THE UNIVERSITY of ADELAIDE

Influence diagnostics in hydrological modeling

David Peter Wright

B.Eng Civil & Environmental Engineering (Honours)

Thesis submitted in fulfilment of the requirements for the degree of Doctorate of Philosophy

The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Civil, Environmental and Mining Engineering

March 2017

Contents

List of Figures

List of Tables

Abstract

Accurate hydrological model predictions play an important role in designing infrastructure for domestic water supply, agriculture, industry, and flood and drought protection. A key step in model development is model calibration where hydrologists fit a model to historical data to make predictions into the future. The original contribution to knowledge of this thesis is to evaluate and develop influence diagnostics 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.

Influence diagnostics can be implemented to describe changes in model predictions, calibrated parameters and model performance. Broadly, these diagnostics can be categorised into two different classes; "case-deletion" influence diagnostics and "regression-theory" influence diagnostics. Although influence diagnostics have previously been applied to a small number of hydrological studies, there is a need to address the following two major limitations with the currently available influence diagnostics before they can be applied to broader hydrological applications:

1. Case-deletion influence diagnostics are too computationally expensive to apply in hydrological modelling applications because of the length of data and therefore number of model recalibrations that are required (e.g. 10 years requires approximately 3650 model re-calibrations).

2. Regression theory influence diagnostics are computationally efficient, but only linear Cook's distance has been applied which has strong assumptions of linear model response and Gaussian residual error that are typrically not valid in hydrological modeling.

This thesis by publication presents three papers in Chapter 2 to Chapter 4. The first paper investigates the application of influence diagnostics in the context of a series of common hydrological case-studies including a rating curve model and a daily hydrological model with two years of calibration data. In the second paper we generalise regression

theory influence diagnostics and evaluate the performance in reproducing the computationally expensive case-deletion influence diagnostics on eleven case studies with a variety of model structures and inference scenarios including: nonlinear model response, heteroscedastic residual errors, data uncertainty and Bayesian priors. Finally, in the third paper we present a hybrid framework for influence assessment that combines the strengths of the two classes of influence diagnostics in order to overcome the key limitations listed above.

The hybrid framework presented in the third paper in this thesis will provide a foundation for all hydrological modellers to have greater insight into the influence of individual data points on model calibration, thereby providing a basis for identifying disinformative points or understanding how sensitive model predictions are to a small proportion of the dataset.

Statement of originality

I, David Peter Wright, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

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

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

30/03/2017

David Peter Wright Date

Acknowledgements

I feel honoured and privileged to have been able to undertake this research. My most sincere gratitude goes to my supervisory team Associate Professor Mark Thyer and Associate Professor Seth Westra who believed in my abilities, provided me with great support and effort in helping me to achieve my goals, patience, and guidance throughout my candidature. Your combined skill, expertise and passion for research have been invaluable to me.

I would also like to thank wholeheartedly the following people.

Dr David McInerney for his support in mathematics and statistics during my doctorate that helped me to see a light at the end of the tunnel.

Dr Benjamin Renard for his support, guidance, and hospitality during my stay in Lyon, France. I extend this gratitude to the staff at IRSTEA especially Dr Valentin Mansanarez and François Tilmant for providing me with accommodation and aiding in my exploration of French gastronomy.

My fellow postgraduate students and academics at the University of Adelaide and in the international hydrological community. Our discussions over the years provided a great deal of insight and support during my candidature.

My parents who gave me enormous opportunity and amongst many other things took me to the library, showed me the beauty of nature, and the importance of kindness.

Ben, Geordan, Samane, Anthony, and other friends and family. Thank you for your friendship and support, for keeping me fascinated by the world, and for reminding me of the simple things that make life beautiful.

David Peter Wright
University of Adelaide
March 2017

Chapter 1

Introduction

Accurate hydrological model predictions play an important role in designing infrastructure for domestic water supply, agriculture, industry, and flood and drought protection. A key step in model development is model calibration where a hydrological model is "fitted" to historical data. This hydrological model is then used to make predictions about the future impact of management options.

Hydrological model calibration is required as parameters generally cannot be inferred directly from catchment measurements but are instead obtained by minimising the differences between observed and simulated streamflow [Beven, 2011]. Studies have increasingly called for the use of influence diagnostics as part of model calibration [e.g. Foglia et al., 2009; Foglia et al., 2007; Hill et al., 2015] to understand the extent to which model calibration outcomes (i.e. model fit, parameter estimates and predictions) are largely determined by a very small number of data points. Generally, influential data points arise in cases when the data may be erroneous and/or the hydrological model miss-represents the catchment behaviour.

It is important to identify data points that have a large influence on hydrological predictions when these data points are erroneous (i.e. have large data errors), as this is likely to lead to sub-optimal model performance when applied to an independent dataset. The importance of identifying such "disinformative" data has been highlighted by Beven and Westerberg [2011], who discuss the need for more formal methods to identify and remove erroneous data prior to model calibration. A key challenge in identifying disinformative data is that examining long hydrological data sets (e.g. 10-20 years of daily data) can be labour intensive. If influence diagnostics can be applied in hydrological applications this would enable a more efficient and focused analysis on a smaller subset of the most influential data to determine if they are erroneous/disinformative. Influence analysis would therefore provide a more efficient approach to identify erroneous/disinformative

data points.

Data points that have a large influence on hydrological predictions but are not erroneous (i.e. they are not erroneous and hence are not 'disinformative') are also important to identify. These influential data may occur in circumstances when the chosen hydrological model does not adequately describe the response between model inputs and outputs, and/or the chosen objective function poorly describes the residual error between observed and simulated streamflow. In these circumstances influence diagnostics would also have potential to be a powerful diagnostic tool that enables focussed analysis on the events where the hydrological model does not capture observed catchment behaviour. Applying influence diagnostics in this manner would make it easier to identify avenues for the development of hydrological model enhancements.

Understanding and quantifying the impact of influential data in hydrological modelling is the key motivation of this thesis. If influence diagnostics can be successfully applied to hydrological examples then they show great potential to augment existing strategies in the literature to identify disinformative data [Beven and Westerberg, 2011], reduce uncertainty in hydrological calibration [Wagener et al., 2003], and couple with existing model analysis tools [Hill et al., 2015]. The original contribution to knowledge of this thesis is to evaluate and develop influence diagnostics in the context of hydrological model calibration and provide a hybrid framework for the future application of influence diagnostics in hydrology.

Before reviewing previous applications of influence diagnostics in hydrological modelling we first review their development and application in the statistics literature and define the various classes of influence diagnostics that are available for application to hydrological modelling.

1.1 Literature review of influence diagnostics in the statistical literature

Influential diagnostics are widely used in the statistical literature for both the detection of erroneous points and for identifying possible model deficiencies in wide range of modelling applications. These include linear regression [Cook, 1977], generalised linear regression [Thomas and Cook, 1989], generalised additive models [Hastie and Tibshirani, 1990] and various other regression-based approaches [Chen et al., 2012; Russo et al., 2009]. In these applications influence diagnostics are applied to identify data that have a disproportionate impact on model calibration so that the modeller may review their inclu-

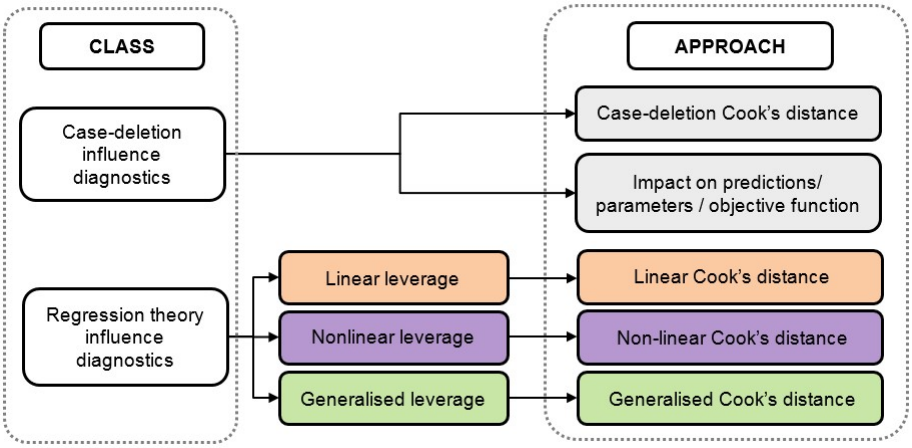

Figure 1.1: Range of available influence diagnostics in the literature. Influence diagnostics are broken up into two classes on the left hand side with the various approaches on the right hand side.

sion or removal in the regression. Influence diagnostics can be implemented to describe changes in model predictions, calibrated parameters and model performance. Broadly, these diagnostics can be categorised into two different classes; "case-deletion" influence diagnostics and "regression-theory" influence diagnostics as illustrated in Figure 1.

Case-deletion influence diagnostics were first developed by Cook [1977] and involve removing ("deleting") a data point ("case") from the set of calibration points, and then recalibrating the model. Parameter estimates and model predictions from the recalibration are compared to the results from the full calibration, and this is repeated for all data points in the calibration set. Once case-deletion has been performed, two different approaches can be used to measure influence. The first approach is to evaluate Cook's distance [Cook, 1977] 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]; however, for the purposes of this thesis we refer to it as "case-deletion" Cook's distance. As an alternative to applying Cook's distance, the modeller may use the case-deletion results to quantify the difference between the original and re-calibrated model predictions, parameters, and performance and/or any metric of their choice. Two additional influence diagnostics that we will not consider in this thesis are DFFITS and DFBETA [see Cook and Weisberg, 1982]. DFFITS is not included in this work as Cook and Weisberg [1982] show that Cook's distance and DFFITS are conceptually identical and provide a closed-form formula to convert one value to the other. DFBETA has not been included in this work as it measures the influence with respect to each individual model parameter separately and therefore does not give a

good measure of the overall influence of data on model calibration.

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 to hydrological applications, where the hydrological models are non-linear and the assumptions of Gaussian independent and identically distributed residuals errors 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, any efforts will compound the computational demands of the case-deletion re-calibrations.

The second class of influence diagnostics are the "regression-theory" influence diagnostics (see Figure 1). They are a more efficient alternative to the exact case-deletion influence diagnostics because they approximate Cook's distance [Cook, 1977] using regression modelling theory to combine the following two components for each observed data point: (1) the leverage, which provides a measure of how far away the independent variable value of an observation is from those of the other observations [Wei et al., 1998] (see Section 3.2.4 for further information), and (2) the standardised residuals, which are based on the raw residuals or the difference between the observed output and the model predicted output (see Section 3.2.3 for further information). By combining these two components to approximate Cook's distance, "regression-theory" influence diagnostics require no additional re-calibrations and are therefore an attractive, more efficient alternative to the computationally demanding case-deletion influence diagnostics. The drawback to this approach is that calculating the leverage can require making assumptions regarding the type of regression model (linear or non-linear) and/or making assumptions about the probabilistic model for residual errors.

Combining linear leverage with the standardised residuals is the most widely used approach to approximate Cook's distance in regression problems [Fox and Weisberg, 2011].

Linear leverage assumes that the regression model is linear and that residual errors are Gaussian, homoscedastic and independent. Hence, the approach of calculating Cook's distance using linear leverage (hereafter referred to as "linear Cook's distance") may not be suitable for identifying the influential points in a hydrological modelling context. This is because the hydrological model calibration violates the assumptions of linear leverage, 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. In a nonlinear regression model predictions are not linear functions of observed responses and therefore calculation of leverage requires first and second order approximations of the nonlinear model response. Calculating Cook's distance using nonlinear leverage (hereafter referred to as "nonlinear Cook's distance) can take into account nonlinear model response, however it is still limited by the assumption that residual errors are Gaussian, homoscedastic and independent.

To overcome the limitations of the assumptions of linear and non-linear 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 a broad range of 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 more broadly environmental modelling. 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 identify influential data points.

1.2 Literature review and current limitations of influence diagnostics in hydrological modelling

The prospect that a small number of data points can exert a high influence on model predictions in hydrological modelling motivates the more widespread implementation of influence diagnostics; however applications have been limited. In the context of ground-

water modelling, Yager [2004] found that models were highly sensitive to small changes in influential data. Foglia et al. [2007] used a series of case-deletion metrics and linear Cook's distance on a groundwater model and found similar performance between the two classes. Foglia et al. [2009] applied linear Cook's distance as part of a suite of diagnostics to a short time series of 37 daily observations in the rainfall-runoff model TOPKAPI and found that some of the low flow observations during small precipitation events were more important than anticipated. Legates and McCabe [1999] discuss the oversensitivity to outliers of correlation based goodness-of-fit measures used in hydrological models and recommend that additional evaluation measures should supplement calibration. Berthet et al. [2010] found a quadratic criterion to be influenced by a very small number of time steps characterised with high runoff variation. Perrin et al. [2007] assess the impact of the quantity and quality of streamflow data on parameter calibration and model robustness and show that a subset of influential points from a larger dataset is sufficient to obtain robust estimates of model parameters. Singh and Bárdossy [2012] pre-process hydrological data using depth functions to identify unusual events and investigate the calibration of the model with only this set of critical data to assess if the subset has enough information to identify model parameters. A recent example in the context of flood frequency analysis is the application of case-deletion to show that low flow outliers can have a disproportionate influence on extreme flood quantile estimates [Lamontagne et al., 2013]. Their technique was based on a generalised Grubbs-Beck test statistic developed by Cohn et al. [2013] that is designed to identify potentially influential low flows. Also recently Hill et al. [2015] highlight the potential value of influence assessment by including Cook's distance in a suite of computationally frugal model analysis tools.

Although influence diagnostics have previously been applied to a small number of hydrological studies, there is a need to address the following two major issues with the currently available influence diagnostics before they can be applied to broader hydrological applications:

1. Case-deletion influence diagnostics are too computationally expensive to apply in hydrological modelling applications because of the length of data and therefore number of model recalibrations that are required (e.g. 10 years requires approximately 3650 model re-calibrations).

2. Regression theory influence diagnostics are computationally efficient, but only linear Cook's distance has been applied which has strong assumptions of linear model response and Gaussian residual error that are typically not valid in hydrological modeling.

These research gaps have prevented further application of influence diagnostics to date, however if the issues can be resolved then there is an opportunity for the development of a framework for influence assessment across hydrological modelling applications.

1.3 Overall research objectives

The overall aim of this research is to explore the application of influence diagnostics in hydrological modelling in a stepwise manner and work towards a general framework for influence assessment in hydrological model calibration. Three specific research objectives have been identified, each of which has a number of sub-objectives:

Objective 1 – Explore the application of influence diagnostics in the context of a series of common hydrological modelling case studies (Paper 1): This objective serves as a starting point to investigate the application of the existing influence diagnostics to a small number of hydrological case studies.

Objective 1.1: Quantify the influence of individual data points on calibrated model predictions and parameters.

Objective 1.2: Evaluate the performance of regression based Cook's distance with linear leverage and nonlinear leverage.

Objective 1.3: Explore the relationship between magnitude of streamflow and influence.

Objective 1.4: Evaluate the computational demand of case-deletion and regression based influence diagnostics.

Objective 2 – Generalise regression theory influence diagnostics to be suitable for a wide range of hydrological modelling scenarios (Paper 2): This objective broadens the regression theory based influence diagnostics to be applicable to case-studies with non-Gaussian residual error.

Objective 2.1: Evaluate the performance of regression based Cook's distance to account for nonlinear model response and heteroscedastic residual error in a series of simple regression models.

Objective 2.2: Evaluate the performance of regression based Cook's distance to account for including nonlinear model response and heteroscedastic residual error.

Objective 2.3: Evaluate the performance of regression based Cook's distance to account for objective functions that include data uncertainty and prior information.

Objective 3 – Develop a hybrid framework for influence assessment in hydrological modelling (Paper 3): This objective provides direction for computationally cheap and accurate application of influence diagnostics to general hydrological examples.

Objective 3.1: Develop a robust and computationally efficient hybrid framework that can be adopted by hydrologists.
Objective 3.2: Determine the number of influential points that have the greatest effect on mean, high and low flows.
Objective 3.3: Understand how the length of calibration data can determine the influence of individual data points on hydrologically relevant metrics.
Objective 3.4: Explore how the choice of objective function can impact the influence of individual data points on hydrologically relevant prediction metrics.

1.4 Thesis overview

The thesis is organised into five chapters where the main contributions are presented in **Chapter 2** to **Chapter 4**. Each of these chapters is presented in the form of a technical paper. The first of these has been published in the *Journal of Hydrology*, the second has been submitted to *Environmental Modelling and Software* for peer review and the third has been submitted to the *Journal of Hydrology* for peer review.

In **Chapter 2** we introduce case-deletion and regression theory based influence diagnostics and apply them to hydrological case studies (Objective 1). For the regression theory methods, both linear and nonlinear estimates of leverage are used to calculate Cook's distance.

In **Chapter 3** we combine the generalised leverage with the standardised residual to produce generalised Cook's distance (Objective 2). We evaluate the performance of linear Cook's distance, nonlinear Cook's distance and generalised Cook's distance on eleven case studies with a variety of model structures and inference scenarios including: nonlinear model response, heteroscedastic residual errors, and objective functions incorprating data uncertainty and Bayesian priors.

The key issues with the current suite of influence diagnostics is that the computationally efficient approaches do not provide hydrologically relevant influence metrics, while the hydrologically relevant influence metrics are computationally expensive to calculate. In **Chapter 4** we introduce a new two-stage hybrid framework that overcomes these challenges, by delivering hydrologically relevant influence metrics in a computationally efficient manner (Objective 3).

Although the manuscripts have been formatted in accordance with the University guidelines, the manuscript material is otherwise unchanged. A copy of Paper 1 is reproduced in Appendix A as published. Paper 2 and Paper 3 have been submitted to peer review.

Conclusions are provided in **Chapter 5**, which summarises the research contributions of **Chapter 2** to **Chapter 4** and includes a discussion of the limitations and recommendations for future work.

Chapter 2

Influential point detection diagnostics in the context of hydrological model calibration (Paper 1)

David P Wright, Mark Thyer, Seth Westra
Journal of Hydrology, 527 (2015) 1161–1172

Statement of Authorship

Title of Paper	Influential point detection diagnostics in the context of hydrological model calibration
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Wright, D.P., Thyer, M., Westra, S, 2015 Influential point detection diagnostics in the context of hydrological model calibration. Journal of Hydrology, 527 (2015) 1161–1172

Principal Author

Name of Principal Author (Candidate)	David Peter Wright		
Contribution to the Paper	Development and implementation of approach, visualisation and interpretation of results, preparation of manuscript and acted as corresponding author.		
Overall percentage (%)	85		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	30/03/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mark Thyer		
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript		
Signature		Date	30/3/17

Name of Co-Author	Seth Westra		
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript		
Signature		Date	30/3/2017

Abstract: Influential data are those that have a disproportionate impact on model performance, parameters and/or predictions. This paper evaluates two classes of diagnostics that identify influential data for hydrological model calibration: (1) numerical "case-deletion" diagnostics, which directly measure the influence of each data point on the calibrated model; and (2) analytical diagnostics based on Cook's distance, which combine information on the model residuals with a measure of the distance of each input point from the centre of the range of the input data (i.e., the leverage). Case-deletion methods rank influence by changes in the model parameters (measured through the Mahalanobis distance), performance (using objective function displacement) and predictions (e.g. mean and maximum streamflow). For the analytical methods, both linear and nonlinear estimates of leverage are used to calculate Cook's distance, which is used to rank influential data. We apply these diagnostics to three case studies and show that a single point could change mean/maximum streamflow predictions by 7%/9% for a rating curve model, and 13%/25%, for a hydrological model (GR4J) in an ephemeral catchment. In contrast, the influence (0.2%/2.3%) was far less in a humid catchment. Assuming the data are of high quality this indicates deficiencies in the ability of the GR4J model structure to reproduce the flow regime in the ephemeral catchment. The linear Cook's distance-based metric produced reasonably similar rankings to the case-deletion metrics at a fraction of the computational cost (300-1000 times faster), but with less flexibility to rank influence using specific aspects of model behaviour. The nonlinear Cooks distance produced rankings that were virtually the same as the case-deletion metrics for all case studies - this highlights the importance of its use for nonlinear hydrological models. Visual assessment was not a reliable method of influence analysis as there was no direct relationship between the most influential data and the highest observed streamflows. The findings establish the feasibility and importance of including influence detection diagnostics as a standard tool in hydrological model calibration.

2.1 Introduction

The process of hydrological model calibration involves the estimation of parameters that maximise the similarity between observed and simulated hydrological response time series such as streamflow. This process requires the optimisation of one or several objective functions [Duan et al., 1992], which provides a summary measure of overall model performance. However in doing so, information on the influence of individual data points in determining the calibrated parameter set (and hence the model predictions) is often ignored.

Identifying data points that have a large influence on hydrological predictions is of particular importance when those data points are erroneous, as this is likely to lead to sub-optimal model performance when applied to an independent dataset. The importance of such "disinformative" data has been highlighted by Beven and Westerberg [2011], who identify the need for more formal methods to identify and remove erroneous data prior to model calibration. They suggest two strategies: firstly that the discrepancies of a water balance time series are evaluated for values outside some acceptable limits of uncertainty, and secondly that likelihood measures are developed that are robust with respect to disinformation. However, examining all of the high residual data can be labour intensive, and focusing only on a smaller subset of influential data is likely to be more feasible in practice. Furthermore, not all influential points are erroneous; in fact, in certain situations it may even be desirable that some data points are more influential than others. For example, objective functions that place a larger weight on high flows may be more desirable if the application is for peak flow prediction (e.g. [Duan et al., 2007]). This paper aims to provide hydrological modellers with the tools to assess relative influence of data points on model calibration.

Influential data points are defined as points that exert a disproportionate impact on the calibrated parameters, performance and/or predictions. Formal influential point detection methods are widely used both for the detection of erroneous points and for identifying possible model deficiencies, with common applications in linear regression [Cook, 1979], generalised linear regression [Thomas and Cook, 1989], generalised additive models [Hastie and Tibshirani, 1990] and various other regression-based approaches [Chen et al., 2012; Russo et al., 2009]. The diagnostics can be grouped into two classes: case-deletion approaches and analytical leverage-based approaches.

Case-deletion methods were first developed by Cook [1977] and involve removing ("deleting") a data point ("case") from the set of calibration points, and then recalibrating the model. Parameter estimates and model predictions from the recalibration are compared to the results from the full calibration, and this is repeated for all data points in the calibration set. A recent example in the context of flood frequency analysis used case-deletion to show that low flow outliers can have a disproportionate influence on extreme flood quantile estimates [Lamontagne et al., 2013]. Their technique was based on a generalised Grubbs-Beck test statistic developed by Cohn et al. [2013] that is designed to identify potentially influential low flows.

Case-deletion approaches can be computationally intensive, as they require the re-estimation of the parameters after deleting each point from the calibration data set. Furthermore, case-deletion involves comparing the optimal parameter sets from each cali-

brated model run, and thus anomalous results are possible for models with complex response surfaces that are prone to local optima [Duan et al., 1992]. As an alternative, Cook's distance [Cook, 1977] provides an analytical measure of the influence of points using only the final calibrated model, and thus does not require multiple re-calibrations. It combines measures of the distance between each observed data point and the fitted model (the residual) and the distance of each data point from the centre of the input space (the leverage). Cook's distance was originally developed for linear regression models, but may also be applied to nonlinear models if the models are approximately linear in the vicinity of the optimum parameter set [Cook and Weisberg, 1982]. Alternatively, nonlinear formulations of the leverage are also available [St. Laurent and Cook, 1992], and may be better suited to the highly nonlinear behaviour of many hydrological models [e.g. see discussion in Kavetski and Kuczera, 2007].

The influence concepts are illustrated in Figure 2.1 by applying case-deletion to a linear regression model. Point A is highly influential, with a significant difference in calibrated parameters when including this point (β_0=2.0,β_1=2.3, compared to β_0=3.4, β_1=1.9). The influence on predictions is also evident by comparing the fitted regression lines, with the greatest differences apparent towards the high and low extremes of the input data. In contrast, although point B has a similar residual to A (i.e. the difference between the data point and the fitted curve is similar), it exerts a much smaller influence on both the parameters (β_0=3.8, β_1=1.9) and the fitted regression line. Although the application of influence diagnostics may appear trivial in this example, the complex mapping from input to output space in hydrological models often precludes visual techniques, so that more formal approaches for the detection of influential points are required.

The prospect that a small number of data points can exert a very high influence on model performance motivates the more widespread implementation of influence diagnostics in hydrology, however applications have been few and recent. In the context of groundwater modelling, Yager [2004] found that models were highly sensitive to small changes in influential data. Foglia et al. [2007] used a series of case-deletion metrics and Cook's distance approaches on a groundwater model and found similar performance between the two metrics. Foglia et al. [2009] applied linear Cook's distance as part of a suite of diagnostics to a short time series of 37 daily observations in the rainfall-runoff model TOPKAPI and found that some of the low flow observations during small precipitation events were more important than anticipated. Legates and McCabe [1999] discuss the oversensitivity to outliers of correlation based goodness-of-fit measures used in hydrological models and recommend that additional evaluation measures should supplement calibration. Berthet et al. [2010] found a quadratic criterion to be influenced by a very

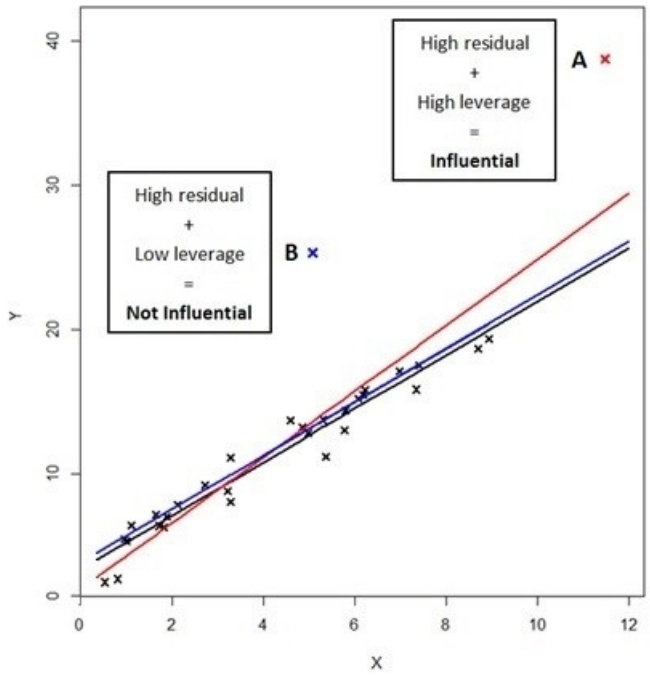

Figure 2.1: A simple linear regression scatter plot illustrates the impact of a highly influential data point on the fitted model. The black line is the prediction curve without point A or B in the calibration data; the red prediction curve is with point A only included, an observation that is both an outlier and a high leverage point; the blue prediction curve is with point B only included, an observation with the same residual as point A but with low leverage.

small number of time steps characterised with high runoff variation. Perrin et al. [2007] assess the impact of the quantity and quality of streamflow data on parameter calibration and model robustness and show that a subset of influential points from a larger dataset are sufficient to obtain robust estimates. Singh and Bárdossy [2012] pre-process hydrological data using depth functions to identify unusual events and investigate the calibration of the model with only this set of critical data to assess if the subset has enough information to identify model parameters. Each of these studies contributes towards the more widespread use of influence assessment, however a comprehensive assessment of the influence of individual data points in the context of hydrological model predictions and parameters is still lacking.

The goal of this paper is to evaluate the use of influence diagnostics in the context of common hydrological modelling case studies: stage/discharge rating curve model and a conceptual hydrological model. Case-deletion, linear and nonlinear Cook's distance will be compared in terms of performance and computational run times. Tailored statistics that are suitable for hydrological model applications will be developed for measuring the effect of data points on the model parameters, performance and/or predictions. This analysis will identify the extent to which the model predictions are influenced by a small number of data points - thereby evaluating the information content of data points and the benefits of including influence diagnostics as a standard tool in the process of hydrological model calibration.

The remainder of this paper is structured as follows. Section 2.2 outlines the methodology of the various approaches to quantifying influence. Section 2.3 introduces the case studies. Section 2.4 applies the numerical case-deletion diagnostics, applies the analytical Cook's distance diagnostics, provides a hydrological interpretation of the influence diagnostics, and evaluates the computational demand of the two classes of diagnostics. Finally, Section 2.5 discusses the importance of understanding the influence of data on hydrological predictions and the advantages and disadvantages of the numerical and analytical approaches.

2.2 Methods for assessing the influence of individual observations

2.2.1 Case-deletion influential point detection diagnostics

The case-deletion approach is widely used in the statistical literature to assess the impact of a deleted observation on the estimated parameters (Cook [1977]; Ross [1987]; and Chen et al. [2012]). It consists of evaluating the effect of excluding observations on the fitted model parameters, predictive performance and/or predictions and does not make strong assumptions on the model structure. As a continuous time series of rainfall inputs is required in hydrological modelling the case-deletion metrics consider the impact of masking an output data point from the objective function and therefore are calculated from model output only. Implementations of the approach differ largely in how the effect of the omitted observations is measured, and in this research we propose a number of measures that are specifically tailored to hydrological modelling applications.

2.2.1.1 Influence on the model predictions

Consider the following representation of a hydrological model:

$$\mathbf{Y} = h(\theta, \mathbf{X}) + \varepsilon \tag{2.1}$$

where $h()$ represents the hydrological model, $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ is an $n \times 1$ vector of observed inputs (such as precipitation and evapotranspiration), $\mathbf{Y}$ is an $n \times 1$ vector of observed responses (streamflow), and $\theta = \{\theta_1, \theta_2, ..., \theta_n\}$ is a $p \times 1$ vector of model parameters, and ε is a residual error models. Calibration of the hydrological model proceeds by finding a set of parameters ($\widehat{\theta}$) such that some measure of the distance between the observed responses $\mathbf{Y}$ and the predicted response and the predicted response $\widehat{\mathbf{Y}} = h\left(\widehat{\theta}, \mathbf{X}\right) + \varepsilon$ is minimised. For the case studies used in this paper, the residual model is assumed to be normally distributed, $\varepsilon \sim N\left(0, {\sigma^2}_{\varepsilon}\right)$, which results in a standard least squares objection function for model calibration.

Influence can be quantified by comparing predictions from a model with the whole data calibrated parameters ($\widehat{\theta}$), and the predictions using parameters estimated from censoring the data point ($\widehat{\theta}^{-i}$) in the objective function used in model calibration. Any model prediction can be compared in this way; for example the mean prediction relative change can be calculated using:

$$RelativeChange\,(\%) = \frac{mean\left(\widehat{\mathbf{Y}}\right) - mean\left(\widehat{\mathbf{Y}}^{-i}\right)}{mean\left(\widehat{\mathbf{Y}}\right)} \times 100 \tag{2.2}$$

Other metrics such as predictions of the median, minimum and maximum flows (or any other quantile of the flow duration curve) can be defined similarly. The choice of metric(s) should be based on the intended modelling objective. In this paper we have chosen to focus on the mean and maximum predictions, to illustrate the impact of influence on the average predictions and extreme predictions, which are commonly metrics used in hydrological modelling (e.g. flood risk).

2.2.1.2. Influence on the model parameters

An alternative influence measure is based on the distance between parameter vectors $\widehat{\theta}$ and $\widehat{\theta}^{-i}$. Given that hydrological model parameters often vary over different scales and can be highly correlated, we use the Mahalanobis distance (MD) [Mahalanobis, 1936] as a measure of the distance measure between the two parameter sets:

$$MD_i = \sqrt{\left(\widehat{\theta} - \widehat{\theta}^{-i}\right)^T \mathbf{C}^{-1} \left(\widehat{\theta} - \widehat{\theta}^{-i}\right)} \tag{2.3}$$

where $\mathbf{C}$ is the parameter covariance matrix. A bivariate example of the Mahalanobis distance is illustrated in Figure 2.2, for each $\widehat{\theta}^{-i}$ for $i = 1,\ldots,n$. The covariance matrix is estimated based on all n parameter sets, and thus the Mahalanobis distance for point i should be viewed as a measure of the influence of that data point relative to the remaining data points, rather than as an absolute measure of influence.

The metric given in Eq. (2.3) assumes that the joint distribution of parameters can be described by a multivariate Gaussian distribution; this may not always be appropriate, with extensions to the Mahalanobis distance beyond the Gaussian distribution given in Ekstrom [2011].

2.2.1.3. Influence on the model predictive performance

The influence of a data point on model performance can be quantified by considering the change in the objective function based on the whole data calibrated parameters $\widehat{\theta}$ and the case-deletion parameters $\widehat{\theta}^{-i}$. We use the term "objective function displacement" (OFD) as general measure of the difference in predictive performance due to including/excluding individual data points.

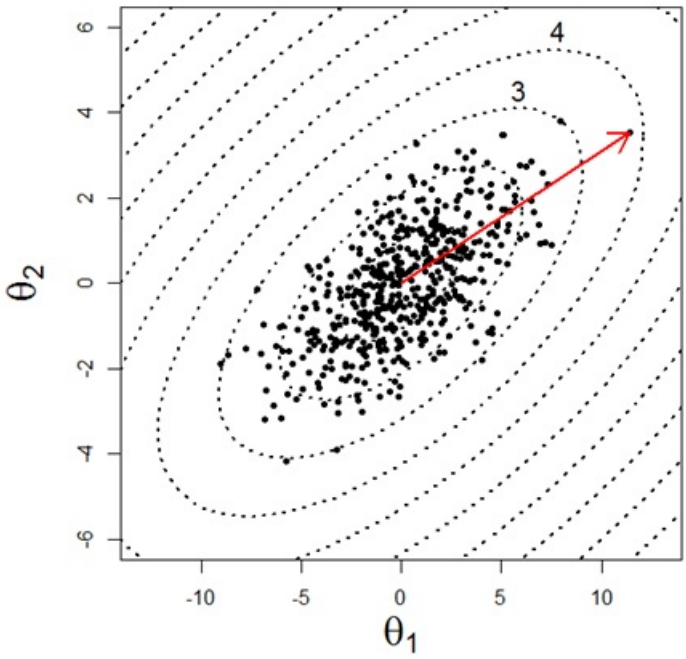

Figure 2.2: Example of Mahalanobis distance in two dimensions. The origin is the parameter set obtained by calibrating to the full calibration dataset (i.e. $\widehat{\theta}$), and the contours represent equal Mahalanobis distance from the origin. The highlighted point has a Mahalanobis distance of 3.8 from the origin (bivariate standard error of 3.8).

The concept of OFD is illustrated in Figure 2.3. If $OF(\widehat{\theta})$ and $OF(\widehat{\theta}^{-i})$ are both evaluated over the entire time series then we can never expect $OF(\widehat{\theta}^{-i})$ to outperform $OF(\widehat{\theta})$ as the parameters $\widehat{\theta}^{-i}$ are calibrated on a different data set to $\widehat{\theta}$. Therefore to isolate the influence of an observation on the remaining observations we consider the OFD evaluated with the i^{th} case excluded:

$$OFD_i = \left| OF_{-i}(\widehat{\theta}^{-i}) - OF_{-i}(\widehat{\theta}) \right| \tag{2.4}$$

Likelihood based methods for model calibration are commonly applied in hydrology [Westra et al., 2014, Evin et al., 2013, Renard et al., 2010] and the OFD is similar to the likelihood displacement (LD) [Cook and Weisberg, 1982] which can be used when using likelihood-based methods for model calibration. Let $L(\theta)$ denote the log-likelihood function. Then the likelihood displacement is defined as:

$$LD_i = 2\left\{ L_{-i}(\widehat{\theta}^{-i}) - L_{-i}(\widehat{\theta}) \right\} \tag{2.5}$$

with the form of the LD defined to be analogous to the residual deviance, which is minus twice its log-likelihood [Hastie et al., 2009].

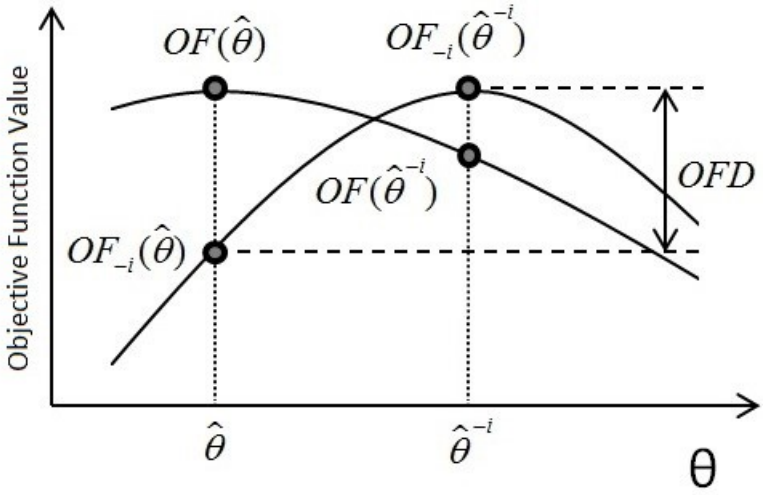

Figure 2.3: Defining the objective function displacement (OFD). Model calibration involves finding the parameter set that maximises (or, analogously, minimise) the objective function value. Any variation in parameters about the optimal set would produce decay in model performance. The OFD is defined as the difference in objective function value when applying parameter sets $\widehat{\theta}$ and $\widehat{\theta}^{-i}$ to the calibrated data with the i^th point removed

In this study we choose to use the OFD (Eq (2.4)) instead of LD in (Eq. (2.5)) to ease interpretability.

2.2.2 Analytical influential point detection diagnostics

The case-deletion method is computationally expensive, requiring n+1 model calibrations: n calibrations with one data point removed in each calibration, and one calibration with all data points included. Cook [1977] developed a distance metric that enabled the estimation of the influence of individual points using only a single full calibration. In this research we consider a linear and nonlinear version of this metric.

2.2.2.1. Linear Cook's distance

Cook's distance (CD) is calculated by accounting for both the leverage of the input data and the residual between the observed and fitted response. A point that is far from the centre of the input range will typically have high leverage, and a higher value of leverage means that the observation will have a greater influence on the model parameters or predictions. In linear regression the vector of predicted values $\widehat{\mathbf{Y}}$ is calculated from:

$$\widehat{\mathbf{Y}} = \mathbf{HY} = \mathbf{X}\widehat{\theta} \tag{2.6}$$

where $\mathbf{H}$ is the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{2.7}$$

The leverage value h_i of an observation corresponds to the i^{th} diagonal element of the $n \times n$ $\mathbf{H}$ matrix and is independent of the model fit. The matrix $\mathbf{H}$ has the following properties: the leverage values are constrained by $0 \leq h_{ii} \leq 1$, and the sum of the diagonal elements is equal to the number of parameters in the model p so that the average leverage value is p/n [Hoaglin and Welsch, 1978; Stuart et al., 2004].

A point with both a large residual and high leverage exerts influence on the regression coefficients, in the sense that if the observation is removed, the parameters change considerably. Cook's distance can be calculated from the standardised model residual r_{Si} and leverage h_i [Fox and Weisberg, 2011]:

$$r_i = Y_i - \widehat{Y}_i \tag{2.8}$$

$$r_{Si} = \frac{r_i}{\widehat{\sigma_\varepsilon}\sqrt{1-h_i}} \tag{2.9}$$

$$\text{Cookdistance}_i = \frac{(r_{Si})^2}{p} \times \frac{h_i}{1-h_i} \tag{2.10}$$

Where $\widehat{\sigma_\varepsilon}^2$ is the estimate of the residual error variance, from Eq. (2.1). Because CD is based on the residuals obtained from one model calibration and the leverage obtained from matrix multiplication, it is computationally far cheaper to calculate than case-deletion.

