameters describing the non-constant standard deviation (i.e. $\boldsymbol{\beta} = \{\beta_1, \beta_2\}$) are unknown they will need to be estimated. Note that (12) is a particular case of the general objective function in equation (2).

2.4.3. Weighted least squares with data uncertainty

In circumstances when independent estimates of data errors are available we may wish to distinguish between heteroscedasticity in hydrological residual errors and uncertainty of observed responses. To implement the WLS method with discharge uncertainty in the WLS objective function (13) we assume that the total errors can be decomposed as the sum of two independent error terms: the “structural errors” that can be described using the WLS standard deviation $\boldsymbol{\sigma}_r = \beta_1 \mathbf{y} + \beta_2$ and the “measurement errors” described using known standard deviations $\boldsymbol{\sigma}_Y$. The latter standard deviations may be derived from an uncertainty analysis of measured responses, which can be performed before and independently from the model calibration. The standard deviation of the total error, combining structural and measurement errors, is therefore equal to $\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}_r^{\,2} + \boldsymbol{\sigma}_Y^{\,2}}$. Hence the σ_i in equation (13) becomes:

$$\sigma_i = \sqrt{\left(\beta_1 y_i + \beta_2\right)^2 + \sigma_{y,i}^2} \qquad (14)$$

where $\sigma_{y,i}$ is the standard deviation of the measurement errors at time step i.

2.4.4. Weighted least squares with priors

In circumstances when prior information about parameter values is available based on previous studies and/or from analysis of physical characteristics that govern the relation between inputs $\mathbf{X}$

and outputs $\mathbf{y}$ we can use an objective function that combines WLS likelihood with priors. Bayes' equation yields the posterior probability distribution of the hydrological and residual error model parameters as follows:

$$\underbrace{\mathrm{p}\left(\boldsymbol{\theta}\middle|\mathbf{X},\mathbf{y}\right)}_{posterior} \propto \underbrace{\mathrm{p}\left(\mathbf{y}\middle|\boldsymbol{\theta},\mathbf{X}\right)}_{likelihood} \underbrace{\mathrm{p}\left(\boldsymbol{\theta}\right)}_{prior} \tag{15}$$

where $\mathrm{p}\left(\boldsymbol{\theta}\middle|\mathbf{X},\mathbf{y}\right)$ is the posterior probability of parameter $\boldsymbol{\theta}$ given $\mathbf{X}$ and $\mathbf{y}$, $\mathrm{p}\left(\boldsymbol{\theta}\right)$ is the joint prior probability density of hydrological and residual error model parameters, and $\mathrm{p}\left(\mathbf{y}\middle|\boldsymbol{\theta},\mathbf{X}\right)$ is the likelihood of $\mathbf{y}$ given $\boldsymbol{\theta}$ and $\mathbf{X}$. Taking the logarithm of equation (15) we obtain:

$$\log\left(\mathrm{p}\left(\boldsymbol{\theta}\middle|\mathbf{X},\mathbf{y}\right)\right) = \log\left(\mathrm{p}\left(\mathbf{y}\middle|\boldsymbol{\theta},\mathbf{X}\right)\right) + \log\left(\mathrm{p}\left(\boldsymbol{\theta}\right)\right) + c \tag{16}$$

where c is a constant. Assuming independence between residuals we can formulate the objective function as:

$$\begin{aligned}\Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X}) &= \log\left(\mathrm{p}\left(\mathbf{y}\middle|\boldsymbol{\theta},\mathbf{X}\right)\right) + \log\left(\mathrm{p}\left(\boldsymbol{\theta}\right)\right) \\ &= \sum_{i=1}^{n} \log\left(\mathrm{p}\left(y_i\middle|\boldsymbol{\theta},\mathbf{X}\right)\right) + \log\left(\mathrm{p}(\boldsymbol{\theta})\right) \\ &= \sum_{i=1}^{n}\left(\log\left(\mathrm{p}\left(y_i\middle|\boldsymbol{\theta},\mathbf{X}\right)\right) + \frac{1}{n}\log\left(\mathrm{p}(\boldsymbol{\theta})\right)\right)\end{aligned} \tag{17}$$

Assuming the residual errors are heteroscedastic with $\boldsymbol{\sigma}$ given by equation (12) and independent priors, we obtain the following objective function:

$$\Phi(\boldsymbol{\theta};\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n}\left\{\log(p_N(y_i - f_i(\boldsymbol{\alpha};\mathbf{X}) \mid 0, \sigma_i^2)) + \frac{1}{n}\sum_{j=1}^{p}\log\left(\mathrm{p}(\theta_j)\right)\right\} \tag{18}$$

where the contributions to the objective function from the priors are split evenly across the n points in the calibration data.

2.4.5. Weighted least squares with data uncertainty and priors

In circumstances when both independent estimates of data errors and prior information about parameter values are available we can use weighted least squares with data uncertainty and priors. Similar to Section 2.4.3, data uncertainty can readily be included in the weighted least squares with priors objective function (18) by simply using $\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}_r{}^2 + \boldsymbol{\sigma}_Y{}^2}$, where $\boldsymbol{\sigma}_r = \beta_1 \mathbf{y} + \beta_2$, and $\boldsymbol{\sigma}_Y$ are known values representing the measurement uncertainty in observed responses.

2.5. Performance metrics

As case-deletion Cook's distance provides a measure of influence with no assumptions regarding the type of model (linear/nonlinear) or the complexity of the residual error model (Gaussian, heteroscedastic, etc.) we use it as a baseline to compare the three formulations of regression-theory influence diagnostics: linear Cook's distance, nonlinear Cook's distance and generalised Cook's distance. We use two metrics to assess the performance of regression-theory influence diagnostics with respect to case-deletion based Cook's distance. These metrics are evaluated on 1) the whole set of influential data points, to show the general ability of regression-theory influence diagnostics to approximate case-deletion Cook's distance; and 2) a subset comprising the 10 most influential data points identified by case-deletion Cook's distance, to highlight the performance with respect to the points that are most influential to calibration. The metrics are:

1. Spearman correlation (Sp. and Sp._{10}), which provides a measure of the performance of the regression-theory influence diagnostics to correctly rank the most influential data points.
2. Coefficient of determination (r^2 and r^2_{10}), which provides a measure of the proportion of the variance in the case-deletion based variable that is accounted for by the regression-theory variable.

The selected performance metrics allow for a thorough comparison of the regression-theory influence diagnostics as approximations of the case-deletion Cook's distance.

3. Case studies

The research objectives of this paper are to evaluate the ability of regression-theory influence diagnostics to identify influential points under the following modelling scenarios:

1. Linear and nonlinear regression models with either homoscedastic or heteroscedastic residual error;
2. A daily hydrological model including nonlinear model response and storage with heteroscedastic residual error; and
3. A stage-discharge rating curve model with Bayesian objective functions that include heteroscedastic residual error, data uncertainty and prior information.

In order to address these objectives we apply case-deletion and regression-theory influence diagnostics to ten different case studies, organised in three distinct case study sets (Table 1). To address the first research objective the first case study set consists of four synthetic regression models, $A_{1\text{-}4}$, are selected to test the performance with linear/nonlinear regression models and homoscedastic/heteroscedastic residual error models. The second research objective is addressed by case study set 2, which tests the performance with daily hydrological models, $B_{1\text{-}2}$, with nonlinear hydrological response, model storage, and heteroscedastic residual errors. Finally, the third objective is addressed by case study set 3, which tests the performance with four different rating curve models, $C_{1\text{-}4}$, with and without data uncertainty and with and without prior knowledge specified using a Bayesian inference approach.

In all cases the objective functions are optimised using the Shuffled Complex Evolution (SCE) search algorithm [*Duan et al.*, 1992; *Duan et al.*, 1994] followed by a Nelder-Mead gradient search from the SCE optimised parameter set to machine precision to ensure convergence to the optima.

3.1. Case study set 1: Synthetic regression models with linear/nonlinear response and homoscedastic/heteroscedastic residual errors

The first case study set uses synthetic regression models that range in complexity from a simple linear model response with homoscedastic residual errors to a nonlinear power model response with heteroscedastic residual errors. The regression models with synthetic data (A_{1-4}; Table 1) are selected to highlight the role of model structure and residual error model on the influence results: A_1 has a linear model response with a standard least squares (SLS) residual error model; A_2 also has a linear model response but with a weighted least squares (WLS) residual error model; and both A_3 and A_4 have a nonlinear model response with SLS and WLS residual error, respectively.

3.2. Case study set 2: Daily hydrological model with synthetic and observed streamflow and heteroscedastic residual errors

The next case study set tests the performance of the regression-theory influence diagnostics in a typical hydrological modelling calibration context. We apply a daily hydrological model that includes nonlinear model response and storage (meaning that inputs at a given time-step can affect outputs many time-steps into the future) and heteroscedastic residual errors. The daily lumped hydrological model GR4J [*Perrin et al.*, 2003] was selected based upon its popularity [e.g. *Andréassian et al.*, 2014; *Evin et al.*, 2014; *Le Moine et al.*, 2007; *Lebecherel et al.*, 2016; *Wright et al.*, 2015] and parsimonious model structure. This allows for computational efficiency in the case-deletion model runs required to calculate case-deletion Cook's distance. The GR4J hydrological model has model parameters $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$, where α_1 is the maximum capacity of the production store, α_2 is the groundwater exchange coefficient, α_3 is the maximum capacity of the routing store, and α_4 is the time base of unit hydrograph.

We apply the GR4J hydrological model to the French Broad River catchment in North Carolina, USA. The French Broad River has a catchment area of 2448 km^2, annual precipitation of 1413 mm and annual streamflow of 800 mm, leading to a runoff coefficient of 0.57.

We explore two alternative modelling scenarios $B_{1\text{-}2}$ (Table 1) that correspond to synthetic streamflow data and real observed streamflow data, respectively. We use three years of calibration data, from 1974 to 1976. The first model B_1 uses the observed rainfall and PET from the French Broad River but has synthetic streamflow data. This synthetic streamflow data is obtained by first using real streamflow data to fit the GR4J parameters, then using the fitted parameters to generate a predicted streamflow time series, and finally adding residual errors to the predicted time series based on the WLS error model. The second hydrological model B_2 also uses observed rainfall and PET from the French Broad River catchment, but is calibrated to the real observed streamflow data. Note that while there are two inputs for GR4J (i.e. rainfall and PET), here we consider only the importance of rainfall data (i.e. don't include PET in $\mathbf{X}$) when calculating leverage, because typically hydrological model response are more sensitive to errors in rainfall, rather than errors in PET [*Oudin et al.*, 2006] .

