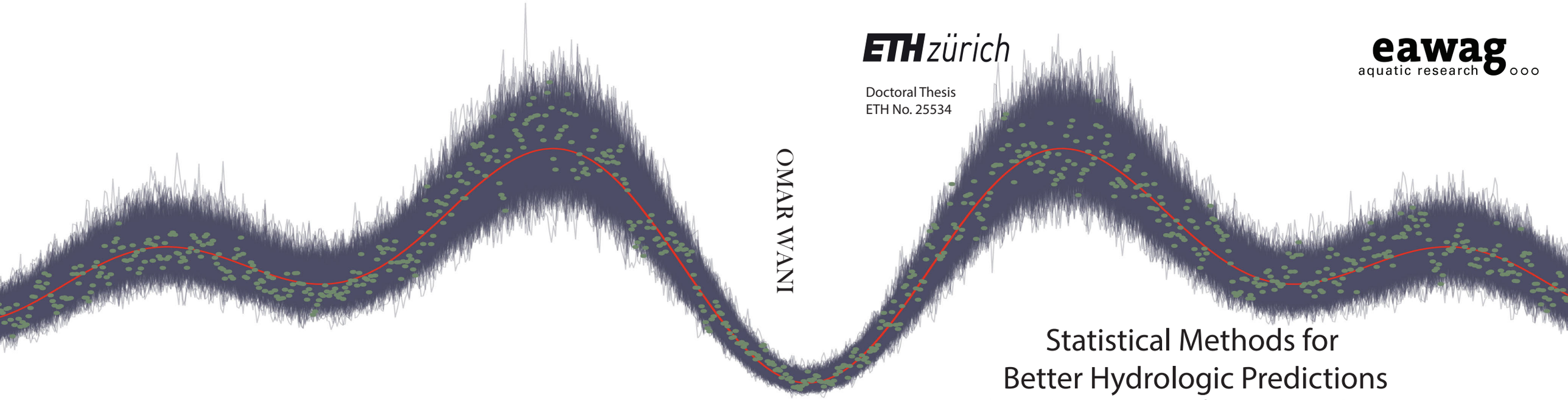


ETH zürich

Doctoral Thesis
ETH No. 25534

eawag
aquatic research

OMAR WANI

Statistical Methods for Better Hydrologic Predictions

Improving parameter and uncertainty estimation

IfU
Institute of
Environmental Engineering

A dissertation submitted to ETH Zürich
for the degree of
DOCTOR OF SCIENCES

QR code for Google Scholar

Statistical Methods for Better Hydrologic Predictions

DISS. ETH NO 25534

Statistical Methods for Better Hydrologic Predictions

Improving parameter and uncertainty estimation

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Omar Wani

M.Sc. Water Sc. & Eng., IHE Delft, the Netherlands
M.Sc. Hydro Sc. & Eng., TU Dresden, Germany

born 08.11.1989
citizen of India

accepted on the recommendation of

Prof. Dr. Max Maurer, examiner
Dr. Jörg Rieckermann, co-examiner
Andreas Scheidegger, co-examiner
Prof. Dr. Jim W. Hall, co-examiner

2018

Cover design by Omar Wani

Cover image: Navigating Uncertainty, 2018
Acrylic on canvas

The painting depicts the flow of a river—capturing the confluence of rainfall-runoff generated from natural and urban landscapes. The graphical overlay represents hydrologic forecasts and the associated uncertainties.

Painting – Sourabh Gupta
Conceptualization – Omar Wani

Statistical Methods for Better Hydrologic Predictions

Improving parameter and uncertainty estimation

For my parents, Shahida and Farooq

The true logic of this world is the calculus of probabilities.

– James Clerk Maxwell

Contents

Abstract

Hydrologic systems are complex—constituting spatially distributed and temporally variable properties and processes. Such systems are generally modelled mathematically after some simplifying assumptions, for example, by aggregating certain variable properties and neglecting many subprocesses. These model structure deficits result in a mismatch between the real and the modelled system response, which entails uncertain predictions. In addition to model structure deficits, the presence of errors in model inputs and in the observations of system response contributes to the uncertainty in the estimation of model parameters. For risk-based decision making and for statistical hypothesis testing using these models, their uncertainties need to be adequately and explicitly stated. Furthermore, attempts need to be made to reduce these uncertainties.

The goal of this thesis is to improve the probabilistic descriptions of hydrologic modelling errors and attain better constrained parameter estimates using such descriptions. To achieve this, several statistical methods are investigated and tailored for hydrologic applications. The research work starts with 1) the analysis of a data-driven method for reliable estimation of predictive uncertainty. It proceeds to 2) reduce uncertainty through improved parameter estimation. And then seeks 3) a more general mathematical description of error time series. Finally, it 4) discusses some philosophical aspects of *unbiased* parameter estimation, given the simplified mathematical descriptions used for complex hydrologic realities.

In part one (chapter 2 and 3), a conceptually simple and easy-to-implement data-driven method is proposed for predictive uncertainty estimation; its applicability is demonstrated for different forms of hydrologic forecasting. This method takes advantage of long time series of observations. It is shown that the performance of this technique is comparable to more sophisticated post-processor uncertainty estimators that employ fuzzy logic and non-linear regression for the same goal. Being a post processor, this method is computationally efficient, as it avoids Monte Carlo simulations. The technique is shown to work well for models with single as well as multiple output variables.

The second part of this work (chapter 4) investigates the reduction of uncertainty by incorporating complementary information in the modelling exercise. A specific method is proposed that can make use of censored and binary observations for the estimation of parameters. This part of the work goes further than just quantifying uncertainties and allows learning from observations. In order to make the uncertainties explicit, the method employs a state-of-the-art description of autoregressive and heteroscedastic error model used in hydrology and shows how to update the prior of the parameters to a posterior, using nonconventional observations. Given the promise of large quantities of data from Internet of Things and the availability of cheaper and more robust sensors, this method enables parameter estimation using such data within a formal Bayesian framework.

In part three (chapter 5), the nature of rainfall-runoff time series, and in turn the error time series, is further explored, disentangling the three main statistical properties a) autocorrelation b) heteroscedasticity and c) non-negativity. This examination seeks to make the probabilistic description of the hydrologic models more representative of the observation generating process. The proposed technique provides more flexibility in constructing probabilistic error models, with the potential for a more reliable parameter estimation compared to the additive Gaussian-process error model. Also, this description precludes the possibility of negative flow predictions.

The fourth part (chapter 6), as an attempt at painting the big picture before the concluding remarks, is a philosophical appraisal of uncertainty quantification for hydrologic models. It consists of a brief discussion on the epistemological paradigm that makes hydrology amenable to probabilistic statements. The discussion, being exploratory in nature, moves on to the utility of uncertainty analysis for both scientific inquiry and practical application in hydrology. It concludes with the expectations of the hydrologic community from such an analysis.

This research work was, in part, motivated by the hydrology community's demand for and enthusiasm towards proper uncertainty analysis tools. The problems of design and forecast in hydrology need solutions that are robust. This requires a framework that does not just rely on the deterministic model output as the *best guess*. Assigning accurate probabilities to the space of all possible outcomes enables such robust solutions. Beyond the technological considerations, there are many scientific questions still open in hydrology. An improved treatment of model uncertainties greatly facilitates finding reliable answers to these questions.

Zusammenfassung

Hydrologische Systeme sind komplex—sie bilden räumlich verteilte und zeitlich veränderliche Eigenschaften und Prozesse ab. Solche Systeme werden mathematisch unter einigen vereinfachenden Annahmen modelliert, beispielsweise durch das Aggregieren bestimmter Variableneigenschaften und das Vernachlässigen einiger Teilprozesse. Diese Defizite in der Modellstruktur führen zu einer Diskrepanz zwischen realem und modelliertem System, was zu Vorhersagen mit grösserer Unsicherheiten führt. Weiterhin trägt das Auftreten von Fehlern in den Input-Parametern sowie in den simulierten Resultaten zu Unsicherheiten bei der Schätzung von Modellparametern bei. Für eine risikobasierte Entscheidungsfindung sowie für statistische Hypothesentests unter Verwendung der Modelle müssen Unsicherheiten angemessen und explizit angegeben und langfristig reduziert werden.

Ziel dieser Arbeit ist es, die probabilistischen Beschreibungen von hydrologischen Modellierungsfehlern zu verbessern und dadurch bessere Parameterschützungen zu erzielen. Dazu werden verschiedene statistische Methoden untersucht und auf hydrologische Gebiete angewandt. Die Forschungsarbeit beginnt mit 1) der Analyse einer datengesteuerten Methode zur zuverlässigen Schätzung der Vorhersageunsicherheit, gefolgt von 2) der Verringerung dieser Unsicherheiten durch eine verbesserte Parameterschätzung. Eine 3) allgemeinere mathematische Formulierung von Fehlerzeitreihen wird vorgestellt. Schliesslich werden 4) einige philosophische Aspekte der unverzerrten Parameterschützung unter Berücksichtigung der vereinfachten mathematischen Beschreibungen, die in komplexen hydrologischen Modellen verwendet werden, diskutiert.

Im ersten Teil (Kapitel 2 und 3) wird eine konzeptionell einfache und leicht zu implementierende datengesteuerte Methode für Vorhersage von Unsicherheiten vorgeschlagen; die Anwendbarkeit wird für verschiedene Formen hydrologischer Vorhersagen demonstriert. Diese Methode nutzt lange Beobachtungszeitreihen. Es wird gezeigt, dass die Methode in ihrer Leistungsfähigkeit mit komplexeren Post-Prozessor-Unsicherheitsschätzern, die Fuzzy-Logik und nichtlineare Regression verwenden, vergleichbar ist. Als Postprozessor ist diese Methode recheneffizient, da Monte-Carlo-Simulationen vermieden werden. Die Methode funktioniert gut für Modelle mit Einzel- als auch Mehrfachausgangsvariablen.

Der zweite Teil dieser Arbeit (Kapitel 4) untersucht die Verringerung von Unsicherheiten durch Einbeziehung komplementärer Informationen in die Modellierung. Eine spezifische Methode wird vorgeschlagen, die zensierte und binäre Beobachtungen zur Schätzung von Parametern verwendet. Dieser Teil der Arbeit geht über die Quantifizierung von Unsicherheiten hinaus und erlaubt es, aus Beobachtungen zu lernen. Zur Verdeutlichung von Unsicherheiten, verwendet die Methode eine Beschreibung des in der Hydrologie üblicherweise angewandten autoregressiven und heteroskedastischen Fehlermodells und zeigt, wie sich A-priori-Wahrscheinlichkeiten an A-posteriori-Wahrscheinlichkeiten anhand nicht-konventioneller Beobachtungen anpassen lassen. Angesichts der zukünftigen Verfügbarkeit grosser Datenmengen aus dem Internet der Dinge sowie kostengünstiger und robusterer Sensoren, ermöglicht diese Methode eine Parameterschätzung unter Verwendung solcher Datenmengen in einem formalen Bayesschen Framework.

Im dritten Teil (Kapitel 5) wird die Art der Niederschlag-Abfluss-Zeitreihen und der Fehlerzeitreihen weiter untersucht, wobei die drei wichtigsten statistischen Eigenschaften a) Autokorrelation b)

Heteroskedastizität und c) Nicht-Negativität genauer betrachtet werden. Dadurch wird versucht, die probabilistische Beschreibung der hydrologischen Modelle repräsentativer für den Beobachtungserzeugungsprozess zu gestalten. Die vorgeschlagene Methode bietet mehr Flexibilität bei der Konstruktion von Wahrscheinlichkeitsfehlermodellen mit dem Potenzial für eine zuverlässigere Parameterschätzung im Vergleich zum additiven Gaussschen Prozessfehlermodell. Ausserdem schliesst diese Beschreibung die Möglichkeit negativer Strömungsvorhersagen aus.

Der vierte Teil (Kapitel 6) dient als Versuch, das Gesamtbild vor den Schlussbemerkungen aufzuzeigen, und umfasst eine philosophische Bewertung der Unsicherheitsquantifizierung für hydrologische Modelle. Es besteht zunächst aus einer kurzen Diskussion über das erkenntnistheoretische Paradigma, das die Hydrologie probabilistischen Aussagen zugänglich macht. Die explorative Diskussion geht auch auf den Nutzen der Unsicherheitsanalyse sowohl für die wissenschaftliche Forschung als auch für die praktische Anwendung in der Hydrologie ein. Sie schliesst mit den sich aus einer solchen Analyse ergebenden Erwartungen der hydrologischen Gemeinschaft.

Diese Forschungsarbeit wurde zum Teil durch das Interesse und die Begeisterung der hydrologischen Gemeinschaft ür geeignete Werkzeuge zur Unsicherheitsanalyse motiviert. Die Probleme der Konzeption und Prognose in der Hydrologie erfordern robuste Lösungen. Dies erfordert ein Framework, das nicht nur auf dem deterministischen Modell-Output als beste Schtzung beruht. Das Zuweisen von genauen Wahrscheinlichkeiten zum Raum aller möglichen Ergebnisse ermöglicht solche robusten Lösungen. Über die technologischen Überlegungen hinaus bleiben in der Hydrologie noch viele wissenschaftliche Fragen offen. Eine verbesserte Behandlung von Modellunsicherheiten erleichtert es, zuverlässige Antworten auf diese Fragen zu finden.

Acknowledgements

There are many people whom I am highly indebted to for making this PhD thesis possible. Many thanks to my direct supervisor, Jörg Rieckermann. He has been a great mentor, making sure that I got all the scientific and social support needed on this journey. His academic supervision, scientific ideas, and constant positivity—with those metaphorical pats on my back—kept me going through the ebb and flow of PhD research. Many thanks to my second supervisor, Andreas Scheidegger. This PhD benefitted immensely from his intellectual investments. He was always accessible and very friendly—it was a pleasure to have all those engaging conversations. He also risked going on holidays with me, even though I could have turned it into a statistics workshop. I didn't. I extend my sincere thanks to my professor-in-charge, Max Maurer, for taking extraordinary care of the personal and professional well-being of his students. Under his guidance, one is encouraged to think creatively and told not to be afraid of making mistakes in the process. Merci Vielmals for instilling such values in us. I am very thankful to Jim Hall for taking out time from his busy schedule for my thesis evaluation.

I thank the European Commission for the Marie Curie Fellowship. The funding provided a great opportunity to work together with scientists from across Europe and beyond. I thank the QUICS (Quantifying Uncertainty in Integrated Catchment Studies) team for creating and running a stimulating research project.

As someone working on the application of probability theory, I learned substantially from the papers of and the discussions with the physicists at Eawag, Peter Reichert and Carlo Albert. I am grateful to them both. I am very thankful to Dario Del Giudice for helping me transition into the topic of Bayesian analysis. I extend my gratitude to the researchers with whom many scientific collaborations ensued during the course of this PhD. Many thanks to Joost Beckers, Frank Blumensaat, Juan Pablo Carbajal, Francesca Cecinati, Austin Chadwick, Lennart Ehlers, Gabriel Espadas, Michael Lamb, Manoranjan Muthusamy, Jens Christian Refsgaard, Miguel Rico-Ramirez, Dimitri Solomatine, and Albrecht Weerts. I am personally thankful to my colleagues—former and current, from SWW, ENG, and SIAM—for providing a very welcoming and inspirational atmosphere at work. A special thanks goes to Philip Beutler, Sanda Dejanic, Sven Eggimann, Matthew Moy de Vitry, Lena Mutzner, Roni Penn, Mariane Schneider, and Dorothee Spuhler for all the fun we had. Many memorable experiences with them—while working, talking, hiking, joking etc.—made these four years much more than just an academic exercise. Additionally, I am very thankful to Canan Aglamaz, Lorenz Ammann, Peter Bach, Andy Disch, Natalia Duque, Ariane Eberhardt, Maike Gaertner, Bruno Hadengue, João Leitão, Liliane Manny, Abishek Narayan, Christoph Ort, Dyllon Randall, Laura Varga, and Kris Villez. Vielen Dank an Liliane für die Übersetzung meines Abstracts ins Deutsche. Apart from all the good times we had together, I thank all my colleagues from ENG for introducing me to many other fascinating aspects of urban water.

I would like to thank all my friends, from all around the world, for their support and encouragement. The list is long, and they know who they are. A shout out to Burooj Ghani and Faisal Wani for all the intellectually stimulating discussions and for being very important constants in my life. Also, many thanks to Candace Chow, Sourabh Gupta, and Azam Iftikhar for their help. I thank all my old and new flatmates here in Zürich for their pleasant company and for creating a very sociable living space.

I thank coffee, the beverage, for existing.

I would like to mention Kashmir here—my roots. Among other things, it inspired me to study hydrology. It has some of the most beautiful catchments in the world.

Finally, I would like to show my gratitude and love to my family. To Kubra, for being the nicest sister and for always taking my side. To my father, Farooq, who showed me, by example, the value of hard work and good routines. To my mother, Shahida, who, apart from teaching me algebra when I was a little boy, taught me the value of cheerfulness and kindness. Shukriya for your love and support.

Omar Wani
August, 2018

Chapter 1

Introduction

Author contributions: Omar Wani wrote the full chapter.

1 The hydrologic process—an apology

Approximately 2.5 percent of the global water budget constitutes of freshwater. And within that small fraction, it is estimated that just about 0.006 percent flows in rivers —the rest mainly constituting of polar ice, glaciers and groundwater (Trenberth et al., 2007). Notwithstanding their small volumetric ratio, the dynamical interaction of surface and channel flows with the biogeosphere is both substantial and vital. Flows in channels form and enable complex ecosystems (Abell et al., 2008). On geological timescales, these flows cause morphodynamic evolution —carrying huge amounts of sediment, sculpting new channel geometries and leading to delta formation (Ganti et al., 2016). Specifically within the human context, rainfall-runoff has been a major actor in shaping societies —influencing the location and structure of cities we live in and influencing their economics. Most of the early civilizations were built in close proximity to rivers (Maisels, 2003). Artificial channel networks came into existence to support irrigation. Similarly, building of drainage networks, considered one the greatest medical milestones (Ferriman, 2007), allowed for the conveyance of waste water and storm water away from human settlements, in most cases into the nearest rivers. Towards the end of the nineteenth century, hydroelectric plants, run by dammed streams, entered the scene of power generation (Chech, 2005).

From an engineering standpoint, it is therefore important to understand the functioning of such natural or engineered hydrologic systems, among other things, for seeking optimality in water resources management, water-related risk management, infrastructure design and operation (Eggimann et al., 2017; Werner et al., 2013). Overflowing drainage systems and swollen river systems annually cost billions of dollars in damages (Ward et al., 2013). Around 40 percent of the economic losses that were incurred from 1900 to 2015 due to natural catastrophes were ascribed to floods (Daniell et al., 2016). And with the changing climate, extreme weather events, which include both floods and droughts, are going to increase in frequency and intensity, depending on the geographic location (IPCC, 2014; Milly et al., 2008). Therefore, unlike Hardy's apology for pure mathematics (Hardy, 2012), we have to conclude that hydrology has to live with the 'burden' of being very urgent and very useful.

Going beyond utility, there is another prominent, a purely scientific, standpoint in researching such hydrologic systems. While the overarching causality that leads to the physical exchange of water, the water cycle, was already discovered in the seventeenth century (Nace, 1974), scientists are still busy trying to explain the myriad subprocesses that happen within this cycle. Understanding such subprocesses in detail may not always be needed for making operational forecasts, however a mechanistic understanding is important to draw causal inferences —to simply push the envelope of knowledge. With process variability at several temporal and spatial scales, the mathematical descriptions used to explain these hydrologic systems and understand their behaviour remain approximate. Additionally, system observations are not available at resolutions that can allow the incorporation of all system variability. This makes hydrologic modelling a work in progress.

2 The hydrologic model

Mathematical models are used to simulate real hydrologic systems to an acceptable degree of accuracy, depending on the task of the modelling exercise. The classical approach to hydrology, in the context of rainfall-runoff, has been to use the basic governing rules of flow for formulating a dependence of discharge on various input variables. The precipitation is generally one of the main input signals. Within a hydrologic model, the geometric properties of the catchment—like its elevation, land cover, soil conductivity etc—*act* on the precipitation signal and return the output as discharge. The governing equations that are usually employed in hydrology can be presented in some hierarchy, where the physics of the flow is captured with a decreasing degree of detail. Some selective didactic examples, with brief descriptions, are mentioned below.

One of the most descriptive ways to model the motion of water in catchments would be to use

Navier-Stokes equations. This set of nonlinear partial differential equations treats fluids as an incompressible continuum, and balances their viscous forces with the variable pressure term, using Newton's second law. Given their reliance directly on the first principles of classical motion, they are considered quite fundamental. Additionally, the conservation of mass is enforced through a continuity equation. One can compute the 3D spatial distribution of the velocity and pressure of a fluid and track the evolution of this field in time. However, this level of detail is usually not needed in hydrologic applications. A simplification of Navier-Stokes equations would be the shallow water equations, which assume a 2D variability of the velocity of the flow and changes in pressure are reflected through the changes in the height of the water column. In 1D these equations are called Saint-Venant equations. (In Eq. 1 and 2, Q is the discharge at time t for the cross sectional area A located at x, y is the depth and q is the lateral inflow. S_o and S_f are gravity and friction terms. These equation capture conservation of mass and momentum, respectively.)

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} + q = 0 \tag{1}$$

$$\left(\frac{1}{A}\right)\frac{\partial Q}{\partial t} + \left(\frac{1}{A}\right)\frac{\partial (Q^2/A)}{\partial x} + g\frac{\partial y}{\partial x} - g(S_o - S_f) = 0 \tag{2}$$

If we are interested in just the discharge coming out of a catchment area, given a certain rainfall input, the often called bucket or reservoir models provide an approximate description of the system. Output of the system is not only dependent on the aggregate rain that falls in a catchment, but also the way it falls. The bucket models consist of a system of ordinary differential equations, which capture the dependence of discharge on the temporal structure of the rain. Also, catchments are modelled as conceptual buckets that store water and give away more water as the storage increases. This mathematical construction provides us an input-output relationship between precipitation and discharge with components that have physical interpretability. (In Eq. 3, 4, and 5, Q and P are discharge and precipitation at the time t. K and a are model parameters. Whereas, S_1 and S_2 are storage variables.)

$$\frac{dS_1}{dt} = aP - \frac{S_1}{K} \tag{3}$$

$$\frac{dS_2}{dt} = \frac{S_1 - S_2}{K} \tag{4}$$

$$Q = \frac{S_2}{K} \tag{5}$$

The next in line simplification to model rainfall-runoff is the unit hydrograph convolution. It assumes that the response of the catchment to instantaneous rainfall events in the past adds up linearly to give the aggregate response. So summation, or in case of continuous variables integration, over time can yield the expected discharge downstream of a catchment. (In Eq. 6, Q and $P(t)$ are the discharge and precipitation at a time t. A and k are model parameters.)

$$Q = A\int_0^t P(t-\tau)ke^{-k\tau}d\tau \tag{6}$$

Many hydrologic software packages employ a combination of the models mentioned above, with some parts represented by conceptual or empirical equations, others by more physically descriptive equations. They also have modules for evapotranspiration, snow melt, infiltration, land use, ground water and flow routing in streams and drainage networks. This results in a highly parameterized overall model description (Butts and Graham, 2005; Werner et al., 2013).

3 Uncertainty in hydrologic models

3.1 Inverse and forward problem

In predictive hydrologic models, we plug the input variables in the model and get the output. This is the forward simulation, where the question being answered is: given a certain input, what is the expected output? However, to constrain the model outputs to our catchment conditions, we first need to determine its parameters, that vary from case to case. Learning about these parameters from observations of the system response is labelled as the inverse problem, where the question being answered is: given a certain input and output, what values of parameters could have produced it?

Deterministic paradigms: Most of the hydrologic modelling till the late twentieth century happened using deterministic equations. For a given set of inputs and parameters, a unique set of outputs is expected. The noise over and above the model, if the model is good enough, is neglected. All we are after is the most likely output that we expect from a system. The model calibration can then be carried out using certain optimization metrics like seeking the least squared error. To prevent fitting the model to the noise, many cross-validation techniques are employed.

Probabilistic paradigms: If we are not only interested in the "best guess" about parameter values and model output, but want to employ the whole distribution of the parameters and in turn obtain the whole distribution of model output, we start with a mathematical framework that can assign probabilities to our modelled system response.

3.2 Sources of uncertainty

While lumped conceptual models lose accuracy due to simplifying assumptions, models that describe the system in greater detail do not necessarily improve the predictive capability. Many other factors play a role. An attempt at adding more subprocesses to the system description also requires that the representation of those subprocesses is correct. Besides, more data is needed to feed these models. So there is a tendency to accumulate errors. Therefore, hydrologic models tend to have uncertainties associated with them irrespective of the detail with which they try to describe the hydrologic system.

The main sources of uncertainty then can be summed up as:

- Model structure deficits: These arise from the fact that the equations used to describe the hydrologic catchment are approximate.

- Observational errors: The measurement of the input, elevation and the system response itself suffer from systematic and random deviations. These deviations may be stationary in time, or show some dependence on external drivers.

- Parameter uncertainty: Given that many combinations of parameter values, the ones we have not measured, can produce a certain time series of observations, to a comparable degree, results in parameter uncertainty. Parameter uncertainty is usually a result of indeterminacy in the deterministic paradigm, i.e. given an algebraic or differential equation, there are n data points, depending on the nature of the equation, necessary to fully define the system. Having fewer observations then results in parameter uncertainty. In the probabilistic paradigm, parameter uncertainty is a consequence of the fact that different parameter values, once a model is defined, can explain the same observational data with varying probabilities because, as mentioned, observations have errors and models are imperfect. In an ideal world, for example, for fitting a line, two observations with no noise will uniquely define its parameters. However, the presence of random error allows us to fit many lines through the same data.

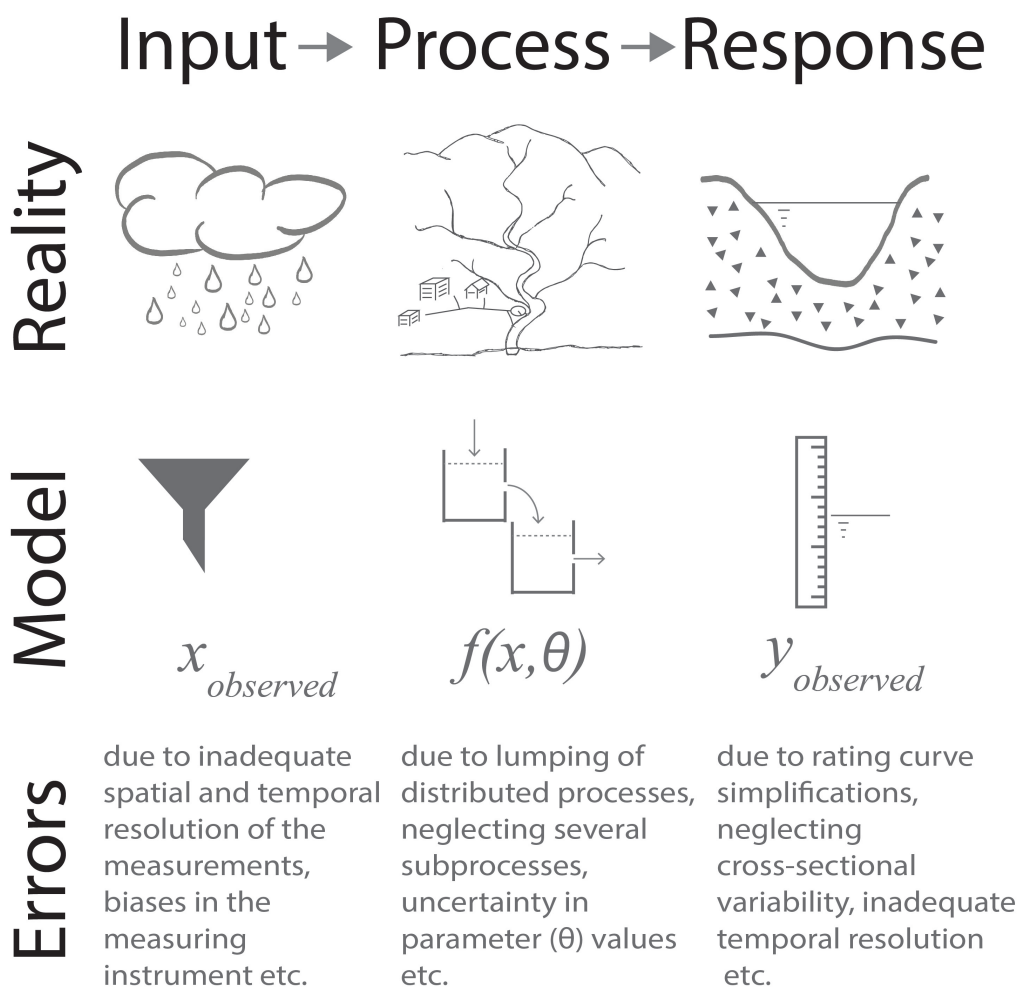


Figure 1: Sources of uncertainty in hydrologic models. (adapted from Kavetski, 2018).

3.3 The value of making uncertainty explicit

Hydrology generally concerns itself with explaining processes happening out in the environment, which many a times do not lend themselves to repeatable controlled experiments. As mentioned before, given the conceptualizations and approximations that enter into the equations, we expect a degree of mismatch between our model output and the true system response. Some contribution to this mismatch can be from the frequentist tendencies of the modelled system e.g. the observation error of the instrument due to an aggregation of several fluctuations in the surroundings. Measuring a time invariant value of a physical quantity using the same instrument should yield the distribution of such a measurement noise. On the other hand, other substantial contributions to uncertainty come from physical realities, like for example the infiltration parameter of the catchment, that do not lend themselves to limits-of-frequency interpretations. Therefore, the Bayesian framework comes to help, where the probabilities are not relative frequencies, but our degrees of belief regarding the true value of physical quantities. In the Bayesian framework, as more information becomes available, the belief gets updated using the calculus of probabilities, which assures the logical consistency of the updating process.

While it is easy to answer the question, given certain parameters, how would the system respond, it is hard to formulate the conditioning the other way round—given a certain number of observations, how are the parameter values distributed. Bayes' Theorem relates these two probabilities in its equation. Knowing one allows us to compute the other. We put out prior understanding of the system behaviour in the probabilistic model. We also have some prior belief about the values of the parameters and, as soon as more observational data comes in, we update our belief. Furthermore, we proceed to make predictions using these updated ranges of parameter values. Prediction then is the evaluation of the probabilistic model given the data.

Evaluating and dealing with such probabilities can help in optimizing decision making (Hall, 2003; Hall and Solomatine, 2008; Verkade and Werner, 2011; Reichert and Schuwirth, 2012). Given the risk aversion of a community and the costs we are willing to incur to avoid a certain risk, design and operation choices can be made accordingly. Besides, safety factors in design, which are hard to justify in terms of their adequacy and economy, get replaced by quantified likelihood of failure. For hydrologic phenomena that have a frequentist characteristic, e.g. annual flood extremes, the evaluation of correct probabilities will lead to predictable aggregate effects. For hydrologic phenomena where probabilities are purely representative of degrees of belief, decisions can be made under consideration of risk tolerance and aversion. For example building an irrigation or drainage network with a failure probability of $p\%$. as opposed to $2p\%$. Similarly, hypothesis testing can be done in more nuanced ways. In cases where the hypotheses closely explain the data, then assigning probabilities to those hypotheses will prevent the false rejection or acceptance of one over the other, till more differentiating data is obtained. In the absence of an explanation for the error, in a strict Popperian sense, all hypotheses should be falsified as none will be able to explain the data with zero error (Nearing et al., 2016; Popper, 2002).

4 Parameter and uncertainty estimation: previous work and research gaps

Going beyond the manual tuning of model parameters, observational data of the hydrologic system response is used to estimate appropriate parameter values. Broadly speaking, there are two paradigms in hydrology within which parameter and uncertainty estimation is performed. 1) Post-processors: In this scheme the best fitting parameter values are identified by minimizing a specific cost function. Once the parameter values are established, uncertainty in model predictions is quantified from model errors in the past. This paradigm generally cares about operational model predictions and does not seek to disentangle uncertainty from various sources. There has been considerable amount of research in identifying the pros and cons of this set of techniques for hydrologic forecasts. Examples of techniques developed under this framework are Quantile Regression (Weerts et al., 2011), UNEEC (Solomatine and Shrestha, 2009), DUMBRAE (Pianosi and Raso, 2012) etc. 2) Joint Inference: In this scheme, probabilities are associated with the whole range of parameter values. Given a time series of observational data, both the parameters of the deterministic hydrologic model and the error model are inferred from the data. Within the Bayesian framework, such a joint inference provides a consistent way to combine expert knowledge and information from the data. Examples of techniques developed under this framework are Bayesian Total Error Analysis (BATEA), bias as an autoregressive order 1 process, bias as a gaussian output dependent process, bias as a result of a stochastic input process etc (Del Giudice et al., 2016, 2013; Kuczera et al., 2006; Schoups and Vrugt, 2010). These techniques generally need Monte Carlo simulations and can thus be computationally expensive.

For operational forecasts, not only is the reliability of the predictions important but the computational and conceptual costs also need to be acceptable. Monte Carlo simulation-based uncertainty analysis techniques require a large number of model evaluations, which may not always be feasible. Therefore, alternatives to Monte Carlo simulations for cheap estimation of predictive uncertainty in situations of computational and operational constraints have always been very attractive for flood forecasting models. The effect on the performance of such techniques due to their simplicity

needs to be analysed in detail.

There are other studies that focus on specific sources of uncertainty in hydrology e.g. due to parameters, model structure deficits or inputs. The formulation of representative error models, which enable reliable prediction intervals and unbiased parameter inference, has been investigated in the past two decades. The model errors are complex, with the properties of memory (i.e. they are autocorrelated) and changing variance (i.e. they are heteroscedastic) (Renard et al., 2010). The benefits of incorporating these properties have been studied. However, their incorporation also comes with other challenges that have been delineated in previous research (Evin et al., 2013). Updates in these descriptions are sought to bypass such problems and make the error models more realistic.

Substantial attention has been paid in hydrology to make error descriptions more representative of the rainfall-runoff process. In the research work mentioned above, several parameterizations for the likelihood function have been proposed, for example, to capture heteroscedasticity (Evin et al., 2013), skewness (Schoups and Vrugt, 2010), autocorrelated structural deficits through time continuous processes (Del Giudice et al., 2013) etc. However, research gaps remain. Models in hydrology are usually used to either hypothesize a phenomenon or make predictions. While for predictive models the system understanding comes second to the performance of the model, for explanatory studies, where we may be interested in the acceptance/rejection of a hypothesis, the underlying process should also be representative. This means if we believe that the parameters in our model correspond to a distributed or aggregated quantity which is actually a property of the system, then we are interested in their "true" values, and not any values that can produce a reasonable prediction. Also, more information about the behaviour of the system should increase our confidence in the value of the parameters. And for predictions, the predictive uncertainty needs to be computed, which is the conditional probability $f(y_{future}|y_{observed}, x_{input})$.

There has also been a paradigm shift in hydrology where data from unconventional sources is being used to improve hydrologic models. Hydrologic data is being collected from CCTV cameras, from crowdsourcing and from alternate sensors. However, the conventional parameter estimation techniques in hydrology have been mostly developed keeping the data from flow meters or gauge readings in mind. Given this data revolution, statistical techniques need to be investigated that can get the maximum information out from hydrologic models used for both explaining and predicting. With the advent of Internet of Things and the proliferation of unconventional data (Scheidegger and Rieckermann, 2014; Eggimann et al., 2017; Ilja Van Meerveld et al., 2017; Mazzoleni et al., 2018; Moy De Vitry et al., 2017), the classical parameter inference for hydrologic models needs to be adapted to the changing nature of the observations. Ideally, if the parameter estimation is done in the Bayesian framework, the likelihood function should be able to handle different types of observational data.

5 Research questions addressed in this work

The objective of this PhD research is to investigate various statistical tools that can improve the descriptions of uncertainties associated with hydrologic models. Besides, reduction in these uncertainties is sought by improving the parameter estimation of models. This thesis addresses three specific questions in the context of hydrologic modelling:

1. For operational hydrology, where Monte Carlo simulations can be computationally prohibitive, evaluating the performance of a simple and easy-to-implement data-driven uncertainty estimation technique that can be applied in the post-processing phase. What is the comparative performance of such a technique?

2. Given the widely employed formulation of an additive random error for hydrologic model outputs, how can we use censored observations for parameter inference? What is the information gain from such censored data? And what is the sensitivity of inference to sensor

placement?

3. As additive error models have some constraints, for example in the choice of marginals that can be used with autocorrelation, possibility of negative flows etc., are there other formulations of error models possible that offer more flexibility? This flexibility can be then used to incorporate properties like autocorrelation, heteroscedasticity, and non-negativity in a better fashion.

6 Thesis outline

This thesis report is divided into seven main chapters and two appendix chapters. Apart from *Introduction* and *Concluding remarks*, the main chapters are structured such that the techniques with more operational value are discussed first. The focus then shifts more towards explanatory analysis in the context of parameter and uncertainty estimation.

Chapter 2 starts with the description of a residual uncertainty estimator that provides reliable prediction intervals, if large time series of observational data is available from the past. This chapter is an investigation into the applicability of one of the easy-to-implement uncertainty estimation techniques. A k nearest neighbour (kNN) search is used to resample errors and generate prediction intervals. The sensitivity of the technique to the choice of search space, value of k, and systematic deficits in model is explored. By comparing kNN resampling to other more sophisticated post processors, its performance and reliability are ascertained.

Chapter 3 investigates the applicability of kNN resampling for distributed models with multiple variable outputs. New metrics are also introduced to choose the components of the search space. The discussion builds on the results and elaborates on the advantages of kNN resampling as a robust technique for generating prediction intervals.

In Chapter 4 a formal likelihood function is suggested that incorporates censored observations, while accounting for model structure deficits and uncertainty in input data. It is shown that the parameter inference can be performed within the Bayesian framework for such nonconventional data. The methodology is implemented on an urban catchment, where parameters of a hydrodynamic rainfall-runoff model are inferred from binary observations of water level. This work demonstrates that censored observations can be valuable for learning about model parameters, while retaining a probabilistic framework during parameter estimation.

Chapter 5 is an exploration into the properties of hydrologic times series i.e autocorrelation, heteroscedasticity, and non-negativity. An alternative to additive error models is suggested. Distributions with a non-negative support are used in conjunction with copulas to define the likelihood function of hydrologic models. Copulas capture autocorrelation in time. This scheme adds flexibility to the formulation of an error description for specific hydrologic models, as mariginals and copulas can be chosen separately. The effects of such a framework on parameter estimation and model prediction are analysed.

Chapter 6 is a short essay on the philosophical basis and implications of employing probability calculus for hydrologic modelling applications. The essay revisits the axiomatic basis of quantifying uncertainty using probability theory. It soon after turns to the implications of using such a framework and delineates some of the expectations of the hydrologic community from such usage.

The final take home messages of this thesis are synthesized and reiterated in chapter 7. The chapter ends with an outlook on statistical tools presented in this thesis. Based on this work, some recommendations for future research are also made in this chapter.

Appendix A and B contain two papers that contribute to the overarching theme of this thesis. Appendix A is a comparative study and analysis of various transformational strategies that make

rain gauge and radar data more normal, before merging them in kriging. The performance of these strategies has a bearing on the performance of the final merged rainfall product, which in turn affects hydrologic model forecasts. Whereas, Appendix B presents a new sediment wash-off model for which parameter and uncertainty estimation is done within the Bayesian framework. The improved performance of both the deterministic model and the associated prediction intervals is demonstrated in the study.

This thesis report, being a cumulative dissertation, consists of published and submitted journal papers. Chapters 2, 4, A, and B are published in peer-reviewed journals. Whereas, chapter 3 and chapter 5 have been submitted. A statement on the contribution of authors is provided at the beginning of each chapter.

Chapter 2

Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting

Omar Wani[1,2,3,4], Joost Beckers[2], Albrecht Weerts[2,5], Dimitri Solomatine[1,6,7]

[1] IHE Delft Institute for Water Education, Delft, the Netherlands
[2] Deltares, Delft, the Netherlands
[3] Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland
[4] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[5] Hydrology and Quantitative Water Management Group, Department of Environmental Sciences, Wageningen University, Wageningen, the Netherlands
[6] Water Resources Section, Delft University of Technology, Delft, the Netherlands
[7] Water Problems Institute of RAS, Moscow, Russia

Journal: Hydrology and Earth System Sciences

Publication date: 10th August 2017

DOI: https://doi.org/10.5194/hess-21-4021-2017

Author contributions: OW, in collaboration with JB, AW, and DS, formulated the method and designed the simulation experiments. OW wrote the code and did the analysis. OW wrote the manuscript, with comments from JB, AW, and DS. All authors reviewed the final manuscript.

User interface: A website has been developed as part of this research to help generate uncertainty intervals using kNN resampling for a given time series of predictions: www.modeluncertainty.com.

Abstract

A non-parametric method is applied to quantify residual uncertainty in hydrologic streamflow forecasting. This method acts as a post-processor on deterministic model forecasts and generates a residual uncertainty distribution. Based on instance-based learning, it uses a k nearest neighbour search for similar historical hydrometeorological conditions to determine uncertainty intervals from a set of historical errors, i.e. discrepancies between past forecast and observation. The performance of this method is assessed using test cases of hydrologic forecasting in two UK rivers: the Severn and Brue. Forecasts in retrospect were made and their uncertainties were estimated using kNN resampling and two alternative uncertainty estimators: quantile regression (QR) and uncertainty estimation based on local errors and clustering (UNEEC). Results show that kNN uncertainty estimation produces accurate and narrow uncertainty intervals with good probability coverage. Analysis also shows that the performance of this technique depends on the choice of search space. Nevertheless, the accuracy and reliability of uncertainty intervals generated using kNN resampling are at least comparable to those produced by QR and UNEEC. It is concluded that kNN uncertainty estimation is an interesting alternative to other post-processors, like QR and UNEEC, for estimating forecast uncertainty. Apart from its concept being simple and well understood, an advantage of this method is that it is relatively easy to implement.

1 Introduction

Hydrologic forecasts for real-life systems are inevitably uncertain (Beven and Binley, 1992; Gupta et al., 1998; Refsgaard et al., 2007). This, among other things, is due to the uncertainties in the meteorological forcing, in the modelling of the hydrologic system response and in the initial state of the system at the time of forecast. It is well accepted that, compared to a simple deterministic forecast, additional information about the expected degree of accuracy of that forecast is valuable and generally leads to better decision making (Krzysztofowicz, 2001). Various techniques have therefore been developed to quantify uncertainties associated with the meteorological model input (van Andel et al., 2013), the initial state of the model (Li et al., 2009) and the hydrologic models themselves (Deletic et al., 2012; Coccia and Todini, 2011). Frameworks and guidelines have been developed to incorporate uncertainty analysis of environmental models effectively in decision making (Arnal et al., 2016; Reichert et al., 2007; Refsgaard et al., 2007). Broadly, there are three basic approaches to uncertainty estimation: (i) explicitly defining a probability model for the system response, e.g. Todini (2008), (ii) estimation of statistical properties of the error time series in the post-processing phase of model forecast, e.g. Dogulu et al. (2015), and (iii) methods using Monte Carlo sampling of inputs and/or parameters, aimed at getting a range of model outputs, e.g. Beven and Binley (1992) and Freer et al. (1996). Other uncertainty estimation techniques may employ a combination of these approaches (Del Giudice et al., 2013). Some techniques focus on one source of uncertainty, such as the model parameter uncertainty (Benke et al., 2008) or the model structure uncertainty (Butts et al., 2004), while others focus on combined uncertainties stemming from model parameters, model structure deficits and inputs (Schoups and Vrugt, 2010; Evin et al., 2013; Del Giudice et al., 2013). In this context, it is important to note that apart from estimating uncertainty of model parameters during calibration, uncertainty estimation for hydrologic forecasting requires quantification of predictive uncertainty, which includes uncertain system response in addition to different combinations of model parameters (Renard et al., 2010; Coccia and Todini, 2011; Dotto et al., 2012).

In this paper, we will restrict ourselves to the class of uncertainty estimators called post-processors. These methods usually do not discriminate between different sources of uncertainty. They "aggregate" all sources into a so-called residual uncertainty. Post-processing methods assume the existence of a single calibrated model with an optimal set of model parameters, and build a statistical or machine learning model of the residual uncertainty. Typically, these techniques relate a combination of model inputs and/or outputs to the model error distribution. Various post-processors have been developed and applied to hydrologic modelling, such as a meta-Gaussian error model (Montanari and Brath, 2004), UNEEC (Solomatine and Shrestha, 2009), quantile regression (Weerts et al., 2011), and DUMBRAE (Pianosi and Raso, 2012). Quantile regression (QR) is a relatively straightforward post-processing technique that relates the probability of residual errors to the model forecast (the predictand) by a regression model that is derived from historical forecasts and observations. QR has been successfully applied for uncertainty quantification in hydrologic forecasts with various modifications (Weerts et al.,2011; Verkade et al.,2013; Roscoe et al.,2012, ;López López et al., 2014; Hoss and Fischbeck,2015), whereas UNEEC involves a machine learning technique for building a non-linear regression model of error quantiles (Solomatine and Shrestha, 2009). UNEEC includes three steps: (1) fuzzy clustering of input data in the space of "relevant" variables; (2) estimating the probability distribution function of residual errors for each cluster and (3) building a machine learning model (e.g. an artificial neural network) of the prediction interval for a given probability (Dogulu et al., 2015). Many other uncertainty estimation techniques, such as DUMBRAE (Pianosi and Raso, 2012), HUP (Krzysztofowicz, 1999), model conditional processor (Coccia and Todini, 2011), Bayesian revision (Reggiani et al., 2009) and Bayesian model averaging (Raftery et al., 2005), make explicit assumptions about the nature of the probability distribution function of error. This is not necessary for QR and UNEEC (López López et al., 2014; Dogulu et al., 2015). Nevertheless, in QR and UNEEC assumptions need to be made about the form of the regression function that is used to calculate the quantiles.

In an attempt to explore the utility of easier-to-implement post-processing techniques, we employ

a simple nonparametric forecast method for residual uncertainty quantification. This method uses kNN search to learn about the past residual errors, which avoids having to make explicit assumptions about the nature of the error distribution and tuning of distribution parameters. Instance-based learning has been used in meteorology and hydrology before for resampling of precipitation and streamflows, most notably by Lall and Sharma (1996), who used the k nearest-neighbour (kNN) method for resampling of monthly streamflow sequences. kNN search has also been used in a non-parametric simulation method to generate random sequences of daily weather variables (Rajagopalan and Lall, 1999). They defined a weighting function for probability where the predictand is resampled from k values. Jules and Buishand (2003) used nearest-neighbour resampling to generate multisite sequences of daily precipitation and temperature in the Rhine basin. Also, instance-based learning has been used as a data-driven model for hydrologic forecasting (Solomatine et al., 2008; Solomatine and Ostfeld, 2008). Beckers et al. (2016) use nearest-neighbour resampling to generate monthly sequences of climate indices and related precipitation and temperature series for the Columbia River basin. Specifically in the context of error modelling, a version of UNEEC that uses kNN instance-based learning as its basic machine learning technique to predict the residual error quantiles was compared to the original ANN-based UNEEC in Shrestha and Solomatine (2008). However, kNN can also be used without the complicated UNEEC procedure that includes fuzzy clustering. The application of kNN has recently been tested for forecast updating by constructing a deterministic error prediction model (Akbari and Afshar, 2014). Similarly, it has been shown that model errors can be resampled using kNN, after explicitly accounting for input and parameter uncertainty, to generate uncertainty intervals (Sikorska et al., 2015). In this paper we extend the simplification of kNN resampling for uncertainty estimation. We present an application of the kNN method to generate residual uncertainty estimates for a predictand, using a fixed time series of input and fixed model parameters, and explore whether this approach, being simpler than many other uncertainty quantification approaches mentioned above, is a useful or even a better alternative.