High leverage is commonly defined as $h_i > 2(p/n)$, a high standardised residual is commonly defined as greater than 2, and a highly influential point is defined $CD_i > 1$ (Stuart et al. [2004], Fox and Weisberg [2011] and Hoaglin and Welsch [1978]).

The different components of CD are illustrated in Figure 2.4 using the same data as in Figure 2.1. The model fit (Fig 2.4a), residuals (Fig 2.4b), leverage (Fig 2.4c) and CD (Fig 2.4d) are all presented, and illustrates that CD accounts for both the residual and the leverage. For example, point 'c' has both a high residual and high leverage value, and therefore has high influence. Point 'b' also has a high residual, however its influence is much lower due to its low leverage value. Point 'a' has high leverage but a low residual and therefore low influence.

2.2.2.2. Nonlinear Cook's distance

In linear models, the definition of leverage only depends on the observed input data, while in nonlinear models it is dependent on the local sensitivity of the model to small perturbations in model parameters [St. Laurent and Cook, 1992]. We use Jacobian leverage,

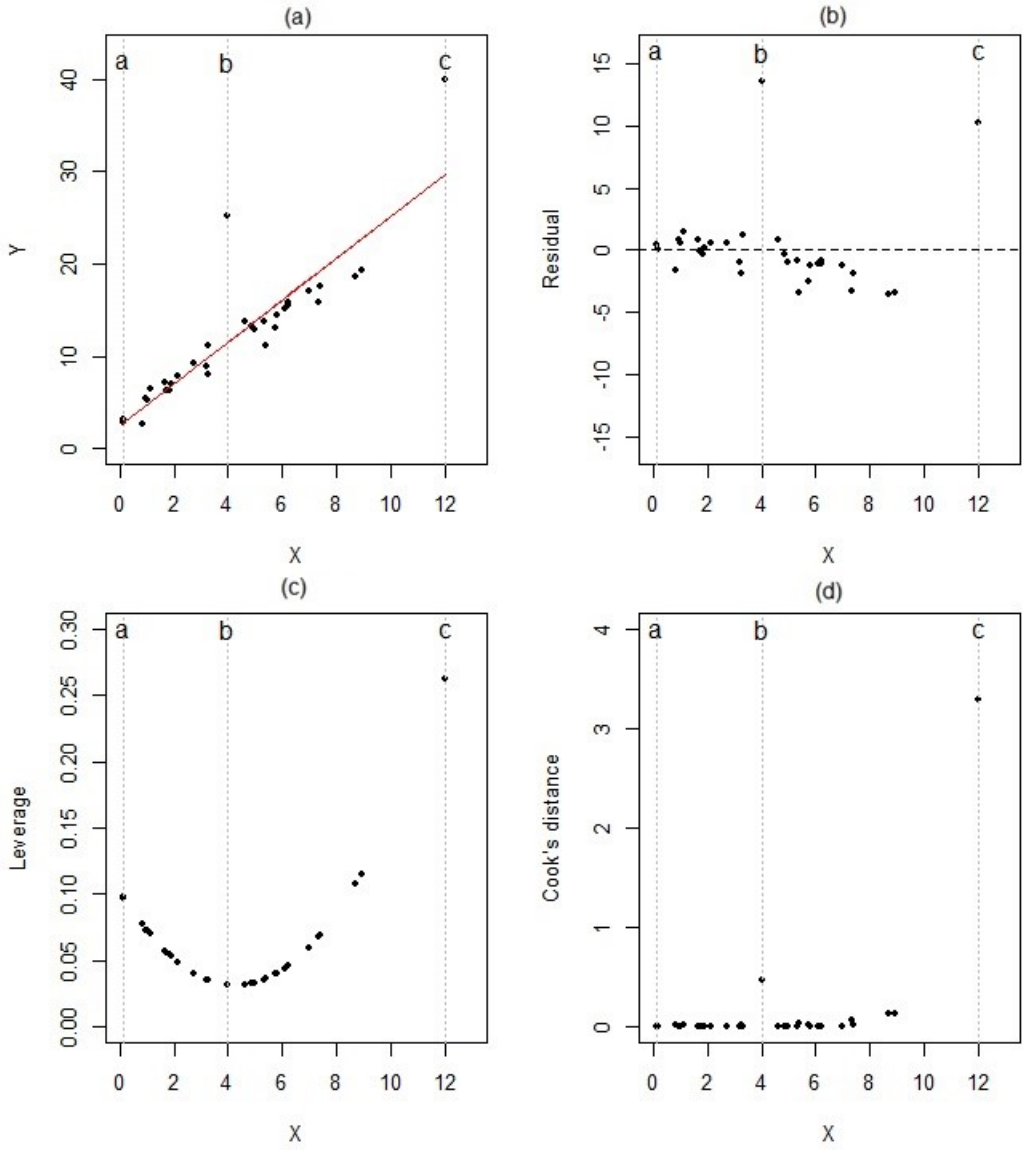

Figure 2.4: Example application of Cook's distance in a simple linear model. (a) a scatter plot generated from $Y = 2X + 3$ with added Gaussian noise and the fitted curve shown in red, (b) raw model residuals against X (c), linear leverage against X, and (d) Cook's distance against X.

a special case of the generalised leverage developed by St. Laurent and Cook [1993], to construct a second-order approximation to the nonlinear response function, where the nonlinear leverage is defined as h^{nl} equal to the diagonal elements of the matrix $\widehat{\mathbf{J}}$

$$\widehat{\mathbf{J}} = \widehat{\mathbf{V}}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}} - [\widehat{\mathbf{r}}][\widehat{\mathbf{W}}])^{-1}\widehat{\mathbf{V}}^T \tag{2.11}$$

Here $\widehat{\mathbf{V}}$ is the $n \times p$ matrix with i^{th} row $\partial f_i(\theta)/\partial(\theta)$, describing the effect of the perturbation of the parameters on the model predictions, $\widehat{\mathbf{W}}$is the $n \times p \times p$ array with n elements $\partial^2 f_i(\theta)/\partial\theta_j\partial\theta_k$ of the $p \times p$ hessian matrix $\widehat{\mathbf{W}}_i$ of $h_i(\theta, \mathbf{X})$, and $\widehat{\mathbf{r}}$ is the $n \times 1$ vector of fitted model residuals (2.8).

Analytical derivatives are typically not available for lumped hydrological models, and therefore estimates of $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{W}}$ can be obtained via finite difference numerical approximation and/or automatic differentiation [Nocedal and Wright, 2006] which is a standard practice in hydrological modelling [Abdulla et al., 1999; Martinez and Gupta, 2011; Mein and Brown, 1978; Vandewiele et al., 1992; Williams and Yeh, 1983]. We use finite difference approximations with a Richardson extrapolation following the procedure of Nocedal and Wright [2006].

Substituting h^{nl} into (2.10) allows us to calculate nonlinear CD which may be more suitable for models with a nonlinear relationship between the predictor and the response. St. Laurent and Cook [1992] show that in nonlinear regression models the constraint $0 \leq {h_i}^{nl} \leq 1$ no longer holds, and cases where ${h_i}^{nl} > 1$ are defined as "superleverage". Therefore, the nonlinear CD metric can have a greater magnitude than the linear CD calculated from the same data.

2.3 Case studies

To evaluate influence diagnostics in the context of hydrological modelling we consider three case studies. Firstly a rating curve model with a short time series and parsimonious model structure is used to demonstrate concepts of influence, and secondly the methods are applied to an ephemeral and a humid catchment with the conceptual hydrological model GR4J [Perrin et al., 2003] to demonstrate the method for a typical hydrological model calibration problem. The case studies are introduced here and the results from applying the influence diagnostics to these case studies are described in Section 2.4.

2.3.1 Case study: Rating Curve Model

A stage/discharge rating curve model was selected due to its simple model structure and short time series to enable visual evaluation of influence diagnostics for individual observations. Site chosen was the experiment catchment, Mahurangi College [Woods et al., 2001], located in the Northland region 50 *km* north of Auckland, New Zealand. The catchment area is 46 km^2, with mean annual rainfall of 1600 *mm*, runoff of 860 *mm* and pan evaporation of 1310 *mm*. There are 27 years of 15 minute interval streamflow data, and 24 stage/discharge gaugings. The rating curve model was a two-part piecewise power function:

$$Y = \begin{cases} \theta_1 X^{\theta_2} & X < 1.5 \\ \theta_1 1.5^{(\theta_2-\theta_3)} X^{\theta_3} & X \geq 1.5 \end{cases} \tag{2.12}$$

where X is river stage (m), and Y is the river discharge (m^3/s).

Calibration was performed using a standard least squares likelihood objective function optimised using the shuffle complex evolution (SCE) algorithm [Duan et al., 1992; Duan et al., 1994]. To reduce the computation burden when undertaking re-calibration for each of the case-deletion points, the SCE algorithm was seeded using bounds of +/- 5% of the optimal parameters from the full data calibration.

2.3.2 Case study: Conceptual Rainfall-Runoff Model

The GR4J model was selected due to its parsimonious model structure and its good performance across a wide range of catchment conditions [Perrin et al., 2003]. GR4J has four calibration parameters: the production store capacity (θ_1 , units of mm), the groundwater exchange coefficient (θ_2, units of mm), the one day-ahead maximum capacity of the routing store (θ_3, units of mm), and the time base of the unit hydrograph (θ_4, units of days).

To facilitate visual inspection of the results, the hydrological time series was restricted to two years with an additional one-year warmup period. GR4J was calibrated using the same prodcedure as the rating curve model.

2.3.2.1. Ephemeral Scott Creek Catchment

The Scott Creek (South Australia) has an area of 29 km^2 and experiences a mean annual rainfall of 992mm, median annual potential evapotranspiration (PET) of 1600 mm, and mean annual runoff of 147 mm. The catchment has a semi-arid climate with a

winter-dominated rainfall regime, due to the low runoff coefficient of 0.14, is classified as ephemeral. The Scott Creek catchment was selected as the GR4J model as it has previously been successfully used by Westra et al. [2014]. The model calibrated with the whole data set obtained an NSE of 0.79 and a prediction bias of 0.09 mm/day.

2.3.2.2. Humid French Broad River Catchment

The French Broad River (North Carolina, USA) was selected from the MOPEX data set [Duan et al., 2006] as a humid catchment (runoff coefficient of 0.57) to contrast the ephemeral Scott Creek catchment. French Broad River has an area of 2448 km^2, mean annual rainfall of 1413 mm, and mean annual runoff of 800 mm. The model calibrated with the whole data set obtained an NSE of 0.86 and a prediction bias of 0.02 mm/day.

2.4 Results: Application of influence diagnostics to case studies

We first evaluate the influence of individual observations by quantifing the impact of case-deletion on the model predictions and parameters. We then apply linear and nonlinear Cook's distance (CD) to the data sets and explore the contributions of the observation's leverage and residual to CD. We compare the linear and nonlinear CD metrics using the case-deletion OFD as a baseline, and then review all three methods in terms of their computational demand and the information they convey about each data point's influence.

In order to understand the relationships between the influence diagnostics we have highlighted specific points in the case study data sets. For the rating curve case study, points 20 and 23 are highlighted because they are the most influential in terms of all metrics. For GR4J Scott Creek, the two most influential points, day 117 and 121, were highlighted, However, as these days were from the same event the third most influential point, day 149, was also highlighted. For GR4J French Broad River, the two highest flow points, day 439 and 438, were highlighted. In addition, point 656 was highlighted because it changed from being identified as highly influential to having little or no influence depending on the metric used.

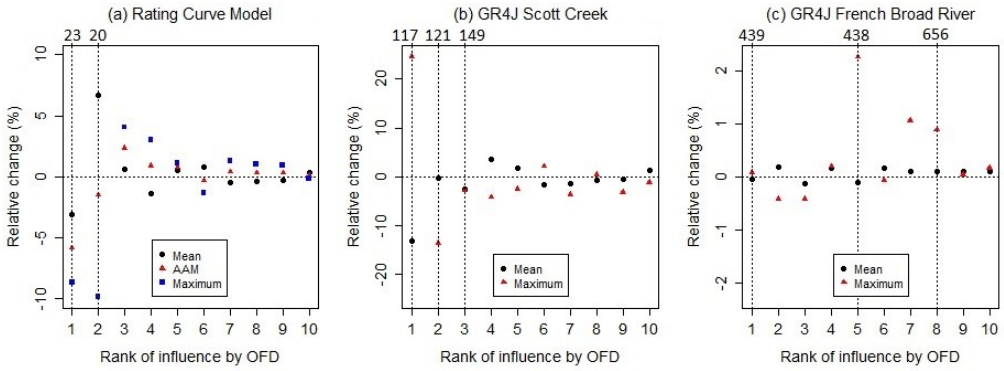

Figure 2.5: Impact of removing the top 10 most influential data points on model predictions.

2.4.1 Case-deletion: quantifying the influence of observations on calibration

We apply the case-deletion approach to the case studies to explore the influence of observations on the mean and maximum predictions and model parameter shifts.

2.4.1.1. Individual observation influence on model predictions

To highlight the impact of individual points, on hydrological predictions, the relative change in mean and maximum predictions (and average annual maximum for the rating curve case study) for the ten most influential points (identified based on the OFD) are plotted for the three case studies are plotted in Figure 2.5.

For the rating curve model the most influential point (point 23) changes the predicted mean flow by 6.7%. As rating curve models are often used to predict streamflow values greater than the highest streamflow gauging, and to illustrate the impact on extrapolated streamflow predictions, we consider two extrapolated values of the extreme predicted streamflow; (1) Average of the annual maximum (AAM) streamflow based the 27 years of streamflow estimated by the rating curve (corresponding to a river stage of 2.86 m) (2) Absolute maximum streamflow based on the highest estimated streamflow for the 27 years, (corresponding to a river stage of 4.2 metres). The most influential point in terms of OFD (point 23) changed the AAM flow by -5.9% and absolute maximum flow by -8.6%. To illustrate the influence of points on the predicted rating curve model, the full set of case-deletion prediction curves is presented in Figure 2.6. Here we see a range of 7.40 m^3/s (9%) for the average annual maximum streamflow prediction, and 35.6 m^3/s (13%) for the absolute maximum streamflow prediction. Here, we see that influence is increased when extrapolating the rating curve beyond the maximum observed value – this type of extrapolation is common in practise [Kuczera, 1996; Leonard et al., 2014].

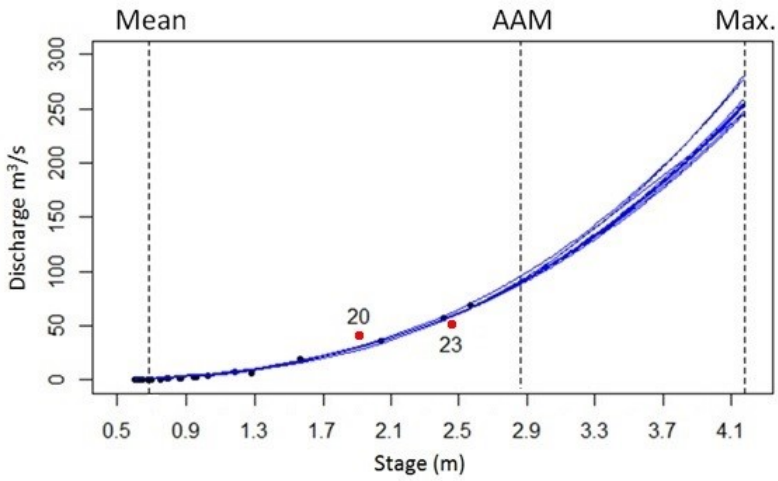

Figure 2.6: Influence of data on the extrapolation of rating curve results. The vertical dashed lines show the mean prediction stage, average annual maximum (AAM) prediction stage and absolute maximum prediction stage. The two most influential points are highlighted in red.

For GR4J Scott Creek, the impact of influential points is high, with the most influential point (day 117) changing the predicted mean and maximum flow by -13.3% and 25.0%, respectively. For GR4J French Broad River the influence of individual points is not as high as the other case studies. The most influential point (day 439) changes the mean and maximum predictions by only 0.2% and -0.8%, respectively. Further analysis is presented in Section 2.5.1, to interpret and understand the reasons for the differences in the magnitude of the influence for these case studies.

2.4.1.2. Individual observation influence on fitted parameters

To illustrate the influence of individual data points on the model parameters, we show example pairwise bivariate plots of of the model parameters for the case studies (Figure 2.7). While the plots only show pairwise relationships, the Mahalanobis distance is based on the full p-dimensional distance of each $\widehat{\theta}$ from $\widehat{\theta}^{-i}$.

Figure 2.7 shows the exclusion of a single observation can produce a significant change on calibrated model parameters. For the rating curve model, day 20 is clearly the most influential, with a Mahalanobis distance of 4.1, inducing a relative change of 10% in θ_2, the power parameter of the rating curve model. For GR4J Scott Creek, the impact on the parameters of the influential points is higher, with day 117 having a Mahalanobis distance of 26.9, inducing a relative change of 35% in θ_1 (production store capacity) and θ_2 (groundwater exchange coefficient). For GR4J French Broad river case study day 439 produces a Mahalanobis distance of 19.3, however, the relative change in parameters is

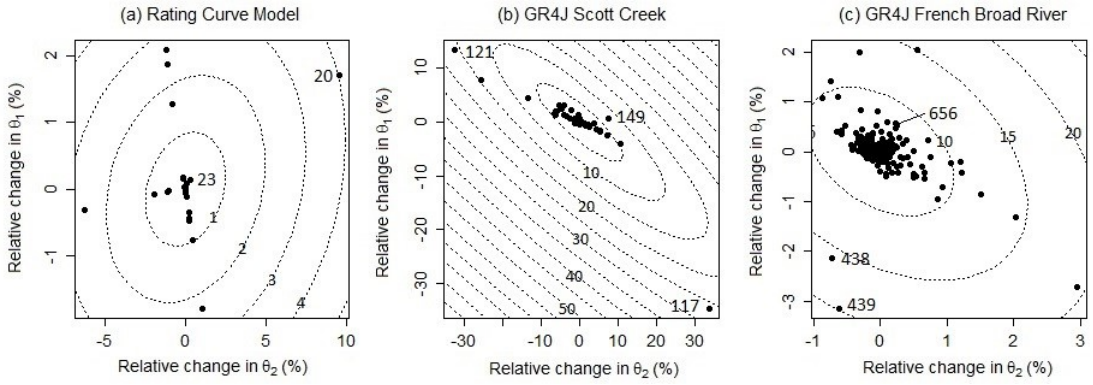

Figure 2.7: Example parameter scatter plots. The axes correspond to shifts in the parameters from the exclusion of a single observation in model calibration and the contour lines represent regions with equal Mahalanobis distance from the origin. Scatterplots for other parameter combinations showed similar trends.

only small, with a maximum relative change of 3% for θ_1 and θ_2.

2.4.2 Linear and nonlinear Cook's distance

The following sections will assess the computationally cheaper CD approaches to identify points and compares against the case-deletion OFD influence diagnostic.

2.4.2.1. Linear Cook's distance

Residuals and leverage both contribute to the CD and the identification of influential points. Figure 2.8 plots contours of linear CD (Eq. (2.10)) for the standardised residuals (Eq. (2.9)) and linear leverage (Eq. (2.7)) for each case study. The CD contours show that both a high magnitude residual and a high magnitude leverage are required to achieve a high magnitude CD. This is a practical demonstration of the principles demonstrated in the linear model in Figure 2.4.

For the rating curve model, point 23 has the highest linear CD value, followed by point 20, although the maximum linear CD is relatively low at 0.014. These points are also the two points with the highest Mahalanobis distance, although by that metric point 20 has a higher influence than point 23. For GR4J Scott Creek, the linear CD values are far higher than the rating curve model. Day 117 is the most influential, with a linear CD of 1.4. followed by day 149, with a linear CD of 0.17. For GR4J French Broad River, the linear CD is lower than GR4J Scott Creek, with day 656 having the highest linear CD of 0.03. Note the most influential point from linear CD (day 656) is not consistent with the most influential points identified by the previous metrics (Mahalanobis distance, OFD,

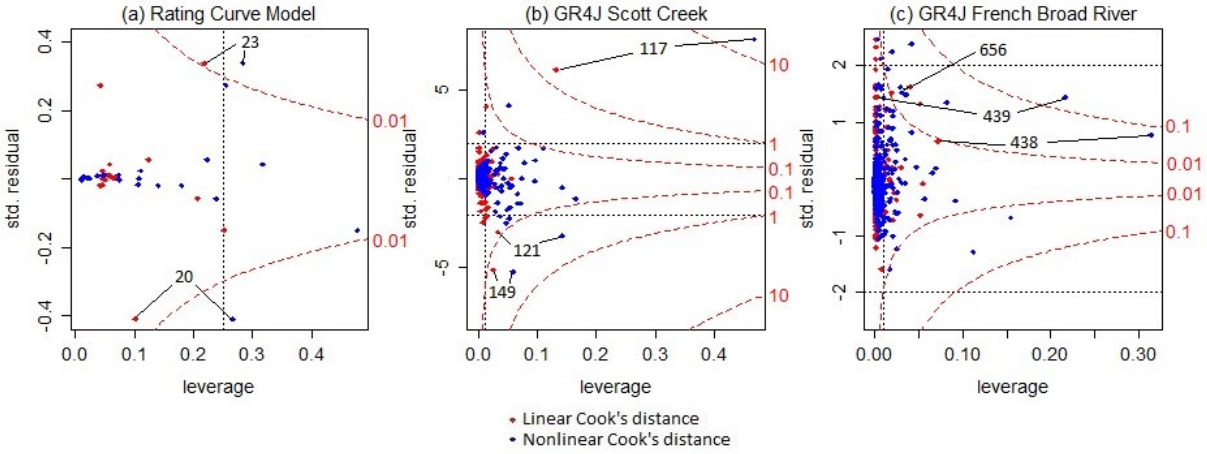

Figure 2.8: Components of Cook's distance. Linear cooks distance points are in red, while nonlinear are blue. The red broken lines represent regions with equal Cook's distance. The vertical dotted line is drawn at 2p/n indicating points with high leverage, and the horizontal dotted lines are drawn at $+/-2$ indicating points with a large standardised residual.

relative change in mean and maximum predictions).

2.4.2.2. Nonlinear Cook's distance

To assess if the nonlinear nature of conceptual rainfall-runoff models affects the identification of influential points, we compare the linear CD metric obtained from linear leverage with nonlinear CD obtained using nonlinear Jacobian leverage.

The nonlinear CD values are superimposed over the linear CD values in Figure 2.8. In general, for all three case studies the nonlinear leverage increases the magnitude of nonlinear CD relative to linear CD, and can lead to identification of influential points that are more consistent with the case deletion metrics.

In the rating curve model and GR4J Scott Creek case studies the nonlinear CD considerably increases compared with linear CD. For GR4J Scott Creek, the most influential point (point 117) increases from a linear CD of 1.4 to a nonlinear CD of 13.6, with several other points, having a significant increase from linear to nonlinear CD. For GR4J French Broad River case study, the use of nonlinear CD actually changes the identification of the most influential points, compared with linear CD. Day 656 was the most influential for linear CD, but with nonlinear CD, day 439 has a significiant increase in nonlinear CD, although it is still relatively low magnitude of 0.14. The most influential point of day 439 from nonlinear CD is also consistent with the case-deletion metrics (Mahalanobis distance, OFD, relative change in mean and maximum predictions).

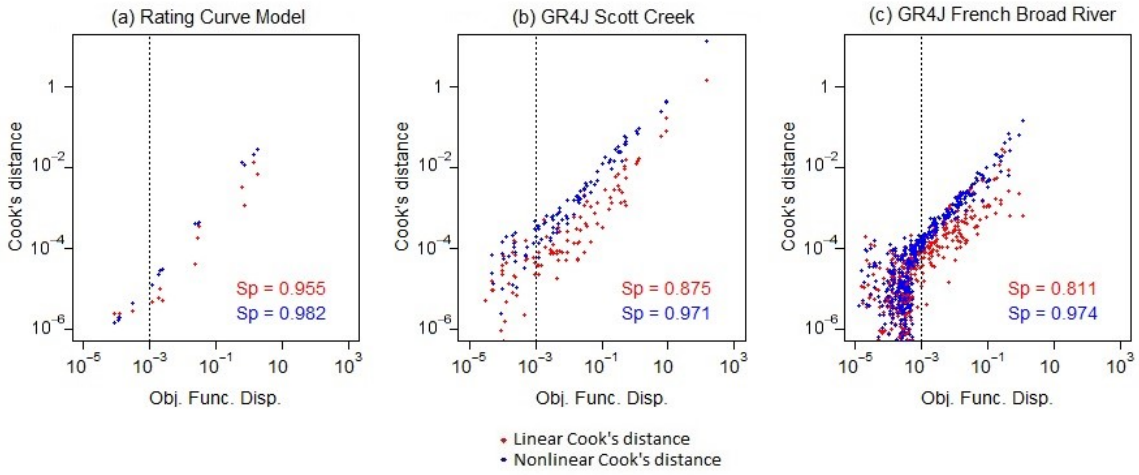

Figure 2.9: Cook's distance (red) and nonlinear Cook's distance (blue) plotted against the case-deletion OFD. Spearman ranking coefficients (Sp) are for above a OFD threshold of $1x10^{-3}$ shown with the vertical dotted line.

2.4.2.3. Comparison with Objective Function Displacement

The linear and nonlinear CD values are compared against OFD in Figure 2.9. The Spearman rank correlation coefficient (Sp) was used to numerically compare the three measures of influence. For GR4J case studies all points with an OFD below the SCE algorithm convergence tolerance of 10^{-5} were removed.

Including nonlinearity in the CD metric increases the consistency of the identified influential points with the case-deletion OFD. For the rating curve model, the OFD, linear and nonlinear CD are all in good agreement (Sp.$>$ 0.98), indicating similar influential points. For the most influential points (OFD $> 10^{-3}$) using the nonlinear CD, the Spearman correlation increases from 0.875 to 0.971 in Scott Creek and 0.811 to 0.974 in French Broad River. The improvement from the use of nonlinear CD is further discussed in Section 2.5.3.

2.4.3 Relationship between hydrological data and influence diagnostics

The hydrological observed and predicted data for the GR4J case studies is shown in Figure 2.10. For GR4J Scott Creek we see the most influential day (117) corresponds to the highest rainfall and streamflow, while the second (day 121) and third (day 149) most influential points are also relatively high rainfall and streamflow. For French Broad River, the most influential point (day 439) from nonlinear CD, MD and has the highest flow,

while the most influential point from linear CD (day 656) has relatively high flow. This suggests that highly influential points are also high flow values.

High flow points are not always the most influential as shown in Figure 2.11 where we compare the top 10 most influential points (ranked using OFD) with the observed streamflow data. We see that there is not a direct one-to-one relationship between the observed streamflow value and the rank of the influence. Similar to that seen in Figure 2.9, we also see that the nonlinear CD better identifies the rank of influential points compared to the OFD than the linear CD measure.

For the rating curve case study neither of the most influential points (20 and 23) are the highest streamflow. In the GR4J Scott Creek, the most influential point is the highest streamflow, however the 2nd and 3rd most influential points, correspond to far lower streamflow values than the highest streamflow. In the GR4J French Broad River, the most influential point is the 2nd highest streamflow according to OFD and the highest streamflow (day 438) is only the 5th most influential according to OFD. Interestingly the, linear CD doies not identify point 439 as influential and instead identifies point 656, the 5th highest streamflow.

Time series of precipitation, streamflow, Mahalanobis distance, OFD, linear and non-linear CD for the three case studies are included in the supplementary material.

2.4.4 Computational demand of influence diagnostics

The computational demand of the influence diagnostics are summarised in Table 2.1 for a general case, and for 10 and 30 years of daily data. We assume a smaller number of runs for each case-deletion re-calibration as the optimisation can be seeded close to the optima from the whole data set to reduce the computational burden.

Case-deletion is the most intuitive approach for influence assessment but also the most computationally intensive as it requires n+1 calibration runs (3 660 000 runs for 10 years). The case studies described in this study use relatively short periods of records (e.g. two years for the GR4J calibration), and the models are relatively parsimonious compared to many other hydrological models. Nevertheless the run-time for applying case-deletion to the GR4J model was about 14 hours on a 2.90 GHz processor. In practical applications we would expect a time series length (n) much larger than the two years used in the GR4J case study and we may choose to apply a less parsimonious conceptual model with more than four parameters (p). The resultant increase in both p and n would increase the computational demand of each model calibration, leading to further increases in computational demand.

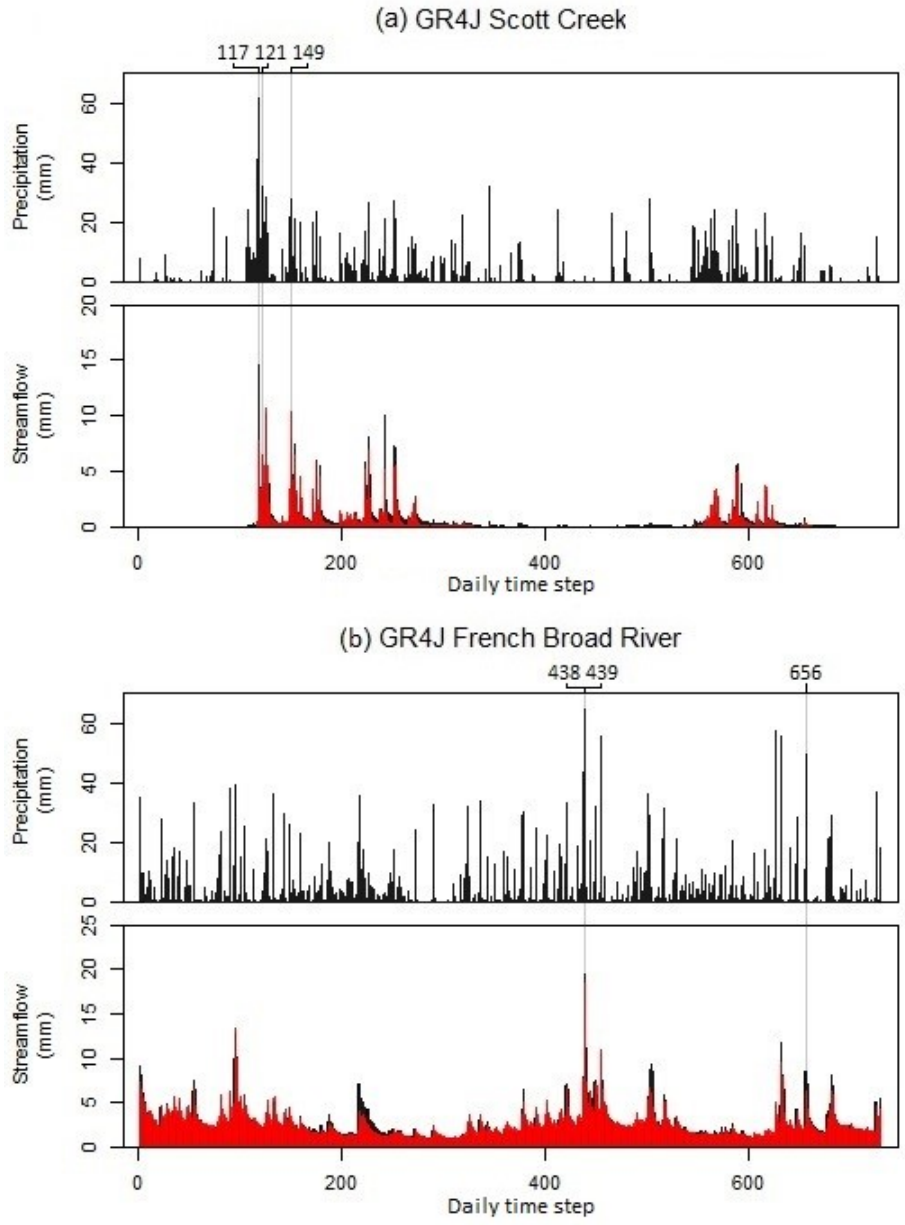

Figure 2.10: Observed precipitation and observed and predicted streamflow for the GR4J case studies.

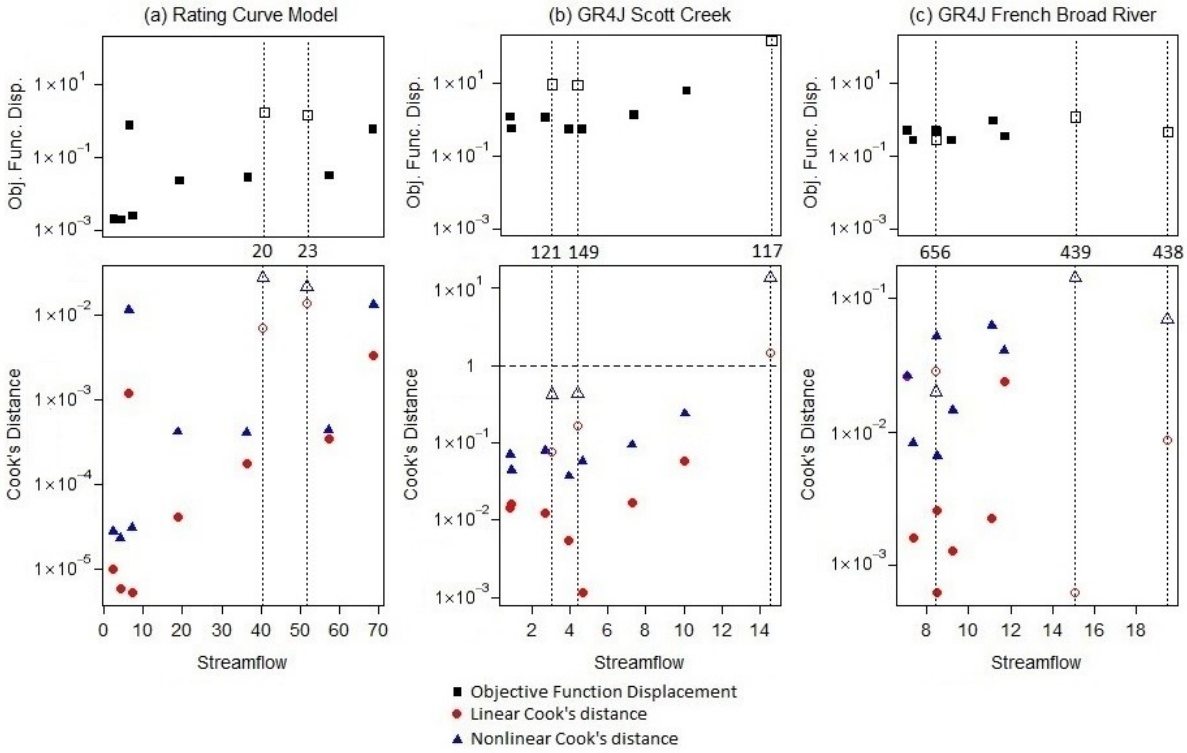

Figure 2.11: Comparison of influence diagnostics against observed streamflow for the top 10 influential points ranked by OFD. A Cook's distance value of 1 is shown with a horizontal broken line, and a selection of the influential points are identified with vertical dotted lines. The points identified with vertical dotted lines are shown with hollow points.

Regardless of the size of the calibration data set and hydrological model structural complexity linear CD requires only one model calibration followed by the application of linear matrix algebra. Nonlinear CD has the additional computational demand of calculating the finite difference approximations for the Jacobian and Hessian matrix to account for nonlinearity in the response surface, however both metrics are much more computationally tractable compared with case-deletion. For 10 years of data, the computational speed of analytical Cook's distance influence diagnostic is 300 times faster than case deletion, while for 30 years, it is 1000 times faster than case deletion.

Table 2.1: Summary of computational demands of influence diagnostics. Computational demands are based on a time series of length n and a model with p=4 parameters which requires 10 000 model runs for initial calibration and an additional 1000 runs for each case-deletion calibration starting from the whole data optima.

Influence diagnostic	**General computational demand**	**10 years of daily data**	**30 years of daily data**
Case-deletion	Repeating model calibration for the $n+1$ case-deletion scenarios	3 660 000 runs	10 960 000 runs
Linear Cook's distance	Single calibration + linear matrix algebra	10 000 runs	10 000 runs
Nonlinear Cook's distance	Single calibration $+2p$ and $2(pxp)$ model runs + linear matrix algebra	10 040 runs	10 040 runs

2.5 Discussion

This section discusses the importance of understanding the influence of data on hydrological predictions, and the advantages and disadvantages of the numerical and analytical influence diagnostics.

2.5.1 Importance of understanding the influence of data on hydrological model predictions

Hydrological models are an essential tool to aid policy makers and businesses to make decisions about water allocations and to design infrastructure. However, the influence of calibration data is seldom assessed during the model development process. Influence assessment based on CD is a computationally cheap addition to the existing model development process, with significant benefits in developing an understanding of which data points have the greatest influence on the calibrated model.

The potential impact of influential data in hydrological model calibration was illustrated in Section 2.4.1, where the Scott Creek case study showed that the exclusion of individual daily observations can have a substantial impact on model predictions. The magnitude of the prediction changes from removing one point (25% influence on the maximum prediction and 13% on the mean prediction) highlights the importance of including influence assessment in the existing hydrological model calibration framework, as changes in model predictions of this magnitude could have a large impact infrastructure design and future flood and drought risk. As Cook's distance is a function of both the residual and the leverage of the data point influence of data on uncertainty in predictions and posterior parameter distributions requires further theoretical development and will be explored in future research.

In the humid French Broad River case study the influence of data was less than the ephemeral Scott Creek catchment. In Scott Creek the smaller number of high rainfall and flow points (compared with the more consistent rainfall and flow events in French Broad River), produced data points with slightly higher leverage - see Figure 2.8. However, the primary contributor to the higher influence in Scott Creek, is the far higher value for the standardised residuals of > 5 for the most influential point (Figure 2.8b). Assuming the data are of high quality, this indicates deficiences in the ability of the GR4J model structure to reproduce the flow regime in this ephemeral catchment – see further discussion in Westra et al. [2014]. This indication that the GR4J model is very susceptible to highly influential flows in ephemeral catchments will need to be investigated with a large range

of varying flow regime case studies.

Once influential data is identified the modeller will need to make a decision as to whether they want to retain the identified data in the calibration set. Despite advances in methods of data collection, input data error is still common in hydrological modelling. Rain gauges may not be representative of total catchment rainfall, and the integrity of individual daily measurements may be impacted by power outages and human reading errors. Potential evapotranspiration are often estimated from pan evaporation measurements which may not be representative of the whole catchment. Flow measurements are often estimated from a rating curve and so may also be prone to human reading error, or stage measurements may be outside the range of measurements used to develop the rating curve. Scrutinising the entire time series for erroneous data is resource intensive. An advantage of the use of influence diagnostics, is that they can identify a smaller number of highly influential points that have the biggest impact on calibration, which can be scrutinised for data errors, thereby reducing effort and costs.

If the highly influential data is erroneous, removing or correcting the data point from the calibration set is likely to be the best course of action to mitigate its effects on model predictions. However, not all influential data is likely to be attributable to measurement errors. As the CD is a function of the residual and leverage, poor model predictions can have high influence, even if there is no error in the data. In such cases, rather than removing the influential data, the modeller may choose to retain the influential points if the data is characteristic of the intended model application. For example, if high flows were found to be influential this would be beneficial to model predictions in the case where the modelling objective is peak flow estimation.