3.3. Case study set 3: Rating curve model incorporating heteroscedastic residual errors, data uncertainty and parameter priors

The final case study set uses a rating curve model, with increasing complexity in the objective function that investigates the impact of data uncertainty and incorporating parameter priors using a Bayesian approach. We apply a piecewise stage-discharge rating curve model to the Ardèche River at Sauze, France. The Ardèche River has a catchment area of 2240 km^2 with a mean annual discharge of 63 m^3/s. We use the reduced subset of 38 stage-discharge gaugings applied in *Le Coz et al.* [2014]. The flow at the hydrometric station is controlled by a rectangular sill at low flows, and a rectangular channel at high flows, leading to a two-part rating curve model with the following stage-discharge relationship:

$$f(\boldsymbol{\alpha}, \mathrm{X}_i) = \begin{cases} a_1 \left(\mathrm{X}_i - b_1\right)^{c_1}, \text{ for } \mathrm{X}_i < k \\ a_2 \left(\mathrm{X}_i - b_2\right)^{c_2}, \text{ for } \mathrm{X}_i \geq k \end{cases} \tag{19}$$

Here $\mathbf{X}$ is stage and $\boldsymbol{\alpha} = \{a_1, b_1, c_1, k, a_2, c_2\}$ are the rating curve model parameters similar to *Le Coz et al.* [2014]. As the rating curve is continuous at the knot (k), the parameter b_2 is computed from the other calibrated parameter values by solving the continuity condition $a_1(k-b_1)^{c_1} = a_2(k-b_2)^{c_2}$, yielding $b_2 = k - \left((a_1/a_2)(k-b_1)^{c_1} \right)^{1/c_2}$. *Petersen-Øverleir* [2004] suggest a heteroscedastic residual error model to take into account the heteroscedasticity of most rating curve errors, and so we use the WLS objective function described in Section 2.4.2. We apply the following four calibration schemes across $C_{1\text{-}4}$: 1) baseline rating curve calibration with WLS in C_1; 2) rating curve calibration with discharge uncertainty in C_2; 3) rating curve with priors in C_3; and 4) rating curve calibration with discharge uncertainty and priors in C_4.

We follow *Le Coz et al.* [2014] who provide gauging uncertainties for the discharge data at Sauze and also a framework for Bayesian inference. In C_3 and C_4 we use the priors from *Le Coz et al.* [2014] for the model parameters that are summarised in Table 2. Perusal of Table 2 shows that the prior standard deviation is smallest for the exponent parameters (c_1 and c_2 in equation (19)), compared with the scaling parameters, a_1 and a_2, and the offset parameters, b_1 and b_2. Hence the priors are more informative for these exponent values because they only depend on the type of hydraulic control (here, rectangular sill and rectangular channel). In the case of the residual error model parameters $\boldsymbol{\beta}$ there is no prior knowledge and so an uninformative uniform distribution is applied.

4. Assessing the ability of regression-theory Cook's distance to reproduce case-deletion Cook's distance

We apply case-deletion and regression-theory influence diagnostics with linear, nonlinear and generalised Cook's distance to the three case studies in Sections 4.1-4.3. In Section 4.4 we summarise the performance of the regression-theory influence diagnostics across the case studies, and we finish in Section 4.5 with an analysis of the computation times of both the regression-theory and case-deletion based influence diagnostics.

4.1. Case study set 1: Synthetic regression models with linear/nonlinear response and homoscedastic/heteroscedastic residual errors

In this section we evaluate the performance of regression-theory Cook's distance based on the three formulations of generalised leverage, using synthetic regression case studies with varying degrees of nonlinear model response and heteroscedastic residual errors ($A_{1\text{-}4}$; Table 1). The synthetically generated "observed" data and fitted models are presented in Figure 2 (row 1) for the four cases. The models are correctly specified, and fit the data well in all cases. This is evidenced by the standardised residuals being independent and normally distributed, with zero mean and unit standard deviation (Figure 2, row 2).

Similarities and differences between the three leverage formulations are shown in Figure 2 (row 3). Linear leverage is smooth and parabolic in all four cases ($A_{1\text{-}4}$), with a minima at the mean of $\mathbf{X}$ (~100). This highlights that linear leverage only depends on input $\mathbf{X}$ (which is identical in all four cases), and therefore does not vary with the case study. Nonlinear leverage is the same as linear leverage for linear response models A_1 and A_2, but differs for nonlinear response models A_3 and A_4. In those cases, the nonlinear model response results in higher leverage for larger values of $\mathbf{X}$, with a slight increase in the midrange of $\mathbf{X}$ for A_3, and with leverage varying smoothly as a function of $\mathbf{X}$. Interestingly, the nonlinear leverage for case A_3 is different to the nonlinear leverage for A_4. This is due to slightly different calibrated parameter values $\hat{\boldsymbol{\alpha}}$ for the nonlinear model in A_3 compared with A_4; if these calibrated parameter values were identical, the nonlinear leverage in equation (9) would be the same, since it is a function of input data $\mathbf{X}$, model response *f*(), and optimal model parameters $\hat{\boldsymbol{\alpha}}$. This highlights the sensitivity of nonlinear leverage to influential data points, despite the observations $\mathbf{y}$ not appearing explicitly in equation (9). Finally, generalised leverage is the same as nonlinear leverage for cases A_1 and A_3, when residuals are homoscedastic. However, when heteroscasticity in residuals is introduced into the "observations" and likelihood functions (cases A_2 and A_4), we see there are two major differences. The first difference is that generalised leverage becomes larger than nonlinear leverage for small values of $\mathbf{X}$. This is because generalised leverage accounts for the higher weights (i.e. smaller standard deviations) placed on low values of $\mathbf{Y}$ in the WLS likelihood function (which correspond to small values of $\mathbf{X}$), while nonlinear leverage applies the same weight to all values of $\mathbf{Y}$. The second differences is that unlike linear and nonlinear leverage, generalised leverage does not vary smoothly as a function of $\mathbf{X}$. This is because for a given point i, the generalised leverage in equation (10) depends on the observation at that point y_i, and the observations $\mathbf{y}$ do not vary smoothly with $\mathbf{X}$.

The magnitude of the case-deletion Cook's distance is presented in Figure 2 (row 4) as grey bubbles, and compared to the regression-theory Cook's distance (which combines the leverage and standardised residuals, equation (5)) in Figure 2 (row 5) as a function of $\mathbf{X}$. The differences between case-deletion Cook's distance and the three regression-theory Cook's distances are also quantified in Figure 3. The three regression-theory Cook's distances are identical for case A_1, as a result of identicial leverages. The errors between the regression-theory Cook's distance and case-deletion Cook's distance are small (green, purple and orange bubbles are all similarly small in Figure 2, column 1, row 5) and the correlations are high (as evidence by r^2 values and Spearman correlations of 1.00 when calculated over all data and the top 10 most influential points in Figure 3, column 1).

When heteroscedastic residual errors are introduced (case A_2), generalised Cook's distance becomes the most accurate approximation (green bubbles show lower errors then purple bubbles in Figure 2, column 2, row 5), with linear and nonlinear Cook's distance being the same (purple bubbles overlay orange bubbles). For linear and nonlinear Cook's distance, performance is worst at the extremes of $\mathbf{X}$, and particularly the lower values of , $\mathbf{X}$ as they do not account for residual heteroscedasticity. The increased accuracy of using generalised Cook's distances is seen in the top 10% of influential points (Figure 3, column 2, row 2) where—relative to the other leverage formulations—the Spearman correlation increases from 0.65 to 0.96, and the r^2 increases from 0.28 to 0.98.

The nonlinear response with homoscedastic residual errors (case A_3) results show identical performance for the nonlinear and generalised Cook's distances, which are typically more accurate than linear Cook's distance (green and purple bubbles have lower errors than orange bubbles in Figure 2, column 3, row 5). Linear Cook's distance performs particular poorly for high values of $\mathbf{X}$, as anticipated based on the leverage results. The largest improvement is obtained by using nonlinear and generalised Cook's distances is seen in the top 10% of influential points (Figure 3, column 3, row 2) where the Spearman correlation increases from 0.75 to 1.00, and the r^2 increases from 0.50 to 1.00.

Finally, the nonlinear model response with heteroscedastic residual errors (case A_4) results show that the generalised Cook's distance is the most accurate of the regression-theory Cook's distances (green bubbles show the lowest error in Figure 2, column 3, row 5). Both Spearman correlation and r^2 values are close to unity in all cases except for the Spearman correlation value for the largest 10% of influential points (Sp. = 0.79), due to a difference in a single point - the largest Cook's distance value. The ranking of the performance linear and nonlinear Cook's distance for this case appears to depend on $\mathbf{X}$ and the accuracy metrix used (abs. errors, correlation or spearman rank on all or top 100 data points). Neither of these leverage approaches, produce the consistent accuracy of generalised Cook's distance.

Overall, the results indicate that for the four synthetic regression model case studies considered, generalised Cook's distance provides a very close approximation of case-deletion Cook's distance, and represents a significant improvement in identifying the influential points compared to the other regression-theory influence diagnostics linear Cook's distance and nonlinear Cook's distance.

4.2. Case study set 2: Daily hydrological model with synthetic and observed streamflow and heteroscedastic residual errors

We now evaluate the performance of regression-theory influence diagnostics in a typical hydrological modelling context where the model has nonlinear response, storage and heteroscedastic errors, with both synthetic and real observed catchment data (models B_1 and B_2, respectively; see Table 1).

Observed and predicted streamflow is shown in the first row in Figure 4 for three representative time periods. For case B_1, when synthetic streamflow data is used for “observations”, the hydrological model provides a good fit to the observations for both low and high flows. This is as expected since the same hydrological and error models are used both for generating the “observations” and for model calibration. When real observed streamflow data is used in case B_2, there are more noticeable differences between observed and simulated streamflow. In particular, simulated peaks consistently under-estimate observed peaks. This indicates that the hydrological model and/or residual error model are miss-specified (i.e. there is evidence of “structural” model error).

The standardised residuals (second row in Figure 4) show large differences between the synthetic data in B_1 and the real hydrological data in B_2. For B_1, standardised residuals are independent and normally distributed with zero mean and unit standard deviation. In contrast, for B_2 the standardised residuals are auto-correlated, skewed (with much larger positive values than negative values), and have large extreme values (~4 standard deviations, c.f. ~3 for B_1). Regression-theory Cook’s distance depends on the magnitude of the standardised residuals (equation (5)), so these differences in standardised residuals may have a large impact on the influence metric.

The three leverage formulations are shown in the third row of Figure 4. Here leverage is plotted against time, rather than inputs $\mathbf{X}$ (rainfall), so that the parabolic relationship between $\mathbf{X}$ and linear leverage is not evident as it was in Figure 2. Linear leverage is high during rainfall events because this leverage formulation depends only on rainfall; at all other times it is zero, including immediately after these rainfall events – this is most clearly seen in Figure 4, Case $B_{1,}$ column 2, row 3. In contrast, nonlinear leverage and generalised leverage remain elevated for a period of time following a rainfall event. Since generalised leverage accounts for heteroscedasticity in residual errors, it is typically smaller than nonlinear leverage during high flow periods, and higher during low flow periods - this is most clearly seen in Figure 4, Case $B_{2,}$ column 2, row 3.