To demonstrate its use, we employ a relatively simple configuration of kNN resampling to produce uncertainty intervals for hydrologic forecasting. The next section explains the method in more detail and describes the validation procedure, i.e. the performance indicators. In Sect. 3, the method is applied to two case studies, each with a different system response (discharge and water level). The performance of kNN uncertainty estimation as a function of forecast lead time is analysed in the first case study. The second case study is used to further validate the performance of kNN uncertainty estimation and analyse its sensitivity to the choice of search space and the value of k. Also, the influence of systematic bias in the hydrologic model on the uncertainty intervals generated by kNN search is explored in the second case study. For both case studies, performance indices of kNN resampling are compared to those of QR and UNEEC. And finally in Sect. 4, we discuss the usability of kNN search as a postprocessor uncertainty estimator in hydrologic forecasting.

2 Method

2.1 kNN error model

The kNN residual uncertainty estimator can be seen as a zero-order local error quantile model built from a kNN search. Let us define a vector $\boldsymbol{v}$ in n-dimensional space of variables (the search space) on which the residual uncertainty is assumed to be statistically dependent.

$$\boldsymbol{v} = [v^1, ..., v^n] \tag{1}$$

The cumulative probability distribution function C of residual errors at prediction time step t conditioned on $\boldsymbol{v} = \boldsymbol{v}_t$ is defined as

$$C_t(e|\boldsymbol{v} = \boldsymbol{v}_t) = P_t(E \leq e|\boldsymbol{v} = \boldsymbol{v}_t) \tag{2}$$

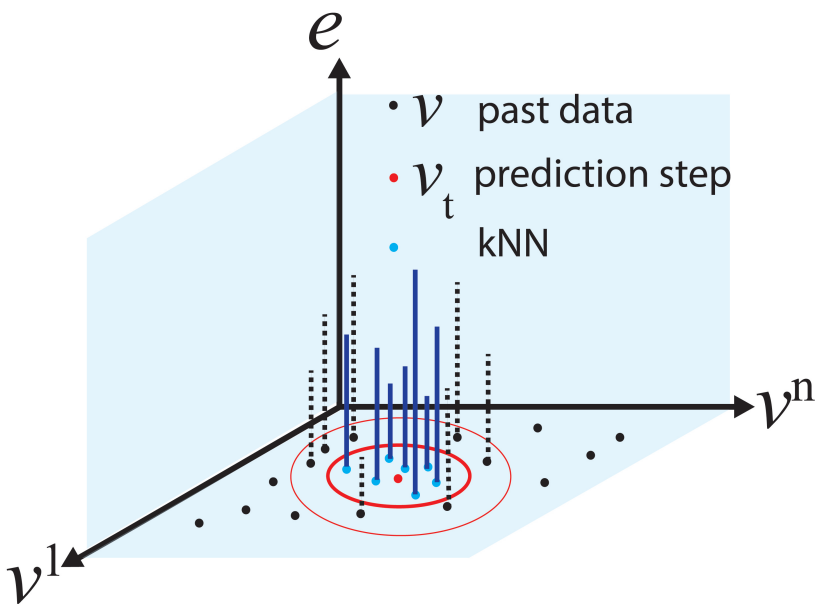


Figure 1: Dependence of error samples on the value of k. For larger values of k, points are at a greater distance from $\boldsymbol{v}_t$ (the prediction step), thus compromising the conditioning of the residual error probability distribution on v_t (Eq. 5).

where P is the probability function and E denotes the random variable for residual errors. Residual error is defined throughout this paper as the difference between the simulated values and the observed values for a hydrologic system response f , like discharge or water level.

$$e = f_{simulated} - f_{observed} \tag{3}$$

We are making the assumption of stationarity in time so that past error distributions are representative of the future:

$$C_t(e|\boldsymbol{v} = \boldsymbol{v}_t) = C_p(e|\boldsymbol{v} = \boldsymbol{v}_t) \tag{4}$$

The subscript p denotes historical time series. Therefore C_p is the cumulative distribution function of residual errors from the past. In Eq. (4), C_p is being conditioned to the input variable vector at time t . Nevertheless, as we only have single realizations of the error variable E for each historical point, we relax the constraint of $\boldsymbol{v} = \boldsymbol{v}_t$. Instead, we assume that the nearby neighbours of $\boldsymbol{v}_t$ in n-dimensional space will have a similar probability distribution of errors to $\boldsymbol{v}_t$ and that these historical errors are samples from $C_p(e|\boldsymbol{v} = \boldsymbol{v}_t)$. An empirical probability distribution can thus be constructed using the kNN historical errors:

$$C_t(e|\boldsymbol{v} = \boldsymbol{v}_t) \approx C_p(e|r_p \leq r_k) \tag{5}$$

where r_p is the Euclidean distance in n-dimensional space of input variables.

$$r_k = \sqrt{[\sum_{i=1}^{n} \frac{(v_p^i - v_t^i)^2}{\sigma_i^2}]} \tag{6}$$

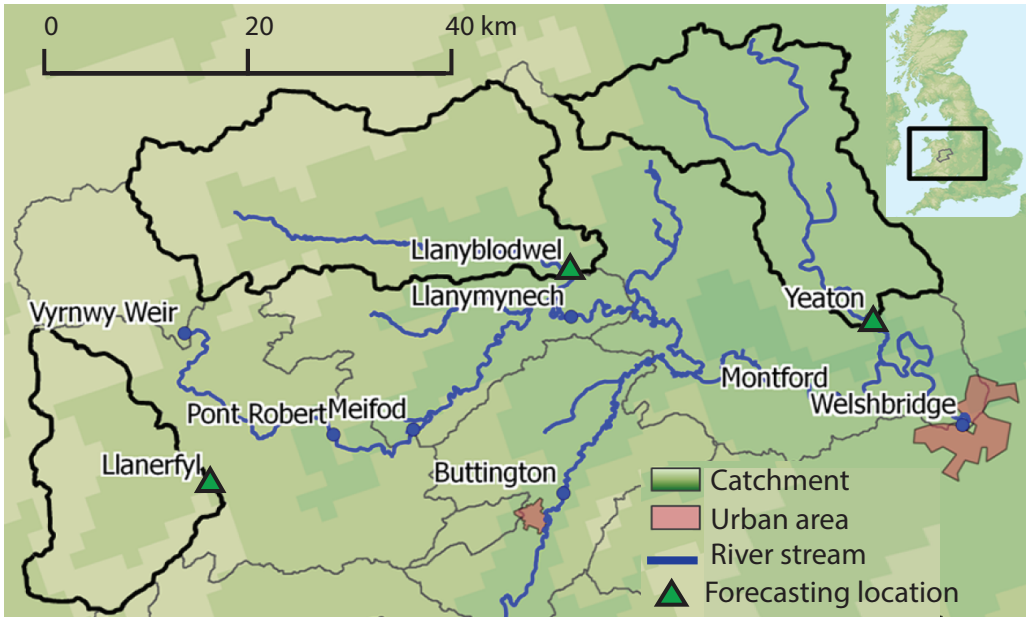


Figure 2: Upper Severn subcatchments with gauging stations (from López López et al., 2014).

$\boldsymbol{v}_p$ is the input variable vector of the past data point in the cloud of such past data points $\boldsymbol{v}$ (Fig. 1) and r_k is the distance to the k^{th} nearest neighbour of $\boldsymbol{v}_t$. The choice of the input variable vector is a problem in itself since it should include only the most relevant variables that determine the forecast uncertainty. In this study, the input variable vector is chosen based on correlation between the candidate variables and the past errors. If the correlation between the error time series and a particular candidate variable is relatively high, then it can be included in the input variable vector space. Other, more sophisticated methods involving the mutual information can be used as well (Fernando et al., 2009). This will be exemplified in the case studies described in the next section. To represent the relative importance of input variables used in the search, dimensions of the input variable vector space can be suitably weighted in. Also, the model-based methods can be used where models are built for each considered candidate input variable set, and the choice is made based on their relative performance. These, however, were not explored in this study; it rather focused on the usability of the kNN search in its most basic implementation for uncertainty quantification. Nevertheless, we do demonstrate the sensitivity of the uncertainty intervals to the choice of input variable vector.
In order to level variables with different magnitudes, they are normalized. If i represents the standard deviation of input variable i calculated using the past data, then

$$r_k = \|\boldsymbol{v}_p - \boldsymbol{v}_t\| = \sqrt{[\sum_{i=1}^{n}(v_p^i - v_t^i)^2]} \tag{7}$$

Once the input variable vector space is decided, the probability of non-exceedance of a forecast error is calculated empirically by sampling from the conditional error distribution:

$$C_t(e|\boldsymbol{v} = \boldsymbol{v}_t) \approx C_p(e|r_p \leq r_k) = j/k \tag{8}$$

where j is the rank of value e (for which the probability of non-exceedance is being computed) in the ascending array of k error values. The kNN search is thus employed to generate a sample and

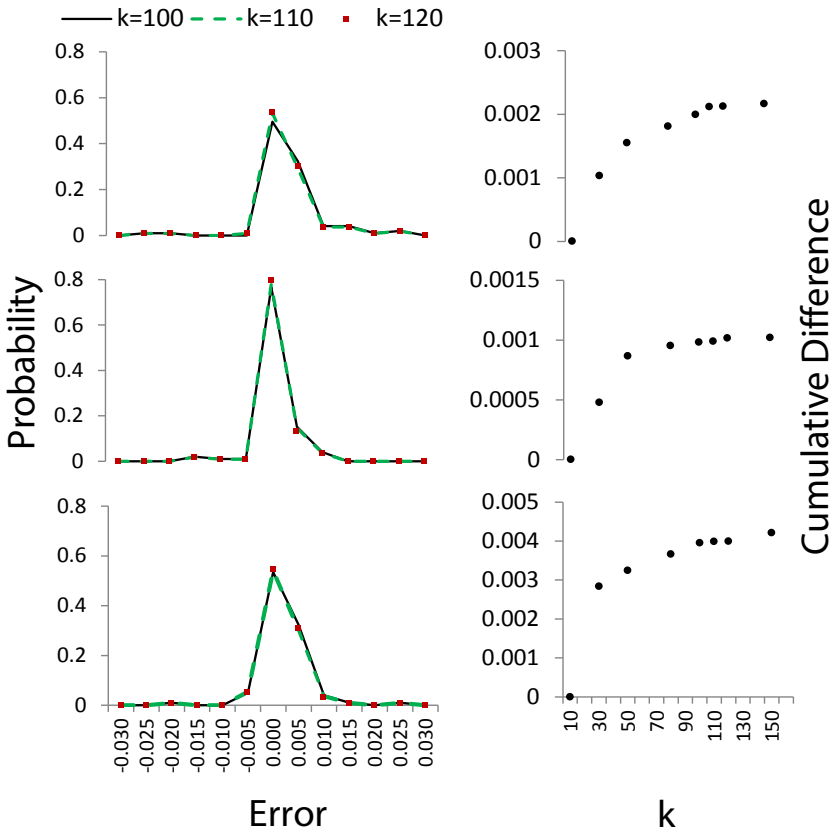


Figure 3: Dependence of the residual error probability function on the value of k for three didactic values of v_t (in each row in this plot). The probability is computed for error bins of size 0.005 units each. The graphs show that for k from around 90 to 120, the corresponding empirical error distributions become almost identical.

to build an empirical error distribution for this predictive uncertainty quantification. Such a mathematical description does not employ explicit regression models for predicting quantiles, which can be seen as a disadvantage in extrapolating outside available data. Also, as this configuration of kNN used in this research generates residual error quantiles, which capture the mismatch between measurement values and simulated values, the uncertainty in observational data is not considered. The generated quantiles are aimed at capturing the measured system response and do not attempt to capture the true response of the hydrologic system.

As one would expect, due to the nature of our sampling approximation (Eq. 8), the number of nearest neighbours, k, will affect the empirical conditional probability distribution of errors. If k is very large, many data points that are quite distant from $\boldsymbol{v}_t$ (Fig. 1) will be selected and the conditioning on the current forecast situation will not be valid. Large values of k will thus yield error distributions with larger uncertainty intervals —resembling the marginal error distribution. If k is small, the set of k errors will be small and subject to sampling error, so this set will not adequately represent the uncertainty distribution at $\boldsymbol{v}_t$. The tail of a distribution is more prone to sampling errors compared to its mean. Thus, to attain an acceptable degree of convergence, many more samples are required for quantiles corresponding to bigger prediction intervals (van der Vaart, 1998). For improved performance, the value of k can be subject to optimization of some cost function: the optimal value of k could be the one that enables a reasonable estimate of the uncertainty quantiles and additionally we may require that the sensitivity of the error distribution to k is small. In this study, we carry out such optimization using quite a simple heuristic guideline —the value

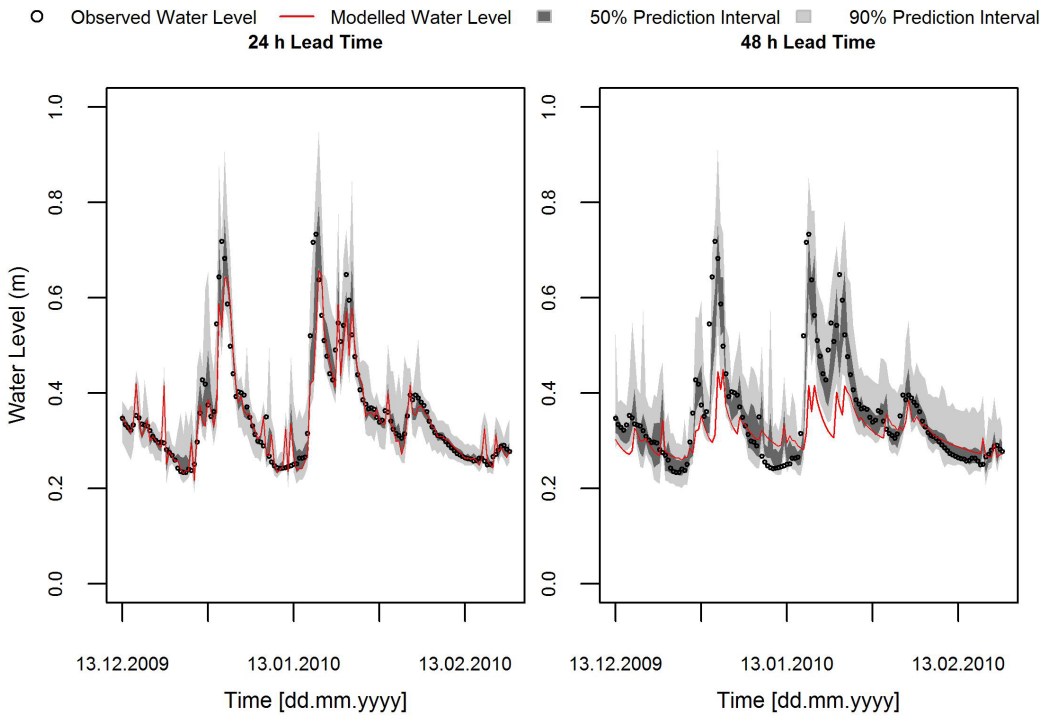


Figure 4: Prediction intervals for the Yeaton catchment using kNN resampling. The hydrographs are shown for the two different lead times. The 50% prediction interval is the interval between the 25 and 75% quantiles of residual error, and the 90% quantile is the interval between the 5 and 95% quantiles. The reporting time interval is 12 h.

of k is varied until the probability distribution of errors stabilizes and becomes less sensitive to the value of k for a few model predictions. We also demonstrate the sensitivity of uncertainty intervals to the value of k in one of the case studies. The choice of this relatively simple procedure for error quantile generation using kNN resampling is a reasonable starting point to assess its potential for residual uncertainty. This study explores the potential of uncertainty estimation using kNN in as simple a way as possible, and then compares its performance to two other residual uncertainty estimators. More advanced application of kNN, for example using fuzzy weights and kNN sampling to assign prediction intervals (Shrestha and Solomatine, 2008) or through explicit consideration of uncertainty in parameter and input by sampling them from their distributions, has been successfully shown (Sikorska et al., 2015). To summarize, the steps for uncertainty quantification using kNN resampling are as follows.

1. Compose the input variable vector space ($\boldsymbol{v}$) on which uncertainty will be conditioned. Correlation analysis can help find the most relevant variables.
2. Set the number of neighbours k.
3. For a forecast at prediction time step t , identify the set of k nearest neighbours to the input vector $\boldsymbol{v}_t$. This set represents the hindcasts (forecasts in retrospect) most similar to $\boldsymbol{v}_t$.
4. Use the residual errors from these k points to build an empirical error distribution for the forecast at time step t .
5. Finally, identify the errors corresponding to the required quantiles (probabilities of non-exceedance) from this empirical distribution (in this paper, we use the 5—95 and 25—75% quantiles).

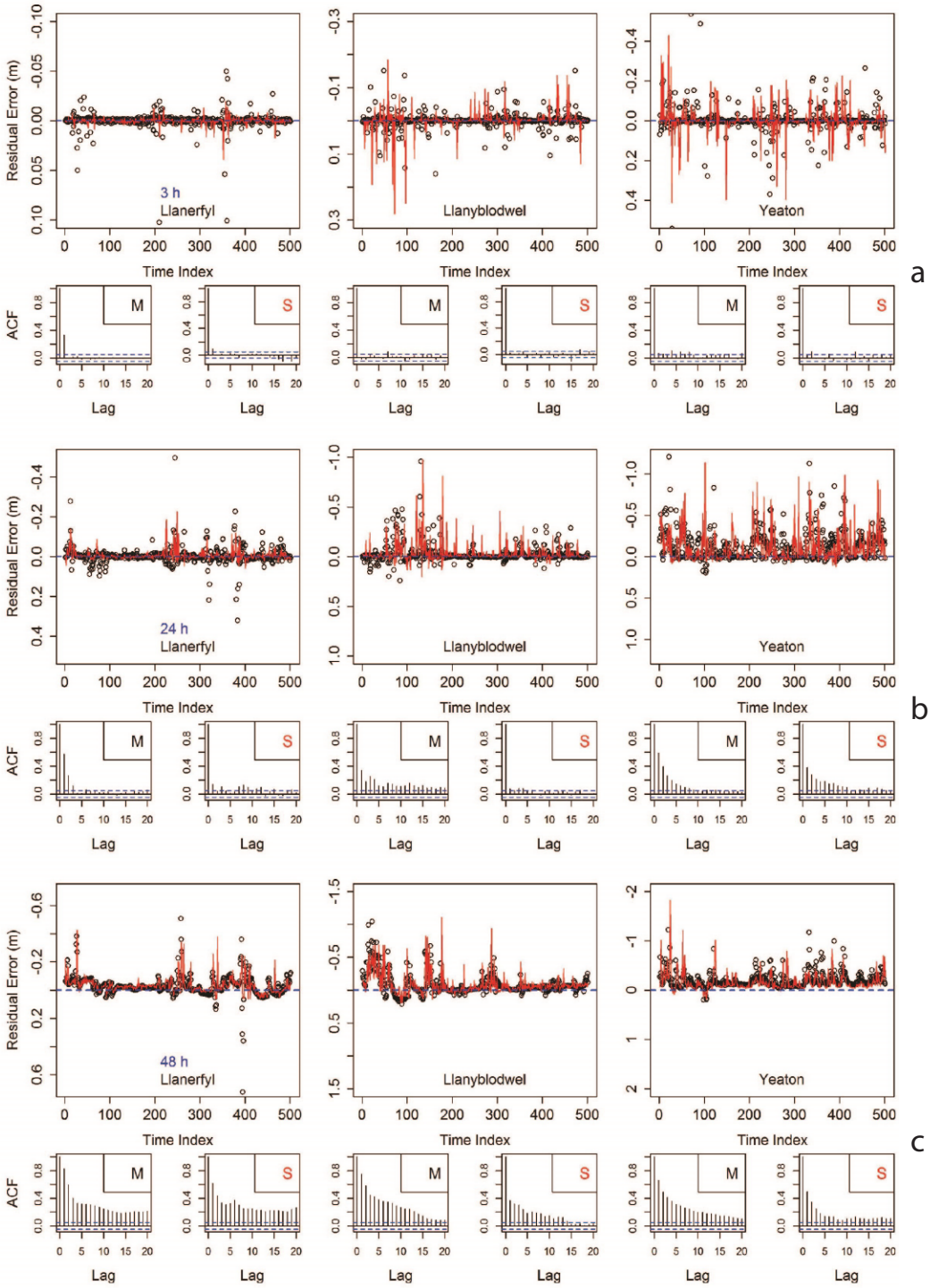


Figure 5: Plots of error samples and their autocorrelation (ACF). The error time series generated using kNN resampling are in red. Black circles represent the observed errors, i.e. obtained after measuring water level and comparing it to the simulated water level. M stands for measured and S for simulated. The lead times for each row are (a) 3 h, (b) 24 h and (c) 48 h.

2.2 Validation methods

Three statistical measures have been employed in this study to check the effectiveness of uncertainty estimation techniques, namely prediction interval coverage probability (PICP_{PI}), the mean prediction interval (MPI_{PI}) (see e.g. Shrestha and Solomatine, 2008; Dogulu et al., 2015) and the Alpha Index (Renard et al., 2010). PICP_{PI} represents the percentage of observations (C) covered by a prediction interval (PI) corresponding to a certain probability of occurrence (in our case 90 and 50 %).

$$PICP_{90} = \frac{N_{in}}{N_{obs}} \cdot 100\% \tag{9}$$

where N_{in} is the number of observations located within the PI and N_{obs} is the total number of observations. These metrics are calculated using the following equations:

$$PICP_{90} = \frac{1}{n}\sum_{i=1}^{n} C_{90} \cdot 100\%,\ PICP_{50} = \frac{1}{n}\sum_{i=1}^{n} C_{50} \cdot 100\% \tag{10}$$

$$C_{90} = \begin{cases} 1 & if\ q_{i,0.05} \leq q_i \leq q_{i,0.95} \\ 0 & else \end{cases},\ C_{50} = \begin{cases} 1 & if\ q_{i,0.25} \leq q_i \leq q_{i,0.75} \\ 0 & else \end{cases} \tag{11}$$

where $q_{i,0.95}$ and $q_{i,0.05}$ are values with 95 and 5% probabilities of non-exceedance at time i. Thus the region bound within these two values will have a confidence interval of 90%. Similarly, $q_{i,0.75}$ and $q_{i,0.25}$ represent the boundaries for 50% C. The MPI is the average width of the confidence intervals corresponding to a particular probability. It is a measure of the magnitude of the uncertainty.

$$MPI_{90} = \frac{1}{n}\sum_{i=1}^{n} q_{i,0.95} - q_{i,0.05},\ MPI_{50} = \frac{1}{n}\sum_{i=1}^{n} q_{i,0.75} - q_{i,0.25} \tag{12}$$

We also quantify the reliability of the predicted error quantiles by comparing it to the observed error quantiles. The mismatch between the observed ($q_{obs,j}$) and predicted ($j/100$) error quantiles can be summarized by the Alpha Index (α).

$$\alpha' = \frac{1}{100}\sum_{j=1}^{100} |q_{obs,j} - j/100| \tag{13}$$

$$\alpha = 1 - 2\alpha' \tag{14}$$

There have been discussions whether an isolated verification index can capture all the aspects that make a probabilistic forecast good or bad (Laio and Tamea, 2007). The choice of a verification index for an uncertainty estimation technique should also be dependent on the purpose of hydrologic forecast. For example, Coccia and Todini (2011) evaluate the performance of model conditional processors for flood forecasting using the predicted and observed probabilities of exceedance over a threshold. Also, in their study predicted error quantiles are compared to observed error quantiles. López López et al. (2014) and Dogulu et al. (2015) use PICP and MPI, among other verification measures, to access the performance of QR and UNEEC. This study will limit the comparison of kNN resampling with other techniques to PICP and MPI only, which give a reasonable assessment of performance. Nevertheless, it does not preclude the possibility that the uncertainty estimation techniques perform differently if evaluated using other indices.

3 Case studies

The performance of kNN resampling was evaluated by applying the technique to hydrological forecasting for several catchments in two different parts of England. The two case studies provide two different hydrologic conditions for testing and include different models for prediction. Also, different kinds of system responses are being predicted in the two case studies —water level and discharge. The accuracy of the quantified prediction intervals was deduced by using validation data sets. Also, the first case study was used to evaluate the impact of changing lead time on uncertainty of hydrologic models and its quantification using kNN resampling.

3.1 Upper Severn catchment

Catchment description

The Upper Severn region is located in the Midlands, UK (Fig. 2). The River Severn, with a total length of 354 km, is the longest river in the UK. Its course acts as a geographic delineation between England and Wales, finally draining into the Bristol Channel. The overall River Severn catchment area is 10 459 km2. Around 2.3 million people live in this region. The area is predominantly rural, but there are also a number of highly urbanized parts. The area covering the upper reaches of the River Severn, from its source on Plynlimon to its confluence with the River Perry upstream of Shrewsbury in Shropshire, is called the Upper Severn catchment. The Upper Severn catchment is predominantly hilly. It is dominated on the western edge by the Cambrian Mountains and a section of Snowdonia National Park (River Severn CFMP; EA, 2009).

The Severn catchment has a diverse geology. The headwaters of the river rise on Silurian mudstones, siltstones and grits and flow eastwards over these same rock formations. These rock formations do not allow water to flow easily through them. Therefore they are classified as non-aquifers with only limited potential for groundwater abstraction. Further west, in the Middle Severn section, the River Severn encounters sandstones, which are classified as a major aquifer and are highly permeable, highly productive and able to support large groundwater abstractions (River Severn CFMP; EA, 2009). The climate of the Severn catchment is generally temperate, experiencing modest to high precipitation depending on topography. Welsh mountains can receive over 2500mm of precipitation per annum, whereas the rest of the catchment receives rainfall similar to the UK average —less than 700mm per annum. The test forecast locations used in this study are Llanerfyl, Llanyblodwel and Yeaton. Table 1 lists the basin and hydrological information for these subcatchments (López López et al., 2014).

Table 1: Basin information for Upper Severn subcatchments (EA, 2013; Marsh and Hannaford, 2008).

Catchment	Area (km^2)	Mean annual rain (mm)	Mean flow (m^3s^{-1})	Max water level (m)
Llanerfyl	125	1077	>10	3.59
Llanyblodwel	229	1267	6.58	2.68
Yeaton	180.8	767	1.6	1.13

Experimental set-up

The flood forecasting system for the River Severn is organized in a sequential manner, being composed of a number of separate systems that are effectively linked. This forecasting system works with a high degree of automation and efforts have been made to involve a minimum amount of human intervention. The UK Environment Agency uses the Midlands Flood Forecasting System (MFFS) to do flood forecasting and to help in warning operation. The MFFS in turn is based on the Delft-FEWS (Flood Early Warning System) platform (Werner et al., 2013).Within the MFFS, there are lumped numerical models for rainfall-runoff (MCRM; Bailey and Dobson, 1981) and

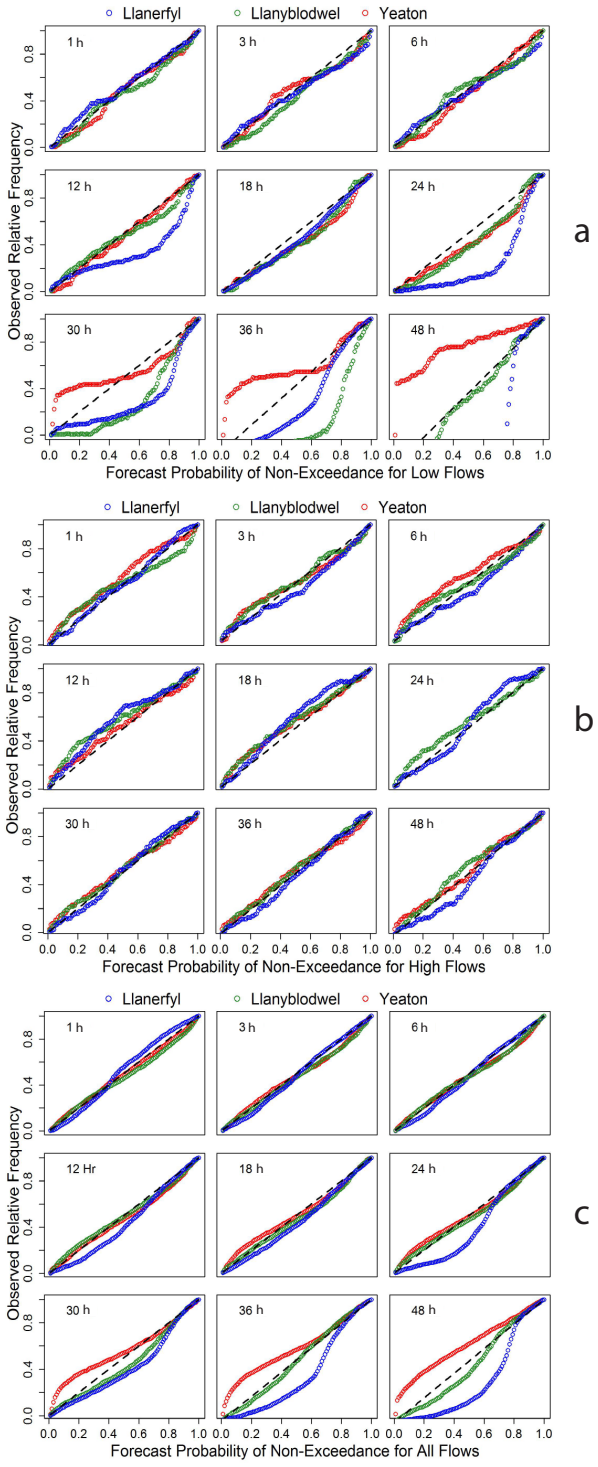


Figure 6: Reliability diagram from Upper Severn subcatchments for (a) low, (b) high and (c) all flows (Llanerfyl: blue; Llanyblodwel: green; Yeaton: red).

models for hydrologic (DODO; Wallingford, 1994) and hydrodynamic routing (ISIS; Wallingford,

1997). The rainfall input for the MFFS is acquired from ground measurements via rain gauges, from radar measurements or from numerical weather prediction data. The MFFS predicts ahead in time the response of the Upper Severn subcatchments but, as expected, the quality of forecast deteriorates with increasing lead time.

To do uncertainty analysis for the MFFS, hindcasting or reforecasting is done and then results are compared to the observed data. All the input time series used for hindcasting are taken from measured data. In this study, the reforecasting period was kept equal to the one employed in the studies of López López et al. (2014) and Dogulu et al. (2015). The chosen period is from 1 January 2006 to 7 March 2013. Data in the period till 6 March 2007 are used for the model spin-up. The remaining period is used for the calibration and validation of the uncertainty estimation techniques. Forecasts are made on a 12-hourly basis —at 08:00 and 20:00 daily, up to a lead time of 48 h. kNN resampling was applied for forecasts at 10 different lead times: 1, 3, 6, 9, 12, 18, 30, 36, 42, and 48 h. To choose an input variable vector for kNN resampling, correlation analysis was done between residual error and contenders for input variable vector space, namely simulated water level (H_{sim}), measured water level (H_{obs}) and residual error (E_{obs}) from various time steps t . The analysis was done to assist in a manual selection of input variable vectors. The correlation between residual error and water level reduces fast with time lag between the two time series. Therefore it is enough to choose relatively simple and small dimensional input variable vector spaces. For lead times, l, up to 6 h we chose

$$\boldsymbol{v} = [H_t^{sim}, H_{t-l}^{obs}, e_{t-l}^{obs}] \tag{15}$$

For higher lead times, uncertainty has only been conditioned on H_t^{sim} as the residual error becomes less and less correlated with variable values as measured several hours behind the prediction time.

$$\boldsymbol{v} = [H_t^{sim}] \tag{16}$$

Ninety-nine values of residual errors were sampled from the nearest neighbourhood to generate an empirical distribution at each prediction step. This allowed us to get the "resolution" of 1 percentile in the generated empirical distribution. To develop confidence in the chosen value of k, we checked for a few prediction steps how sensitive the generated empirical distribution is to the value of k. Four different instances of $\boldsymbol{v}_t$ were chosen. Each instance represents a prediction step in the input variable vector space (the red circle in Fig. 1), with different hydrologic conditions. The plots of the cumulative mean square difference between probability density functions (pdfs) of varying k were generated. Cumulative mean square difference (Eq. 17) serves as an index to show how much the empirical pdfs change with changing k. We get a decreasing slope with increasing k. It shows that the pdfs become almost identical for values of k around 100. If $P_{k_i}(e)$ is the probability density for a residual error e calculated through k_i nearest neighbours using kNN resampling, for probability functions corresponding to discrete bin size δe, the cumulative difference is defined as cumulative difference

$$\text{cumulative difference} = \sum_{k_i=10}^{k_i=k} \sum_{first\ e\ bin}^{last\ e\ bin} [\Delta e \cdot P_{k_i}(e) - \Delta e \cdot P_{k_{i-1}}(e)] \tag{17}$$

The various values of k_i that were tested are 10, 30, 50, 70, 90, 100, 110, 130 and 150. Using the information from Fig. 3, a value of k=99 does not seem to be heavily affected by sampling errors. Nevertheless, it is not a mathematically calibrated value of k and therefore is likely to be sub-optimal. However, it should still be able to provide reasonably representative samples from the error distribution, as is suggested by Fig. 3.

Results

Figure 4 shows two hydrographs for the same event, where model predictions were made at different lead times. From the graph of lead time 48 h it is evident that the error quantiles that

kNN resampling produces are not forced to have zero mean. Therefore the model prediction can sometimes lie outside the predicted quantiles. This is because kNN resampling learns from past instances where the model has consistently underpredicted or overpredicted the flow, so it corrects for this bias. The hydrographs capture the low flows and the peaks well. It can also be seen that for high flows the errors are usually higher than for medium and low flows. The residual error distribution is thus heteroscedastic, i.e. the variance depends on the magnitude of the predicted flow. The autocorrelation can be checked by plotting errors versus time, whereas performance of an error model with regard to heteroscedasticity can be estimated by plotting reliability diagrams for different magnitudes of flow, which would mean different water levels in this case.The plotting of error time series (Fig. 5) for various lead times shows some recurring trends across all three subcatchments. The errors are small for small lead time forecasts and the spread of error time series increases with increasing lead time. Moreover, the errors do not look autocorrelated for smaller lead times, whereas for the higher lead times autocorrelation becomes more prominent. This can be ascribed to the memory of the hydrologic system. If the system response is higher than what the model simulates for a particular lead time, then the system response is likely to be higher for the next time step as well. As the errors become larger, they tend to lose their independence property. This is captured by the error samples generated by kNN resampling as well. The rate at which autocorrelation deteriorates for observed residual errors corresponds well to the kNN resampling error samples‘ autocorrelation. It can be seen that kNN resampling preserves the autocorrelation in the error time series without using an autoregressive model.

To check the performance of kNN resampling for various flow magnitudes, the simulation values were divided into low and high flows —the lowest and highest 10% of water levels simulated in the validation phase respectively. The reliability diagrams (Fig. 6) show that the overall performance of error quantiles for all water levels is good for low and medium lead times. The reliability decreases with high lead times (24 h and above). The reliability plots show that kNN resampling performs better for high flows compared to low flows, even for higher lead times. For low flows and high lead times, the forecast probability of non-exceedance is higher than the observed relative frequency. Nevertheless, from 0.90 probability of non-exceedance and above, the reliability curve comes back to the desired 45 line. For flood forecasting it is important to model the high and medium flows well. kNN resampling delivers quite reliable quantiles for such flow regimes. The deteriorating model performance with higher lead times gets reflected in the performance of kNN resampling quantiles as well.

To assess the performance of kNN resampling relative to other established post-processor uncertainty estimation techniques, comparisons with QR and UNEEC have been carried out. The results for QR have been taken from López López et al. (2014) and the results for UNEEC from Dogulu et al. (2015). QR results for uncertainty estimation were available for all the lead times as done using kNN resampling and, from UNEEC, only for lead times of 1, 3, 6, 9, 12, and 24 h. Values of PICP and MPI are shown in Fig. 7, together with results from UNEEC and QR. The Alpha Index (α) is reported for several lead times in Table 2. As expected, the MPI of all the uncertainty estimation techniques increases with increasing lead time. Comparison between kNN resampling and QR has been made for three locations and 10 lead times in the validation period. Model simulations were run two times each day. Verification indices for uncertainty analysis were calculated separately for each lead time and each location. Considering the 90 and 50% quantiles as two prediction intervals, this allowed for the evaluation of PICP and MPI 60 times (Fig. 7). kNN resampling has a higher PICP in 67% of the cases and a smaller MPI for 73% of the cases. A comparison between kNN resampling and UNEEC was made for three locations and five lead times for the validation. For each location and each lead time, the 90 and 50% quantiles were generated, which allowed for the evaluation of PICP and MPI 30 times (Fig. 7). The PICP of kNN resampling is higher in 60% of the cases and the MPI is smaller in 36% of the cases. Based on these results we concluded that, for this case study, kNN resampling generally produces narrower confidence bands and provides a better coverage of the probability distribution than the other methods in the majority of forecasts, especially showing better performance for the larger lead times.

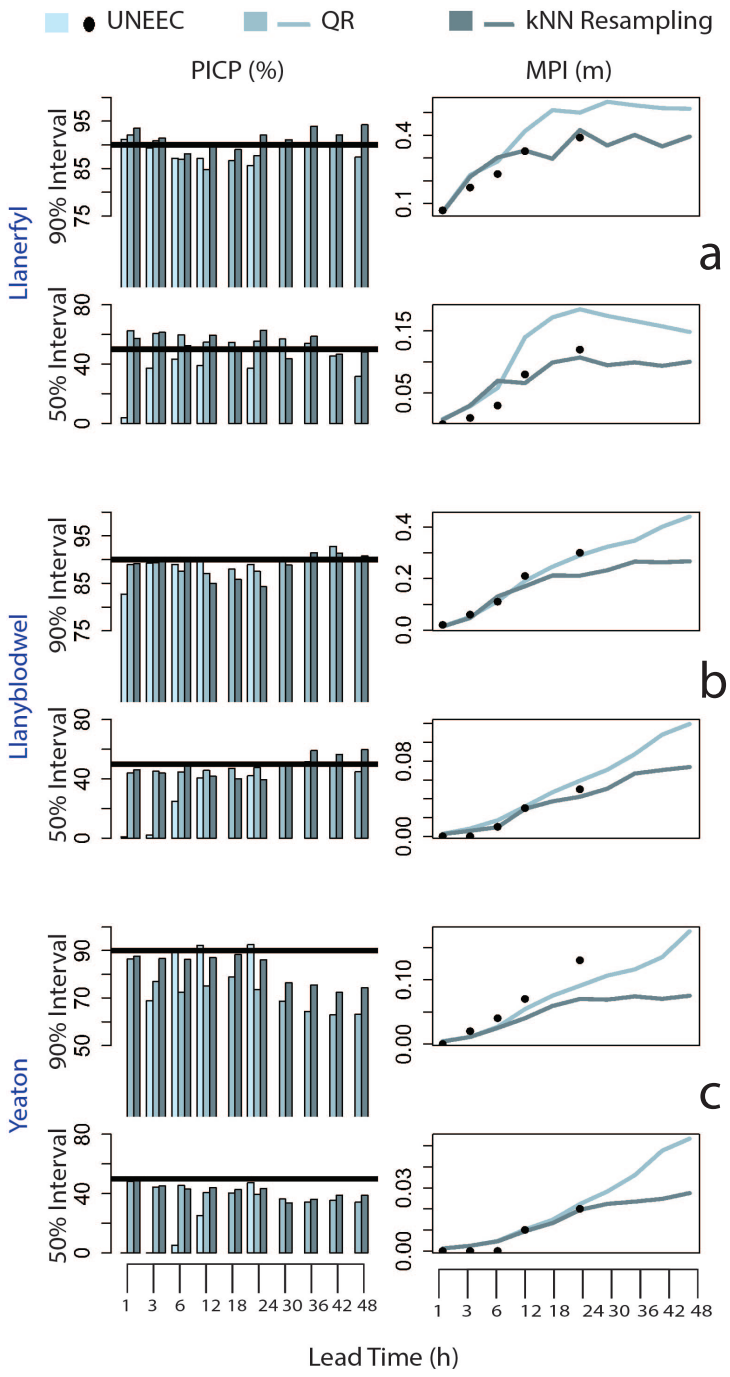


Figure 7: Reliability diagram from Upper Severn subcatchments for (a) low, (b) high and (c) all flows (Llanerfyl: blue; Llanyblodwel: green; Yeaton: red).

3.2 River Brue

Catchment description

The River Brue, located in the south-west of England, has a history of severe flooding. The test forecast location used in this study is Lovington, where the upstream catchment area is 135 km^2 (Fig. 8). The catchment is predominantly rural and the soil consists of clay and sand. This kind of

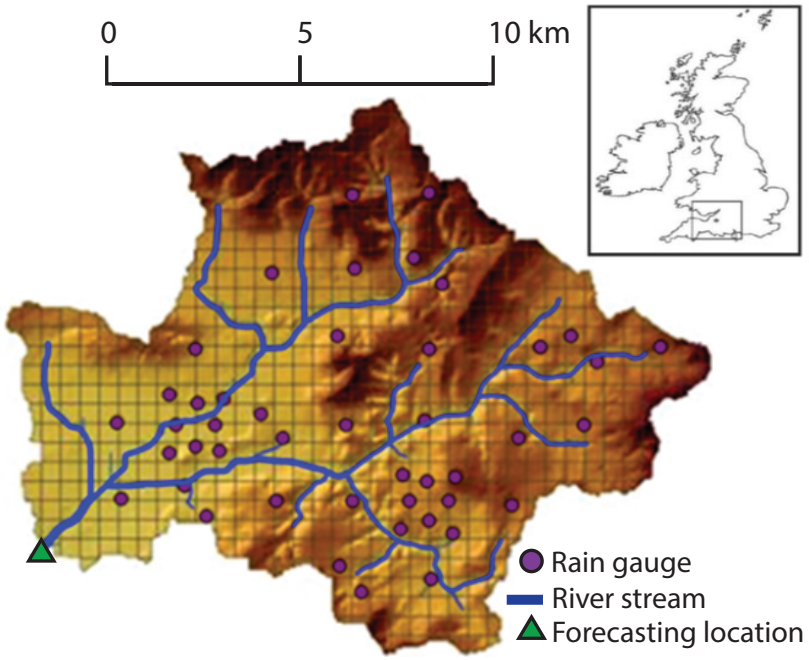


Figure 8: Brue catchment (from Shrestha and Solomatine, 2008).

soil and the modest relief give rise to a slowly responsive flow regime. The mean annual rainfall in the catchment is 867 mm; the mean river flow is 1.92 m^3 s1 and has a maximum flow of 39.58m3 s1. This catchment has been extensively used for research on weather radar, quantitative precipitation forecasting and hydrologic modelling.

Experimental set-up

For the Brue catchment the simplified version of the HBV rainfall-runoff model has been used (Bergström, 1976). The HBV-96 model is a lumped conceptual model (Lindström et al., 1997). Like most other conceptual models, HBV consists of subroutines for snow accumulation and melt, soil moisture accounting and surface runoff, and employs a simple routing scheme. The input for the HBV model consists of precipitation (basin average), air temperature and potential evapotranspiration (estimated by the modified Penman method using automatic weather data available). Historical input data are available for a period of 1994-1996. Predictions are only made for 1 h lead time. Uncertainty analysis is done for a chosen period from 24 June 1994 to 31 May 1996. Hindcasts were made on a daily basis, using a warm state from a historical run. The hindcasts were split into calibration and validation set at 24 June 1995 for the uncertainty estimation techniques. The calibration data set was used to calibrate (train) UNEEC and QR, and for the resampling of

Table 2: Alpha Index (α) for high flows corresponding to different lead times of Upper Severn subcatchments.

Lead time (h)	1	12	24	48
Llanerfyl	0.92	0.87	0.79	0.64
Llanyblodwel	0.93	0.95	0.93	0.90
Yeaton	0.97	0.94	0.94	0.75

errors using the kNN algorithm. The resampled errors were used to estimate prediction intervals for the predictions from the validation data set. Each of the two data sets represents almost a full year of observations. Three input variable vectors were chosen based on the results of correlation analysis, from simple to complex. This allows us to study the dependence on the choice of search space. Input variable vector (ivv) 3 for kNN resampling and UNEEC is the same for this comparison, whereas QR only uses Qsim (Dogulu et al., 2015). The three input variable vectors used are where R is the effective rainfall, Q is the discharge and e is the residual error. Considering t as the prediction time, then the subscripts of the various input variables represent the time and the superscripts sim and obs mean they are simulated and observed values respectively. The number of nearest neighbours was chosen to be 99 and 199, to analyse its influence on uncertainty quantification. Uncertainty analysis was done for a calibrated HBV model as well as a model with a unit systematic bias. The bias was introduced to the simulation results of the calibrated model by simple addition. The aim of a biased model for uncertainty quantification using kNN resampling is to assess the performance of kNN resampling when the residuals are not zero mean.

$$\boldsymbol{v}_{(ivv1)} = [Q_t^{sim}] \tag{18}$$

$$\boldsymbol{v}_{(ivv2)} = [Q_t^{sim}, e_{t-l}^{obs}] \tag{19}$$

$$\boldsymbol{v}_{(ivv3)} = [R_{t-8}^{sim}, R_{t-9}^{sim}, R_{t-10}^{sim}, Q_{t-1}^{obs}, Q_{t-2}^{obs}, Q_{t-3}^{obs}, e_{t-1}^{obs}, e_{t-2}^{obs}] \tag{20}$$

Results

Table 3: Performance of various uncertainty estimation techniques for the Brue catchment. For kNN resampling and UNEEC the same input variable vector is used (Eq. 20). For QR only Q_{sim} is used.