Finally, in addition to the decision of whether to remove or retain data from the calibration set, knowledge of which points are likely to be influential can have value for experimental design. For example, identifying the types of data that are most influential in model calibration may allow for the targeted collection of informative data, rather than investing resources in collecting data that will have a minimal effect on the model predictions. Furthermore, for situations where influential data arises due to model misspecification (i.e., structural error, which often arises due to the highly abstracted nature of hydrological models relative to the system being investigated) and/or issues with the calibration strategy, it may be possible to use influence measures as a diagnostic to compare the performance of multiple alternative hydrological model structures.

2.5.2 Advantages of influence diagnostics over a visual assessment of the time series

In some cases influential data points can be identified by visual inspection of the hydrological time series alone, i.e. the highest streamflow/rainfall values are the most influential. This study showed this was not always the case - Section 2.4.3 showed there was not a clear one-to-one relationship between observed streamflow value and OFD in the three case studies (Figure 2.11). In hydrological model calibration visual inspection is more difficult due to longer time series, nonlinear model response, and correlation in predictions due to model storage. It is likely that influence diagnostic will become even more valuable with increasing hydrological model complexity, multi-catchment studies (e.g. > 200 in [Coron et al., 2012]), longer calibration data sets, more complex objective functions such as the generalised likelihood [Schoups and Vrugt, 2010], and consideration of the effects of persistence and heteroscedasticy [Evin et al., 2013]). Both the numerical and analytical influence diagnostics considered in this study quantify influence regardless of the magnitude of data or the complexity of model calibration strategy and can provide insights in cases where calibration information is otherwise limited.

2.5.3 Advantages and disadvantages of the different classes of influence diagnostics

The numerical case-deletion and analytical Cook's distance (CD) influence diagnostics vary in the way they quantify influence and in the complexity and computational demand of their application. Here we compare the advantages and disadvantages of the two classes of methods and discuss the impact of including nonlinearity in the CD formulation.

Case-deletion can be used on wide range of measures that quantify influence on model parameters, performance and/or predictions. This flexibility to tailor the case-deletion diagnostics to specific hydrological measures makes them an intuitive method of influence detection. Case-deletion has the additional advantage in that it makes no strong assumptions on hydrological model or assumed residual error model structure and can therefore be applied to any case study regardless of complexity. The major drawback with case-deletion is that for complex model calibration strategies combined with long calibration data time series, the use of high performance computing becomes essential if the method is to be feasibly applied in hydrological modelling. An additional concern is that individual calibration runs may find different local optima, potentially leading to significant differences in the optimised parameter sets. The response surface geometry

of many hydrological models is often highly complex, with common features including curving ridges, microscale and macroscale discontinuities and multiple optima [Duan et al., 1992]. The possibility of calibrating each case-deletion dataset to local optima may cause misleading results where two case-deletion calibrated parameter sets appear different from each other even if the data points have low influence on the actual model calibration. If the case-deletion metrics are to be implemented then the modeller should ensure that parameter optimisation approach can robustly handle complex response surface.

CD is a computationally cheap addition to existing hydrological model diagnostics and can be applied to any model calibration. The major drawback of CD is that its ranking of influence may be less interpretable when compared to ranking using the hydrologically orientated case-deletion diagnostics. For the three case studies considered there were significant benefits from including nonlinear leverage in the CD metric. The ranking of the influential points using nonlinear CD was far more consistent with case deletion metrics than linear CD. Figure 2.9 showed using nonlinear CD (cf linear CD) increased the Spearman ranking coefficient with OFD from 0.875 to 0.971 in the ephemeral Scott Creek case study, and from 0.811 to 0.974 in the humid French Broad River case study. Additionally, in Figure 2.11 we see that in French Broad River the linear CD metric identifies point 656 instead of 439, the most influential by case-deletion OFD and nonlinear CD.

The impact of model nonlinearity in hydrological models was highlighted by Duan et al. [1992] who discuss nonlinear parameter interaction in hydrological models, and Kavetski and Kuczera [2007] who highlight the difficulties in model calibration due to the highly complex and nonlinear nature of conceptual hydrological models. Similarly insights were found in this study – that influence diagnostics need to take into account the nonlinearity in of hydrological models. Further case studies will be needed to verify this conclusion.

The challenges with using CD is that it is unknown how it will be affected by more complex objective function formulations, and further research will need to be undertaken to understand how the choice of objective function impacts the diagnostic. As CD is a measure of the leverage and residual at a single time step it will not account for the "memory" or time lag of hydrological models, with storage errors propagating across multiple consecutive time steps. The consistency of the results of the nonlinear CD with the case deletion metrics indicated this was not a major problem, however more case studies are needed to verify this and may prompt the incorporation of time lag into CD in future studies.

2.6 Conclusions

Influential point detection diagnostics are not commonly used in hydrological modelling despite being regularly applied in the computational statistics literature since the introduction of Cook's distance (CD) in linear regression models in Cook [1977]. This paper evaluates the application of numerical case-deletion and analytical CD influence diagnostics in the context of three case studies: a stage/discharge rating curve model and the conceptual hydrological model GR4J over two contrasting datasets. We found that individual influential data points can have a substantial impact on model predictions in the ephemeral Scott Creek catchment but a relatively small influence on predictions in the humid French Broad River catchment. Assuming the data are of high quality this indicates deficiencies in the ability of the GR4J model structure to reproduce the flow regime in the ephemeral catchment, however this will need to be investigated with a large range of varying flow regime case studies.

Case-deletion approaches are capable of ranking the influence of individual data points using a wide variety of metrics, including the impact on:

1. model parameters, either individually or through an aggregated measure such as the Mahalanobis distance;

2. model performance, based on the difference between objective function values when a point is either included or excluded from the calibration set; and

3. model predictions, such as the mean, minimum or maximum flow or any other metric of interest.

Limitations of the case-deletion approaches include the significant run times associated with the large number of model re-calibrations, and the possibility of finding local rather than global optima in each re-calibration.

In contrast, methods based on CD are much more computationally tractable (300-1000 times faster) compared with case deletion and are less sensitive to the presence of local optima. However, specific values of CD are more difficult to interpret compared to the hydrologically orientated case-deletion measures. This study found that nonlinear CD provided a ranking of the influential points that was more consistent with the case deletion metrics compared with linear CD. This is likely due to the nonlinear structure of hydrological models. Further case studies will investigate the efficacy of nonlinear CD.

Regardless of the specific choice of metric, it is clear that influential point detection diagnostics can provide important insights into the mechanics of hydrological model

calibration, as well as providing a better understanding of the impact of individual data points on the calibrated model. Visual assessment was not a reliable method of influence assessment as there was no direct relationship between the most influential data and the highest observed streamflows. Application of influence diagnostics will allow the modeller to make informed decisions about including or excluding influential data and fine tune their adopted calibration strategy to ensure robust predictions for the intended model application.

2.7 Supplementary material

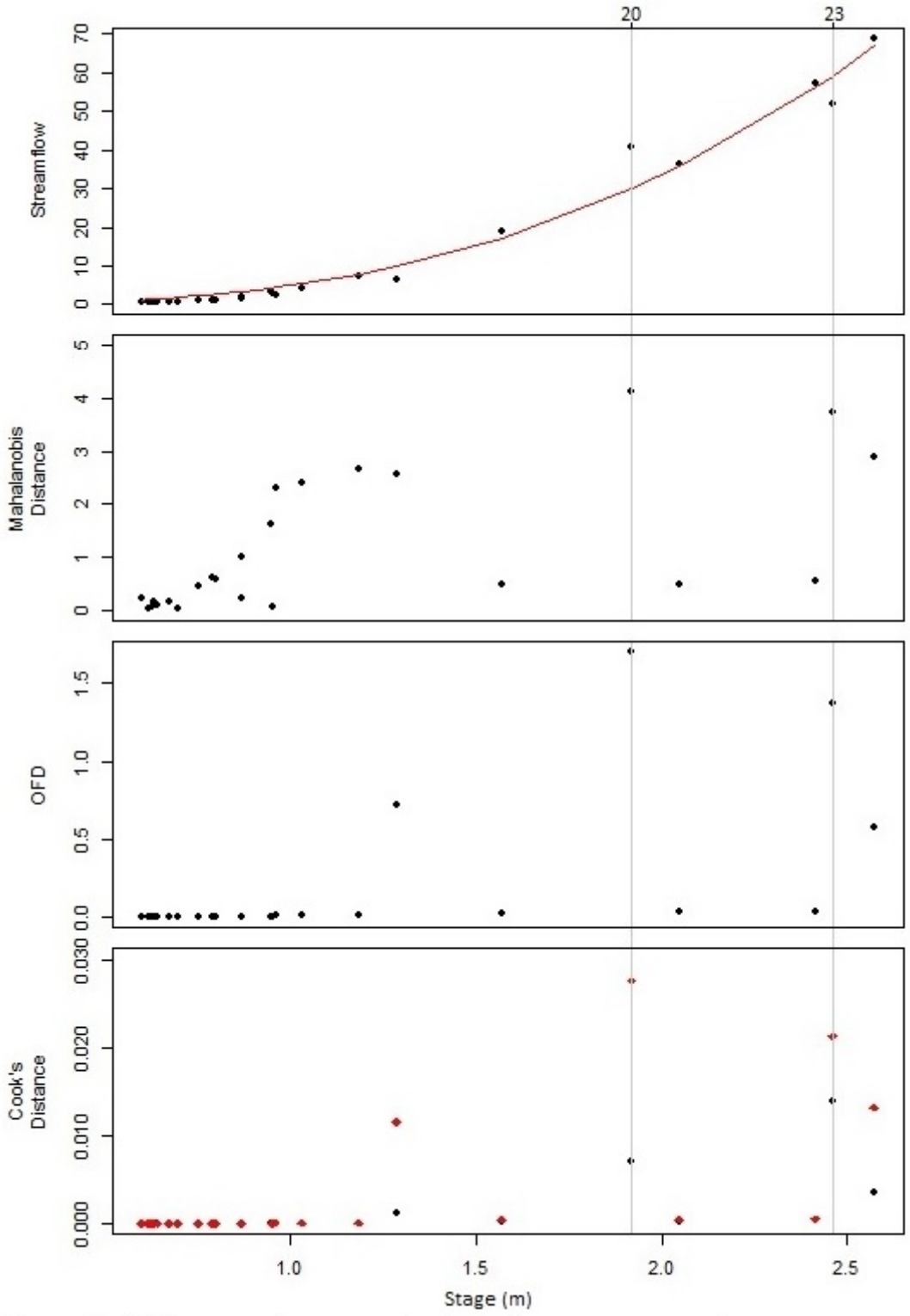

Figure S1 – Influence against stage for the rating curve case study

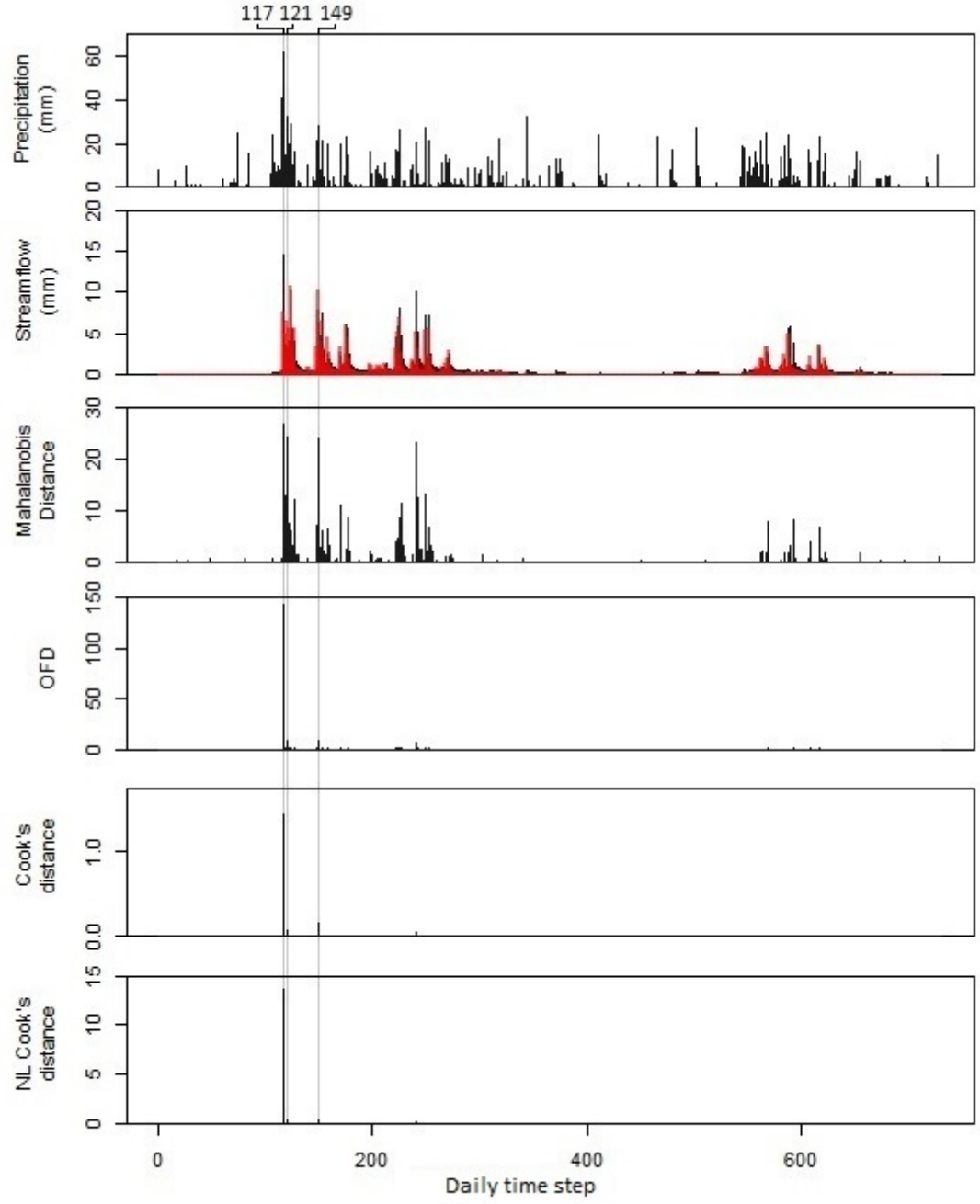

Figure S2 – Influence time series for the GR4J Scott Creek case study

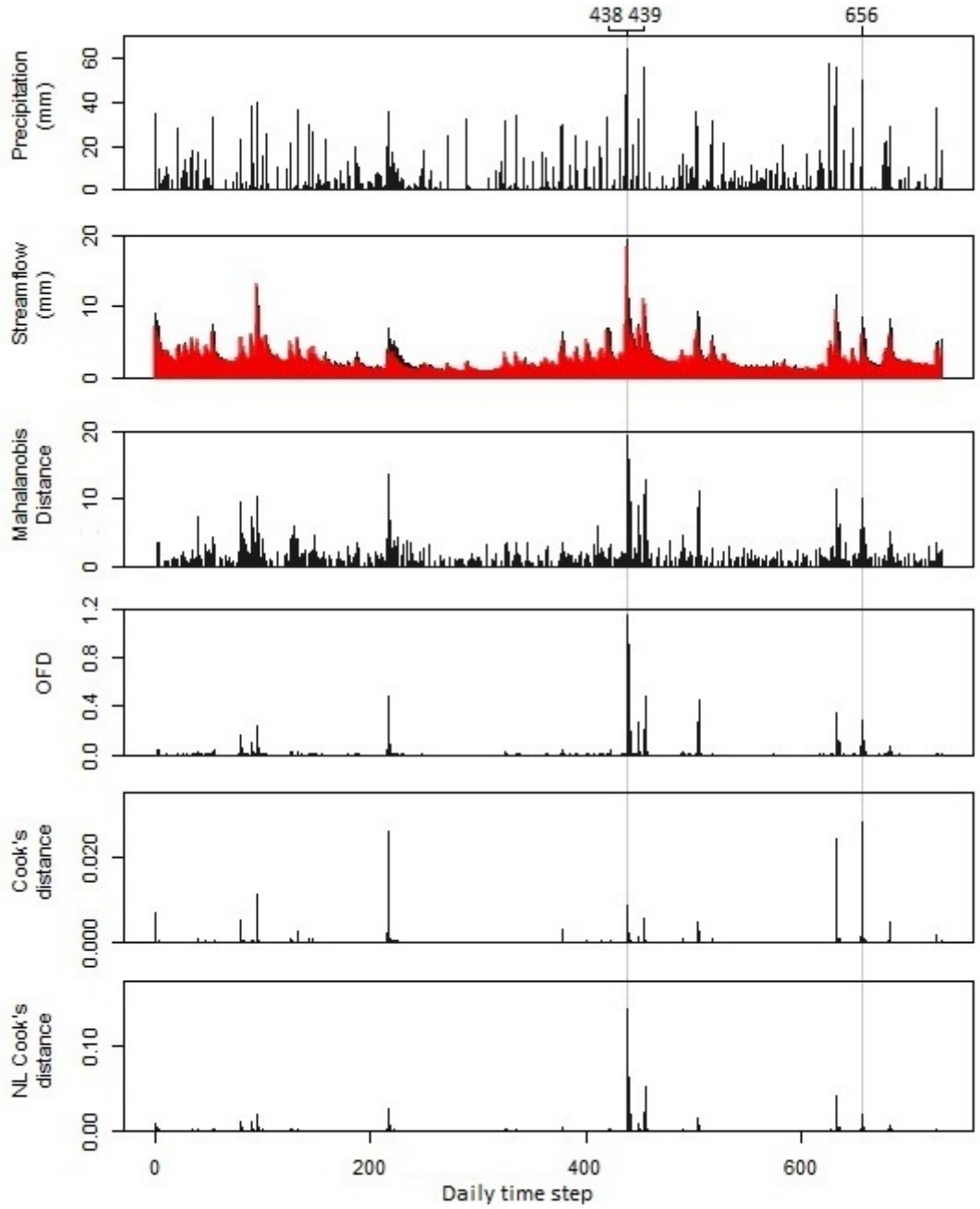

Figure S3 – Influence time series for the GR4J French Broad River case study

2.8 References

Abdulla, F. A., D. P. Lettenmaier, and X. Liang (1999), Estimation of the ARNO model baseflow parameters using daily streamflow data, Journal of Hydrology, 222(1-4), 37-54.

Berthet, L., V. Andréassian, C. Perrin, and C. Loumagne (2010), How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value

of a quadratic criterion, Hydrological Sciences Journal, 55(6), 1063-1073.
Beven, K., and I. Westerberg (2011), On red herrings and real herrings: disinformation and information in hydrological inference, Hydrological Processes, 25(10), 1676-1680.
Chen, X. D., N. S. Tang, and X. R. Wang (2012), Local influence analysis for semiparametric reproductive dispersion nonlinear models, Acta Math Appl Sin-E, 28(1), 75-90.
Cohn, T. A., J. F. England, C. E. Berenbrock, R. R. Mason, J. R. Stedinger, and J. R. Lamontagne (2013), A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series, Water Resources Research, 49(8), 5047-5058.
Cook, R. D. (1977), Detection of Influential Observation in Linear-Regression, Technometrics, 19(1), 15-18.
Cook, R. D. (1979), Influential Observations in Linear-Regression, J Am Stat Assoc, 74(365), 169-174.
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 Resour Res, 48(5), W05552.
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.
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.
Duan, Q. Y., N. K. Ajami, X. G. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources, 30(5), 1371-1386.
Duan, Q. Y., et al. (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320(1-2), 3-17.
Ekstrom, O. (2011), Mahalanobis' distance beyond normal distibutions, UCLA Statistics.
Evin, G., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resour Res, 49(7), 4518-4524.

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(6), W06427.
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.
Hastie, T., and R. Tibshirani (1990), Generalized Additive Models, CRC Press.
Hastie, T., R. Tibshirani, and J. Friedman (2009), Elements of Statistical Learning: Data Mining, Inference and Prediction (Second Edition), New York.
Hoaglin, and Welsch (1978), The Hat Matrix in Regression and ANOVA, The American Statistician, 32, 17-22.
Hsu, K. L., H. V. Gupta, and S. Sorooshian (1995), Artificial Neural-Network Modeling of the Rainfall-Runoff Process, Water Resources Research, 31(10), 2517-2530.
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), W03411.
Kuczera, G. (1996), Correlated rating curve error in flood frequency inference, Water Resources Research, 32(7), 2119-2127.
Lamontagne, J., J. Stedinger, T. Cohn, and N. Barth (2013), Robust National Flood Frequency Guidelines: What Is an Outlier?. : , World Environmental and Water Resources Congress 2013, 2454-2466.
Legates, D. R., and G. J. McCabe (1999), Evaluating the use of ”goodness-of-fit” measures in hydrologic and hydroclimatic model validation Water resources Research, 35(1), 233-241.
Leonard, J., M. Mietton, H. Najib, and P. Gourbesville (2014), Rating curve modelling with Manning’s equation to manage instability and improve extrapolation, Hydrological Sciences Journal, 45(5), 739-750.
Mahalanobis, P. C. (1936), On the generalized distance in statistics, Proceedings National Institute of Science, India, 2(1), 49-55.
Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, Water Resources Research, 47(12), W12540.
Mein, R. G., and B. M. Brown (1978), Sensitivity of Optimized Parameters in Watershed Models, Water Resources Research, 14(2), 299-303.

Nocedal, J., and S. J. Wright (2006), Numerical Optimization, Springer. Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279(1-4), 275-289.
Perrin, C., L. Oudin, V. Andreassian, C. Rojas-Serna, C. Michel, and T. Mathevet (2007), Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models, Hydrological Sciences Journal, 52(1), 131-151.
Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour Res, 46.
Ross, W. H. (1987), The Geometry of Case Deletion and the Assessment of Influence in Nonlinear-Regression, Can J Stat, 15(2), 91-103.
Russo, C. M., G. A. Paula, and R. Aoki (2009), Influence diagnostics in nonlinear mixed-effects elliptical models, Comput Stat Data An, 53(12), 4143-4156.
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.
Singh, S. K., and A. Bárdossy (2012), Calibration of hydrological models on hydrologically unusual events, Advances in Water Resources, 38, 81-91.
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
Stuart, A., K. Ord, and S. Arnold (2004), Kendall's Advanced Theory of Statistics: Volume 2A: Classical Inference and the Linear Model, 6th edition, Wiley.
Thomas, W., and R. D. Cook (1989), Assessing Influence on Regression-Coefficients in Generalized Linear-Models, Biometrika, 76(4), 741-749.
Vandewiele, G. L., C. Y. Xu, and N. Larwin (1992), Methodology and Comparative-Study of Monthly Water-Balance Models in Belgium, China and Burma, Journal of Hydrology, 134(1-4), 315-347.
Westra, S., M. Thyer, M. Leonard, D. Kavetski, and M. Lambert (2014), A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resources Research, 50(6), 5090-5113.
Williams, B. J., and W. W. G. Yeh (1983), Parameter-Estimation in Rainfall Runoff Models, Journal of Hydrology, 63(3-4), 373-393.
Woods, R. A., R. B. Grayson, A. W. Western, M. J. Duncan, D. J. Wilson, R. I. Young, R. P. Ibbitt, R. D. Henderson, and T. A. McMahon (2001), Experimental Design and

Initial Results from the Mahurangi River Variability Experiment: MARVEX, Observations And Modeling Of Land Surface Hydrological Processes, pp. 201-213.
Yager, R. M. (2004), Effects of model sensitivity and nonlinearity on nonlinear regression of ground water flow, Ground Water, 42(3), 390-400.

Chapter 3

A generalised approach for identifying influential data in hydrological modelling (Paper 2)

David P Wright, Mark Thyer, Seth Westra, Benjamin Renard, David McInerney
Submitted to Environmental Modelling & Software

Statement of Authorship

Title of Paper	A generalised approach for identifying influential data in hydrological modelling
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Wright, D.P., Thyer, M., Westra, S., Renard, B., McInerney, D. 2017 A generalised approach for identifying influential data in hydrological modelling. Environmental Modelling & Software, (submitted)

Principal Author

Name of Principal Author (Candidate)	David Peter Wright
Contribution to the Paper	Development and implementation of approach, visualisation and interpretation of results, preparation of manuscript and acted as corresponding author.
Overall percentage (%)	85
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 30/03/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mark Thyer
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript
Signature	Date 30/3/2017

Name of Co-Author	Seth Westra
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript
Signature	Date 30/3/2017

Name of Co-Author	Benjamin Renard
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript
Signature	Date 28/03/2017

Name of Co-Author	David McInerney
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript
Signature	Date 30/3/2017

Abstract: Influence diagnostics are able 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 without the computational demand of case-deletion influence diagnostics. This study evaluates the performance of a range of regression-theory influence diagnostics on eleven case studies with a variety of model structures and inference scenarios including: nonlinear model response, heteroscedastic residual errors, data uncertainty and Bayesian priors. Generalised Cook's distance, which uses a generalised leverage formulation, clearly outperformed linear and non-linear leverage formulations to identify the most influential points (Spearman rank correlation: 0.93-1.00) at a fraction of the computational demand of case-deletion (99.6% saving). Computationally efficient generalised Cook's distance has the potential to enable influential data to be identified on a wide variety of hydrological and environmental modelling problems.

3.1 Introduction

Hydrological model calibration is a critical component of model development as parameters generally cannot be easily determined directly from measurements but are instead indirectly inferred by calibrating the hydrological model to observed hydrological responses (e.g. daily streamflow) [Beven, 2011]. Studies have increasingly 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 range of "influence diagnostics" available 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 categorized into two different classes; "case-deletion" influence diagnostics and "regression-theory" influence diagnostics (see Figure 3.1). "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, two different approaches can be used to measure influence. The first approach is to evaluate Cook's distance [Cook, 1977], which is a commonly used

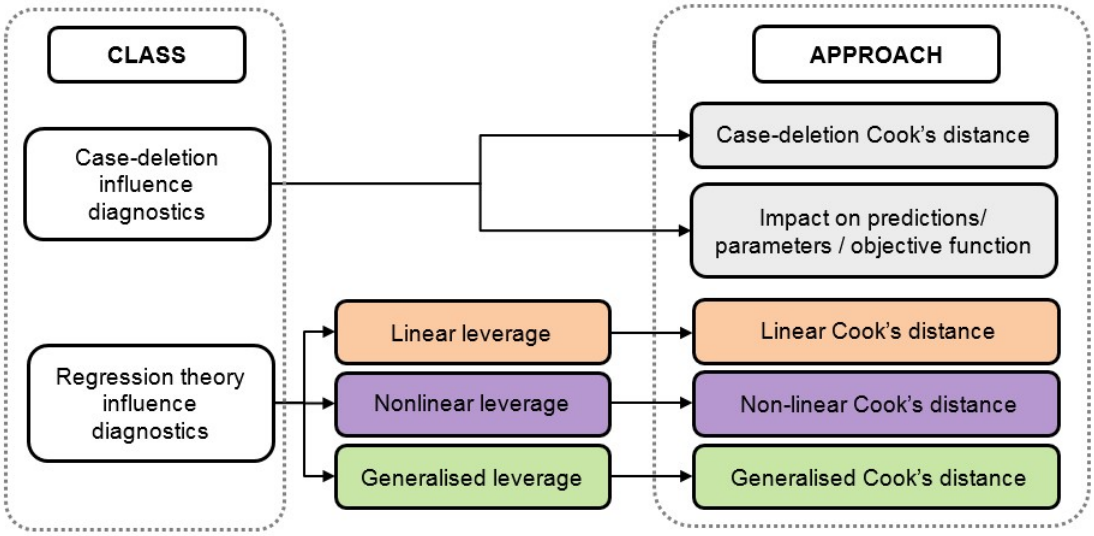

Figure 3.1: Range of available influence diagnostics in the literature. Influence diagnostics are broken up into two classes on the left hand side with the various approaches on the right hand side. The three regression theory approaches are colour coded based on the leverage formulation that they use 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)

measure of influence of data [Cook, 1977] and has been used in a large range of regression problems [Fox and Weisberg, 2011]. 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]; however, for the purposes of this paper we refer to it as "case-deletion" Cook's distance. The second approach to measure influence using the case-deletion class is to quantify the difference between the original and re-calibrated model parameters, model performance (such as objective function displacement) and/or model predictions of interest (see Wright et al, 2015).

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 to hydrological applications, where the hydrological models are non-linear and the assumptions of the residuals errors 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, any efforts will compound the computational demands of the case-deletion recalibrations.

The second class of influence diagnostics are the "regression-theory" influence diagnostics (see Figure 3.1). They are a more efficient alternative to the exact case-deletion influence diagnostics, because they approximate Cook's distance [Cook, 1977] using regression modelling theory to combine the following two components for each observed data point: (1) the leverage, which is used to assess the potential importance of individual observations [Wei et al., 1998], and (2) the standardised residuals. By combining these two components to approximate Cook's distance, "regression-theory" influence diagnostics require no additional re-calibrations and are therefore an attractive, more efficient alternative to the computationally demanding case-deletion influence diagnostics. The drawback is that calculating the leverage can require making assumptions regarding the type of regression model (linear or non-linear) and/or making assumptions about the probabilistic model for residual errors.

For example, linear leverage, which is arguably the most widely used approach to approximate Cook's distance in regression problems [Fox and Weisberg, 2011] assumes that the regression model is linear and that residual errors are Gaussian, homoscedastic and independent. Hence, the approach of calculating Cook's distance using linear leverage (hereafter referred to as "linear Cook's distance") may not be suitable for identifying the influential points in a hydrological modelling context because the hydrological model calibration violates the assumptions of linear leverage, 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 using 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 with Wright et al. [2015] is that the hydrological models were calibrated using a standard least squares residual error model, which is known to perform poorly in a hydrological modelling context [see McInerney et al., 2017], when the residual errors are non-Gaussian and/or heteroscedastic.

To overcome the limitations of the assumptions of linear and non-linear 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 a broad range of 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 more broadly environmental modelling. 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 identify influential data points.

Given the substantial computational advantages of regression-theory influence diagnostics over case-deletion influence diagnostics this study will assess the performance of the different approaches within the class of regression-theory influence diagnostics (i.e. linear Cook's distance, non-linear Cook's distance, and generalised Cook's distance) to reproduce the 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. Nonlinear model response and heteroscedastic residual errors in a series of simple regression models: This didactical modelling scenario provides an opportunity to learn how linear, nonlinear and generalised leverage impact on the identification of influential points using Cook's distance, by comparing linear/nonlinear regression with both homoscedastic and heteroscedastic residual errors.

2. Hydrological model structure including nonlinear model response and storage: This modelling scenario will test the performance of the regression theory influence diagnostics in a typical hydrological modelling context: the calibration of a lumped conceptual rainfall-runoff model assuming heteroscedastic residual errors.

3. Bayesian objective functions that include data uncertainty and prior information: This modelling scenario will test the performance of the regression theory influence

diagnostics in a case study with a more complex objective function, based on using a Bayesian approach that incorporates data uncertainty and prior information.

For all three objectives, the Cook's distance obtained using the linear, non-linear 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. Section 3.2 describes the methodology, in Section 3.3 we introduce the three case studies selected to address the study objectives, and in Section 3.4 we apply the influence diagnostics to these case studies. In Section 3.5 we discuss the advantages and disadvantages of case-deletion and regression-theory influence diagnostics, and the suitability of applying generalised Cook's distance to a broader class of hydrological and environmental models.

3.2 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 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 different approaches to determining the leverage which produces three different estimates of Cook's distance:
   i. Linear Cook's distance, which uses linear leverage,
   ii. Non-linear Cook's distance, which uses non-linear leverage,
   iii. Generalised Cook's distance, which uses generalised leverage.

In this section we first introduce the general modelling framework applied throughout the study, and then define the objective functions, standardised residuals, leverage and Cook's distance influence diagnostics. We finish by describing the metrics that we will use to evaluate the performance of the regression-theory influence diagnostics.

3.2.1 General model framework

We define the general model response as:

$$\mathbf{y} = f(\alpha; \mathbf{X}) + \varepsilon \tag{3.1}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)$ is a vector of n observed responses, $f(.)$ is the model structure, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{p_\alpha})$ is a vector of p_α model parameters,$\mathbf{X}$ is an $n \times k$ matrix of $k$observed inputs, (e.g., precipitation, PET), and ε is a vector of n residual errors. Residuals are further assumed to be realizations from a given probability distribution, parameterized with some unknown parameters $\beta = \left(\beta_1, \beta_2, ..., \beta_{p_\beta}\right)$ (e.g. a centred Gaussian distribution with unknown standard deviation).

3.2.2 Objective functions

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

$$\Phi(\theta; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \rho_i \left(f_i(\alpha; \mathbf{X}), \beta; y_i\right) \tag{3.2}$$

where $\rho_i(.)$ is a function that describes the contribution of the i^{th} data point to $\Phi(\theta; \mathbf{y}, \mathbf{X})$, $f_i(\alpha; \mathbf{X})$ is the i^{th} model prediction, and $\Phi(\theta; \mathbf{y}, \mathbf{X})$ and $f(\alpha; \mathbf{X})$ are assumed to be twice differentiable with respect to $\theta$and $\mathbf{y}$. With model parameters α and the residual error model parameters β the whole set of p parameters to be optimised is $\theta = \{\alpha, \beta\}$. We will denote $\widehat{\theta}$ as the model parameters which maximise Φ in equation 3.2), and $\widehat{\mathbf{y}}$ as the predicted response associated with $\widehat{\theta}$, i.e. $\widehat{\mathbf{y}} = f(\hat{\alpha}; \mathbf{X})$.

The generalised form in equation (3.2) can be adapted to a number of well-known objective functions in hydrological modelling as outline below.

3.2.2.1. 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:

$$\Phi(\theta; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \log(p_N(y_i - f_i(\alpha; \mathbf{X}) | 0, \sigma^2)) \tag{3.3}$$

where $p_N(x|\mu,\sigma^2)$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 $\beta = \{\sigma\}$.

3.2.2.2. Weighted least squares

Due to heteroscedasticity in hydrological residual errors [McInerney et al., 2017; Thyer et al., 2009] it is common to replace the constant standard deviation σ in equation (3.3) with a standard deviations σ that varies in time. 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

$$\sigma = \beta_1 \widehat{\mathbf{y}} + \beta_2 \tag{3.4}$$

The objective function becomes:

$$\Phi(\theta;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \log(p_N(y_i - f_i(\alpha;\mathbf{X})\,|0,{\sigma_i}^2)) \tag{3.5}$$

As the parameters describing the non-constant standard deviation are unknown they will need to be estimated (i.e. $\beta = \{\beta_1,\beta_2\}$).

3.2.2.3. Weighted least squares with data uncertainty

When uncertainty estimates are available for observed responses, we have the option to take them into account in the WLS method. To implement the WLS method with discharge uncertainty in the WLS objective function (5) 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 $\sigma_r = \beta_1 \widehat{\mathbf{y}} + \beta_2$ And the "measurement errors" described using known standard deviations σ_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 $\sigma = \sqrt{{\sigma_r}^2 + {\sigma_Y}^2}$. Hence the σ_i in equation (3.5) becomes:

$$\sigma_i = \beta_1 \widehat{y_i} + \beta_2 + \sigma_y^i \tag{3.6}$$

where σ_y^i is the standard deviation of the measurement errors at time step i.

3.2.2.4. Weighted least squares with priors

When prior information on model parameters is available we can use Bayes' equation to yield the posterior probability distribution of the hydrological and residual error model parameters given inputs $\mathbf{X}$ and outputs $\mathbf{y}$, as follows:

$$\underbrace{\mathrm{p}(\theta|\mathbf{X},\mathbf{y})}_{posterior} \propto \underbrace{\mathrm{p}(\mathbf{y}|\theta,\mathbf{X})}_{likelihood}\underbrace{\mathrm{p}(\theta)}_{prior} \tag{3.7}$$

where $\mathrm{p}(\theta|\mathbf{X},\mathbf{y})$is the posterior probability of parameter θ given $\mathbf{X}$ and $\mathbf{y}$, $\mathrm{p}(\theta)$ is the joint prior probability density of hydrological and residual error model parameters, and $\mathrm{p}(\mathbf{y}|\theta,\mathbf{X})$ is the likelihood of $\mathbf{y}$ given θ and $\mathbf{X}$. Taking the logarithm of equation (3.7) we obtain:

$$\log(\mathrm{p}(\theta|\mathbf{X},\mathbf{y})) = \log(\mathrm{p}(\mathbf{y}|\theta,\mathbf{X})) + \log(\mathrm{p}(\theta)) + c \tag{3.8}$$

where$c$ is a constant, so that we can formulate the objective function as:

$$\begin{aligned}\Phi(\theta;\mathbf{y},\mathbf{X}) &= \log(\mathrm{p}(\mathbf{y}|\theta,\mathbf{X})) + \log(\mathrm{p}(\theta)) \\ &= \sum_{i=1}^{n} \log(\mathrm{p}(y_i|\theta,\mathbf{X})) + \log(\mathrm{p}(\theta)) \\ &= \sum_{i=1}^{n} \left(\log(\mathrm{p}(y_i|\theta,\mathbf{X})) + \tfrac{1}{n}\log(\mathrm{p}(\theta))\right)\end{aligned} \tag{3.9}$$

Assuming the residual errors are heteroscedastic with σ given by equation (3.4) and independent priors, we obtain the following objective function:

$$\Phi(\theta;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n}\left\{\log(p_N(y_i - f_i(\alpha;\mathbf{X})|0,{\sigma_i}^2)) + \frac{1}{n}\sum_{j=1}^{p}\log\left(\mathrm{p}(\theta_j)\right)\right\} \tag{3.10}$$

where the contributions to the objective function from the prior are split evenly across the n points in the calibration data. Note that this corresponds to the general objective function in equation (3.2) with

$$\rho_i(f_i(\alpha;\mathbf{X}),\beta;y_i) = \left\{\log(p_N(y_i - f_i(\alpha;\mathbf{X})|0,{\sigma_i}^2)) + \tfrac{1}{n}\sum_{j=1}^{p}\log(\mathrm{p}\left(\theta_j\right))\right\}.$$

3.2.2.5. Weighted least squares with data uncertainty and priors

As explained in Section 3.2.2.3, data uncertainty can readily be included in the objective function (10) by simply using $\sigma = \sqrt{{\sigma_r}^2 + {\sigma_Y}^2}$, where $\sigma_r = \beta_1\mathbf{y} + \beta_2$ and σ_Y are known values representing the measurement uncertainty in observed responses.

3.2.3 Standardised residuals

The standardised residuals, v, required to estimate the regression-theory influence diagnostics described in Section 3.2.5.2, are obtained by dividing the raw residuals ε by their calibrated standard deviations:

$$v = \frac{\mathbf{y} - \widehat{\mathbf{y}}}{\widehat{\sigma}} \tag{3.11}$$

For the SLS objective function (3), $\widehat{\sigma}$ is constant across the dataset. For all other WLS objective functions, $\widehat{\sigma}$ differs at each data point (i.e. following equation (3.4) for WLS and with the addition of data uncertainty following equation (3.6)).

3.2.4 Leverage

Leverage is a key component of regression-theory influence diagnostics and is measured by the rate of the change of the i^{th} predicted value $\widehat{y}_i$ with respect to the j^{th} observed value y_j [Wei et al., 1998]:

$$L_{ij} = \partial \hat{y}_i / \partial y_j \tag{3.12}$$

Or in matrix notation:

$$\mathbf{L} = \frac{\partial \widehat{\mathbf{y}}}{\partial \mathbf{y}^T} \tag{3.13}$$

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 will be used for calculating regression theory Cook's distance in Section 3.2.5.2.