The magnitude of the case-deletion Cook's distance is presented in row 4 of Figure 4 as size of the grey bubbles. This influence metric is typically larger for case B_2 when observed streamflow is used, compared with when synthetic "observations" are used in B_1. This is likely due to the impact of model mis-specification for case B_2, as seen in rows 1 and 2. The accuracy of regression-theory Cook's distance compared with case-deletion Cook's distance is shown in Figure 4 (row 5). Generalised Cook's distance is the most accurate (green bubbles show the smallest absolute errors) for both cases B_1 and $B_{2.}$ For case B_1, with synthetic observations, linear Cook's distance has the highest absolute errors (orange bubbles), while for case B_2, real observations, nonlinear Cook's distance has the largest absolute errors.

Figure 5 confirms these findings when it evaluates regression-theory Cook's distance over the entire 3 years of data (~1100 points). Generalised Cook's distance provides the best performance of all three regression-theory influence diagnostics, with the smallest spread about the 1:1 line and very high performance metrics (ranging from 0.93-1.00 for all metrics). Linear Cook's distance captures neither the ranking nor the values of the case-deletion Cook's distance – as reflected by the lower metrics (e.g. r^2 values ranging from 0.01 to 0.23), with the sole exception of the Sp. having relatively high values (values of 0.93 and 0.90 for models B_1 and B_2). Nonlinear Cook's distance performs a little better than linear Cooks Distance for some metrics (e.g. $Sp._{10}$ improves from -0.30 to 0.95) for case B_1 (with synthetic observations); however, for case B_2 (with real observations) the performance is still relatively poor (e.g. $Sp._{10}$ is 0.19 and r^2 is 0.05).

These results indicate generalised Cook's distance is successfully able to capture the impact on leverage of the nonlinear and storage components of the hydrological model response as well as the heteroscedastic distribution of the model errors.

4.3. Case study set 3: Rating curve model with heteorscedastic residual errors, data uncertainty and parameter priors

The third case study set evalutes regression-theory influence diagnostics when using objective functions that account for data uncertainty and prior parameter information as part of a Bayesian inference. The magnitude of the case-deletion Cook's distance for the four rating curve cases ($C_{1\text{-}4}$) are shown in Figure 6. Each panel shows observed data (with uncertainties for cases C_2 and C_4), the fitted model and the 38 case-deletion fitted models, and the relative magnitude of case-deletion Cook's distance for each data point. We provide extrapolated axes in Figure 6 to highlight the impact of influential data on the model predictions that correspond to historical evidence of the largest floods for the Ardèche River at Sauze exceeding 6000 m^3/s [*Naulet et al.*, 2005].

In each case, the most influential data are typically extreme (both high and low) stage-discharge observed data. Accounting for discharge uncertainty in C_2 (Figure 6b) slightly reduces the magnitude of the most influential data, as seen in a slight reduction of Cooks' distance influence metric, and in a more practical sense in terms of reducing the variability in the case-deletion rating curves. Accounting for priors in C_3 (Figure 6c) leads to a larger reduction in the influential data, while the combined effect of accounting for discharge uncertainty and priors in C_4 (Figure 6d) results in an even larger reduction in the influential data, as seen by a significant reduction in case-deletion Cook's distance and a tight spread in the case-deletion rating curves. This demonstrates the value of using data uncertainty and parameter priors in reducing the impact of influential data.

Comparing the influence diagnostic results in Figure 7, the standardised residuals (second row of Figure 7) for the four rating curve models in cases $C_{1\text{-}4}$ are quite similar, hence the leverage will largely control differences in regression-theory Cook's distance between the four cases.

The third row in Figure 7 shows the different leverage formulations for cases $C_{1\text{-}4}$. For linear leverage, we see the expected parabolic shape for the leverage values as a function of $\mathbf{X}$ across the four cases

$C_{1\text{-}4}$. As $\mathbf{X}$ is not uniform the minima is off centre unlike the synthetic regression models case study sets (see Figure 2). For nonlinear leverage, since we have different objective functions between the cases, there are different calibrated model parameters, and hence different curves for the nonlinear leverage. Consistently the highest magnitude leverage is the highest stage-discharge value across the four cases, but the main difference in leverage occurs in the region of the knot where there is an increase in leverage as we go from C_1 to C_2, but a decrease in leverage for C_3 and C_4.

For generalised leverage there is an increase in leverage for low magnitude stage-discharge data and a decrease in leverage for high magnitude data relative to nonlinear leverage. This is because generalised leverage accounts for the heteroscedastic residual errors, which place higher weight on low vale of the stage-discharge data. There are also distinctive differences between the four cases $C_{1\text{-}4}$. In C_1 we have higher generalised leverage than linear and nonlinear leverage with the exception of the highest stage-discharge data point where nonlinear leverage is slightly higher. Including discharge uncertainty (C_2, column 2) and including prior information (C_3, column 3) both result in a decrease in generalised leverage across most data points except the smallest stage measurements – with prior information especially reducing the leverage on the highest stage value. Accounting for both discharge uncertainty and priors in C_4 (column 3) reduces the magnitude of the generalised leverage compared to C_1 for all but the minimum stage measurement.

Figure 8 shows the performance of the three regression-theory influence diagnostics across the four rating curve models, where we see the following patterns:

1. Linear Cook's distance generally performs poorly for all data points in terms of absolute correlation (r^2 range is 0.03-0.42, except for case C_1) but has good performance in terms of rank correlation (Sp. range is 0.90-0.94). For the top 10 most influential points the performance is lower ($Sp._{10}$ range is -0.16-0.54, r^2 range is 0.01-0.33, except for C_1). This indicates that the diagnostic has identified the ranking of the influential points moderately well, but does not identify the top 10 influential points.

2. Nonlinear Cook's distance has mixed performance with some mid to high range performance metrics (e.g. r^2 and r^2_{10} range is 0.88-0.90 for cases C_1 and C_2) but much lower performance once the priors are incorporated (e.g. r^2 and r^2_{10} range is 0.01-0.37 for cases C_3 and C_3).
3. Generalised Cook's distance has consistently high Sp. (ranging from 0.97-1.00) and performs relatively well with respect to the other metrics with lowest performance in the case of C_4 ($Sp._{10}$. of 0.66, minimum r^2 of 0.60, and minimum r_{10}^2 of 0.42).

4.4. Performance summary of regression-theory influence diagnostics

The performance metrics Sp. $Sp._{10}$, r^2 and r^2_{10}. for all ten cases ($A_{1\text{-}4}$, $B_{1\text{-}2}$, and $C_{1\text{-}4}$) in Sections 4.1 to 4.3 are summarised in Figure 9. The results for linear Cook's distance (Figure 9, top row, columns 1 and 2) show it does a reasonable job at ranking the most influential data across all data points (very high Sp. values) However, in terms of the top-ten influential points there is a significant degradation in performance ($Sp._{10}$, is lower than Sp. for all but the linear SLS model (A_1) with some negative $Sp._{10}$ for several cases meaning that the top 10 influential points are completely different to those identified by case-deletion Cook's distance. The absolute correlations (Figure 9, top row, columns 3 and 4) show that with exception of the linear SLS model**,** linear Cook's distance struggles to reproduce the magnitude of the case-deletion Cook's distance values.

Nonlinear Cook's distance (Figure 9, middle row of panels) show good performance at ranking the influential points for all data and the top 10 in synthetic cases, $A_{1\text{-}4}$ and B_1**.** However for the real data case studies (B_2 and $C_{1\text{-}4}$) there is a sharp decrease in the performance of ranking the top 10 influential points. This is maybe because in the real case studies, the impact of the heteroscedastic residual errors comes into play, which is not accounted for by nonlinear leverage.

Finally we see that generalised Cook's distance (Figure 9, bottom row of panels) produces the highest performance of the regression-theory influence diagnostics across the four performance metrics. For nine of the ten case studies, all performance metrics are above 0.75. The exception being the rating curve model with data uncertainty and priors (C_4), where generalised Cook's distance, still outperforms the linear and nonlinear Cook's distance.

4.5. Computational efficiency of influence diagnostics

An important reason for evaluating regression-theory influence diagnostics is to reduce the computational burden associated with case-deletion Cook's distance. A summary of computational demands of the different formulations is provided in Table 3, and shows that although case-deletion Cook's distance may be the most exact approach for influential point identification, it is also the most computationally intensive, requiring $n+1$ calibration runs. In contrast, all three regression theory Cook's distance are substantially more efficient, on average requiring <1% of the computational effort of case-deletion Cook's distance.

Linear Cooks Distance is the fastest because regardless of the size of the calibration data set (n) and number of model and residual error parameters (m), it requires only one model calibration followed by the application of linear matrix algebra. Nonlinear Cook's distance has the additional computational demand of calculating the finite difference approximations for the Jacobian and Hessian matrices in the leverage formulation (equation (9)). Generalised Cook's distance has the additional computational demand of calculating the finite difference approximations for the Jacobian and Hessian matrices in the leverage formulation (equation (10)). However, surprisingly, due to the number of finite difference calculations required by each formulation, generalised leverage requires fewer model runs (~140,000 in the example in Table 3) than nonlinear leverage (~270,000 runs in the example in Table 3) despite making fewer assumptions about the residual errors and therefore being broader in potential applications.

5. Discussion

5.1. Advantages and disadvantages of case-deletion and regression-theory influence diagnostics

The case-deletion and regression-theory influence diagnostics have varying assumptions and computational demands. Here we discuss the advantages and disadvantages of implementing the two classes of influence diagnostics in hydrological applications.

Case-deletion Cook's distance represents the most reliable measure of the influence as it provides a direct measure of the impact that a particular data point has on a model's predictions. Furthermore, hydrological models typically have nonlinear responses, including time-dependences in the predictions (and residuals) as a result of storage, and the residual errors are typically heteroscedastic and non-Gaussian. Therefore, case-deletion Cook's distance is attractive because it does not make any assumptions and can handle a wide range of modelling scenarios. However, the computational demand associated with re-calibrating the parameters for every data point in the observed record renders case-deletion influence analysis infeasible for anything but the simplest models with small datasets. For example, for a four parameter hydrological model with a decade of daily data, case-deletion required a run-time of 675 hours (~28 days) - see Table 3. A secondary concern with the implementation of case-deletion approaches is the repeated optimisation on complex response surfaces that are prone to multiple local optima [Duan et al., 1992; Kavetski et al., 2006].