	PICP (expected 90%)			MPI (m^3s^-1)		
	UNEEC	QR	kNN	UNEEC	QR	kNN
Calibration	91.19	90.00	86.30	1.58	1.69	0.51
Validation	88.29	82.33	84.42	1.37	1.39	0.21

kNN resampling was applied to a single historical simulation and compared to observations. The simulated hydrographs for the highest discharge event with 50 and 90% prediction intervals are shown in Fig. 9. The residual distribution of kNN resampling is generally a non-zero mean. Therefore we see that the prediction intervals may sometimes deviate from the deterministic model prediction quite significantly. The ability of kNN resampling to search for similar hydrologic conditions, like rainfall, and discharge in the past, and to learn from the residuals, allows it to make more representative error distributions. For example, in Fig. 9, the falling limb of the hydrograph shows that the prediction band generated by kNN resampling captures the observed flow for input variable vectors 2 and 3, even though the model shows a noticeable mismatch with the measurements. This can be explained by considering the history of errors that the model made during such hydrologic conditions in the past. And as kNN resampling learns that the model consistently underestimates in such cases, the corresponding error distribution corrects for this bias. The results of the PICP and MPI are shown in Table 3 together with results from UNEEC and QR (Dogulu et al., 2015). As can be seen from the table, kNN resampling's performance is comparable to that of UNEEC and QR for this case study. The prediction intervals generated by kNN resampling are smaller, compared to the other two uncertainty estimation techniques, while

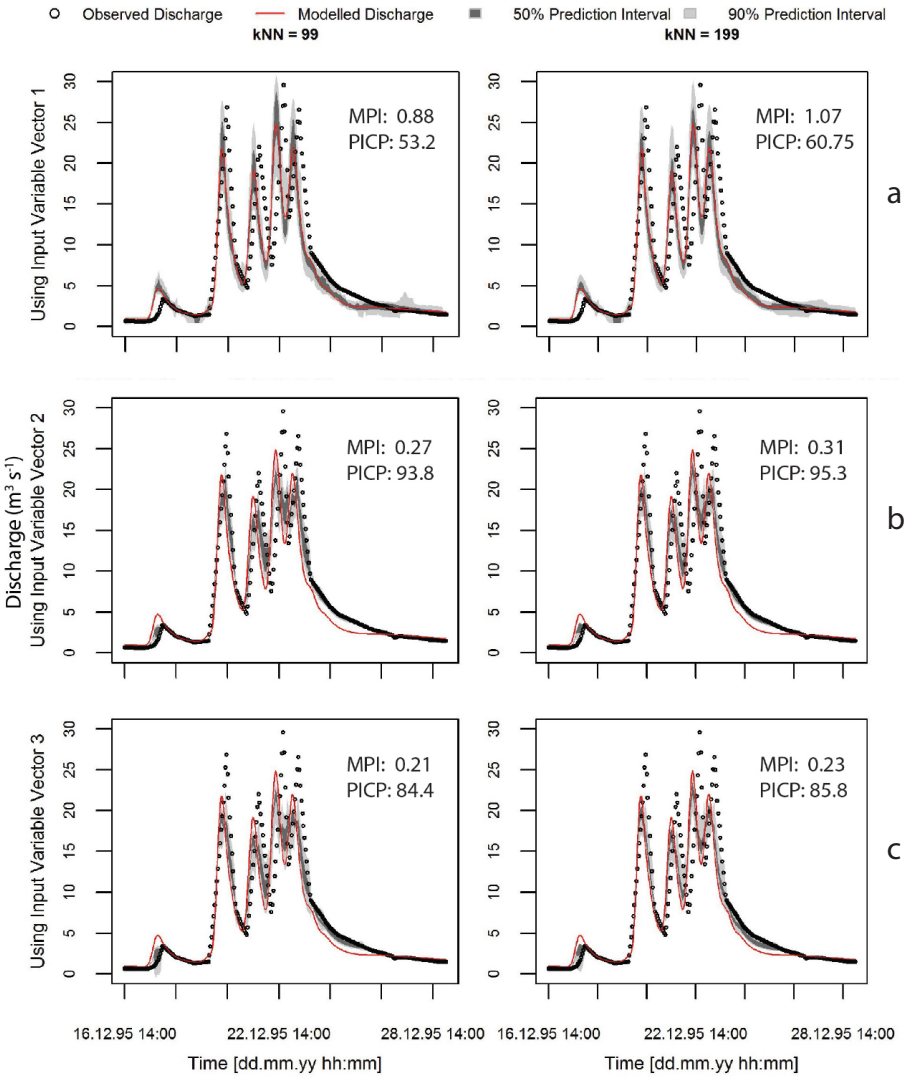


Figure 9: The 50 and 90% prediction intervals for the Brue catchment using kNN resampling. The hydrographs are shown for two different k values (99, 199) and three different input variable vectors given by (a) Eq. (18), (b) Eq. (19) and (c) Eq. (20). This is the largest event in the validation time series. (The 50% prediction interval is the interval between the 25 and 75% quantiles of residual error, and the 90% quantile is the interval between the 5 and 95% quantiles. MPI and PICP correspond to the whole validation time series.)

the coverage probability is similar. It indicates that kNN resampling is able to learn well from past data and condition the probability of residual errors well. The Alpha Index for the validation phase is also high (0.96). We also notice that three different input variable vectors show different degrees of performance (Fig. 9). The past errors, e_{t-1}^{obs}, seem to be informative in this case, providing very narrow conditional error probabilities.

Apart from evaluating the usability of kNN resampling for calibrated models, the performance of kNN resampling quantiles generated by kNN resampling for a model with systematic bias was also checked. Figure 10 shows that the performance of kNN resampling does not diminish under systematic bias. The reliability of the generated quantiles remains almost unfazed. As a systematic bias will not affect the autocorrelation structure of the residual errors, the autocorrelation of error

samples generated through kNN resampling also remains unchanged. Nevertheless, we see a shift in the mean of the sample time series, which is roughly equal to unity. The reliability of quantiles generated using kNN resampling for high flows (highest 10% in the validation period) is poorer than for all flows. The invariance of kNN resampling performance to model bias makes it a robust postprocessor technique; however, unlike in the case of Upper Severn subcatchments, the technique's performance diminishes for high flows.

4 Discussion and conclusions

The application of kNN resampling to two case studies shows that the forecast uncertainty intervals are relatively narrow and still capture the observations well. The expected increase in uncertainty for longer lead times is also reproduced well and the probability coverage of kNN resampling remains good, as verified from historical observations. This is in accordance with previous research (Sikorska et al., 2015). The error samples generated by kNN resampling reproduce two important characteristics of residual errors in hydrologic models, namely autocorrelation and heteroscedasticity. Also, for applications to flood modelling, the high flows are most important and the uncertainty quantification by kNN resampling for the Upper Severn shows reasonable reliability for this high-flow regime. For the Brue, the performance is poorer. This can be attributed to the inadequacy of representative high flows in the calibration phase in combination with the choice of the input variable vector. The highest flow in calibration time series is $15.4m^3/\text{s}$, whereas in validation time series it is $29.9m^3/\text{s}$. It is also shown that the technique is generally robust to the performance of the underlying deterministic model. If the model has systematic biases, kNN resampling learns from the past errors of the model and recreates the systematic bias in the empirical error distribution mean, thus maintaining the performance of prediction intervals. Our results on systematic error correction by kNN resampling substantiate the findings from previous research on forecast updating using kNN (Akbari and Afshar, 2014). These findings from this study are confirmed by three quantitative indicators of forecast reliability. The comparison of kNN resampling uncertainty estimates to those generated by QR and UNEEC shows that the mean prediction intervals (MPIs) generated by kNN resampling are generally smaller. A significantly smaller MPI using kNN resampling, as in the case of the Brue, is in part due to the conditioning on input variable vectors as compared to UNEEC and QR. As the values of k in this study have been restricted to 99 and 199, the error distribution tends to be much narrower than the marginal error distribution. The conditional distribution will turn into a marginal distribution when the number of k is equal to the time steps in the calibration time series. A more quantitative dependence on the k value and MPI will need further research. Apart from a narrow MPI, we also find that kNN resampling is generally able to capture the expected ratio of observations within its intervals (PICP) most of the time, or at least be close to the expected value.As in the case of all other data-driven methods, the applicability of kNN resampling depends on the availability of sufficiently long and representative historical forecasts and observations. The historical series should include several occurrences of forecasting situations that are similar to the current situation. In extreme cases, the kind of kNN search proposed here will select the most similar historical situations which may or may not be representative of the current situation. In contrast to the methods like QR and UNEEC that build explicit predictive regression models which are able to extrapolate for the data which are beyond the limits of the calibration (training set), kNN resampling does not extrapolate. This could be seen as a disadvantage. On the other hand, however, the extrapolation that is done by regression techniques could also be seen as doubtful. It is not a given that the most extreme historical situations are less representative of the uncertainty of an extremely high flow than an extrapolated result. The results in this paper show that kNN resampling has a good or poor reliability for the highest values in the validation set, depending on the case study and the choice of input variable vector. Due to the non-parametric nature of kNN resampling, the increasing variance of residual errors for higher values of predictands is generally adequately taken into account. As kNN resampling, like other post-processors, learns about the residual error process from the past, the historical records should be representative of the current forecast conditions. In changing condi-

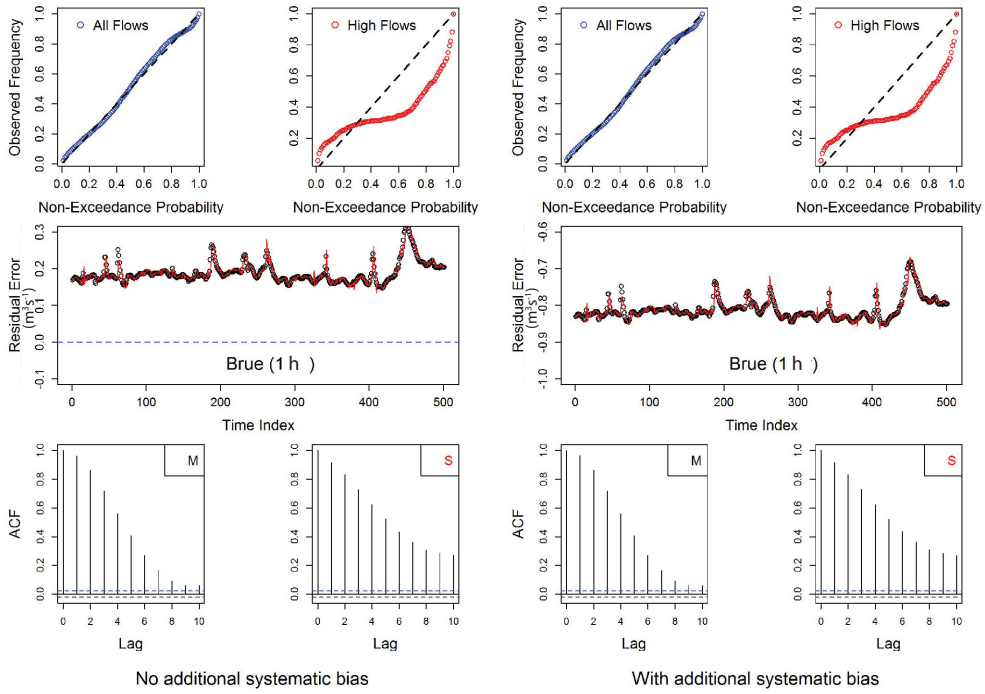


Figure 10: Effect on reliability of quantiles and autocorrelation of error samples on adding a systematic bias to the model artificially. kNN samples, generated using input variable vector 3 (Eq. 20), are plotted in red, and observed errors in black circles. M stands for measured and S for simulated.

tions, this may not be true. Changing conditions may be caused, for example, by climate change or more local changes in the catchment like deforestation and dam building. This is a common problem for all data-driven statistical estimators and is not unique to kNN resampling. Care needs to be taken to use data time series which do not outright violate the assumptions regarding the invariance of catchment and climate.

One of the few calibration parameters of kNN resampling is the number of nearest neighbours k. In this study, k has been chosen by a simple heuristic technique. For optimal performance, it would be advisable to calibrate k for each application in a more systematic way. We do show for Brue that the sensitivity of the uncertainty intervals to the value of k is not significant, when changing it from 99 to 199. However, we also expect that the optimal value of k will depend on the length of the historical data series and on the uncertainty quantiles of interest. In the context of search space, in this research, the input variable vector has been chosen by correlation analysis. It can be recommended to use more sophisticated procedures for real-life applications, which can capture the non-linear dependence between the error process and input variable vector candidates. Improvements in performance can possibly be achieved by seeking a better set of input variables for each forecast location and lead time of interest.

In conclusion, kNN resampling can be considered a relatively simple machine learning technique to predict hydrologic residual uncertainty. The errors from the similar hydrologic conditions in the past are used as samples for the residual error probability distribution and the samples are collected by a k nearest-neighbour search. The application of this technique to case studies Brue and Upper Severn subcatchments has shown promising results. In comparison to many other data-driven techniques, kNN resampling has the advantage of avoiding assumptions about the nature of the residual error distribution: the instance-based learning approach is non-parametric and non-

regressive and requires little calibration. The method was shown to be able to quantify hydrologic uncertainty to an accuracy that is comparable to other techniques like QR and UNEEC. Given the relatively small effort in setting up the method, the performance of kNN resampling in uncertainty quantification is more than acceptable when compared to other post-processor error models.

Acknowledgement: The authors would like to acknowledge Bonneville Power Administration, Portland, USA, for supporting this research. The UK Environment Agency is acknowledged for provision of the data for the case studies described in this paper. Many thanks to Nilay Dogulu, Patricia López López, Marijn Swenne, and Azam Iftikhar for their help during the course of this research. We are thankful to Sven Eggimann and Heuning Badger in helping structure the paper. Finally, we are grateful to the editor, Bettina Schaefli, and two reviewers for their valuable comments. Part of this study was supported by EC FP7 project WeSenseIt (Citizen Observatory of Water), grant agreement no. 308429, by the Russian Science Foundation (grant no. 17-77-30006), and project QUICS (Quantifying Uncertainty in Integrated Catchment Studies), grant agreement no. 607000.

Chapter 3

Using a simple post-processor to predict residual uncertainty for multiple hydrological model outputs

Lennart B. Ehlers[1,2], Omar Wani,[3,4], Julian Koch,[1,2], Torben Sonnenborg,[1,2],Jens Christian Refsgaard[1,2]

[1] Geological Survey of Denmark and Greenland, Department of Hydrology, Copenhagen, Denmark
[2] Copenhagen University, Department of Geosciences and Natural Resource Management, Copenhagen, Denmark
[3] Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland
[4] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

Journal: Advances in Water Resources

Submission date: 29th April 2018 (under review)

Author contributions: LBE and OW designed the simulation experiments. OW and LBE wrote the code for uncertainty analysis. LBE conducted the analysis, in consultation with OW and JCR. LBE wrote and revised the manuscript with contributions from OW. JK helped with model calibration. All authors reviewed and commented on the manuscript and the revision.

Abstract

Regardless of the complexity of the hydrological model employed, uncertainty assessment (UA) is predominantly performed for the aggregated catchment response discharge. For coupled integrated models that simulate various hydrological states and fluxes on a grid cell basis, this represents a severe shortcoming. In order to provide water resource management and decision makers with an uncertainty estimate for spatially distributed hydrological variables, their uncertainty needs to be assessed at interior catchment locations. In recent years, advances in terms of computer power and algorithm efficiency have led to an increasing usage of UA techniques based on Monte Carlo simulation that allow the disaggregation of uncertainty into its contributing sources. In spite of this development, application of these techniques to the most complex coupled integrated hydrological models commonly employed in operational hydrology remains a challenge because of their exceptionally high computational demands. Therefore, we test a simple data-driven technique (k-NN resampling) to evaluate its ability to provide reliable residual uncertainty estimates for the multi-variable (discharge, hydraulic head, soil moisture and actual evapotranspiration), deterministic output of two coupled groundwater-surface water models with different complexities. K-NN resampling can be applied to both manually calibrated and uncalibrated hydrological models. Being a nonparametric method, no explicit prior assumptions about the error distribution of different hydrological variables are required. When conditioning the algorithm, we propose to limit the number of error lags to be included based on inspection of the partial autocorrelation function (PACF). Our results corroborate previous findings regarding reliability and robustness of the k-NN technique: The 90% prediction intervals (PI) capture the observations in the testing period satisfactorily for all hydrological variables (92.6-97.3%), while Alpha indices (0.84-0.95) indicate very reliable PIs for all error quantiles. Differences in error structure between hydrological variables are successfully inferred from historical data and reflected in the results. We conclude that k-NN resampling represents a potent, cost-efficient UA technique for applications in operational hydrology, facilitating a near-simultaneous, easy uncertainty assessment for various outputs of computationally heavy hydrological models.

1 Introduction

The intricate dynamics of natural hydrological processes are difficult to reproduce using even the most sophisticated, physically based hydrological models and model equations. As hydrological models represent mere approximations of real-world systems, it follows that model predictions are uncertain, cf. e.g. Keith Beven (2012). In order to provide a reliable basis for decision makers, model outputs need to be accompanied by an estimate of the uncertainties associated with them. As research conducted in the past two decades ascertains, effectiveness and sustainability of water resource management critically depend on a thorough uncertainty assessment (UA), see for example K. Beven (2002), Refsgaard and Henriksen (2004), Wagener and Gupta (2005) and Ajami, Hornberger, and Sunding (2008). Accordingly, quantification of uncertainties in hydrological modeling is considered good scientific practice, e.g. Refsgaard, van der Sluijs, Hojberg, and Vanrolleghem (2007).

Among the first to recognize the necessity of acknowledging the presence of uncertainty in hydrological model simulations were K. Beven and Binley (1992). They developed the Generalized Likelihood Uncertainty Estimator (GLUE), which uses Monte Carlo (MC) simulation. The authors coined the term "equifinality", assuming that model imperfection and incomplete process understanding entail that a potentially endless number of hydrological models and parameter combinations will result in equally acceptable model performance. However, GLUE has been shown to have limitations due to the usage of an informal likelihood function for model parameter identification, cf. Mantovan and Todini (2006) and Stedinger, Vogel, Lee, and Batchelder (2008).

With the aim to quantify the contributions of different sources of uncertainty to predictive uncertainty individually, a number of formal UA approaches based on Bayesian inference were introduced. In this context, particular attention was given to the explicit consideration of input uncertainty (rainfall). A prominent example is the Bayesian Total Error Analysis (BATEA) (Kavetski, Kuczera, & Franks, 2006a, 2006b), in which uncertainty in rainfall is accounted for by means of a rainfall multiplier. While often neglected, model structural uncertainty is a major source of predictive uncertainty, see for example Refsgaard, van der Sluijs, Brown, and van der Keur (2006) for a comprehensive review on this matter. In contrast, Ajami, Duan, and Sorooshian (2007) developed the Integrated Bayesian Uncertainty Estimator (IBUNE) which allows combining the rainfall multiplier concept with consideration of model structural deficiencies by using a variety of model structures by means of Bayesian Model Averaging (BMA), see Hoeting, Madigan, Raftery, and Volinsky (1999). However, finding an adequately representative likelihood function for hydrology capturing autocorrelated, non-normal and heteroscedastic errors is a matter of ongoing research, e.g. Schoups and Vrugt (2010), McInerney, Thyer, Kavetski, Lerat, and Kuczera (2017) and Han and Zheng (2018).

Bayesian approaches and more generally speaking MC-based UA techniques require a vast number of model evaluations (orders of magnitude of $10^3 - 10^5$). The parameter samples generated during inference need to be propagated through the hydrological model until convergence is diagnosed. In operational hydrology, this can represent a serious obstacle for several reasons. For example, when commercial software is employed to perform hydrological simulations, parallelization may be limited by license availability. In addition, while Bayesian methods are often tested using fast, conceptual rainfall-runoff models, computational costs are incrementally larger when applied to complex coupled integrated hydrological models. Although likely only being a matter of time, this currently still restricts their applicability, e.g. Juston et al. (2013).

For this reasons, computationally less demanding UA methods represent an attractive alternative, especially for applications in operational hydrology. To this end, one possibility is to resort to meta- modelling in which outputs of a complex hydrological models (simulator) are being approximated through a surrogate model (emulator), cf. O'Hagan (2006). The emulator is obtained by establishing a statistical relationship between the model outputs (Reichert, White, Bayarri, & Pitman, 2011). A recent example is given by Carbajal, Leitao, Albert, and Rieckermann (2017), who

compared the performance of data-driven and mechanistic emulators in an urban drainage context. Similarly, Miller, Berg, Davison, Sudicky, and Forsyth (2018) use polynomial chaos expansion (conceptually being an emulator) for uncertainty quantification of an integrated surface-subsurface hydrological model. Another possibility for more time-efficient UA is to apply post-processing techniques to (deterministic) hydrological model runs. Post-processors are based on the assumption that the model parameters are identifiable using optimization of a cost function, which entails an implicit likelihood function. Uncertainty estimation is then based on the application of data-driven techniques that require comprehensive historical data to generate anticipated model errors. This would be equivalent to producing samples from a likelihood function, for a defined set of model parameters. Post-processors have been applied in various hydrological modeling studies, mostly in conjunction with short-term streamflow forecasts. Prominent examples include Quantile Regression (QR) (Weerts, Winsemius, & Verkade, 2011), Estimation Based on Local Errors and Clustering (UNEEC) (Solomatine & Shrestha, 2009) or the Dynamic Uncertainty Model by Regression on Absolute Error (DUMBRAE) method (Pianosi & Raso, 2012). Despite their relative conceptual simplicity, post-processors are generally robust, see for example Pokhrel, Robertson, and Wang (2013). Evin, Thyer, Kavetski, McInerney, and Kuczera (2014) even report a post-processor approach to outperform a joint Bayesian technique in terms of robustness in the presence of error heteroscedasticity and autocorrelation.

Among post-processing techniques, k-nearest neighbor (k-NN) resampling is comparatively simple and easy to implement. Its univariate predecessor was introduced to the hydrological modeling community by Karlsson and Yakowitz (1987), who used it as a tool to predict streamflow. K-NN resampling is a nonparametric method, which implies that no explicit assumptions (Gaussian, zero mean) about the nature of model error are made, see for example Hollander, A. Wolfe, and Chicken (2015). Instead, the error structure is reproduced by sampling realizations from the space of historical model errors, resulting in multiple local approximations of the complex target function during prediction (Solomatine, Maskey, & Shrestha, 2008). Recent applications of k-NN in hydrology include work by Anna E. Sikorska, Montanari, and Koutsoyiannis (2015), Beckers, Weerts, Tijdeman, and Welles (2016) and Wani, Beckers, Weerts, and Solomatine (2017). Anna E. Sikorska et al. (2015) used k-NN resampling for uncertainty estimation while accounting for input uncertainty (by means of an ensemble) and parameter uncertainty separately. The authors found the method to outperform a formal statistical meta-Gaussian technique devised by Montanari and Brath (2004). Beckers et al. (2016) adopted k-NN to select ensemble streamflow prediction (ESP) traces from original ESPs based on climate index values, prior to generating new ensemble traces in a resampling procedure. In Wani et al. (2017), k-NN resampling was applied to estimate residual uncertainty for streamflow forecasting. A comparison with two more complex post-processors (QR and UNEEC) showed that k-NN, in spite of its simplicity, performed well in terms of accuracy and reliability of its uncertainty estimation. This work represent an application of the k-NN configuration detailed in Wani et al. (2017).

In contrast to the majority of cases in which k-NN has been applied as a means of generating short-term streamflow predictions in a forecasting context, we employ k-NN resampling in a hindcasting exercise, in order to perform uncertainty analysis for a complex coupled groundwater-surface water model. A major shortcoming common to the examples listed above is that UA is generally restricted to the aggregated catchment response discharge. While this may suffice for flood forecasting applications, this is unsatisfactory for coupled integrated models that generate a multitude of outputs. It is obvious that, in order to gain a deeper understanding of hydrological processes that occur within the catchment, calibration and evaluation have to be extended by suitable hydrological variables that tell something about the model performance in the spatial domain, cf. Pokhrel and Gupta (2011) and Koch, Siemann, Stisen, and Sheffield (2016). Accordingly, this notion applies to UA as well.

Therefore, the main objective of this study is to test the potential of k-NN resampling regarding the generation of reliable residual uncertainty intervals for the multi-variable output of a coupled integrated hydrological model. The hydrological variables incorporated in the UA were discharge,

hydraulic heads, soil moisture and actual evapotranspiration, which in addition represents model response at interior catchment locations. In doing so, we take a step further towards fully exploiting the full potential of this class of hydrological models. This will furthermore allow the communication of the associated uncertainties to public and end-users, which in turn may lead to an increased level of trust in scientific results (Juston et al., 2013). Moreover, we contribute to a formalization of the process of composing the k-NN input variable vector (ivv) by suggesting a hierarchical approach that is primarily based on analysis of (partial) autocorrelation of model errors. Using a flexible and fast post-processing technique coded in freely available software for uncertainty quantification (R Development Core Team, 2018), we do our part to further establish UA as a routine in hydrology (Pappenberger & Beven, 2006), especially with regard to complex hydrological models whose computational costs preclude MC-simulation.

2 Study area

The study area is the 1055 km^2 Ahlergaarde catchment on the Jutland peninsula in Western Denmark (Fig. 1). The catchment is the headwater of the Skjern River, which has its source at the Jutland Ridge towards the eastern boundary of the catchment and discharges into the North Sea via RingkÃ¸bing Fjord. The annual mean discharge is 37 m^3/s (Ovesen et al., 2000), making the Skjern the largest river in Denmark in terms of discharge volume. Geologically, the Ahlergaarde catchment is a former glacial outwash plain. The surface is largely made up of Quaternary sandy sediments, while the subsurface geology predominantly consists of pre-Quaternary clay which in several locations contains Miocene sand lenses (Scharling, Rasmussen, Sonnenborg, Engesgaard, & Hinsby, 2009). The topography is mostly flat, with a maximum elevation at 130 m.a.s.l. in the east of the catchment. The climate is temperate maritime with an annual mean temperature of 8.2°C, annual precipitation of 1050 mm and annual reference evapotranspiration of 570 mm (Stisen, Sonnenborg, Hojberg, Troldborg, & Refsgaard, 2011). Due to its topographical, geological and climate characteristics, the Ahlergaarde catchment exhibits high groundwater recharge rates and a hydrogeological system that is mostly groundwater controlled. The Ahlergaarde catchment is part of the Danish Hydrological Observatory HOBE (Jensen & Illangasekare, 2011).

3 Methodology

3.1 Hydrological models

We employ two hydrological models in this study, both of them based on the MIKE SHE code (Abbott, Bathurst, Cunge, Oconnell, & Rasmussen, 1986; Graham & Butts, 2005). Both models belong to the class of fully integrated, spatially distributed hydrological models and feature fully coupled modules describing 3D groundwater flow, overland flow, river routing and land surface processes. Their model setup is largely based on the National Water Resources model of Denmark (DK-Model), which was developed and is maintained by the Geological Survey of Denmark (Hojberg, Troldborg, Stisen, Christensen, & Henriksen, 2013). Climate data are provided as 10×10 km grids for rainfall and 20×20 km grids for reference evapotranspiration and air temperature. The gridded rainfall product represents dynamically corrected rainfall, accounting for undercatch (Stisen et al., 2012). The two models (henceforth referred to as "Model 1" and "Model 2") are characterized by two different levels of complexity, translating into notably diverging simulation run times (Tab. 1). Moreover, the models represent two different calibration paradigms: a traditional (calibration based on discharge and hydraulic heads) versus a multi-objective calibration scheme (calibration additionally based on spatially distributed soil moisture, remote sensing based evapotranspiration and soil surface temperature, Tab. 1). While testing the capability of k-NN resampling to perform uncertainty assessment for multiple hydrological model outputs, we also intend to answer the question of whether or not k-NN performs equally good for both hydrological model set-ups. Specifically, we are interested in comparing k-NN performance for hydrological variables that were not part of calibration in Model 1.

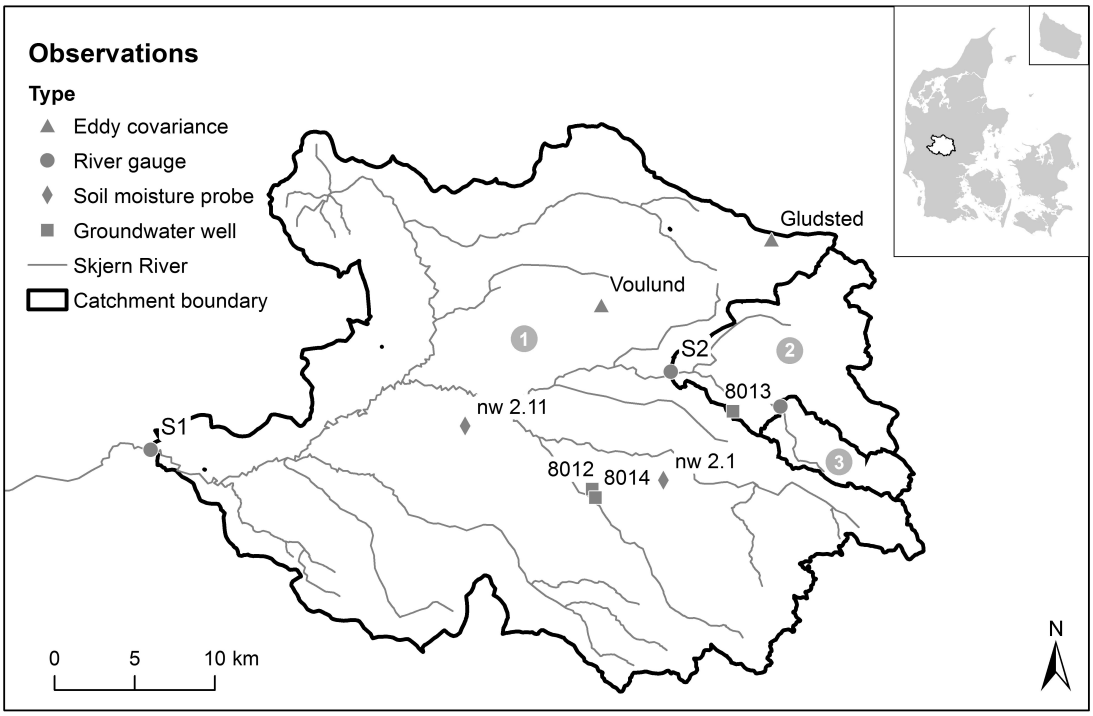


Figure 1: Ahlergaarde catchment and locations of the observational sites used in the present study.

Model 1 (simple)

In Model 1, the unsaturated zone is assumed 1D and described by MIKE SHE's gravity flow solver, where only wilting point and field capacity as given by the soil water retention curves are utilized. The effects of capillary forces on the movement of water through the soil column are ignored. The saturated zone is described by 3D flows using a vertical discretization of the permeable part of the subsurface (around 500 m) into seven layers. Actual evapotranspiration is calculated adopting the approach presented by Kristensen and Jensen (1975), where reference evapotranspiration and leaf area index (LAI) are used as forcing data and soil moisture as a depending variable. In order to avoid a long warm-up period and to ensure stable initial conditions, a hot start file was generated by performing a simulation covering the years 1990-2006. The hydrological model was then subject to a sensitivity analysis. Seventeen parameters were included in the sensitivity analysis; the initial parameter values were taken from a structurally similar and previously optimized hydrological model. Sensitivity analysis was performed using PEST (Doherty, 2016). To reduce dimensionality, we selected only the five most sensitive parameters (cf. Tab. 2) for subsequent calibration.
Calibration was performed for the years 2007 - 2010 using PEST's "SCEUA_P" utility, which is an implementation of the SCE (Shuffled Complex Evolution) global optimization algorithm (Duan, Sorooshian, & Gupta, 1992). Calibration targets were discharge from a set of gauges and hydraulic head in the layers of the geological model. RMSE (root mean square error) and PBIAS (percent bias, cf. section 3.3) were used as objective functions.

Model 2 (complex)

In Model 2, unsaturated zone dynamics were calculated using the full 1D Richards' equation as implemented in MIKE SHE. In contrast to Model 1, actual evapotranspiration was simulated using the SW-ET model, an energy balance module requiring hourly climate input data (Overgaard,

Table 1: Differences in model conceptualization, calibration and simulation between Model 1 and Model 2.

	Model 1 (simple)	Model 2 (complex)
Conceptualization		
Unsaturated zone	Gravity flow Reduced vertical discretization of soil profiles (9 cells between 0-2 m depth	Richards' equation Original vertical discretization of soil profiles (17 cells between 0-2 m depth)
Saturated zone	7 layers	9 layers
Evapotranspiration	Based on reference evapotranspiration after Kristensen and Jensen (1975)	Simulated by the energy-balance model SW-ET (Overgaard, 2005) using hourly climate data
Calibration		
Period	2007 - 2010	2007 - 2010
Algorithm	Shuffled Complex Evolution (SCE), implemented in PEST (Doherty, 2016)	CMA-ES (Covariance Matrix Adaptation Estimation Strategies), implemented in PEST (Doherty, 2016), calibrated by Stisen et al. (2017)
Dimensionality	5 parameters (cf. Tab. 2)	11 parameters
Targets	Discharge, hydraulic head	Discharge, hydraulic head, spatially distributed soil moisture, remotely sensed ET patterns and land surface temperature
Simulation		
Period	2011 - 2016	2011 - 2016
Run time	ca. 5 min	ca. 150 min

Table 2: MIKE SHE model parameters that were optimized using the PEST SCEUA_P utility, Model 1.

Name	**Description**	**Module**
Kx1_ss	Horizontal conductivity of sand units	Saturated zone
Drain	Drainage time constant	Drainage
Khsat_a_JB1	Saturated hydraulic conductivity for soil type JB1, top layer	Unsaturated zone/soil surface parameters
vG_n_a_JB1	van Genuchten parameter n for soil type JB1, top layer	Unsaturated zone/soil surface parameters
Kx2_ler	Horizontal conductivity of clay units	Saturated zone

2005). Due to this modification, Model 2 is classified as a coupled subsurface-surface-atmosphere model.

Especially the coupling with the SW-ET model are the reason for the significantly longer run times of Model 2 (Tab. 1). Representing a rare example of a rigorous inverse calibration (not least due to the immense computational demands), the model had been calibrated in a previous study (Stisen et al., 2017) using a multi-objective calibration approach. Calibration was carried out using a global optimizer (CMA-ES, Covariance Matrix Adaptation Evolution Strategy), five different calibration targets (Tab. 1) and eleven objective functions. This ensured both a balanced calibration and the finding of a Pareto optimal solution as defined in e.g. Madsen (2003).

3.2 Uncertainty estimation: k-NN resampling

k-NN resampling, as the name suggests, resamples historical errors made by the model. This technique is based on the assumption that the response of a hydrological system is documented by a sufficiently long record of observations and thus, if the system is stationary, that the statistical properties of errors can be inferred from a comparison between historical data and model predictions (Anna E. Sikorska et al., 2015). The foundations of the approach used in this study were recently presented by Wani et al. (2017). We adopt this approach while suggesting improvements to the workflow, which involves the consideration of error autocorrelation upon choosing the variables that control the identification of nearest neighbors by the k-NN algorithm. The k-NN resampling procedure is illustrated by the flow chart given in Fig. 2 and described systematically in the following. For a detailed mathematical description the reader is referred to Wani et al. (2017).

Step 1. Split data record: The data record is split into two periods. The first period represents a repository for the sampling of past errors. In other words, the k-NN algorithm "learns" about the typical response of the hydrological system under different hydrological conditions. The second period is a testing period for which the k-NN algorithm predicts residual uncertainty based on the error samples collected from the repository. The error e is defined as the difference between simulated and observed hydrological system response at any time step.

Step 2. Inspect ACF & PACF: In order to establish which variables are suitable to condition the estimation of residual uncertainty on, we suggest a hierarchical approach. We propose to first focus on the error time series and check whether there is evidence for an autoregressive process (AR) of order p. The order p would then be equivalent to the number of error lags to be included in the input variable vector (ivv). This can be achieved by plotting and (visually) examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) (Box, Jenkins, & Reinsel, 2008). For an AR of order p, the ACF will exhibit an exponential decay, while the PACF will have significant nonzero correlation coefficients for lags smaller than or equal p and cut off thereafter, being essentially zero (Box et al., 2008). It is worth noting that we are not actually suggesting to fit an AR model to the data but merely use ACF and PACF as guidance regarding the selection of ivv candidates. Subsequently, dependence between the error time series and other hydrological variables is investigated. This can be useful if the analysis of ACF and PACF is inconclusive or in order to find supplementary variables for the ivv. To also be able to capture nonlinear association between error and other variables (which manifests in heteroscedastic residuals), we additionally employ the distance correlation metric (Szekely, Rizzo, & Bakirov, 2007). Another possible solution to choose ivv candidates for k-NN is to employ the concept of partial information (Sharma & Mehrotra, 2014) but will not be further addressed here. To ensure that potential differences in units or amplitude between the error and hydrological variables do not affect the analysis, standard scores need to be calculated a priori. Viable ivv candidates are then selected based on the strength of their interdependence with the error time series, only including those with a high statistical dependence (here: user defined threshold of Pearson's $r > 0.5$) and keeping the number of ivv elements to a minimum.

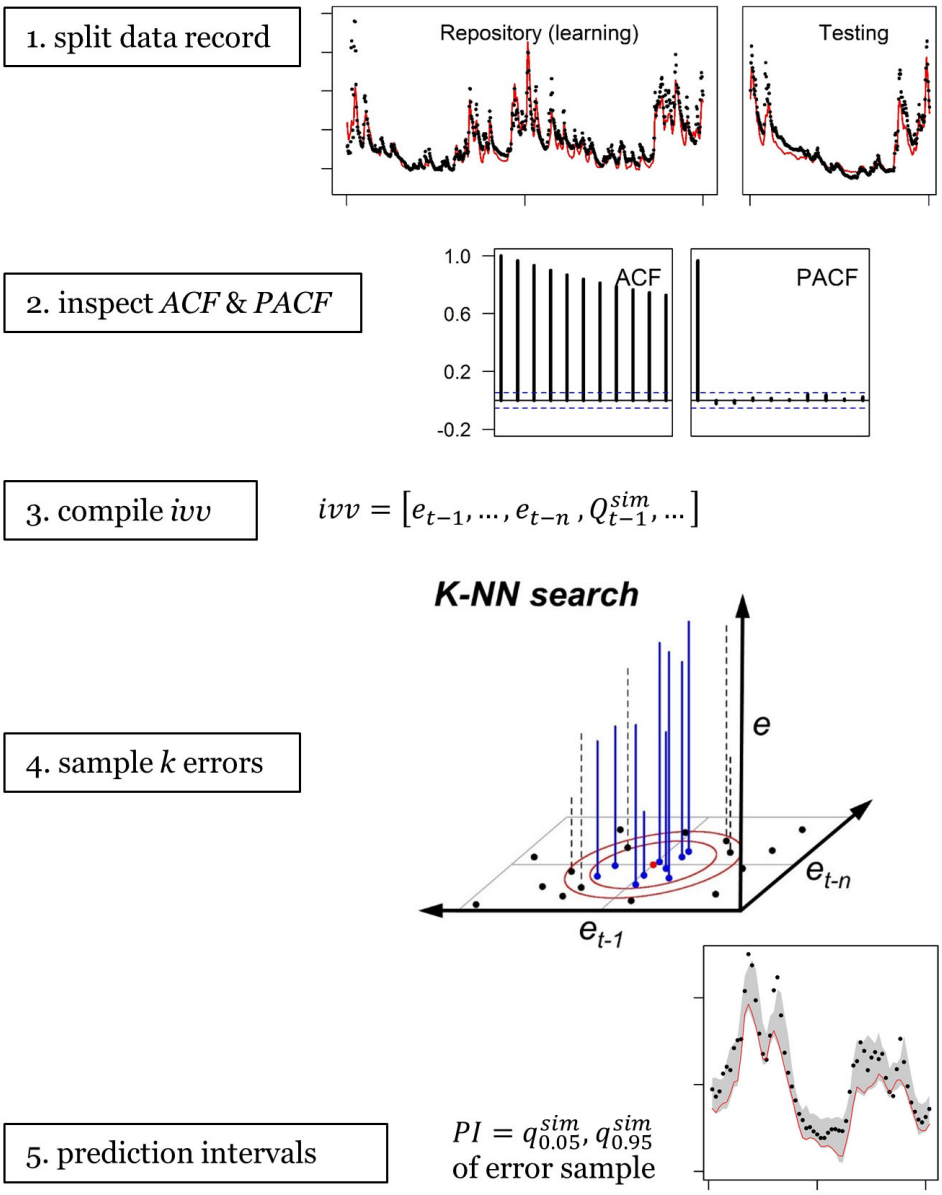


Figure 2: Flow chart illustrating the five steps in k-NN resampling.

Step 3. Compile ivv: assemble the input variable vector using error (number of lags included equals p) 266 and possibly auxiliary hydrological variables such as simulated discharge (Q^{sim}) etc. with high 267 explanatory power as identified in the step before.

Step 4. Sample k errors: The k-NN algorithm is used to sample k historical errors from the repository to estimate residual uncertainty for the testing period. Determining an appropriate number of k is not straightforward and will be addressed in the discussion. The error sampling works as follows: for each day within the testing period, the input variable vector points to a specific

location in n-dimensional ivv-space. The algorithm then searches the repository and identifies k days with the most similar conditions, therefore being closest to that location in n-dimensional ivv-space (nearest neighbors). Usually, the Euclidean distance is used to determine proximity. To avoid having to standardize variables, usage of the Mahalanobis distance represents an alternative, cf. Yates, Gangopadhyay, Rajagopalan, and Strzepek (2003).

Step 5. Prediction intervals: The k historical errors associated with the k nearest neighbors define an error distribution, which - under the assumption of stationarity and hence considering them as being representative for the testing period - provides an estimate of residual uncertainty for each day of the testing period. Mathematically, this can be expressed as

$$C_t(e|v = v_t) \approx C_p(e|r_p \leq r_k) = j/k \tag{1}$$

where C_t is the cumulative distribution function (CDF) of residual error at time step t conditioned on $ivv = ivv_t$ (similarity of input variable vector ivv). C_t is being approximated empirically by C_p, the CDF composed of k historical errors that satisfy $r_p \leq r_k$, where r is the Euclidean distance and p stands for "past". Then, by calculating the 5^{th} and 95^{th} quantiles $(q_{i,0.05}, q_{i,0.95})$ of the error sample, 90% prediction intervals for each day of the testing period can be generated, their width being defined by the difference between the percentiles.

3.3 Metrics

Metrics for assessing hydrological model performance

- Nash Sutcliffe efficiency (NSE), Nash and Sutcliffe (1970)

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_{sim}^i - Q_{obs}^i)^2}{\sum_{i=1}^{n}(Q_{obs}^i - \bar{Q}_{obs})^2} \tag{2}$$

  where Q_{sim}^i is the simulated discharge, Q_{obs}^i is the observed discharge at time step i and $\bar{Q}_{obs}$ is the mean observed discharge.

- Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{sim}^i - y_{obs}^i)}{n}} \tag{3}$$

  where n is the number of data pairs, y_{obs}^i are the observations, y_{sim}^i the simulations at each time step i.

- Mean error

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_{obs}^i - y_{sim}^i) \tag{4}$$

  where n is the number of data pairs, y_{obs}^i are the observations, y_{sim}^i the simulations at each time step i.

- Pearson's r

$$\rho_{y_{obs}^i y_{sim}^i} = \frac{cov(y_{obs}^i, y_{sim}^i)}{\sigma_{y_{obs}^i}\sigma_{y_{sim}^i}} \tag{5}$$

  where cov is the covariance of y_{obs}^i (observations) and y_{sim}^i (simulations), while σ is the standard deviation.

- Percent bias (PBIAS)

$$PBIAS = \left[\frac{\sum_{i=1}^{n} (y_{obs}^{i} - y_{sim}^{i}) * 100}{\sum_{i=1}^{n} y_{obs}^{i}} \right] \tag{6}$$

where n is the number of data pairs, y_{obs}^{i} are the observations, y_{sim}^{i} the simulations at each time step i.

Metrics for verification of uncertainty estimate

In order to assess the performance of k-NN resampling for uncertainty prediction, and analogous to Wani et al. (2017), three validation metrics were calculated: the prediction interval coverage probability ($PICP_{PI}$), the mean prediction interval (MPI_{PI}) see for example Shrestha and Solomatine (2008), and the Alpha Index (α), cf. Renard, Kavetski, Kuczera, Thyer, and Franks (2010).

- Prediction interval coverage probability ($PICP_{90}$): The $PICP_{90}$ is the percentage of instances (C_{90}) for which an observation is contained within the limits of the 90% prediction interval:

$$PICP_{90} = \frac{1}{n} \sum_{i=1}^{n} C_{90} \cdot 100\% \tag{7}$$

$$C_{90} = \begin{cases} 1 & if\ q_{i,0.05} \leq q_i \leq q_{i,0.95} \\ 0 & else \end{cases} \tag{8}$$

- Mean prediction interval (MPI_{90}): The mean prediction interval allows expressing uncertainty as the average width of the 90 % prediction interval based on the k-NN error distribution. The 90% prediction interval is defined by the difference between the 5^{th} and 95^{th} quantile ($q_{i,0.05}$ and $q_{i,0.95}$ respectively) of the k-NN error distribution at every time step i:

$$MPI_{90} = \frac{1}{n} \sum_{i=1}^{n} q_{i,0.95} - q_{i,0.05} \tag{9}$$

- We further test the reliability of quantiles of k-NN-based error prediction by plotting them against the quantiles of observed error in a reliability diagram. Additionally, the Alpha Index (α) provides a numerical measure of reliability, as it relates to the area α' that is defined by calibration function and identity line.

$$\alpha' = \frac{1}{100} \sum_{j=1}^{100} |q_{obs,j} - j/100| \tag{10}$$

$$\alpha = 1 - 2\alpha' \tag{11}$$

where α' is the area between calibration function and identity line, $q_{obs,j}$ are the quantiles of observed errors and $j/100$ gives the quantiles of predicted errors.

3.4 Experimental set-up

k-NN resampling was applied to model outputs from deterministic simulations by Model 1 and Model 2, both covering the historical period January 2011 to August 2016. Repository for the k-NN resampling was the entire period (provided data were available) apart from the year 2014 which was selected as the testing period. This enables direct comparability of results with a related study

Table 3: Input variable vector composition by hydrological variable and hydrological model

Hydrological variable	Observations from-to (n in days)			Model 1	Model 2
Discharge (gauge 250082)	01-01-2011	31-08-2016	(2070)	e_{t-1}, e_{t-2}, e_{t-3}	e_{t-1}, e_{t-2}, e_{t-3}
Soil moisture (nw_2.1)	01-01-2011	31-08-2016	(2070)	e_{t-1}, e_{t-2}	e_{t-1}, e_{t-2}
Hydraulic head (well 8014)	08-11-2012	31-08-2016	(1393)	e_{t-1}	e_{t-1}
Actual ET (Gludsted)	01-01-2011	07-09-2015	(1711)	e_{t-1}, e_{t-2}	e_{t-1}

[reference] involving the propagation of input and paramater uncertainty through a hydrological model equal to Model 2. Despite excluding 2014 from the k-NN resampling, the repository was treated as if it was continuous. K-NN resampling was carried out using the RANN package (Arya, Mount, Kemp, & Jefferis, 2015) implemented in the scientific progamming environment R (R Development Core Team, 2018).

Eight individual k-NN input variable vectors were established, one for each of the four hydrological variables discharge, hydraulic head, soil moisture and evapotranspiration and each of them for the two hydrological models. The input variable vectors for the k-NN models were composed based on analysis of ACF and PACF. Additionally, correlations and distance correlations were calculated in order to test whether there were (non) linear dependencies between the error time series and a set of additional hydrological variables which potentially could impact errors. The hydrological variables that were analyzed in this context were observed and simulated discharge (soil moisture, hydraulic head, actual evapotanspiration, respectively), observed rainfall (both at catchment and at single grid cell scale), observed soil moisture, hydraulic head and actual evapotraspiration. In all cases, (distance) correlations (not shown here) with other hydrological variables were found to be relatively weaker as compared to error autocorrelation. Potential commensurability issues (Keith Beven, 2012) will be addressed in the discussion. Therefore, we decided to restrict the conditioning to observed error, including as many lags as the analysis of ACF and PACF suggested. An overview of the input variable vector composition derived based on the considerations detailed in section 3.2 is given in Table 3.
For the period of interest regarding the application of k-NN resampling (2011 - 2016), observational data were retrieved from the HOBE database (locations of observational sites are shown in Fig. 1). Discharge observations (daily mean in m^3/s) were collected by a gauge located at the outlet of Ahlergaarde catchment, while the soil moisture observations were made by a Decagon ECH2O data logger (ID nw_2.1, depth of 22.5 cm) (Bircher, Skou, Jensen, Walker, & Rasmussen, 2012). Hydraulic head observations stem from a groundwater well (ID 8014) and actual evapotranspiration from an eddy covariance flux tower (Ringgaard et al., 2011) at the forest site in Gludsted (Fig. 1). While the HOBE database contains a vast amount of data for a variety of hydrological variables measured at numerous sites within the catchment, the measurement sites used in this study were selected based on the criterion of longest possible gap-free observational record; however, there are differences in the lengths of these records (cf. Table 3). The number of k (i.e. the number of nearest neighbors to be sampled from the repository) was chosen to be 199.

4 Results

4.1 Hydrographs and residual uncertainty - Model 1 (simple)

Hydrographs (simulated by Model 1) for discharge as well as equivalents for soil moisture, hydraulic head and actual evapotranspiration in the testing year 2014 are presented in Fig. 3. While the k-NN prediction intervals look different depending on the hydrological variable, PIPC90 is generally close to the expected value of 90 % (88.2 / 87.4 / 93.4 / 90.7 % for Q / SM / h / ET). With regard to discharge, the resulting prediction interval, while being overall relatively constant (MPI: 4.03 m^3/s, Table 5), is slightly wider during low-flow conditions and slightly narrower during high-flow conditions. This corresponds well with the fact that the largest mismatch between observations and simulation is indeed found during low-flow conditions throughout the summer period, while

the model reproduces observed peak flow relatively well. Around day 200, the prediction interval departs from the simulation, successfully encapsulating most of the observations. This is exemplary for a situation where k-NN resampling leads to an error distribution that effectively corrects for systematic over- or underprediction by the model assocated with similar hydrological conditions in the past. Compared to discharge, the prediction interval for soil moisture is highly unsteady, exhibiting an erratic structure and a partly extremly wide prediction interval (MPI: 0.104, Table 5). Although PICP90 (87.9 %) is near the expected value, k-NN was clearly unable to obtain informative error samples from the repository. This may be linked to the choice of using "gravity flow" as unsaturated zone solver, resulting in a failure of the model to reproduce soil moisture content and dynamics for most parts of the testing year (crude overestimation up to day 100 and past day 225, underestimation between day 125 and 225; especially depletion of soil moisture during spring / summer is not reproduced). As for hydraulic head, the prediction interval is very narrow (MPI: 0.32 m, Table 5) with a number of local expansions. The $PICP_{90}$ (93.4 %) slightly exceeds the expected value. The substantial bias between observations and simulation is successfully corrected for by k-NN resampling. A visual assessement of the prediction interval for actual evapotranspiration is impeded by the large fluctuations of simulation and observations from day to day, especially during the summer period. Generally, the prediction interval is relatively wide (2.2 mm/d), while the $PICP_{90}$ is very good (90.7%).