3.2.4.1. Linear leverage

In linear regression, leverage can be calculated from the hat matrix [Fox and Weisberg, 2011]

$$\mathbf{L} = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}} \tag{3.14}$$

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 SLS residual errors, regression-based Cook's distance is equivalent to case-deletion Cook's distance [see Cook, 1977].

3.2.4.2. Nonlinear leverage

In nonlinear regression, 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 from the diagonal elements of $\mathbf{L}(\widehat{\alpha})$ [St. Laurent and Cook, 1992; 1993; Wei et al., 1998; Wright et al., 2015]:

$$\mathbf{L}(\widehat{\alpha}) = \frac{\partial f(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}} \left(\left(\frac{\partial f(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}} \right)^T \frac{\partial f(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}} - \sum_{i=1}^{n} \left((y_i - \widehat{y_i}) \frac{\partial^2 f_i(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}^2} \right) \right)^{-1} \left(\frac{\partial f(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}} \right)^T \tag{3.15}$$

where $\frac{\partial f(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}}$ is the $n \times p_\alpha$ Jacobian matrix with i^{th} row $\frac{\partial f_i(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}}$, and $\frac{\partial^2 f_i(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\alpha}^2}$is the $p_\alpha \times p_\alpha$ 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 finite 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].

3.2.4.3. Generalised leverage

In order for leverage to be applied to a general class of regression models, including a larger range of objective functions, the generalised leverage takes into account the curvature of the objective function about the whole set of calibrated parameters $\widehat{\theta}$. In this case, leverage is equal to the diagonal elements of $\mathbf{L}(\widehat{\theta})$ given by:

$$\mathbf{L}(\widehat{\theta}) = \frac{\partial f(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\theta}} \left(-\frac{\partial^2 \Phi(\widehat{\theta};\mathbf{y},\mathbf{X})}{\partial \widehat{\theta}^2} \right)^{-1} \frac{\partial^2 \Phi(\widehat{\theta};\mathbf{y},\mathbf{X})}{\partial \widehat{\theta} \partial \mathbf{y}^T} \tag{3.16}$$

where $\frac{\partial f(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\theta}}$ is the $n \times p$ Jacobian matrix with i^{th} row $\frac{\partial f_i(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\theta}}$ (note that $\frac{\partial f_i(\widehat{\alpha};\mathbf{X})}{\partial \widehat{\beta}} = 0$), $\frac{\partial^2 \Phi(\widehat{\theta};\mathbf{y},\mathbf{X})}{\partial \widehat{\theta}^2}$ is a $p \times p$ Hessian matrix and $\frac{\partial^2 \Phi(\widehat{\theta};\mathbf{y},\mathbf{X})}{\partial \widehat{\theta} \partial \mathbf{y}^T}$ is a $p \times n$ matrix. Generalised leverage can be applied to any objective function that takes the general form in equation (3.2). 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].

3.2.5 Influence diagnostics

3.2.5.1. Case-deletion Cook's distance

Case-deletion based Cook's distance measures influence by comparing model predictions $\widehat{\mathbf{y}}$ based on using all of the calibration data and model predictions $\widehat{\mathbf{y}}^{(-i)}$ with the i^{th}point

masked from calibration data. For a given point, case-deletion based Cook's distance is given by:

$$CD_i = \sum_{j=1}^{n} \frac{\left(\hat{y}_j - \hat{y}_j^{(-i)}\right)^2}{p \times \hat{\sigma}_j^2} \tag{3.17}$$

where $\widehat{\sigma}_j$ is the calibrated standard deviation for the j^{th}data point, estimated based using all calibration data (i.e. $\widehat{\mathbf{y}}$). For the SLS objective function in equation (3.3), $\widehat{\sigma}$ is constant across the dataset.

3.2.5.2. Regression-theory influence diagnostics

Regression-theory influence diagnostics calculate Cook's distance [Cook and Weisberg, 1982; Fox and Weisberg, 2011] by combining the standardised residual of the i^{th}point (v_i) with the leverage of i^{th} observation on the i^{th} prediction (L_{ii}) to give

$$CD_i = \frac{{v_i}^2}{p} \frac{L_{ii}}{1 - L_{ii}} \tag{3.18}$$

We implement equation (3.18) with the three forms of leverage; i.e. linear leverage in equation (3.14) to give linear Cook's distance, nonlinear leverage in equation (3.15) to given nonlinear Cook's distance and generalised leverage in equation (3.16) to give generalised Cook's distance.

3.2.6 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 provide a measure of the proportion of the variance in the regression-based variable that is accounted for by the case-deletion based 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.

3.3 Case studies

In order to evaluate the performance of regression-theory influence diagnostics we apply them to three case studies (Table 3.1). To address the first research objective, four synthetic regression models, A_{1-4}, were selected to test the performance with regression model linearity/non-linearity and homoscedastic/heteroscedastic residual error models. The second research objective was addressed by testing the performance with hydrological models, B_{1-2}, with nonlinear hydrological response, model storage, and heteroscedastic residual errors. Finally, the third objective was addressed by testing the performance with four different rating curve models, C_{1-4}, with and without data uncertainty and with and without prior knowledge specified using a Bayesian inference approach.

In all cases the objective functions were optimized 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.

3.3.1 Case study 1: Regression models with linear/nonlinearity and homoscedastic/heteroscedastic residual errors

We start with four studies that range in complexity from a simple linear model with homoscedastic residual errors to a nonlinear power model with heteroscedastic residual errors. The regression models with synthetic data A_{1-4} (Table 3.1) were selected to highlight the role of model structure and residual error model on the influence results; A_1 has a linear model response and standard least squares (SLS) residual error model; A_2 introduces the heteroscedastic weighted least squares (WLS) residual error model; A_3 and A_4 have a nonlinear model response and SLS and WLS residual error, respectively.

Table 3.1: Details of the case studies.

Case study	Input	Model	Residual error model	"Observed" output Y
A_1	$X \sim \mathrm{U}(1,200)$	$f(\mathbf{X},\alpha_1,\alpha_2) = \alpha_1\mathbf{X}+\alpha_2$	$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1$	$f(\mathbf{X},10,500)+\varepsilon(100)$
A_2		$f(\mathbf{X},\alpha_1,\alpha_2) = \alpha_1\mathbf{X}+\alpha_2$	$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1\mathbf{y}+\beta_2$	$f(\mathbf{X},10,500)+\varepsilon(0.2,10)$
A_3		$f(\mathbf{X},\alpha_1,\alpha_2,\alpha_3) = \alpha_1+\alpha_2\mathbf{X}^{\alpha_3}$	$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1$	$f(\mathbf{X},500,0.1,2.3)+\varepsilon(100)$
A_4		$f(\mathbf{X},\alpha_1,\alpha_2,\alpha_3) = \alpha_1+\alpha_2\mathbf{X}^{\alpha_3}$	$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1\mathbf{y}+\beta_2$	$f(\mathbf{X},500,0.1,2.3)+\varepsilon(0.1,0.5)$
B_1	Observed	$\mathrm{GR4J}(\mathrm{P},\mathrm{PET},\alpha)$	$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1\mathbf{y}+\beta_2$	$\mathrm{GR4J}(\mathrm{P},\mathrm{PET},\alpha=\{2200,1.15,87,0.55\})+\varepsilon(0.1,0.5)$
B_2		$\mathrm{GR4J}(\mathrm{P},\mathrm{PET},\alpha)$	$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1\mathbf{y}+\beta_2$	Observed
C_1	Observed	$f(X_i,\alpha) = \begin{cases} \alpha_1(X_i-\alpha_2)^{\alpha_3}, X_i < \alpha_4 \\ \alpha_5(X_i-b_2)^{\alpha_6}, X_i \geq \alpha_4 \end{cases}$	$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1\mathbf{y}+\beta_2$	Observed
C_2			$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \sqrt{{\sigma_r}^2+{\sigma_Y}^2}, \sigma_r = \beta_1\mathbf{y}+\beta_2$	
C_3			$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \beta_1\mathbf{y}+\beta_2$	
C_4			$\varepsilon(\sigma) \sim \mathrm{N}(0,\sigma^2), \sigma = \sqrt{{\sigma_r}^2+{\sigma_Y}^2}, \sigma_r = \beta_1\mathbf{y}+\beta_2$	

3.3.2 Case study 2: Daily hydrological model with synthetic and observed streamflow and heteroscedasticity

The next case study was chosen to test of 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 widespread use [e.g. Evin et al., 2014; Le Moine et al., 2007; 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 $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$, 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 GR4J 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 use three years of calibration data, from 1974 to 1976.

We explore two alternative modelling scenarios B_{1-2} (Table 3.1) that correspond to synthetic streamflow data and real observed streamflow data, respectively. The first model B_1 uses the observed rainfall and PET from the French Broad River but has synthetic heteroscedastic residual errors generated using known GR4J parameters obtained from the B_2 calibration with the French Broad River catchment. The second hydrological model B_2 also uses observed rainfall and PET from the French Broad River catchment, but is calibrated to observed streamflow data.

3.3.3 Case study 3: Rating curve model incorporating discharge uncertainty and parameter priors

The final case-study uses a rating curve model, with increasing complexity in the objective function that investigates the impact of discharge 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 hydro-

metric 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:

$$f(\alpha, \mathrm{X}_i) = \begin{cases} a_1(\mathrm{X}_i - b_1)^{c_1}, \text{for} \mathrm{X}_i < k \\ a_2(\mathrm{X}_i - b_2)^{c_2}, \text{for} \mathrm{X}_i \geq k \end{cases} \tag{3.19}$$

Here $\mathbf{X}$ is stage and $\alpha = \{a_1, b_1, c_1, k, a_2, c_2\}$ are the rating curve model parameters. As the rating curve is continuous at the knot k, the parameter b_2 is inferred 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 $b_2 = k - ((a_1/a_2)(k-b_1)^{c_1})^{1/c_2}$. Petersen-Øverleir [2004] suggest a heteroscedastic residual error model to take into account the usually observed heteroscedasticity of rating curve errors and so we use the WLS objective function described in Section 3.2.2.2. We apply four calibration schemes across C_{1-4}, as follows: 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 3.2. Perusal of Table 3.2, shows that the prior standard deviation is smallest for the exponent parameters [c_1 and c_2 in equation (3.18)), 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 β there is no prior knowledge and so an uninformative uniform distribution is applied.

Table 3.2: 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)

3.4 Performance evaluation of regression-theory influence diagnostics

We apply regression-theory influence diagnostics with linear, nonlinear and generalised Cook's distance in Sections 3.4.1-3.4.3. In Section 3.4.4 we summarise the performance of the regression-theory influence diagnostic across the case studies, and we finish in Section 3.4.5 with an analysis of the computation times of both the regression-based and case-deletion based influence diagnostics.

3.4.1 Case study 1: Regression models with increasing model nonlinearity and residual error complexity

We start by evaluating the performance of regression-based Cook's distance under a series of models with synthetic data that have varying degrees of nonlinear model response and heteroscedastic residual errors. The results from applying the Cook's distance influence diagnostics to the synthetic regression case studies A_{1-4} (Table 3.1) are presented in Figure 3.2. The first row for each synthetic model compares "observed data" and model predictions, followed by the standardised residuals in the second row and the three formulations of leverage in the third row. The final row shows the Cook's distance influence metrics for the four different approaches, case-deletion, linear, nonlinear and generalised Cook's distance respectively.

We start by comparing the points that are identified as influential by linear Cook's distance with those identified by case-deletion Cook's distance across A_{1-4}. Figure 3.2 (fourth row) shows that linear Cook's distance performs well for linear SLS regression model (A_1), but its ability to identify the same influential points as case-deletion Cook's distance reduces for the heteroscedastic errors (A_2 linear WLS) and nonlinear models (A_3 nonlinear SLS, A_4, nonlinear WLS). This is confirmed in Figure 3.3 where for linear Cook's distance we see an almost perfect correlation with case-deletion Cook's distance across all performance metrics (A_1, first column). For the linear WLS case A_2 (second column) for linear Cook's distance (Figure 3.3, top row, 2nd column) we see a small reduction in performance (with correlations 0.98 and 0.70 for Sp. and r^2 respectfully). However, for the top ten influential points (Figure 3.3, bottom row, 2nd column) we see a larger reduction in performance with correlations far lower (0.65 and 0.28 for $Sp._{10}$ and r^2_{10} respectfully). For the nonlinear SLS case (A_3, third column) a similar trend as linear WLS is seen, with a small reduction in performance for the entire dataset, but a larger reduction for the top ten influential points. Finally in the nonlinear WLS case A_4

(fourth column) we also see a reduction in performance, in particular a systematic under prediction of the case-deletion Cook's distance by linear Cook's distance.

Inspecting the leverage values for linear leverage (Figure 3.2, third row) across the four models, A_{1-4} shows that irrespective of model nonlinearity and/or homoscedastic/heteroscedastic residuals errors the linear leverage is smooth and parabolic in shape, with a minima at the mean of $\mathbf{X}$ and higher leverage values at extreme values of $\mathbf{X}$. This is because the linear leverage is independent of $\mathbf{y}$ and calculated from the input $\mathbf{X}$ only, and as the four models A_{1-4} have identical $\mathbf{X}$, the linear leverage is also identical across all four models. This means, that when using linear leverage the leverage of a single point is not impacted by the type of model response or the type of residual error which results in poor performance in identifying the influential points when the regression model becomes nonlinear and/or the residual errors become heteroscedastic.

Next, we evaluate the performance for nonlinear Cook's distance (purple points) across the four models A_{1-4} against case-deletion Cook's distance. Applying nonlinear Cook's distance we see the same high performance as linear leverage in the linear SLS case (A_1), and also in the nonlinear SLS case (A_3) (Figure 3.3, bottom row, first and third columns). However, when a WLS error model is used (linear WLS, A_2 and nonlinear WLS) this lowers the performance of nonlinear Cook's distance. For both linear WLS (A_2) and nonlinear WLS (A_3) this degradation in performance is most obvious for the top ten influential points with large decrease in performance metrics ($Sp._{10}$ =0.65 and r^2_{10}=0.28 for A_2, $Sp._{10}$ =0.65 and r^2_{10}=0.95 for A_4). This can be explained by inspecting the difference in the leverage values for nonlinear leverage across the four models (Figure 3.2, third row). For the linear SLS model A_1 (first column) and the linear WLS model A_2 (second column) we see that nonlinear leverage is exactly equivalent to linear leverage as expected, since nonlinear leverage simplifies to linear leverage under a linear regression model. In contrast, in the nonlinear SLS model A_3 (third column) and nonlinear WLS model A_4 (fourth column), nonlinear leverage differs from linear leverage but remains a smooth curve across the values of $\mathbf{X}$. The nonlinear model structure results in higher nonlinear leverage in the upper range of $\mathbf{X}$ with a slight increase in the midrange of $\mathbf{X}$ for A_3. This comparison shows that nonlinear leverage is only impacted by the linear/nonlinear regression model response, but not impacted by the differences (homoscedastic versus heteroscedastic) in residual errors, hence this results in poorer performance in identifying influential points using nonlinear Cook's distance when the residual errors are heteroscedastic.

Finally we evaluate generalised Cook's distance (green points) across the four models. Applying generalised Cook's distance, we see that the majority of identified influential points are similar to case-deletion Cook's distance (Figure 3.2, fourth row). Similarly,

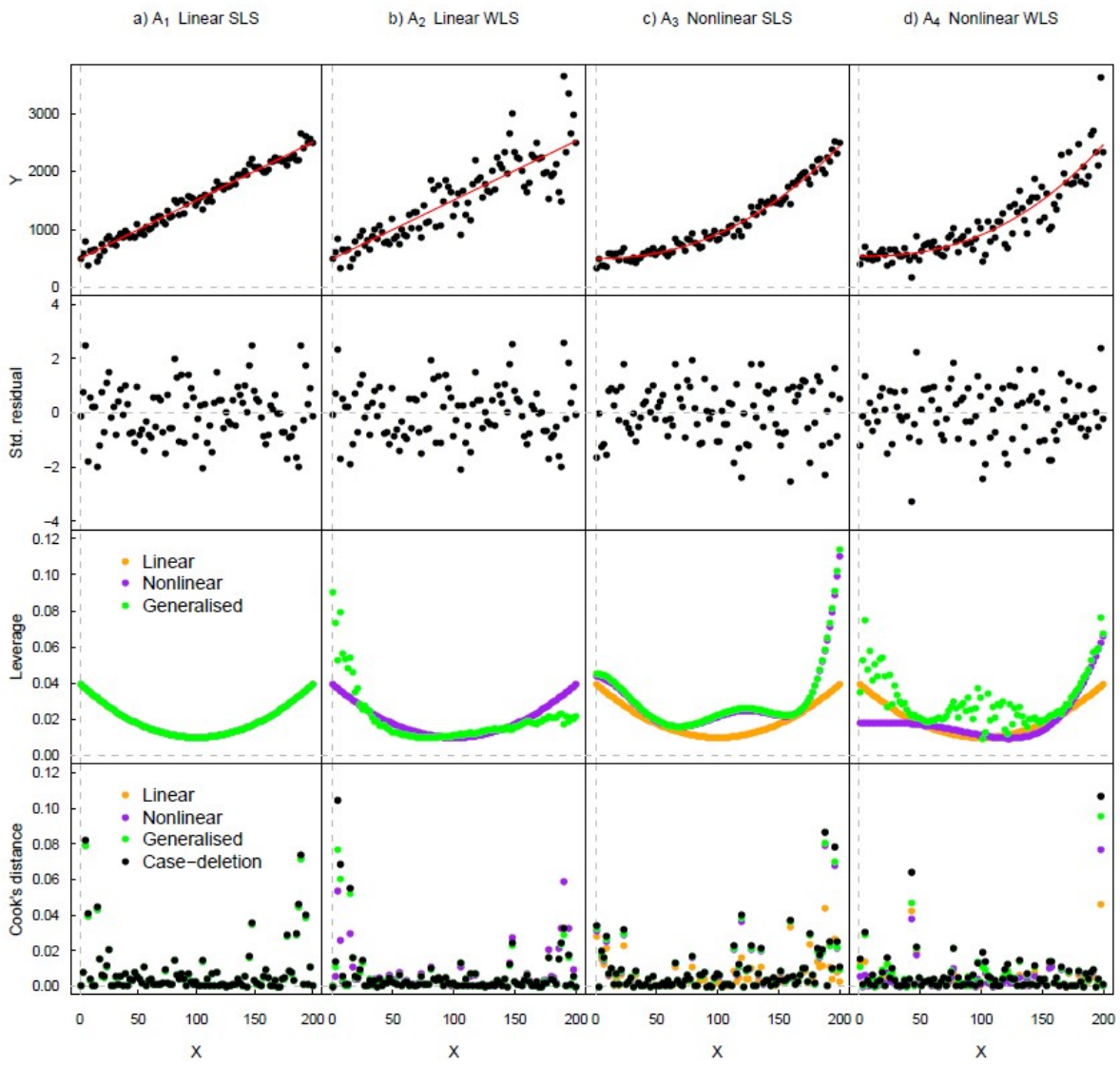

Figure 3.2: Synthetic regression model case-study results. Observed data (black), and predicted model (red) in the top row, followed by standardised residuals in the second row. Leverage is shown in the third row with: linear leverage, nonlinear leverage, generalised leverage. In the case of A_1 the 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 final row shows regression-based Cook's distance with linear, nonlinear and generalised leverage, and case-deletion Cook's distance. Note that in the third row for A_1 linear leverage and nonlinear leverage are hidden by the generalised leverage, for A_2 linear leverage is hidden by nonlinear leverage, for A_3 nonlinear leverage is partially hidden by generalised leverage. Additionally, in the fourth row, case-deletion is superimposed over the linear, nonlinear and generalised Cook's distance obscuring the points that match closely.

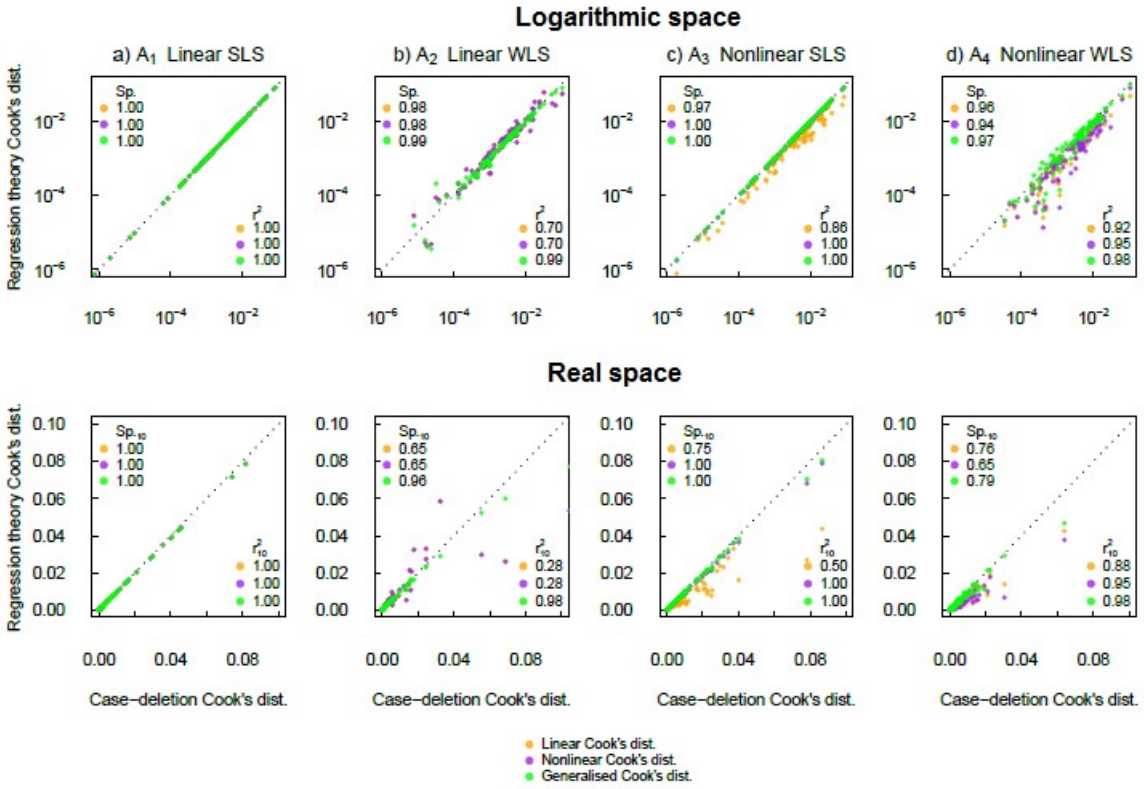

Figure 3.3: Synthetic regression model case-study comparison of case-deletion Cook's distance and regression-theory influence diagnostics. In the first row we compare the performance in logarithmic space and use the Sp. and r^2 to highlight performance across the whole dataset. In the second row we compare the performance in real space and use the $Sp._{10}$ and r^2_{10} to compare the subset of the ten most influential data points.

Figure 3.3 shows very high performance across all four cases where all metrics are either perfect (1.00) or higher than the linear leverage and nonlinear leverage performance.

We investigate the reasons why generalised Cook's distance has the highest performance by comparing the leverage values provided by the generalised leverage (Figure 3.2, third row). In the linear SLS model A_1 (first column) we see that generalised leverage is exactly equivalent to linear leverage and nonlinear leverage, as expected, since generalised leverage and non-linear leverage simplify to linear leverage under a homoscedastic residual error and linear model. In the linear WLS model A_2 (second column) generalised leverage is non-smooth due to heteroscedastic WLS residual error in the synthetic $\mathbf{y}$. Applying the WLS log likelihood increases generalised leverage for low $\mathbf{X}$ and decreases generalised leverage compared to linear and nonlinear leverage for high $\mathbf{X}$ indicating the increased potential influence of the lower range $\mathbf{X}$. In the nonlinear SLS model A_3 (third column) generalised leverage is similar to nonlinear leverage, which shows that under the SLS hypothesis (i.e. homoscedastic errors), both leverage formulations treat model nonlinearity in the same way. Finally, in the nonlinear WLS model A_4 (fourth column) generalised leverage can account for the WLS heteroscedasticity in $\mathbf{y}$ and so the generalised leverage is non-smooth as well as accounting for the change in shape from the model non-linearity observed in the nonlinear leverage.

Overall, this indicates that for the four simplified regression models considered, generalised Cook's distance provides a very close approximation of case-deletion Cook's distance, and a significant improvement in identifying the influential points compared to the other regression-theory influence diagnostics.

3.4.2 Case study 2: Daily hydrological model with synthetic and observed streamflow and heteroscedastic residual errors

Next we evaluate the performance of regression-theory influence diagnostics in a typical hydrological modelling context where the model has nonlinear response, storage and heteroscedastic errors. The results from the hydrological model with both synthetic and real observed catchment data (models B_1 and B_2, respectively; see Table 3.1) are presented in Figure 3.4. In Figure 3.4 we display a subset of the hydrological data in the form of three 100 day hydrograph windows to aid visual interpretation of the influence diagnostics. Across the two models leverage is now plotted against the daily time step rather than input value $\mathbf{X}$, so that the parabolic nature of linear leverage when plotted against $\mathbf{X}$ is not evident as in Figure 3.2. Examining the standardised residuals (second row) we see a large difference between the synthetic data in B_1 (Figure 3.4a) and the real hydrological

data in B_2 (Figure 3.4b). Interestingly, this difference does not always lead to a large difference in the magnitude of influence in the fourth row. This indicates the importance of the leverage (third row) in the regression-based Cook's distance. We see that the most influential data typically corresponds to those points that have both large standardised residuals and high leverage values.

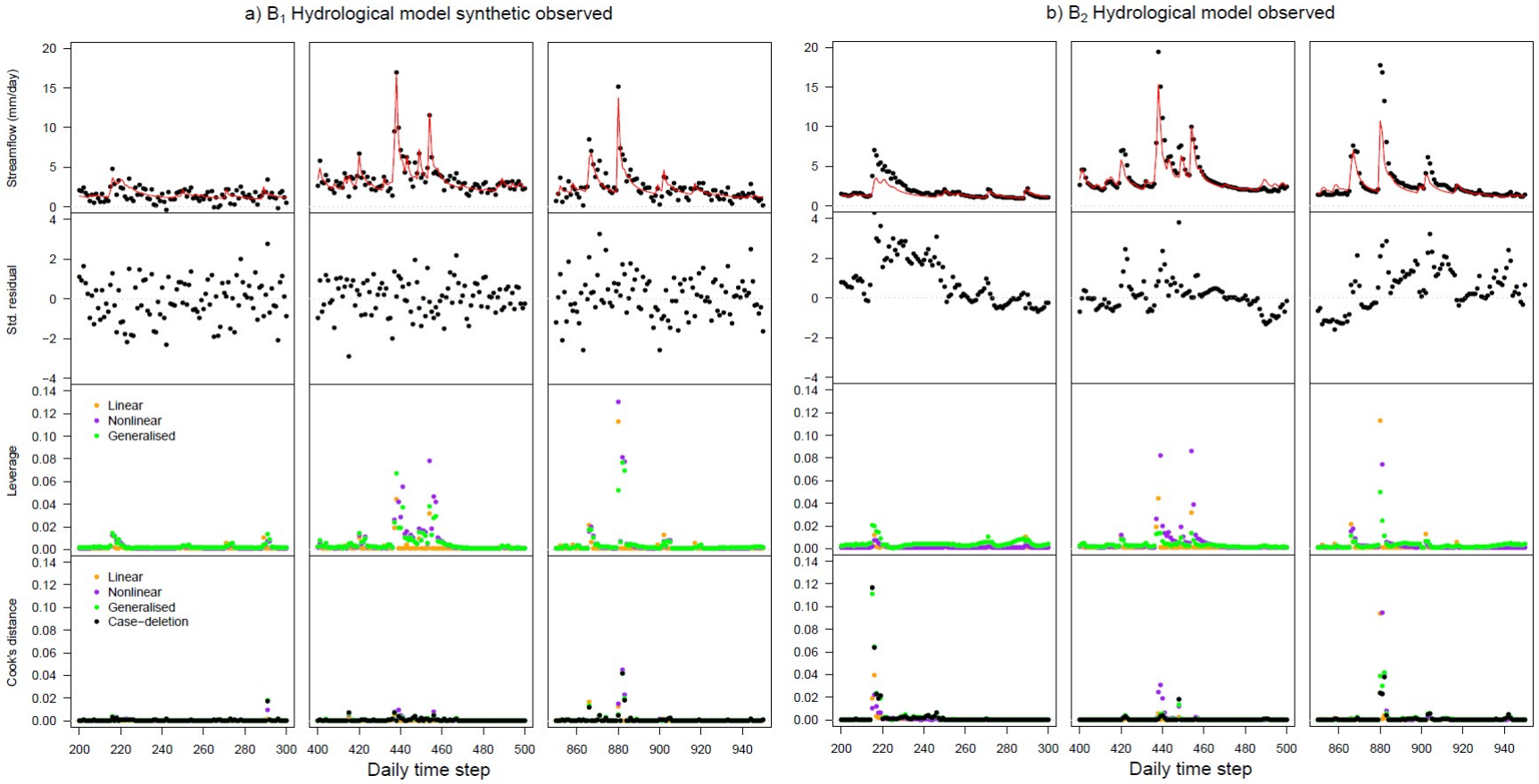

Figure 3.4: Representative hydrographs from the hydrological model case-study. Observed streamflow (black), and predicted streamflow (red) in the top row, followed by standardised residuals in the second row. Leverage is shown in the third row with: linear leverage, nonlinear leverage, generalised leverage. The final row shows regression-based Cook's distance with linear, nonlinear and generalised leverage, and case-deletion Cook's distance.

We next test the performance of the three regression-theory influence diagnostics to reproduce case-deletion based Cook's distance in Figure 3.5. Starting with linear Cook's distance we see a spread about the 1:1 line in both B_1 and B_2, indicating that linear Cook's distance captures neither the ranking nor the value of the influential data points identified by case-deletion Cook's distance. This low performance is reflected by the 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, respectively).

Next we examine nonlinear Cook's distance across the two cases. In terms of the performance metrics for the synthetic hydrological case (B_1) there are improvements in some metrics, compared with linear leverage, (e.g. $Sp._{10}$ improves from -0.30 to 0.95) but for the real hydrological case study (B_2) the performance is still relatively poor (e.g. $Sp._{10}$ is 0.19 and r^2=0.05).

Finally we examine generalised Cook's distance across the two cases (Figure 3.5). This provides by far the best performance of all three regression-theory influence diagnostics, with tight spread about the 1:1 line in both B_1 and B_2, very high performance metrics (ranging from 0.93-1.00 for all metrics for both B_1 and B_2). This indicates generalised Cook's distance is successfully able to capture the impact on leverage of the nonlinear and storage components of the hydrological model response and heteroscedastic errors.

3.4.3 Case study 3: Rating curve model incorporating discharge uncertainty and parameter priors

Finally, we explore the ability of regression-theory influence diagnostics to identify influential points when using objective functions that account for data uncertainty and prior parameter information in Bayesian inference. We start by examining the magnitude of the case-deletion Cook's distance across the four rating curve cases (C_{1-4}) in Figure 3.6. In each panel the observed data are shown with uncertainties (models C_2 and C_4 only) in a scatterplot, with the fitted model and 38 case-deletion fitted models, with the point size proportional to the magnitude of the case-deletion Cook's distance.

Comparing across the four rating curve models, the influential data are typically the extreme (both high and low) stage-discharge observed data. Accounting for discharge uncertainty in C_2 (Figure 3.6b) slightly reduces both the magnitude of the most influential data, and the variability in the case-deletion rating curves. Accounting for priors in C_3 (Figure 3.6c) leads to a larger reduction in the influential data, significantly reducing both the magnitude of the most influential data and the variability in the case-deletion rating curves. Finally the combined effect of accounting for discharge uncertainty and priors

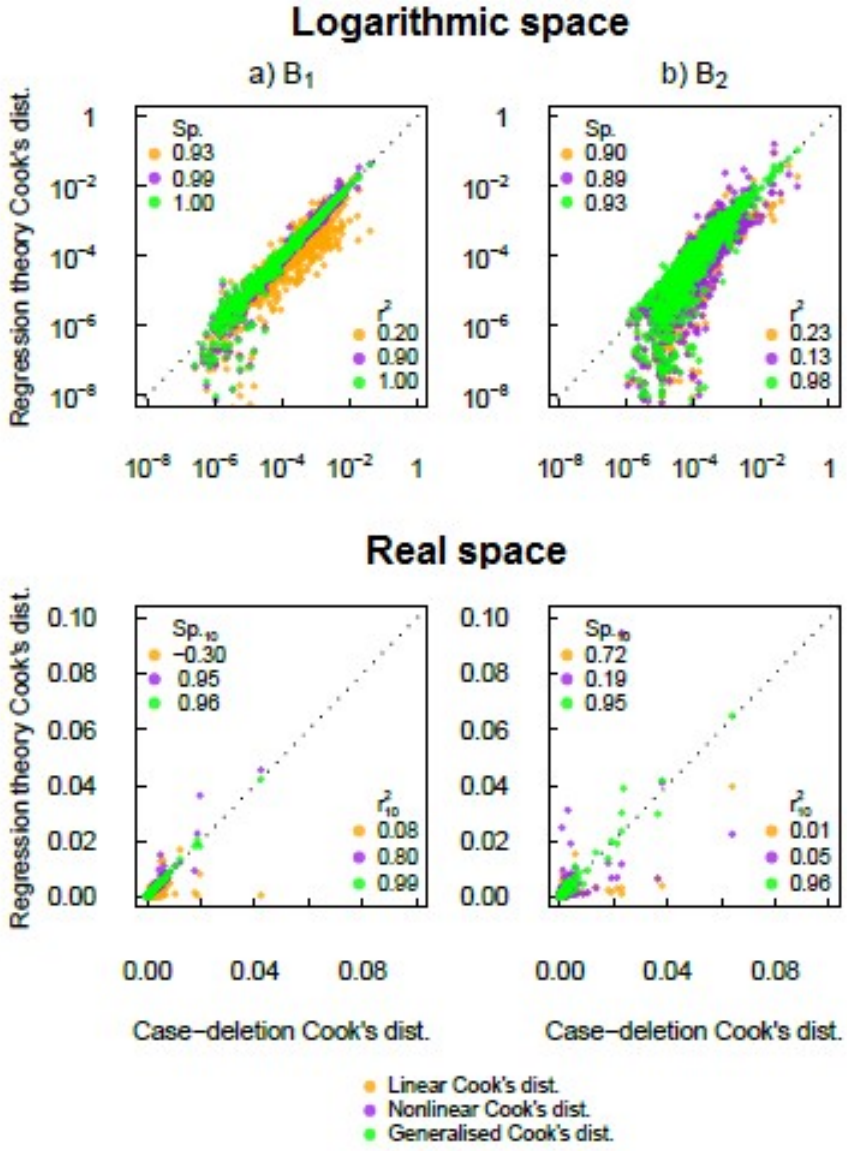

Figure 3.5: Hydrological model case-study comparison of case-deletion and regression-theory influence diagnostics. In the first row we compare the performance in logarithmic space and use the Sp. and r^2 to highlight performance across the whole dataset. In the second row we compare the performance in real space and use the $Sp._{10}$ and r^2_{10} to compare the subset of the ten most influential data points.

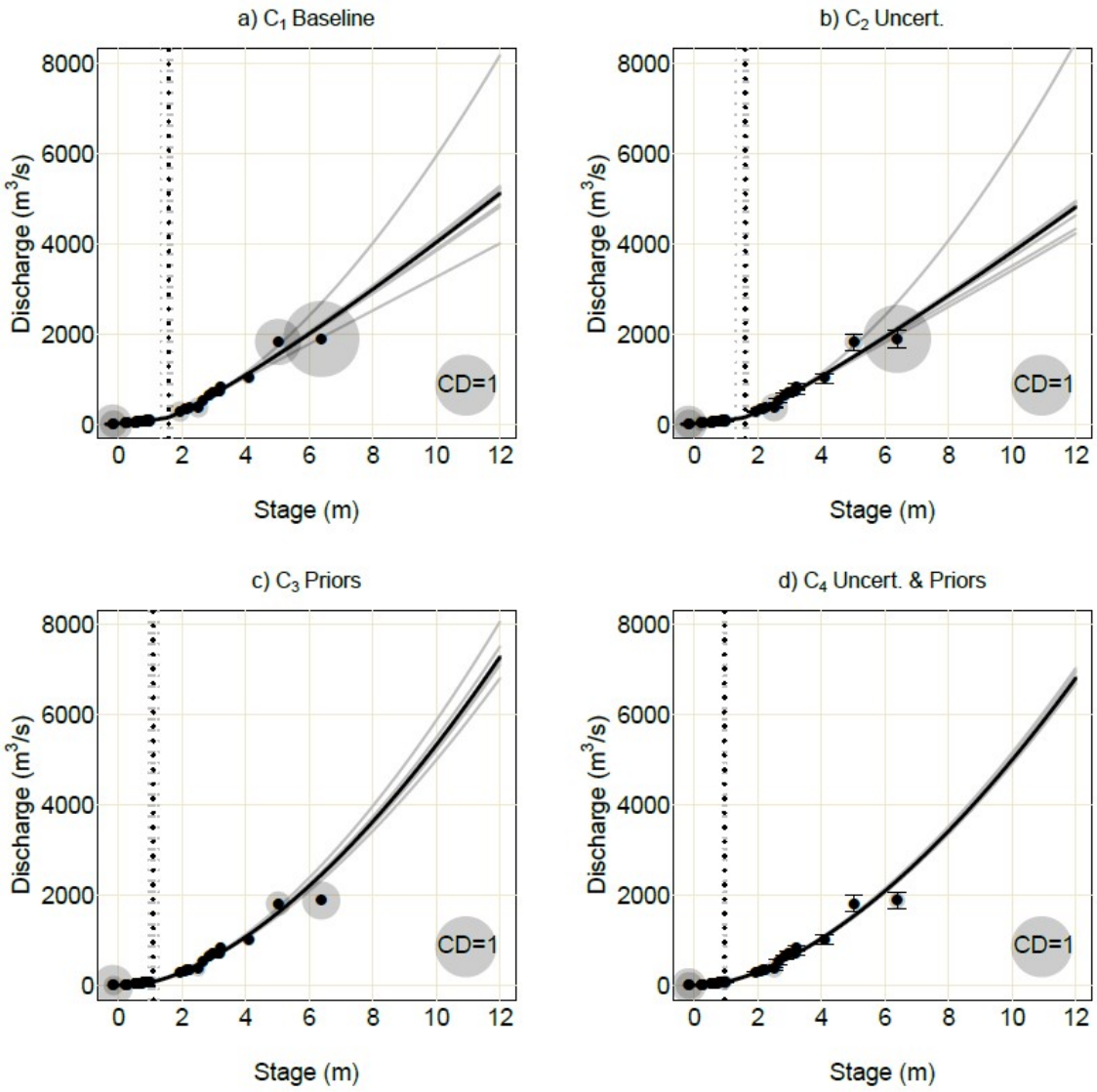

Figure 3.6: Stage-discharge rating curves for the Ardèche River at Sauze. The four rating-curves presented are a) baseline rating curve without accounting for discharge uncertainty and priors, b) Rating curve with discharge uncertainty, c) Rating curve with parameter priors, d) Rating curve with both 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. Magnitude of case-deletion Cook's distance is shown by the grey bubble size.

in C_4 (Figure 3.6d) has an even larger reduction in the influential data, both significantly reducing the magnitude of influential data and leading to a tight spread in the case-deletion rating curves.