Another drawback to applying the case-deletion Cook's distance is the loss of additional information supplied by the leverage. Cook's distance indicates which points are influential, but it does not tell us why they are influential. Analysing both the leverage and the standardised residual contribution to the magnitude of the Cook's distance therefore provides more detailed information on the nature of influential data points. Examining the standardised residuals in the case studies we see only slight variability across the four rating curve models, indicating that in some cases (such as $C_{1\text{-}4}$) the leverage contribution can be the dominant factor influencing regression-theory influence diagnostics. The additional insight from examining generalised leverage is clear from a broad range of examples from the statistical literature [e.g. *Leiva et al.*, 2014; *Lemonte and Bazán*, 2015; *Osorio*, 2016; *Rocha and Simas*, 2011]. This is evident in the hydrological model cases $B_{1\text{-}2}$ where there is a clear discrepancy between the magnitude of the standardised residual and the magnitude of Cook's distance, indicating the importance of the leverage in the influence of data points in the time series. In hydrological examples, points with high leverage can provide direction to the modeller in terms of where to focus additional data collection efforts. This is because these points will be highly influential in circumstances when high leverage is combined with high residual error.

Regression-theory influence diagnostics therefore have the following key advantages: (1) they are more efficient, due to the minimal additional computational requirements compared to a standard hydrological model calibration (99.6% fewer runs than case-deletion Cook's distance as indicated in Table 3), and (2) they provide additional diagnostic information in the form of the leverage and standardised residuals. The key limitations of regression-theory influence diagnostics are (1) they cannot evaluate case-deletion impact on predictions, parameters or objective function values (see Figure 1), and (2) they have assumptions required in the regression model structure and residual errors to formulate the leverage. In the empirical results of this study, the impact of these assumptions was illustrated with the low performance of linear and nonlinear Cook's distance on real data case studies, which had both model nonlinearity and heteroscedastic residual errors.

The development of generalised Cook's distance, which uses generalised leverage, to efficiently identify influential data points demonstrates considerable promise. For the ten case studies with a broad range of modelling scenarios (i.e. nonlinear model response, heteroscedastic residual error, data uncertainty and Bayesian inference) we saw generally high performance in terms of its ability to identify the same influential points as case-deletion Cook's distance at a fraction of the overall computational cost. This demonstrates that calculating generalised Cook's distance using generalised leverage provides a promising avenue to evaluate influential points in complex hydrological and environmental modelling scenarios. For future applications of influence diagnostics an attractive alternative to case-deletion and regression-theory influence diagnostics is to apply a hybrid framework for influence assessment [*Wright et al.*, 2018] that combines the strengths of the two existing classes; namely 1) computational efficiency, and 2) flexibility to quantify influence using hydrologically relevant metrics.

5.2. Application of generalised Cook's distance to a broader class of hydrological and environmental modelling scenarios

An important advantage of generalised Cook's distance is that the formulation of generalised leverage on which it is based can be applied to a very broad class of objective functions, as long they can be written in the general form in equation (2). Examples of suitable objective functions are: (1) those that account for autocorrelation in the residual error [see *Wei et al.*, 1998], which is common in hydrological modelling [see *Evin et al.*, 2014], and (2) alternative methods to account for heteroscedasticity such as logarithmic and Box-Cox transformations, also common in hydrological modelling [see *McInerney et al.*, 2017]. The additional challenges in applying generalised Cook's distance to environmental models outside of the model classes described herein could include: increased model structure complexity, increased computation time for model simulations, increased size of the parameter space, and potential challenges in numerically differentiating the objective function. A number of these challenges are in common with case-deletion approaches (e.g. the increased computational time), whereas others are unique to regression-based approaches (e.g., numerical differentiation issues).

Furthermore, an extension to this work would be to examine whether removing influential data in the model calibration period can improve predictions on an independent model validation time series. This would further demonstrate the impact of influential data, given the importance of model validation in hydrology [*Biondi et al.*, 2012]

5.3. Understanding the key drivers of influential data is key to reducing their impact on model calibration

Due to complex interactions between the chosen data, model and objective function, it can be difficult to identify influential data without undertaking an influence analysis post model calibration. Future work could endeavour to understand the key drivers of influential data by identifying situations where data are influential due to drivers independent of the choice of response model and objective function (e.g. rainfall and streamflow from an extreme weather event) and those situations where influential data are driven by the choice of response model (e.g. the response model poorly describes the response between $\mathbf{y}$ and $\mathbf{X}$) and/or choice of objective function (e.g. the assumed residual error model poorly describes the residual error structure). Understanding these key drivers of influential data and determining whether influential data follow a particular pattern (e.g. they tend to be the largest observed model input and/or output values, or they correspond to a specific input range, etc.) will enable the modeller to determine if additional targeted data collection (e.g. collection of more high or low flows) and/or changes to the response model and/or objective function are needed to reduce the impact of influential data. The computationally efficient regression-theory influence diagnostics developed in this study will enable future investigation towards this long term goal.

6. Conclusions

Influence diagnostics identify data points that have a disproportionate impact on model parameters, performance and/or predictions, and are therefore useful tool as part of the model calibration process. Case-deletion influence diagnostics provide an exact measure of influence; however, they have a large computational demand due to the requirement for re-calibration of the model parameters for every data point in the calibration dataset. Regression-theory influence diagnostics provide an approximation of case-deletion Cook's distance by combining two regression components for each observed data point: 1) the leverage which is used to assess the potential importance of individual observations, and 2) the standardised residuals. These are more computationally efficient than case-deletion influence diagnostics, but require making assumptions about the response model and the residual error.

We evaluate the performance of the regression-theory influence diagnostics for three different approaches 1) linear Cook's distance, which uses linear leverage, 2) nonlinear Cook's distance, which uses nonlinear leverage, and 3) generalised Cook's distance, which uses generalised leverage. This study is the first time that generalised leverage has been combined with the standardised residual to produce generalised Cook's distance in this manner. The performance in identifying the most influential data points was evaluated against case-deletion Cook's distance on a wide range of modelling scenarios (ten case studies) that included linear/nonlinear model responses, homoscedastic/heteroscedastic residual errors, and Bayesian approaches that include data uncertainty and prior information. The performance evaluation looked at correlations (rank and absolute) with the entire dataset and the top 10 influential points identified by case-deletion Cook's distance.

The key outcome of this study is that generalised Cook's distance has a high performance in approximating case-deletion Cook's distance (measured by the rank and absolute correlations) for the following modelling scenarios :

1. Nonlinear regression model with heteroscedastic residual error (Sp. 0.97, r^2 0.92),
2. Daily hydrological model including nonlinear model response and storage with heteroscedastic residual error (Sp. 0.93, r^2 0.98),
3. Rating curve model calibrated using a Bayesian framework that includes heteroscedastic residual error, data uncertainty and prior information (Sp. 0.98, r^2 0.60).

Importantly, generalised Cook's distance was able to achieve this high performance at identifying influential points at a fraction of the computational cost (<1%) of case-deletion Cook's distance.

As hydrological modelling complexity increases (i.e. more complex model structures [*Fenicia et al.*, 2011], multi-catchment datasets (e.g. >200 catchments [*Coron et al.*, 2012]), and complex objective functions [*Schoups and Vrugt*, 2010], hydrological modellers are increasingly reliant on methods to detect and diagnose the impact of modelling decisions on whether a realistic representation of the catchment response has been achieved [*Gupta et al.*, 2008]. Influential data could be significant impediment towards this goal, as their presence indicates heightened sensitivity of model outputs to a small number of data points. The development of generalised Cook's distance enables influential points to be identified without the computational demand of undertaking the numerous re-calibrations required by case-deletion Cook's distance.

7. References

Andréassian, V., F. Bourgin, L. Oudin, T. Mathevet, C. Perrin, J. Lerat, L. Coron, and L. Berthet (2014), Seeking genericity in the selection of parameter sets: Impact on hydrological model efficiency, *Water Resources Research*, *50*(10), 8356-8366.

Beven, K. (2011), *Rainfall-runoff modelling: the primer*, John Wiley & Sons.

Biondi, D., G. Freni, V. Iacobellis, G. Mascaro, and A. Montanari (2012), Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice, *Physics and Chemistry of the Earth, Parts A/B/C*, *42*, 70-76.

Cook, R. D. (1977), Detection of Influential Observation in Linear-Regression, *Technometrics*, *19*(1), 15-18.

Cook, R. D., and S. Weisberg (1982), *Residuals and influence in linear regression*, Chapman and Hall, New York.

Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, *Water Resources Research*, *48*(5).

Das, S. (2008), *Generalized linear models and beyond: An innovative approach from Bayesian perspective*, ProQuest.

Duan, Q. Y., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research*, *28*(4), 1015-1031, doi:1010.1029/1091WR02985.

Duan, Q. Y., S. Sorooshian, and V. K. Gupta (1994), Optimal Use of the Sce-Ua Global Optimization Method for Calibrating Watershed Models, *Journal of Hydrology*, *158*(3-4), 265-284.

Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resources Research*, *50*(3), 2350-2375.

Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, *Water Resources Research*, *47*(11).

Foglia, L., M. C. Hill, S. W. Mehl, and P. Burlando (2009), Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function, *Water Resources Research*, *45*.

Foglia, L., S. W. Mehl, M. C. Hill, P. Perona, and P. Burlando (2007), Testing alternative ground water models using cross-validation and other methods, *Ground Water*, *45*(5), 627-641.

Fox, J., and S. Weisberg (2011), *An R Companion to Applied Regression, Second Edition*, Sage Publications, Inc.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: elements of a diagnostic approach to model evaluation, *Hydrological Processes*, *22*(18), 3802-3813.

Hill, M. C., D. Kavetski, M. Clark, M. Ye, M. Arabi, D. Lu, L. Foglia, and S. Mehl (2015), Practical Use of Computationally Frugal Model Analysis Methods, *Groundwater*.

Hoaglin, and Welsch (1978), The Hat Matrix in Regression and ANOVA, *The American Statistician*, *32*, 17-22.

Kavetski, D., and G. Kuczera (2007), Model smoothing strategies to remove microscale discontinuities and spurious secondary optima im objective functions in hydrological calibration, *Water Resources Research*, *43*(3).

Kavetski, D., G. Kuczera, and S. W. Franks (2006), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, *Journal of Hydrology*, *320*(1-2), 173-186.

Le Coz, J., B. Renard, L. Bonnifait, F. Branger, and R. Le Boursicaud (2014), Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach, *Journal of Hydrology*, *509*, 573-587.

Le Moine, N., V. Andréassian, C. Perrin, and C. Michel (2007), How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments, *Water Resources Research*, *43*(6).

Lebecherel, L., V. Andréassian, and C. Perrin (2016), On evaluating the robustness of spatial-proximity-based regionalization methods, *Journal of Hydrology*, *539*, 196-203.

Leiva, V., E. Rojas, M. Galea, and A. Sanhueza (2014), Diagnostics in Birnbaum-Saunders accelerated life models with an application to fatigue data, *Applied Stochastic Models in Business and Industry*, *30*(2), 115-131.