4.2 Hydrographs and residual uncertainty - Model 2 (complex)

Hydrographs (simulated by Model 2) for discharge as well as equivalents for soil moisture, hydraulic head and actual evapotranspiration in the testing year 2014 are presented in Fig. 4. The PIPC90 is close to but consistently exceeds the expected value of 90% (94.0 / 97.3 / 96.2 / 92.6% for Q / SM / h / ET). With regard to discharge, the resulting prediction interval is narrower (MPI: 3.28 m3/s, Table 5), varies notably more and generally seems to be better fitted than its counterpart obtained for Model 1. Contrary to what was observed before, the prediction interval is narrowest during low-flow conditions, wider for high-flow conditions and has its maximum width in the transitionary period between winter and spring (ca. day 75 to 100), coinciding with the largest difference between observations and simulation. Apart from a short period in summer (around day 200), the prediction interval effectively only extends in an upward direction. Heteroscedasticity is captured, as illustrated by the increasing width in MPI for different discharge flow categories (cf. Tab. 5). When looking at soil moisture, the prediction interval is distinctly less erratic and, on average, narrower (MPI: 0.048, Table 5) than the one obtained using Model 1. However, the prediction interval shows little differences in width thoughout the testing year. In comparison to Model 1, the usage of the Richards' equation as unsaturated zone solver clearly led to an improved, although biased, simulation of the soil moisture dynamics. While k-NN resampling results in error prediction intervals that contain more than the expected 90% of all observations ($PICP_{90}$ of 97.2%), they do correct for the systematic deviation of simulation and observations (overestimation by the model). With regard to hydraulic head, the k-NN- based prediction interval is yet again very narrow (MPI: 0.34 m, Table 5). However, it appears to be systematically wider for the first, systematically narrower for the second half of the testing year. The $PICP_{90}$ (96.2%) moderately exceeds the expected value. As seen for Model 1, the bias between observations and simulation is successfully corrected for by k-NN resampling. When comparing the graphs for Model 1 and 2, it appears that the hydraulic head simulation for Model 2 shows somewhat larger fluctuations; this is very likely connected to the different unsaturated zone solvers. As was the case for Model 1, the prediction interval for actual evapotranspiration is difficult to assess as the large day-to-day fluctuations of simulation and observations prevents this. It is worth noting that simulated actual evapotranspiration for Model 2 is derived using the SW-ET model. The prediction interval is relatively wide (3.05 mm/d), while the $PICP_{90}$ is rather good (92.6%).

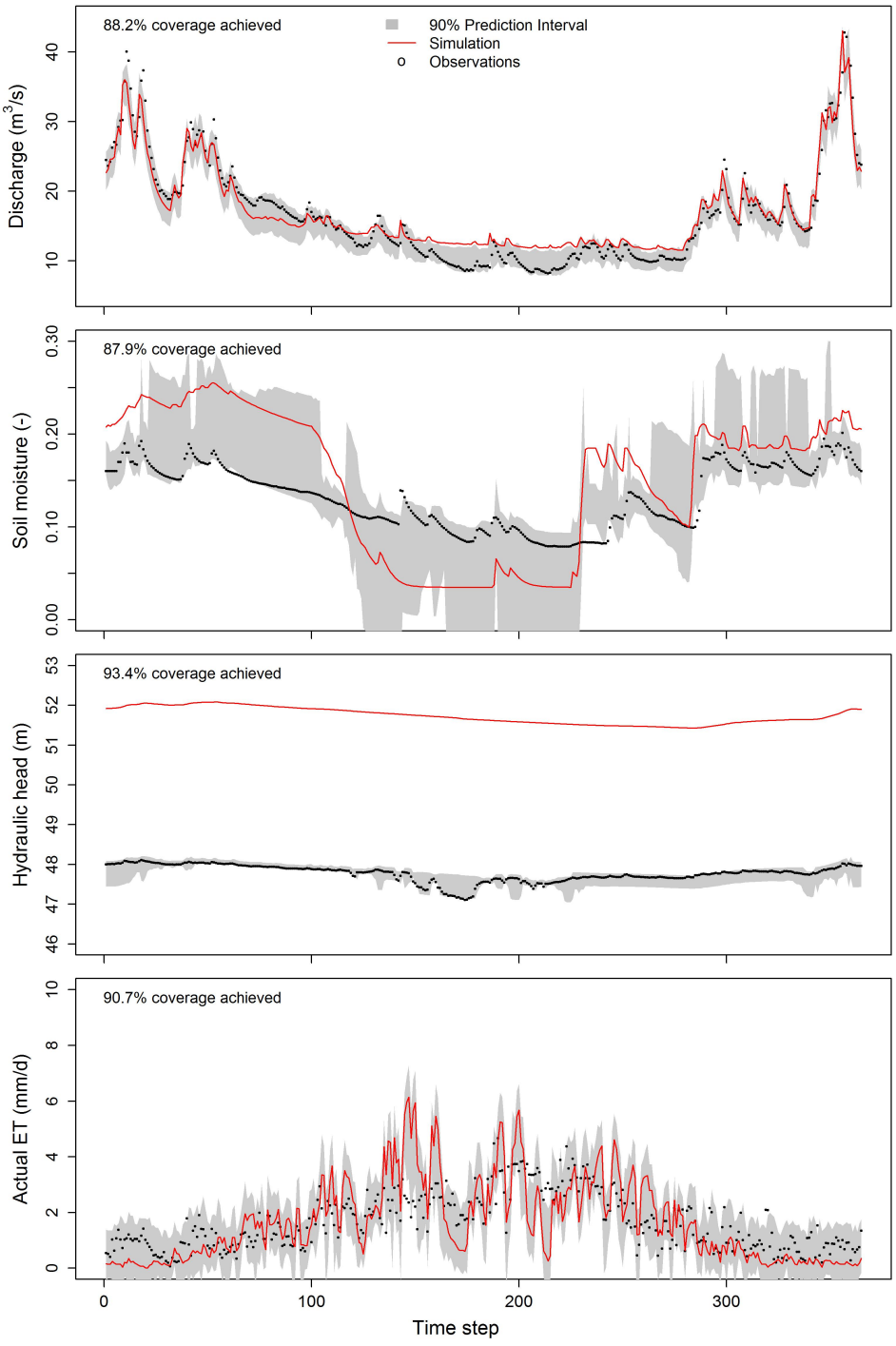


Figure 3: Hydrograph for discharge and equivalents for soil moisture, hydraulic head and actual evapotranspiration in the testing year 2014, Model 1. The residual uncertainty is characterised by 90% prediction intervals using k-NN resampling (k = 199).

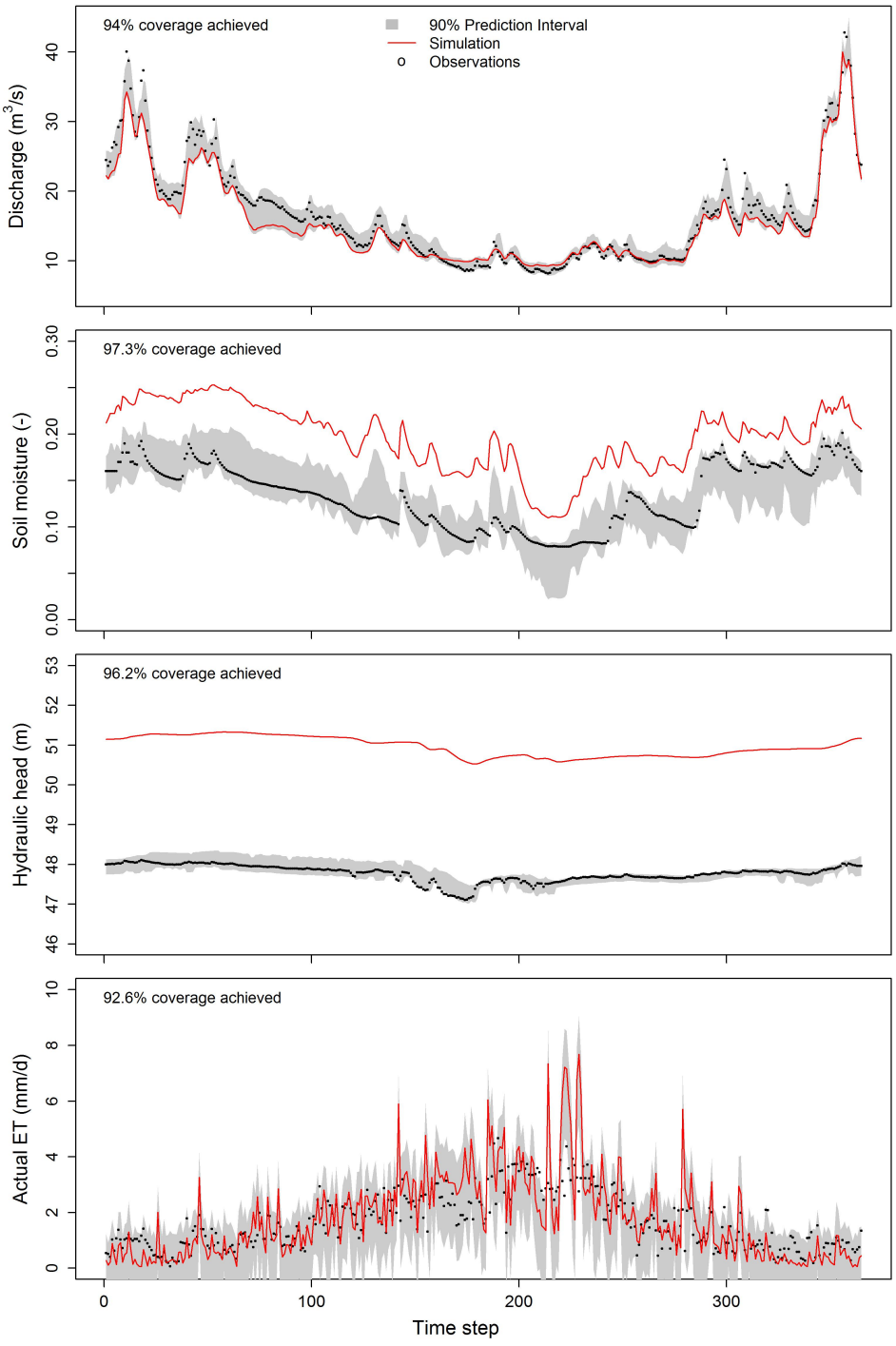


Figure 4: Hydrograph for discharge and equivalents for soil moisture, hydraulic head and actual evapotranspiration for Ahlergaarde catchment in testing year 2014, Model 2. Residual uncertainty as given by 90% prediction intervals using k-NN resampling (k = 199). Input variable vector composition is presented in Table 3.

4.3 Reliability diagrams

The performance of k-NN resampling regarding the reliability of the error quantiles for all flows is evaluated by generating realiability diagrams (Fig. 5) and calculating the Alpha Index (Tab. 5).

Table 4: Performance metrics for hydrological simulation in the testing year 2014 linked to Figure 2 (Model 1) and Figure 3 (Model 2).

Hydrological variable	**NSE**	**RMSE**		**ME**		**Pearson's r**	**PBIAS (%)**
			Model 1				
Discharge	0.91	-	m^3/s	0.57	m^3/s	0.96	3.41
Soil moisture	-	0.06	-	0.02	-	0.81	14.17
Hydraulic head	-	3.97	m	3.96	m	0.65	8.30
Actual evapotranspiration	-	0.99	mm/d	-0.001	mm/d	0.72	-0.08
			Model 2				
Discharge	0.94	-	m^3/s	-1.01	m^3/s	0.98	-6.07
Soil moisture	-	0.07	-	0.06	-	0.83	47.52
Hydraulic head	-	3.19	m	3.19	m	0.76	6.67
Actual evapotranspiration	-	0.95	mm/d	0.10	mm/d	0.77	-6.07

With regard to discharge, the reliability diagram for Model 1 suggests high reliability of the error quantiles, as nearly all points fall along the identity line. The Alpha Index of 0.94 supports this conclusion. Above a probability of non-exceedance of ca. 0.5, a slight trend towards overprediction (probabilities too high; deviation of calibration function to the right of the 1:1 line) is detectable. The corresponding reliability diagram for Model 2 shows a slight downward deviation of the calibration function from the 1:1 line for probabilities of non-exceedance of ca. 0.05 to 0.3, yet again indicating overprediction. However, the graph returns back to the identity line thereafter. Overall, the reliability of the error quantiles can be considered very high (Alpha Index of 0.93).

As for soil moisture, the points of the calibration function for the error quantiles of Model 1 are close to the identity line up to a non-exceedance probability of 0.5. For larger probabilities, the graph begins to deviate from the identity line in an upward direction (underprediction), with the largest deviation found at a probability of non-exceedance around 0.7. The graph eventually returns to the desired 45 degree line around a probability of non-exceedance of 1. The Alpha Index is 0.89. In comparison, the reliability diagram for error quantiles of Model 2 despicts a calibration function that shows comparatively smaller but constant deviations from the identity line between probabilities of non- exceedance of > 0.2 and up to nearly 1, before returning to the 1:1 line; the Alpha Index is equal to 0.93 and thus notably better than that of Model 1.

The reliability diagrams for error quantiles of hydraulic head represent the ones with the strongest deviations from the identity line, and therefore lowest reliability. For both Model 1 and Model 2, there is an overprediction up to a probability of non-exceedance of about 0.4, above which there is a transition towards underprediction (with peaks at a probability of non-exceedance of 0.7 and 0.6, respectively). The graph eventually returns to the identity line around 1. According to Wilks (2006), the reliability diagrams have good resolution but are underconfident. The Alpha Index is equal to 0.86 and 0.84 for Model 1 and Model 2, respectively, suggesting reasonable reliability.

Finally, the reliability diagrams for error quantiles of actual evapotranspiration suggest a very reliable error prediction for both Model 1 and Model 2. As for Model 1, a slight underprediction (at probability of non-exceedance of 0.1 to 0.2, respectively) is followed by an equally minor overprediction (probability of non-exceedance of 0.5 to 0.8), while the scatter is still located very closely to the identity line. Regarding Model 2, there is a slight overprediction around a probability of non- exceedance of 0.3 while the remaining calibration function largely falls along the identity line. Alpha Indices of 0.93 and 0.95 indicate high reliability.

In summary, the reliability diagrams in Figure 4 show that the k-NN error quantiles are indeed reliable for all hydrological variables and both models. When looking at the Alpha Indices, reliability of the k-NN error quantiles appears to be slightly higher for Model 2 in the case of soil moisture and actual evapotranspiration, while being higher for Model 1 in the case of discharge

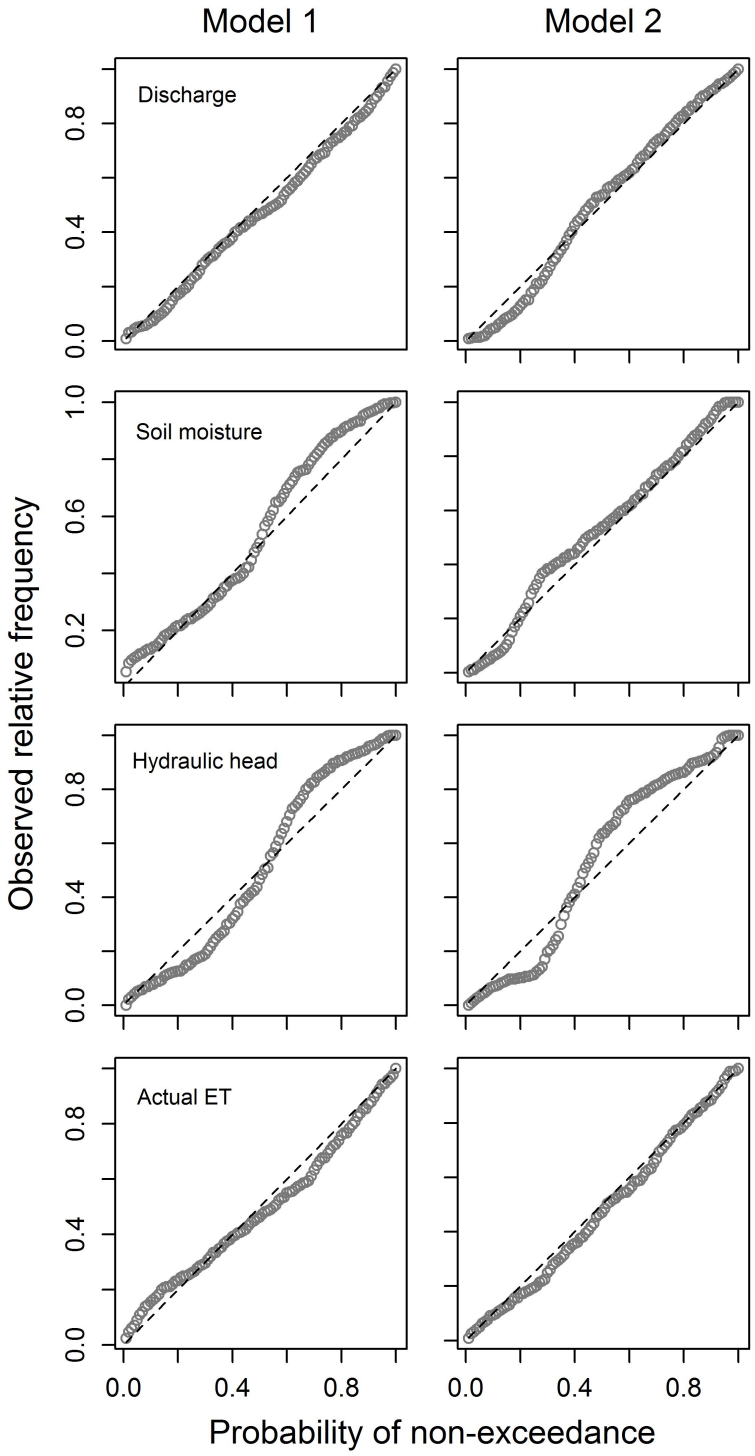


Figure 5: Hydrograph for discharge and equivalents for soil moisture, hydraulic head and actual evapotranspiration for Ahlergaarde catchment in testing year 2014, Model 2. Residual uncertainty as given by 90% prediction intervals using k-NN resampling (k = 199). Input variable vector composition is presented in Table 3.

Table 5: Alpha Index (α) and mean prediction interval width (MPI) for discharge (all flows), soil moisture (entire variable range), hydraulic head (all states) and actual ET (all fluxes) of Ahlergaarde catchment.

					Model 1			
Variable	α	MPI for different flow categories						**Unit**
		All flows	<10%	10-25%	25-75%	75-90%	>90%	
Discharge	0.94	4.03	3.22	3.28	3.87	4.83	5.48	m^3/s
Soil moisture	0.89	0.103	0.143	0.138	0.091	0.084	0.092	-
Hydraulic head	0.86	0.32	0.33	0.35	0.35	0.26	0.22	m
Actual ET	0.93	2.34	2.16	2.00	2.23	2.57	3.19	mm/d
					Model 2			
Discharge	0.93	3.28	1.80	2.25	3.28	4.12	4.49	m^3/s
Soil moisture	0.93	0.048	0.071	0.051	0.050	0.040	0.039	-
Hydraulic head	0.84	0.34	0.27	0.24	0.33	0.43	0.43	m
Actual ET	0.95	3.05	3.31	3.31	3.07	3.32	3.88	mm/d

and hydraulic head. However, the numerical differences are minimal.

4.4 Reproduction of error ACF

Figure 5 compares autocorrelation functions for observed and k-NN predicted errors in the testing year 2014. Given that the respective ivv (cf. Tab. 3) informed the k-NN algorithm well and the historical record held sufficient examples of relevant hydrological conditions, the ACF of observed errors should be reproduced by the ACF of k-NN sampled errors. A look at the ACF for discharge (Fig. 6, first row) shows that the autocorrelation function is only roughly repoduced for both models. In both cases, the k-NN based ACF exhibits correlation coefficients with a higher magnitude than found in the observed ones. However, the overall decline in autocorrelation with increasing lag is reflected to some extent. For Model 2, the increase in correlation at lag 6 to 8 can be found in the k-NN based ACF. When looking at the ACFs of soil moisture, it seems that the k-NN based ACF agrees well with the ACF of observed errors. For Model 1, the continuous decline in correlation is well reproduced; for Model 2, both the decline as well as the increase of correlation around lag 25 are mirrored in the k-NN based ACF. As for hydraulich head, k-NN resampling appears to have performed better with regard to Model 2: the k-NN based ACF matches the one of observed errors quite well in terms of slope of the decline in correlation and magnitude of correlation coefficients; for Model 1, this is clearly not the case. Finally, the k-NN based ACFs for actual evapotranspiration are well reproduced for both models even though they are very different in appearance; while the ACFs associated with Model 1 exhibits correlation coefficients, which are significant for all lags, Model 2 shows considerably lower correlation, including insignificant peaks from at e.g. lag 2 to 5. Hence, it can be concluded that, on average, k-NN resampling was able to reproduce observed error autocorrelation well. This is true for both models, albeit that the k-NN based ACFs obtained for Model 2 appear to agree a bit better with the observed error autocorrelation than those of Model 1.

5 Discussion and conclusion

In this study, we employed a nonparametric method called k-NN resampling to perform uncertainty assessment for two deterministic simulations carried out using two configurations of a coupled integrated hydrological model. The two model configurations were characterized by differences in structural complexity, process description and calibration data. While previous applications in hydrological literature are limited to prediction of streamflow (Lall & Sharma, 1996; Souza & Lall, 2003) or e.g. weather variables (Rajagopalan & Lall, 1999), our study presents an application with a multi-site, multi-variable prediction. We simultaneously quantify residual uncertainty for

multiple model outputs (soil moisture, hydraulic head and actual evapotranspiration in addition to discharge). Despite a number of differences between the two employed models, the results show that k-NN resampling is capable of producing reliable prediction intervals for this class of hydrological models. Consideration of error autocorrelation led to smooth PIs for discharge (highest order p, longest memory), which is in line with observations made by Evin et al. (2014). For all four examined hydrological variables, k-NN based prediction intervals achieved coverage probabilities that were close to but mostly exceeded the expected value of 90%. Similar results were obtained by Anna E. Sikorska et al. (2015) (using k-NN; PICP coverage of 97.9% compared to expectation of 95%) and Montanari and Koutsoyiannis (2012) (using a Meta-Gaussian approach; PICP coverage of 90.3% compared to expectation of 95%). As demonstrated by Fig. 6, k-NN rather successfully mimicked the autocorrelation of errors observed in the testing period (year 2014). Systematic biases in the hydrological model simulation such as observed for hydraulic head (cf. Fig. 3 and Fig. 4) are accurately estimated resulting in narrow prediction intervals that encapsulate most of the observations and error quantiles that achieve high reliability. This is in line with findings by Wani et al. (2017) who showed that the reliability was not impeded when imposing an artificial bias on the simulation. Learning from past errors, k-NN accounts for systematic deviations, resulting in prediction intervals that may depart significantly from the deterministic model prediction. This would not be possible if e.g. using a parametric method, which assume zero mean normally distributed residuals. We proposed to exploit information contained in the error time series by analyzing ACF and PACF to determine the order of the autoregressive process and formalize input variable vector composition for cases where conclusive. Demonstrably, the performance based on error alone was very good.

As noted before, the identification of suitable ivv candidates might be problematic due to commensurability issues (Keith Beven, 2012). Good examples to illustrate this are soil moisture and hydraulic heads: While the observations are point measurements representative of very limited volumes, the simulated values are grid cell averages; therefore, the two quantities are technically not comparable. When it comes to error itself, K-NN successfully compensates for this mismatch prediction because it learns from past errors. However, this might still be problematic in case of e.g. rainfall. Owing to a lack of rain gauge observations, rainfall (based on the 10 km input grid) was extracted from the grid cell that contained the respective observation. This is certain to fail representing the pronounced small-scale spatial variability of rainfall (Wood, Jones, & Moore, 2000), which, however, will likely be reflected in the soil moisture measurement. Therefore, it is not surprising that the correlation between the extracted rainfall and error (the difference of a point measurement observation and a grid cell average) was found to be negligible.
Part of this study was dedicated to the comparison of k-NN performance for two different configurations of the same hydrological model. Both PICP and Alpha Index allowed no clear conclusion regarding the question for which model k-NN performs better. However, the MPIs for discharge and soil moisture are considerably smaller for Model 2 than for Model 1 (cf. Tab. 5) and are generally better defined; reliability for the error quantiles is higher for Model 2 (0.93 vs. 0.89). As for soil moisture, this is likely a result of the superior simulation using the Richards' equation (Model 2) instead of gravity flow (Model 1) and possibly, because soil moisture had been included into the calibration of Model 2. While correlation is only slightly better for Model 2 (r = 0.83 vs. 0.81), a visual inspection indicates that Model 2 recreates soil moisture dynamics significantly better than Model 1. While PICP is good for both models, the resulting PI of Model 1 looks very inconsistent, while the one of Model 2 gives a notably better impression. As for hydraulic head, a notably improved correlation can be seen for Model 2 (r = 0.76 vs 0.65). This is probably a consequence of the more 563 complex, physically more realistic soil moisture simulation, therefore leading to simulated hydraulic heads that are in better accordance with the observed heads. However, as compared to other hydrological variables, the Alpha Index suggests an inferior reliability of the predicted error quantiles 566 (0.86 and 0.84, the only ones below 0.9). Possibly, the k-NN resampling is at a disadvantage for this particular variable, due to comparatively short length of the observational data set for hydraulic heads (n=1393) and the fact that both observation and

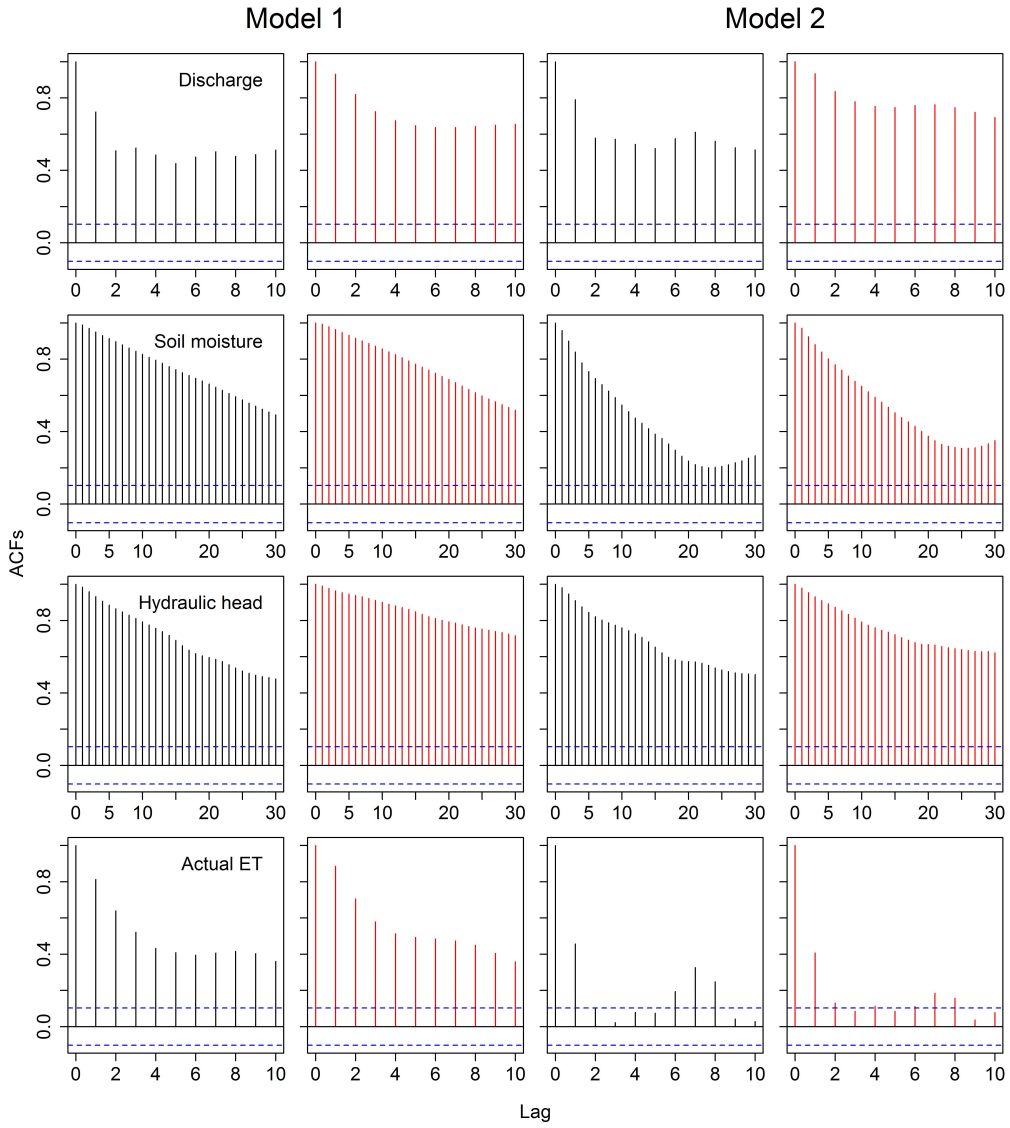


Figure 6: Autocorrelation functions of observed error (black lines) and k-NN resampling based error (red lines) for testing year 2014 for discharge, soil moisture, hydraulic head and actual evapotranspiration.

simulation time series do not fluctuate much (i.e. errors tend to be very similar basically everywhere). When looking at actual ET, the reliability of the error quantiles is better for Model 2 (Alpha Index = 0.93 vs. 0.95), yet the MPI is significantly wider. As the energy balance model SWET is involved here, the ET simulation might still be superior in Model 2 (higher correlation, 0.72 vs 0.77). The quality of the prediction intervals, while PICP is good, is difficult to judge in this case.

With regard to k-NN resampling, we kept the value of k in this study at 199 for all four hydrological variables. A rule of thumb to obtain a robust number of k (Sivakumar, 2017) is to take the square root of the number of available observations, $\sqrt{N}$ which would have resulted in a value of k around 45. Given the relatively short length of the observational records (ranging from 1393 to 2070 days, Tab. 3) as compared to other applications of k-NN and the PICP coverage exceeding the expected

value, this number is possibly too high, hence introducing a high number of not-so-similar error samples into the uncertainty prediction. For example, Anna E. Sikorska et al. (2015) use 10 years for error sampling and 10 years for testing, while Wani et al. (2017) use 6 years (but at a resolution of 12 hours) for error sampling and testing for their Upper Severn catchment case study. Anna E. Sikorska et al. (2015) found the value of k (when varied between 5 and 100) to have little impact on the prediction intervals. As k is effectively a smoothing parameter (Sivakumar, 2017), the resulting empirical error distributions might be close to the marginal distribution of the error associated with the respective hydrological variable. This, in turn, might result in prediction intervals that have a rather constant width, regardless of hydrological conditions. Indeed, in order to provide the k-NN algorithm with sufficient information to learn from, the historical record should contain several examples for all hydrological conditions found in the testing period and for which error prediction is desired (Wani et al., 2017). Particular characteristics of each hydrological variable regarding recurrence of certain signals should also be considered. While seasonality might be factoring into this equation as well, there is a distinct difference in the dynamics of e.g. discharge or hydraulic head and actual ET, ergo possibly requiring individually adjusted k values. As Wani et al. (2017) points out, the determination of a more rigorous approach of finding an optimal k value would require further research efforts. The application of k-NN resampling requires the data to be stationary. However, the stationarity assumption for catchment response holds only if environmental conditions remain stable - an illusory assumption (K. Beven, 2016). Anthropogenic alterations that affect e.g. river flow volume (extractions, temporary damming) or the timing of flow peaks (weirs) but also soil moisture content (drainage) and changing climatic conditions would render data-driven techniques such as k-NN effectively worthless. The whole notion of errors observed in the past giving an indication of errors in the period for which uncertainty prediction is carried out would lose its justification. In this context, it was of advantage that the current project used data from a multi-disciplinary research project (HOBE) collected in a relatively small catchment. Therefore, continuity and quality of observational data can be considered relatively constant. In addition, no change in human impacts (land use groundwater abstraction) occurred during the six years study period.

The application of k-NN in this paper demonstrated that the technique works well with a complex coupled integrated hydrological model. The method is very flexible and allows near-simultaneous UA (time-efficient, < 1 min per hydrological variable as opposed to several weeks for MC-based UA on a standard desktop computer). Theoretically, k-NN can also be used in combination with other methods 610 to explicitly account for specific uncertainty sources expicitly, e.g. input and parameter uncertainty (Anna E. Sikorska et al., 2015) or rating curve uncertainty (A. E. Sikorska & Renard, 2017). However, as Verkade, Brown, Davids, Reggiani, and Weerts (2017) showed, source-based UA is not necessarily superior to its lumped counterpart. Furthermore, the choice of a suitable UA method should depend on the purpose of the application (Refsgaard et al., 2007). While knowing the relative contributions of individual sources of uncertainty to overall uncertainty might be insightful when aiming for an enhanced process understanding or the devleopment of strategies to reduce predictive ucnertainties, an aggregated estimate may be sufficient when its purpose is to provide orientation for decision making. It needs to be emphasized that the communication of uncertainty and limitations in modeling results to end users has the potential to increase trust in scientific products (Juston et al., 2013). The well-known limitation of being unable to predict outside the observed range can be overcome using innovations to perturb the observations (Rajagopalan & Lall, 1999) or using a Kernel density estimator (Sharma, Tarboton, & Lall, 1997).

While k-NN resampling is a comparative lightweight amongst UA methods in terms of mathematical sophistication, its strengths are its robustness in the face of model bias and the ease with which in can be implemented and set-up to carry out UA for multiple model outputs. The latter property renders it a potent tool for applied hydrology and situations where uncertainty assessment is to be performed for multiple hydrological model outputs individually, thus providing decision makers with valuable information about uncertainty of secondary model outputs. Shifting focus

away from discharge-only oriented UA will help fully exploiting the enormous potential of complex hydrological models. This paper is seen as a contribution to further advance the establishment of UA in operational hydrology.

Acknowledgement: This study has been carried out as part of the HOBE project (Hydrological Observatory, www.hobecenter.dk) which is funded by the Villum Foundation. All data used in this study is stored in the project database (www.hobedata.dk); to obtain access, contact Frederik Uldall (fu@ign.ku.dk).

Chapter 4

Parameter estimation of hydrologic models using a likelihood function for censored and binary observations

Omar Wani[1,2], Andreas Scheidegger[2], Juan Pablo Carbajal[2], Jörg Rieckermann[2], Frank Blumensaat[1,2]

[1] Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland
[2] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

Journal: Water Research

Publication date: 20th May 2017

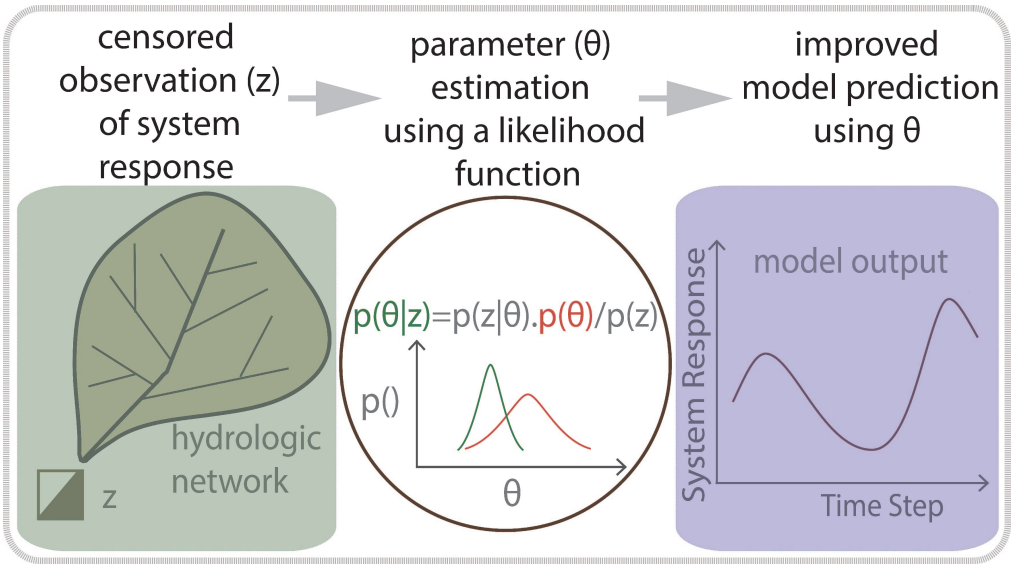
DOI: http://dx.doi.org/10.1016/j.watres.2017.05.038

Author contributions: OW, AS, FB, and JR wrote and structured the main text. OW conceptualized the simulation experiments, wrote the code, and carried out the analysis. OW and AS formulated the likelihood function. JPC built and trained the emulator, and also wrote the supplementary section. FB planned and conducted the observations, and developed the rainfall-runoff model. JR and AS envisioned this project, whereas JR and FB actively supervised the research. All authors reviewed the paper.

Award: This work received the Best Paper Award at the 10th International Conference on Urban Drainage Modelling, 20-23 September, 2015, Quebec City, Canada.

Abstract

Observations of a hydrologic system response are needed to accurately model system behaviour. Nevertheless, often very few monitoring stations are operated because collecting such reference data adequately and accurately is laborious and costly. It has been recently suggested to use observations not only from dedicated flow meters but also from simpler sensors, such as level or event detectors, which are available more frequently but only provide censored information. Binary observations can be considered as extreme censoring. It is still unclear, however, how to use censored observations most effectively to learn about model parameters. To this end, we suggest a formal likelihood function that incorporates censored observations, while accounting for model structure deficits and uncertainty in input data. Using this likelihood function, the parameter inference is performed within the Bayesian framework. We demonstrate the implementation of our methodology on a case study of an urban catchment, where we estimate the parameters of a hydrodynamic rainfall-runoff model from binary observations of combined sewer overflows. Our results show, first, that censored observations make it possible to learn about model parameters, with an average decrease of 45% in parameter standard deviation from prior to posterior. Second, the inference substantially improves model predictions, providing higher Nash-Sutcliffe efficiency. Third, the gain in information largely depends on the experimental design, i.e. sensor placement. Given the advent of Internet of Things, we foresee that the plethora of censored data promised to be available can be used for parameter estimation within a formal Bayesian framework.



1 Introduction

Traditionally, mathematical models of river catchments and urban drainage systems are calibrated on uncensored observations of physical variables, such as discharge or water levels (Gupta et al., 1998; Madsen, 2003; Refsgaard, 1997; Westerberg et al., 2011). By uncensored observations we mean that the monitoring device used has an adequately large measurement range i.e. it can measure any physically possible value of the hydrologic variable. However, the installation and management of such monitoring devices for hydrologic systems is generally laborious and costly (Maheepala et al., 2001). Therefore, typically only a few locations are equipped with such sensors. There are other types of sensors and observation equipment that collect data only within a certain range, and the value of any system response beyond an interval cannot be measured (Borup et al., 2015). Also, recent advances in sensor and data transfer technology development reveal a trend towards more cost-efficient sensors, which simply detect the occurrence of overflow events (Rasmussen et al., 2008) or detect only the exceedance after some critical water levels, thus providing binary information. Within this context, it has even been suggested to specifically develop binary monitoring devices based on robust and low-cost sensors such as temperature probes (Hofer et al., 2014; Montserrat et al., 2016, 2013), motion detectors (Siemers et al., 2011) and electrical switches (Rasmussen et al., 2008). Also, even though learning about the depth and velocity of flow from the footage of CCTV cameras during floods may be challenging, they easily capture the inundation and the duration of an event (Le Coz et al., 2016; Lo et al., 2015). In the context of Internet of Things, we are witnessing a growing trend where more real-time information related to hydrologic systems is available from hitherto unconventional sources (Eggimann et al., 2017; Kerkez et al., 2016).

Availability of any kind of relevant information related to the output of a modelled system should theoretically be able to reduce uncertainty in the parameters, allowing us to update our prior assumptions about their values (Riggelsen, 2006; Rinderknecht et al., 2014). As model predictions for hydrologic systems can be highly sensitive to various parameters (Gamerith et al., 2013; Song et al., 2015), the reduction in parametric uncertainty and identification of better parameter values has a direct bearing on the performance of models. It is therefore desired that in the future all types of data from different sensors, a) uncensored and b) censored, should help to calibrate hydrologic models within one consistent and general mathematical framework.

The underlying idea that many less accurate sensors may provide as much information about a complex hydrologic system as a few very accurate ones is very compelling. The first step to test this hypothesis can be to evaluate the value of binary observations in isolation, which can be seen as the most extreme case of censored data. While there has been research on modifying data assimilation techniques, like ensemble Kalman filtering, in order to handle censored data for updating the system state (Borup et al., 2015), the next logical step is to explore techniques that can infer model parameters from censored observations as well. However, so far only ad-hoc approaches to model calibration and parameter estimation using censored data have been suggested and it is currently not clear what the information content of such observations is and how we can use them most efficiently to learn about the system. Rasmussen et al. (2008) used the binary information measured during the occurrence of combined sewer overflows (CSO) to estimate the hydrologic reduction coefficient based on the mismatch between the predicted duration of the overflow and its observed duration. Similarly, measurements of overflow duration have been successfully used for calibration in other studies (Montserrat et al., 2016). Aronica et al. (2002) used a scalar performance measure and mapped it to the parameter space of the model with an informal likelihood function. The performance measure gives an aggregated indication of the performance of the inundation model. Thorndahl et al. (2008) applied similar mathematical framework for calibration of an urban drainage model and the quantification of parametric uncertainty using data on CSOs. In these two studies, those parameters that perform well on the chosen performance measure were identified as good and those which perform below a defined value were termed non-behavioural and not considered for model simulations. However, in the authors' opinion, this inference procedure is limited by the usage of an informal likelihood function (Stedinger et al., 2008). First, the cri-

terion of acceptance/rejection of various parameter values is ad-hoc (Dotto et al., 2012). Second, having more observations does not reduce parametric uncertainty (Mantovan and Todini, 2006). And third, it assumes a perfect model structure and error-free input data, which is often not the case in rainfall-runoff studies (Del Giudice et al., 2015). To incorporate relevant error-generating processes, we need a formal likelihood function, which generally is lacking in the current treatment of censored observation. In this paper, we therefore suggest a likelihood function which makes it possible to estimate parameters from censored signals, while accounting for uncertainties in input variables, such as rainfall or land use, and model structure deficits. Based on the results from a case study, we demonstrate that censored data is surprisingly effective in reducing parametric uncertainty of rainfall-runoff models. Although uncensored observations are comparably more informative than censored data, our results are still quite promising, because they show the way towards using all forms of available information in model calibration, which is often not done.

In the following sections, we first present the mathematical formulation of our likelihood function. We then demonstrate its application in inferring parameters and discuss why it requires a Bayesian framework (Section 2.1). The proof of concept is given through parameter estimation experiments for a real-life case study (described in Section 2.3). For this, we first infer parameters for a hydrologic model using binary data collected in our dedicated measurement campaign and then test the sensitivity of inference to experimental design and choice of priors (Section 2.4). In the same section, we also conduct tests on synthetic data from a simple linear model to facilitate reproducibility of this research. Finally, we present the results from simulation experiments (Section 3), provide a critical overview of this technique's limitations and interpret the results (Section 4). At the end, we draw our main conclusions (Section 5).

2 Method and material

In this section we lay out the mathematical formulation of our likelihood function. Once the likelihood function is defined, we discuss how Bayesian inference can be used to combine censored information and prior belief about the parameter values to update the probabilities of parameters. We then describe the case study and the simulation experiments carried out on it. The simulation experiments first show the performance of the inference procedure on binary observations. Finally we describe the sensitivity analysis to the threshold and prior.

2.1 Parameter estimation

Likelihood function

The explicit consideration of structural deficits and input errors in computer models, during the inference procedure of parameters, has been extensively discussed in literature (Higdon et al., 2004; Kennedy and O'Hagan, 2001; Todini, 2008). Reichert and Schuwirth (2012) propose to describe true system response Y_t at time t of an environmental system as the sum of a deterministic model m_t and a stochastic process B_t (see also Fig. 1):

$$Y_t = m_t(\boldsymbol{x}, \boldsymbol{\theta^m}) + B_t(\boldsymbol{\theta^m}) \tag{1}$$

We use the convention to represent random variables with capital letters and write vectors in bold. In Eq. (1), the stochastic process $\boldsymbol{B}$ represents the bias of the model due to structural limitations and input errors. The model inputs are denoted with x and parameters with $\boldsymbol{\theta} = (\boldsymbol{\theta^m}, \boldsymbol{\theta^B})$, where the superscript $\boldsymbol{m}$ denotes the model parameters and B stochastic process parameters. It is reasonable to assume that observations $Y^o = (Y_1^o, Y_2^o, ...)$ of the system response contain observation errors, which are modelled by an independent additive error term (E_t):

$$Y_t^o = Y_t + E_t(\boldsymbol{\theta^E}) \tag{2}$$

For parameter inference, the probability (density) distribution of the observations given the parameters and input, $p_{\boldsymbol{Y_o}}(\boldsymbol{y_o}|\boldsymbol{\theta}, \boldsymbol{x})$ the so called likelihood function, is required. It has been suggested to represent the bias B as a Gaussian process and consider independent normally and identically distributed observation errors E (Kennedy and O'Hagan, 2001). This makes it possible to write the likelihood function as a multivariate normal distribution with mean m and covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \boldsymbol{m}(\boldsymbol{x}, \boldsymbol{\theta}) \tag{3}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma_B} + (\theta_1^{\boldsymbol{E}})^2 \boldsymbol{I} \quad with \quad (\theta_1^{\boldsymbol{B}})^2 \cdot exp\Big(- \frac{1}{2\theta_2^{\boldsymbol{B}}} |t_i - t_j| \Big) \tag{4}$$

where I is the identity matrix, $\theta_2^{\boldsymbol{E}}$, $\theta_1^{\boldsymbol{B}}$ are standard deviations of the two error processes $\boldsymbol{E}$, $\boldsymbol{B}$ and $\theta_2^{\boldsymbol{B}}$ captures the degree of error correlation in time. However, the observation model in Eq. (2) is only adequate for uncensored observations. This prohibits using sensors providing censored observations, for example, measuring the system response only within lower and upper thresholds $[\underline{y}, \overline{y}]$. An observation of such a sensor can be described as (see also Fig. 1):

$$Z_t = \begin{cases} AT & if\ Y_t^o \geq \overline{y} \\ Y_t^o & if\ \underline{y} < Y_t^o < \overline{y} \\ BT & if\ Y_t^o \leq \underline{y} \end{cases} \tag{5}$$

i.e. the observation is either a measurement of the system response or it is "censored" and takes one of the special values: AT (above threshold) or BT (below threshold). Instead of discarding this information, the likelihood function defined by a multivariate normal distribution can be adapted to censored observations $\boldsymbol{Z}$ with an approach similar to what is used in survival analysis (e.g. Klein and Moeschberger, 2003). This is achieved by integrating the likelihood function over all possible values corresponding to either AT or BT (shaded areas in Fig. 1)

$$p_{\boldsymbol{Z}}(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{x}) = \iiint_{-\infty}^{\infty} p_{Y^o}(y_1^o, ..., y_n^o|\boldsymbol{\theta}, \boldsymbol{x}) g(z_1, y_1^o) ... g(z_n, y_n^o) \partial y_1^o ... \partial y_n^o \tag{6}$$

where, n is the number of observations and

$$g(z, y) = \begin{cases} H(y - \overline{y}) & if\ z = AT \\ 1 - H(y - y) & if\ z = BT \\ \delta(y - z) & else \end{cases} \tag{7}$$

is a helper function to ensure the correct limits of integration. The Heaviside function H(x) equals to zero for negative x and one otherwise,δ denotes the Dirac delta function. Although the integrals in the likelihood function Eq. (6) are known to be analytically intractable for normal distributions, a very efficient numerical solution has been suggested (Genz, 1992). In this formulation, binary observations, such as those which we later use in our case study, are a special case of censored observations for which the upper and the lower threshold fall together, i.e. $\underline{y} = \overline{y}$.