Next we investigate the impact of the leverage formulation on the regression-theory influence diagnostics for the four cases. Analogous to the results in Figure 3.2 and 3.4, in Figure 3.7 we first present the scatterplot of the stage and discharge measurements with the fitted rating curve, followed by the standardised residuals, the leverage, and the Cook's distance. Examining the standardised residuals (second row) we see only slight variability across the four rating curve models, so that it is the different leverage values (third row) that have the greatest effect on the differences in Cook's distance.

Starting with linear leverage we see the expected parabolic shape for the leverage values as a function of $\mathbf{X}$. Turning our attention to nonlinear leverage we see different nonlinear leverage across the four cases C_{1-4} as we have different objective functions and therefore different calibrated model parameters. Consistently the highest magnitude leverage is the highest stage-discharge value and across the four cases the main difference in leverage occurs in the region of the knot where there is an increase in leverage for C_2 but a decrease in leverage for C_3 and C_4.

Finally we examine generalised leverage where in all cases the application of the WLS objective function increases leverage for low magnitude stage-discharge data. There are also distinctive differences between the four cases C_{1-4}. In C_1 we have generally higher generalised leverage than linear and nonlinear leverage with the exception of the highest stage-discharge data point where nonlinear leverage is slightly higher. Focusing our attention on the discharge uncertainty in C_2 (column 2) we see a slight decrease in the magnitude of generalised leverage with the exception of the maximum stage value. Accounting for priors in C_3 (column 3) leads to a decrease in generalised leverage across the data especially for the maximum 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 where there is a slight increase.

We now test the performance of the three regression-theory influence diagnostics to reproduce case-deletion based Cook's distance in Figure 3.8. Across the four rating curve models we see the following trends:

1. Linear Cook's distance performs well across all models in terms of Sp. (0.90), but generally poorly in terms of r^2 for models C_2-4 (maximum of 0.42 in column 3). There is generally low performance for the top 10 most influential data point metrics indicating that the diagnostic has identified the ranking of the influential

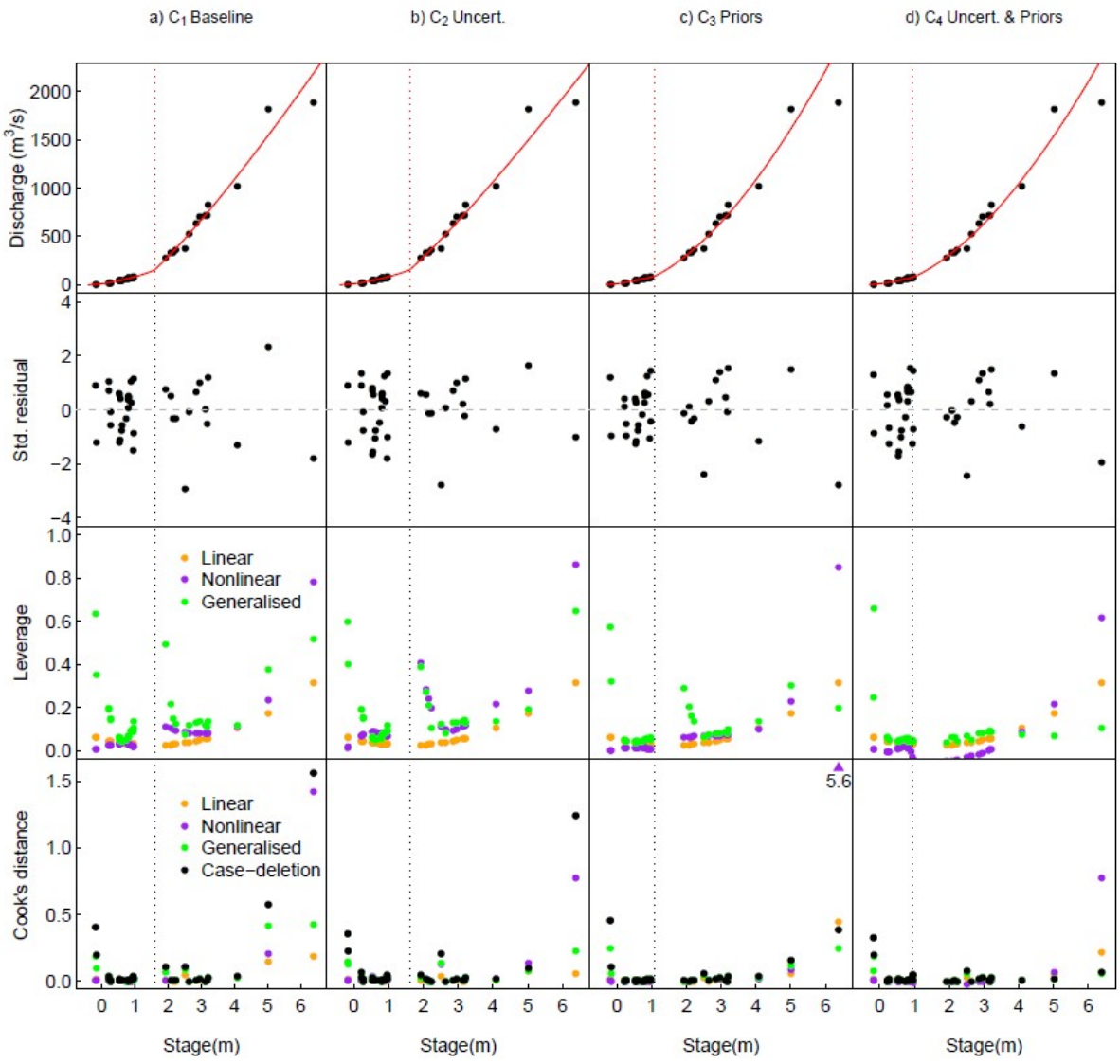

Figure 3.7: Rating curve case-study results. The computed transition level (knot) between section and channel controls is marked with a vertical dashed line. Observed data (black), and predicted model (red) in the top row, followed by standardised residuals in the second row. Leverage is shown in the third row with: linear leverage, nonlinear leverage, generalised leverage. The final row shows regression-based Cook's distance with linear, nonlinear and generalised leverage, and case-deletion Cook's distance.

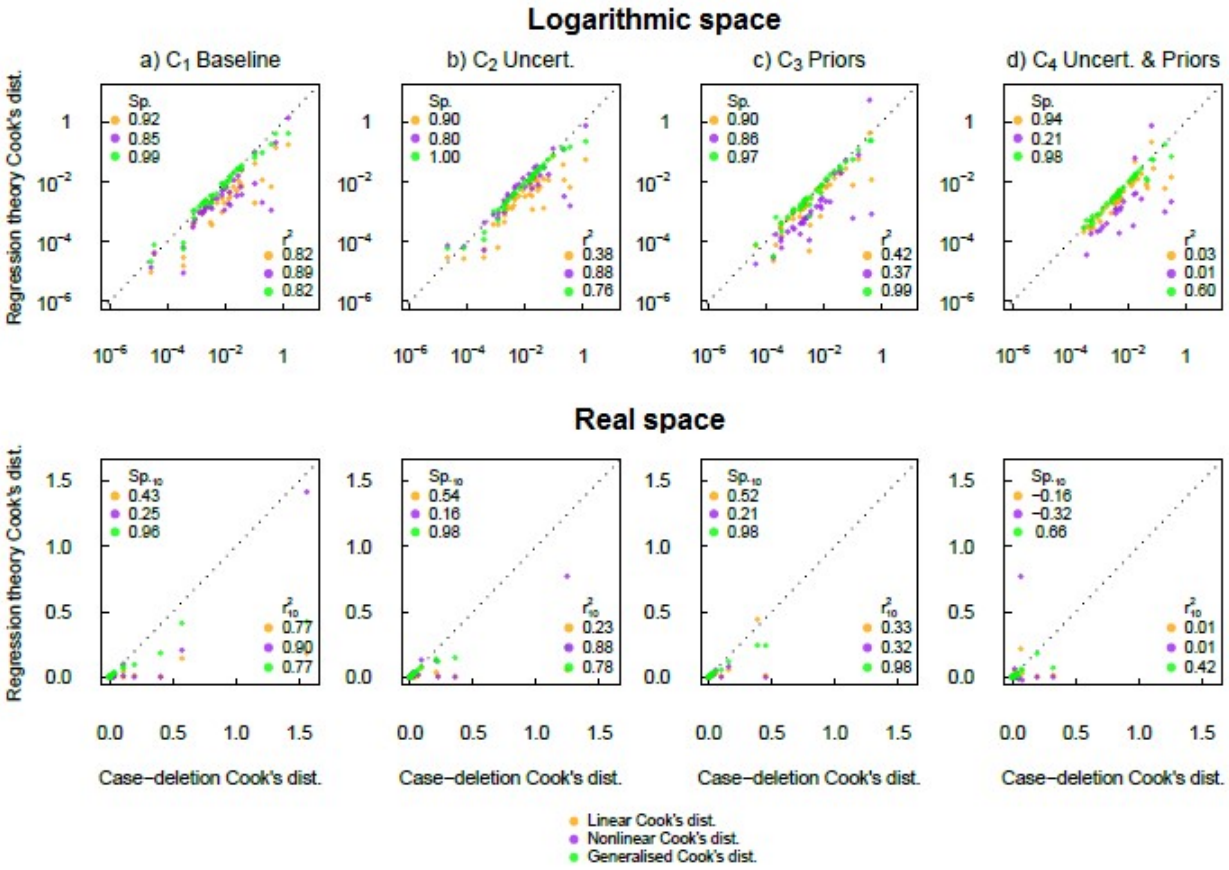

Figure 3.8: Rating curve case-study comparison of case-deletion and regression-based Cook's distance. In the first row we compare the performance in logarithmic space and use the Sp. and r^2 to highlight performance across the whole dataset. In the second row we compare the performance in real space and use the $Sp._{\cdot 10}$ and r^2_{10} to compare the subset of the ten most influential data points.

points moderately well, but has does not perform well in terms of the rest of the performance metrics,

2. Nonlinear Cook's distance has mixed performance with some mid to high range performance metrics (e.g. r^2 of 0.89 in the baseline rating curve model) but much lower performance once the priors are incorporated (e.g. r^2 of 0.37 in the third rating curve case),

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^2_{10} of 0.42).

3.4.4 Performance summary of regression-theory influence diagnostics

The performance criteria Sp. $Sp._{10}$, r^2 and r^2_{10}. results for all ten cases (A_{1-4}, B_{1-2}, and C_{1-4}) discussed in Sections 4.1 to 4.3 are summarised in Figure 3.9. The results for linear Cook's distance, (Figure 3.9, top panel) show very high $Sp.$ values, indicating that it does a reasonable job at ranking the most influential data across all data points. 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 points are completely at odds with the top 10 points identified by case-deletion Cook's distance. Examining the r^2 and r^2_{10} we see that with the exception of case linear SLS model, linear Cook's distance struggles to reproduce the case-deletion Cook's distance values.

Examining nonlinear Cook's distance (Figure 3.9, middle panel) we see that this measure does a good job at ranking the influence across all data and the top 10 influential points in synthetic cases, A_{1-4} and B_{1-2}. However for the real data case studies (B_2 and C_{1-4}) there is a sharp decrease in the performance of ranking the top 10 influential points. This is likely to be because in the real case studies, the impact of the heteroscedastic residuals errors comes into play, which is not accounted for by nonlinear leverage.

Finally we see that generalised Cook's distance (Figure 3.9, bottom panel) clearly produces the highest performance of the regression-theory influence diagnostics, across the four performance metrics. For ten of the eleven case studies, all performance metrics are above 0.9. The only case study, where generalised Cook's distance has performance below 0.9 is for C_4 the rating curve model with data uncertainty and priors.

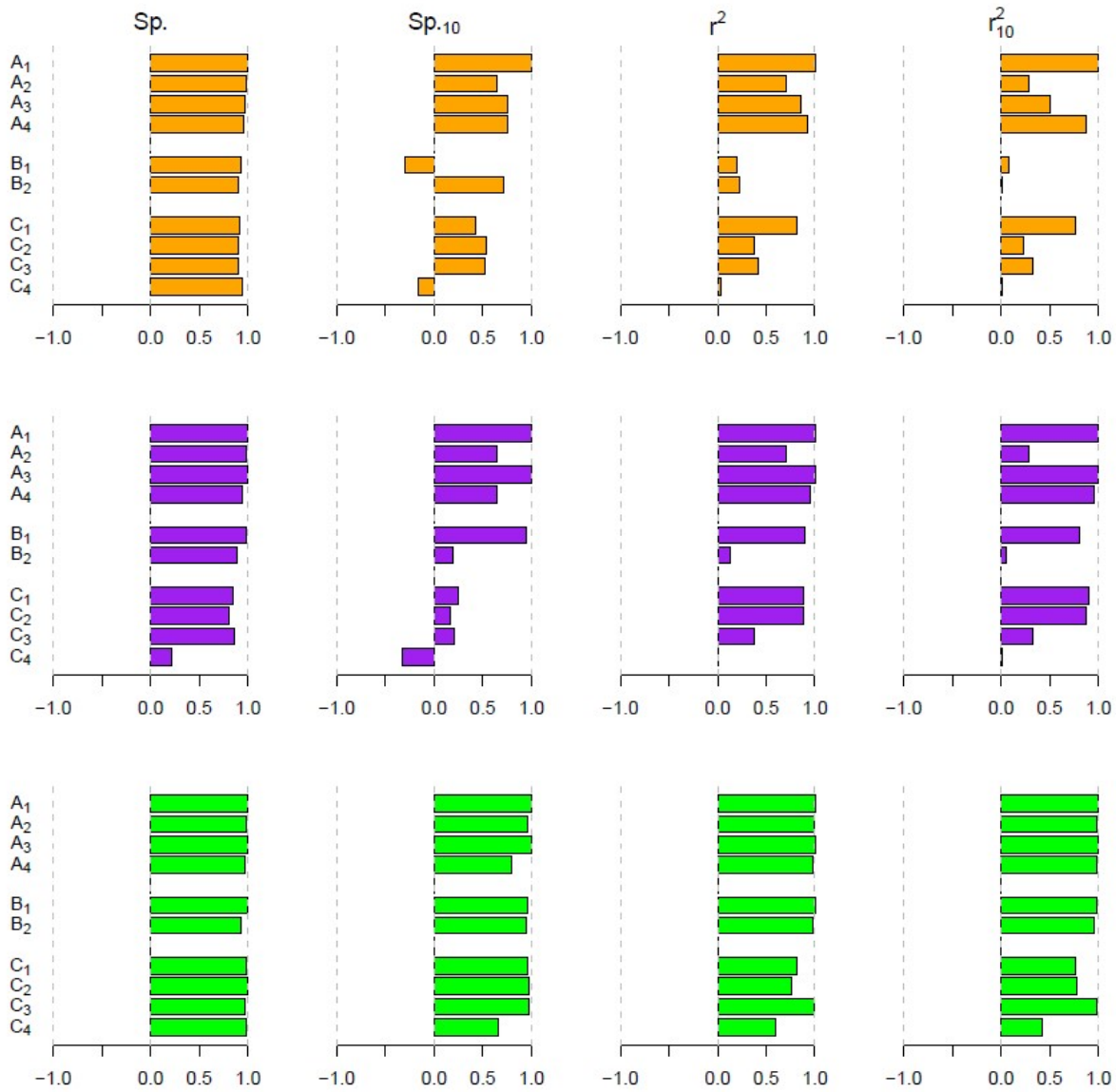

Figure 3.9: Performance metrics for regression-theory influence diagnostics across the ten models in the three case studies. Linear Cook's distance is shown in the first row (orange), nonlinear Cook's distance in the second row (purple) and finally generalised Cook's distance in the bottom row (green).

3.4.5 Computational efficiency of influence diagnostics

As described in the introduction, 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 calibration runs. Regardless of the size of the calibration data set (n) and number of model and residual error parameters (p), linear Cook's distance 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 matrix the leverage formulation (equation (3.15)). Furthermore, 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 (3.16)). Surprisingly, generalised leverage requires fewer model runs (140,000 in the example in Table 3.3) than nonlinear leverage (270,000 runs in the example in Table 3.3) despite making fewer assumptions about the residual errors and therefore being broader in potential applications.

Table 3.3: Summary of the computational demand of case-deletion and regression-based Cook's distance. The example case-study corresponds to the daily hydrological model (i.e. $p_\alpha = 4$,$p = 6$) with 10 years of data (i.e. $n = 3650$) where each model calibration requires 10 000 model runs. The example runtime is calculated with a 2.90GHz processor.

Approach	Leverage	General computation demand	Model runs
Case-deletion Cook's distance	-	n+1 model recalibration	$10000 \times (n+1)$
Linear Cook's distance	Linear	Single calibration	10000
Nonlinear Cook's distance	Nonlinear	Single calibration + finite difference calculations	$10000 + 2\,(n \times p_\alpha) + 4\,(n \times p_\alpha \times p_\alpha)$
Generalised Cook's distance	Generalised	Single calibration + finite difference calculations	$10000 + 2\,(n \times p) + 4\,(p \times p) + 4\,(n \times p)$
Approach	Example computational demand	Example runtime (hours)	Reduction from case-deletion
Case-deletion Cook's distance	36,510,000 runs	675.37	-
Linear Cook's distance	10,000 runs	0.18	99.97%
Nonlinear Cook's distance	272,800 runs	5.05	99.25%
Generalised Cook's distance	141,544 runs	2.62	99.61%

3.5 Discussion

3.5.1 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 non-linear responses, include storage, and the residuals errors are typically heteroscedastic and non-Gaussian, so that 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]. If case-deletion approaches are to be implemented, then the modeller should ensure that parameter optimisation approach can robustly handle complex response surfaces, otherwise they are at risk of falsely identifying influential data points.

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 how 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 the influential data point. 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_{1-4}, 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_{1-2} 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 provide direction of where to focus additional data collection because these points are most susceptible to influence from high residuals.

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 3.3), 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 3.1), 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 non-linear 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 eleven 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.

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

A further 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 (3.2). For example, this includes objective functions that account for autocorrelation in the residual error [see Wei et al., 1998], which is common in hydrological modelling [see Evin et al., 2014]. 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.

3.6 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 recalibration of the model parameters for every data point in the calibration dataset. Regression-theory influence diagnostics provide an approximation of 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 between the model response and residual error structure.

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. The performance in identifying the most influential data points was evaluated against case-deletion Cook's distance on a wide range of modelling scenarios (eleven case studies) which included linear/nonlinear model responses, homoscedastic/heteroscedastic residual errors, and Bayesian approaches that include data uncertainty and prior information. 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 under the following modelling conditions:

1. Nonlinear model response and heteroscedastic residual error ($Sp.$ 0.97, r^2 0.92),

2. Hydrological model structure including nonlinear model response and storage ($Sp.$ 0.93, r^2 0.98),

3. Posterior probability distributions that include data uncertainty and prior informa-

tion (Sp. 0.98, r^2 0.60).

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 these 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. 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.

3.7 References

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

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), W05552.

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.

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), W11510.
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(6), W06427.
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), W03411.
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), W06428.
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, 53(3), pp.

2199-2239.
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.
Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279(1-4), 275-289.
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.
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. P., M. Thyer, and S. Westra (2015), Influential point detection diagnostics in the context of hydrological model calibration, Journal of Hydrology, 527, 1161-1172.

Chapter 4

A hybrid framework for quantifying the influence of data in hydrological model calibration (Paper 3)

David P Wright, Mark Thyer, Seth Westra, David McInerney
Submitted to Journal of Hydrology

Statement of Authorship

Title of Paper	A hybrid framework for quantifying the influence of data in hydrological model calibration
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Wright, D.P., Thyer, M., Westra, S., McInerney, D. 2017 A hybrid framework for quantifying the influence of data in hydrological model calibration. Journal of Hydrology, (submitted)

Principal Author

Name of Principal Author (Candidate)	David Peter Wright		
Contribution to the Paper	Development and implementation of approach, visualisation and interpretation of results, preparation of manuscript and acted as corresponding author.		
Overall percentage (%)	85		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	30/03/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mark Thyer		
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript		
Signature		Date	30/3/2017

Name of Co-Author	Seth Westra		
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript		
Signature		Date	30/3/2017

Name of Co-Author	David McInerney		
Contribution to the Paper	Supervised research, helped to evaluate and edit the manuscript		
Signature		Date	30/3/2017

Abstract: Influence diagnostics identify data that have a disproportionate impact on model calibration. Although an important and commonly used diagnostic in regression, they have had limited applications in hydrological modelling. The key issues with current diagnostic approaches are that the computationally efficient approaches do not provide hydrologically relevant influence metrics, while the hydrologically relevant influence metrics are computationally expensive to calculate. This study introduces a new two-stage hybrid framework that overcomes these challenges, by delivering hydrologically relevant influence metrics in a computationally efficient manner. The first stage uses the computationally efficient generalised Cook's distance influence diagnostic to identify the most influential points. The second stage then uses the case-deletion influence diagnostics to quantify the influence of points using hydrologically relevant metrics. To illustrate the application of the hybrid framework, we conducted three experiments on 11 hydroclimatologically diverse Australian catchments using the GR4J hydrological model. The first experiment evaluated how many influential points are needed from stage one for further scrutiny in stage two of the framework. We found that in general fewer than five points were needed, irrespective of data length or hydrological influence metric (i.e. change in high flows, low flows, or mean flows). This clearly justifies the development of this hybrid framework. The second experiment found that the impact of influence generally decreased as the data length increased from one to 10 years; however even for 10 years, the impact of influential points can sometimes still be high. The third experiment compared two different objective functions and found that the WLS objective function identified data points with higher influence than SLS based on the case-deletion metrics. The hybrid framework can be applied to a wide range of hydrological and environmental models.

4.1 Introduction

Hydrological model calibration is a critical component of model development as parameters generally cannot be inferred directly from catchment measurements but are instead obtained by minimising the differences between observed and simulated streamflow [Beven, 2011]. Studies have increasingly called for the use of influence diagnostics as part of model calibration [e.g., Foglia et al., 2009; Foglia et al., 2007; Hill et al., 2015; Wright et al., 2015; Wright et al., 2017], 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 could affect the

predicted streamflow by more than 25% in a semi-arid catchment. However, the current approaches to identifying influential points (i.e. influence diagnostics) have significant drawbacks that limit their ability to identify points across a large number of hydrological studies. This paper therefore develops a two stage hybrid framework for quantifying the impact of influential points, which combines the strengths and overcomes the weakness of the currently available influence diagnostics.

Influence diagnostics can be classified into two different classes: "case-deletion" influence diagnostics and "regression-theory" influence diagnostics. "Case-deletion" influence diagnostics measure the influence of censoring ("deleting") a data point ("case") from the set of calibration points, re-calibrating, and then providing a measure of the difference between the original and re-calibrated model parameters and/or predictions. Case-deletion diagnostics are common in the statistics literature [e.g. Chen et al., 2012; Cook, 1977; Ross, 1987]. In hydrological applications they are particularly attractive as they can be tailored to provide interpretable measures of influence with respect to streamflow or any metric that the modeller wishes to investigate, such as the change in maximum or total predicted flow [Wright et al., 2015]. Additionally, case-deletion makes no assumptions regarding the type of model (linear/nonlinear) or the complexity of the residual error model (Gaussian, heteroscedastic, etc.). Although flexible, the computational demand associated with re-estimating the parameters for every data point in the observed data renders influence analysis using case-deletion influence diagnostics infeasible for anything but the simplest models (e.g. for a decade of daily data case-deletion requires 3650 model re-calibrations). A secondary issue with the case-deletion approach is that anomalous results may arise when calibrating to complex response surfaces with multiple local optima [Duan et al., 1992; Kavetski et al., 2006b], 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, any efforts will compound the computational demands of the case-deletion recalibrations.

The second class of influence diagnostic is "regression-theory" influence diagnostics. Here, the goal is to approximate the value for Cook's distance [Cook, 1977], which is a commonly used measure of influence and has been used in a large range of regression problems [Fox and Weisberg, 2011]. This is achieved by combining two regression components for each observed data point: (1) the leverage, which is used to assess the potential importance of individual observations [Wei et al., 1998], and (2) the standardised residuals. By combining these two components, regression-theory influence diagnostics

require no additional re-calibrations and are therefore an attractive alternative to the computationally demanding case-deletion influence diagnostics. Regression-theory influence diagnostics can be calculated using a range of formulations for the leverage—some with strict assumptions. Linear leverage [Cook, 1977] assumes that the regression model is linear and that residual errors are Gaussian, homoscedastic and independent. Nonlinear leverage [St. Laurent and Cook, 1992; 1993] can take into account nonlinear model response, and is suitable for nonlinear models with Gaussian residuals. Lastly, generalised leverage [Wei et al., 1998] can also take into account nonlinear model response, and can be applied to a broad range of objective functions, including those with heteroscedastic and/or non-Gaussian residual error assumptions. Wright et al. [2017] compared all three leverage formulations and found that "generalised Cook's distance", which was based on generalised leverage, to be the most suitable approach for identifying influential data for wide range of hydrological model calibration problems, including heteroscedastic residual errors and Bayesian objective functions which incorporate data uncertainty and prior information. However, a key drawback of applying "generalised Cook's distance" as a sole influence diagnostic is that despite providing a measure of the relative influence ranking across the calibration data it provides no information on the quantitative impact of influential data points on hydrologically relevant predictions, such as mean flow, high flows (relevant to flood risk assessment) and low flows (relevant for environmental flow assessment).

The aim of this study is to develop a hybrid framework for quantifying the impact of influential data in hydrological model calibration. This new framework seeks to overcome the limitations of current approaches by combining the strengths of the two classes of influence diagnostics; namely computational efficiency, and the quantification of the influence of data points on hydrologically relevant metrics. The hybrid framework achieves this through a two-stage approach, where the most influential points are first identified using the computationally efficient generalised Cook's distance, and then case-deletion influence diagnostics are applied so that influence can be presented in terms of hydrologically relevant metrics. This hybrid framework paves the way for quantifying the influence of individual data points for wide range of hydrological modelling applications. After developing the hybrid framework this paper then illustrates its use by investigating the following research questions:

1. *How many Cook's distance influential points are required to identify the points that have the greatest effect on mean, high and low flows?* This hybrid framework relies on the identification of the "most influential" data points in its first stage.

This leads to the question of how many data points to retain for the second stage to calculate case-deletion influence using prediction metrics such as mean, high and low flows. This is currently unknown because the generalised Cook's distance quantifies influence on the entire data series and does not consider the influence on the specific components of hydrological time series that are relevant to hydrologists.

2. *How does the length of calibration data determine the influence of individual data points on hydrologically relevant metrics?* There have been some studies that have examined how the length of the calibration data impacts on hydrological model predictions. Perrin et al. [2007] found that in general 350 days of calibration data sampled from a larger data set including dry and wet conditions are sufficient to obtain robust estimates of model parameters. Li et al. [2010] showed that in general eight years of data are sufficient to obtain steady estimates of model performance and parameters for the hydrological model SIMHYD [Chiew et al., 2002]. However, no study has yet evaluated how the length of calibration data impacts on the extent to which a single data point will have a high influence on model predictions.

3. *How can the choice of objective function impact on the influence of individual points on hydrologically relevant prediction metrics?* There have been a large number of objective functions developed for hydrological modelling applications. Recently there has been an increased focus on objective functions based on likelihood theory as these can be used to estimate predictive uncertainty (e.g. BATEA [Kavetski et al., 2006a] and DREAM [Vrugt and Ter Braak, 2011] frameworks). Of these, the simplest likelihood is standard least squares (SLS). However, it has long been known that residuals in hydrological applications are generally heteroscedastic and autocorrelated [Sorooshian and Dracup, 1980]. This can lead to direct treatments of heteroscedasticity using weighted least squares (WLS) [Evin et al., 2013; Evin et al., 2014; Schoups and Vrugt, 2010] or through the use of transforms such as logarithmic or Box-Cox transforms [Cheng et al., 2014; Del Giudice et al., 2013; Li et al., 2015; Smith et al., 2015; Wang et al., 2012]. Despite the widespread use of different objective functions, no study has evaluated how different objective functions affect the influence of individual points. This paper will investigate this issue by comparing the influence metrics when using the SLS or WLS objective functions.

Addressing these three research questions will provide modellers with guidance to apply computationally efficient, flexible and accurate influence diagnostics to their own studies. The remainder of this paper is structured as follows. In Section 4.2 we introduce the hybrid framework for influence assessment, and describe the components of the

proposed two-stage approach. In Section 4.3 we describe the experimental approach to answering the three questions outlined above, including the hydrological model calibration procedure and the hydro-climatologically diverse case study catchments. In Section 4.4 we apply the hybrid framework to answer the three research questions. Section 4.5 discusses the interpretation and implications of the experiments and future extension to the hybrid framework, and we finish with the conclusions in Section 4.6.

4.2 Hybrid framework

Until now in both the statistical and hydrological literature, influence diagnostics have been classified into two different classes: case-deletion influence diagnostics and regression-theory influence diagnostics. In hydrology we are particularly interested in quantifying the magnitude of influence with respect to specific components of hydrological time series that are relevant to hydrologists—such as the mean flow, high flows and low flows—so we require case-deletion diagnostics that quantify influence calculated with respect to the variables of interest. Here we develop a hybrid framework for quantifying the impact of influential data on hydrological model calibration. This new framework aims to overcome the weakness of current influence assessment approaches by combining the strengths of the two classes of influence diagnostics: 1) computationally efficiency, and 2) the flexibility to calculate influence with respect to any hydrologically relevant metric of interest. The framework proceeds by first calculating influence using generalised Cook's distance, and then applying case-deletion on only those points that have the highest generalised Cook's distance. We also require a choice of a suitable sample of influential data points identified by generalised Cook's distance, as this sample will not necessarily be identical to the points identified by each individual flow metric. This leads us to development of an innovative two-stage hybrid framework:

Stage one: Identifying the most influential points using regression based generalised Cook's distance

Once hydrological model calibration has been undertaken, the first stage of the hybrid framework is to calculate generalised Cook's distance (see Section 4.2.1) using generalised leverage and the standardised residuals. We then use this generalised Cook's distance to rank the data points and choose a subset of the most influential data points denoted as N_I. The advantage of applying generalised Cook's distance in stage one is that it is much more computationally efficient than applying case-deletion influence assessment to the whole calibration data set (e.g. 250 times faster in Wright et al. [2017]). The disadvantage is that generalised Cook's distance is not hydrologically interpretable and

therefore cannot be used to meaningfully quantify influence in a hydrological context, therefore necessitating stage two of the hybrid framework.

Stage two: Quantifying influence using case-deletion hydrologically relevant metrics

The outcome of stage one is to identify the N_I most influential data points identified by generalised Cook's distance. Stage two re-calibrates the hydrological model an additional N_I times in order to calculate the case-deletion influence diagnostics for each of these 'high influence' data points (see Section 4.2.2). This enables the influence to be described using hydrologically relevant variables such as mean, high and low flows, while reducing the runtime of the case-deletion component as it is only applied N_I times, rather than for all the data points in the calibration series.

The disadvantage of stage two of the framework is that influence is only quantified on the subset of the N_I most influential data points identified in stage one, and the points identified by generalised Cook's distance may not correspond exactly to the points that are influential with respect to the different flow metrics. Therefore we face a trade-off in that N_I should be large enough to capture the most influential data points with respect to the metrics, but small enough to minimise the computational demand. As a result, appropriate choice of N_I is crucial to the successful application of the two stage hybrid framework. In the experiments in this study we start by applying a conservative N_I of 30, and then we subsequently optimise N_I in experiment one (Sections 3.1 and 4.1).

4.2.1 Stage one: Identifying the most influential points using regression based generalised Cook's distance

Influential data points are typically those that have both high standardised residuals and high leverage (i.e. how far the data point is from the centre of the input space) [Wright et al., 2015]. Wright et al. [2017] developed generalised Cook's distance by combining the generalised leverage [Wei et al., 1998] with the standardised residuals and showed that it had high performance in approximating case-deletion influence diagnostics in a range of modelling conditions including linear/nonlinear model responses, homoscedastic/heteroscedastic residual errors, and Bayesian objective functions that include data uncertainty and prior information.

Before we define generalised Cook's distance, we need to first define a general model response as:

$$\mathbf{y} = f(\alpha; \mathbf{X}) + \varepsilon \tag{4.1}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)$ is a vector of n observed responses, $f(.)$ is the model structure, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{p_\alpha})$ is a vector of p_α model parameters, $\mathbf{X}$ is an $n \times k$ matrix of

*k*observed inputs, (e.g., precipitation, PET), and ε is a vector of n residual errors. Residuals are further assumed to be realisations from a given probability distribution, parameterised with some unknown parameters $\beta = \left(\beta_1, \beta_2, ..., \beta_{p_\beta}\right)$ (e.g. a centred Gaussian distribution with unknown standard deviation).

We calculate generalised Cook's distance by combining the standardised residual of the i^{th} point (ν_i) with the generalised leverage of i^{th} observation on the i^{th} prediction (L_{ii}):

$$CD_i = \frac{{\nu_i}^2}{p} \frac{L_{ii}}{1 - L_{ii}} \tag{4.2}$$

The standardised residuals ν are obtained by dividing the raw residuals ε by their calibrated standard deviations:

$$\nu = \frac{\mathbf{y} - \widehat{\mathbf{y}}}{\widehat{\sigma}} \tag{4.3}$$

where $\widehat{\sigma}$ can be a vector or a scalar $\widehat{\sigma}$ enabling the standardised residual to be calculated with respect to a wide range of heteroscedastic or homoscedastic objective functions.

Leverage is a key component of generalised Cook's distance and is measured by the rate of change of the i^{th} predicted value $\widehat{y}_i$ with respect to the j^{th} observed value y_j [Wei et al., 1998]:

$$L_{ij} = \partial \widehat{y}_i / \partial y_j \tag{4.4}$$

Or in matrix notation:

$$\mathbf{L} = \frac{\partial \widehat{\mathbf{y}}}{\partial \mathbf{y}^T} \tag{4.5}$$

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 will be used to calculated generalised Cook's distance. In order for leverage to be applied to a general class of regression models, including a larger range of objective functions, the generalised leverage takes into account the curvature of the objective function about the whole set of calibrated parameters $\widehat{\theta}$. In this case, leverage is equal to the diagonal elements of $\mathbf{L}(\widehat{\theta})$ given by:

$$\mathbf{L}(\widehat{\theta}) = \frac{\partial f(\widehat{\alpha}; \mathbf{X})}{\partial \widehat{\theta}} \left(-\frac{\partial^2 \Phi(\widehat{\theta}; \mathbf{y}, \mathbf{X})}{\partial \widehat{\theta}^2} \right)^{-1} \frac{\partial^2 \Phi(\widehat{\theta}; \mathbf{y}, \mathbf{X})}{\partial \widehat{\theta} \partial \mathbf{y}^T} \tag{4.6}$$

where $\frac{\partial f(\widehat{\alpha}; \mathbf{X})}{\partial \widehat{\theta}}$ is the $n \times p$ Jacobian matrix with i^{th} row $\frac{\partial f_i(\widehat{\alpha}; \mathbf{X})}{\partial \widehat{\theta}}$ (note that $\frac{\partial f_i(\widehat{\alpha}; \mathbf{X})}{\partial \widehat{\beta}} = 0$), $\frac{\partial^2 \Phi(\widehat{\theta}; \mathbf{y}, \mathbf{X})}{\partial \widehat{\theta}^2}$ is a $p \times p$ Hessian matrix and $\frac{\partial^2 \Phi(\widehat{\theta}; \mathbf{y}, \mathbf{X})}{\partial \widehat{\theta} \partial \mathbf{y}^T}$ is a $p \times n$ matrix.

The final component of stage one is to choose a suitable size N_I that balances the trade-off between minimising computational demand and ensuring that the N_I is large

enough to capture the influence with respect to the case-deletion flow metrics. This trade-off will be explored in experiment one in Section 4.3.1 and 4.1. Once a suitable N_I is selected we rank all of the points identified by generalised Cook's distance in equation (4.2), and extract the N_I most influential points by generalised Cook's distance in which to undertake case-deletion influence assessment in stage two.

4.2.2 Stage two: Quantifying influence using case-deletion hydrologically relevant metrics

In stage two we calculate case-deletion influence diagnostics for the N_I most influential points identified by generalised Cook's distance. Case-deletion influence diagnostics quantify influence by comparing predictions from a model calibrated using the full calibration dataset ($\widehat{\mathbf{y}}$), and the predictions from a model calibrated after censoring the ith data point ($\widehat{\mathbf{y}}^{-i}$). The case-deletion diagnostic for each flow measure is given as:

$$\Delta metric\,(\%) = \frac{metric\,(\widehat{\mathbf{y}}) - metric\,\left(\widehat{\mathbf{y}}^{-i}\right)}{metric\,(\widehat{\mathbf{y}})} \times 100 \tag{4.7}$$

In the experiments in this study we use three hydrologically relevant case-deletion influence metrics that focus directly on different flow measures; namely 1) mean flow prediction, 2) maximum flow prediction, and 3) the 10th percentile low flow by volume of the flow duration curve. It is noted that although the focus of this study is on understanding the impact of data on model predictions using three prediction measures described above, the case-deletion approach can be applied to any performance or prediction metric of interest, such as the objective function displacement metric or change in model parameters in Wright et al. [2015].

4.3 Methodology

The hybrid framework described in Section 4.2 paves the way for identifying influential points in stage one, and then quantifying the influence using hydrologically meaningful measures in stage two. To test the hybrid framework we conduct three experiments. Firstly, we investigate how many influential points are needed for stage one of the hybrid framework. Secondly, we investigate the impact of data length on the magnitude of influence when using hydrologically relevant flow metrics. Finally, we investigate the impact that objective function choice can have on the influence magnitude of data points.

4.3.1 Experiment one: Determining how many influential points are needed for stage one of the hybrid framework

A key element of stage one of this hybrid framework is to use generalised Cook's distance to identify the most influential data points. We must also make a choice on how many influential points, N_I, are needed to be confident that the identified data points capture the most influential points when focusing on the case-deletion flow metrics. We face a trade-off in our choice of N_I in that it needs to be large enough to sufficiently capture the most influential data points, but also small enough in order to limit the computational demand of case-deletion model recalibrations required by stage two of the hybrid framework.

To investigate a suitable choice of N_I, we start by quantifying the case-deletion influence metrics for the 30 points with the highest generalised Cook's distance influence metrics. Our goal is to provide guidance on a suitable choice for N_I (assumed to be less than 30) based on the likely magnitude of influence calculated by the hydrological metrics. To explore this we apply the GR4J hydrological model [Perrin et al., 2003] (see Section 4.3.1.1) to 11 Australian catchments (see Section 4.3.1.2) and examine the number of influential data points above an influence threshold (see Section 4.3.1.3). We use calibration data lengths of one and 10 years. To facilitate the comparison between the two data lengths, we use the calendar year within the full 10 year period that has the (1) same maximum value as the full 10 year period; (2) closest matching mean compared to the full 10 year period; and (3) closest matching low flow percentile to the 10 year period, depending on which case-deletion metric is being considered.

4.3.1.1. Hydrological model and calibration procedure

The GR4J model has a simple structure that represents interception, infiltration, and percolation. It was selected based upon its widespread use [e.g. Evin et al., 2014; Le Moine et al., 2007; Wright et al., 2015] and parsimonious model structure allowing for computational efficiency in the case-deletion model runs required to calculate case-deletion changes in flow metrics. The GR4J hydrological model has parameters $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$, 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 the unit hydrograph.

As part of this experiment, we apply the SLS objective function (see later Section 4.3.3.1) to the catchment data optimised using the Shuffled Complex Evolution (SCE) search algorithm [Duan et al., 1992; Duan et al., 1994] followed by a polish to machine precision using a Nelder-Mead gradient search.