Lemonte, A. J., and J. L. Bazán (2015), New class of Johnson SB distributions and its associated regression model for rates and proportions, *Biometrical Journal*.

McInerney, D., M. Thyer, D. Kavetski, J. Lerat, and G. Kuczera (2017), Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, *Water Resources Research*.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I - A discussion of principles, *Journal of Hydrology*, *10*(3), 282-290.

Naulet, R., M. Lang, T. B. M. J. Ouarda, D. Coeur, B. Bobee, A. Recking, and D. Moussay (2005), Flood frequency analysis on the Ardeche river using French documentary sources from the last two centuries, *Journal of Hydrology*, *313*(1-2), 58-78.

Nocedal, J., and S. J. Wright (2006), *Numerical Optimization*, Springer.

Osorio, F. (2016), Influence diagnostics for robust P-splines using scale mixture of normal distributions, *Annals of the Institute of Statistical Mathematics*, *68*(3), 589-619.

Oudin, L., C. Perrin, T. Mathevet, V. Andreassian, and C. Michel (2006), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *Journal of Hydrology*, *320*(1-2), 62-83.

Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, *Journal of Hydrology*, *279*(1-4), 275-289, doi:210.1016/S0022-1694(1003)00225-00227.

Petersen-Øverleir, A. (2004), Accounting for heteroscedasticity in rating curve estimates, *Journal of Hydrology*, *292*(1-4), 173-181.

Rocha, A., and A. Simas (2011), Influence diagnostics in a general class of beta regression models, *TEST*, *20*(1), 95-119.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resources Research*, *46*(10), W10531.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resources Research*, *16*(2), 430-442.

St. Laurent, R. T., and R. D. Cook (1992), Leverage and Superleverage in Nonlinear-Regression, *J Am Stat Assoc*, *87*(420), 985-990.

St. Laurent, R. T., and R. D. Cook (1993), Leverage, local influence and curvature in nonlinear regression, *Biometrika Trust*, *80*(1), 99-106

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resources Research*, *45*(12), W00B14.

Wei, B. C., Y. Q. Hu, and W. K. Fung (1998), Generalized leverage and its applications, *Scandinavian Journal of Statistics*, *25*(1), 25-37.

Wright, D., M. Thyer, and S. Westra (2015), Influential point detection diagnostics in the context of hydrological model calibration, *Journal of Hydrology*, *527*, 1161-1172.

Wright, D., M. Thyer, S. Westra, and D. McInerney (2018), A hybrid framework for quantifying the influence of data in hydrological model calibration, *Journal of Hydrology*.

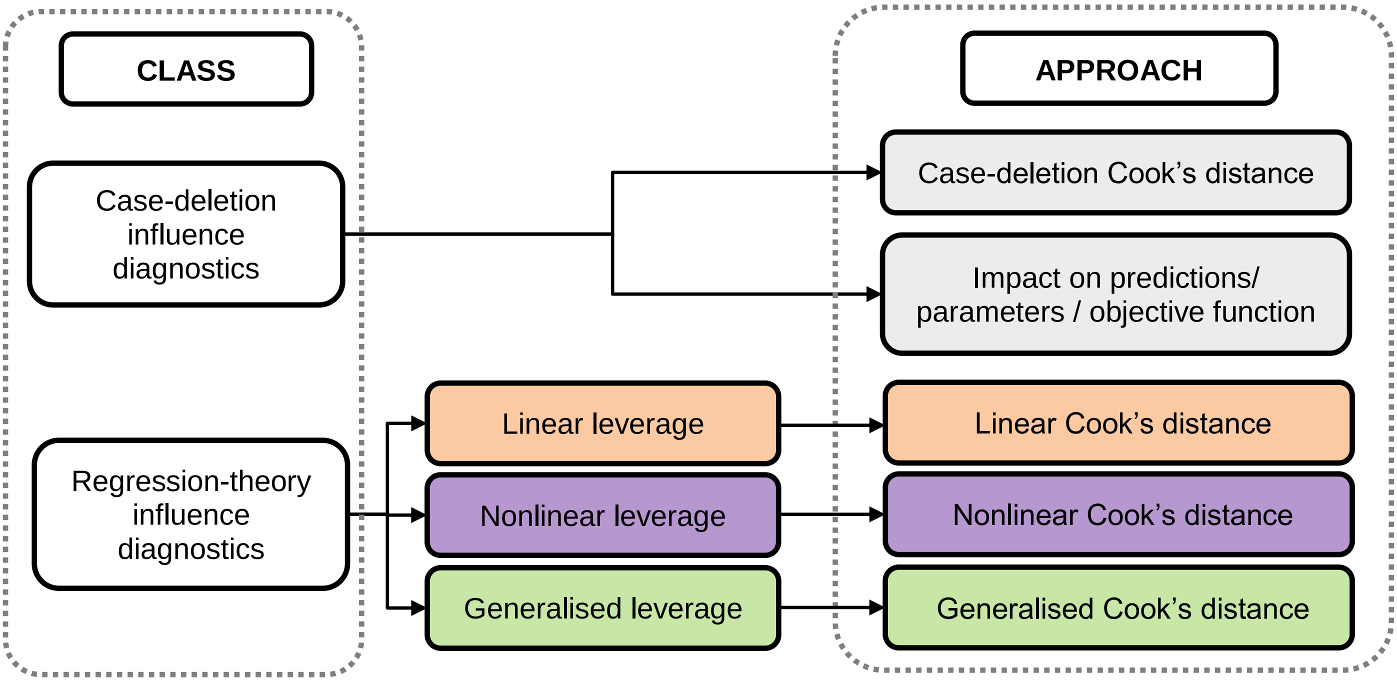

Figure 1 – Range of available influence diagnostics in the literature. Influence diagnostics are broken up into two classes on the left hand side with the various approaches on the right hand side. The three regression-theory approaches are colour coded based on the leverage formulation that they use and as they appear in the latter figures with linear Cook's distance (orange), nonlinear Cook's distance (purple), and generalised Cook's distance (green).

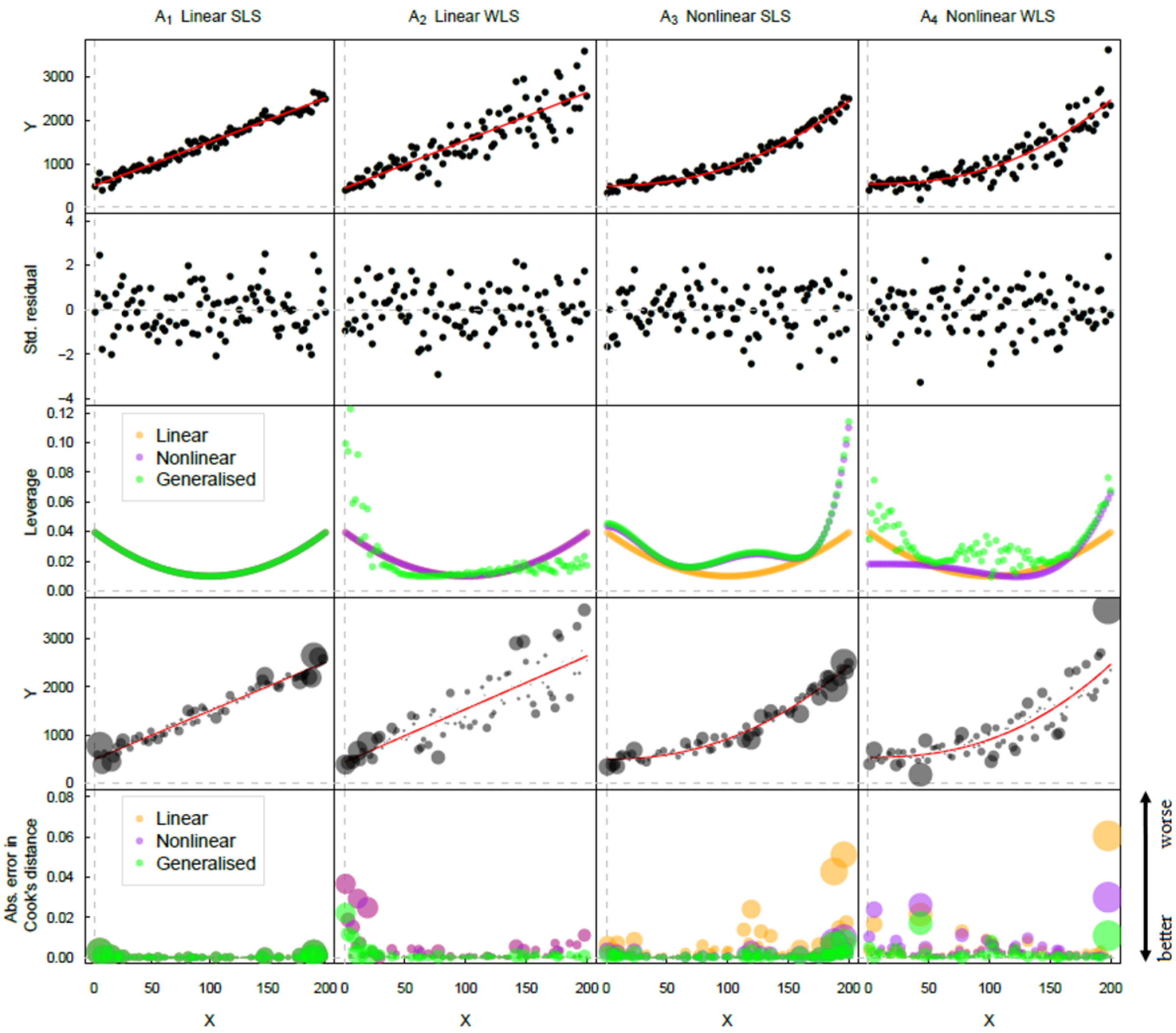

Figure 2 – Results for case study set 1: Synthetic regression models. "Observed" data (black), and model predictions (red) in the top row, followed by standardised residuals in the second row. Leverage is shown in the third row with linear leverage, nonlinear leverage and generalised leverage. In the case of A_1 the three leverage formulations are exactly equal and so are superimposed over each other, as is the case in A_2 with linear and nonlinear leverage. The fourth row highlights the distribution of the most influential data in the context of the observed data (black) and model predictions (red) where the size of the bubbles is scaled to the value of case-deletion Cook's distance giving a relative indication of influence. For actual case-deletion Cook's distance values refer to Figure 3. The final row shows the absolute error between regression-theory Cook's distance and case-deletion Cook's distance where the size of the bubbles is scaled to the value of case-deletion Cook's distance to highlight the absolute error for the most influential data points. Note that in the final row the relative errors are superimposed over each other.