Bayesian inference

When the likelihood function is comprised of a deterministic model and an additive stochastic process to capture the remaining bias, a maximum likelihood estimation is not advisable. This is because the data can, in principle, be modelled equally well with both components, which leads to

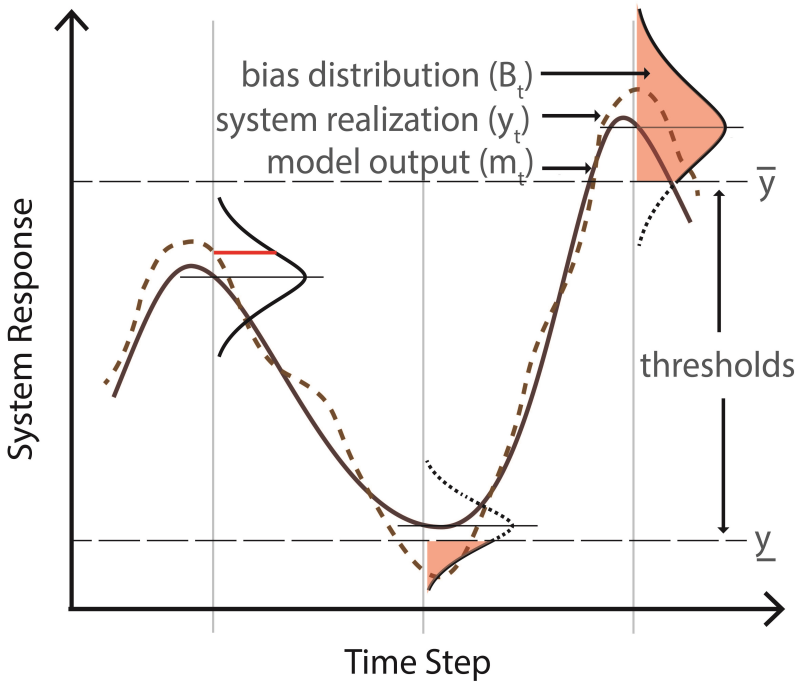


Figure 1: Schematic representation of likelihood estimation for censored data. Integrals of bias distribution (shaded orange) need to be computed in the censored regions (below y and above y), whereas the probability density is computed in between (red line).

identifiability problems (Reichert and Schuwirth, 2012). This dilemma can be solved by introducing our prior beliefs on the parameters, for example that the bias has a high probability for a zero mean. This is possible via Bayesian inference, where our belief is formally expressed as the prior probability distribution of the parameters. In inference the posterior is proportional to the product of the prior probability and likelihood of observing the data, given the hypothesized model:

$$p_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{x}) \propto p_{\boldsymbol{Z}}(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{x})p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{8}$$

2.2 Performance assessment

To gauge the performance of the best parameter estimates prior to inference and posterior to inference, we compute their corresponding Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970). This performance index, which is a popular assessment tool for model performance in the hydrologic modelling community, captures the degree of mismatch between model prediction and uncensored observations. The absolute value of NSE after inference is limited by the performance of the underlying deterministic model. An inference technique can perform well even when the deterministic model is not capturing the system response with adequate accuracy. However, the improvement in NSE from prior parameters to posterior parameters does capture the performance of an inference technique. A performance index, based on similar philosophy as NSE, can be calculated for the binary system response, which tells us to what degree do the binary observations match with the binary predictions. Percentage Matching Binary Observation can be defined as the number of time steps where binary predictions match binary observations divided by the total number of time steps. As we calibrate the model using binary observations, this index is expected to improve in the calibration phase. To quantify the reduction in the parametric uncertainty from the prior to the posteriors, we check the reduction in the spread on the marginal distribution of parameters. For informative data, posteriors are expected to be narrower than priors. Therefore, we

analyse the ratio of the standard deviation of prior to posterior parameter marginal distributions. This contraction in the probability distribution of parameters gives an indication about how much has been learned about the parameter values using the new data of the system response. For a deterministic model, the posterior standard deviation of parameters should decrease as the time series of system response observation becomes longer.

2.3 Case study

We demonstrate the potential of censored observations by using them to calibrate a hydrodynamic rainfall-runoff model for a small urban subcatchment in the Wartegg city district of Lucerne, Switzerland.

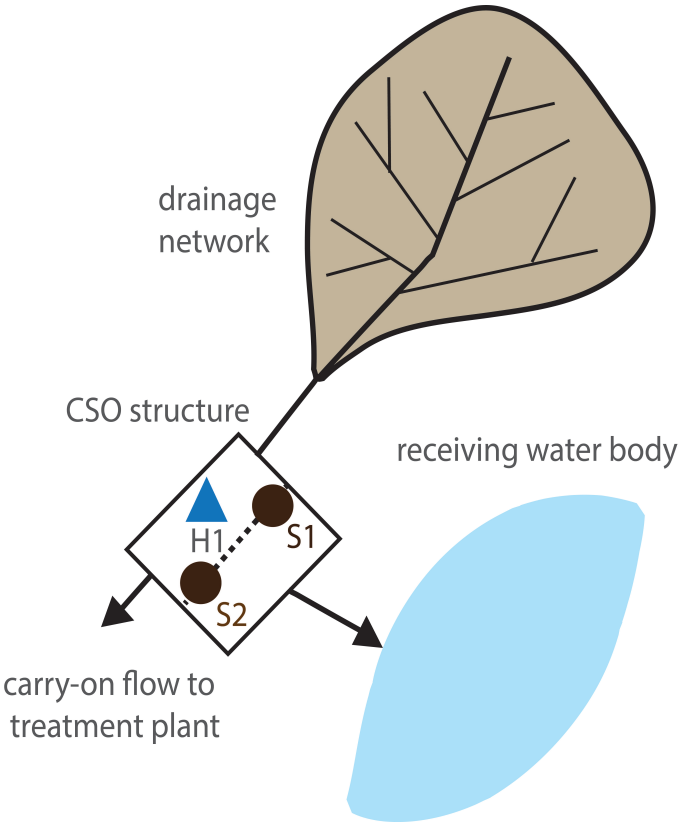


Figure 2: Schematic outline of the experimental setup in the Wartegg subcatchment in the city of Lucerne, CH.

Catchment description

The investigated catchment is a typical Swiss residential area with single to two-story housing of moderate density, including typical public infrastructure, such as a shopping centre, parking lots, and a school complex. The catchment area is drained by separate and combined sewers with a network length of 11.2 km, where an area of 30.2 ha is connected to the combined system. A CSO structure consisting of a side-flow weir and a retention tank with a storage capacity of about 100

m3 is located in the lower part of the catchment (see Figs. 2 and 3). In case of heavy rainfall, the CSO structure diverts excess discharge of combined sewage into Lake Lucerne. In this study, only the combined sewer system is considered.

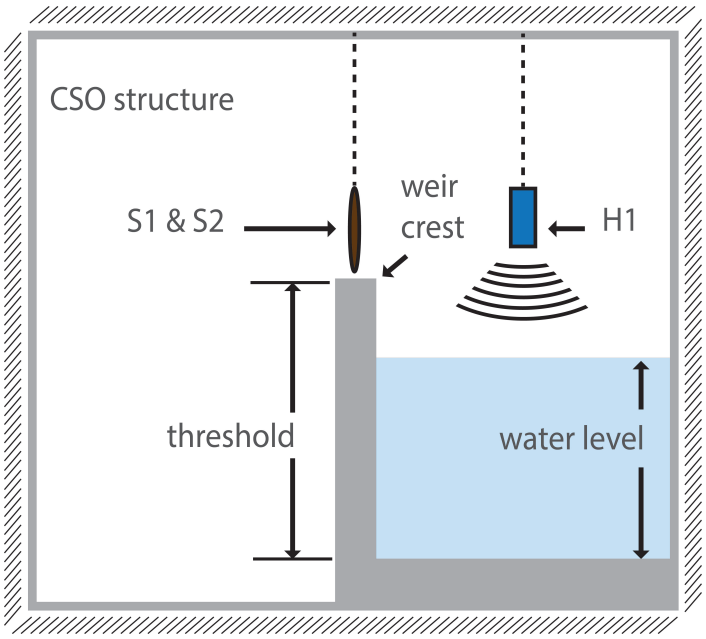


Figure 3: Schematic of the experimental setup at the CSO structure in the Wartegg subcatchment: 1) H1 sensor: ultrasonic measurements (uncensored observations). 2) S1 and S2 TOMST moisture probe: binary observations. 3) Threshold is the weir crest. When the water level overtops the threshold, the S1 and S2 sensors capture the signal.

Uncensored data

We obtained the routine surveillance data from the utility, which operates an ultrasonic water level sensor positioned in the flow-through chamber of the overflow structure (H1, Fig. 2). The recordedwater level H1, hereinafter referred to as uncensored data, is used as a benchmark for the binary data (S1 and S2, Fig. 2). In order to back up the operators' water level data H1, we additionally performed a targeted flow monitoring with two independent flow sensors about 200 m downstream of the overflow structure (carryon flow) from July 2014eJune 2015. Further details on this monitoring setup and instrumentation can be found in Peleg et al. (2017).

Rainfall data is obtained from a rain gauge located about 2 km west of the pilot catchment. The rainegauge is operated by MeteoSwiss, the Swiss Federal Office of Meteorology and Climatology. It records rainfall heights at 10 min intervals with a precision of 0.1 mm, which we validated by hourly data provided through MeteoSwiss.

Binary data

At the overflow structure, two identical low-cost, multiparameter sensors (TMS-3 data logger; TOMST, 2017) were installed, perpendicularly hanging above the weir crest to record temperature, moisture, and movement (S1 and S2 in Figs. 2 and 3). The moisture signal (capacity-related measurement technique) clearly shows elevated levels when the sensor head is submerged, i.e.

when the water level exceeds the threshold (Fig. 3). In the present study, the raw moisture signal was classified into either 'dry' (BT e no overflow) or 'submerged' (AT − overflow) conditions. The autarkic battery-powered sensors were permanently operated for two months from April until the end of May 2015, after their performance had been thoroughly tested and optimized at various locations over a period of several weeks. The recording interval was set to 1 min to sufficiently cover process dynamics.

Rainfall-runoff model

A hydrodynamic sewer model was developed based on infrastructure data from the municipality's cadaster. The drainage system is modelled with EPA SWMM, version 5.1.009 (EPA, 2015) employing a conceptual hydrologic surface runoff model and routing the discharge in sewer pipes using the full set of Saint-Venant equations. The model physically consists of 158 subcatchment units, 167 junction nodes, and 174 conduit links, which overall corresponds to a rather detailed catchment representation. For parameter estimation, we first performed a sensitivity analysis to identify the most influential parameters: i) multiplicative factor for subcatchment width (θ_1^m), ii) maximum infiltration rate (θ_2^m), iii) and infiltration decay (θ_3^m), (both for HORTON infiltration method), which is in agreement with previous work e.g. (Liong et al., 1991). We then inferred these SWMM parameters together with those of the error model. The SWMM reporting step was chosen to be 5 min.

2.4 Simulation experiments

To be able to better interpret the performance and applicability of our inference technique, we performed simulation experiments for parameter estimation using the likelihood function described in Section 2.1 on 1) a simple linear model 2) and then on the case study described in Section 2.3. 3) We also conduct analysis to understand the dependence of inference on the value of threshold and the choice of priors.

2.5 Preliminary proof of concept using linear regression

As a preliminary test for the validity of this technique, we used a simple linear model with known "true" parameters. This model has been used before to compare the performance of various likelihood functions (Stedinger et al., 2008):

$$\boldsymbol{Q} = a\boldsymbol{P} + b \tag{9}$$

where $\boldsymbol{Q}$ is the model output, $\boldsymbol{P}$ is the model input and (a,b) are the model parameters. Using the bias description from Eq. (1) to Eq. (4), we defined a likelihood function where the value of parameter vector $(a, b, \theta_1^{\boldsymbol{E}}, \theta_1^{\boldsymbol{B}}, \theta_2^{\boldsymbol{B}})$ is (1,1,0.1,0.3,3). To generate synthetic uncensored observations, we used a sinusoidal $\boldsymbol{P}$, $sin(t/10) + 1$, and sampled from the likelihood function. We then get binary observations by fixing a threshold of 2.5 for the uncensored observations. Data was generated for a thousand t values, from 1 to 1000. The prior belief for $(a, b, \theta_1^{\boldsymbol{E}}, \theta_1^{\boldsymbol{B}}, \theta_2^{\boldsymbol{B}})$was defined using normal truncated distributions, having (1.5,1.5,0.08,0.2,2) as means, (1,1,0.1,0.3,1) as standard deviations and with the lower and upper bounds of (0.25,0.25,0,0,0.1) and (4,4,0.2,2,4) respectively. Parameter inference was done for the first five hundred binary observation values as the calibration event, and the inferred parameters were verified on the remaining time series.

Parameter estimation for the case study

We conduct parameter inference using real binary data obtained from our dedicated measurement campaign (see Section 2.3). These binary observations can contain false positives and negatives, and allow studying the quality of inference for real measurements as well as the robustness of the technique to noisy data. Based on the duration of our monitoring campaign for binary data, three

rainfall events were used to assess the ability of binary likelihood function in inferring the model parameters, for which binary observations were available. The model was calibrated on one event, from 03/05/ 2015 12:00 to 05/05/2015 00:00, and then tested by predicting water levels for two different events, from 25/05/2015 03:00 to 27/ 05/2015 00:00.

For simplicity, while doing the inference, we only focus on $\boldsymbol{B}$ (Eq. (1)) and do not consider $\boldsymbol{E}$ (Eq. (2)). In our view, this is reasonable because i) model structure deficits and input errors are supposedly much more influential than random observation errors, i.e. detect/ non-detects, ii) detectors of threshold exceedance are very simple devices which in turn are rather robust and, at least in our case, show very little randomness. While the optimization is done on posterior probability density of the parameters, we also compared the Nash-Sutcliffe efficiency (NSE) of the model that has been calibrated on the binary dataset to the NSE of prior parameters. As variables in hydrological systems, such as water levels in this case study, are often non-negative and errors are often positively correlated with the magnitude of the variable, e.g. large uncertainties with high flows/water levels, it is more realistic to first transform the system response with a function f, and then to apply Eq. (6) in the transformed space. This also requires transforming the thresholds $f(\underline{y})$, $f(\overline{y})$. For simplicity, we here suggest the Box-Cox transformation (Box and Cox, 1964)

$$f(y) = a\frac{y^{\lambda}-1}{\lambda} \tag{10}$$

which is commonly applied in many hydrologic case studies (Bates and Campbell, 2001; Del Giudice et al., 2013; Frey et al., 2011; Sikorska et al., 2012). Depending on the individual application, other transformations might be more realistic (Wang et al., 2012). The values for the Box-Cox transformation parameter was suitably chosen from the past experience in similar urban hydrologic studies (Del Giudice et al., 2013). The value of 0.35 produced satisfactory results for inference in the preliminary analysis. In this simulation, the model parameters that we calibrate are i) multiplicative constant for the width of the subcatchments (mult.width) and ii) the maximum infiltration rate (maxrate).We assume a prior distribution for our parameter space, capturing our best guess about the system. However, the prior distribution is kept wide so that we can learn from our measured data.

Table 1: Information on priors used in the sensitivity analysis. All the priors of elements of q are normal truncated. The vector in the table give the values for mean, standard deviation, lower limit and upper limit of the prior for each parameter in our inference. $\theta^{\boldsymbol{m}}$ represents model parameters and $\theta^{\boldsymbol{B}}$ the bias parameters in the transformed space (Eqs. (1) and (4)).

	$\theta_1^{\boldsymbol{m}}(mult.width)$	$\theta_2^{\boldsymbol{m}}(max.rate)$	$\theta_3^{\boldsymbol{m}}(decay)$	$\theta_1^{\boldsymbol{B}}(sd.B\_Q)$	$\theta_2^{\boldsymbol{B}}(corrlen)$
prior 1	[3,1,0.01,7]			[0.54,0.54,0,10^8]	[0.05,0.05,0.01,0.01]
prior 2	[0.1,0.5,0.01,7]			[0.54,0.54,0,10^8]	[0.05,0.05,0.01,0.01]
prior 3	[3,1,0.01,7]	[90,30,10,100]		[0.54,0.54,0,10^8]	[0.05,0.05,0.01,0.01]
prior 4	[0.1,0.5,0.01,7]	[37,5,30,10,100]		[0.54,0.54,0,10^8]	[0.05,0.05,0.01,0.01]
prior 5	[3,1,0.01,7]	[90,30,10,100]	[2,2,1,7]	[0.54,0.54,0,10^8]	[0.05,0.05,0.01,0.01]
prior 6	[0.1,0.5,0.01,7]	[37,5,30,10,100]	[2,2,1,7]	[0.54,0.54,0,10^8]	[0.05,0.05,0.01,0.01]

Sensitivity to threshold and prior

The effect of different priors was studied when inferring parameters. The standard deviation and mean of the prior was changed to gauge its impact on the inference procedure. The prior distributions should describe the knowledge of the analyst on model and error parameters, which also has to satisfy physical constraints, e.g. that negative pipe roughness or imperviousness values are impossible. Therefore, normal truncated distribution was used as the probability density function for priors, and five sets were chosen for the bounds, mean and the standard deviation for

each simulation. We do not analyse the effect of using different types of probability distributions as priors e.g. lognormal, uniform etc. As in practice, different analysts often do not have the same level of expertise or available databases for different parameters, so we also alter the dimensionality of parameter space, by increasing the number of calibration parameters from 1 to 2 and then to 3, with different specifications of prior distributions (Table 1). The prior 1, 3 and 5 overestimate the peaks, whereas prior 2, 4 and 6 underestimate them. The uncertainty in the prior belief about parameter values is captured by the spread of the prior distribution. Wide as well as narrow priors were tested. The use of priors can help eliminate unrealistic parameter estimates and allows for the incorporation of accumulated knowledge about a system like a river catchment or a drainage network. As it is challenging to formulate the priors of the error model parameters, we have good experiences with defining them based on preliminary analysis of residuals of past observation periods. Besides, it is safe to assume that bias is unlikely to be higher than the standard deviation of the system response itself.

Apart from doing a sensitivity analysis on priors, the general ability of the likelihood function to infer parameters for different thresholds was also tested using synthetic binary data. Simulations were conducted for the Lucerne catchment and the synthetic binary observations were constructed from the uncensored observations of water level H1 and five different thresholds (Fig. 3). Our prior knowledge on the model parameters was formulated as prior 1 (Table 1), then parameter inference was done using synthetic binaries corresponding to various thresholds. The effect of the threshold choice on parameter inference was studied using this preliminary analysis.

Implementation

We implemented the inference in the programming language R (CRAN, 2015) and used the pmvnorm function of the package mvtnorm (Genz et al., 2015) to compute integrals of the likelihood function. The mode of the posterior distribution in Eq. (8) can be estimated with an effort comparable to the use of an informal objective function. If the whole distribution is of interest, Markov Chain Monte Carlo (MCMC) methods are used to sample from the posterior (Vihola, 2012).

To speed up the inference, we constructed a fast surrogate of the SWMM model, a so-called "emulator" and used it instead of the original SWMM model in the calibration phase (Carbajal et al., 2017). Although a single run of our SWMM simulator only takes about 20 s, this translates into an inference time of more than 55 h for an MCMC chain of 10,000 runs. In addition, the evaluation of the likelihood function requires numerical integration for each data point, which costs additional 10 s for each evaluation. To alleviate this cost, we followed the recipes in Carbajal et al. (2017) and built a data-driven emulator using the non-negative matrix factorization approach (more information in supplementary material). We produced 500 simulations with parameter values randomly drawn from the prior distributions and use the first 250 as training examples. The emulator prediction was evaluated in the remaining 250 simulations, achieving relative root mean square errors below 2%. The emulator demands 0.04 s per simulation, 500 times faster than the simulator, and reduces the simulation cost of the mentioned MCMC chain from about 30 h to 5 h, given the same overhead computational costs for both simulator and emulator to evaluate the integral in Eq. (6).

3 Results

3.1 Parameter estimation for the linear model

We found that the synthetic binary data was able to guide the inference procedure towards the true parameter values. The maximum prior probability density parameter values that had a poor NSE (0.94) were updated to parameter values that produce a significantly better NSE (0.84), after inference where five hundred data points were used (Fig. 4). The maximum posterior density parameter vector produced after inference is closer to its true values ($a = 1.03, b = 0.97, \theta_1^{\boldsymbol{E}} =$

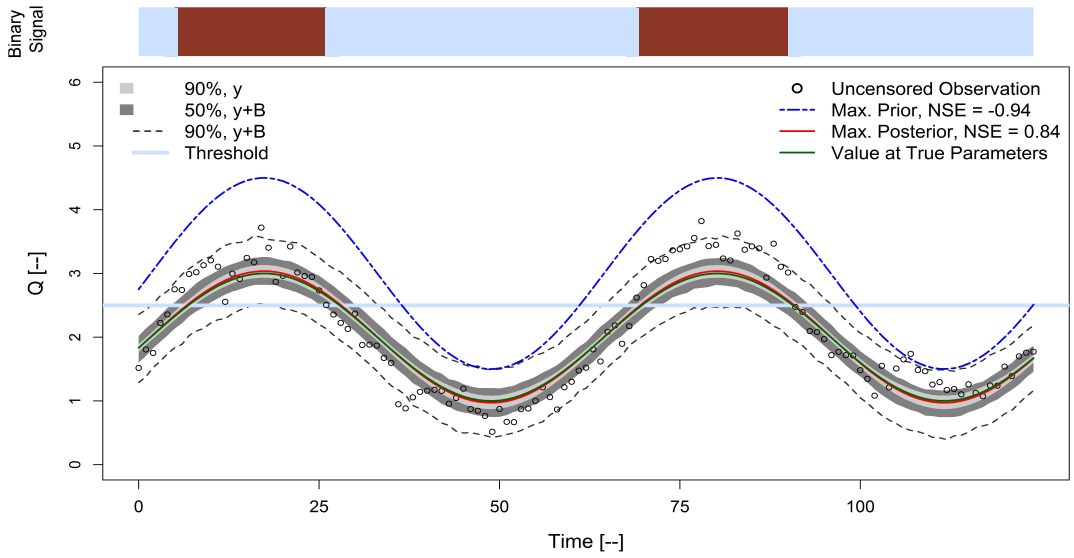


Figure 4: The improvement in the model performance in the validation phase after inference using five hundred binary data points. The dashed blue line is the model prediction with the prior best estimate of parameters and the red line with the posterior best estimate. The y band (light grey area) represents the 90% interval generated after sampling from the posterior parameter distribution. The dark grey band and dashed black lines represent the 50% and 90% intervals respectively for $y + B$. Refer to Eq. (1).

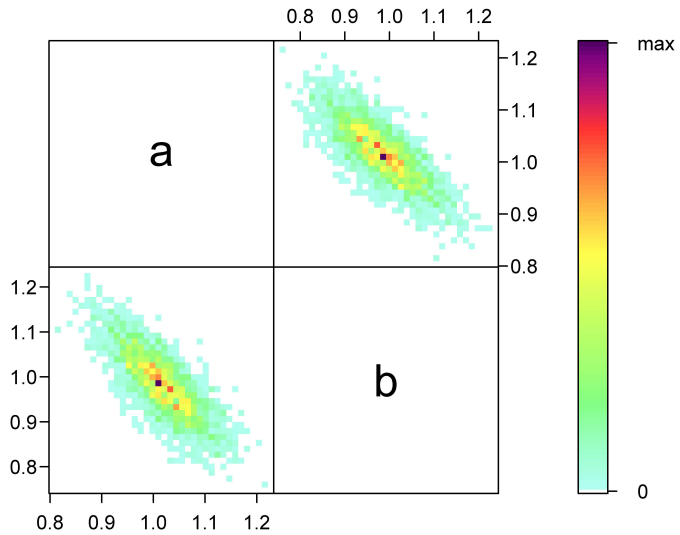


Figure 5: Bivariate posterior distribution of model parameters a and b. The negative correlation that is not assumed by the prior gets captured in the posterior.

$0.075, \theta_1^B = 0.29, \theta_2^B = 2.03$). We found that the dependence in the joint density of parameters a and b was captured, showing a negative correlation (Fig. 5). We were also able to reproduce the decrease in parametric uncertainty with increasing time series of binary data used for inference (from hundred to thousand data points), with a decrease of around 20% in the standard deviation of posterior marginal distribution of the model parameters.

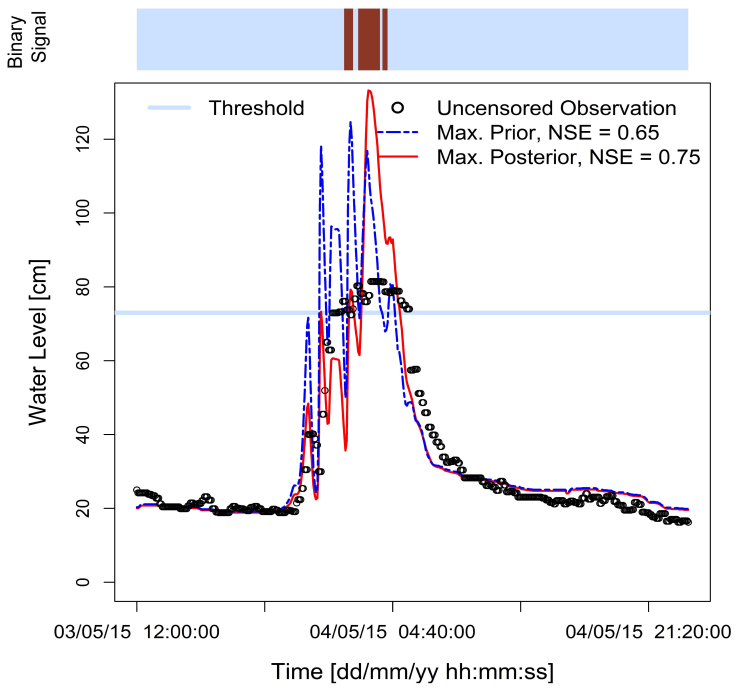


Figure 6: The improvement in the model performance for the calibration phase after inference using binary data.

3.2 Parameter estimation using the case study

The main aim of this simulation experiment was to analyse the potential of binary data in directing us from poor prior estimates of parameters (fully uncalibrated model) towards better posterior estimates using real binary data (Fig. 6).

Table 2: Comparison of NSE for maximum prior and maximum posterior density parameter vectors (Table 1). Comparing the prior and posterior NSEs for the validation period shows the improvement in model performance after inference.

	Calibration Period		Validation Period	
	NSE (Max. Prior)	NSE (Max. Posterior)	NSE (Max. Prior)	NSE (Max. Posterior)
prior 1, one parameter	0.56	0.83	-0.61	0.32
prior 2, one parameter	0.47	0.82	-0.01	0.55
prior 3, two parameter	0.65	0.75	-0.61	0.46
prior 4, two parameter	0.65	0.64	-0.02	0.04
prior 5, three parameter	0.65	0.78	-0.61	0.43
prior 6, three parameter	0.52	0.74	0.00	0.33

For this model, we see that the parametric uncertainty reflects itself much more pronouncedly in the peaks of the water levels, and does not influence the base levels much. The prior parameters grossly overestimate the water level peaks in the validation phase (Fig. 7).We see that the posterior obtained using uncensored data is narrower than the posterior using binary data, which is in turn narrower than the prior (Fig. 8a). The shrinkage in posterior captures the information gain in parameter values, and we also see that the joint distribution captures the dependence between various parameters (Fig. 8b). After calibration, the parameter estimates are taken away from

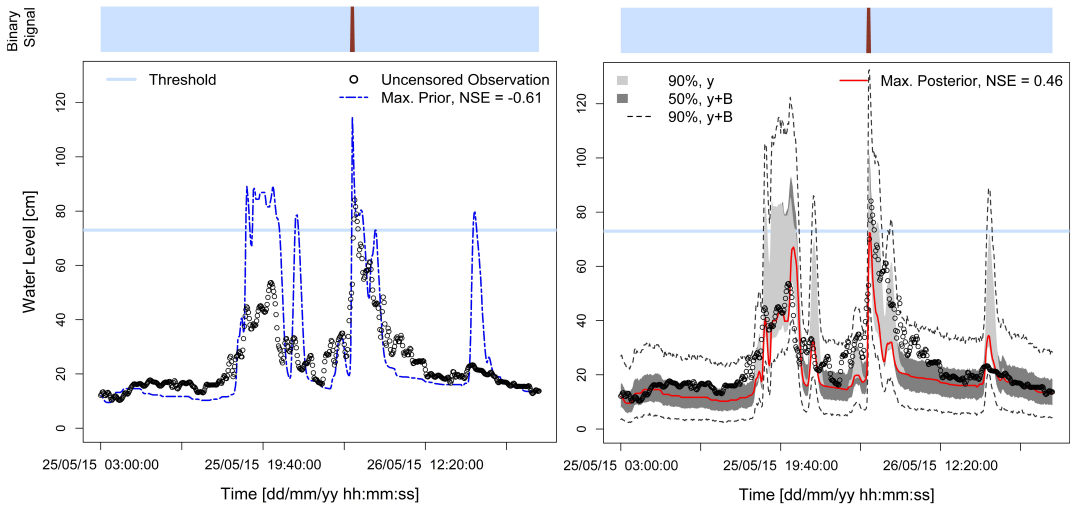


Figure 7: The improvement in the model performance for the validation phase after inference using binary data.

regions which grossly overestimate the water level and higher probability is given to sections of parameter space which make model predictions that are in conformity with binary observations. We also see that the uncertainty bands stay wide as the inference is not provided with the uncensored data, therefore staying underconfident in defining the uncertainty intervals. Parametric uncertainty can be further reduced by utilizing newly obtained observations. The posterior distribution from one calibration event can be used as a prior during further inference on new events.

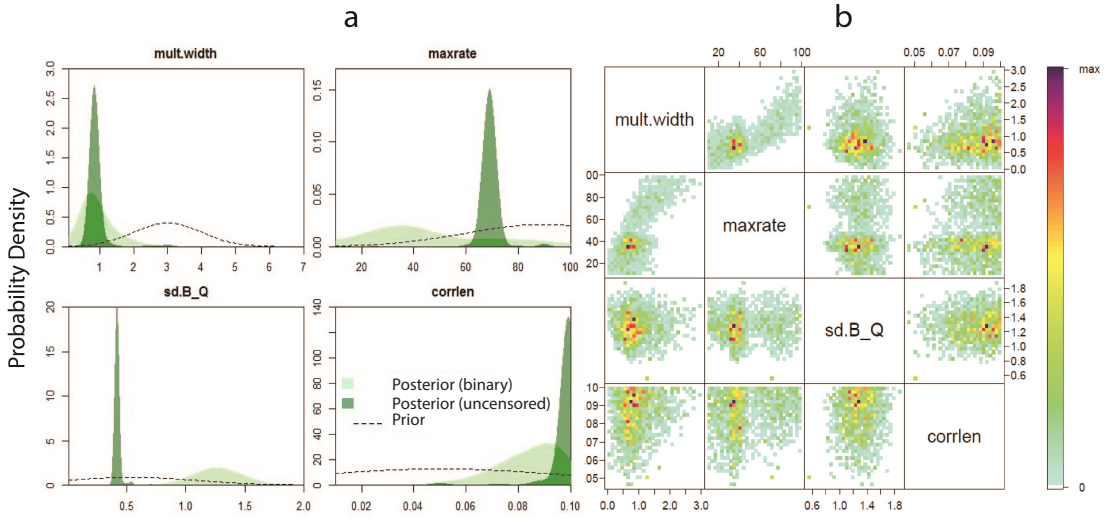


Figure 8: Posterior distribution of model and bias parameters; (a) Marginal from binary and uncensored data, (b) Bivariate from binary data.

3.3 Sensitivity to threshold and prior

We find that for both calibration and validation periods, predictions based on the posterior outperform predictions based on prior knowledge. However, the quality of parameter inference is sensitive

to the prior, i.e. the knowledge of or information available to the analyst. This is an expected result in Bayesian inference, as posterior probability distribution is, by definition, proportional to the product of prior and the likelihood.We find that the parameter posteriors are always narrower than their priors, thus greatly reducing the uncertainty in parameters (Fig. 9).

The inference leads to better parameter values, even when multiple parameters are inferred simultaneously (Table 2). This dependence on priors is expected to be pronounced for short calibration periods; theoretically, as longer and longer time series of data is used for inference, the impact of the likelihood function should outweigh the impact of priors in determining the posterior. Also, the posterior distribution generated gets narrower with increasing time series. However, as expected, inference using uncensored data for the same calibration period is much more informative than the binary data, especially in narrowing down the posteriors. Given the fact the binary information does not capture the detailed dynamics of a system response, the uncertainty intervals stay underconfident.

For varying threshold, the maximization of posterior distribution from the calibration event using synthetic binary data shows that the NSE of the posterior is higher for those thresholds which have a proportionate ratio of positives and negatives (Figs. 1 and 11). Thresholds which always (or never) indicate elevated water levels are less informative, or, in other words, virtually "blind" and thus do not contain much information to discriminate between likely and unlikely parameter values. The experiment was done using six different threshold values of water level; the observed maximum, minimum, and average water level for the event being 81 cm,16 cm, and 32 cm respectively. This means that the threshold value below 16 cm will have only AT as binary observations and the threshold value above 81 cm will have only BT as binary observations. Given the fact that we have a fixed model structure, and we introduce a specific rainfall time series as input, the range of parameter combinations that can produce a particular sequence of binary signals is not wide. Therefore, while the information content of the signal containing only BT or AT is limited (Eq. (5)), the likelihood function is surprisingly useful to learn some information about the parameter values.

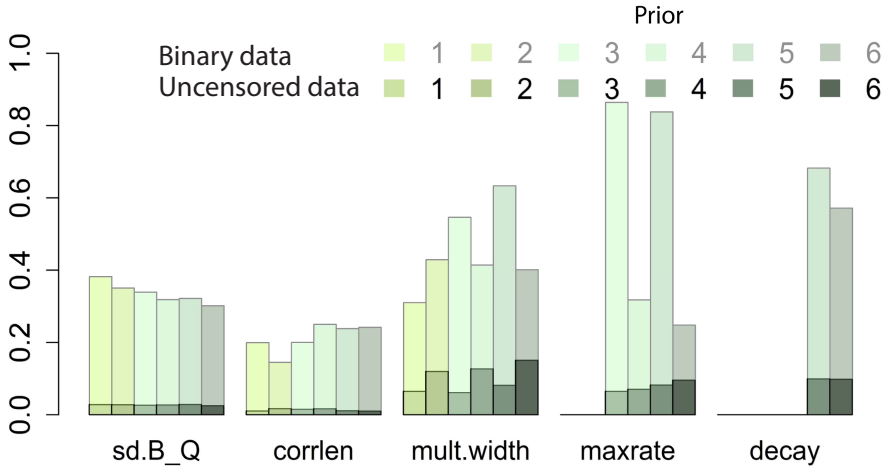


Figure 9: Sensitivity to different priors and the gain of information from binary observations. The graph depicts the ratio of the posterior standard deviation to the prior standard deviation of parameters.

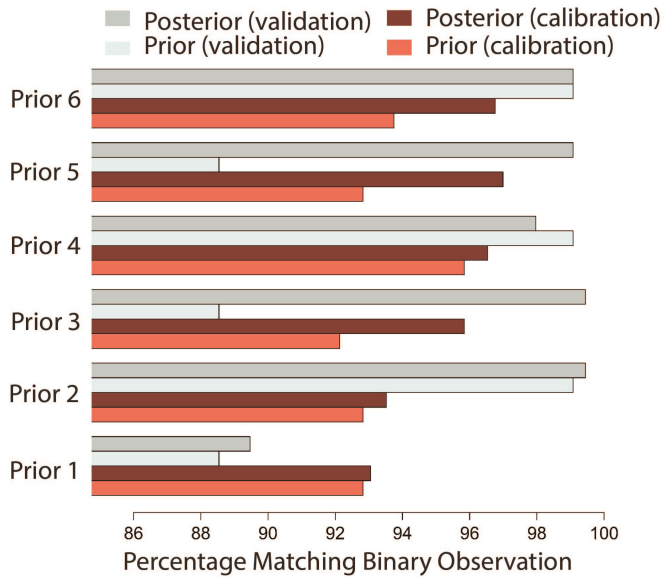


Figure 10: Percentage Matching Binary Observation using the maximum probability density prior parameters and the maximum probability density posterior parameters. Inference in the calibration phase yields a noticeable increase in the model output matching with the binary observations.

4 Discussion

The results of this research are in accordance with the theoretical framework that we employed. We expected noticeable improvements in model performance and our results confirm this. The main aim of this research was to formulate a formal Bayesian framework in which censored data can be used for parameter inference. It is difficult to properly associate probabilities to parameter values using ad-hoc methods. Also, it has not yet been suggested how to combine uncensored and censored observations using such non-Bayesian inference procedures for hydrologic models. In the following sections, we examine the performance of a likelihood function that can integrate the inference procedure for uncensored and censored data in one consistent framework. i) We specifically discuss the information gained with respect to parameter values and the performance of the model with maximum posterior parameters. ii) We discuss the ability of censored data in the quantification of model uncertainty. iii) We interpret our findings vis-a-vis the dependence of inference on the choice of the threshold.We discuss the effect of priors on inference. iv) We also discuss the accuracy of binary observations.Towards the end of the discussion, we talk about the big picture and make some recommendations on data collection and sensor placement so that data collected is better suited for calibration exercises.

4.1 Information gain from binary observations

As a simple validation of the likelihood function, Figs. 4 and 5 show that, for a linear model, inference using synthetic binary observations is able to reproduce the true parameter values reasonably well. For the real case study, where the assumptions of the likelihood function are only approximately valid, inference is still able to provide improved parameter estimates. Table 2 and Fig. 9 show that, despite the fact that the information content of censored data is inferior to uncensored observations, censored data in combination with a rigid model structure (SWMM in this study) and a well-defined input in the form of a rainfall time series allow us to learn about better parameter values. From all the different parameter combinations that the prior distribution allow, only a limited range of parameter values would produce the model output that agrees with the observed censored signal. Thus, the maximum posterior parameter values perform noticeably

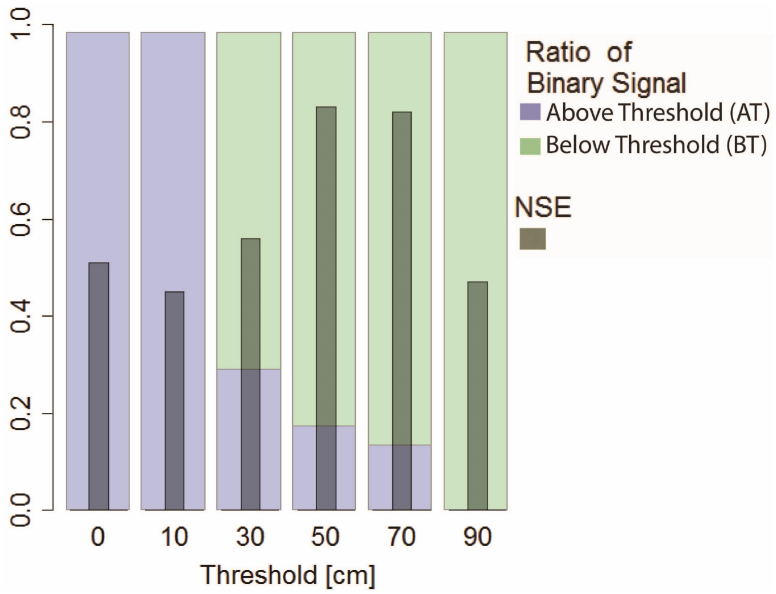


Figure 11: The sensitivity of NSE to placing a binary sensor at different threshold levels in the cross-section. Results correspond to the parameter values at maximum posterior probability density in the calibration phase. Performance for thresholds that only contain AT or BT signals is poorer compared to other thresholds. Synthetic binary data corresponding to different thresholds were generated from the original uncensored observations.

better than the maximum prior parameter values. In the presence of uncensored observations, our description of likelihood function evaluates the probability of observing a particular series of data, given the chosen SWMM and error model, for a specific combination of parameters. Whereas, when we have censored signals, the same description of likelihood evaluates the probabilities while taking into account a) the amount of exceedance/non-exceedance and b) the duration of that exceedance/non-exceedance. For example, the top of Fig. 6 shows an example of a binary signal and its duration. We see that the best estimate of parameter values after inference reproduce the observed binary signal more accurately than the uncalibrated model (Figs. 6 and 10). We also see that the calibration phase narrows down the parameter range after exploring the parameter space (Fig. 8). Parameters that cause longer durations of exceedance or shorter durations of exceedance are assigned smaller probabilities than those which correspond closely to the observed binary signal. Therefore, rainfall events that result in multiple instances of flow over the threshold will lead to better parameter estimation.

4.2 Ability of censored data in the quantification of model uncertainty

Given a stochastic model, identification of the "true" parameters using a finite set of samples of the system response always comes with uncertainty. As the number of observations increases, the posterior probability distribution becomes narrower. This was also observed during the inference of the linear model parameters. The availability of censored data, in the absence of uncensored observations of the system response, still allows us to narrow down the parameter space and, interestingly, capture their dependence with each other, as it can be seen in Fig. 8. For hydrologic forecasting, the parametric uncertainty is usually found to be smaller than the model structure and input uncertainty when there is adequate reference data to calibrate the model (Del Giudice et al., 2015). Fig. 7 shows the uncertainty bands generated by sampling the parameters from the posterior distribution and running simulations using those parameter samples, labelled "90% y", referring to the quantile and the deterministic model without bias respectively. As the binary observations

are insufficient to assess the mismatch of model predictions and the actual system response, the standard deviation of the bias process is not reduced drastically through the inference procedure. This is expected, as binary data provides very little information about the actual system response, thus forcing the error model to prevent overconfident intervals. This results in uncertainty bands that are wide. In our view, it therefore seems best to combine binary and uncensored data to reduce prediction uncertainty or utilize longer time series of data for inference.

4.3 Sensitivity to prior knowledge and threshold

The analysis also shows that the posteriors are sensitive to priors. Theoretically, this should be the case, especially if the event duration used for calibration is small. Bayesian probabilities represent the degree of belief and the posterior represents the combined effect of prior beliefs and the updating as a result of new information. Fig. 9 shows the reduction in parametric uncertainty after inference, for various priors. We see that different priors lead to different amounts of shrinkage in the density function of parameters, however, it does not hamper the inference procedure and the binary data is still able to guide us towards better posterior distribution. As far as deciding the prior distributions isconcerned, we recommend not to exclude regions of parameter space that can potentially contain the sought parameter vector even though it is not highly likely, while choosing truncated distributions. As far as the sensitivity to the choice of threshold is concerned, it is clear that signals with only BTs or ATs are the least informative. However, it is difficult to analytically formulate an exact relationship between the "value" of the censored signal in providing better parameter estimates and the ratio of BT and AT. This is because other factors, such as proximity of the actual system response to the threshold or the number of threshold crossovers, play a role. Our results show that the ability of the likelihood to learn about model parameters does depend on the threshold (Fig. 11). This can be explained, in part, by the high information content that crossing the threshold has. As the model has to be near the threshold when the observations flip from one side of the threshold to the other, the model learns more about the parameters which allow for such a model response. Thus, the ratio of positives and negatives and the number of cross overs in a binary signal will have a bearing on the information content of the data. In a real hydrologic network, different thresholds would correspond to installing an electrical switch or motion sensor, at different heights in the cross-section. Although in most real applications, binary sensors would detect overflow events at a certain weir with given height or a predetermined critical levels that, for example, would lead to local flooding, we conclude from our first set of simulation experiments that the placement is most informative when a proportionate amount of BTs and ATs are detected. In real systems, the parallel installation of several sensors at different threshold heights may address this dependency. This seems most relevant regarding the experimental design for an entire network, where the information content from multiple threshold detectors can be maximized by installing them at different points and at different levels. In practical applications, the information content also depends on the state, i.e. level. For a utility manager, it is probably interesting to install all sensors at levels that correspond to 80% or 90% pipe capacity. Thus, continuously observing "no detects" in the entire network shows her/him that the network has sufficient hydraulic capacity.

4.4 Accuracy of binary observations

The performance of the binary sensors applied in this case study must be considered imperfect regarding sensor operation and accuracy. This holds true despite the fact that we installed two sensors of the same type in parallel (redundant application). The given binary observations are afflicted with noise, which is expected to influence the calibration performance. In this research, we did not include a specific error model of the binary observation instrument $\boldsymbol{E}$, because a normal distribution is not necessarily a good error model for the noise. A dedicated error model incorporating an explicit false positive or a false negative rate for binary signals is relatively easy to formulate, however, the computational effort scales with 2^n, where n is the number of observations. Also, because the bias process $\boldsymbol{B}$ is generally capable of capturing remaining uncertainty, for

example from noisy output observations, we do not see a practical need for greater detail in the error model. Ongoing tests with alternative low-cost sensor applications show that more accurate observations can be obtained compared to the accuracy of the data used in this study. This underlines the future potential of censored signals, in particular, when combined with low-power wireless data transfer techniques. For measurement campaigns, we recommend obtaining a combination of uncensored and censored data. That can be a good trade-off between having only expensive accurate observations or only less informative censored signals. An optimal experimental design with binary sensors, however, has to consider their spatial location in the hydrologic network and the desired threshold levels. This goes beyond the scope of the article and requires further investigation.

5 Conclusion

We see that monitoring the dynamic response of hydrologic systems is often constrained by inadequately available manpower, technology and financial resources. Data from traditional flow and water level observation devices is hence limited. In the future, Internet-of-Things type communication platforms will also make it possible to operate low-power sensors that can transfer comparably little and often aggregated or censored data, such as the duration of an overflow or whether a threshold was exceeded or not. In this study, we suggest a suitable likelihood function to optimally use such sensor data in model calibration. This likelihood function can handle both uncensored and censored observations, without having to alter the underlying mathematical description of errors due to model structure and input. Based on theoretical considerations and the results from a real case study, where we predicted urban rainfall runoff, we draw our main conclusions:

- Given current developments in sensing, increasingly more sensors are becoming available, which often provide only censored observations. We suggest a likelihood function that is able to reliably learn about model parameter values from censored data.
- As our models and input data are always prone to deficiencies, we recommend choosing a likelihood function that can capture such deficiencies. One possibility is the usage of additive autocorrelated error models, which capture remaining uncertainties. Furthermore, using a Bayesian inference framework will help to avoid identifiability problems.
- Our results from a real-world case study in the city of Lucerne demonstrate that parameter estimation using censored data is beneficial, because the predictions are significantly better than the prediction with priors, with a change in NSE from 0.61 to 0.46.We also find that the posterior parameter distributions are substantially narrower than the priors, thus reducing the model parametric uncertainty (standard deviation) on average by around 45%.
- We demonstrate that the inference procedure is sensitive to the choice of priors, which is expected when little or comparably less informative data is used.
- Using a likelihood function makes it possible to construct prediction intervals, which have a sound probabilistic interpretation. Although the uncertainty bands of the calibrated model are seemingly wide, they can be much narrower than those of the uncalibrated model. This is because a relatively short time series of binary signals should not be able to reduce the uncertainty in the model predictions significantly as only little information about the system is available. The likelihood takes that into account and keeps the bands wide.
- We find that, for having informative censored observations, a proper experimental design is crucial. This is because the location of the sensor, in the hydrologic network as well as within the cross-section (i.e. the resulting level of the threshold), affects the ability of the model to learn from such data. For model calibration, thresholds are advised to be chosen such that proportionate ratio of positive and negative signals can be observed.