4.3.1.2. Catchments

We use 11 Australian catchments applied by McInerney et al. [2017] (Table 4.1) that are broadly representative of Australian conditions with Köppen-Geiger climate classifications ranging from tropical savannah in the North-East, to temperate regions with warm and dry summers in the South-West and Central regions [Peel et al., 2007] as illustrated in Figure 4.1.

Table 4.1: The Australian hydrological reference station catchments used in this analysis.

Catchment	ID	Annual prec. (mm/yr)	Annual runoff (mm/yr)	Runoff/prec.	Calibration time period
Coen River at Racecourse	922101B	1570	905	0.58	1970-1979
Wild River at Silver Valley	116014A	1198	246	0.21	1990-1999
Namoi River at North Cuerindi	419005	822	94	0.11	1990-1999
Abercrombie River at Hadley No. 2	412066	768	101	0.13	1978-1987
Queanbeyan River at Tinderry	410734	773	95	0.12	1978-1987
Cotter River at Gingera	410730	1136	306	0.27	1978-1987
Murray River at Biggara	401012	1090	328	0.30	1978-1987
Hellyer River at Guilford Junction	61	2051	1340	0.65	1990-1999
Lerderderg River at O'Brien Crossing	231213	957	185	0.19	1978-1987
Rocky River upstream Gorge Falls	A5130501	756	96	0.13	1978-1987
Deep River at Teds Pool	606001	869	68	0.08	1978-1987

4.3.1.3. Thresholds to determine if a point is influential

For the purposes of this study, we choose a minimum threshold of 5% as the minimum value for categorising whether a point is highly influential to model calibration. This threshold was chosen to correspond with the level of typical uncertainties associated with streamflow gaugings [e.g. 5-20% in Le Coz et al., 2014] and is therefore a good measure when the influence of an individual data point on flow metrics is of significant concern.

4.3.2 Experiment two: Investigating the impact of data length on magnitude of influence on hydrologically relevant flow metrics

The key purpose of experiment two is to evaluate how the length of calibration data impacts on the influence of individual data points on the model calibration. To investigate this relationship we use the same catchment data and hydrological model calibration procedure as in experiment one, however this time we apply four different calibration data lengths: one year, two years, five years and the full 10 years of calibration data. To ensure consistency of the comparison, we use the one year, two year and five year data periods within the 10 year period that have the corresponding maximum and closest matching mean and low flow percentile compared to the full 10 year period for our three case-deletion flow metrics.

4.3.3 Experiment three: Investigating the impact of objective functions on magnitude of influence on hydrologically relevant flow metrics

The key purpose of experiment three is to evaluate how the choice of objective function can affect the influence of individual data points on the model calibration. To investigate this we use the same catchment data (focusing on the full 10 years of calibration data for this experiment) and hydrological model calibration procedure as in experiments one and two, however this time we apply two different objective functions: standard least squares (SLS), and weighted least squares (WLS).

3.3.1. Standard Least Squares (SLS)

The SLS objective function is equivalent to Nash-Sutcliffe efficiency (NSE) [Nash and Sutcliffe, 1970]. Assuming independent and identically distributed Gaussian residual er-

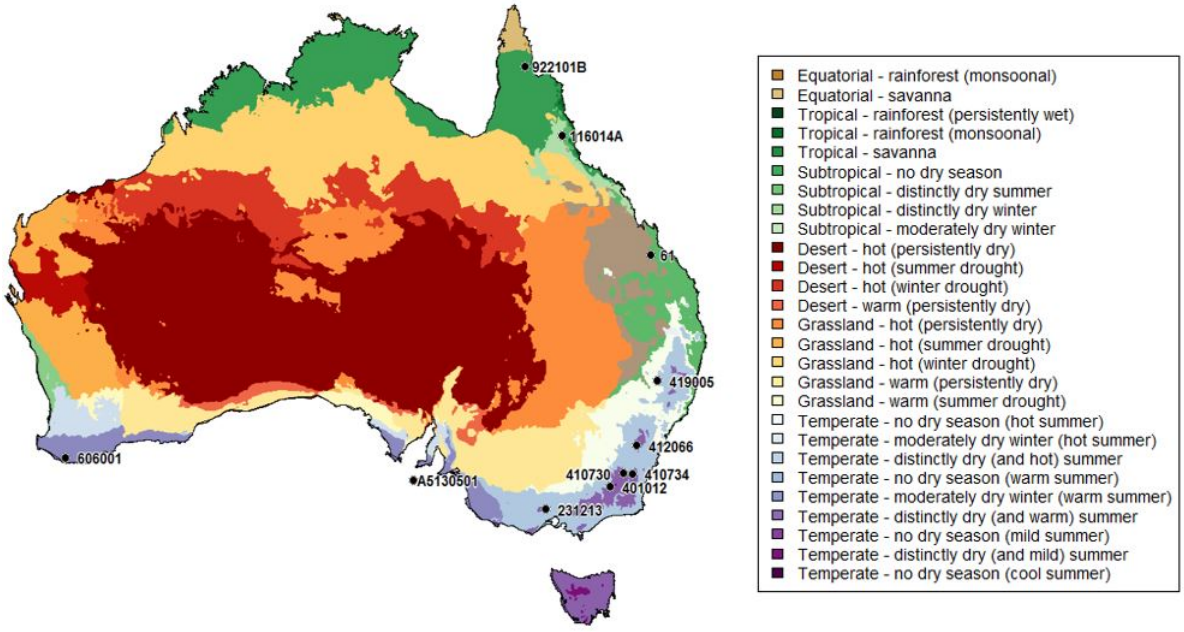

Figure 4.1: Locations of the Australian catchments considered in this study including with Köppen climate classification. Catchment IDs and properties are detailed in Table 1.1.

rors, the following log likelihood can be used as the objective function:

$$\Phi(\theta;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \log(p_N(y_i - f_i(\alpha;\mathbf{X})\,|0,\sigma^2)) \tag{4.8}$$

where $p_N(x|\mu,\sigma^2)$ is the Gaussian probability density of x assuming constant mean μ and variance σ^2. For the SLS objective function (8), $\hat{\sigma}$ is constant across the dataset. As the standard deviation σ is unknown it will need to be estimated, and therefore we have the additional parameter $\beta = \{\sigma\}$.

3.3.2. Weighted Least Squares (WLS)

Due to well-known heteroscedasticity in hydrological residual error [Thyer et al., 2009] we can replace the constant standard deviation σ in equation (4.8) with an n vector of standard deviations σ. 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:

$$\sigma = \beta_1 \widehat{\mathbf{y}} + \beta_2 \tag{4.9}$$

The WLS objective function then becomes:

$$\Phi(\theta;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \log(p_N(y_i - f_i(\alpha;\mathbf{X})\,|0,{\sigma_i}^2)) \tag{4.10}$$

As the parameters describing the non-constant standard deviation are unknown they will need to be estimated (i.e.$\beta = \{\beta_1, \beta_2\}$).

4.4 Results

4.4.1 Experiment one: Determining how many influential points are needed for stage one of the hybrid framework

The first experiment investigates the number of influential points N_I required to successfully apply the hybrid framework. In Figure 4.2 we demonstrate the viability of the hybrid framework, whereby we show how the top 30 most influential data points identified using generalised Cook's distance are able to reflect the most influential data points using the case-deletion based metrics. For this experiment the SLS objective function was used, and the results are assessed using both one year and 10 year calibration periods.

Comparing the medians across columns we can see a significant difference between the one year and the 10 year calibration data. In the 10 year calibration data we have a sharper drop-off in the boxplot medians moving from the most influential data point (left) to the least influential data point (right). This indicates that there are typically a larger number of influential data points in the case of the one year than the 10 year calibration data. Interestingly, in the case of the mean flows and low flows we see larger variability in the 90% confidence limits in the case of the one year calibrations, indicating a greater variability in the influential data with respect to the three flow metrics.

Having examined the trends across the two calibration data lengths and the three flow metrics, we now evaluate the number of influential points that lie above the median 5% influence threshold in Figure 4.2 as the basis for recommending a suitable size of N_I. In the case of both one year and 10 year calibrations, the only case-deletion flow metric with a median above the 5% influence threshold is in the change in maximum flows (top row). Also, in both cases it is only the most influential point identified by generalised Cook's distance that has a median above the 5% threshold. This result indicates that regardless of the calibration data length, for all three metrics a value of $N_I > 5$ is more than sufficient to identify the points that are likely to have the greatest case-deletion influence metrics.

The results of experiment one therefore indicates that the hybrid framework introduced in Section 4.2 is likely to be viable using $N_I > 5$, and indicates that the application

of $N_I = 30$ throughout the three experiments in this study is more than sufficient to capture the most influential case-deletion changes in flow metrics.

4.4.2 Experiment two: Investigating the impact of data length on magnitude of influence on hydrologically relevant flow metrics

Next we investigate the impact of data length on the magnitude of influence on the hydrologically relevant flow metrics. In Figure 4.3 we apply four different calibration data lengths, in each case comparing the shorter calibration periods (i.e. one, two and five years) that have the closest maximum, mean and low flow to the full 10 years calibration period. Instead of plotting all 30 influential points in each case in Figure 4.3, we plot boxplots of maximum change in flow metric (i.e. the highest magnitude in each 30 cases) for the four different calibration data lengths.

We start by comparing the change in maximum flows (first row) across calibration data lengths. Comparing the median values we see that the magnitude of influence typically decreases with increasing calibration data length. The one year of calibration median is highest (15.7%), followed by the five year (13.4%), two year (12.2%) and finally the 10 year (6.7%) periods. In addition to a drop in the median, we see that the 75th percentile drops from 28.8% in the one year period to 19.2% in the 10 year period. We see a similar pattern in the change in mean flows (second row) and the change in low flows (third row) where in both cases we see general decrease in influence moving from one year of calibration data to the 10 years of calibration data. Surprisingly, even in the case with 10 years of data a single point out of 3650 data points can have a large magnitude of influence on the predicted flows (6.7%, 1.2% and 3.2% median influence for maximum, mean and low flows, respectively).

The results from experiment two therefore show that there is a general decrease in the magnitude of case-deletion influence as the length of the calibration data increases. However, even when applying 10 years of calibration data it is still possible to observe large magnitude case-deletion influence values, indicating that individual data points can strongly influence predictions even when calibrating on relatively long data lengths.

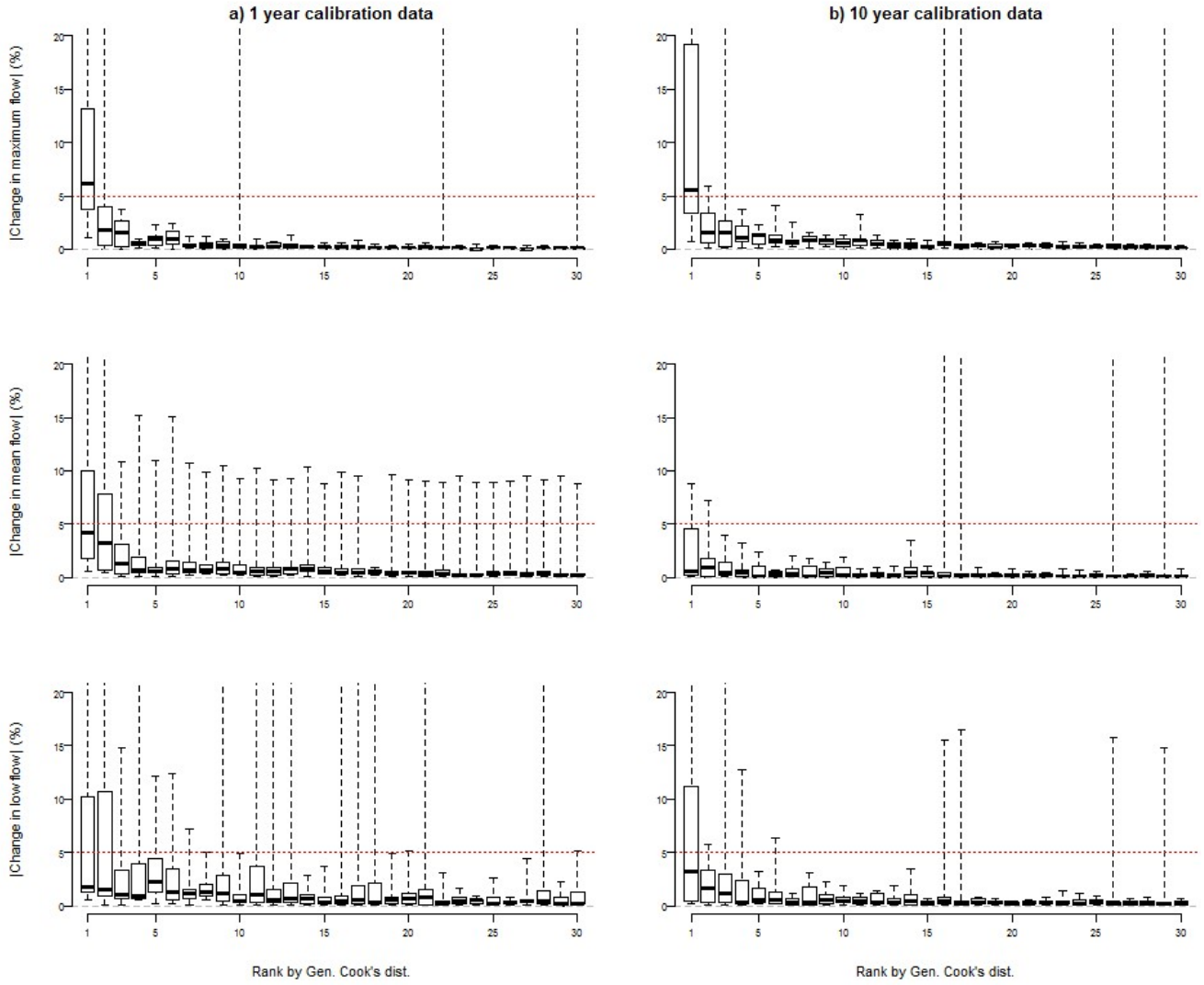

Figure 4.2: Determining a suitable N_I for the hybrid framework based upon the magnitude of case-deletion influence metrics for the top 30 most influential data points identified by generalised Cook's distance. Column a) shows the results from a 1 year calibration period for the 11 catchments, and column b) shows the results from the 10 calibration period for the 11 catchments. The whiskers represent the 90% confidence intervals. The horizontal red broken line indicates the threshold case-deletion influence of 5%.

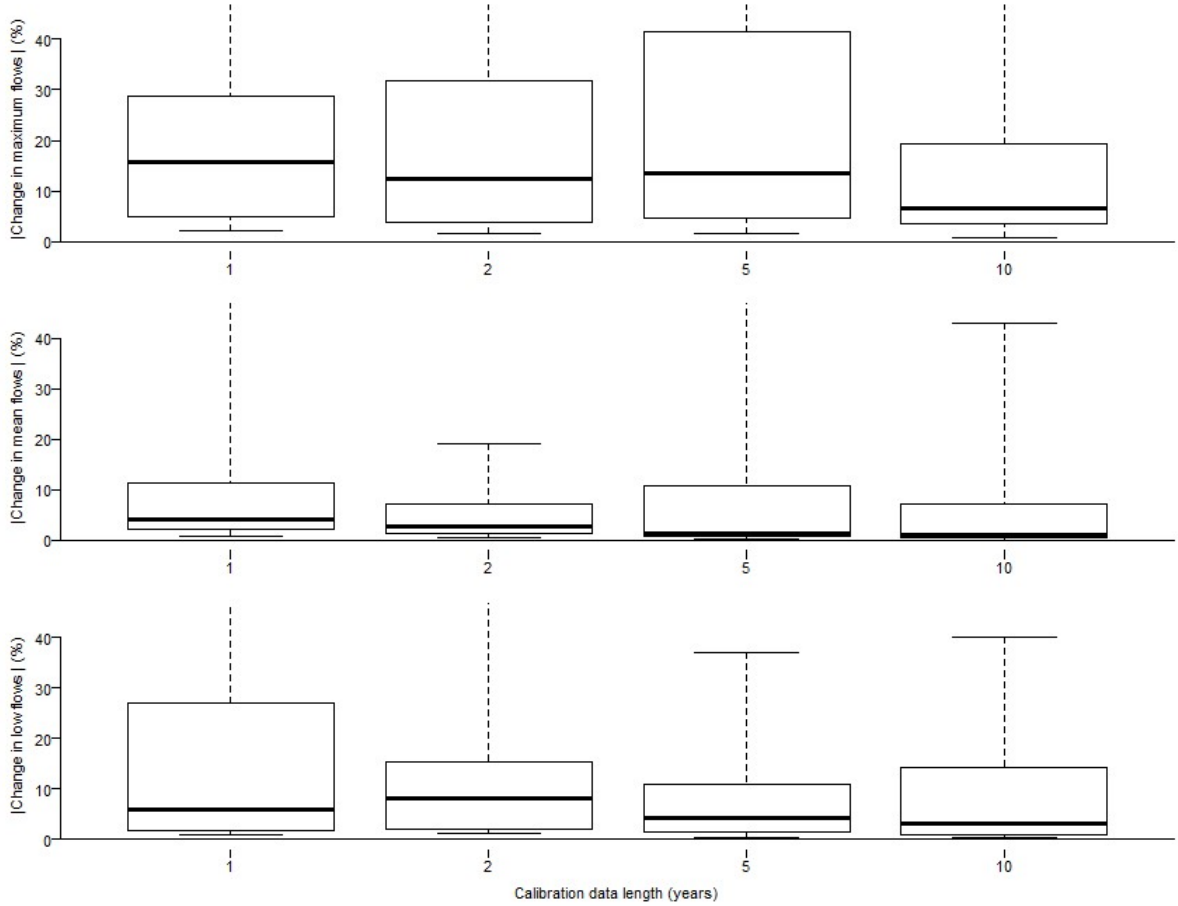

Figure 4.3: Comparing the magnitude of case-deletion influence metrics across the four calibration data lengths across the 11 catchments for the top 30 most influential data points identified by generalised Cook's distance. We apply four different calibration data lengths in each case comparing the 1, 2 and 5 year period with the closest maximum, mean and low flow to the full 10 years of calibration data. The whiskers represent the 90% confidence intervals.

4.4.3 Experiment three: Investigating the impact of objective functions on magnitude of influence on hydrologically relevant flow metrics

In the third experiment, we investigate the impact objective function choice can have on the magnitude of influence on the hydrologically relevant flow metrics. In Figure 4.4 we plot boxplots for the SLS and WLS objective functions against the three different case-deletion flow metrics and the generalised Cook's distance.

Comparing across the first three rows in Figure 4.4 for all three cases we see that there is a much larger influence in the case of WLS than SLS in the maximum flow metric (median 58.2% for WLS compared to 6.7% for SLS), mean flow metric (median 28.6% for WLS compared to 1.2% for SLS) and low flow metric (median 35.3% for WLS compared to 3.2% for SLS). These results indicate that the WLS objective function is more susceptible to highly influential data across the three case-deletion flow metrics than SLS.

Moving to the fourth row of Figure 4.4 we examine the magnitude of generalised Cook's distance for the SLS and WLS objective functions. Surprisingly, we see a different result to the first three rows where according to generalised Cook's distance the SLS objective function is more influential (median 8.1%) compared to WLS (median 0.6%). This result indicates that despite being able to identify the most influential points when the points are ranked in an internally consistent way, the generalised Cook's distance clearly fails at quantifying the magnitude of influence of individual data points on hydrological predictions as its magnitude may not be compared across two different models.

In summary, the results from experiment three show that the WLS objective function is more susceptible to single influential data points than the SLS objective function in terms of all three case-deletion flow metrics. We see the opposite trend when examining generalised leverage, indicating that despite performing well in identifying the most influential data points in a ranked sense, it is not suitable to compare the magnitude of influence across the two objective functions.

4.5 Discussion

4.5.1 Interpretation and implications of the experiments

In experiment one, we found that typically a low number of points are influential to model predictions. The reason why only a small number of data points are influential is because

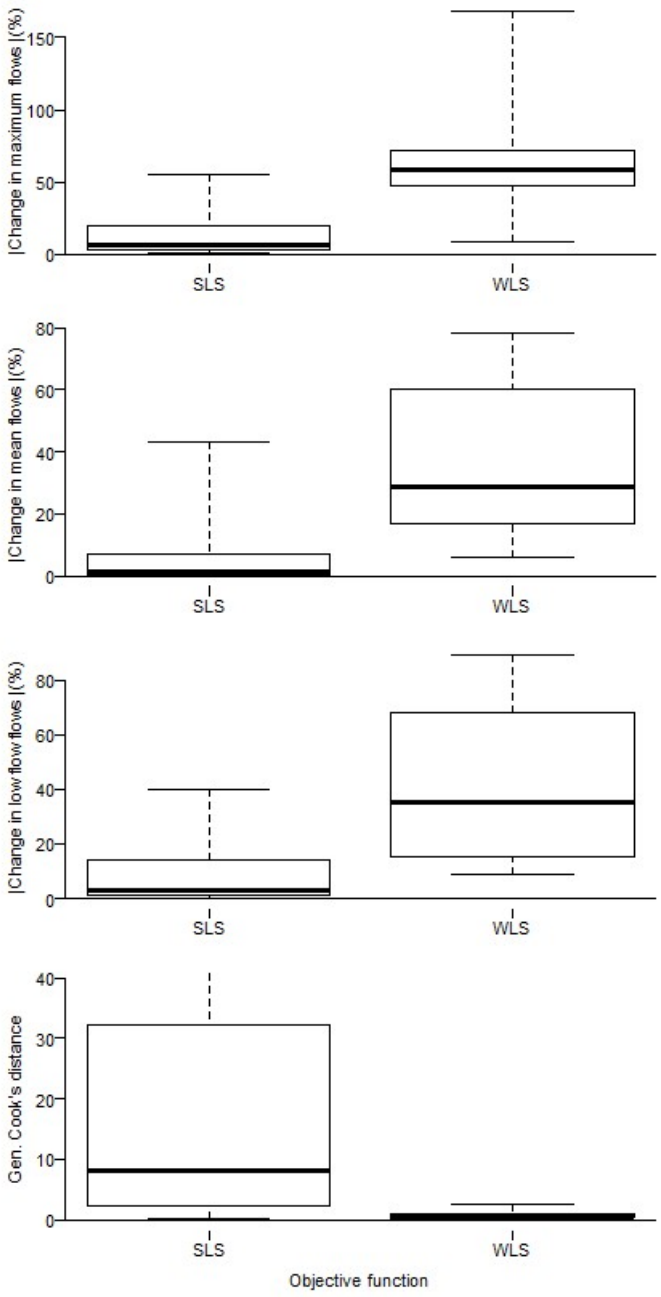

Figure 4.4: Comparing the magnitude of case-deletion influence metrics and generalised Cook's distance across the two objective functions (SLS and WLS respectively) for 10 year calibration data period across the 11 catchments for the top 30 most influential data points identified by generalised Cook's distance. The whiskers represent the 90% confidence intervals.

in order for a point to be influential it must have both a very high standardised residual and high leverage, and typically very few data points satisfy both of these criteria. This result highlights the importance of undertaking influence assessment in hydrological modelling to ensure that this small number of influential data points are true representations of the catchment response and are not erroneous or disinformative [see Beven and Westerberg, 2011] to model calibration.

In experiment two, we found that the influence typically decreases with increasing data length, a trend that is not surprising since an increase in the amount of data would be expected to reduce the possibility of a single point having a significant influence on the overall calibration. Surprisingly, we did see that even in the case of 10 years of daily hydrological data (i.e. 3650 data points), a single daily data point can still have a significant influence on case-deletion flow metrics, thereby reinforcing the need for influence assessment even with in the case of long calibration data periods.

In experiment three we saw that applying the WLS objective function leads to generally much larger influence across the three case-deletion flow metrics (typically 10 times the magnitude of SLS influence when comparing the medians across the three case-deletion metrics). This result could be because WLS puts higher weight on high leverage low flows, and in cases where these low flows also have very high standardised residuals they have high influence. The implication of this result is that when using heteroscedastic objective functions the modeller may need to be more wary of the impact of influential points. Furthermore, we saw that despite being able to identify the most influential data points, the generalised Cook's distance failed to quantify influence as we saw an opposite trend than the three case-deletion influence diagnostics. This result illustrates a key issue with applying generalised Cook's distance; namely, that the magnitude of Cook's distance is case study specific and cannot be applied across case-studies. This highlights the danger of applying generalised Cook's distance as a standalone influence diagnostic and reinforces the need for the hybrid framework that combines both generalised Cook's distance and case-deletion influence diagnostics to provide comparable influence metrics.

4.5.2 Future extension to the hybrid framework

The hybrid framework provides a computationally cheap and flexible approach to influence assessment that can be applied to a broad range of hydrological and environmental models. However, it should be noted that generalised Cook's distance requires an objective function that can be summed over the individual data points [see Wright et al., 2017], so that the approach is particularly well suited to objective functions that are expressed

via a likelihood. In practice most objective functions commonly used in hydrology can be expressed in this suitable form; however these assumptions may not be appropriate for all hydrological or environmental modelling applications. Where this is not feasible, it is not possible to proceed with the first stage of the hybrid approach, and it becomes necessary to apply the case-deletion metrics on the full data set.

Finally, the hybrid approach described here is particularly beneficial for situations where re-calibrating across all data points in the model presents a significant computational burden. For cases where the optimal parameter set is readily identified (such as linear regression), and/or for situations where the calibration data set is relatively small (such as rating curves [see Wright et al., 2017]), the modeller may choose to apply the case-deletion influence diagnostics to the whole set of calibration data rather than following the two stage approach.

4.6 Conclusions

Influence diagnostics identify data points that have a disproportionate impact on model calibration, and are therefore useful to identify possible erroneous data points or to scrutinise the sensitivity of the model predictions to a small portion of the overall calibration dataset. Current methods for influence assessment have drawbacks that have limited their application: case-deletion based influence diagnostics have a large computational demand, whilst generalised Cook's distance is not hydrologically interpretable.

In this study we present a two stage hybrid framework for quantifying influence in hydrological model calibration that combines the strengths of the two existing classes of influence diagnostics; namely 1) computationally efficiency, and 2) quantifying influence using hydrologically relevant metrics. Stage one of the hybrid framework identifies the influential data using the computationally cheap generalised Cook's distance. The modeller then choses a suitable subset of influential data points N_I. Stage two of the hybrid framework involves undertaking case-deletion re-calibration of the hydrological model on the subset of N_I influential data points in order to quantify influence with respect to any hydrological metric of choice.

After describing the two stage hybrid framework for influence assessment, we apply the new approach to three experiments that use 11 Australian catchment data and the GR4J hydrological model. The experiments have the following key outcomes:

1. In general $N_I > 5$ is a suitable choice as fewer than five points have large magnitude influence, irrespective of the data length (up to calibration periods of 10 years), or

hydrological influence metric (change in high, low, or mean flows). The result justifies the development of the two stage hybrid framework, as applying case-deletion to the subset of $N_I > 5$ data points is clearly much more computationally feasible compared to applying case-deletion to the full set of 3650 data points (in the case of the 10 year calibration period) that would be required in the absence of the hybrid approach.

2. A general pattern of decreasing impact of influence as the data length increased from 1 to 10 years; however even for 10 years, the impact of influential points can still be high (>5%). The application of the hybrid approach is therefore recommended even when calibrating to relatively long data points, as single data points can still have a disproportionate impact of hydrological model predictions for these cases.

3. The WLS objective function had far higher case-deletion change in flow metrics than the SLS objective function (e.g. case-deletion change in maximum flows median 58.2% for WLS and 6.7% for SLS), suggesting that particular caution is needed when calibrating using heteroscedastic objective functions such as WLS.

As hydrological modelling complexity increases (e.g. 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 these modelling decisions on whether a realistic representation of the catchment response has been achieved [Gupta et al., 2008]. The hybrid framework presented in this paper is suitable for a wide range of hydrological case studies and will provide modellers with new insights to aid in model calibration. These experimental results highlight that regardless of data length or objective function choice, there is a clear need to undertake influence assessment as a regular part of the model calibration and model evaluation process.

4.7 Acknowledgements

Data for the Australian catchments was obtained from the Hydrologic Reference Stations database provided by the Australian Bureau of Meteorology (http://www.bom.gov.au/water/hrs).

4.8 References

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

Beven, K., and I. Westerberg (2011), On red herrings and real herrings: disinformation and information in hydrological inference, Hydrological Processes, 25(10), 1676-1680.

Chen, X. D., N. S. Tang, and X. R. Wang (2012), Local influence analysis for semiparametric reproductive dispersion nonlinear models, Acta Math Appl Sin-E, 28(1), 75-90.

Cheng, Q.-B., X. Chen, C.-Y. Xu, C. Reinhardt-Imjela, and A. Schulte (2014), Improvement and comparison of likelihood functions for model calibration and parameter uncertainty analysis within a Markov chain Monte Carlo scheme, Journal of Hydrology, 519, Part B, 2202-2214.

Chiew, F. H. S., M. C. Peel, and A. W. Western (2002), Application and testing of the simple rainfall-runoff model SIMHYD, 335-367 pp.

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), W05552.

Del Giudice, D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann (2013), Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, Hydrol. Earth Syst. Sci., 17(10), 4209-4225.

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.

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., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resources Research, 49(7), 4518-4524.

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), W11510.
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(6), W06427.
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., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resources Research, 42(3), W03407.
Kavetski, D., G. Kuczera, and S. W. Franks (2006b), 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), W06428.
Li, C., H. Wang, J. Liu, D.-h. Yan, F.-l. Yu, and L. Zhang (2010), Effect of calibration data series length on performance and optimal parameters of hydrological model, Water Science and Engineering, 3(4), 378-393.
Li, M., Q. J. Wang, J. C. Bennett, and D. E. Robertson (2015), A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts, Hydrol. Earth Syst.

Sci., 19(1), 1-15.
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, 53(3), pp. 2199-2239.
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.
Peel, M. C., B. L. Finlayson, and T. A. McMahon (2007), Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11(5), 1633-1644.
Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279(1-4), 275-289.
Perrin, C., L. Oudin, V. Andreassian, C. Rojas-Serna, C. Michel, and T. Mathevet (2007), Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models, Hydrological Sciences Journal, 52(1), 131-151.
Ross, W. H. (1987), The Geometry of Case Deletion and the Assessment of Influence in Nonlinear-Regression, Can J Stat, 15(2), 91-103.
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.
Smith, T., L. Marshall, and A. Sharma (2015), Modeling residual hydrologic errors with Bayesian inference, Journal of Hydrology, 528, 29-37.
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.
Vrugt, J. A., and C. J. F. Ter Braak (2011), DREAM(D): an adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems, Hydrol. Earth Syst. Sci., 15(12), 3701-3713.
Wang, Q. J., D. L. Shrestha, D. E. Robertson, and P. Pokhrel (2012), A log-sinh

transformation for data normalization and variance stabilization, Water Resources Research, 48(5), W05514.
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. P., 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. P., M. Thyer, S. Westra, B. Renard, and D. McInerney (2017), A generalised approach for identifying influential data in hydrological modelling, Environmental Modelling & Software, In Review submitted February 2017.

Chapter 5

Conclusion

5.1 Research contributions

The overall contribution of this research is the development and application of influence diagnostics to hydrological modelling. Influence diagnostics have been applied in the statistical literature for decades since first introduced by Cook [1977], but have not been broadly applied to the hydrological literature due to two key issues with the current approaches:

1. Case-deletion influence diagnostics are too computationally expensive to apply in hydrological modelling applications because of the length of data and therefore number of model recalibrations that are required (e.g. 10 years requires approximately 3650 model re-calibrations).

2. Regression theory influence diagnostics are computationally efficient, but only linear Cook's distance has been applied which has strong assumptions of linear model response and Gaussian residual error that are typrically not valid in hydrological modelling.

This thesis has applied influence diagnostics to hydrological models, addressed the limitations of the current methods, and provided a hybrid framework for influence assessment hydrological modelling. The specific research contributions to address the research objectives are as follows.

Objective 1 – Explore the application of influence diagnostics in the context of a series of common hydrological modelling case studies (Paper 1): In Chapter 2 we evaluated the two classes of influence diagnostics with some hydrological examples. Case-deletion was applied to changes in the model parameters (measured through the Mahalanobis distance), performance (using objective function displacement) and predictions

(e.g. mean and maximum streamflow). For the regression theory methods, both linear and nonlinear estimates of leverage are used to calculate Cook's distance, which is used to rank influential data. We applied the diagnostics to three case studies and showed that a single point could change mean/maximum streamflow predictions by 7%/9% for a rating curve model, and 13%/25%, for a hydrological model (GR4J) calibrated using 2 years of data in an ephemeral catchment. In contrast, the influence (0.2%/2.3%) was far less in a humid catchment. The linear Cook's distance-based metric produced reasonably similar rankings to the case-deletion metrics at a fraction of the computational cost (300-1000 times faster). The nonlinear Cooks distance produced rankings that were virtually the same as the case-deletion metrics for all case studies - this highlights the importance of its use for nonlinear hydrological models. Visual assessment was not a reliable method of influence analysis as there was no direct relationship between the most influential data and the highest observed streamflows. The paper presented establishes the feasibility and potential importance of including influence diagnostics as a standard tool in hydrological model calibration.

Objective 2 – Generalise regression theory influence diagnostics to be suitable for a wide range of hydrological modelling scenarios (Paper 2): In Chapter 3 we evaluated the performance of a range of regression-theory influence diagnostics on eleven case studies with a variety of model structures and inference scenarios including: nonlinear model response, heteroscedastic residual errors, data uncertainty and Bayesian priors. Generalised Cook's distance, which uses a generalised leverage formulation, clearly outperformed linear and non-linear leverage formulations to identify the most influential points (Spearman rank correlation: 0.93-1.00) at a fraction of the computational demand of case-deletion (99.6% saving). The paper presented establishes the feasibility of applying the computationally efficient generalised Cook's distance on a wide variety of hydrological and environmental modelling problems.

Objective 3 – Develop a hybrid framework for influence assessment in hydrological modelling (Paper 3): The key issues with the current suite of influence diagnostics are that the computationally efficient regression theory approaches do not provide hydrologically relevant influence metrics, while the hydrologically relevant case-deletion influence metrics are computationally expensive to calculate. In Chapter 4 we introduce a new two-stage hybrid framework that overcomes these challenges, by delivering hydrologically relevant influence metrics in a computationally efficient manner. The first stage uses the computationally efficient generalised Cook's distance influence diagnostic to identify the most influential points. The second stage then uses the case-deletion influence diagnostics to quantify the influence of points using hydrologically relevant metrics.

To illustrate the application of the hybrid framework, we conducted three experiments on 11 hydro-climatologically diverse Australian catchments using the GR4J hydrological model on calibration data lengths of up to 10 years. The first experiment evaluated how many influential points are needed from stage one for further scrutiny in stage two of the framework. We found that in general fewer than five points were needed, irrespective of data length or hydrological influence metric (i.e. change in high flows, low flows, or mean flows). This clearly justifies the development of this hybrid framework. The second experiment found that the impact of influence generally decreased as the data length increased from one to 10 years; however even for 10 years, the impact of influential points can sometimes still be high ($> 5\%$). The third experiment compared two different objective functions and found that the WLS objective function identified data points with higher influence than SLS based on the case-deletion metrics. The hybrid framework developed in Chapter 4 is suitable for a wide range of hydrological and environmental models.

5.2 Research limitations

This work evaluated and developed influence diagnostics that can be adopted in hydrological modelling applications. However, there are three key limitations in the application of the influence diagnostics in this research as follows:

Once the influential data points have been identified their impact on model calibration will also need to be interpreted: The hybrid framework provides a computationally cheap and flexible approach to influence assessment that can be applied to a broad range of hydrological and environmental models. However a key limitation of this thesis is the interpretation of the role of influential data in model calibration and guidance to hydrological modellers on whether to include or remove influential data during model calibration.

Regression based Cook's distance calculation requires calculation of generalised leverage using objective functions that can be summed over the individual data points: The hybrid framework provides a computationally cheap and flexible approach to influence assessment that can be applied to a broad range of hydrological and environmental models. However, it should be noted that generalised Cook's distance requires an objective function that can be summed over the individual data points so that the approach is particularly well suited to objective functions that are expressed via a likelihood. In practice most objective functions commonly used in hydrology can be expressed in this form; however these assumptions may not be appropriate for all hydrological or environmental modelling applications. Where this is not feasible, it is not possible to proceed with the first stage of the hybrid approach, and it becomes necessary to apply the case-deletion

metrics on the full data set.

Case-deletion based influence diagnostics may create false influential data: Anomalous influence results may arise when calibrating to complex response surfaces with multiple local optima [see 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, any efforts will compound the computational demands of the case-deletion recalibrations.

5.3 Recommendations for future work

The culmination of this thesis resulted in a hybrid framework for influence assessment that can be adopted by hydrological practitioners. The opportunities for future development of this work are classified into two areas: 1) Determining the practical impact of identifying influential data, and 2) Enabling influence assessment to be applied to a broader range of hydrological problems.

5.3.1 Determining the practical impact of identifying influential data

Exploring the value of identifying influential points: This thesis has focused on methods for identifying influential data; however a key extension of this work would be to provide hydrological modellers with direction on whether influential data should be retained or discarded from the calibration dataset. A key issue in hydrological modelling is the presence of error in input data and examining a multi-decadal daily time series for input data error is an exhaustive exercise. Examining a much smaller subset of the most influential data points that are most informative/disinformative to model calibration gives a shortlist of data points that can be checked for error.

Exploring the impact of influential data on model validation: This research has focused on model calibration however the importance of model validation in hydrological modelling is also well known [Coron et al., 2012]. Examining the influence of calibration data on validation performance will allow the modeller to understand whether influential data increase or decrease performance when the calibrated model parameters are applied to an independent validation data set.

Investigating the relationship between influential data and hydrological model structure, catchment properties, and objective function choice: There is a large scope for in-

vestigating what elements of model development cause influential data. For example, when addressing Objective 1 we saw higher influence than in the case of the ephemeral catchment than in the case of the humid catchment. Also, when addressing Objective 3 we saw that the WLS objective function had typically higher influence than the SLS objective function. Further work exploring the impact of isolating the individual elements such as catchments, hydrological model structure, and objective functions using a range of synthetic tests will allow for a better understanding of the key mechanisms that produce influential data.

5.3.2 Enabling influence assessment to be applied to a broader range of hydrological problems

Extending the framework to measure influence for multiple responses: The case studies in this thesis are all based on single-response data. It is also common in hydrological modelling to calibrate to multiple responses such as multiple streamflow sources, streamflow and groundwater data [e.g. Rödiger et al., 2014], and streamflow and evapotranspiration. Furthermore, distributed models are commonly applied alongside or in place of lumped hydrological models [e.g. Vansteenkiste et al., 2014]. There is a significant scope for future work in extending the influence diagnostics to a fully distributed hydrological model with multiple streamflow sources. Extending influence diagnostics to handle multi-response data will require the development of case-deletion influence metrics that apply appropriate weightings to measure the change across the multiple responses.

5.4 Concluding remarks

Despite having an integral role in model calibration in the statistical literature for several decades there has been limited application of influence diagnostics in hydrological modelling and more broadly environmental modelling. The prospect that a small number of data points can exert a very high influence on model predictions motivated the investigation of influence diagnostics in hydrological applications in this thesis. The hybrid framework presented in the third paper in this thesis will provide a foundation for all hydrological modellers to have greater insight into the influence of individual data points on model calibration, thereby providing a basis for identifying disinformative points or understanding how sensitive model predictions are to a small proportion of the dataset.