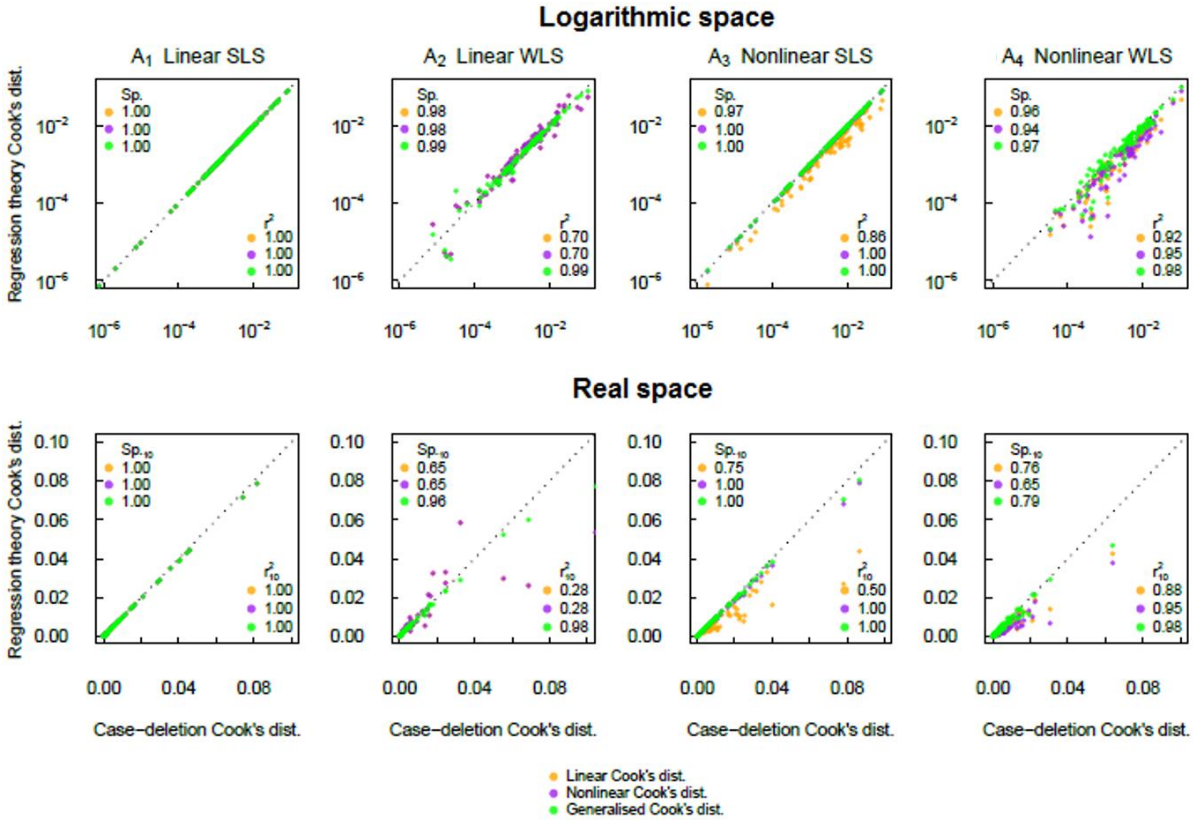

Figure 3 –Comparison of case-deletion Cook's distance and regression-theory influence diagnostics for case study set 1: Synthetic regression models. In the first row we compare the performance in logarithmic space and use the Spearman rank correlation (Sp.) and Pearson correlation (r^2) to highlight performance across the whole dataset. In the second row we compare the performance in real space and use the $Sp._{10}$ and r_{10}^2 to compare the subset of the ten most influential data points.

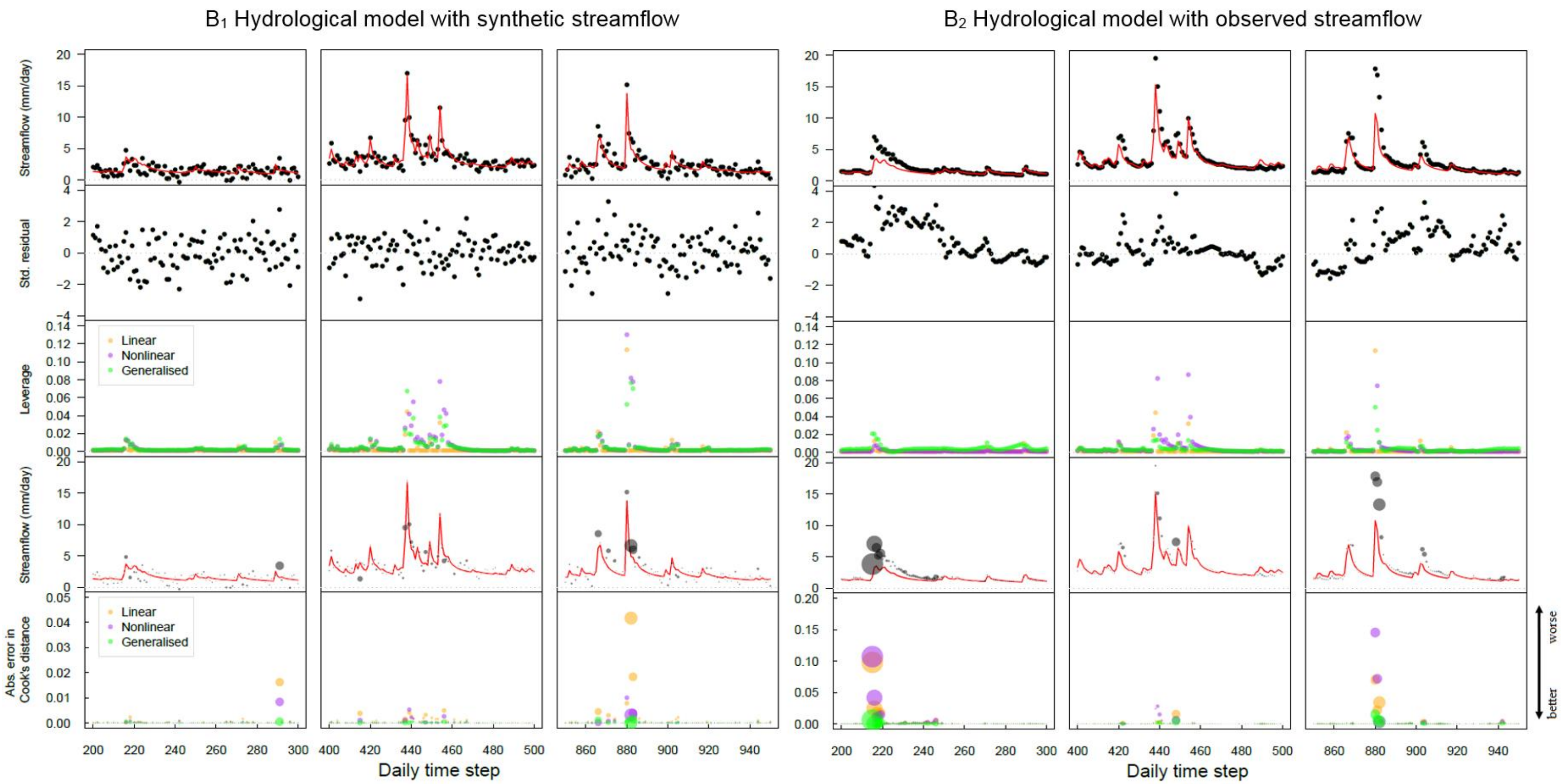

Figure 4 – Results from case study set 2: Daily hydrological modelling case studies B_1 and B_2, presented in an analogous manner to Figure 2. Observed streamflow (black), and predicted streamflow (red) are shown in the top row for three different representative 100 day time periods, followed by standardised residuals in the second row. Leverage is shown in the third row. The fourth row highlights the distribution of the most influential data, where the size of the bubbles is scaled to the value of case-deletion Cook's distance. The final row shows the absolute error between regression-theory Cook's distance and case-deletion Cook's distance. Note that in the final row the relative errors are superimposed over each other.

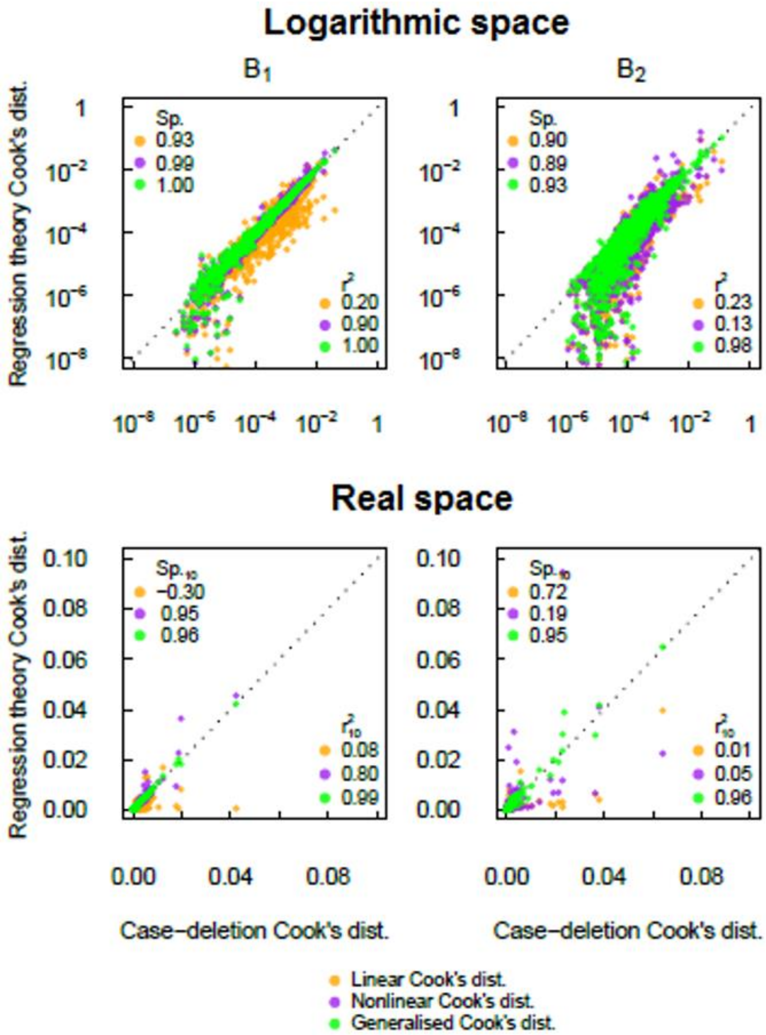

Figure 5 –Comparison of case-deletion and regression-theory influence diagnostics for case study set 2: Daily hydrological modelling cases B_1 and B_2, presented in the same manner as Figure 3