To summarize, this research delineates a formal Bayesian framework so that we can utilize censored information, representing a variety of data types, for model parameter inference. We validate this concept using a synthetic and a real case study. The results show that the parameter estimation using censored data works quite well.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.watres.2017.05.038.

Acknowledgement: We gratefully acknowledge the help of Tobias Doppler (Eawag, VSA) during the field monitoring in Lucerne.We thank MeteoSwiss, the Swiss Federal Office of Meteorology and Climatology and the city of Lucerne for providing us with the precipitation and infrastructure data. We furthermore would like to thank the Engineering Consultants from HOLINGER AG, Bern for assisting us with details on the hydraulic model and extracting operation data from the central operator database. We acknowledge the work of a group of master's students from ETH Zürich who supported the initial phase of the case study analysis. We thank H. Badger and Candace Chow for proofreading. Finally, we extend our thanks to the editor and two anonymous reviewers whose comments helped to improve the manuscript. Part of this study was supported by the EC FP7 project QUICS (Quantifying Uncertainty in Integrated Catchment Studies), grant agreement no. 607000.

Chapter 5

Exploring a copula-based alternative to additive error models—for non-negative and autocorrelated time series in hydrology

Omar Wani[1,2], Andreas Scheidegger[2], Francesca Cecinati[3],Gabriel Espadas[1,4], Jörg Rieckermann[1]

[1] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[2] Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland
[3] Department of Architecture and Civil Engineering, University of Bath, Bath, UK
[4] (formerly at) Seminar for Statistics, ETH Zürich, Zürich, Switzerland

Journal: Journal of Hydrology

Submission date: 27$^{\text{th}}$ August 2018

Author contributions: OW, with contributions from AS, FC, and JR, wrote and structured the main text. OW, with inputs from AS, GE, FC, and JR, conceptualized the methodology. OW, AS, FC, and GE ran simulation experiments. JR and AS supervised the research. All authors reviewed the paper.

Abstract

Inaccurate description of uncertainty in the error models can cause biases in parameter estimation. For example poor accounting of low flow errors can lead to poor prediction for high and medium flows. When the parameters of the deterministic model and the error model are inferred jointly from the observations, the posterior converges to regions which reflect the processes in both high and low flows. If the nature of errors in low and high flows is different to the extent that the same error description cannot be used for both, biases in inference are introduced. In such cases, the parameter posterior will adjust to region of the hydrograph with longer proportionate presence in the calibration time series. In this paper we demonstrate that the autoregressive order 1 (AR1) description of errors can lead to sub-optimally performing predictive models if the calibration period has substantial sections of inadequately modelled flows. Inference is performed within the Bayesian framework. We show this for a synthetic example as well as a case study. We also see that the predictive uncertainty bands that we get using the AR1 description can be overconfident and also admit negative values. To mitigate this, we analyze an alternative to additive error models. We use a distribution with a non-negative support, gamma in this study, reflecting the uncertainty in the system response at every time step. The gamma distribution is conditioned on the deterministic model output, which determines its mode and standard deviation. We capture autocorrelation in time using copulas. Given that copulas can capture dependence between different marginals, we use different specifications of the marginal distribution for high and low flows. The results show: 1) Biases in parameter estimation can be reduced if a representative error description is attained using the flexibility of a copula-based likelihood 2) the non-negative support allows to make more realistic uncertainty intervals for low flows. 3) We see that the autocorrelation parameter in copulas severely interacts with the model and heteroscedasticity parameters. 4) While the formulation, in principle, should be of added value for parameter inference, in case of less informative priors, the flexibility of this description can produce non-robust inference.

1 Introduction

Due to the complexity of hydrologic processes like rainfall-runoff, deterministic modelling proves to be inadequate for efficient decision making (Krzysztofowicz, 2001; Verkade and Werner, 2011). Use of probability theory helps quantify uncertainty in such predictions (Gupta et al., 1998; Honti et al., 2013; Kuczera et al., 2006; Refsgaard et al., 2007). Specifically within the Bayesian framework, to further constrain the range of parameters values that seem feasible for a catchment a prioiri, observations of the system response are used for parameter probability density updating. The physical understanding of the system, as well as the understanding of errors is put in formulating the conditional probability density of system response, $p(y_o|x, \theta)$, given parameters (θ) and input forcing (x). This reflects our assumption about the system as the observation generating random processes. Bayes' theorem is then used to invert the conditioning and get the probability density of parameters given some observations of the system i.e. $p(\theta|x, y_o)$ (Hall et al., 2011; Kavetski, 2018; Kennedy and O'Hagan, 2001).

However, like the assumptions about the deterministic model, our assumptions about the errors should be representative to an adequate degree so that inference gives meaningful parameter estimates. The rainfall-runoff time series usually has these properties: 1) scaling with rainfall, 2) autocorrelation in time, and 3) heteroscedasticity i.e. changing variance and 4) non-negativity. While the deterministic rainfall-runoff models provide information about the aggregate tendencies of the flow, for an unbiased parameter estimation and accurate uncertainty estimation, the hydrologic error model should be able to reproduce the other stochastic properties to an adequate degree (Del Giudice et al., 2016). However, it is not straight forward to reproduce all of these properties in the formulation of the likelihood function without some further simplifying assumptions, for example allowing the possibility of negative flow predictions in additive error models (Sun et al., 2017). Also, the structural deficits in many hydrologic models may come from very systematic biases which may not lend themselves adequately to one single probabilistic description for the entire hydrologic time series. Data from different hydrologic regimes, like low and high flows, may appear to be a result of different data generating random processes that need their own unique probabilistic model, e.g different autocorrelation (Ammann et al., 2018). Generally the modeler prioritizes one property over the other depending on the task for which the model is being designed.

To make error descriptions more representative of the rainfall-runoff, several parameterizations for the likelihood function have been proposed, like for example, to capture heteroscedasticity (Sorooshian, 1981) , skewness (Schoups and Vrugt, 2010), autocorrelated structural deficits through time continuous processes (Del Giudice et al., 2013) etc. Beyond using probabilistic description of errors to perform parameter inference for the model, these descriptions have also been utilized to just quantify predictive uncertainty, given a hydrologic model with fixed parameters (Weerts et al., 2011).

Flexibility in the description of errors has been sought for a more representative parameter estimation and a more reliable predictive uncertainty estimation, be it within a formal Bayesian framework or as a post-processor where parametric uncertainties are neglected e.g (Dogulu et al., 2015; Pianosi and Raso, 2012; Wani et al., 2017). However, especially for inference, it has been challenging to suggest a probabilistic description of errors that is effective for all different catchments and all different models (McInerney et al., 2017). One of statistical tools to capture more complex multivariate probability densities in hydrology, be it data or modelling, is copulas (Bárdossy and Hörning, 2016; Sadegh et al., 2017; Salvadori and De Michele, 2004). Especially in the context of uncertainty analysis, copulas have been used to model the residual uncertainty in the post-processing phase of the model (Klein et al., 2016; Liu et al., 2018). The ability of copulas to model dependence between random variables, regardless of the nature of their marginal distributions, gives us the choice to construct the description of marginals and dependence separately. One of the limitations of exploring more representative marginals of errors to construct likelihood functions is the loss of temporal autocorrelation. Copulas can then be used to model such time

series dependence e.g (Borgomeo et al., 2015; Salvadori and De Michele, 2004).

In this paper, we describe and analyze a relatively general formulation of likelihood function for hydrologic models. This formulation employs copulas, which offer much more flexibility, for example by allowing the parameterization of errors separately for low and high flow, while having the possibility to represent autocorrelation and non-negativity. The additive AR1 Gaussian error model can be seen a special case of this general formulation. Using results from simulation experiments, we move on to show the effects of inadequately modeled low flows on parameter inference. We do this using a synthetic data and real data. We also show that the flexibility in the proposed formulation allows the modeler to choose different parameterized marginals for high and low flows and discuss the benefits and shortcoming of this formulation.

2 Method and material

2.1 Error model

If we expect to observe a hydrologic system response Y_o, given input time series x and a model output y_m with parameters θ, the classical description of additive autoregressive Gaussian error B gives us (vectors in bold, and random variables in capital letters):

$$Y_o(x, \theta, \psi) = y_m(x, \theta) + B_M(x, \psi) \tag{1}$$

Where ψ is the error model parameter vector. The conditional probability density of t number of observations, given a certain input and given parameters can be then written as:

$$\boldsymbol{p(y_o|x, \theta, \psi) = \frac{(2\pi)^{-\frac{t}{2}}}{\sqrt{\det(\sum(x, \psi))}} \exp\Big(-\frac{1}{2}[y_o - y_m(x, \theta)]^T \sum(x, \psi)^{-1}[y_o - y_m(x, \theta)]\Big)} \tag{2}$$

Samples from this probability density give us "candidate" time series of observations. We expect the observation to be *like* those samples. And the whole range of the samples constitutes the predictive distribution of the hydrologic system response, for a certain value of inputs and parameters. Then suitable quantiles from this distribution can be chosen to provide uncertainty bands.

One of the short comings of having additive normal noise in the description of $Y_o(x, \theta, \psi)$ is the choice it offers for the marginal densities of errors. Only few such analytical descriptions are capable of capturing dependence in time. Also, as it is additive, one cannot guarantee non-negative model realizations. On the other hand, using gamma or lognormal distribution as marginal of the observations makes the incorporation of the autocorrelation structure difficult. Ammann et al., 2018, address this problem by transforming the real observation space, which does not admit negative values, to a space in which the distribution is expected to be normally distributed and then they model the autocorrelation using the standard Gaussian process. In this study, we used a generalized formulation of an autocorrelated time series, by employing copulas that capture the rank correlation, and make the usage of gamma and similar non-negative distributions feasible as marginals.

An m-dimensional copula is a cumulative probability distribution function defined over a unit cube in some m-dimensional space, such that the marginal density over each dimension is uniform. The idea is that the deterministic model for rainfall-runoff can provide the mean or mode of the probability density of observations at each time step and the temporal correlation between consecutive steps can be captures using suitable copulas. If c(u,v) is the copula density corresponding to the copula C(u,v), Eq. (3) follows from Sklar's theorem, which states that for all multivariate distributions, we can find a copula that captures the dependence structure, in the integral transformed space.

$$\boldsymbol{p(y_o \mid y_m) = c\Big(F_0(y_{o,1} \mid y_{m,1}), ..., F_t(y_{o,t} \mid y_{m,t})\Big) \prod_{i=1}^{t} f_i(y_{o,i} \mid y_{m,i}, \psi)} \tag{3}$$

$f_i()$ in Eq.(3) is any univariate probability density, and $F_i()$ is its corresponding cumulative probability distribution. The copula can be simplified if a Markov property is assumed for the time series. For example with an order one Markov property we can write Eq. (3) as.

$$\boldsymbol{p(y_o \mid y_m) = \prod_{i=2}^{t} c\Big(F_i(y_{o,i} \mid y_{m,i}), F_{i-1}(y_{o,i-1} \mid y_{m,i-1})\Big) \prod_{i=1}^{t} f_i(y_{o,i} \mid y_{m,i}, \psi)} \tag{4}$$

Eq. (4) is effectively an AR1 process, if the marginal are taken as normally distributed and a Gaussian copula is used to model the autocorrelation. In this study Frank copula, which is symmetric for low and high quantiles, is used to capture autocorrelation. Archimedean copulas, like in this case Frank, have few parameters, and therefore do not increase the dimensionality of parameter space by much. Frank copula is defines as:

$$C(u,v) = \frac{1}{\alpha} ln\Big[1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{(e^{-\alpha} - 1)}\Big] \tag{5}$$

And its density function is defined as:

$$c(u,v) = \frac{\partial C(u,v)}{\partial u \partial v} \tag{6}$$

$$c(u,v) = \alpha \frac{(1 - e^{-\alpha})(1 - e^{-\alpha(u+v)})}{((1 - e^{-\alpha}) - (1 - e^{-\alpha u})(1 - e^{-\alpha v}))^2} \tag{7}$$

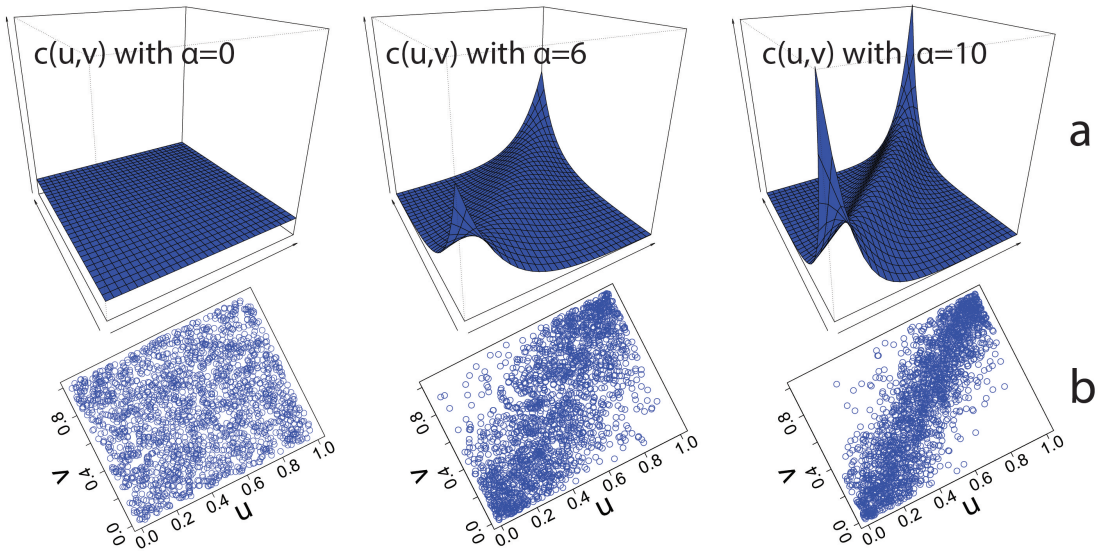


Figure 1: (a) Copula density of Frank copula with different alpha values. (b) Samples from the corresponding copula.

To assure non-negative output y_o as defined by Eq. (4), we use a gamma distribution as marginals ($f_i()$), given by Eq. (8), where $\Gamma()$ is the gamma function.

$$f_i(y) = gamma(y|shape, scale)) = \frac{1}{scale^{shape}\Gamma(shape)} y^{(shape-1)} e^{-(\frac{y}{scale})} \tag{8}$$

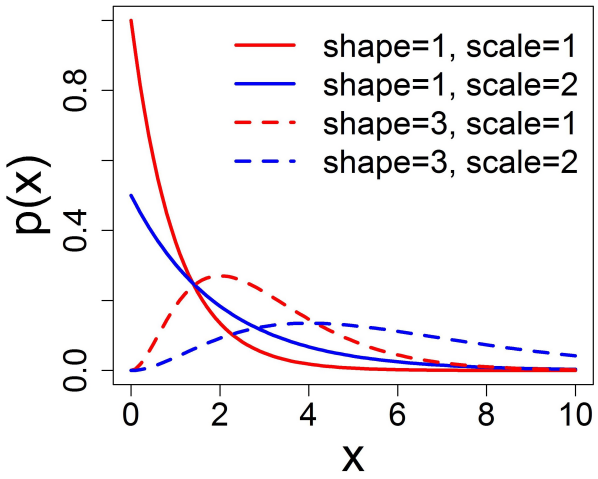


Figure 2: Univariate gamma distribution with different *shape* and *scale* parameter values.

We assume a threshold that separates low flows and high flows called *base* (as a single valued parameter). For low flows, we assume that the model has deficits that cause it to have errors such that it follows exponential distributions. So for this regime of flow we assign a shape factor of 1 and get the scale from the model output. For flows higher than *base*, the assumption is made that the deterministic model output is the mode of the observational distribution (Fig. 2). This differentiates the observation generating process for low and other flows.

In this research, the change in standard deviation (sd) of Y_o as a function of the deterministic model output is incorporated by the explicit equivalent of Box-Cox with $\lambda = 0.5$ (Eq. (5.9)).

$$sd \approx sd.transform \times (y_m)^{1-\lambda} \tag{9}$$

Where $sd.transform$ is the standard deviation of the random variable in the Box-Cox transformed space. This result can be arrived using the truncated approximation of a Taylor series (McInerney et al., 2017). Numerical results from repeated sampling also confirm this (Fig. 3). Eq. (10) defines how the variance should scale in the untransformed scale, if a random variable, E, is normally distributed with a standard deviation $sd.transform$ in the transformed space.

$$\frac{Y_o^\lambda - 1}{\lambda} - \frac{y_m^\lambda - 1}{\lambda} = E \quad gives \quad var[Y_o - y_m] = var[(y_m^\lambda + \lambda E)^{1/\lambda} - y_m] \tag{10}$$

Given this equivalence, we define the standard deviation of Y_o as:

$$sd \approx rate.sd \times (y_m + sd.min)^{0.5} \tag{11}$$

Defining the standard deviation dependence explicitly makes the formulation of priors for the error model more intuitive, compared to the transformed space. Where $rate.sd$ and $sd.min$ are constants to be inferred (The name 'rate.sd' as a name reflects the multiplicative nature of this parameter and 'sd.min' reflects the fact that it will be the minimum standard deviation of the process). Once the gamma is defined this way, we get:

$$Y_o \sim \begin{cases} gamma\big(y|shape = 1, scale = sd\big) & if\ y_m < base \\ gamma\big(y|shape = \big(\frac{y_m}{scale} + 1\big), scale = \big(\frac{\sqrt{4sd^2 + y_m^2}}{2} - y_m\big)\big) & otherwise \end{cases} \tag{12}$$

The definition of *shape* parameter and *scale* parameter for $y_m \geq base$ in Eq. (12) assures that the mode of the distribution is y_m and its standard deviation is sd.

Given the general formulation as described in Eq. (4), we are free to pick and choose different autocorrelation structures (similar to Eq. (6)), heteroscedasticity structures (similar to Eq. (11)) and marginal structures (similar to Eq. (12)). Each of these three components can be employed independently of the other, with its own parameterization.

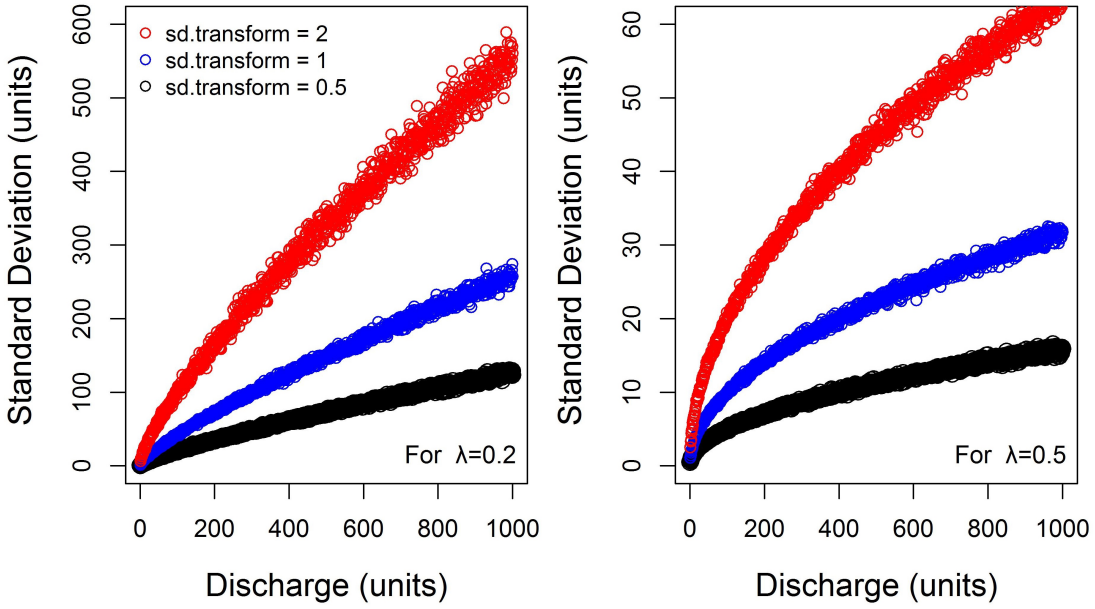


Figure 3: The change in standard deviation of the error in untransformed space, when the error is assumed to be homoscedastic (with a fixed standard deviation = sd.transform) in the Box-Cox transformed space. (Eq. 9 and 10) This graphs helps understand the implicit dependence of standard deviation of the output when using box cox. Using Taylor approximation this dependence is $sd = sd.transform(y_m)^{(1-\lambda)}$.

2.2 Bayesian inference

The expressions in Eqs. (2) and (4) turn into a likelihood function when a given set of observations y'_o are put into them and the parameter values are varied. These expression is not a probability density in the parameter space (θ), but only in the observation space (y_o), hence referred to as a function of (θ). (As likelihood function gets defined uniquely once we define the probability model for Y_o, therefore, in this paper 'copula-based probability model' and 'copula-based likelihood function' refer to the same mathematical description as defined in section 2.1).Once the likelihood function is specified, in combination with the prior, we use Bayes' theorem to get the posterior. An adaptive Markov Chain Monte Carlo (MCMC) sampling scheme is employed to get the parameter samples (Vihola, 2012). The samples represent the updated belief in the parameter values, given the observed data.

$$\boldsymbol{p(\theta, \psi | y'_o, x) = \frac{p(y'_o | x, \theta, \psi) p(\theta, \psi)}{\int p(y'_o | x, \theta, \psi) p(\theta, \psi) d\theta d\psi}} \tag{13}$$

These samples are then run through the probabilistic model to generate the prediction bands.

2.3 Rainfall-runoff model

For analysis using synthetic data, we use a linear model and with input P and (a,b) as model parameters.

$$y_m = aP + b \tag{14}$$

This model gives flexibility and speed to facilitate preliminary analysis and provide proof of concept.

To analyze the ability of this likelihood formulation on real data, we use a simple conceptual model with deficits which do not take infiltration, evapotranspiration and rainfall variability into account. A unit hydrograph convolution is used which defines the input-output relationship between discharge and precipitation. Such models have been used before for capturing discharge relationship of catchments (Betterle et al., 2017). The model assumes an exponentially decaying response to a unit of instantaneous rainfall. Given this formulation, the response of a time series of rainfall can be obtained by integrating the response corresponding to each time slice of the rainfall.

$$y_m(t) = A \int_0^t P(t-\tau) k e^{-k\tau} d\tau \tag{15}$$

Parameter A represents the effective area of the catchment, and k represents the dependence on the rainfall at past time steps. The motivation here is to use fast and simple rainfall-runoff models with overt deficits so that the performance of the error description can be evaluated.

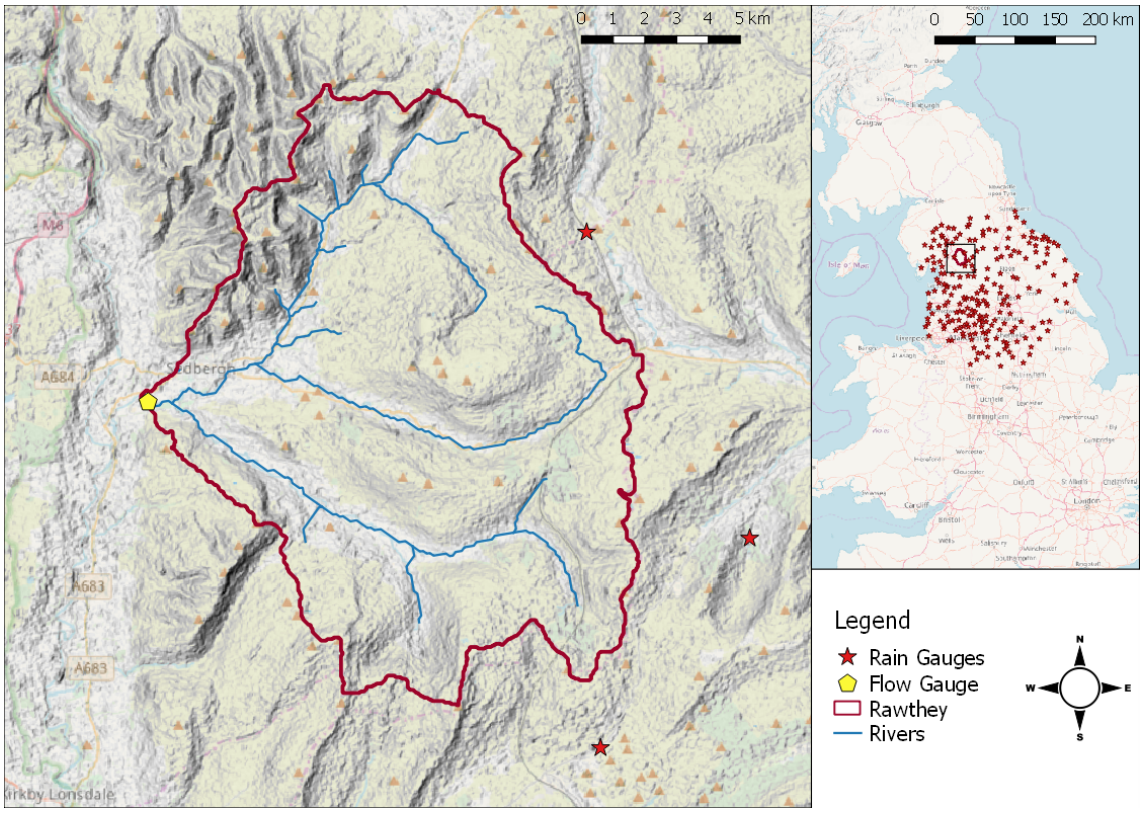


Figure 4: The Rawthey catchment in England.

2.4 Case study

The study area is the catchment of the River Rawthey, North of England, in the Yorkshire Dale National Park, an affluent of the River Lune. It covers 219 sq.km, and collects waters from a terrain that ranges from 675m to 85m of elevation. The catchment is primarily a Natural Park, with very limited human intervention in terms of agriculture and urban land use. The rainfall data used in this work is obtained from a rain gauges from the Environment Office. Data was available from a set of 203 tipping bucket devices with a resolution of 0.2 mm, available at 15 minutes resolution, then accumulated at hourly resolution for two years, from January 2009 to December 2010. The study area, the rain gauges and the flow gauge are represented in Fig. 4. Rainfall time series are obtained with block kriging (Chiles and Delfiner, 2014). Block kriging is a technique that integrates the values obtained through ordinary kriging over an area. Apart from this, we also use a rainfall multiplier, named b, as a parameter for inference. This is done to alleviate the systematic errors due to the fact that none of the rain gauges used to generate the rainfall time series lie within the catchment area.

2.5 Simulation experiments

Parameter estimation for synthetic data:
A) We first show using a didactic example why incorporating autocorrelation and heteroscedasticity is important for unbiased parameter estimation and predictive uncertainty estimation. We do this using synthetic data generated from a linear model (Eq. (14)), feeding it with a sinusoidal rainfall. The error is assumed to be heteroscedastic, following Eq.(11), and autocorrelated (AR1), with a correlation parameter rho. Inference is done using flat priors. The parameter values for $(a, b, rate.sd, min.sd, rho)$ used to produce data are (1,5,2.5,2,0.9). Inference is done using P as, $30sin(x)+30$, with x from 0 to 100, spaced at 0.1. Verification time series is generated as a second sample using same probability model specification.

B) Preliminary proof of concept: To facilitate reproducibility, we again do parameter inference using the new copula-based formulation of likelihood and synthetic data generated through same linear model. Inference is done using flat priors. Eqs. (11) and (14) is used to generate synthetic data. We use two precipitation time series

$$P.high = 30sin(x) + 30, \quad P.low = 10sin(x) + 10 \tag{16}$$

A sequence of x from 0 to 100, spaced at 0.1, is used to generate P at discreet time steps. We then combine first 220 points from P.low and the other 780 from P.high to get the precipitation for the calibration time series. And combine first 420 points from P.low and the other 580 from P.high to get the precipitation for the validation time series. The parameters (a,b) used for the high flow are (1,5) and for the low flow (0.1,5). This gives us observation generated by two separate processes. rate.sd and *rho* is 0.3 and 0.9 respectively.

Parameter estimation for real case study: For the case study we use two time series of observation for calibration and validation. As there are substantial sections of the time series that are low flows, we see the effect of such flows on the inference of the parameters and on the predictive uncertainty. The inference is done using two priors. Informative priors and wide priors. We use an hourly time series of discharge and precipitation. We calibrate the model on 2000 data points, from 01.01.2009 00:00 to 25.03.2009 07:00, and then validate the results for 3000 data points, from 08.09.2009 00:00 to 10.01.2010 23:00. Both the calibration and validation time series contain sections of high and low flows. As the model used is in this study has substantial deficits, we expect the prediction intervals to be wide. The model has two inference parameters A and k (Eq. (15)). The priors used for the inference are mentioned are presented in table 1. And the ratio high and low flows in the calibration and validation time series is presented in Table 2.

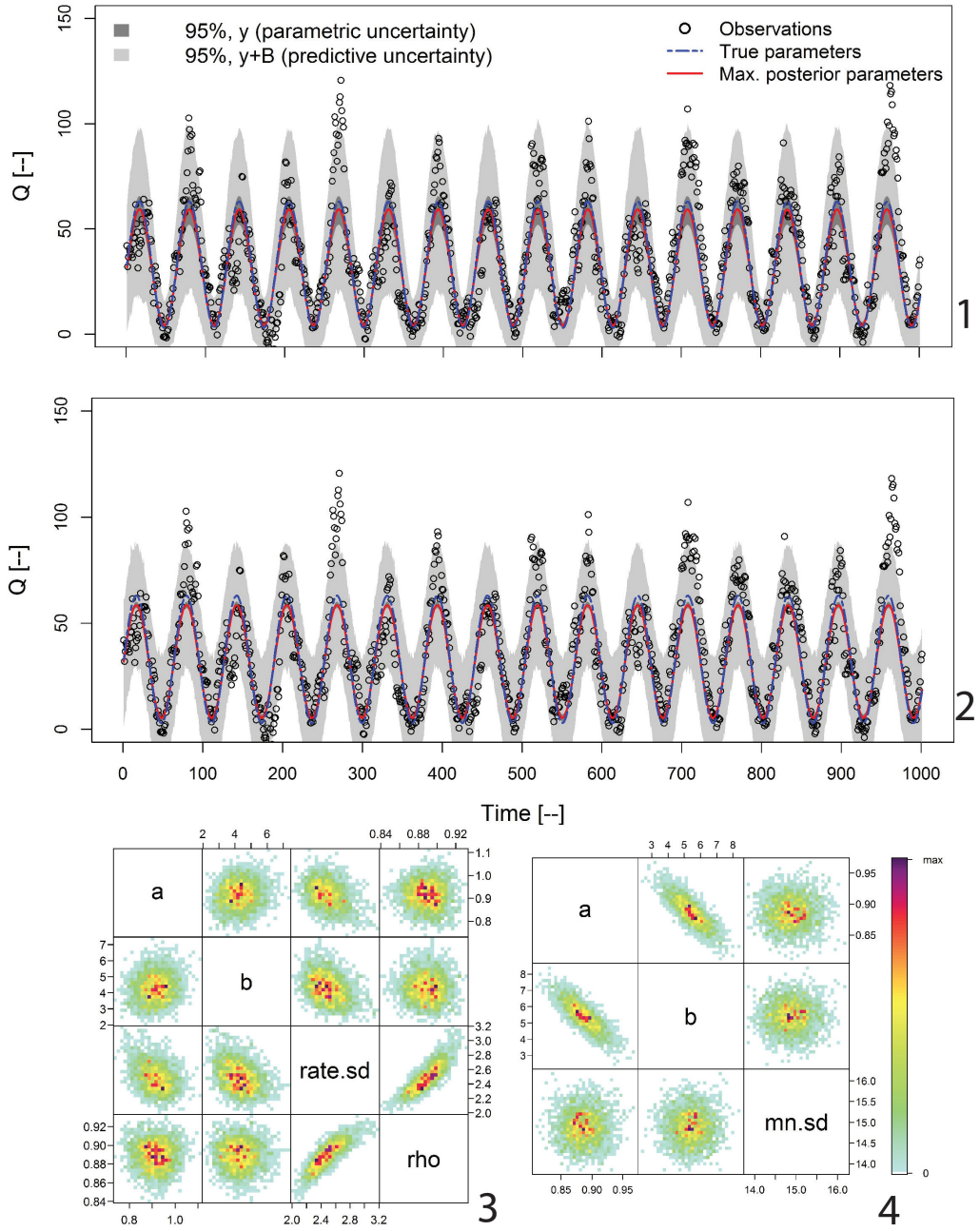


Figure 5: Effect of various assumptions regarding the likelihood on the inference of linear model parameters when the error generating process is alike for low and high flows. 1) Predictive uncertainty estimated using correlated and heteroscedastic (True) Likelihood 2) Predictive uncertainty estimated using uncorrelated and homoscedastic (i.i.d) Likelihood. 3) and 4) are the corresponding posteriors.

3 Results

As it is evident from Fig. 5, when data that is generated from an autocorrelated and heteroscedastic process is used to infer parameters for a linear model, the inference results in underestimation of parametric (given by the spread in the posterior) and predictive uncertainties (given by the width

Table 1: Normal truncated distributions were used to define the priors for the parameters in simulation experiments for the case study. The vector in the table give the values for mean, standard deviation, lower limit and upper limit of the prior for each parameter in our inference.

For copula-based likelihood function

	mn.sd	*rate.sd*	*Alpha*(α)	*b*	*k*	*base*
(5,2,0,30)	(0.4,0.3,0,5)	(5,10 ,0,50)	(16000,10000, 12000,100000)	(1.5,1, 1,5)	(0.2,0.5, 0.08,3)	(50,3, 40,60)

For AR1 likelihood function

	mn.sd	*rate.sd*	*rho*	*b*	*k*	*base*
(5,2,0,30)	(0.4,0.3,0,5)	(0.9,0.3,0,1)	(16000,10000, 12000,100000)	(1.5,1, 1,5)	(0.2,0.5, 0.08,3)	

Table 2: High and low flows in calibration and validation time series.

Calibration			Validation		
Highest flow (m^3/s)	No. of observations below 30 m^3/s	Number of observations above 150 m^3/s	Highest flow (m^3/s)	No. of observations below 30 m^3/s	No. of observations above 150 m^3/s
259	1901	12	279	2607	14

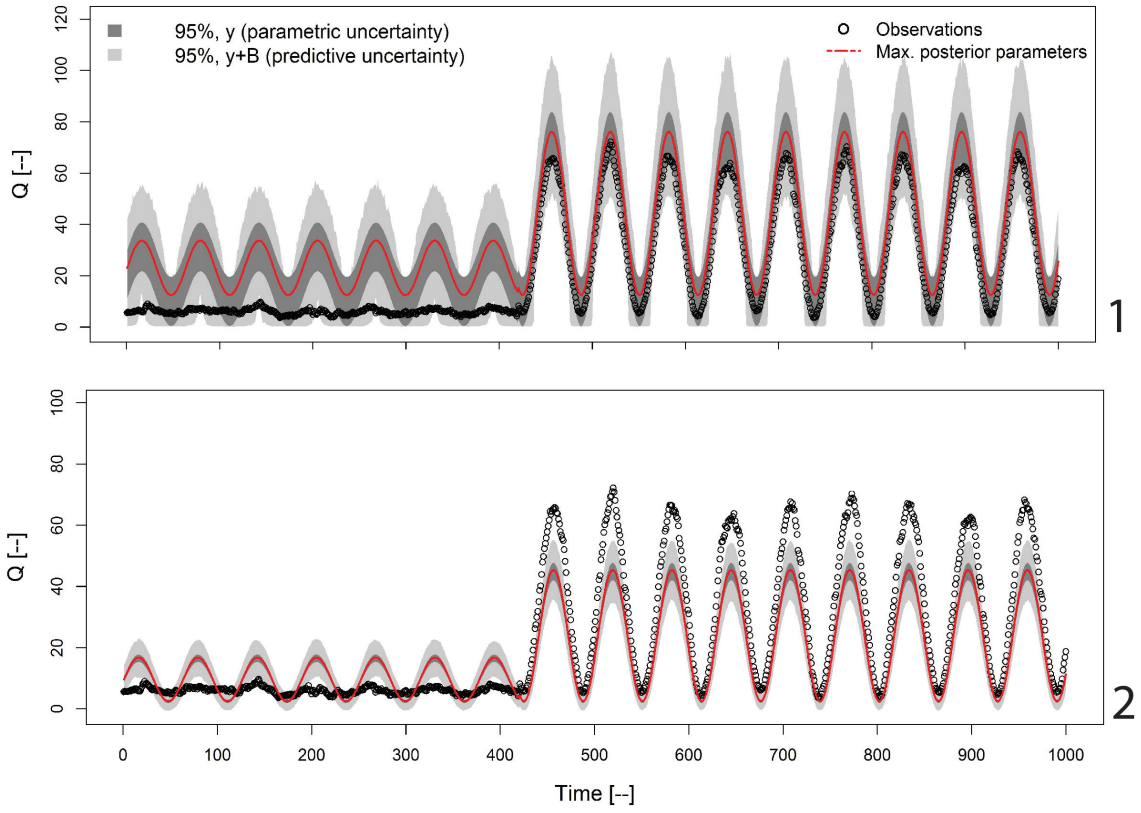


Figure 6: Model deficits in low flows inducing biased parameter inference can be avoided by using the flexibility of copula based likelihood during inference (synthetic data). 1) Prediction using the copula based probability model. 2) using AR1 probability model.

of prediction intervals for high flows). Also, in the predictive phase the high and low flows do not get captured adequately. For example in this case, the prediction intervals using independent, identically distributed (i.i.d) model are narrower for high flows and wider for lower flows, compared

to actual distribution of observation in these flow regimes. Also, the posteriors of parameters from i.i.d are narrower than the posterior from autocorrelated model. This didactic example establishes the operational need to capture the autocorrelation and heteroscedasticity of a hydrologic time series.

Apart from heteroscedasticity and autocorrelation in the hydrologic time series, errors in high and low flows can be generated by different deficits in the model, and therefore it may not be reasonable to treat them with the same marginal error distribution. Fig. 6 depicts the inference of the linear model which systematically overestimates low flows. This bias in the model causes severe biases in inference if we use the AR1 processes. The model parameters tend to underestimate flows during high rainfall, so that they can fit better the base flows. One way to overcome this problem is use different nature of marginal distribution, which can capture such a tendency of the modelling errors. The error description, as suggested in Eq. (12), is more tolerant towards the biases in the low flows and thus allows the model to fit for high flows Also, we see that the non-negativity of the flow is avoided.

Similar underestimation of high flows is seen for the real case study. Whereas copulas-based likelihood description is able to save the inference procedure from the underestimation of high flows (Fig. 7). We use the maximum posterior density parameters to generate the plots. Given we anticipate that model does not capture the infiltration process well, and may overestimate runoff from small rainfall events, we put this understanding in Eq. (12), where low flow predictions are more likely to overestimate flow. This makes the probability model lenient towards such errors in low flows, giving the inference procedure the flexibility to search for parameters that take the prediction closer towards the high flow observations. Many other formulation, different than Eq. (12), can be used to incorporate more complex error structures.

From the analysis of the real case study, we find that around 62% of observational time steps for the validation time series have prediction intervals going into negative flows. No such negative flows are seen using the copula-based likelihood function description. Also, for 95% prediction interval, the mean width of the copula-based likelihood is higher, and it thus captures more observations within it. AR1 captures 81.7% and copula-based probability model 92.1% observation in the validation phase. For the high flow regime (flows higher that 150 m3/s), the relative mean error of model prediction corresponding to maximum posterior parameter values is 33.2% for AR1 and 14.2% for copula-based likelihood function.

4 Discussion

The results of the simulation experiments demonstrate that, akin to the accuracy of the deterministic hydrologic model, our description of the errors should also be representative of the modelled process. A departure from this can lead to non-trivial biases in the parameter estimation (Fig. 5 and 6). We also show that the disproportionate presence of low and high flows affects how the parameter inference would perform (Fig. 6 and 7). Manually choosing only high events from the past time series to calibrate the model is not always a desirable alternative. We may also be interested in the prediction of low flows using the same model. Then the error description needs to be flexible enough to capture both high and low flows adequately. There can be cases where the errors in the low flows are skewed, with the model having some systematic tendencies to depart from the observations. It is intuitive that the tendency of the model to overestimate and underestimate would not be symmetric for flows closer to zero. The errors have to truncate such that the real flow is always positive. In order to prevent such complex errors undesirably influencing the whole inference procedure, we can describe these errors with heavy-tail or exponential distributions (Eq. (12)). The results show that AR1, due to its inflexibility with the autocorrelation is not able to capture non-negative support and skewed distributions simultaneously. We use copulas to capture this and see a noticeable improvement in inference. Copulas allow us to capture the temporal

dependence and we are free to choose the marginals. However, we also found that this flexibility, while in principle desirable, in practice still does not guarantee unbiased parameters. As we can see from Fig. 7 (3), the alpha ($< alpha$) parameter tends to go towards the upper extreme value, and is only bound by the hard prior. While this is not the case for the synthetic case studies, where the error structures are known, real case studies have more complex errors and these interactions between error model parameters and deterministic model parameters become more severe. So this description of error, as defined in Eq. (4), just makes an unbiased inference potentially achievable, if we are able to define the adequately representative marginal distributions using this flexibility, and have informative priors, for example in the case of Fig. 6. More concrete guidelines to choose such representative marginals still need to be researched. And it is not possible to formulate a unique description of errors that is representative of all the hydrologic time series.

We find that copula-based likelihood functions prefer a high correlation value,as such parameter values produce peaky densities at the edges of the unit cube (Fig. 1). These edges correspond to high and low quantiles. In principle, high copula density values at these edges should be compensated by corresponding low values of marginal densities, hence avoiding parameter biases. As the likelihood is a product of copula densities and the marginal densities (Eq. (4)), if copula densities have high value for extreme correlations, the low values of marginal should bring down the product and we should still have low likelihood values. However, this is not always achieved - the observations are not always described adequately by the selected probability model, so the compensation of the marginal is not guaranteed to suffice in avoiding very high parameters values of copulas. In many cases the posterior can converge to a high autocorrelation, at the expense of taking the deterministic model away from the observational data. One way to avoid this is by having informative priors where we know for sure that the autocorrelation of the process cannot be more than a certain amount and the standard deviation of the error cannot be more than a certain amount, representing the model understanding and the local hydrology of that particular case study. Nonetheless, more robust schemes which allow the model to go closer to the observations even when the priors are uninformative need to be formulated. The desirable formulation an error model would be such that small inaccuracies in the error models only lead to small inaccuracies in inference. Priors, however, can be used to exclude unreasonable parameter spaces.

As far as the improved performance of copula-based likelihood function over the AR1 is concerned, one of the reasons why this is seen in both synthetic and the real case studies of this work is that the peaky low flows are poorly represented by the chosen hydrologic models. The rainfall produces a flow signal in the model which is not seen in the real catchment. This effect is a combination of several sub-effects. First, the rainfall itself is measured by rain gauges that do not lie within the catchment (Fig. 4). The biases in the precipitation input, with overestimation and underestimation of rainfall induce high relative errors in the low flows. Also, in the absence of an infiltration mechanism, all the rainfall is converted into surface runoff by the model. These deficits lead to peaky errors in low flows every once in a while. As these errors will have very low probability to occur, in case of an AR1 process, they get penalized heavily, distorting the parameter inference. Hence, it is recommended, before choosing the marginal of this coupla-based probability model, to ascertain the regions of the hydrograph where high errors can be tolerated and where relatively better predictive capacity of the model is desired. This way during inference, the likelihood function will not produce low probability values for parameters that perform poorly only in low flows but perform adequately well for other sections of the hydrograph.

In case of systematic model deficits, for example Fig. 6, the AR1 likelihood function tends to either assume that the observations have been produced by a process with high standard deviation of errors and high autocorrelation, instead of a process that has relatively low autocorrelation. To mitigate the effect of such interaction between the parameters, we assume that low and the high flows are generated by different marginal, and demonstrate that such an assumption improves parameter estimation. Such flexibility allows the error model to ascribe a different variance to the base flows, as opposed to the high flows, when there is rainfall (Fig. 7).

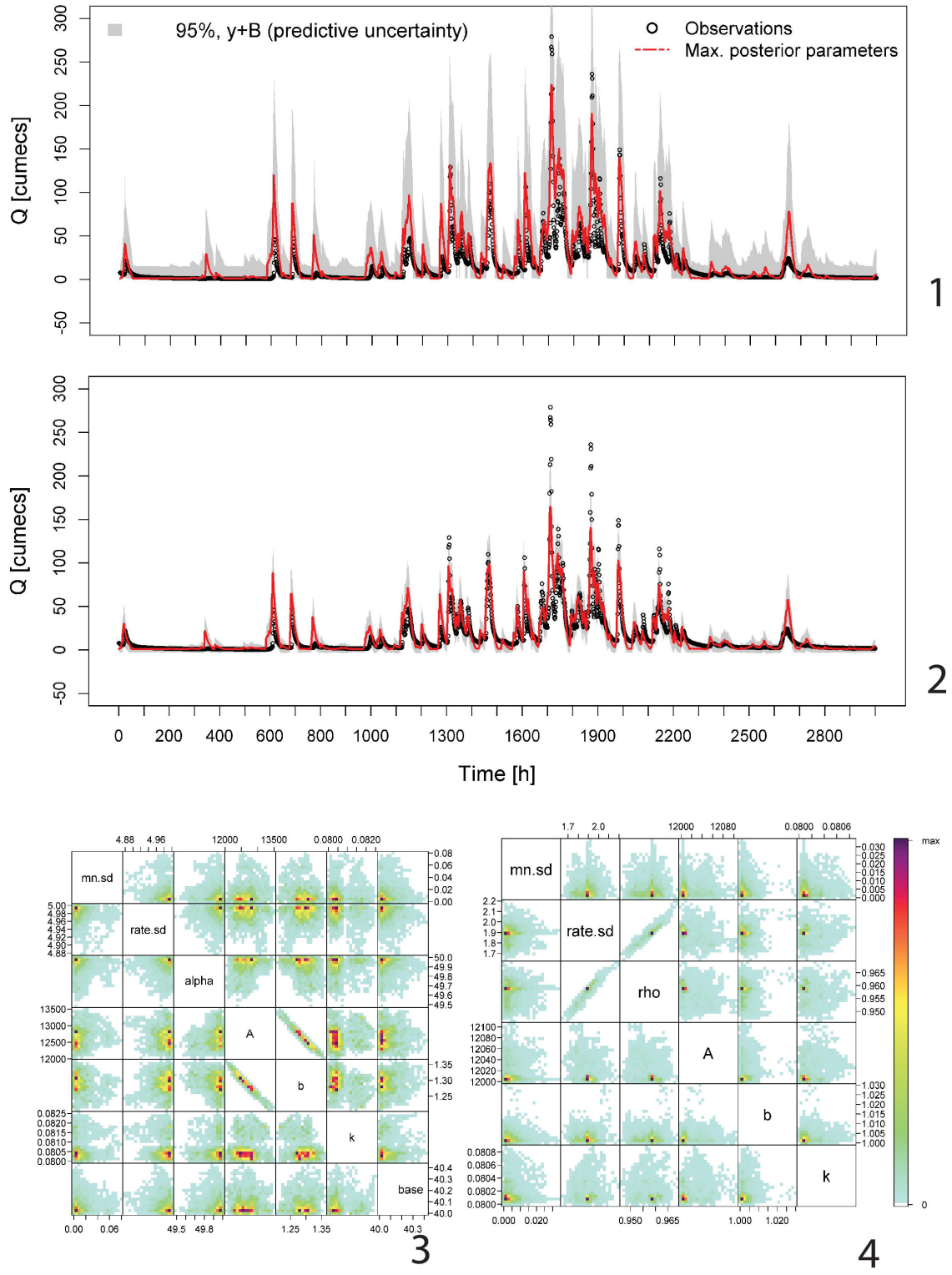


Figure 7: Comparing inference using copula based likelihood and AR 1 likelihood. 1) and 2) are the predictions. And 3) and 4) are the bivariate posteriors for copula based likelihood and AR1 respectively.