Appendix A

Copy of paper from Chapter 2:

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

Journal of Hydrology 527 (2015) 1161–1172

Contents lists available at ScienceDirect

Journal of Hydrology

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

Influential point detection diagnostics in the context of hydrological model calibration

David P. Wright *, Mark Thyer, Seth Westra

School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide 5005, Australia

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form 22 April 2015
Accepted 23 May 2015
Available online 30 May 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Vazken Andréassian, Associate Editor

Keywords:
Hydrologic calibration
Case-deletion
Cook's distance
Influence diagnostics
Mahalanobis distance

SUMMARY

Influential data are those that have a disproportionate impact on model performance, parameters and/or predictions. This paper evaluates two classes of diagnostics that identify influential data for hydrological model calibration: (1) numerical "case-deletion" diagnostics, which directly measure the influence of each data point on the calibrated model; and (2) analytical diagnostics based on Cook's distance, which combine information on the model residuals with a measure of the distance of each input point from the centre of the range of the input data (i.e., the leverage). Case-deletion methods rank influence by changes in the model parameters (measured through the Mahalanobis distance), performance (using objective function displacement) and predictions (e.g. mean and maximum streamflow). For the analytical methods, both linear and nonlinear estimates of leverage are used to calculate Cook's distance, which is used to rank influential data. We apply these diagnostics to three case studies and show that a single point could change mean/maximum streamflow predictions by 7%/9% for a rating curve model, and 13%/25%, for a hydrological model (GR4J) in an ephemeral catchment. In contrast, the influence was far less for GR4J in a humid catchment (0.2%/2.3%). Assuming the data are of high quality this indicates deficiencies in the ability of the GR4J model structure to reproduce the flow regime in the ephemeral catchment. The linear Cook's distance-based metric produced reasonably similar rankings to the case-deletion metrics at a fraction of the computational cost (300–1000 times faster), but with less flexibility to rank influence using specific aspects of model behaviour. The nonlinear distance produced rankings that were virtually the same as the case-deletion metrics for all case studies – this highlights the importance of its use for nonlinear hydrological models. Visual assessment was not a reliable method of influence analysis as there was no direct relationship between the most influential data and the highest observed streamflows. The findings establish the feasibility and importance of including influence detection diagnostics as a standard tool in hydrological model calibration.

1. Introduction

The process of hydrological model calibration involves the estimation of parameters that maximise the similarity between observed and simulated hydrological response time series such as streamflow. This process requires the optimisation of one or several objective functions (Duan et al., 1992), which provide a summary measure of overall model performance. However in doing so, information on the influence of individual data points in determining the calibrated parameter set (and hence the model predictions) is often ignored.

Identifying data points that have a large influence on hydrological predictions is of particular importance when those data points are erroneous, as this is likely to lead to sub-optimal model performance when applied to an independent dataset. The importance of such "disinformative" data has been highlighted by Beven and Westerberg (2011), who identify the need for more formal methods to identify and remove erroneous data prior to model calibration. They suggest two strategies: firstly that the discrepancies of a water balance time series are evaluated for values outside some acceptable limits of uncertainty, and secondly that likelihood measures are developed that are robust with respect to disinformation. However, examining all of the high residual data can be labour intensive, and focusing only on a smaller subset of influential data is likely to be more feasible in practice. Furthermore, not all influential points are erroneous; in fact, in certain situations it may even be desirable that some data points are more influential than others. For example, objective functions that place a larger weight on high flows maybe more desirable if the application is for peak flow prediction (e.g. Duan et al., 2007). This paper aims to provide

* Corresponding author.
E-mail addresses: david.p.wright@adelaide.edu.au (D.P. Wright), mark.thyer@adelaide.edu.au (M. Thyer), seth.westra@adelaide.edu.au (S. Westra).

http://dx.doi.org/10.1016/j.jhydrol.2015.05.047
0022-1694/

hydrological modellers with the tools to assess relative influence of data points on model calibration.

Influential data points are defined as points that exert a disproportionate impact on the calibrated parameters, performance and/or predictions. Formal influential point detection methods are widely used both for the detection of erroneous points and for identifying possible model deficiencies, with common applications in linear regression (Cook, 1979), generalised linear regression (Thomas and Cook, 1989), generalised additive models (Hastie and Tibshirani, 1990) and various other regression-based approaches (Chen et al., 2012; Russo et al., 2009). The diagnostics can be grouped into two classes: case-deletion approaches and analytical leverage-based approaches.

Case-deletion methods were first developed by Cook (1977) and involve removing ("deleting") a data point ("case") from the set of calibration points, and then recalibrating the model. Parameter estimates and model predictions from the recalibration are compared to the results from the full calibration, and this is repeated for all data points in the calibration set. A recent example in the context of flood frequency analysis used case-deletion to show that low flow outliers can have a disproportionate influence on extreme flood quantile estimates (Lamontagne et al., 2013). Their technique was based on a generalised Grubbs-Beck test statistic developed by Cohn et al. (2013) that is designed to identify potentially influential low flows.

Case-deletion approaches can be computationally intensive, as they require the re-estimation of the parameters after deleting each point from the calibration data set. Furthermore, case -deletion involves comparing the optimal parameter sets from each calibrated model run, and thus anomalous results are possible for models with complex response surfaces that are prone to local optima (Duan et al., 1992). As an alternative, Cook's distance (Cook, 1977) provides an analytical measure of the influence of points, and thus does not require multiple re-calibrations. It combines measures of the distance between each observed data point and the fitted model (the residual) and the distance of each data point from the centre of the input space (the leverage). Cook's distance was originally developed for linear regression models, but may also be applied to nonlinear models if the models are approximately linear in the vicinity of the optimum parameter set (Cook and Weisberg, 1982). Alternatively, nonlinear formulations of the leverage are also available (St. Laurent and Cook, 1992), and may be better suited to the highly nonlinear behaviour of many hydrological models (e.g. see discussion in Kavetski and Kuczera, 2007).

The influence concepts are illustrated in Fig. 1 by applying case-deletion to a linear regression model. Point A is highly influential, with a significant difference in calibrated parameters when including this point ($\beta_0 = 2.0$, $\beta_1 = 2.3$, compared to $\beta_0 = 3.4$, $\beta_1 = 1.9$). The influence on predictions is also evident by comparing the fitted regression lines, with the greatest differences apparent towards the high and low extremes of the input data. In contrast, although point B has a similar residual to A (i.e. the difference between the data point and the fitted curve is similar), it exerts a much smaller influence on both the parameters ($\beta_0 = 3.8$, $\beta_1 = 1.9$) and the fitted regression line. Although the application of influence diagnostics may appear trivial in this example, the complex mapping from input to output space in hydrological models often precludes visual techniques, so that more formal approaches for the detection of influential points are required.

The prospect that a small number of data points can exert a very high influence on model performance motivates the more widespread implementation of influence diagnostics in hydrology, however applications have been few and recent. In the context of groundwater modelling, Yager (2004) found that models were highly sensitive to small changes in influential data. Foglia et al. (2007) used a series of case-deletion metrics and Cook's distance

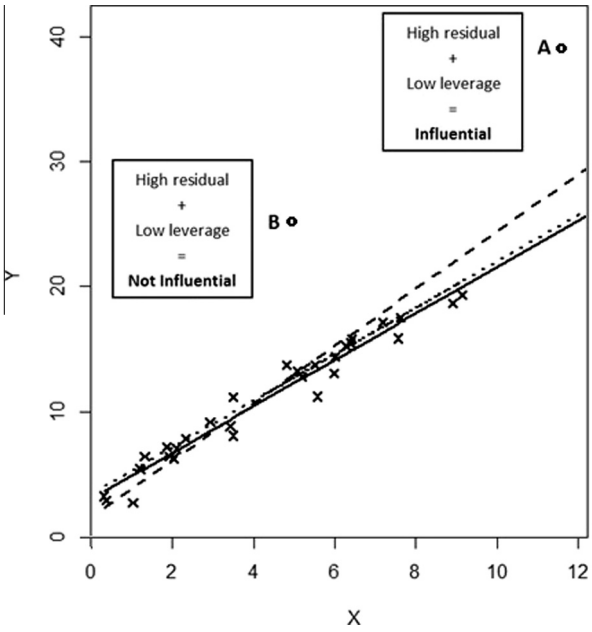

Fig. 1. A simple linear regression scatter plot illustrates the impact of a highly influential data point on the fitted model. The solid line is the prediction curve without point A or B in the calibration data; the broken prediction curve is with point B only excluded, as A is an observation that is both an outlier and a high leverage point; the dotted prediction curve is with point A only excluded, as B is an observation with the same residual as point A but with low leverage.

approaches on a groundwater model and found similar performance between the two metrics. Foglia et al. (2009) applied linear Cook's distance as part of a suite of diagnostics to a short time series of 37 daily observations in the rainfall-runoff model TOPKAPI and found that some of the low flow observations during small precipitation events were more important than anticipated. Legates and McCabe (1999) discuss the oversensitivity to outliers of correlation based goodness-of-fit measures used in hydrological models and recommend that additional evaluation measures should supplement calibration. Berthet et al. (2010) found a quadratic criterion to be influenced by a very small number of time steps characterised with high runoff variation. Perrin et al. (2007) assess the impact of the quantity and quality of streamflow data on parameter calibration and model robustness and show that a subset of influential points from a larger dataset are sufficient to obtain robust estimates. Singh and Bárdossy (2012) pre-process hydrological data using depth functions to identify unusual events and investigate the calibration of the model with only this set of critical data to assess if the subset has enough information to identify model parameters. Each of these studies contributes towards the more widespread use of influence assessment, however a comprehensive assessment of the influence of individual data points in the context of hydrological model predictions and parameters is still lacking.

The goal of this paper is to evaluate the use of influence diagnostics in the context of common hydrological modelling case studies: stage/discharge rating curve model and a conceptual hydrological model. Case-deletion, linear and nonlinear Cook's distance will be compared in terms of performance and computational run times. Tailored statistics that are suitable for hydrological model applications will be developed for measuring the effect of data points on the model parameters, performance and/or predictions. This analysis will identify the extent to which the model predictions are influenced by a small number of data points – thereby evaluating the information content of data points

and the benefits of including influence diagnostics as a standard tool in the process of hydrological model calibration.

The remainder of this paper is structured as follows. Section 2 outlines the methodology of the various approaches to quantifying influence. Section 3 introduces the case studies. Section 4 applies the numerical case-deletion diagnostics, applies the analytical Cook's distance diagnostics, provides a hydrological interpretation of the influence diagnostics, and evaluates the computational demand of the two classes of diagnostics. Finally, Section 5 discusses the importance of understanding the influence of data on hydrological predictions and the advantages and disadvantages of the numerical and analytical approaches.

2. Methods for assessing the influence of individual observations

2.1. Case-deletion influential point detection diagnostics

The case-deletion approach is widely used in the statistical literature to assess the impact of a deleted observation on the estimated parameters (Cook, 1977; Ross, 1987; Chen et al., 2012). It consists of evaluating the effect of excluding observations on the fitted model parameters, predictive performance and/or predictions and does not make strong assumptions on the model structure. As a continuous time series of rainfall inputs is required in hydrological modelling the case-deletion metrics consider the impact of masking an output data point from the objective function and therefore are calculated from model output only. Implementations of the approach differ largely in how the effect of the omitted observations is measured, and in this research we propose a number of measures that are specifically tailored to hydrological modelling applications.

2.1.1. Influence on the model predictions

Consider the following representation of a hydrological model:

$$\boldsymbol{Y} = h(\boldsymbol{\theta}, \boldsymbol{X}) + \boldsymbol{\varepsilon} \tag{1}$$

where $h()$ represents the hydrological model, $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ is an $n \times 2$ vector of observed inputs (such as precipitation and evapotranspiration), $\boldsymbol{Y}$ is an $n \times 1$ vector of observed responses (streamflow), and $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}$ is a $p \times 1$ vector of model parameters, and $\boldsymbol{\varepsilon}$ is a residual error model. Calibration of the hydrological model proceeds by finding a set of parameters ($\widehat{\boldsymbol{\theta}}$) such that some measure of the distance between the observed responses $\boldsymbol{Y}$ and the predicted response $\widehat{\boldsymbol{Y}} = h(\widehat{\boldsymbol{\theta}}, \boldsymbol{X})$ is minimised. For the case studies used in this paper, the residual model is assumed to normally distributed, $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma_\varepsilon^2)$, which results in a standard least squares objection function for model calibration.

Influence can be quantified by comparing predictions from a model with the whole data calibrated parameters ($\widehat{\boldsymbol{\theta}}$), and the predictions using parameters estimated from censoring the ith data point ($\widehat{\boldsymbol{\theta}}^{-i}$) in the objective function used in model calibration. Any model prediction can be compared in this way; for example the mean prediction relative change can be calculated using:

$$\text{Relative Change } (\%) = \frac{mean(\widehat{\boldsymbol{Y}}) - mean(\widehat{\boldsymbol{Y}}^{-i})}{mean(\widehat{\boldsymbol{Y}})} \times 100 \tag{2}$$

Other metrics such as predictions of the median, minimum and maximum flows (or any other quantile of the flow duration curve) can be defined similarly. The choice of metric(s) should be based on the intended modelling objective. In this paper we have chosen to focus on the mean and maximum predictions, to illustrate the impact of influence on the average predictions and extreme predictions, which are commonly metrics used in hydrological modelling (e.g. flood risk).

2.1.2. Influence on the model parameters

An alternative influence measure is based on the distance between parameter vectors $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^{-i}$. Given that hydrological model parameters often vary over different scales and can be highly correlated, we use the Mahalanobis distance (MD) (Mahalanobis, 1936) as a measure of the distance measure between the two parameter sets:

$$\text{MD}_i = \sqrt{(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^{-i})^{\mathrm{T}} \boldsymbol{C}^{-1} (\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^{-i})} \tag{3}$$

where $\boldsymbol{C}$ is the parameter covariance matrix. A bivariate example of the Mahalanobis distance is illustrated in Fig. 2, for each $\widehat{\boldsymbol{\theta}}^{-i}$ for $i = 1, \ldots, n$. The covariance matrix is estimated based on all n parameter sets, and thus the Mahalanobis distance for point i should be viewed as a measure of the influence of that data point relative to the remaining data points, rather than as an absolute measure of influence.

The metric given in Eq. (3) assumes that the joint distribution of parameters can be described by a multivariate Gaussian distribution; this may not always be appropriate, with extensions to the Mahalanobis distance beyond the Gaussian distribution given in Ekstrom (2011).

2.1.3. Influence on the model predictive performance

The influence of a data point on model performance can be quantified by considering the change in the objective function based on the whole data calibrated parameters $\widehat{\boldsymbol{\theta}}$ and the case-deletion parameters $\widehat{\boldsymbol{\theta}}^{-i}$. We use the term "objective function displacement" (OFD) as general measure of the difference in predictive performance due to including/excluding individual data points.

The concept of OFD is illustrated in Fig. 3. If $\text{OF}(\widehat{\boldsymbol{\theta}})$ and $\text{OF}(\widehat{\boldsymbol{\theta}}^{-i})$ are both evaluated over the entire time series $\boldsymbol{Y}$ then we can never

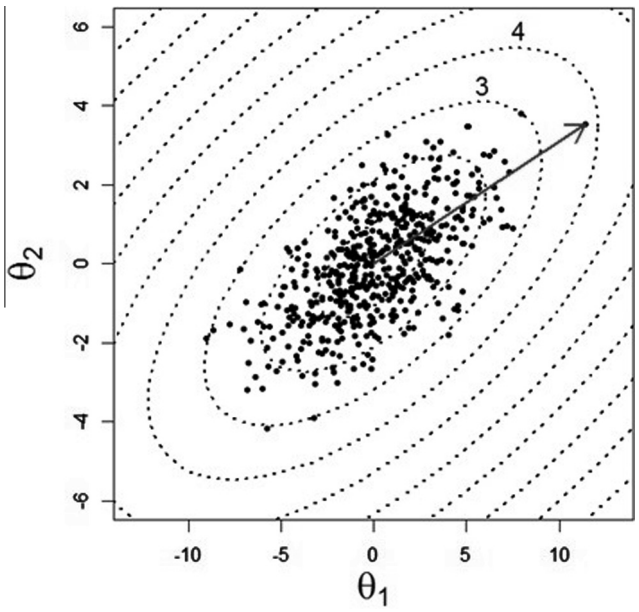

Fig. 2. Example of Mahalanobis distance in two dimensions. The origin is the parameter set obtained by calibrating to the full calibration dataset (i.e. $\widehat{\boldsymbol{\theta}}$), and the contours represent equal Mahalanobis distance from the origin. The highlighted point has a Mahalanobis distance of 3.8 from the origin (bivariate standard error of 3.8).

expect $OF(\widehat{\boldsymbol{\theta}}^{-i})$ to outperform $OF(\widehat{\boldsymbol{\theta}})$ as the parameters $\widehat{\boldsymbol{\theta}}^{-i}$ are calibrated on a different data set to $\widehat{\boldsymbol{\theta}}$. Therefore to isolate the influence of an observation on the remaining observations we consider the OFD evaluated with the ith case excluded:

$$OFD_i = \left|\left(OF_{-i}(\widehat{\boldsymbol{\theta}}^{-i}) - OF_{-i}(\widehat{\boldsymbol{\theta}})\right)\right| \tag{4}$$

Likelihood based methods for model calibration are commonly applied in hydrology (Westra et al., 2014; Evin et al., 2013; Renard et al., 2010) and the OFD is similar to the likelihood displacement (LD) (Cook and Weisberg, 1982) which can be used when using likelihood-based methods for model calibration. Let $L(\boldsymbol{\theta})$ denote the log-likelihood function. Then the likelihood displacement is defined as:

$$LD_i = 2\left\{L_{-i}\left(\widehat{\boldsymbol{\theta}}^{-i}\right) - L_{-i}(\widehat{\boldsymbol{\theta}})\right\} \tag{5}$$

with the form of the LD defined to be analogous to the residual deviance, which is minus twice its log-likelihood (Hastie et al., 2009).

In this study we choose to use the OFD (Eq. (4)) instead of LD in (Eq. (5)) to ease interpretability.

2.2. Analytical influential point detection diagnostics

The case-deletion method is computationally expensive, requiring $n+1$ model calibrations: n calibrations with one data point removed in each calibration, and one calibration with all data points included. Cook (1977) developed a distance metric that enabled the estimation of the influence of individual points using only a single full calibration. In this research we consider a linear and nonlinear version of this metric.

2.2.1. Linear Cook's distance

Cook's distance (CD) is calculated by accounting for both the leverage of the input data and the residual between the observed and fitted response. A point that is far from the centre of the input range will typically have high leverage, and a higher value of leverage means that the observation will have a greater influence on the model parameters or predictions. In linear regression the vector of predicted values $\widehat{\boldsymbol{Y}}$ is calculated from:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{\theta}} \tag{6}$$

where $\boldsymbol{H}$ is the hat matrix

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}} \tag{7}$$

The leverage value h_i of an observation corresponds to the ith diagonal element of the $n \times n$ $\boldsymbol{H}$ matrix and is independent of the model fit. The matrix $\boldsymbol{H}$ has the following properties: the leverage values are constrained by $0 \leqslant h_{ii} \leqslant 1$, and the sum of the diagonal elements is equal to the number of parameters in the model(p) so that the average leverage value is p/n (Hoaglin and Welsch, 1978; Stuart et al., 2004).

A point with both a large residual and high leverage exerts influence on the regression coefficients, in the sense that if the observation is removed, the parameters change considerably. Cook's distance can be calculated from the standardised model residual r_{Si} and leverage h_i (Fox and Weisberg, 2011):

$$r_i = Y_i - \widehat{Y}_i \tag{8}$$

$$r_{Si} = \frac{(r_i)}{\widehat{\sigma_\varepsilon}\sqrt{1-h_i}} \tag{9}$$

$$\text{Cook's distance}_i = \frac{{r_{Si}}^2}{p} x \frac{h_i}{1-h_i} \tag{10}$$

where $\widehat{\sigma}_\varepsilon^2$ is the calibrated residual error variance, from Eq. (1). Because CD is based on the residuals obtained from one model calibration and the leverage obtained from matrix multiplication, it is computationally far cheaper to calculate than case-deletion.

High leverage is commonly defined as $h_i > 2(p/n)$, a high standardised residual is commonly defined as greater than 2, and a highly influential point is defined $CD_i > 1$ (Stuart et al., 2004; Fox and Weisberg, 2011; Hoaglin and Welsch, 1978).

The different components of CD are illustrated in Fig. 4 using the same data as in Fig. 1. The model fit (Fig. 4a), residuals (Fig. 4b), leverage (Fig. 4c) and CD (Fig. 4d) are all presented, and illustrates that CD accounts for both the residual and the leverage. For example, point 'c' has both a high residual and high leverage value, and therefore has high influence. Point 'b' also has a high residual, however its influence is much lower due to its low leverage value. Point 'a' has high leverage but a low residual and therefore low influence.

2.2.2. Nonlinear Cook's distance

In linear models, the definition of leverage only depends on the observed input data, while in nonlinear models it is dependent on the local sensitivity of the model to small perturbations in model parameters (St. Laurent and Cook, 1992). We use Jacobian leverage, a special case of the generalised leverage developed by St. Laurent and Cook (1993), to construct a second-order approximation to the nonlinear response function, where the nonlinear leverage is defined as h^{nl} equal to the diagonal elements of the matrix $\widehat{\boldsymbol{J}}$

$$\widehat{\boldsymbol{J}} = \widehat{\boldsymbol{V}}\left(\widehat{\boldsymbol{V}}^{\mathrm{T}}\widehat{\boldsymbol{V}} - [\widehat{\boldsymbol{e}}][\widehat{\boldsymbol{W}}]\right)^{-1}\widehat{\boldsymbol{V}}^{\mathrm{T}} \tag{11}$$

Here $\widehat{\boldsymbol{V}}$ is the $n \times p$ matrix with ith row $\partial f_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, describing the effect of the perturbation of the parameters on the model predictions, $\widehat{\boldsymbol{W}}$ is the $n \times p \times p$ array with n elements $\partial^2 f_i(\boldsymbol{\theta})/\partial\theta_j\partial\theta_k$ of the $p \times p$ hessian matrix $\widehat{\boldsymbol{W}}_i$ of $h_i(\boldsymbol{\theta}, \boldsymbol{X})$, and $\widehat{\boldsymbol{e}}$ is the $n \times 1$ vector of fitted model residuals.

Analytical derivatives are typically not available for lumped hydrological models, and therefore estimates of $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{W}}$ can be obtained via finite difference numerical approximation and/or automatic differentiation (Nocedal and Wright, 2006) which is a standard practice in hydrological modelling (Abdulla et al., 1999; Martinez and Gupta, 2011; Mein and Brown, 1978; Vandewiele et al., 1992; Williams and Yeh, 1983). We use finite difference approximations with a Richardson extrapolation following the procedure of Nocedal and Wright (2006).

Substituting h^{nl} into (10) allows us to calculate nonlinear CD which may be more suitable for models with a nonlinear

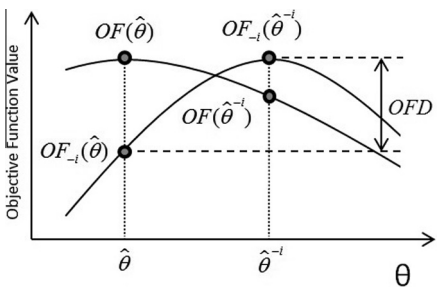

Fig. 3. Defining the objective function displacement (OFD). Model calibration involves finding the parameter set that maximises (or, analogously, minimise) the objective function value. Any variation in parameters about the optimal set would produce decay in model performance. The OFD is defined as the difference in objective function value when applying parameter sets $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^{-i}$ to the calibrated data with the ith point removed.

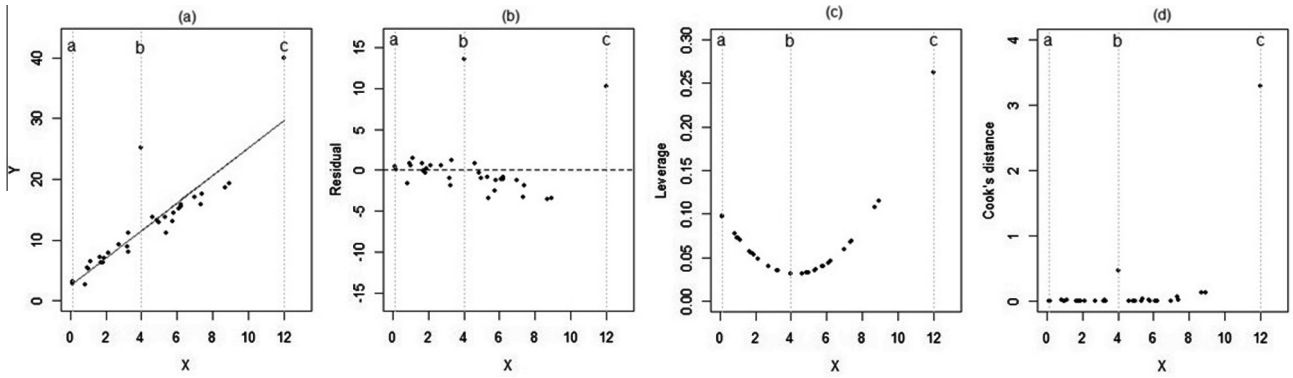

Fig. 4. Example application of Cook's distance in a simple linear model. (a) a scatter plot generated from $Y = 2X + 3$ with added Gaussian noise and the fitted curve, (b) raw model residuals against X (c), linear leverage against X, and (d) Cook's distance against X.

relationship between the predictor and the response. St. Laurent and Cook (1992) show that in nonlinear regression models the constraint $0 \leqslant h_i^{nl} \leqslant 1$ no longer holds, and cases where $h_i^{nl} > 1$ are defined as "superleverage". Therefore, the nonlinear CD metric can have a greater magnitude than the linear CD calculated from the same data.

3. Case studies

To evaluate influence diagnostics in the context of hydrological modelling we consider three case studies. Firstly a rating curve model with a short time series and parsimonious model structure is used to demonstrate concepts of influence, and secondly the methods are applied to an ephemeral and a humid catchment with the conceptual hydrological model GR4J (Perrin et al., 2003) to demonstrate the method for a typical hydrological model calibration problem. The case studies are introduced here and the results from applying the influence diagnostics to these case studies are described in Section 4.

3.1. Case study: Rating curve model

A stage/discharge rating curve model was selected due to its simple model structure and short time series to enable visual evaluation of influence diagnostics for individual observations. Site chosen was the experiment catchment, Mahurangi College (Woods et al., 2001), located in the Northland region 50 km north of Auckland, New Zealand. The catchment area is 46 km^2, with mean annual rainfall of 1600 mm, runoff of 860 mm and pan evaporation of 1310 mm. There are 27 years of 15 min interval streamflow data, and 24 stage/discharge gaugings. The rating curve model was a two-part piecewise power function:

$$Y = \begin{cases} \theta_1 X^{\theta_2} & X < 1.5 \\ \theta_1 1.5^{(\theta_2-\theta_3)} X^{\theta_3} & X \geqslant 1.5 \end{cases} \quad (12)$$

where X is river stage (m), and Y is the river discharge (m^3/s).

Calibration was performed using a standard least squares likelihood objective function optimised using the shuffle complex evolution (SCE) algorithm (Duan et al., 1992, 1994). To reduce the computation burden when undertaking re-calibration for each of the case-deletion points, the SCE algorithm was seeded using bounds of ±5% of the optimal parameters from the full data calibration.

The model calibrated with the full data set obtained a NSE of 0.79 and a prediction bias of 0.09 m^3/s.

3.2. Case study: Conceptual rainfall-runoff model

The GR4J model was selected due to its parsimonious model structure and its good performance across a wide range of catchment conditions (Perrin et al., 2003). GR4J has four calibration parameters: the production store capacity (θ_1, units of mm), the groundwater exchange coefficient (θ_2, units of mm), the one day-ahead maximum capacity of the routing store (θ_3, units of mm), and the time base of the unit hydrograph (θ_4, units of days).

To facilitate visual inspection of the results, the hydrological time series was restricted to two years with an additional one-year warmup period. GR4J was calibrated using the same procedure as the rating curve model.

3.2.1. Ephemeral Scott Creek catchment

The Scott Creek (South Australia) has an area of 29 km^2 and experiences a mean annual rainfall of 992 mm, median annual potential evapotranspiration (PET) of 1600 mm, and mean annual runoff of 147 mm. The catchment has a semi-arid climate with a winter-dominated rainfall regime, due to the low runoff coefficient of 0.14, is classified as ephemeral. The Scott Creek catchment was selected as the GR4J model as it has previously been successfully used by Westra et al. (2014). The model calibrated with the data set obtained an NSE of 0.79 and a prediction bias of 0.09 mm/day.

3.2.2. Humid French Broad River catchment

The French Broad River (North Carolina, USA) was selected from the MOPEX data set (Duan et al., 2006) as a humid catchment (runoff coefficient of 0.57) to contrast the ephemeral Scott Creek catchment. French Broad River has an area of 2448 km^2, mean annual rainfall of 1413 mm, and mean annual runoff of 800 mm. The model calibrated with the data set obtained an NSE of 0.86 and a prediction bias of 0.02 mm/day.

4. Results: Application of influence diagnostics to case studies

We first evaluate the influence of individual observations by quantifying the impact of case-deletion on the model predictions and parameters. We then apply linear and nonlinear Cook's distance (CD) to the data sets and explore the contributions of the observation's leverage and residual to CD. We compare the linear and nonlinear CD metrics using the case-deletion OFD as a baseline, and then review all three methods in terms of their computational demand and the information they convey about each data point's influence.

In order to understand the relationships between the influence diagnostics we have highlighted specific points in the case study data sets. For the rating curve case study, points 20 and 23 are

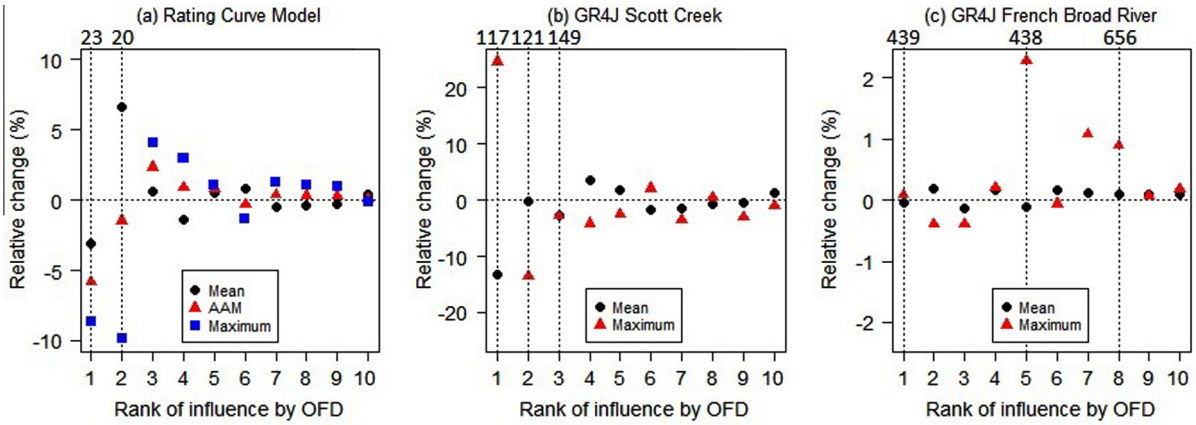

Fig. 5. Impact of removing the top 10 most influential data points on model predictions.

highlighted because they are the most influential in terms of all metrics. For GR4J Scott Creek, the two most influential points, day 117 and 121, were highlighted, However, as these days were from the same event the third most influential point, day 149, was also highlighted. For GR4J French Broad River, the two highest flow points, day 439 and 438, were highlighted. In addition, point 656 was highlighted because it changed from being identified as highly influential to having little or no influence depending on the metric used.

4.1. Case-deletion: quantifying the influence of observations on calibration

We apply the case-deletion approach to the case studies to explore the influence of observations on the mean and maximum predictions and model parameter shifts.

4.1.1. Individual observation influence on model predictions

To highlight the impact of individual points, on hydrological predictions, the relative change in mean and maximum predictions (and average annual maximum for the rating curve case study) for the ten most influential points (identified based on the OFD) are plotted for the three case studies in Fig. 5.

For the rating curve model the most influential point (point 23) changes the predicted mean flow by 6.7%. As rating curve models are often used to predict streamflow values greater than the highest streamflow gauging, and to illustrate the impact on extrapolated streamflow predictions, we consider two extrapolated values of the extreme predicted streamflow; (1) Average of the annual maximum (AAM) streamflow based the 27 years of streamflow estimated by the rating curve (corresponding to a river stage of 2.86 m) (2) Absolute maximum streamflow based on the highest estimated streamflow for the 27 years, (corresponding to a river stage of 4.2 m). The most influential point in terms of OFD (point 23) changed the AAM flow by −5.9% and absolute maximum flow by −8.6%. To illustrate the influence of points on the predicted rating curve model, the full set of case-deletion prediction curves is presented in Fig. 6. Here we see a range of 7.40 m^3/s (∼9%) for the average annual maximum streamflow prediction, and 35.6 m^3/s (∼13%) for the absolute maximum streamflow prediction. Influence is increased when extrapolating the rating curve beyond the maximum observed – this type of extrapolation is common in practise (Kuczera, 1996; Leonard et al., 2014).

For GR4J Scott Creek, the impact of influential points is high, with the most influential point (day 117) changing the predicted mean and maximum flow by −13.3% and 25.0%, respectively. For GR4J French Broad River the influence of individual points is not as high as the other case studies. The most influential point (day 439) changes the mean and maximum predictions by only 0.2% and −0.8%, respectively. Further analysis is presented in Section 5.1, to interpret and understand the reasons for the differences in the magnitude of the influence for these case studies.

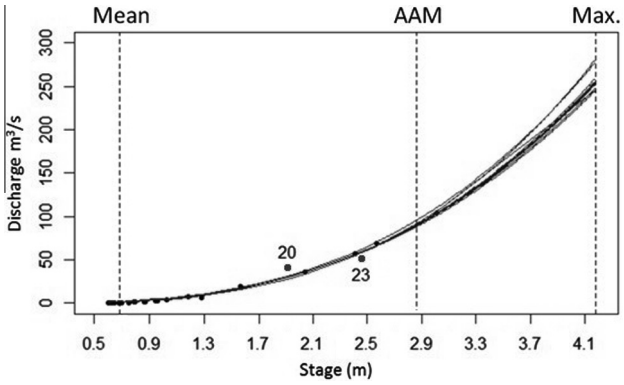

Fig. 6. Influence of data on the extrapolation of rating curve results. The vertical dashed lines show the mean prediction stage, average annual maximum (AAM) prediction stage and absolute maximum prediction stage. The two most influential points are highlighted.

4.1.2. Individual observation influence on fitted parameters

To illustrate the influence of individual data points on the model parameters, we show example pairwise bivariate plots of the model parameters for the case studies (Fig. 7). While the plots only show pairwise relationships, the Mahalanobis distance is based on the full p-dimensional distance of each $\widehat{\theta}^{-i}$ from $\widehat{\theta}$.

Fig. 7 shows the exclusion of a single observation can produce a significant change on calibrated model parameters. For the rating curve model, day 20 is clearly the most influential, with a Mahalanobis distance of 4.1, inducing a relative change of 10% in θ_2, the power parameter of the rating curve model. For GR4J Scott Creek, the impact on the parameters of the influential points is higher, with day 117 having a Mahalanobis distance of 26.9, inducing a relative change of 35% in θ_1 (production store capacity) and θ_2 (groundwater exchange coefficient). For GR4J French Broad river case study day 439 produces a Mahalanobis distance of 19.3, however, the relative change in parameters is only small, with a maximum relative change of 3% for θ_1 and θ_2.

4.2. Linear and nonlinear Cook's distance

The following sections will assess the computationally cheaper CD approaches to identify points and compares against the case-deletion OFD influence diagnostic.

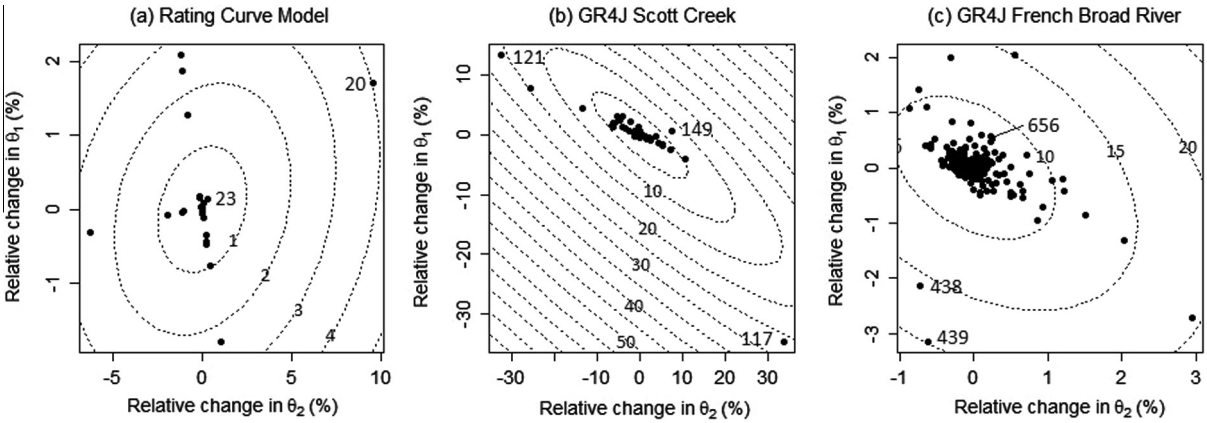

Fig. 7. Example parameter scatter plots. The axes correspond to shifts in the parameters from the exclusion of a single observation in model calibration and the contour lines represent regions with equal Mahalanobis distance from the origin. Scatterplots for other parameter combinations showed similar trends.

4.2.1. Linear Cook's distance

Residuals and leverage both contribute to the CD and the identification of influential points. Fig. 8 plots contours of linear CD (Eq. (10)) for the standardised residuals (Eq. (9)) and linear leverage (Eq. (7)) for each case study. The CD contours show that both a high magnitude residual and a high magnitude leverage are required to achieve a high magnitude CD. This is a practical demonstration of the principles demonstrated in the linear model in Fig. 4.

For the rating curve model, point 23 has the highest linear CD value, followed by point 20, although the maximum linear CD is relatively low at 0.014. These points are also the two points with the highest Mahalanobis distance, although by that metric point 20 has a higher influence than point 23. For GR4J Scott Creek, the linear CD values are far higher than the rating curve model. Day 117 is the most influential, with a linear CD of 1.4. followed by day 149, with a linear CD of 0.17. For GR4J French Broad River, the linear CD is lower than GR4J Scott Creek, with day 656 having the highest linear CD of 0.03. Note the most influential point from linear CD (day 656) is not consistent with the most influential points identified by the previous metrics (Mahalanobis distance, OFD, relative change in mean and maximum predictions).

4.2.2. Nonlinear Cook's distance

To assess if the nonlinear nature of conceptual rainfall-runoff models affects the identification of influential points, we compare the linear CD metric obtained from linear leverage with nonlinear CD obtained using nonlinear Jacobian leverage.

The nonlinear CD values are superimposed over the linear CD values in Fig. 8. In general, for all three case studies the nonlinear leverage increases the magnitude of nonlinear CD relative to linear CD, and can lead to identification of influential points that are more consistent with the case deletion metrics.