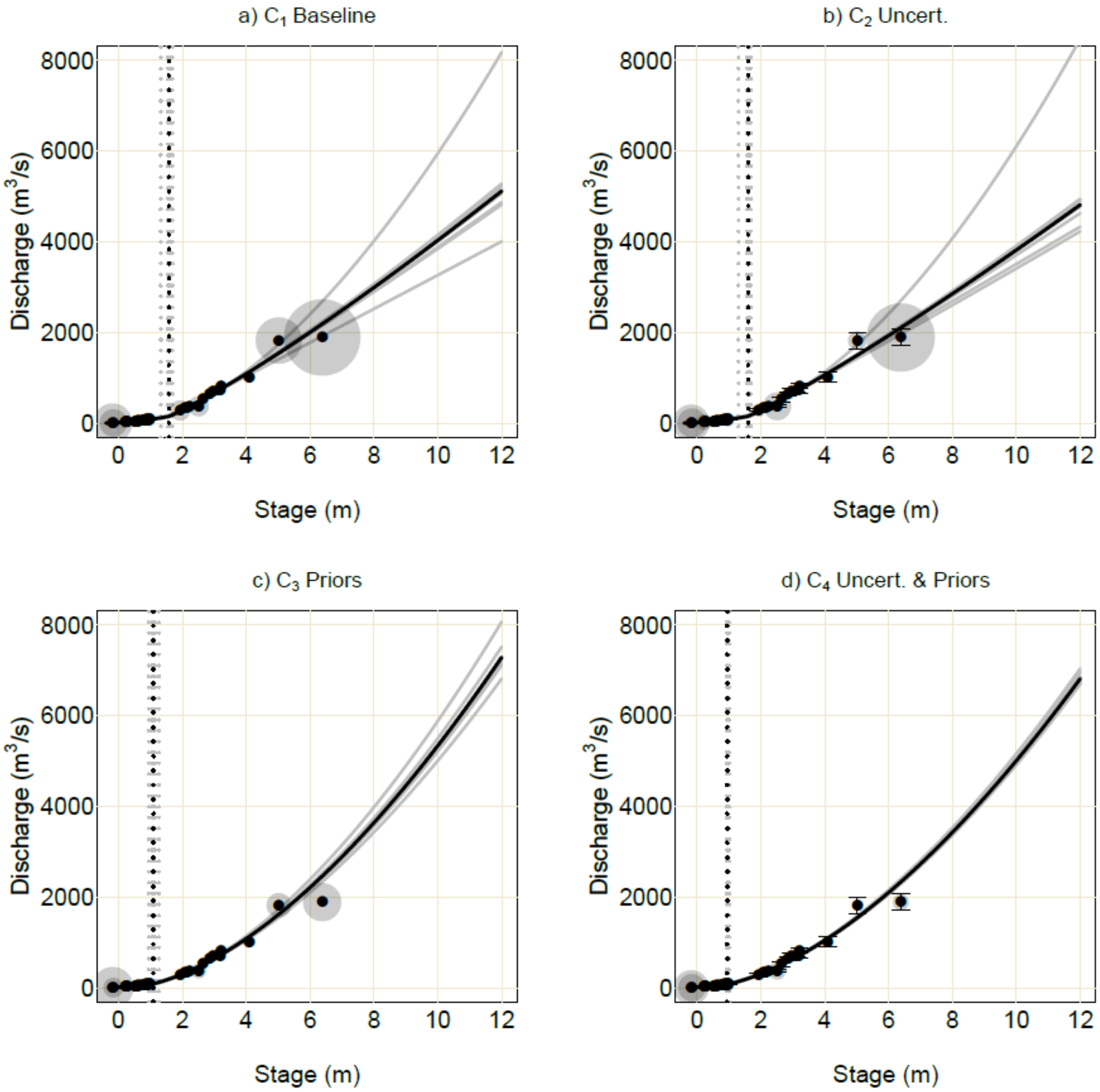

Figure 6 – Stage-discharge rating curves for the Ardèche River at Sauze. The four rating-curves presented are a) baseline rating curve without accounting for discharge uncertainty and priors, b) Rating curve with discharge uncertainty, c) Rating curve with parameter priors, d) Rating curve with both discharge uncertainty and parameter priors. Corresponding computed transition levels between section and channel controls is marked with vertical broken lines. The 38 case-deletion rating-curves and computed transition levels are shown in grey. The size of the points correspond to the relative magnitude of the case-deletion Cook's distance.

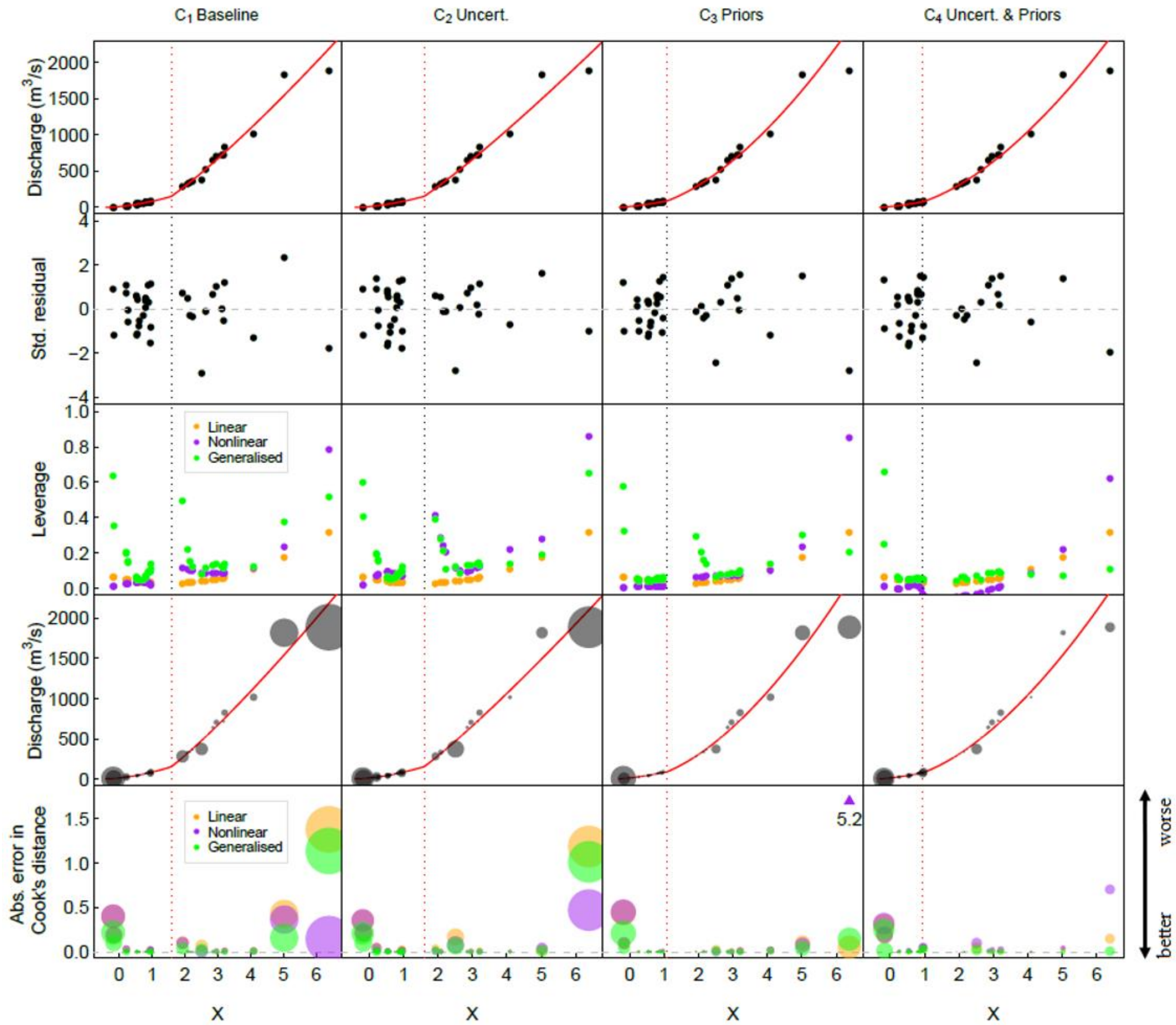

Figure 7 – Results for case study set 3: Rating curve models. The computed transition level (knot) between section and channel controls is marked with a vertical dashed line. Observed data (black), and model predictions (red) in the top row, followed by standardised residuals in the second row. Leverage is shown in the third row. The fourth row highlights the distribution of the most influential data, where the size of the bubbles is scaled to the value of case-deletion Cook's distance. The final row shows the absolute error between regression-theory Cook's distance and case-deletion Cook's distance. Note that in the final row the relative errors are superimposed over each other.

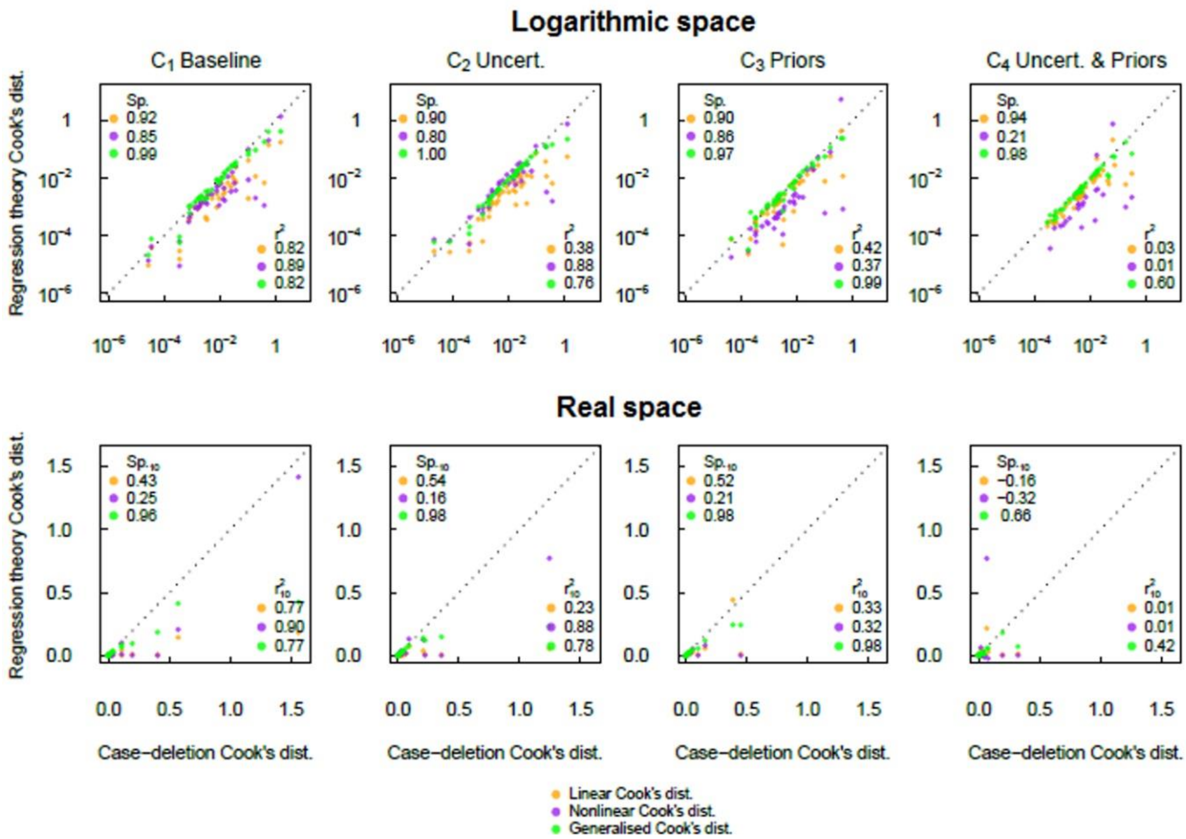

Figure 8 Comparison of case-deletion and regression-theory influence diagnostics for case study set 3: Rating curve models, presented in the same manner as Figure 3.