We also find that non-negativity in the predictions is easily achieved using the formulation of marginals used in this work. Even though there are substantial deficits in the low flows, we know that the observations cannot be negative. Therefore, the errors are generally from a skewed distribution. The use of shape parameter as 1 in our case reflects that the error tend to be skewed in low flows, and the model, as in our case, is more likely to overestimate the flow than to underestimate it.

Regarding the robustness of this error model formulation, even though this description is flexible, and can be used to formulate more sophisticated error models, with different marginals, and still be able to capture the autocorrelation, the flexibility is not a guarantee for convergence to meaningful posteriors. The description of the error model still needs to incorporate the relevant knowledge of the modeler about the kind of deficits they expect in their models. Also, the method can perform poorly in cases where the autocorrelation parameter and the heteroscedasticity parameters do not have proper prior constraints. There is no guarantee that a small approximation in the choice of copula or marginal will not diverge the inference from parameter spaces that actually reflect the hydrology of the modeled system. To attain this robustness in inference using copula-based likelihood function, more research is needed.

5 Conclusions

Biases in parameter and predictive uncertainty estimation arise from poorly modelled errors in hydrologic time series. We show that the parameter inference can be skewed in many situations if we always stick to the state-of-the-art AR1 process as the error model. To curtail that we propose a flexible formulation of the likelihood function that allows for different choices of marginal for the time series and takes care of the autocorrelation using copulas. From the results and discussion we conclude:

1. Capturing both non-negativity of the hydrologic system response and the autocorrelation in time is not straight forward.
2. We can use copulas with non-negative marginals to define a probability model for hydrologic time series.
3. Given Sklar's theorem, this description allows for more flexibility in choosing the marginals. For example, we can have different error structures for low and high flows.
4. If the right marginal densities are chosen, we show that this likelihood formulation provides better parameter and predictive estimates, where additive error models would otherwise perform poorly.
5. However, if the copula likelihood function is not fully reflecting the underlying data generating process, for example, by choosing wrong marginals or wrong heteroscedastic dependence, or having uninformative priors, we again run into issues of non-robustness during inference.
6. Just like in the case of AR1 probability model, the posterior can converge to unrealistic parameter spaces.

This research extends the suite of mathematical descriptions available in hydrology to model time series. Given that such a description is flexible, we foresee that it can be useful for many problems related to parameter inference and model prediction. The analysis is intended to add to the body of literature on representative likelihood functions. It brings forth some benefits and challenges of using a copula-based likelihood function.

Acknowledgement: We thank Lorenz Ammann for many fruitful discussions. We thank Miguel Rico-Ramirez for his support during the stay at the University of Bristol. Part of this study was supported by the EC FP7 project QUICS (Quantifying Uncertainty in Integrated Catchment Studies), grant agreement no. 607000.

Chapter 6

A philosophical appraisal of hydrologic uncertainty

Author contributions: Omar Wani wrote the full chapter.

Given the unique trajectory and assimilation of uncertainty analysis in the field of hydrology, there are community-specific expectations about what should such an analysis cater to. One may call it a scientific (sub)paradigm that reflects a somewhat aggregate attitude towards and a somewhat aggregate interpretation of uncertainty analysis within the hydrologic community. Having this in mind, the motivation of this essay is, first, to provide a didactic axiomatic sketch of probability theory and its usage for uncertainty quantification, with hydrologic modeling as a working example. After the theoretical considerations, a brief overview of the current expectations of uncertainty analysis in the hydrologic community is presented. This chapter contains speculative and subjective aspects, which should be read as such.

1 Probability theory and uncertainty quantification

Axioms: Probability theory, devoid of any interpretation, can be seen as just a collection of axioms, and the resulting theorems, about a mapping between two sets. The mapping goes from a set F, called the event space, into another set P, which contains elements from the set of real numbers. The event space in turn is the set of all subsets that can be formed using the elements of another set called the sample space, denoted by Ω. The sample space can for example contain finite categorical elements like letters of the alphabet, countably infinite elements like integers, or uncountably infinite elements like the real numbers. This gives us three distinct sets (Ω,F,P), collectively called the probability space. The definitions and axioms in probability theory put special constraints on these three sets and the mapping F$\rightarrow$P. This mapping assigns an element in P to every element in F. The usual notation of the mapping is P(ω), which signifies the element in set P corresponding to the element ω in F. The three probability axioms, as formulated by Kolmogorov, can then be stated as:

$$P(\omega) \geq 0 \qquad \forall \omega \in F \tag{1}$$

(called the axiom of non-negativity)

$$P(\Omega) = 1 \tag{2}$$

(called the axiom of unitarity)

$$P\big(\bigcup_i \omega_i\big) = \sum_i P(\omega_i) \qquad \omega_i \cap \omega_j = \phi \qquad \forall i, j \tag{3}$$

(called the axiom of σ-additivity)

Given these axioms, the total probability mass of 1 gets divided over all the elements of the sample space. Also, using these axioms, a whole suite of corollaries and theorems can be derived (Terenin and Draper, 2015). So far nothing has been said that makes probability theory a fitting mathematical framework for uncertainty. It is just a collection of deductive statements about some abstract mathematical object. However, given its deductive nature, once we accept the axioms, probability calculus is the only consistent way to evaluate changes within this framework (Nearing et al., 2016).

Adding interpretation to probabilities: Probability theory rendered as a complete mathematical abstraction is one of the modern developments and was done to formalize it. It is, however, quite ahistorical, as the earlier developments of probability theory were all motivated by the need to explain real world problems, e.g. games of chance (Mlodinow Leonard, 2008).

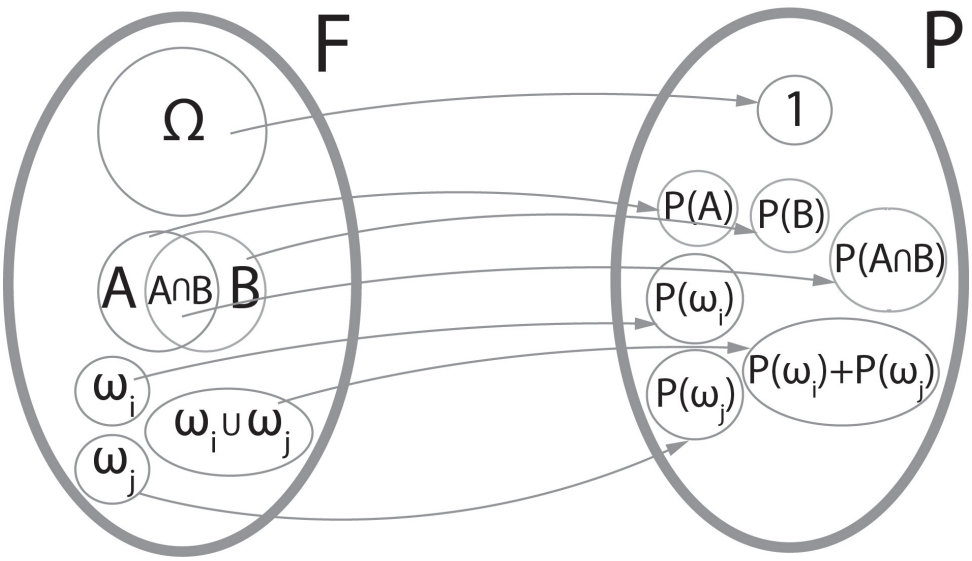


Figure 1: Schematic for the mapping between the set F and the set P.

Kolmogorov also presented the justification of his axioms when probability is viewed as a limit of the relative frequency of an event. This lays the groundwork for the frequentist perspective. One of the Bayesian interpretations of probability, as objective belief, is attributed to Cox. Cox, and then Jaynes (Cox, 1946; Jaynes, 2003), extend the reasoning of Boolean algebra, where propositions are either true or false, to cases where there is uncertainty about the truth-value of such propositions. Given the formulation of Cox and Jaynes, any proposition is still either true or false but we don't have complete information which one it is. It has been shown that if the plausibility of a proposition is defined such that it:

1. Is a scalar
2. Reflects common sense (i.e (a) if the belief in the truth of a proposition increases, its associated plausibility increases and (b) the plausibility of two propositions, A and B, being true should be a function of the plausibility that A is true given B is true and the plausibility that B is true.)
3. Is consistent with Boolean algebra

then Kolmogorov's axioms (apart from some formal caveats) can be derived from this logical system. This, called Cox's Theorem, shows that true-false logic under uncertainty is isomorphic to probability theory (Terenin and Draper, 2015). Given 2 (b) in Cox's postulates and the property of associativity from Boolean algebra, conditional probability, $P(A|B)$ (i.e. probability of A given B), is defined as

$$P(A \cap B) = P(B \cap A) = P(A|B)P(B) = P(B|A)P(A) \qquad \forall A, B \in F \tag{4}$$

Equation 4 can be rewritten to get the famous Bayes' theorem:

$$P(A|B) = P(A \cap B)/P(B) = P(B|A)P(A)/P(B) \tag{5}$$

2 Relevance of Bayesian probability in hydrology

In case of hydrology, we may be interested in hypothesis testing, perhaps in the form of two competing model structures, or parameter inference once a model structure is decided. Given the

understanding of the model, and given the data, we would like to know where do our parameter values lie? While sensitivity analysis of parameters has been a part of hydrologic modelling for decades now, it only gives some quantitative understanding of the robustness of model results, if the parameter values are tweaked. However, the use of highly parameterized models to make design and forecast decisions, when one is not sure about the parameter values used, confronted the modeller with the challenge of uncertainty. The next logical step by the modelling community was to use the entire range of parameters and produce the entire range of model outputs. Ensemble outputs are produced this way. This helped capture the non-linear interactions of the model parameters in the model for forward simulations. Apart from this, once the observations of the system response increase, we would like this information to be reflected in parameter ranges that are admitted in Monte Carlo simulations. This updating procedure, as mentioned earlier in the axiomatic basis of probability theory, can only happen consistently using Bayes' theorem, given probability density of parameters represents our beliefs.

Unless we are willing to reformulate the axiomatic structure of the theory or report plausibility by metrics other than probability, formal Bayesian updating cannot be avoided (Nearing et al., 2016). Therefore, when there were efforts in hydrologic community to capture the uncertainty of parameters using informal likelihood functions, formally speaking, such evaluations did not yield posterior parameter distributions, $p(\theta|y_{observed}, x_{input})$, or the final predictive distributions $p(y_{future}|y_{observed}, x_{input})$, where y represents the system response and θ the model parameters (Mantovan and Todini, 2006; Stedinger et al., 2008). However, it has been pointed out that the probability models used to describe hydrologic modelling errors are often simplistic and many a times could lead to inadequate description of uncertainty, and, notwithstanding the formality of Bayesian inference, lead to biased parameter estimation and biased model prediction. While this is certainly the case, it does not qualify as an argument against Bayesian updating per se. As an analogy, the deficit in the simplistic differential equations used to model hydrologic time series does not qualify as an argument against calculus. Both Bayesian updating and calculus are based on deductive results, given some axioms. Either we replace the axioms with others that may be more representative of the reality we are trying to model, or we have to accept the theorems that ensue from the default axiomatic structure.

3 Expectations in hydrologic community from uncertainty analysis

One of the recurrent expectations from the uncertainty bands is that the observations should lie within them with a certain frequency. However, by the very definition of Bayesian probability, that is neither guaranteed nor even expected. What we are after is our updated belief in parameter values given the data, and our belief in how the system will respond in the future given the data (the conditional predictive distribution). The only case where that would correspond to bands that capture observations with a certain frequency is if we have converged to a dirac posterior in parameter space and the remaining uncertainties are aleatoric i.e random.

Another expectation from an inference scheme is that it should penalize the parameter values proportionate to the errors they produce in the model output. However, the convergence of the posterior to a certain (premeditatively) *desirable* parameter space is contingent on the choice of the likelihood function. Also, if the data that we throw at the inference problem is not generated by the process we expect it to be generated from, then we don't get, for all practical purposes, any meaningful posteriors. What we even lose is the pragmatic robustness of a least squared error (LSE) estimate. While the least squares would guarantee to take the model towards the data, no such guarantees exist if the probability model is complicated and flexible enough that it produces high probabilities for regions of parameter space that are not even close to the true parameter values. So the danger is that small approximations in the likelihood functions may not always

lead to small approximations in meaningful posteriors. Thus, if the likelihood function is chosen carefully, we have the great benefit of assigning probabilities to parameter space, and not just accept the best guess value of the parameter vector. However, in cases of a poorly representative likelihood function even the consolations of having a predictive model, which functions just well, can be taken away. And then one may ask, what is the point of the added complication anyway, what Klemeš calls mathematical prestidigitations (Klemeš, 2000). To avoid this (1) we can restrict unreasonable parameter spaces using priors, (2) not use observations during inference from sections of time series that may not be representative of the error model or (3) devise likelihood functions that are robust and only approximately digress from the sought solution, when there are meta uncertainties.

The critiques against Bayesian updating are therefore implicit critiques against the simplistic probabilistic description of errors. It seems that there is no unique description of hydrologic errors that can be used off the shelf. While we put the system understanding in formulating the deterministic model for hydrologic systems, we also have to invoke our understanding to formulate a representative description of the modeling errors. And just like *adequate* deterministic models vary from case study to case study, *adequate* error models vary too.

Post script: This essay should not be misconstrued as a claim on the superiority of Bayesian techniques in hydrology. Depending on the task of the modeling exercise, Bayesian techniques might not be needed or may even prove to be inefficient. For example, if a modeler is interested in having a predictive model, has large amounts of data, and does not plan to extrapolate beyond the past observations, fitting an artificial neural network may perform very well for the task. Additionally, if the modeler wants to report the results by assigning probabilities to the model output, then he or she needs to define a probability model, which is possible using post-processing techniques. It is only when he/she wants to assign probabilities to elements of his/her model that do not have a frequentist interpretation, like assigning probabilities to parameters, he/she naturally enters the domain of Bayesian statistics. Bayesian statistics is necessitated when one wishes to talk about beliefs as probabilities and update those beliefs as new data comes in. If one wishes not to use Bayesian statistics, then, as explained in this chapter, epistemic uncertainties need to be reported with a mathematical framework other than probability theory. If one wants to work with the latter for these uncertainties, one has to employ the former.

Chapter 7

Concluding remarks

Author contributions: Omar Wani wrote the full chapter.

Reporting hydrologic model outputs along with the associated uncertainties allows for risk-based decision making, where the whole range of possibilities is considered. Also, this paradigm of hydrologic modelling, where model outputs are assigned probabilities, allows for statistical hypothesis testing. Such a paradigm works with degrees of belief in different hypotheses, updating the belief once more data becomes available. Akin to the proper formulation of the deterministic elements of hydrologic models, research is needed on the stochastic elements of these models so that all the representative properties of these elements get reproduced by the model. Additionally, for any meaningful application of hydrologic models, the associated uncertainties should not be so large that no information can be extracted from model outputs. In this thesis several methods were proposed to both quantify and reduce uncertainties in hydrologic modelling, thus, overall, leading to improved predictions. The findings of this thesis can be summarized as follows:

1 For uncertainty estimation using kNN resampling

1. For models that have been calibrated in the deterministic paradigm or the parameters have been manually set, prediction intervals can be generated in the post-processing phase. A simple, computationally inexpensive method that uses instance-based learning can be used to provide reliable estimates of residual uncertainty. The method relies on the established k nearest neighbour search algorithm. It looks for hydrometeorological conditions similar to the model output in the past and learns from the errors made by the model in those instances.

2. This method does not employ parametric distributions and relies on empirical distributions. This allows for flexibility in the face of different error structures. The technique is also able to capture more systematic errors and correct the deterministic model for such deviations. We show that the method produces reliable prediction intervals, and the heteroscedasticity of errors is reflected in the intervals. Furthermore, this technique performs well for both single and multiple output models.

3. The method is sensitive to the choice of search space and the value of k. While we have shown that the method works well, even when the simplest search space is chosen and the value of k is calibrated, further research is needed to establish guidelines for optimal selection of the search space and the value of k.

2 For parameter estimation using censored and binary observations

1. Substantial research has gone into the formulation of likelihood functions that capture the error properties of hydrologic models. Nevertheless, these likelihood formulations are only used when uncensored observational data is available for parameter inference. We extended the description of a likelihood function so that censored observations can be used to estimate parameters. We showed its application on a synthetic and a real case study, where a rainfall-runoff model was calibrated using binary signals, which can be considered as an extreme form of censoring.

2. The method allows the usage of censored observations which otherwise would have been discarded. The inference is done in the Bayesian paradigm, so the technique is amenable to sequential updating. As longer time series of binary data becomes available, the parameter uncertainty can be further reduced. Given this framework, the observational data from unconventional sources, like CCTV cameras, binary sensors etc, can be used to update the parameter values. The method allows to disentangle the contribution of uncertainty from the parameters and from the model structure. Given this method, accurate measurement instruments can be supplemented with cheap and robust binary sensors to attain an optimal combination of observations for efficient model calibration.

3. We were able to show that the inference improves the model prediction in cases where the parameter priors produce poor model outputs. The added utility of binary observations in improving a model which already has an acceptable performance has not been researched yet. Besides, in cases where the formulation of the probabilistic model that is supposed to generate the observations is incorrect, the nature of biases that inference can impose on parameters need to be researched further. For instance, the autocorrelation parameter interacts heavily with the deterministic model parameters. For exploratory analysis we have constrained the autocorrelation parameters in the prior, however, in cases where we do not have any prior understanding of the autocorrelation of the error process, this might lead to biases in the parameter estimation.

3 For non-negative heteroscedastic autocorrelated process using copulas

1. The probabilistic description of the error in hydrologic models has generally been additive. Apart from that, the heteroscedasticity is modelled either by assuming a dependence explicitly on input/output or first stabilizing the error variance through transformations. To move from error generating process towards observation generating process, we investigated the use of marginal distributions of the observations that do not result in negative flows. The temporal dependence was captured by using copulas.

2. We demonstrate that the flexibility of this description can be used to capture complex error structures, leading to better predictions. However, we also find that there is a strong interaction between the marginal distribution and hydrologic model parameters. Results show that copula parameters have a tendency to go to extreme values during the inference process.

3. The challenge to make the inference of the copula parameters robust still remains open. In hydrographs with long sections of flows that do not fit the probabilitic description of our model, and in the absence of informative priors, the likelihood function would try to evaluate high probabilities based on the fit for that section, in many cases leading to incorrect model predictions for other more interesting sections of the flow regime. Therefore, robust inference using the flexibility of a copula-based likelihood function needs further scientific investigation.

Outlook and recommendations for future research

The work presented in this thesis is intended to become an enabling factor for other questions posed in hydrology, and facilitate the research with better calibrated models and reliable uncertainty estimates. It is important to have the big picture in sight, and ask what purpose does uncertainty analysis serve. In the context of scientific inquiry, the interesting questions about hydrologic systems per se, beyond the modelling exercise itself, need to be posed and answers sought. However, a proper statistical treatment of uncertainties puts these answers on a firmer ground.

For such a proper treatment, many open questions and challenges remain. The probabilistic models, that attempt to capture the uncertainty in deterministic models, are themselves approximate. This introduces meta-uncertainties in model predictions. There need to be further investigations into the nature of hydrologic modelling errors. The likelihood functions need to be robust in the sense that if there are small errors in the description of the likelihood function, they should only lead to small errors in the inferred parameters. Highly biased parameter estimation should be prevented, like the one resulting from the interactions of deterministic model and error model elements during inference. Also, while we deal with the censored data of a specific kind in this thesis, more generic likelihood functions, that can capture other kinds of censored data need to be formulated and analyzed. As an example, CCTV cameras not only capture the inundation during floods, they also provide information on whether the water depth is increasing, decreasing or stagnant. Such

observations should contribute to parameter inference. In the context of data-driven hydrologic modelling, research is needed to enhance the spatial accuracy of models by utilizing censored data. Distributed hydrologic models are mostly calibrated on observational data gathered from one of the downstream outlets of the entire network. As it is not possible to decompose the contributions to discharge from various subsections of the network, it leads to identifiability problems during parameter inference. In many cases, the parameters of the distributed models have a physical interpretation. Inference using only spatially limited data can lead to unrealistic or simply incorrect values of these physical realities. To avoid the model predicting the *right* output for the wrong reasons, i.e with biased parameters, further research is needed to quantify the benefits of sampling campaigns that collect spatially distributed data. A detailed investigation into understanding the nature of biases that a single observation source can induce during calibration will have high scientific value. Besides, the added utility of such spatially-explicit censored data to the calibration exercise needs to be quantified in the form of some easily interpretable metrics.

As suggested by the overarching theme of this work as well as the specific techniques presented here, we believe that hydrologic models benefit from the employment of the probabilistic paradigm. Notwithstanding the diversity of the cases—be they, for example, from catchment-scale hydrology, operational flood forecasting or urban water management—availability of observational data should enable us to improve our models. Additionally, we should have a quantitative awareness of the uncertainties involved. We hope that this work is a step forward in this direction.

Appendix A

Comparing approaches to deal with non-gaussianity of rainfall data in kriging-based radar-gauge rainfall merging

Francesca Cecinati[1], Omar Wani[2,3], Miguel Angel Rico-Ramirez[1]

[1] Department of Civil Engineering, University of Bristol, Bristol, UK
[2] Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland
[3] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

Journal: Water Resources Research

Publication date: 12th November 2017

DOI: https://doi.org/10.1002/2016WR020330

Author contributions: FC, with inputs from OW, conceptualized the research questions, methodology and performed the calculations. OW contributed to the analysis on time variant optimization and to the formulation of the discussion section. MRR supervised the work and provided expertise on weather radar data processing. All authors have contributed to write and review the paper.

Abstract

Merging radar and rain gauge rainfall data is a technique used to improve the quality of spatial rainfall estimates and in particular the use of Kriging with External Drift (KED) is a very effective radar-rain gauge rainfall merging technique. However, kriging interpolations assume Gaussianity of the process. Rainfall has a strongly skewed, positive, probability distribution, characterized by a discontinuity due to intermittency. In KED rainfall residuals are used, implicitly calculated as the difference between rain gauge data and a linear function of the radar estimates. Rainfall residuals are non-Gaussian as well. The aim of this work is to evaluate the impact of applying KED to non-Gaussian rainfall residuals, and to assess the best techniques to improve Gaussianity. We compare Box-Cox transformations with λ parameters equal to 0.5, 0.25, and 0.1, Box-Cox with time-variant optimization of λ, normal score transformation, and a singularity analysis technique. The results suggest that Box-Cox with λ=0.1 and the singularity analysis are not suitable for KED. Normal score transformation and Box-Cox with optimized λ, or λ=0.25 produce satisfactory results in terms of Gaussianity of the residuals, probability distribution of the merged rainfall products, and rainfall estimate quality, when validated through cross-validation. However, it is observed that Box-Cox transformations are strongly dependent on the temporal and spatial variability of rainfall and on the units used for the rainfall intensity. Overall, applying transformations results in a quantitative improvement of the rainfall estimates only if the correct transformations for the specific dataset are used.

1 Introduction

Accurate spatially distributed estimates of precipitation are desirable in many water-related fields, like hydrology, water quality monitoring and water resources management. Merging weather radar rainfall estimates and rain gauge data has proved to be a successful solution compared to the use of a single source of information [Wilson, 1970; Haberlandt, 2007; Schuurmans et al., 2007; Goudenhoofdt & Delobbe, 2009; Jewell & Gaussiat, 2015]. Several techniques have been developed for merging radar and rain gauge data, many of which are based on kriging interpolation methods [Goudenhoofdt & Delobbe, 2009; Li and Heap, 2011; McKee and Binns, 2015]. In many comparative studies, kriging with external drift (KED - also known as universal kriging, or regression kriging) is reported to be one of the best performing merging methods [Goudenhoofdt & Delobbe, 2009; Velasco-Forero et al., 2009; Li and Heap, 2011; Jewell & Gaussiat, 2015; Nanding et al., 2015; Boudevillain et al., 2016]. Other methods that perform comparably well, usually require considerably more computational effort, making use of large covariance matrices or Monte Carlo methods [e.g. Todini, 2001; or Scheidegger & Rieckermann, 2014] and this high computational cost is an obstacle for their operational usage, especially for large areas, or for long time series. In spite of the good performance of KED, the method has some limitations that may introduce errors in the rainfall estimates. One of the limiting factors of many kriging-based merging techniques is their applicability only to Gaussian variables [Diggle and Ribeiro Jr, 2007; MÃ¼ller, 2007]. In particular KED requires the residuals between the rainfall process and its mean (or drift, modelled as a linear function of the radar estimate) to be Gaussian. Although in many rainfall data merging applications the Gaussian approximation is adopted [e.g. Schiemann et al., 2011; Delrieu et al., 2014; Verdin et al., 2015], rainfall intensity and accumulation have a strongly non-Gaussian, skewed, probability distribution, with only positive values and a discontinuous peak for the value zero. The residuals are not known a priori, but can be approximated by the difference between rain gauge values (assumed as the true rainfall at point locations) and a linear function of the radar values at rain gauge locations, derived from a linear regression [Delrieu et al., 2014]. Since the difference between two Gaussian variables results in a Gaussian variable, often Gaussianity of residuals is achieved applying Gaussian transformations on rain gauge and radar rainfall data. This leads to the use of transGaussian kriging methods, i.e. to the application of kriging to variables transformed with analytical or numerical Gaussian transformations [Cressie, 1993]. Various transformations can be applied to rainfall, like square root [Sideris et al., 2014], fourth root [Van De Beek et al., 2011], Box-Cox with variable λ parameter [Erdin et al., 2012], logarithmic [Sinclair and Pegram, 2005], or normal score transformation [de Wit et al., 2008; Germann et al., 2009; Villarini et al., 2014]. Beyond transGaussian kriging, some other approaches can be used to address non-Gaussianity of rainfall data. A possible approach is to separate the wet areas from the dry ones at each time step and do the merging only in the wet ones, in order to eliminate the discontinuity in the probability distribution corresponding to the zero values [Barancourt et al., 1992; Braud et al., 1994; Schleiss et al., 2014]. The extension of this approach is to separate different rain intensity intervals and krige them separately. Indicator kriging transforms the continuous variable in a categorical one, allowing to krige separately the values in different intervals, thus approximately reproducing the original cumulative density function [Remy et al., 2012]. It has been applied to KED, but an improvement is not always observed [Haberlandt, 2007; Berndt et al., 2014]. Disjunctive kriging uses a non-linear formulation, based on Hermite polynomials, to include a non-linear bijective relationship between the actual variable and a normal one [Yates et al., 1986a, 1986b, 1986c]. It has been applied to co-kriging merging, with good results, but the method is complex, computationally intensive, and its application to KED is not straight-forward, given the different ways that radar and rain gauges are integrated in KED [Azimi-Zonooz et al., 1989]. The local singularity analysis (SA) proposed by Wang et al., [2015], separates the non-Gaussian features in the spatial field (local singularities), and performs kriging only on the non-singular part, approximately Gaussian, recovering the singularities on the merged result. It has been applied to the Kalman filter - Bayesian merging technique proposed by Todini, [2001], but it has not been applied to Kriging with External Drift so far. The methodology is used to deal with the smoothing in rainfall fields due to the application of Gaussian merging methods.

Although many methods of different complexity are available, not much work has been done on the comparison between them and on the uncertainty reduction deriving from their application. Kirstetter et al., [2010], showed that the probability distribution of rainfall residuals conditioned on the rainfall intensity is not as far from Gaussian as the probability distribution of rainfall, suggesting that a transformation of rainfall may be not necessary at all for KED application. Erdin et al., [2012] studied the effects of using Box-Cox transformations in KED merging, comparing the use of different λ parameters. The work acknowledges that the rainfall is best transformed to a normal variable using a transformation close to logarithmic ($\lambda \to 0$), but, at the same time, an almost logarithmic transformation introduces biases in the back transformation, therefore the study recommends a parameter close to λ=0.2, and proposes a Bayesian approach to infer the best parameter at each time step. It must also be noticed that the logarithmic transformation cannot be applied to rainfall data, because of the impossibility of transforming the zero values, therefore low values of λ need to be tested instead. Although Erdin et al.'s findings are particularly significant to the aim of this work and are compared to the outcome of it, the study does not compare Box-Cox transformations with numerical transformations or alternative methods. The Normal Score Transformation (NST), also known as Normal Quantile Transformation (NQT), is a technique widely used in hydrology [Montanari, A., & Brath, 2004; Weerts et al., 2011; Bogner et al., 2012; Villarini et al., 2014]. It is based on the idea of finding an empirical transformation to convert the quantiles of the available dataset probability distribution, to the quantiles of a standard normal distribution. Its application to rainfall presents some difficulties that need to be addressed. The application of NST should be done on variables with a strictly increasing, and continuous distribution [Kelly and Krzysztofowicz, 1994; Herr and Krzysztofowicz, 2005]. When this is not the case, like in the case of rainfall, numerical approximations need to be introduced. For example, the discontinuity of the rainfall probability distribution in zero is addressed by Lien et al., [2013], or the back-transformation of values outside the observed domain is addressed by Bogner et al., [2012]. Similar solutions are addressed and discussed in the methodological part of this paper.
The objectives of this work are:

1. To compare different rainfall products generated using different approaches to address the rainfall non-Gaussianity problem:
    1.1. Radar without merging,
    1.2. Ordinary Kriging,
    1.3. KED without transformations,
    1.4. KED with Box-Cox with λ=0.1 (almost logarithmic),
    1.5. KED with Box-Cox with λ=0.25 (fourth root),
    1.6. KED with Box-Cox with λ=0.5 (square root,
    1.7. KED with Box-Cox with time-variant optimization of λ,
    1.8. KED with Normal Score Transformation,
    1.9. KED with an adaptation of the Singularity Analysis technique approximation problem in radar-gauge rainfall merging.
2. To understand the advantages and the issues in addressing the Gaussian approximation problem in radar-gauge rainfall merging.

This work provides a comprehensive comparison of the most popular methodologies to address the Gaussian approximation in radar-gauge rainfall merging and the results are of practical use for the scientific community and operational agencies. The work confirms and expands the limited literature existing in this field. A broader comparison, both in terms of methodology and investigation techniques, is implemented. The results reinforce the validity of previous findings, also thanks to the application to a different case study and to different datasets, and show more insight into the applicability of the different methodologies. Additionally, some important considerations on the variability of the optimal transformation with time and space variant rainfall intensity and used

units are included.

The document is organized as follow: the case study and the used datasets are presented in Section 2; Section 3 presents all the tested methods to correct for Gaussianity; Section 4 introduces the techniques used for the evaluation of the results; the results obtained by the application of the methods to the case study and the quantitative and qualitative evaluations are presented in Section 5 and discussed in Section 6; the conclusions are drawn in Section 7.

2 Datasets and case study

The proposed case study is based on 9 months of data, from January 2016 to September 2016, in the UK. The data represents different weather conditions: In January, February and the beginning of March the rainfall events were mainly stratiform; in June, July and August, strongly convective events occur; in April, May and September rainfall is characterized by a mix of frontal rainfall and convective precipitation. The analysis is conducted only on "wet hours" defined as those hours in which both the rain gauges and the radar recorded rainfall in the studied area, the mean of all the radar pixels is greater than 0.001 mm/h and at least the 1% of pixels record rainfall. The thresholds are empirical, based on the specific case study and resulted in 3968 hours of data available for the study.

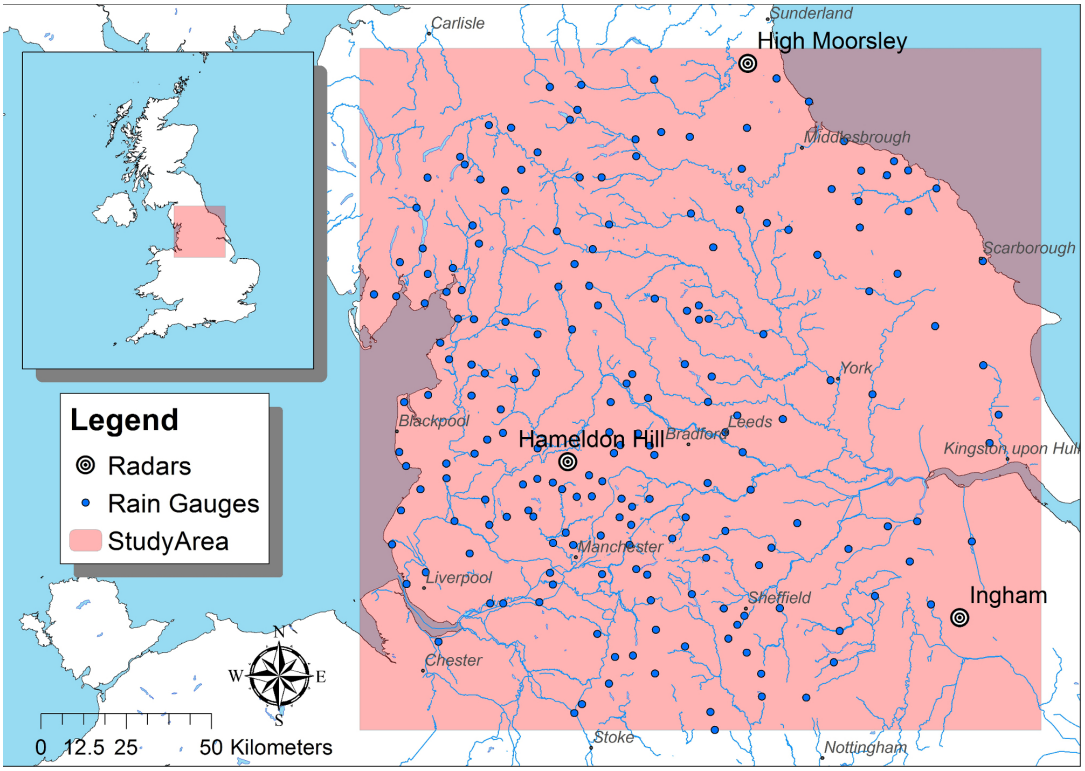


Figure 1: Study area. The figure shows the position of the rain gauges, the radars, and the considered study area extent.

The considered area is a large portion of Northern England, 200 km by 200 km wide. The area covers flat and hilly orography, as well as more urban environment in the South and more natural in the North. The area contains 212 tipping bucket rain gauges from the Environment Agency, with a tipping resolution of 0.2 mm. The dataset is freely available upon request(national.requests@environment-agency.gov.uk) under the Open Government Licence. The dataset was accumulated to hourly time steps. The dataset was quality checked, to eliminate the stations

that reported an excessive number of missing values, and to eliminate those presenting suspicious behaviors (signs of blocking, strong and consistent disagreement with neighboring stations and the radar, or unrealistic prolonged dry spells). This operation was performed manually and expert judgment was used to recognise errors. The Environment Agency rain gauge network is represented in Figure 1. The radar rainfall data mosaic is provided by the UK Met Office at 5 minutes resolution on a 1 km by 1 km grid [Met Office, 2003]. The study area is covered by three C-band weather radars: Hameldon Hill, High Moorsley, and Ingham. The radar rainfall product is provided by the UK Met Office through the BADC (British Atmospheric Data Centre) portal already corrected for beam blocking, clutter, attenuation, anomalous propagation, vertical reflectivity profile, bright band, and orographic enhancement; it is also bias-corrected with rain gauge measurements [Harrison et al., 2000, 2009]. The rain gauge network used by the UK Met Office for bias correction is different and independent from the one used in this work, managed by the Environment Agency. In 2016, the Hameldon Hill and the Ingham radars were already converted to polarimetric, while the High Moorsley radar still used single-polarisation. The data, as provided by the UK Met Office, is at 5-minute resolution and presents some gaps over time. Any missing 5-minute resolution radar scans were interpolated by advecting the rainfall radar field with a nowcasting model (maximum of 3h), and the 5-minute radar data are accumulated at hourly time steps. However, the nowcasting interpolations represent less than the 0.8% of the data. A residual 0.6% of radar data remains not available.

3 Methods

Starting from the same dataset, different rainfall products are obtained by applying different techniques to correct the rainfall non-Gaussianity. The transformed rain gauge and radar data are then merged and back transformed. The original radar data and an ordinary kriging interpolation of rain gauges without the use of radar data are added to the comparison as references. The methods and the products are then compared under different aspects, presented in Section 4. The compared products are:

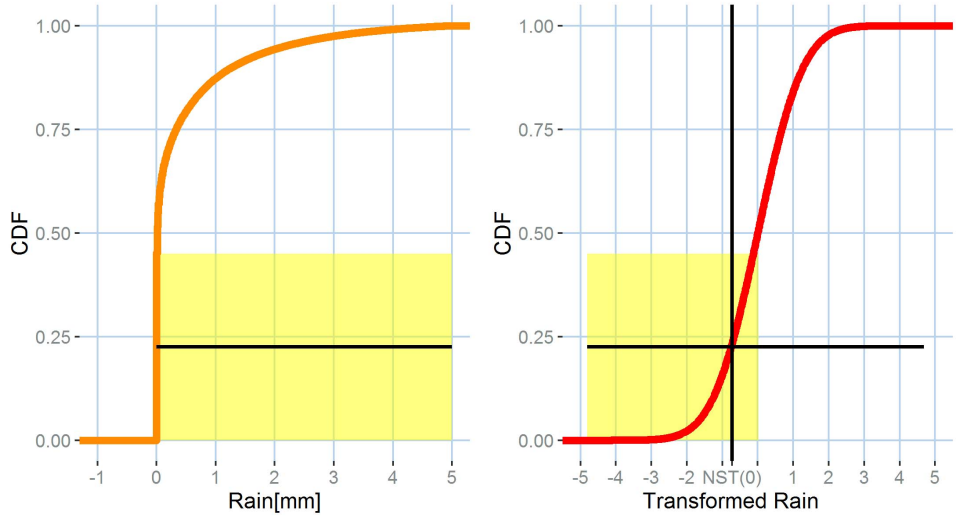


Figure 2: Example of the method used to transform with NS the nonunique values, in this case applied to the zero value. All the values equal to zero in the data set are transformed to NST (0), which is the median of the corresponding values in the Gaussian space.

1. Radar only —RD
2. Ordinary Kriging (using only rain gauges) —OK

3. KED with no transformation applied —NO
4. KED applying Box-Cox transformation with λ=0.1 (almost logarithmic) —BC 0.1
5. KED applying Box-Cox transformation with λ=0.25 (fourth root) —BC 0.25
6. KED applying Box-Cox transformation with λ=0.5 (square root) —BC 0.5
7. KED applying Box-Cox transformation with time-variant optimization of λ —BC opt
8. KED with Normal Scores Transformation —NS
9. KED with Normal Scores Transformation —NS
10. KED with adapted Singularity Analysis —SA

The proposed methods aim at transforming the rainfall data sets to be merged and back transform the results of the KED merging. The KED merging algorithm is therefore not affected and is the same for all the products. We use the formulation included in the gstat R package, based on Cressie (1993).

3.1 Ordinary kriging (OK)

Ordinary Kriging (OK) is one of the most popular interpolation methods and can be summarized as a weighted average interpolation, where the weights are a function of the distance between the rain gauge observations and the estimated point, and of the correlation characteristics of the modelled phenomenon as observed from the available data. OK does not use radar information for the interpolation. In this work, a global average is used, i.e., all the rain gauge observations are used to estimate each point on a 1 km by 1 km grid. The rainfall is estimated in a generic point x_0 as follow:

$$\hat{R}_{OK}(x_0) = \sum_{\alpha=1}^{n} w_\alpha \cdot R(x_\alpha), \tag{1}$$

where $\hat{R}_{OK}(x_0)$ is the estimated rainfall in a generic point x_0, $R(x_\alpha)$ are the measured values in the rain gauge locations x_α, n is the number of observations, and w_α are the kriging weights, estimated minimizing the variance under unbiased conditions, which results in solving the kriging system (Cressie, 1993):

$$\begin{cases} \sum_{\alpha=1}^{n} w_\alpha = 1 \\ \sum_{\alpha=1}^{n} w_\alpha \cdot C(x_\beta - x_\alpha) + \mu_1 = C(x_\beta - x_0) \quad \beta = 1, ..., n \end{cases} \tag{2}$$

Where $C(d)$ is the covariance function at distance d, x_α and x_β are generic rain gauge locations, and μ_1 is a Lagrange parameter (Cressie, 1993). In this paper, the covariance function is estimated at each time step as illustrated in section 3.7. When the kriging system is solved, the optimized kriging variance for each point x_0 is equal to:

$$\sigma^2(x_0) = c - \mu_1 - \sum_{\alpha=1}^{n} w_\alpha \cdot C(x_\alpha - x_0), \tag{3}$$

where c is the sill parameter, explained in section 3.3.

3.2 Kriging with external drift (KED)

Universal kriging, as opposed to ordinary kriging, considers the mean of the studied field $R(x)$ nonstationary in space:

$$R(x) = m(x) + \delta(x), \tag{4}$$

where $m(x)$ is the mean and $\delta(x)$ is a zero-mean stationary random process (Cressie, 1993). Kriging with External Drift (KED) is a special case of universal kriging, in which the mean is considered a linear function of external factors. In the presented case, the mean of the rain gauge interpolation is considered a linear function of the radar rainfall estimate:

$$m(x) = a + b \cdot r(x), \tag{5}$$

where $r(x)$ is the radar rainfall estimate at location x, whereas a and b are linear coefficients to be determined.
KED equations are very similar to the ones for OK. The rainfall estimation at location x_0, $\hat{R}_{KED}(x_0)$, is derived from equation (3) but the weights are calculated as follows:

$$\begin{cases} \sum_{\alpha=1}^{n} w_\alpha = 1 \\ \sum_{\alpha=1}^{n} w_\alpha \cdot C(x_\beta - x_\alpha) + \mu_1 + \mu_2 \cdot r(x_\beta) = C(x_\beta - x_0) \quad \beta = 1, ..., n \\ \sum_{\alpha=1}^{n} w_\alpha \cdot r(x_\alpha) = r(x_0), \end{cases} \tag{6}$$

where $r(x)$ is the radar observation in a point x and μ_2 is a second Lagrange parameter. Similarly, the kriging variance equation for each point x_0 is modified as follows:

$$\sigma^2(x_0) = c - \mu_1 - \mu_2 \cdot r_0 - \sum_{\alpha=1}^{n} w_\alpha \cdot C(x_\alpha - x_0). \tag{7}$$

3.3 Variogram and covariance functions

The correlation characteristics of the modelled processes are estimated at each time step after the data set transformation. However, KED requires the variogram to be estimated on the rainfall residuals, which are not known a priori, therefore a four-passage process is followed to estimate the residual variogram.

1. The rainfall variogram is estimated at each time step using a subset of the wet radar pixels. Radar estimates are used instead of rain gauges, to have a sufficient number of observations at each time step. In particular, if less than 1,000 wet pixels are available, all of them are used (however, at least 400 pixels - 1% of the total - need to be wet to consider the time step "wet"), otherwise the subset is limited to 1,000 pixels randomly selected, to limit the required computational time. The empirical variogram is calculatedusing 1 km bins, and then fitted with an exponential function:

$$\gamma(d) = \begin{cases} 0, & \textit{for } d = 0 \\ c_0 + c\left(1 - \exp\left(-\frac{3d^2}{h^2}\right)\right), & \textit{for } d > 0. \end{cases} \tag{8}$$

   where h is the range parameter, c_0 is the nugget, and c is the sill.

2. OK is performed on the rain gauges as illustrated in section 3.1, obtaining the field $\hat{R}_{OK}(x)$.

3. The residuals are estimated as the difference between the OK field and a linear function of the radar estimate:

$$y(x) = \hat{R}_{OK}(x) - (a \cdot r(x) + b), \tag{9}$$

   where a and b are estimated fitting a linear model between the OK field $\hat{R}_{OK}(x)$ and the radar field $r(x)$.

4. An empirical variogram is calculated as in point 1, using the residual field $y(x$ instead of the radar field.

The covariance function $C(d)$ is directly derived from the variogram function $\gamma(d)$, thanks to the assumption of stationary random function:

$$C(d) = c + c_0 - \gamma(d). \tag{10}$$

3.4 Box-Cox with fixed λ

The Box-Cox transformation is one of the most well-known analytical transformations used to improve the Gaussianity of a given data set. The transformation is dependent on only one parameter λ:

$$y^* = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}, \tag{11}$$

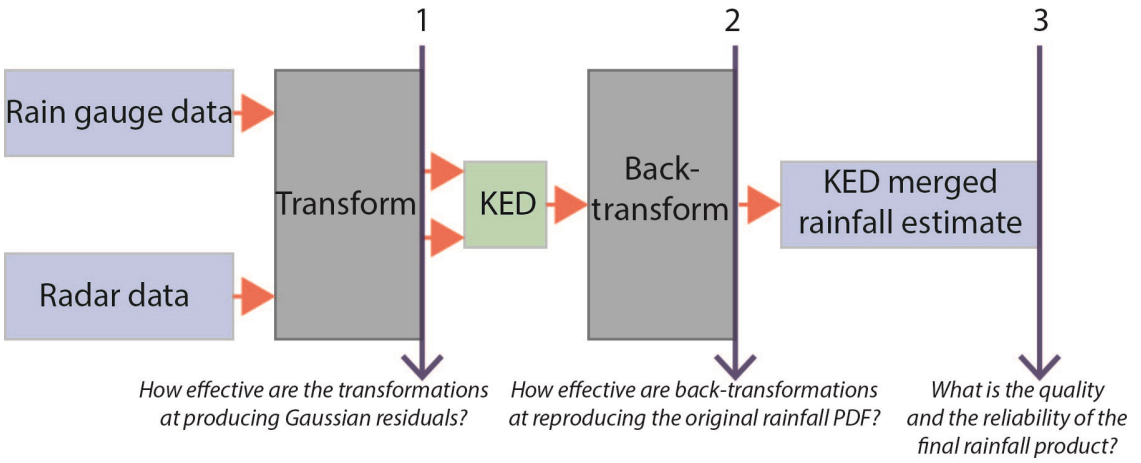


Figure 3: The evaluation of the studied methods is done in three stages, at three different phases of the radar-gauge rainfall merging process.

where y is a nontransformed variable of interest and y^* is the variable after transformation.
In this work, three fixed λ parameters are tested. λ=0.1 is used to test an almost-logarithmic transformation, since the logarithmic one with λ=0 cannot be applied to the rainfall data set, because of the presence of zero values. λ=0.25 corresponds to a fourth root of the variable, i.e., to a square root - square root transformation. Finally, λ=0.5 corresponds to a square root transformation of the variable. The same transformation is applied to both the radar and the rain gauge data set before merging. KED provides us with mean and variance of the merged field, in the transformed domain. For the square root and for the fourth root transformations, there is an analytical formula to obtain the mean and the variance in the untransformed space. Nevertheless, for consistency with all the other methods that do not have an analytical back-transformation, the back-transformation is accomplished applying the inverse of the analytical transformation function to 99 quantiles and calculating the mean of the back-transformed quantiles. This is done for each pixel at each time step, based on the kriging mean and variance.

3.5 Box-Cox with time-variant optimal λ

The Box-Cox transformations with a fixed λ parameter assume that the probability distribution of rainfall is stationary in space and time. While we have no means to estimate separately the

probability distribution of the different points in space, it is possible to assess at each time step what is the best transformation to apply. Erdin et al. (2012), propose a method to estimate the optimal λ parameter at each time step using Bayesian inference.

The assumption that in the transformed space rain gauge measurements should be equal to a linear function of the radar measurements (drift) plus an additive normal noise that is independent and identically distributed is used to construct a likelihood function. This likelihood function is used to arrive at the posterior distribution of λ; and the value at the maximum posterior probability density is chosen.

In our analysis, we prefer to use a different objective function for two reasons. First, maximizing the posterior density does not ensure to maximize the Gaussianity of the sample. Second, posterior probability tends to be highly sensitive to the choice of our prior distribution, given the limited number of spatial observations at each time step.