In the rating curve model and GR4J Scott Creek case studies the nonlinear CD considerably increases compared with linear CD. For GR4J Scott Creek, the most influential point (point 117) increases from a linear CD of 1.4 to a nonlinear CD of 13.6, with several other points, having a significant increase from linear to nonlinear CD. For GR4J French Broad River case study, the use of nonlinear CD actually changes the identification of the most influential points, compared with linear CD. Day 656 was the most influential for linear CD, but with nonlinear CD, day 439 has a significant increase in nonlinear CD, although it is still relatively low magnitude of 0.14. The most influential point of day 439 from nonlinear CD is also consistent with the case-deletion metrics (Mahalanobis distance, OFD, relative change in mean and maximum predictions).

4.2.3. Comparison with objective function displacement

The linear and nonlinear CD values are compared against OFD in Fig. 9. The Spearman rank correlation coefficient (Sp) was used to numerically compare the three measures of influence. For GR4J

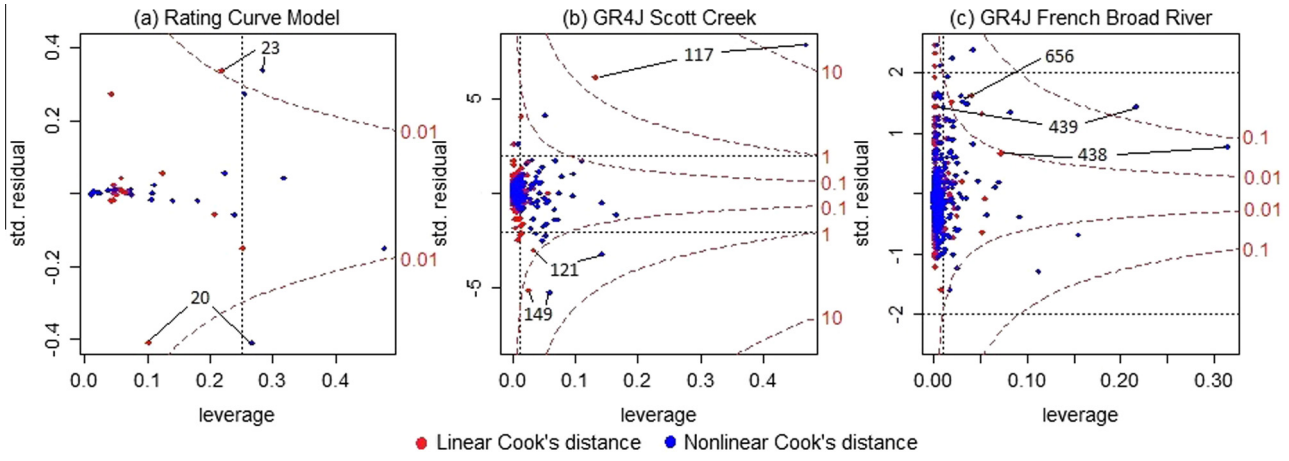

Fig. 8. Components of Cook's distance. Linear cooks distance points are in red, while nonlinear are blue. The red broken lines represent regions with equal Cook's distance. The vertical dotted line is drawn at $2p/n$ indicating points with high leverage, and the horizontal dotted lines are drawn at ±2 indicating points with a large standardised residual.

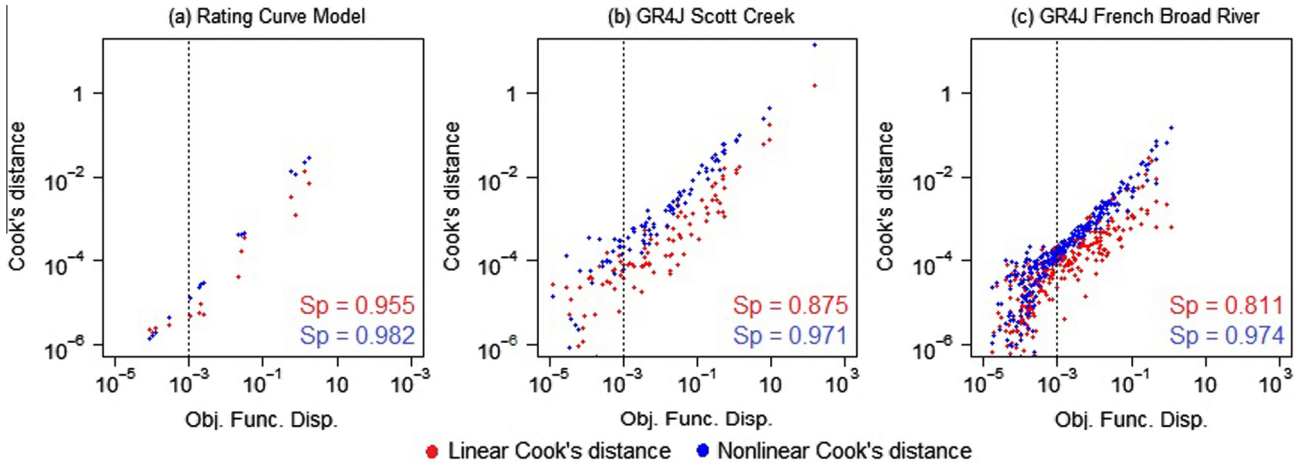

Fig. 9. Cook's distance (red) and nonlinear Cook's distance (blue) plotted against the case-deletion OFD. Spearman ranking coefficients (Sp) are for above a OFD threshold of 1×10^{-3} shown with the vertical dotted line.

case studies all points with an OFD below the SCE algorithm convergence tolerance of 10^{-5} were removed.

Including nonlinearity in the CD metric increases the consistency of the identified influential points with the case-deletion OFD. For the rating curve model, the OFD, linear and nonlinear CD are all in good agreement (Sp > 0.98), indicating similar influential points. For the most influential points (OFD > 10^{-3}) using the nonlinear CD, the Spearman correlation increases from 0.875 to 0.971 in Scott Creek and 0.811 to 0.974 in French Broad River. The improvement from the use of nonlinear CD is further discussed in Section 5.3.

4.3. Relationship between hydrological data and influence diagnostics

The hydrological observed and predicted data for the GR4J case studies is shown in Fig. 10. For GR4J Scott Creek we see the most influential day (117) corresponds to the highest rainfall and streamflow, while the second (day 121) and third (day 149) most influential points are also relatively high rainfall and streamflow. For French Broad River, the most influential point (day 439) from nonlinear CD, MD and has the highest flow, while the most influential point from linear CD (day 656) has relatively high flow. This suggests that highly influential points are also high flow values.

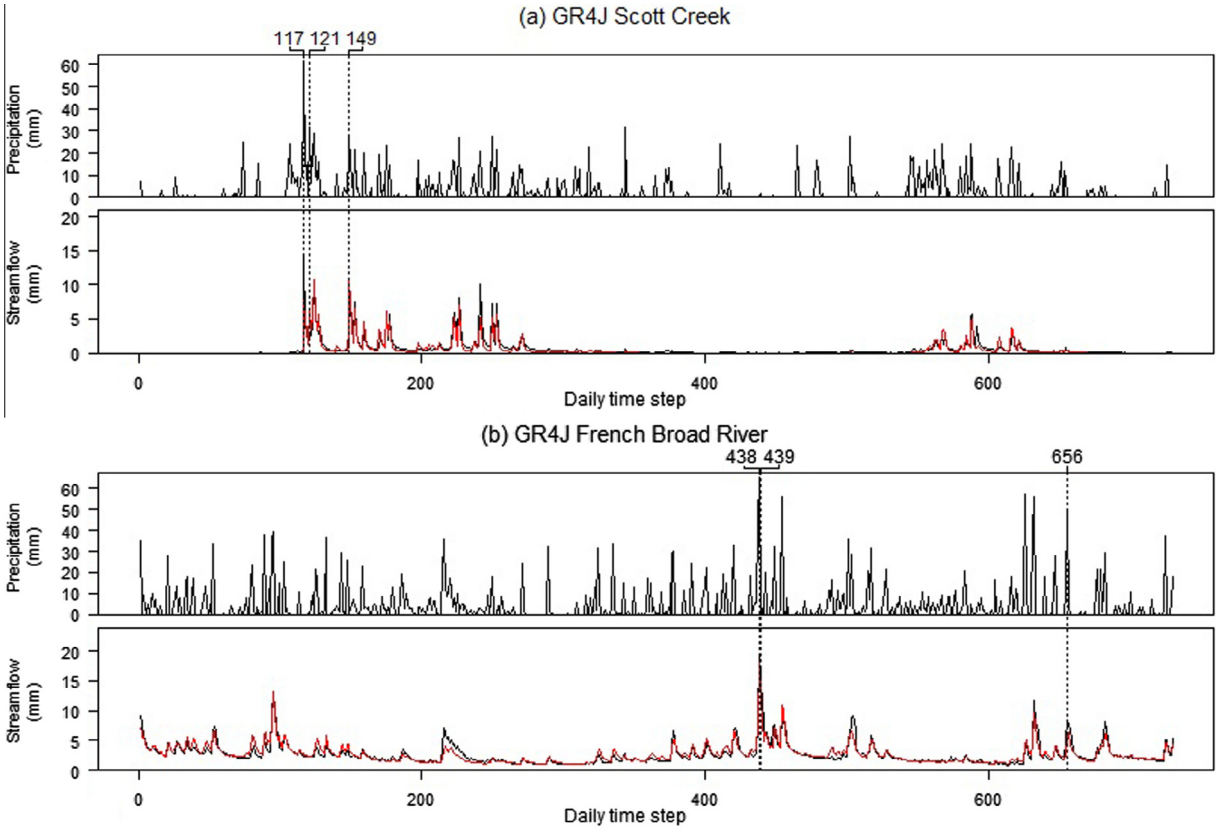

Fig. 10. Observed precipitation and observed (black) and predicted (red) streamflow for the GR4J case studies.

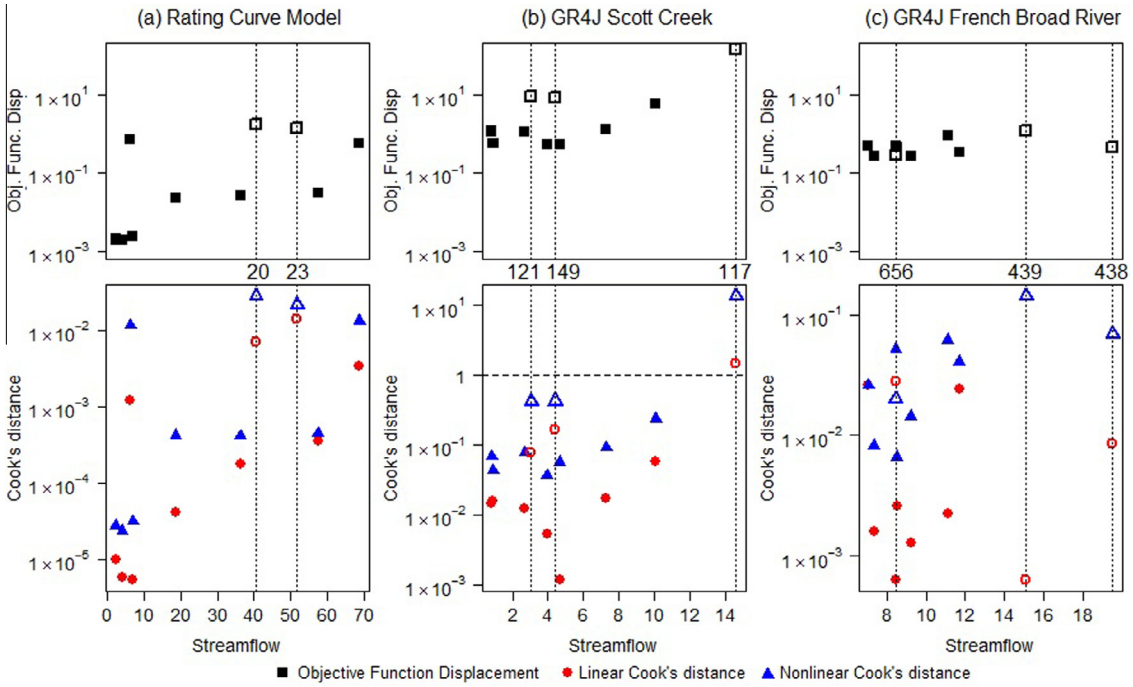

Fig. 11. Comparison of influence diagnostics against observed streamflow for the top 10 influential points ranked by OFD. A Cook's distance value of 1 is shown with a horizontal broken line, and a selection of the influential points are identified with vertical dotted lines. The hollow points are those corresponding to the vertical dotted lines.

High flow points are not always the most influential as shown in Fig. 11 where we compare the top 10 most influential points (ranked using OFD) with the observed streamflow data. We see that there is not a direct one-to-one relationship between the observed streamflow value and the rank of the influence. Similar to that seen in Fig. 9, we also see that the nonlinear CD better identifies the rank of influential points compared to the OFD than the linear CD measure.

For the rating curve case study neither of the most influential points (20 and 23) are the highest streamflow. In the GR4J Scott Creek, the most influential point is the highest streamflow, however the 2nd and 3rd most influential points, correspond to far lower streamflow values than the highest streamflow. In the GR4J French Broad River, the most influential point is the 2nd highest streamflow according to OFD and the highest streamflow (day 438) is only the 5th most influential according to OFD. Interestingly the, linear CD does not identify point 439 as influential and instead identifies point 656, the 5th highest streamflow.

Time series of precipitation, streamflow, Mahalanobis distance, OFD, linear and nonlinear CD for the three case studies are included in the supplementary material.

4.4. Computational demand of influence diagnostics

The computational demand of the influence diagnostics are summarised in Table 1 for a general case, and for 10 and 30 years of daily data. We assume a smaller number of runs for each case-deletion re-calibration as the optimisation can be seeded close to the optima from the whole data set to reduce the computational burden.

Case-deletion is the most intuitive approach for influence assessment but also the most computationally intensive as it requires $n+1$ calibration runs (~3660000 runs for 10 years). The case studies described in this study use relatively short periods of records (e.g. two years for the GR4J calibration), and the models are relatively parsimonious compared to many other hydrological models. Nevertheless the run-time for applying case-deletion to the GR4J model was about 14 h on a 2.90 GHz processor. In practical applications we would expect a time series length (n) much larger than the two years used in the GR4J case study and we may choose to apply a less parsimonious conceptual model with more than four parameters (p). The resultant increase in both p and n would increase the computational demand of each model calibration, leading to further increases in computational demand.

Table 1
Summary of computational demands of influence diagnostics. Computational demands are based on a time series of length n and a model with $p=4$ parameters which requires 10000 model runs for initial calibration and an additional 1000 runs for each case-deletion calibration.

Influence diagnostic	General computational demand	10 years of daily data	30 years of daily data
Case-deletion	Repeating model calibration for the $n+1$ case-deletion scenarios	3660000 runs	10960000 runs
Linear Cook's distance	Single calibration + linear matrix algebra	10000 runs	10000 runs
Nonlinear Cook's distance	Single calibration + $2p$ and $2(p \times p)$ model runs + linear matrix algebra	10040 runs	10040 runs

Regardless of the size of the calibration data set and hydrological model structural complexity linear CD requires only one model calibration followed by the application of linear matrix algebra. Nonlinear CD has the additional computational demand of calculating the finite difference approximations for the Jacobian and Hessian matrix to account for nonlinearity in the response surface, however both metrics are much more computationally tractable compared with case-deletion. For 10 years of data, the computational speed of the analytical influence diagnostic is 300 times faster than case deletion, while for 30 years, it is 1000 times faster than case deletion.

5. Discussion

This section discusses the importance of understanding the influence of data on hydrological predictions, and the advantages and disadvantages of the numerical and analytical influence diagnostics.

5.1. Importance of understanding the influence of data on hydrological model predictions

Hydrological models are an essential tool to aid policy makers and businesses to make decisions about water allocations and to design infrastructure. However, the influence of calibration data is seldom assessed during the model development process. Influence assessment based on CD is a computationally cheap addition to the existing model development process, with significant benefits in developing an understanding of which data points have the greatest influence on the calibrated model.

The potential impact of influential data in hydrological model calibration was illustrated in Section 4.1, where the Scott Creek case study showed that the exclusion of individual daily observations can have a substantial impact on model predictions. The magnitude of the prediction changes from removing one point (25% influence on the maximum prediction and 13% on the mean prediction) highlights the importance of including influence assessment in the existing hydrological model calibration framework, as changes in model predictions of this magnitude could have a large impact infrastructure design and future flood and drought risk. As Cook's distance is a function of both the residual and the leverage of the data point influence of data on uncertainty in predictions and posterior parameter distributions requires further theoretical development and will be explored in future research.

In the humid French Broad River case study the influence of data was less than the ephemeral Scott Creek catchment. In Scott Creek the smaller number of high rainfall and flow points (compared with the more consistent rainfall and flow events in French Broad River), produced data points with slightly higher leverage – see Fig. 8. However, the primary contributor to the higher influence in Scott Creek, is the far higher value for the standardised residuals of >5 for the most influential point (Fig. 8b). Assuming the data are of high quality, this indicates deficiencies in the ability of the GR4J model structure to reproduce the flow regime in this ephemeral catchment – see further discussion in Westra et al. (2014). This indication that the GR4J model is very susceptible to highly influential flows in ephemeral catchments will need to be investigated with a large range of varying flow regime case studies.

Once influential data is identified the modeller will need to make a decision as to whether they want to retain the identified data in the calibration set. Despite advances in methods of data collection, input data error is still common in hydrological modelling. Rain gauges may not be representative of total catchment rainfall, and the integrity of individual daily measurements may be impacted by power outages and human reading errors. Potential evapotranspiration are often estimated from pan evaporation measurements which may not be representative of the whole catchment. Flow measurements are often estimated from a rating curve and so may also be prone to human reading error, or stage measurements may be outside the range of measurements used to develop the rating curve. Scrutinising the entire time series for erroneous data is resource intensive. An advantage of the use of influence diagnostics, is that they can identify a smaller number of highly influential points that have the biggest impact on calibration, which can be scrutinised for data errors, thereby reducing effort and costs.

If the highly influential data is erroneous, removing or correcting the data point from the calibration set is likely to be the best course of action to mitigate its effects on model predictions. However, not all influential data is likely to be attributable to measurement errors. As the CD is a function of the residual and leverage, poor model predictions can have high influence, even if there is no error in the data. In such cases, rather than removing the influential data, the modeller may choose to retain the influential points if the data is characteristic of the intended model application. For example, if high flows were found to be influential this would be beneficial to model predictions in the case where the modelling objective is peak flow estimation.

Finally, in addition to the decision of whether to remove or retain data from the calibration set, knowledge of which points are likely to be influential can have value for experimental design. For example, identifying the types of data that are most influential in model calibration may allow for the targeted collection of informative data, rather than investing resources in collecting data that will have a minimal effect on the model predictions. Furthermore, for situations where influential data arises due to model miss-specification (i.e., structural error, which often arises due to the highly abstracted nature of hydrological models relative to the system being investigated) and/or issues with the calibration strategy, it may be possible to use influence measures as a diagnostic to compare the performance of multiple alternative hydrological model structures.

5.2. Advantages of influence diagnostics over a visual assessment of the time series

In some cases influential data points can be identified by visual inspection of the hydrological time series alone, i.e. the highest streamflow/rainfall values are the most influential. This study showed this was not always the case – Section 4.3 showed there was not a clear one-to-one relationship between observed streamflow value and OFD in the three case studies (Fig. 11). In hydrological model calibration visual inspection is more difficult due to longer time series, nonlinear model response, and correlation in predictions due to model storage. It is likely that influence diagnostic will become even more valuable with increasing hydrological model complexity, multi-catchment studies (e.g. >200 in Coron et al., 2012), longer calibration data sets, more complex objective functions such as the generalised likelihood (Schoups and Vrugt, 2010), and consideration of the effects of persistence and heteroscedasticity (Evin et al., 2013). Both the numerical and analytical influence diagnostics considered in this study quantify influence regardless of the magnitude of data or the complexity of model calibration strategy and can provide insights in cases where calibration information is otherwise limited.

5.3. Advantages and disadvantages of the different classes of influence diagnostics

The numerical case-deletion and analytical Cook's distance (CD) influence diagnostics vary in the way they quantify influence and in the complexity and computational demand of their application. Here we compare the advantages and disadvantages of the two classes of methods and discuss the impact of including nonlinearity in the CD formulation.

Case-deletion can be used on wide range of measures that quantify influence on model parameters, performance and/or predictions. This flexibility to tailor the case-deletion diagnostics to specific hydrological measures makes them an intuitive method of influence detection. Case-deletion has the additional advantage in that it makes no strong assumptions on hydrological model or assumed residual error model structure and can therefore be applied

to any case study regardless of complexity. The major drawback with case-deletion is that for complex model calibration strategies combined with long calibration data time series, the use of high performance computing becomes essential if the method is to be feasibly applied in hydrological modelling. An additional concern is that individual calibration runs may find different local optima, potentially leading to significant differences in the optimised parameter sets. The response surface geometry of many hydrological models is often highly complex, with common features including curving ridges, microscale and macroscale discontinuities and multiple optima (Duan et al., 1992). The possibility of calibrating each case-deletion dataset to local optima may cause misleading results where two case-deletion calibrated parameter sets appear different from each other even if the data points have low influence on the actual model calibration. If the case-deletion metrics are to be implemented then the modeller should ensure that parameter optimisation approach can robustly handle complex response surface.

CD is a computationally cheap addition to existing hydrological model diagnostics and can be applied to any model calibration. The major drawback of CD is that its ranking of influence may be less interpretable when compared to ranking using the hydrologically orientated case-deletion diagnostics. For the three case studies considered there were significant benefits from including nonlinear leverage in the CD metric. The ranking of the influential points using nonlinear CD was far more consistent with case deletion metrics than linear CD. Fig. 9 showed using nonlinear CD (cf linear CD) increased the spearman ranking coefficient with OFD from 0.875 to 0.971 in the ephemeral Scott Creek case study, and from 0.811 to 0.974 in the humid French Broad River case study. Additionally, in Fig. 11 we see that in French Broad River the linear CD metric identifies point 656 instead of 439, the most influential by case-deletion OFD and nonlinear CD.

The impact of model nonlinearity in hydrological models was highlighted by Duan et al. (1992) who discuss nonlinear parameter interaction in hydrological models, and Kavetski and Kuczera (2007) who highlight the difficulties in model calibration due to the highly complex and nonlinear nature of conceptual hydrological models. Similarly insights were found in this study – that influence diagnostics need to take into account the nonlinearity in of hydrological models. Further case studies will be needed to verify this conclusion.

A challenge with using CD is that it is unknown how it will be affected by more complex objective function formulations, and further research will need to be undertaken to understand how the choice of objective function impacts the diagnostic. As CD is a measure of the leverage and residual at a single time step it will not account for the "memory" or time lag of hydrological models, with storage errors propagating across multiple consecutive time steps. The consistency of the results of the nonlinear CD with the case deletion metrics indicated this was not a major problem, however more case studies are needed to verify this and may prompt the incorporation of time lag into CD in future studies.

6. Conclusions

Influential point detection diagnostics are not commonly used in hydrological modelling despite being regularly applied in the computational statistics literature since the introduction of Cook's distance (CD) in linear regression models in Cook (1977). This paper evaluates the application of numerical case-deletion and analytical CD influence diagnostics in the context of three case studies: a stage/discharge rating curve model and the conceptual hydrological model GR4J over two contrasting datasets. We found that individual influential data points can have a substantial impact on model predictions in the ephemeral Scott Creek catchment but a relatively small influence on predictions in the humid French Broad River catchment. Assuming the data are of high quality this indicates deficiencies in the ability of the GR4J model structure to reproduce the flow regime in the ephemeral catchment, however this will need to be investigated with a large range of varying flow regime case studies.

Case-deletion approaches are capable of ranking the influence of individual data points using a wide variety of metrics, including the impact on:

1. model parameters, either individually or through an aggregated measure such as the Mahalanobis distance;
2. model performance, based on the difference between objective function values when a point is either included or excluded from the calibration set; and
3. model predictions, such as the mean, minimum or maximum flow or any other metric of interest.

Limitations of the case-deletion approaches include the significant run times associated with the large number of model re-calibrations, and the possibility of finding local rather than global optima in each re-calibration.

In contrast, methods based on CD are much more computationally tractable (300–1000 times faster) compared with case deletion and are less sensitive to the presence of local optima. However, specific values of CD are more difficult to interpret compared to the hydrologically orientated case-deletion measures. This study found that nonlinear CD provided a ranking of the influential points that was more consistent with the case deletion metrics compared with linear CD. This is likely due to the nonlinear structure of hydrological models. Further case studies will investigate the efficacy of nonlinear CD.

Regardless of the specific choice of metric, it is clear that influential point detection diagnostics can provide important insights into the mechanics of hydrological model calibration, as well as providing a better understanding of the impact of individual data points on the calibrated model. Visual assessment was not a reliable method of influence assessment as there was no direct relationship between the most influential data and the highest observed streamflows. Application of influence diagnostics will allow the modeller to make informed decisions about including or excluding influential data and fine tune their adopted calibration strategy to ensure robust predictions for the intended model application.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jhydrol.2015.05.047.

References

Abdulla, F.A., Lettenmaier, D.P., Liang, X., 1999. Estimation of the ARNO model baseflow parameters using daily streamflow data. J. Hydrol. 222 (1–4), 37–54.
Berthet, L., Andréassian, V., Perrin, C., Loumagne, C., 2010. How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. Hydrol. Sci. J. 55 (6), 1063–1073.
Beven, K., Westerberg, I., 2011. On red herrings and real herrings: disinformation and information in hydrological inference. Hydrol. Process. 25 (10), 1676–1680.
Chen, X.D., Tang, N.S., Wang, X.R., 2012. Local influence analysis for semiparametric reproductive dispersion nonlinear models. Acta Math. Appl. Sin. – Engl. 28 (1), 75–90.
Cohn, T.A., England, J.F., Berenbrock, C.E., Mason, R.R., Stedinger, J.R., Lamontagne, J.R., 2013. A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. Water Resour. Res. 49 (8), 5047–5058.
Cook, R.D., 1977. Detection of influential observation in linear-regression. Technometrics 19 (1), 15–18.
Cook, R.D., 1979. Influential observations in linear-regression. J. Am. Stat. Assoc. 74 (365), 169–174.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Linear Regression. Chapman and Hall, New York.
Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., Hendrickx, F., 2012. Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. Water Resour. Res. 48 (5), W05552.
Duan, Q.Y., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28 (4), 1015–1031.
Duan, Q.Y., Sorooshian, S., Gupta, V.K., 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. J. Hydrol. 158 (3–4), 265–284.
Duan, Q.Y. et al., 2006. Model Parameter Estimation Experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J. Hydrol. 320 (1–2), 3–17.
Duan, Q.Y., Ajami, N.K., Gao, X.G., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resour. 30 (5), 1371–1386.
Ekstrom, O. (2011). Mahalanobis' distance beyond normal distributions. UCLA Statistics.
Evin, G., Kavetski, D., Thyer, M., Kuczera, G., 2013. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resour. Res. 49 (7), 4518–4524.
Foglia, L., Mehl, S.W., Hill, M.C., Perona, P., Burlando, P., 2007. Testing alternative ground water models using cross-validation and other methods. Ground Water 45 (5), 627–641.
Foglia, L., Hill, M.C., Mehl, S.W., Burlando, P., 2009. Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. Water Resour. Res. 45.
Fox, J., Weisberg, S., 2011. An R Companion to Applied Regression, second ed. Sage Publications Inc.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. CRC Press.
Hastie, T., Tibshirani, R., Friedman, J., 2009. Elements of Statistical Learning: Data Mining, Inference and Prediction, second ed., New York.
Hoaglin, Welsch, 1978. The hat matrix in regression and ANOVA. Am. Stat. 32, 17–22.
Kavetski, D., Kuczera, G., 2007. Model smoothing strategies to remove microscale discontinuities and spurious secondary optima im objective functions in hydrological calibration. Water Resour. Res. 43 (3).
Kuczera, G., 1996. Correlated rating curve error in flood frequency inference. Water Resour. Res. 32 (7), 2119–2127.
Lamontagne, J., Stedinger, J., Cohn, T., Barth, N., 2013. Robust national flood frequency guidelines: what is an outlier? World Environ. Water Resour. Congress 2013, 2454–2466.
Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Leonard, J., Mietton, M., Najib, H., Gourbesville, P., 2014. Rating curve modelling with Manning's equation to manage instability and improve extrapolation. Hydrol. Sci. J. 45 (5), 739–750.
Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. National Inst. Sci., India 2 (1), 49–55.
Martinez, G.F., Gupta, H.V., 2011. Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. Water Resour. Res. 47.
Mein, R.G., Brown, B.M., 1978. Sensitivity of optimized parameters in watershed models. Water Resour. Res. 14 (2), 299–303.
Nocedal, J., Wright, S.J., 2006. Numerical Optimization. Springer.
Perrin, C., Michel, C., Andreassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279 (1–4), 275–289.
Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., Mathevet, T., 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. Hydrol. Sci. J. 52 (1), 131–151.
Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. Water Resour. Res. 46.
Ross, W.H., 1987. The geometry of case deletion and the assessment of influence in nonlinear-regression. Can. J. Stat. 15 (2), 91–103.
Russo, C.M., Paula, G.A., Aoki, R., 2009. Influence diagnostics in nonlinear mixed-effects elliptical models. Comput. Stat. Data Anal. 53 (12), 4143–4156.
Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46 (10), W10531.
Singh, S.K., Bárdossy, A., 2012. Calibration of hydrological models on hydrologically unusual events. Adv. Water Resour. 38, 81–91.
St. Laurent, R.T., Cook, R.D., 1992. Leverage and superleverage in nonlinear-regression. J. Am. Stat. Assoc. 87 (420), 985–990.
St. Laurent, R.T., Cook, R.D., 1993. Leverage, local influence and curvature in nonlinear regression. Biometrika Trust 80 (1), 99–106.
Stuart, A., Ord, K., Arnold, S., 2004. Kendall's Advanced Theory of Statistics: Volume 2A: Classical Inference and the Linear Model, sixth ed., Wiley.
Thomas, W., Cook, R.D., 1989. Assessing influence on regression-coefficients in generalized linear-models. Biometrika 76 (4), 741–749.
Vandewiele, G.L., Xu, C.Y., Larwin, N., 1992. Methodology and comparative-study of monthly water-balance models in Belgium, China and Burma. J. Hydrol. 134 (1–4), 315–347.
Westra, S., Thyer, M., Leonard, M., Kavetski, D., Lambert, M., 2014. A strategy for diagnosing and interpreting hydrological model nonstationarity. Water Resour. Res. 50 (6), 5090–5113.
Williams, B.J., Yeh, W.W.G., 1983. Parameter-estimation in rainfall runoff models. J. Hydrol. 63 (3–4), 373–393.
Woods, R.A., Grayson, R.B., Western, A.W., Duncan, M.J., Wilson, D.J., Young, R.I., Ibbitt, R.P., Henderson, R.D., McMahon, T.A., 2001. Experimental design and initial results from the Mahurangi River Variability Experiment: MARVEX. Observ. Model. Land Surface Hydrol. Process., 201–213
Yager, R.M., 2004. Effects of model sensitivity and nonlinearity on nonlinear regression of ground water flow. Ground Water 42 (3), 390–400.

Bibliography

Abdulla, F. A., D. P. Lettenmaier, and X. Liang (1999), Estimation of the ARNO model baseflow parameters using daily streamflow data, Journal of Hydrology, 222(1-4), 37-54.

Berthet, L., V. Andréassian, C. Perrin, and C. Loumagne (2010), How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion, Hydrological Sciences Journal, 55(6), 1063-1073.

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

Beven, K., and I. Westerberg (2011), On red herrings and real herrings: disinformation and information in hydrological inference, Hydrological Processes, 25(10), 1676-1680.

Chen, X. D., N. S. Tang, and X. R. Wang (2012), Local influence analysis for semiparametric reproductive dispersion nonlinear models, Acta Math Appl Sin-E, 28(1), 75-90.

Cheng, Q.-B., X. Chen, C.-Y. Xu, C. Reinhardt-Imjela, and A. Schulte (2014), Improvement and comparison of likelihood functions for model calibration and parameter uncertainty analysis within a Markov chain Monte Carlo scheme, Journal of Hydrology, 519, Part B, 2202-2214.

Chiew, F. H. S., M. C. Peel, and A. W. Western (2002), Application and testing of the simple rainfall-runoff model SIMHYD, 335-367 pp.

Cohn, T. A., J. F. England, C. E. Berenbrock, R. R. Mason, J. R. Stedinger, and J. R. Lamontagne (2013), A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series, Water Resources Research, 49(8), 5047-5058.

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

Cook, R. D. (1979), Influential Observations in Linear-Regression, J Am Stat Assoc, 74(365), 169-174.

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), W05552.
Das, S. (2008), Generalized linear models and beyond: An innovative approach from Bayesian perspective, ProQuest.
Del Giudice, D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann (2013), Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, Hydrol. Earth Syst. Sci., 17(10), 4209-4225.
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.
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.
Duan, Q. Y., N. K. Ajami, X. G. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources, 30(5), 1371-1386.
Duan, Q. Y., et al. (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320(1-2), 3-17.
Ekstrom, O. (2011), Mahalanobis' distance beyond normal distibutions, UCLA Statistics.
Evin, G., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resources Research, 49(7), 4518-4524.
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), W11510.
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(6), W06427.
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.
Hastie, T., and R. Tibshirani (1990), Generalized Additive Models, CRC Press.
Hastie, T., R. Tibshirani, and J. Friedman (2009), Elements of Statistical Learning: Data Mining, Inference and Prediction (Second Edition), New York.
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.
Hsu, K. L., H. V. Gupta, and S. Sorooshian (1995), Artificial Neural-Network Modeling of the Rainfall-Runoff Process, Water Resources Research, 31(10), 2517-2530.
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), W03411.
Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resources Research, 42(3), W03407.
Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, Journal of Hydrology, 320(1-2), 173-186.
Kuczera, G. (1996), Correlated rating curve error in flood frequency inference, Water Resources Research, 32(7), 2119-2127.
Lamontagne, J., J. Stedinger, T. Cohn, and N. Barth (2013), Robust National Flood Frequency Guidelines: What Is an Outlier?. : , World Environmental and Water Resources Congress 2013, 2454-2466.
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), W06428.
Legates, D. R., and G. J. McCabe (1999), Evaluating the use of ”goodness-of-fit”

measures in hydrologic and hydroclimatic model validation Water resources Research, 35(1), 233-241.
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.
Leonard, J., M. Mietton, H. Najib, and P. Gourbesville (2014), Rating curve modelling with Manning's equation to manage instability and improve extrapolation, Hydrological Sciences Journal, 45(5), 739-750.
Li, C., H. Wang, J. Liu, D.-h. Yan, F.-l. Yu, and L. Zhang (2010), Effect of calibration data series length on performance and optimal parameters of hydrological model, Water Science and Engineering, 3(4), 378-393.
Li, M., Q. J. Wang, J. C. Bennett, and D. E. Robertson (2015), A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts, Hydrol. Earth Syst. Sci., 19(1), 1-15.
Mahalanobis, P. C. (1936), On the generalized distance in statistics, Proceedings National Institute of Science, India, 2(1), 49-55.
Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, Water Resources Research, 47, W12540.
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, 53(3) 2199-2239.
Mein, R. G., and B. M. Brown (1978), Sensitivity of Optimized Parameters in Watershed Models, Water Resources Research, 14(2), 299-303.
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.
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.
Peel, M. C., B. L. Finlayson, and T. A. McMahon (2007), Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11(5), 1633-1644.
Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279(1-4), 275-289.
Perrin, C., L. Oudin, V. Andreassian, C. Rojas-Serna, C. Michel, and T. Mathevet

(2007), Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models, Hydrological Sciences Journal, 52(1), 131-151.
Petersen-Øverleir, A. (2004), Accounting for heteroscedasticity in rating curve estimates, Journal of Hydrology, 292(1-4), 173-181.
Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour Res, 46.
Rocha, A., and A. Simas (2011), Influence diagnostics in a general class of beta regression models, TEST, 20(1), 95-119.
Rödiger, T., S. Geyer, U. Mallast, R. Merz, P. Krause, C. Fischer, and C. Siebert (2014), Multi-response calibration of a conceptual hydrological model in the semiarid catchment of Wadi al Arab, Jordan, Journal of Hydrology, 509, 193-206.
Ross, W. H. (1987), The Geometry of Case Deletion and the Assessment of Influence in Nonlinear-Regression, Can J Stat, 15(2), 91-103.
Russo, C. M., G. A. Paula, and R. Aoki (2009), Influence diagnostics in nonlinear mixed-effects elliptical models, Comput Stat Data An, 53(12), 4143-4156.
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.
Singh, S. K., and A. Bárdossy (2012), Calibration of hydrological models on hydrologically unusual events, Advances in Water Resources, 38, 81-91.
Smith, T., L. Marshall, and A. Sharma (2015), Modeling residual hydrologic errors with Bayesian inference, Journal of Hydrology, 528, 29-37.
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
Stuart, A., K. Ord, and S. Arnold (2004), Kendall's Advanced Theory of Statistics: Volume 2A: Classical Inference and the Linear Model, 6th edition, Wiley.
Thomas, W., and R. D. Cook (1989), Assessing Influence on Regression-Coefficients in Generalized Linear-Models, Biometrika, 76(4), 741-749.
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.
Vandewiele, G. L., C. Y. Xu, and N. Larwin (1992), Methodology and Comparative-Study of Monthly Water-Balance Models in Belgium, China and Burma, Journal of Hydrology, 134(1-4), 315-347.
Vansteenkiste, T., M. Tavakoli, V. Ntegeka, F. De Smedt, O. Batelaan, F. Pereira, and P. Willems (2014), Intercomparison of hydrological model structures and calibration approaches in climate scenario impact projections, Journal of Hydrology, 519(PA), 743-755.
Vrugt, J. A., and C. J. F. Ter Braak (2011), DREAM(D): an adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems, Hydrol. Earth Syst. Sci., 15(12), 3701-3713.
Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrological Processes, 17(2), 455-476.
Wang, Q. J., D. L. Shrestha, D. E. Robertson, and P. Pokhrel (2012), A log-sinh transformation for data normalization and variance stabilization, Water Resources Research, 48(5), W05514.
Wei, B. C., Y. Q. Hu, and W. K. Fung (1998), Generalized leverage and its applications, Scandinavian Journal of Statistics, 25(1), 25-37.
Westra, S., M. Thyer, M. Leonard, D. Kavetski, and M. Lambert (2014), A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resources Research, 50(6), 5090-5113.
Williams, B. J., and W. W. G. Yeh (1983), Parameter-Estimation in Rainfall Runoff Models, Journal of Hydrology, 63(3-4), 373-393.
Woods, R. A., R. B. Grayson, A. W. Western, M. J. Duncan, D. J. Wilson, R. I. Young, R. P. Ibbitt, R. D. Henderson, and T. A. McMahon (2001), Experimental Design and Initial Results from the Mahurangi River Variability Experiment: MARVEX, Observations And Modeling Of Land Surface Hydrological Processes, pp. 201-213.
Wright, D. P., 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. P., M. Thyer, S. Westra, B. Renard, and D. McInerney (2017), A generalised approach for identifying influential data in hydrological modelling, Environmental Modelling & Software, In Review submitted February 2017.
Yager, R. M. (2004), Effects of model sensitivity and nonlinearity on nonlinear regression of ground water flow, Ground Water, 42(3), 390-400.