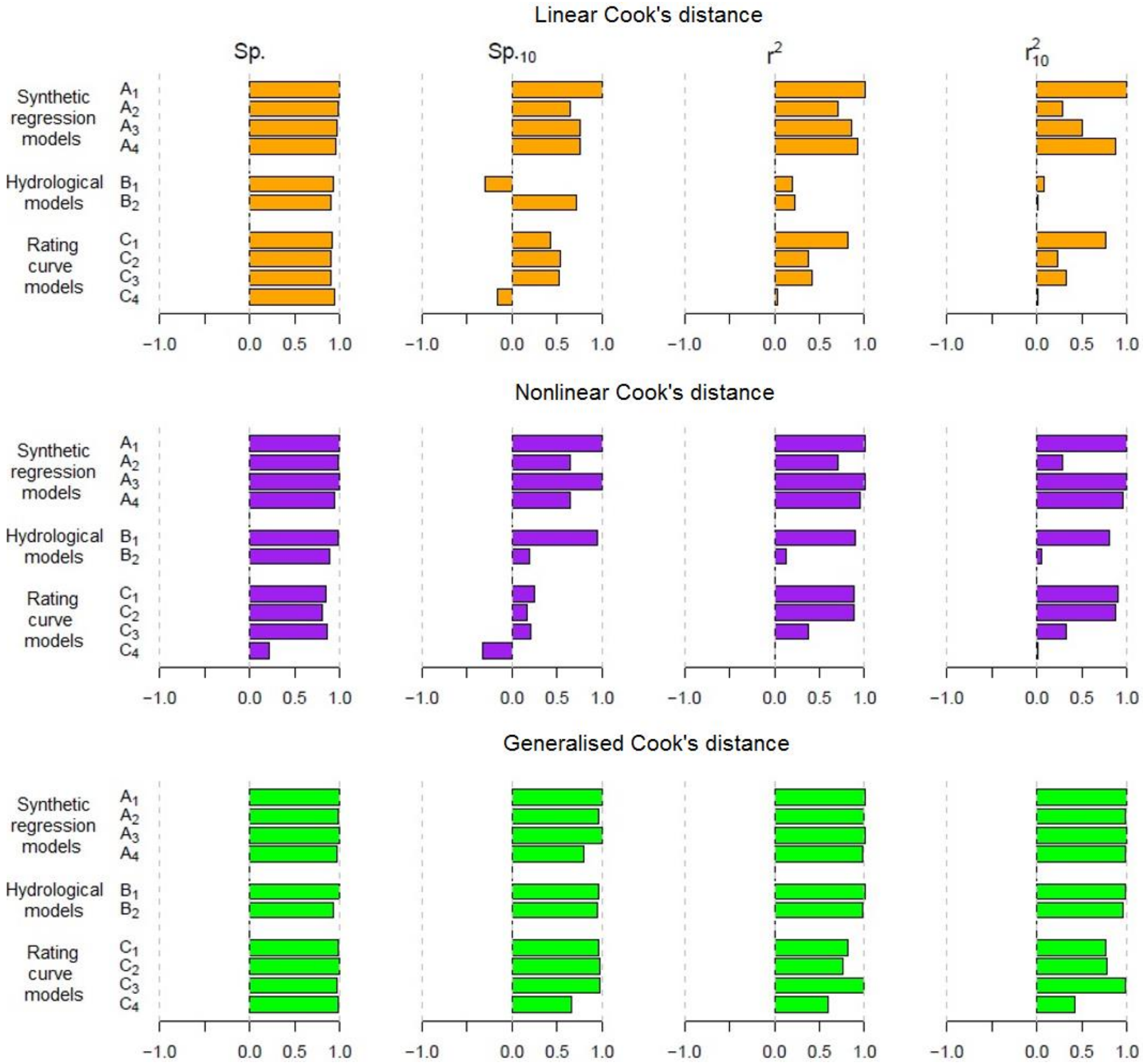

Figure 9 – Performance metrics for regression-theory influence diagnostics across the ten case studies in the three case study sets. We apply the Spearman rank correlation and Pearson correlation to: (1) the whole set of data points (Sp. and r^2, respectfully), and (2) the top 10 most influential data points identified by case-deletion Cook's distance ($Sp._{10}$ and r^2_{10}, respectfully). Linear Cook's distance is shown in the first row (orange), nonlinear Cook's distance in the second row (purple) and finally generalised Cook's distance in the bottom row (green).

Table 1 – Details of the case studies.

Case study	Response model	Residual error model	"Observed" output Y	Objective function
Case study set 1: Synthetic regression models, **Input:** $X \sim \mathrm{U}(0,200)$				
A_1: Linear regression, homoscedastic residuals	$f(\mathbf{X},\alpha_1,\alpha_2)=\alpha_1\mathbf{X}+\alpha_2$	$\varepsilon(\sigma)\sim \mathrm{N}(0,\sigma^2),\sigma=\beta_1$	$f(\mathbf{X},10,500)+\varepsilon(100)$	2.2.1
A_2: Linear regression, heteroscedastic residuals	$f(\mathbf{X},\alpha_1,\alpha_2)=\alpha_1\mathbf{X}+\alpha_2$	$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\beta_1\mathbf{y}+\beta_2$	$f(\mathbf{X},10,500)+\varepsilon(0.2,10)$	2.2.2
A_3: Nonlinear regression, homoscedastic residuals	$f(\mathbf{X},\alpha_1,\alpha_2,\alpha_3)=\alpha_1+\alpha_2\mathbf{X}^{\alpha_3}$	$\varepsilon(\sigma)\sim \mathrm{N}(0,\sigma^2),\sigma=\beta_1$	$f(\mathbf{X},500,0.1,2.3)+\varepsilon(100)$	2.2.1
A_4: Nonlinear regression, heteroscedastic residuals	$f(\mathbf{X},\alpha_1,\alpha_2,\alpha_3)=\alpha_1+\alpha_2\mathbf{X}^{\alpha_3}$	$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\beta_1\mathbf{y}+\beta_2$	$f(\mathbf{X},500,0.1,2.3)+\varepsilon(0.1,0.5)$	2.2.2
Case study set 2: Daily Hydrological models, **Input:** Observed rainfall measurements, All models have heteroscedastic residuals				
B_1: GR4J, synthetic output	$\mathrm{GR4J}(\mathrm{P},\mathrm{PET},\boldsymbol{\alpha})$	$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\beta_1\mathbf{y}+\beta_2$	$\mathrm{GR4J}(\mathrm{P},\mathrm{PET},\boldsymbol{\alpha}=\{2200,1.15,87,0.55\})+$ $\varepsilon(0.1,0.5)$	2.2.2
B_2: GR4J, observed output	$\mathrm{GR4J}(\mathrm{P},\mathrm{PET},\boldsymbol{\alpha})$	$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\beta_1\mathbf{y}+\beta_2$	Observed	2.2.2
Case study set 3: Rating curve models, Input: Observed stage measurements, All models have heteroscedastic residuals				
C_1: Rating curve model,	$f(X_i,\boldsymbol{\alpha})=\begin{cases}\alpha_1(X_i-\alpha_2)^{\alpha_3}, X_i<\alpha_4\\ \alpha_5(X_i-b_2)^{\alpha_6}, X_i\ge\alpha_4\end{cases}$	$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\beta_1\mathbf{y}+\beta_2$	Observed	2.2.2
C_2: Rating curve model, data uncertainty		$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\sqrt{\boldsymbol{\sigma}_r^{\,2}+\boldsymbol{\sigma}_Y^{\,2}},\boldsymbol{\sigma}_r=\beta_1\mathbf{y}+\beta_2$		2.2.3
C_3: Rating curve model, parameter priors		$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\beta_1\mathbf{y}+\beta_2$		2.2.4
C_4: Rating curve model, data uncertainty, parameter priors		$\varepsilon(\boldsymbol{\sigma})\sim \mathrm{N}(0,\boldsymbol{\sigma}^2),\boldsymbol{\sigma}=\sqrt{\boldsymbol{\sigma}_r^{\,2}+\boldsymbol{\sigma}_Y^{\,2}},\boldsymbol{\sigma}_r=\beta_1\mathbf{y}+\beta_2$		2.2.5

Table 2 – Selected prior mean (standard deviation) for the two-part rating curve model taken from Le Coz [2014]. An uninformative uniform distribution was used for the residual error model parameters. Control 1 is the rectangular sill at low flows, and Control 2 is to the rectangular channel at high flows.

Control 1					Control 2	
α	a_1	b_1	c_1	k_1	a_2	c_2
	50 (100)	-0.5 (2)	1.5 (0.025)	1 (1)	100(200)	1.67 (0.025)

Table 3 – Summary of the computational demand of case-deletion and regression-theory Cook's distance. The example case study corresponds to the daily hydrological model (i.e. $m_\alpha = 4$, $m = 6$) with ~10 years of data (i.e. $n = 3650$) where a fixed number of model runs is assumed per calibration ($r = 10000$ model runs). The example runtime is calculated with a 2.90GHz processor.

Approach	Leverage	General computation demand	Model runs	Example computational demand	Example runtime (hours)	Reduction from case-deletion
Case-deletion Cook's distance	-	n+1 model re-calibration	$r \times (n+1)$	36,510,000 runs	675.37	-
Linear Cook's distance	Linear	Single calibration	r	10,000 runs	0.18	99.97%
Nonlinear Cook's distance	Nonlinear	Single calibration + central difference calculations	$r + 2(n \times m_\alpha) + 4(n \times m_\alpha \times m_\alpha)$	272,800 runs	5.05	99.25%
Generalised Cook's distance	Generalised	Single calibration + central difference calculations	$r + 2(n \times m) + 4(m \times m) + 4(n \times m)$	141,544 runs	2.62	99.61%