We use an approximation of negentropy as objective function. Negentropy tests are based on the information theory principle that, given a certain variance, entropy is maximum for the Gaussian distribution (Cover & Thomas, 1991). Negentropy, defined as differential entropy, is zero for Gaussian variables, and larger than zero otherwise, but it requires knowledge of the variable probability distribution to be calculated. Approximations of negentropy are therefore used instead. Here, we use an approximation based on the use of nonquadratic functions (Hyvärinen & Oja, 2000):

$$J(y) \propto J_a(y) = (E\{G(y)\} - E\{G(z)\})^2, \tag{12}$$

$$G(u) = -\exp\left(-\frac{u^2}{2}\right). \tag{13}$$

Where $E\{-\}$ represents the expected value operation, $J(y)$ is the negentropy of the normalized residual variable y; $J_a(y)$ is its approximation, z is a normal variable, and G is a nonquadratic function, in this case exponential. The approximation of negentropy $J_a(y)$ ranges between zero and $+\infty$, and for Gaussian distributions it is equal to zero. Therefore, we select the λ that, applied to rain gauge and radar data, minimizes the approximation of negentropy of the residuals. As in Erdin et al. (2012), the tested λ values are limited to the interval [0.2, 1.5]. To avoid skewing the residual distribution using an ordinary kriging interpolation of the rain gauges, the residuals are calculated only at rain gauge locations. However, to ensure that at each time step a sufficient number of points is available to optimize λ; a time window centered in the time step of interest is used, so that at least 500 points are available. The time window for this case study is never larger than 11 h.

3.6 Normal score transformation

Normal score transformation (NST) is an empirical transformation method, that associates each quantile of a given probability distribution to the corresponding quantiles of a standard normal probability distribution, thanks to multiplicative factors (scores, Bogner et al., 2012). The method is simple and very effective, but requires the data set to transform to have a continuous strictly increasing cumulative distribution. Rainfall data sets do not respect these conditions and therefore the method has to be modified, and some approximations need to be done.
In particular, two problems are here addressed:

1. The data sets contain a large number of zeros, meaning that the cumulative distribution has an initial jump for the value zero. Additionally, the same problem appears for values multiples of 0.2 mm because the rain gauges have a tipping bucket resolution of 0.2 mm; therefore, the data set is not strictly continuous. This results in the fact that equal values in the nontransformed domain are associated to multiple different values in the transformed

domain. The method used to solve the problem is similar to the one proposed by Lien et al. (2013): to all the equal values in the nontransformed domain is associated the median of all the different values that would have been obtained in the transformed domain applying NST without correction. Figure 2 illustrates the problem and the adopted solution for the zero value as an example.

2. The KED interpolation may generate values outside the observation domain. For those values, an extrapolation method is necessary. In the presented case, a different approach is used for the values falling outside the domain on the negative side, and the ones falling outside the domain on the positive side. For the former issue, all the values below the zero threshold are converted to zero. For the latter issue, the values are extrapolated with a simple linear regression using the last two observed values.

Considering both solutions, the normal score transformation is applied as follow:

1. At each time step, both the radar and the rain gauge data sets are considered, and all the values are sorted in increasing order.

2. The corresponding number of quantiles from a standard normal distribution are calculated.

3. A lookup table is created matching the quantiles from the original distribution to the corresponding quantiles of the standard normal distribution.

4. Both the radar and the rain gauge data sets are transformed using the lookup table.

5. The maximum normal quantile corresponding to zero $NS_{0,max}$ is recorded, before applying the correction method for repeated values presented above. All the values in the transformed domain corresponding to equal values in the original domain are substituted with their median.

At this point the KED merging can be performed, resulting in a probabilistic estimation of rainfall on a grid in the transformed domain, described by a mean value and a variance for each pixel. The back transformation is calculated as follow:

1. For each pixel, given the mean and the variance obtained from the KED merging, 99 quantiles are calculated.

2. For each quantile, if it is lower than the zero threshold $NS_{0,max}$, it is back-transformed to zero, otherwise, the two values above and below are found in the lookup table. If the value is larger than all the scores in the lookup table, the two maximum values in the lookup table are used. The two values are used to linearly interpolate/extrapolate the back-transformed value for that quantile.

3. The mean of the back-transformed quantiles is calculated and assigned to the pixel.

3.7 Singularity analysis

The singularity analysis method, introduced by Cheng et al. (1994) to identify geochemical anomalies, was adapted to identify and separate singularities in rainfall by Wang et al. (2015). According to Wang et al. (2015 the term *singularity* refers to "an anomalous amount of energy release or mass accumulation [$\cdots$] often associated with structures depicting fractality or multifractality." It must be noted that the aim of the method is not strictly to obtain a Gaussian field, but to separate singularities, which are characteristic of non-Gaussian structures and carry information relative to moments beyond the second. However, removing singularities in the merging phase has also the effect of obtaining a field closer to Gaussian; recovering the singularities in the merged products should help reconstruct a field with a probability distribution similar to the one of the original data.

Rainfall singularities are characterized by the fact that the areal averaged rainfall centered in x follows a power function of the area (Wang et al., 2015):

$$\rho(x,\epsilon) = c_{SA}(x)\epsilon^{\alpha_{SA}(x)-E_{SA}}, \tag{14}$$

where $\epsilon = I/L$, i.e., the side of the square under consideration I divided by the side of the maximum considered square L; $\rho(x,\epsilon)$ is the average rainfall intensity in a square centered in x with side I, $c_{SA}(x)$ is a constant intensity value, $\alpha_{SA}(x)$ is the singularity index, and E_{SA} is the Euclidean dimension. The squares considered for each pixel are selected to have I=1, 3, 5, 7, 9 km.
$c_{SA}(x)$ and $(\alpha_{SA}(x) - E_{SA})$ can be calculated as the intercept and the slope of a line fitting $\log(\rho(x,\epsilon))$ as function of $\log(\epsilon)$. Then, $c_{SA}(x)$ can be considered the nonsingular rainfall value. In reality, to reach stability, the process needs to be iterated

$$c_{SA}^{k-1}(x) = c_{SA}^{k}(x)\epsilon^{\alpha_{SA}^{k}(x)-E_{SA}}, \tag{15}$$

where $c_{SA}^{-1}(x)$ is the original value and k=0, 1, ...n. A number of iteration equal to 5 (k=4) is selected, as it should be sufficient to remove the singularities (Wang et al., 2015).
Wang et al. (2015) applied the singularity analysis to the Kalman filter, Bayesian radar-gauge merging technique (Todini, 2001). The Bayesian merging technique does not require to remove the singularities from rain gauge data, because a block kriged rain gauge field is used to be merged with the nonsingular radar field, which already produces a smoother and more Gaussian field. In kriging with external drift, instead, the rain gauge point measurements are used directly, without previous interpolation. The main limiting factor for the application of the singularity analysis method to KED is that the singularity removal cannot be applied to rain gauge data, because they lack the areal characteristics necessary to apply the method. The singularity identification and removal is therefore only applied to radar. An attempt to modify rain gauge values using the same factor of the radar pixel containing the rain gauge was made, but the results were not satisfactory, due to the very large representativeness difference between the point measurement and the 1 km by 1 km radar pixel value, especially for the convective events that generate the most intense singularities. The removed singularities are reintroduced by multiplying the nonsingular KED rainfall estimation $\hat{R}_{KED,SA}(x)$ by the ratio between the calculated nonsingular radar field $r_{SA}(x)$, and the original radar field r(x):

$$\hat{R}_{KED} = \hat{R}_{KED,SA}(x) \cdot \frac{r_{SA}(x)}{r(x)}. \tag{16}$$

4 Evaluation techniques

The studied methods are evaluated under three points of view, as reported in Figure 3. The first evaluation stage is based on the ability to transform the rainfall residuals in Gaussian variables. The second evaluation stage is based on the ability to reproduce the original probability distribution after back-transformation of the merged products. The true probability distribution of rainfall is not known, but it is assumed that the radar Probability Density Function (PDF) is proportional to it, and therefore this test is done comparing the radar data and the back-transformed KED product with QQ-plots. Finally, the most important evaluation stage assesses the quality of the final product. The KED rainfall estimates are validated with "leave one out" cross-validation. Figure 3 illustrates when each evaluation technique is used in the merging process, as well as what questions it aims at answering.

4.1 Gaussianity of the residuals

Although transformations are applied to radar and rain gauge rainfall, KED requires rainfall residuals to be Gaussian. A Gaussianity test is applied at each time step to the residuals, as calculated in section 3.5. There are several methods used in literature to assess the Gaussianity of a data set.

Two of the most used indicators are kurtosis and skewness, which give a measure of the "tailedness" and asymmetry of a distribution, respectively. Both are informative measures, but are not univocally indicators of Gaussianity (Joanes & Gill, 1998), since they measure the fourth and the third moment, respectively. A third method, the approximation of negentropy, is based on the notion that entropy is maximum for a Gaussian variable, for a given variance. The presentation of the three indicators together gives a more complete overview of the data set Gaussianity. The indicators are calculated at each time step, using a time window as illustrated in section 3.5. The values are then averaged over each month.

Kurtosis

Kurtosis is a measurement of the fourth standardized moment, and has been proposed in different forms. In this work we adopted the formulation used in many statistical software packages, like SAS, SPSS, and EXCEL (Joanes & Gill, 1998):

$$G_2(y) = \left((n+1) \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^2} - 3 \right) + 6 \right) \cdot \frac{(n-1)}{(n-2)(n-3)}, \tag{17}$$

where y_i are the available residual values, $\bar{y}$ is their mean, and n is the number of samples. Kurtosis $G_2(y)$ ranges between $-\infty$ and $+\infty$, and for Gaussian distributions it is equal to zero.

Skewness

Skewness is a measure of asymmetry of the distribution and is defined as the third standardized moment. The formulation adopted here is the one adopted in statistical packages, similarly to the kurtosis one, (Joanes & Gill, 1998):

$$G_1(y) = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{3/2}} \right) \cdot \frac{\sqrt{(n(n-1))}}{(n-2)}. \tag{18}$$

Skewness $G_1(y)$ ranges between $-\infty$ and $+\infty$, and for Gaussian distributions it is equal to zero.

Approximation of negentropy

The approximation of negentropy is a fast and robust method to estimate the Gaussianity of a variable. The formulation in section 3.5 is used to evaluate the Gaussianity of the residuals as well.

4.2 Coefficient of determination R^2 of QQ-plots

The second mean of evaluation of the studied methods is the comparison between the original rainfall probability distribution, and the probability distribution of the merged rainfall products after backtransformation. Since the probability distribution of the true rainfall is not known, the probability distribution of the original radar data set is compared to the probability distribution of the KED back-transformed rainfall estimates. Although rain gauges are assumed to be representative of the true rainfall, we decided to use the radar data for two reasons:

1. Radar rainfall is a continuous measurement, while rain gauge data is discretized due to the bucket resolution equal to 0.2 mm.

2. Rain gauge data points are available at given locations only, and they may miss the extremes of the distribution.

The 99 quantiles of the radar PDF are plotted against 99 quantiles of the KED products. Since bias may be present in the radar data and may have been adjusted through the merging process,

we do not expect the data to lay on a $x = y$ line, but rather on a generic straight line. For this reason, a linear model is fit to the QQ-plots and a coefficient of determination R^2 is calculated using the fitted line. R^2 ranges from 0 to 1, with 1 indicating perfect correlation. If the R^2 is close to 1, regardless of the slope and intercept of the fitted line, the method is assumed to reproduce the probability distribution of the original rainfall data.

4.3 Rain gauge validation

The ability of a method to produce Gaussian residuals and to back transform the final merged rainfall products to the original probability distribution are important for the kriging application, given the Gaussianity assumption underlying the method. Nevertheless, the primary purpose of the Gaussianity correction is to improve the quality and the reliability of the final rainfall estimates. For this reason, the most important evaluation is the validation of the final merged products.

A cross-validation exercise is carried out. The KED estimation is performed leaving one rain gauge out, and estimating the value in its location. Then the exercise is repeated for each rain gauge. The KED estimations at rain gauge locations are compared to the rain gauge values. Two indicators are used to quantitatively compare the KED estimates and the rain gauge measurements: mean root transformed error and bias. The estimators are calculated at each time step and presented averaged monthly, to compare the different performance over the change of season and rainfall conditions

Mean root transformed error

The Mean Root Transformed Error (MRTE) is a measurement of the difference between observations and modeled data that is preferred in meteorology to the more common Root Mean Square Error (RMSE) because of the lower weight given to high intensity rainfall errors. Its values range from 0 to $+\infty$, with optimal value equal to zero. It is calculated as:

$$MRTE = \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\hat{R}_{KED}(x_i)} - \sqrt{R(x_i)}\right)^2, \tag{19}$$

where $\hat{R}_{KED}(x_i)$ is the KED prediction at rain gauge location x_i, $R(x_i)$ is the corresponding rain gauge observation, and n is the number of rain gauges.

Bias

To measure the bias in the estimated rainfall, the mean of the difference between the estimated value and the observed value is calculated. The bias value defined this way can range from $-\infty$ to $+\infty$, with optimal value equal to 0:

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}\hat{R}_{KED}(x_i) - R(x_i) \tag{20}$$

5 Results

The three evaluation passages shown in Figure 3 and explained in section 4 are used to evaluate the performance of all the tested kriging methods applied to the case study. The results are reported in this section. Additionally, a qualitative evaluation is reported.

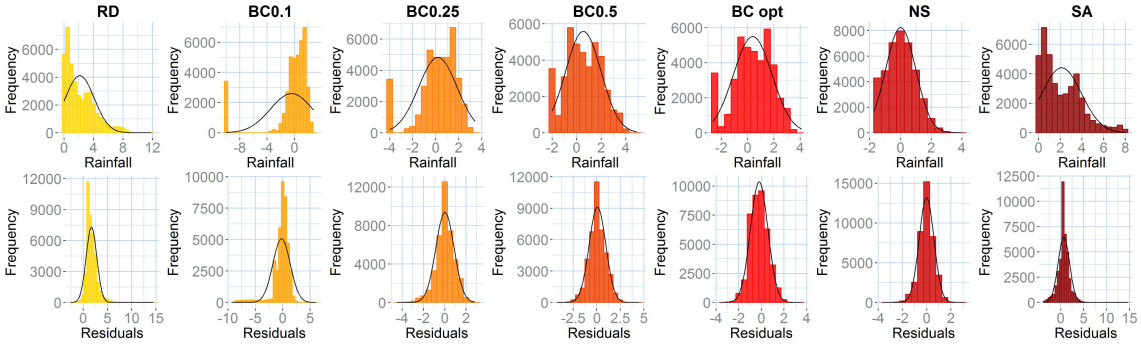


Figure 4: The histograms of rainfall, as estimated by the radar, and the corresponding residuals, approximating the empirical probability distributions, are presented with and without transformations for a case study at 04:00 of the 28 March 2016. The black line represents the corresponding Gaussian distribution with same mean and variance. The optimal k for this time step is equal to 0.38. Note that the units of rainfall and residuals after transformation are no longer (mm/h).

5.1 Gaussianity of the residuals

The first evaluation stage assesses how well the methods perform in transforming rainfall residuals to a Gaussian variable. Although the transformations are applied to radar and rain gauge rainfall data, their effects are expected to improve also the Gaussianity of the residuals. For a qualitative understanding of the rainfall and the residuals probability distributions, Figure 4 shows the histograms of radar rainfall and residuals, approximated as explained in section 3.5, with and without transformations, for an example time step at 04:00 of the 28 March 2016. For the specific time step, the Box-Cox transformation with optimized parameter used an estimated $\lambda = 0.38$. To quantitatively evaluate how effective the transformations are in producing Gaussian residuals, tests of Gaussianity are applied to the residuals. Figure 5 shows how the methods compare to each other and to the untransformed data, in terms of the three selected indicators: kurtosis, skewness, and approximation of negentropy.

The results show that all the methods improve the Gaussianity of the rainfall residuals. The SA is performing worst, probably affected by the impossibility of transforming rain gauge data. The best performing method is the NS, but satisfying results are obtained also with Box-Cox with a fixed $\lambda = 0.1$, $\lambda = 0.25$, and with the optimal λ. The optimized Box-Cox transformation does not perform as well as $\lambda = 0.1$, because it is limited in the range [0.2, 1.5]. In accordance to the results published by Erdin et al. (2012), the majority of the parameters are close to the lower limit of $\lambda = 0.2$ as shown in Figure 6. This explains why the results for the fixed $\lambda = 0.25$ and the time-variant λ are very similar.

5.2 Coefficient of determination R^2 of QQ-Plots

The second evaluation stage assesses the ability of each method to reproduce the original rainfall probability distribution, approximated by a linear function of the radar probability distribution. The 99 quantiles from the KED distribution are plotted against 99 quantiles from the original radar distribution. To account for possible radar biases, the methods are compared using the coefficient of determination R^2 relative to a linear regression fitting the QQ-plots, regardless of the slope and intercept. Example QQ-plots for the same time step considered in Figure 4, 04:00 of the 28 March 2016, are presented in Figure 7, while the quantitative comparison of the R^2 for the whole study, averaged on each month, is reported in Figure 8.

The results of the second evaluation stage show that all methods achieve a satisfactory reproduction of the original rainfall probability distribution, resulting in R^2 values for the QQ-plots above 0.8 in all analyzed cases, and very little differences can be noticed across the studied methods. Ordinary

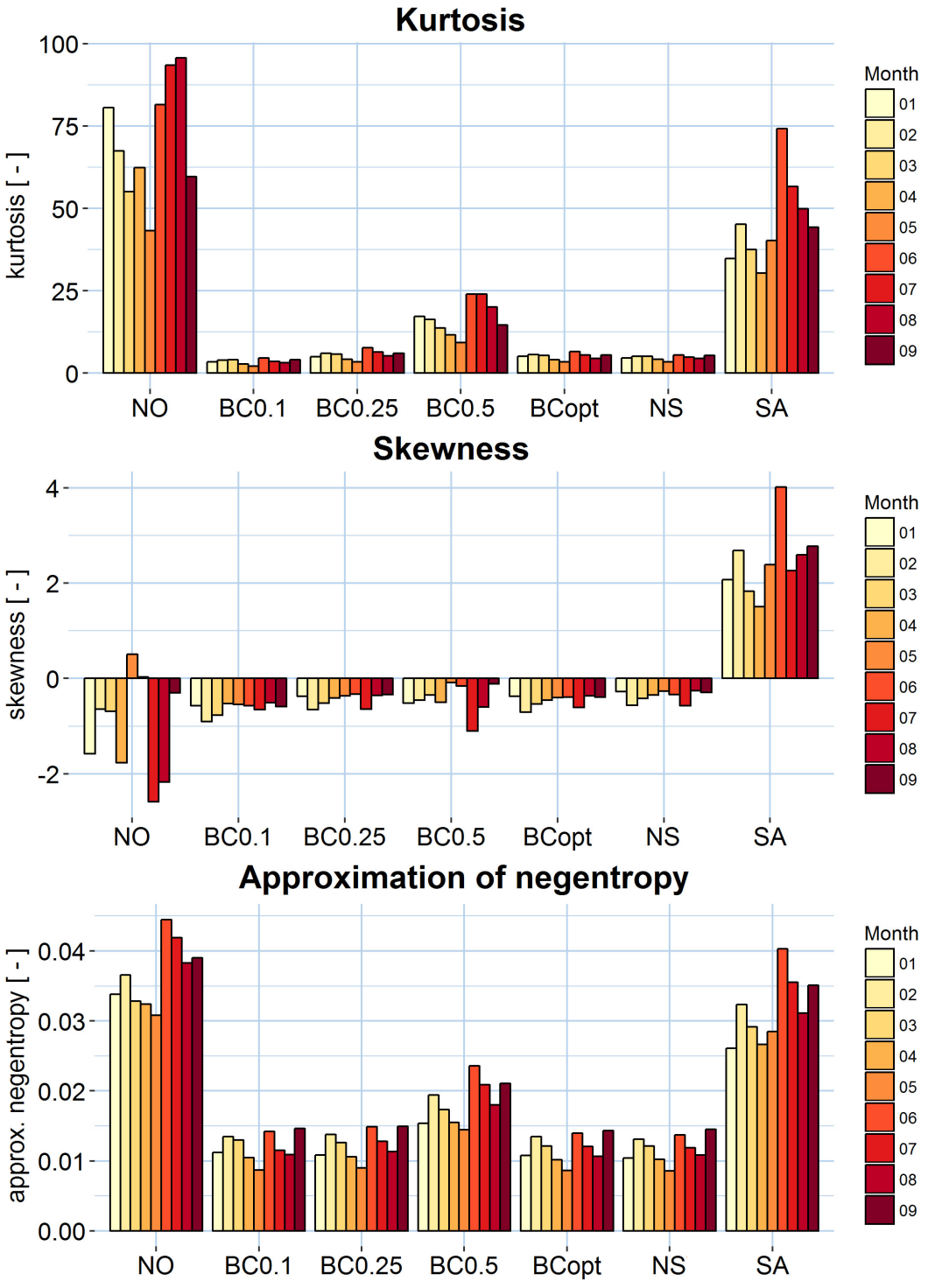


Figure 5: Performance of the tested methods in terms of the ability to transform the rainfall residuals into a Gaussian variable. Three indicators are shown: Kurtosis, Skewness, and Approximation of Negentropy, for the studied 9 months. The methods' acronyms are the ones introduced in section 3.

Kriging is included in the comparison as well. The Box-Cox transformation with $\lambda = 0.1$ shows

worse results overall, while the use of no transformation achieves comparably good results in this evaluation stage, meaning that the use of transformations does not bring particular benefits in terms of reproducing the original rainfall probability distribution. The reproduction of the original probability distribution by KED products seem more effective in winter months, when more stratiform precipitation is observed.

5.3 Rain gauge validation

The last, but most important stage of evaluation assesses the quality and the reliability of the merged products. Validation of the results is performed with cross-validation. The performance is evaluated with two indicators, namely the Mean Root Transformed Error (MRTE) and the bias. The comparison between the performance of the radar without merging, the OK results, the KED merged product without Gaussianity corrections, and the KED with the different tested correction methods are reported in Figure 9.

The results confirm that KED, even without transformations, produces better results compared to the noncorrected radar rainfall estimates. Nevertheless, OK performs better than KED without transformation. The use of a Box-Cox transformation with $\lambda = 0.1$, produces significantly worse results. The use of Box-Cox with: $\lambda = 0.5$ and of the Singularity Analysis produce nonsatisfactory results as well, comparable to the use of no transformation. Improvements are instead observed using the optimized Box-Cox, Box-Cox with fixed $\lambda = 0.25$ and the normal score transformation, both in terms of MRTE and bias.

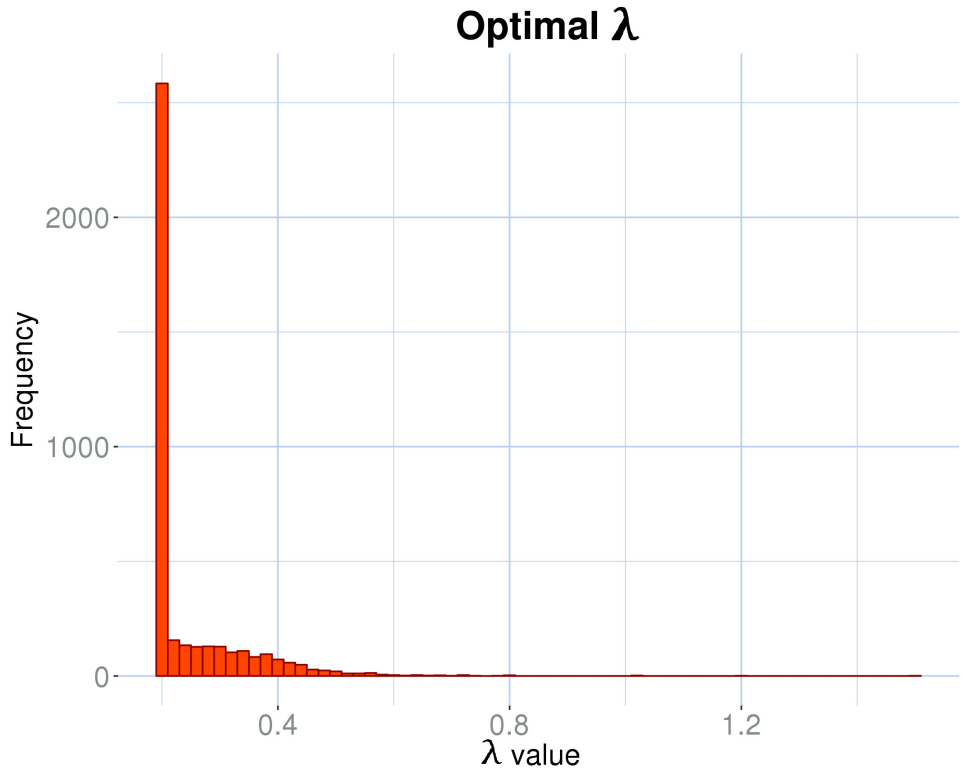


Figure 6: Frequency of the optimal lambda values over time.

5.4 Qualitative evaluation

In addition to the skill scores used to quantitatively assess the performance of the studied methods, an example of a visual comparison of the different rainfall estimates is reported in Figure 10. The

presented rainfall estimations are relative to the same example presented in Figures 4 and 7, on 04:00 of the 28 March 2016. The visual comparison of the results provides additional information about the features of each method that may not be obvious looking only at the performance indicators. The original radar and OK are compared to the KED results without transformations and with the application of the analyzed methods. In addition, the rain gauge values are superimposed, to understand how well the merging process reproduces the observed rainfall. The example was selected because it presents at one time many characteristics that appear in other cases as well and can be discussed: radar artifacts, evidence of spatially heterogeneous radar bias (the radar overestimates the rainfall intensity in the south part of the image and underestimates it in the north part), and spatial variability of rainfall. In the radar image, some artifacts are evident: it is possible to identify some partial radar beam blockage in the north east part of the domain; there is also high reflectivity values in the south-east part of the image, not observed by the rain gauge network; on the contrary, the intensity in the north part is underestimated; there are patches of lower reflectivity in the north part of the image, and finally it is possible to identify where the acquisitions from the three radars are joined in one composite, resulting in a sharp change of reflectivity along three straight lines cutting the image in three sectors. These effects are partially corrected in most of the KED results.

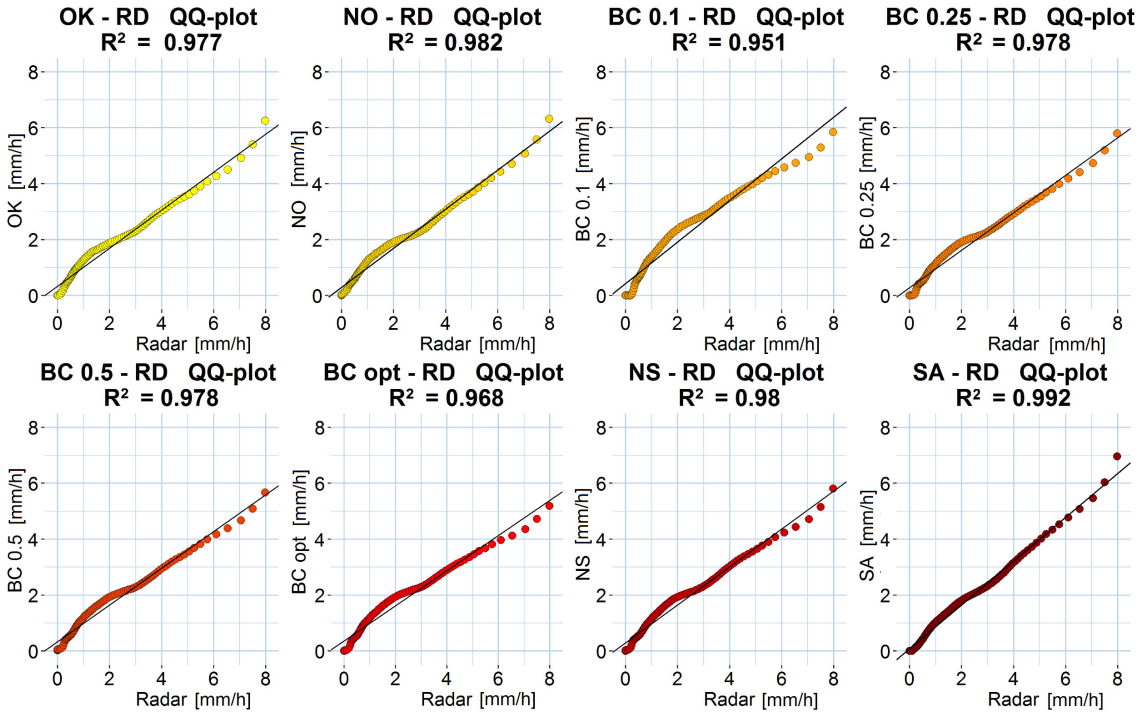


Figure 7: Example of QQ-plots between the merged product distribution and the radar distribution for an example event at 04:00 of the 28 March 2016.

As expected, all the KED rainfall estimates with and without transformations show a better agreement with rain gauges than the radar. The ordinary kriging is designed to show perfect agreement where the rain gauges are available, and shows a smoother behavior elsewhere. However, the high rain gauge density allows OK to reproduce remarkably well the spatial features of the rainfall field. The KED with SA shows enhanced spatial features compared to the rest of the products. BC 0.1 presents higher intensity values, compared to the other products.

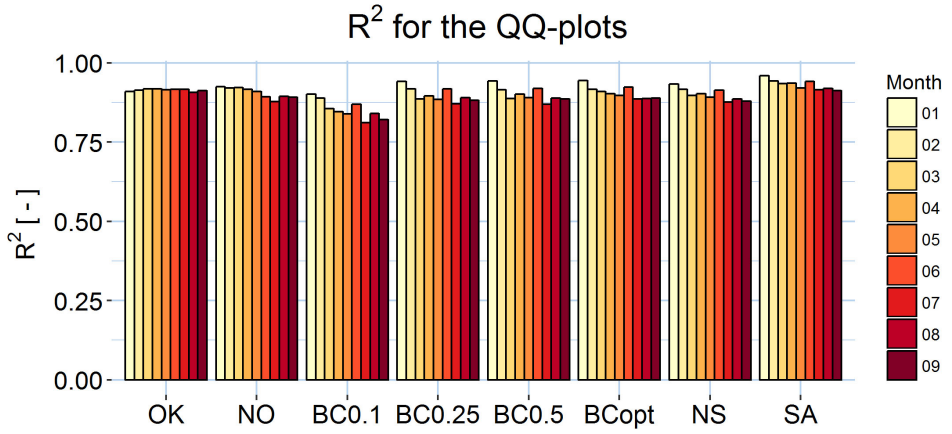


Figure 8: Coefficients of determination R^2 relative to the linear models fitting the QQ-plots. The R^2 is considered an indicator of the ability for each method to reproduce the original rainfall probability distribution. The methods' acronyms are the ones introduced in section 3.

6 Discussion

Figure 4 shows that residuals do have a more Gaussian distribution compared to rainfall, as observed by Kirstetter et al. (2010), but the application of transformations to rainfall can benefit the residuals as well. In particular, for the specific time step, the necessary transformation is well addressed by the optimized BoxCox transformation and the normal score transformation. From the results in section 5.1, showed in Figure 5, the ability of the different methods to transform the rainfall residuals in a Gaussian variable is assessed. All the tested methods improve the Gaussianity of the residuals. However, poor results are achieved by KED with the Singularity Analysis (SA). In the SA, the residuals are calculated between the untransformed rain gauge data and the transformed radar data and this results in a non-Gaussian distribution, which affects all the KED rainfall methods and the quality of the results. The results in section 5.3 do not show a significant improvement in using SA compared to no transformations. For this reason, the SA is considered not appropriate for KED merging, because of the impossibility to apply a transformation to point measurements, and also because the method is computationally very expensive.

As concerns the other tested methods, they all significantly improve the Gaussianity of the residuals. The results confirm that the Box-Cox transformations tend to perform better in transforming rainfall data to Gaussian with a lower λ. Nevertheless, transformations with a low λ present problems in the backtransformation phase, and therefore in the evaluation phases presented in sections 5.2 and 5.3 in this paper. The results are in agreement with the conclusions of Erdin et al. (2012). The normal score transformation, being explicitly designed to produce Gaussian residuals, achieves the best results overall.

The second evaluation stage assesses the ability of the methods to reproduce the original rainfall probability distribution back-transforming the KED merged products. The results are presented in Figure 8, section 5.2. It is remarkable that all the methods perform well, with R^2 values above 0.8, especially OK and KED without transformations. Only the Box-Cox with $\lambda = 0.1$ shows an appreciable lower performance due to the exponential "stretch" in the back transformation. In fact, what is called an introduced bias in Erdin et al. (2012), is actually not uniform on the rainfall intensity scale, and therefore noticeable in the QQ-plot linear fit. In Figure 7, the QQ-plots for an example time step are reported. The plot corresponding to BC 0.1 has the lowest R^2 values and accentuates the deviations from the linear regression. As concerns the other methods, it seems that KED, especially when transformations are applied, is particularly beneficial in winter months

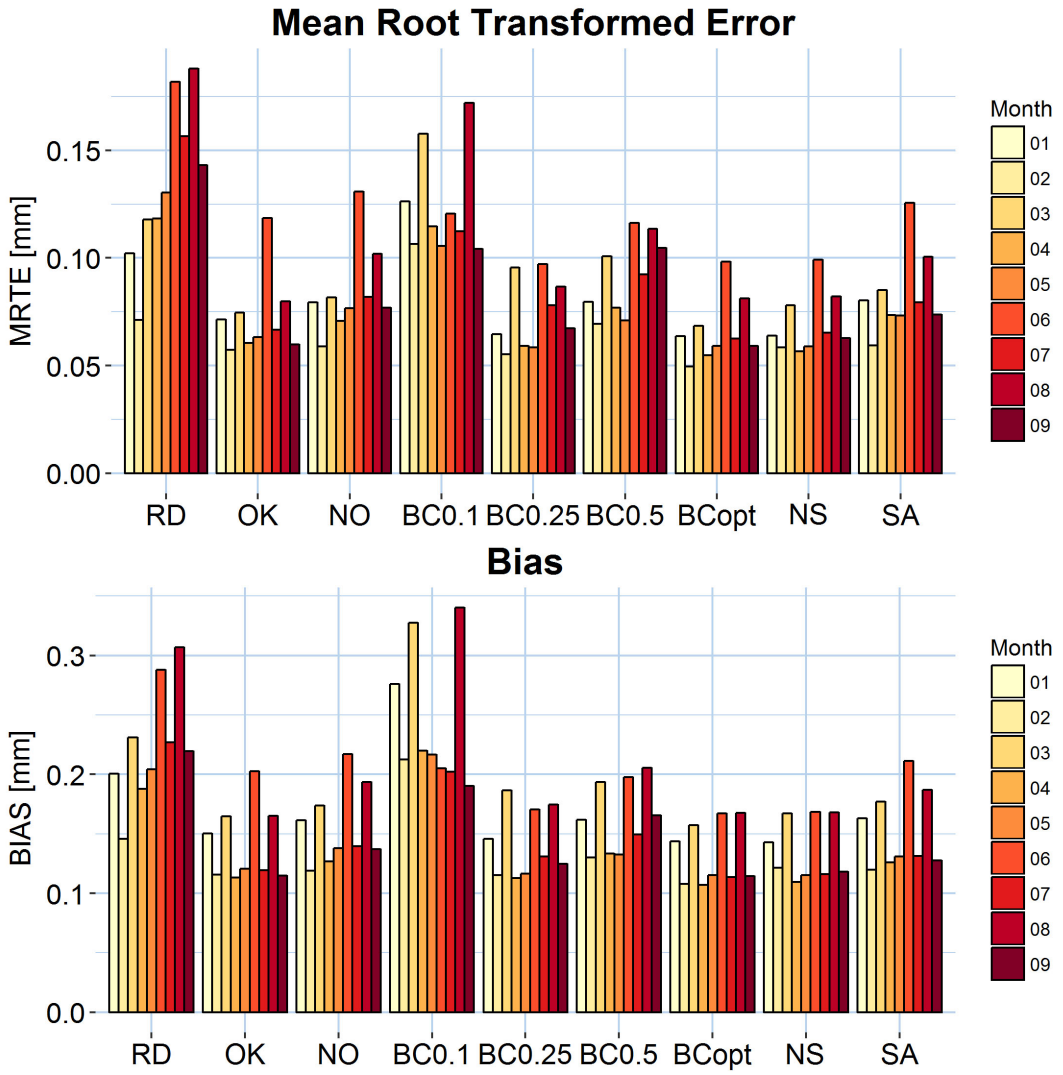


Figure 9: Comparison between the different products, in terms of the quality of the rainfall estimates, based on the cross validation. The skills are measured with two indicators: Mean Root Transformed Error (MRTE) and bias.

rather than in the summer. This is probably due to the fact that during convective events, the spatial variability of rainfall at subpixel scale causes the rain gauges and the radar to disagree often, therefore the KED finds a weaker correlation between the rain gauges and the radar. When this happens, the radar spatial information is not used to improve the estimates correctly. The SA performs well in terms of reproducing the radar probability distribution, being able to reintroduce all the singularities that the KED merging process would remove. Nevertheless, this is not always a positive feature, as illustrated later in this section.

The exponential stretch effect in Box-Cox with $\lambda = 0.1$ is evident in the results of the third evaluation stage too, presented in Figure 9 and section 5.3. The third evaluation stage is the most important, because it assesses the quality of the resulting merged products: the purpose of merging is indeed to improve the quality and the reliability of rainfall estimates, while the probability distributions of the transformed and back-transformed variables are important only if they lead to a higher-quality final product. Indeed, transforming radar and rain gauge data sets with a Box-Cox

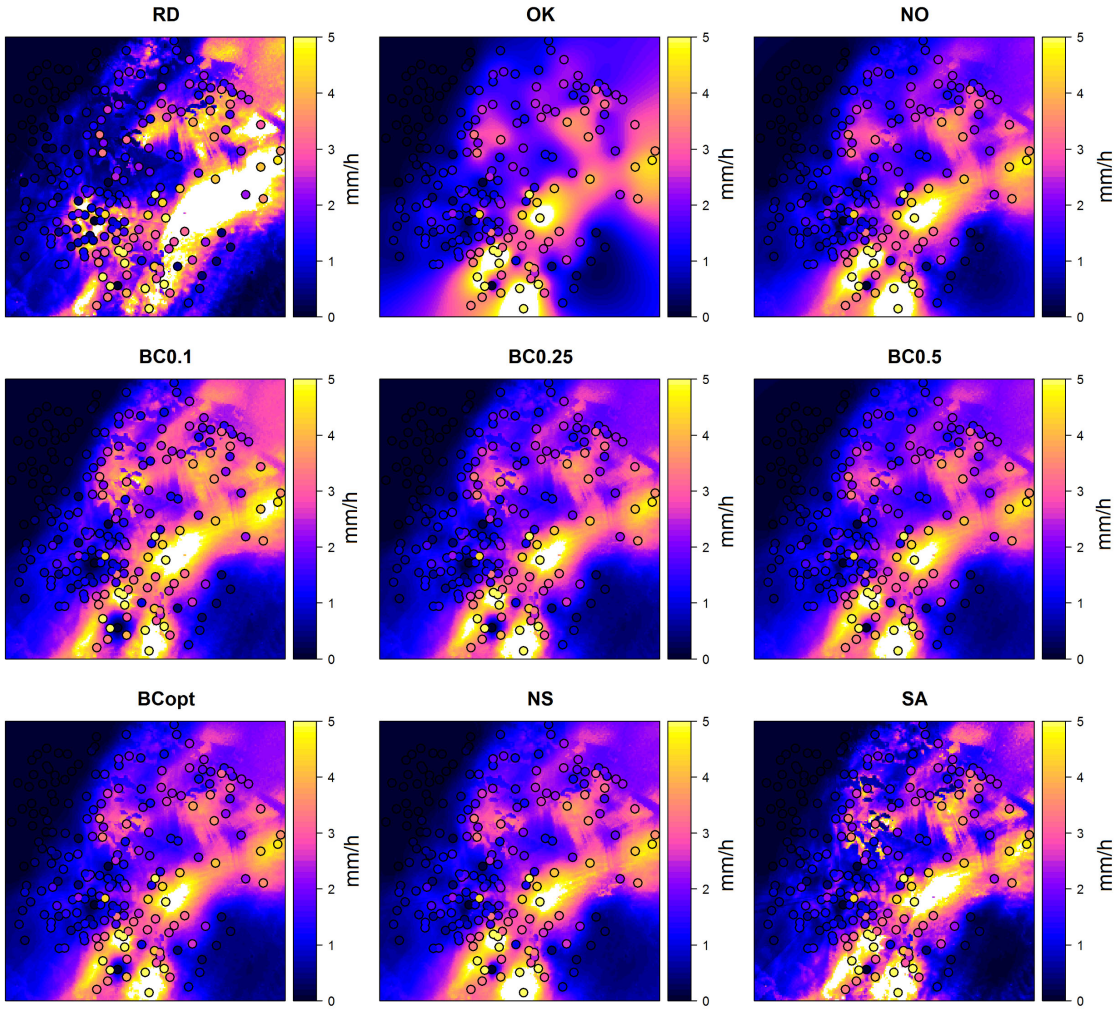


Figure 10: The original radar, the KED product without transformations, and the KED products with different tested transformations are compared for an example hourly time step at 04:00 of the 28 March 2016. The values of the rain gauges are superimposed.

transformation with $\lambda = 0.1$ leads to a good transformation of the variables to Gaussian, but fails in producing a high-quality back-transformed rainfall product, as can be observed in the results of both the second and third evaluation stages; the KED products obtained with the Box-Cox 0.1 transformation shows larger error (MRTE) and larger bias. KED with square root transformation (BC 0.5) or with the Singularity Analysis performs better, but still worse than the other presented methods and do not bring significant advantages compared to the use of KED without transformations. Ordinary Kriging produces outstanding results in this evaluation stage. This may be due to the high-density EA rain gauge network, and also partially due to the rain gauge-based cross-validation method. The KED without transformations performs well, as shown by all the indicators used in the third phase of evaluation (MRTE and BIAS) and it improves the results of the uncorrected radar, which confirms that KED merging, even without transformations, is a positive and reliable practice to improve the quality of radar rainfall estimates. Some transformation methods do improve the results, compared to KED without transformations. The results for the optimized Box-Cox and Box-Cox with $\lambda = 0.25$, show remarkably similar results, with a lower MRTE and a lower bias compared to KED without transformations. In fact, as shown by Figure 6, the optimal λ is close to the 0.2 lower boundary in the majority of the cases. However, it is

important to note that the optimal Box-Cox transformation is very case-specific. In particular, the Box-Cox function has the power to stretch the probability distribution for values below 1 and compress the probability distribution of values above 1. This has two important consequences: (1) the optimal λ is a function of the units and (2) the optimal λ is a function of the temporal accumulation. It is important therefore to be very careful in selecting the optimal Box-Cox transformation for the specific case study. The use of the Box-Cox optimization method is; therefore, recommended over the use of the $\lambda = 0.25$ case, although in this specific case study they work comparably well.

The Normal Score transformation, instead, is not affected by the used units or temporal accumulation. It produces interesting results overall: it shows low MRTE and bias, and performed very well in the transformation and the back-transformation evaluation stages. The qualitative comparison in Figure 10 gives an overview of the methods' characteristics. In particular, it is easy to note the differences between the Box-Cox methods, the Normal Score Transformation, and the Singularity Analysis. As mentioned before, the Singularity Analysis is very good in reconstructing and enhancing the small-scale features. Nevertheless, it is not able to discern between the rainfall field features and the radar artifacts. Methods like the Box-Cox with $\lambda = 0.5$ and $\lambda = 0.25$, and the Normal Score, have a different way to preserve the spatial structures, which can better eliminate the radar artifacts and maintain the rainfall field structures (for example they partially mitigate a lower intensity sector in the top-right corner, or they improve the sharp intensity jump between the different radar composite). The way Box-Cox and Normal Score methods preserve the spatial features is using the transformed radar image as a drift, and adjusting it with the rain gauge observations. The spatial features that have a good agreement with rain gauge measurements are maintained, while the ones that are in disagreement with the ground measurements (like the artifacts) are smoothed. In the singularity analysis, instead, the radar field is smoothed before merging, and all the singularities are reintroduced after, without the possibility to compare them with the ground reference.

7 Conclusion

Kriging merging methods are among the most popular and effective methods to improve radar rainfall estimates and, among them, KED proved to be effective and reliable. Kriging methods are based on the assumption of Gaussianity, and one could argue that their application to rainfall data sets could introduce errors. The objectives of this paper are to understand how significant the uncertainty introduced by the Gaussian approximation is, and what are the best methods to correct it.

Starting from the latter, some methods are identified to be not suitable for KED applications. In particular, Box-Cox transformations with a k close to zero, approximation of the logarithmic transformation, introduces a "stretch" in the back transformed value, resulting in unrealistic high intensities of rainfall. This result is consistent with the results of Erdin et al. (2012). The Singularity Analysis does not produce particularly positive results as well, and two reasons are identified. On the one hand, the method relies on a transformation based on areal rainfall characteristics; since the KED merging scheme uses point rain gauge measurements, the method cannot be applied to correct rain gauge data and is applied only on the radar fields, which are used only as a drift. This may limit its effectiveness in improving the merging process. On the other hand, the transformed radar images that are used to preserve the rainfall spatial characteristics in KED are smoothed by the transformation, and therefore during the merging process; the local singularities, i.e., the spatial details, are then reintroduced only after the merging process. This does not allow to compare the spatial features with the ground measurements and therefore to distinguish between the real rainfall features and the radar artifacts, due for example to radar beam blockage, attenuation, clutter, or radar compositing, which are preserved and even enhanced in the merged product.

It is fair to say that the Met Office is currently upgrading all weather radars in the UK to dual-

polarization weather radars. This will enable better identification of nonmeteorological echoes (Rico-Ramirez & Cluckie, 2008), improvements of attenuation correction algorithms (Bringi et al., 2001; Rico-Ramirez, 2012), and better rainfall estimates in the presence of partial beam blockage (Lang et al., 2009). The SA method could be more effective in the absence of these weather radar artifacts.

The Normal Score Transformation produces overall positive results in all the stages of the analysis. It is effective in transforming the variables to Gaussian and back-transforming the KED products. The final KED rainfall estimates have a low MRTE and a low bias compared to other methods. However, the method is not as simple to implement for rainfall data as it is when applied to different data sets. The application to rainfall data, with a noncontinuous, not strictly increasing cumulative distribution introduces some challenges, and the back transformation requires a linear interpolation/extrapolation for each pixel, for a sufficient number of quantiles (in this case 99), which results in a computationally slow method. Nevertheless, the Normal Score Transformation has the advantage of being independent on the variable values, and therefore independent on the units or on the temporal accumulation of rainfall.

In terms of Box-Cox transformation, instead, it is necessary to be careful in identifying the optimal transformation for the specific case study. In case units in millimeters per hour are used, like in this case, the majority of the rainfall values fall below the threshold of 1 mm/h in most of the considered time steps, therefore the optimal Box-Cox parameter tends to be low. The Box-Cox transformation with $\lambda = 0.25$ produce positive results, under all the examined points of view. It is effective in transforming the rainfall residuals to more Gaussian variables before merging, and to back-transform the merging results to the original distribution; it generates good-quality merged rainfall estimates, with low MRTE and bias. The advantage in selecting this transformation method is that it has an analytical back transformation, although in this work the back transformation is done using the mean of the back transformation of 99 quantiles in order to be consistent with the other studied methods (Sideris et al., 2014).

Therefore, the use of Box-Cox transformation with $\lambda = 0.25$ is usually a good solution, when [mm/h] are used as rainfall units, but this transformation needs to be used carefully, due to its dependency on the rainfall numerical values. An optimization of the λ parameter can be beneficial and the method proposed in this paper is simpler and faster than the Bayesian approach by Erdin et al. (2012). Nevertheless, the Normal Score Transformation is preferable to Box-Cox transformations, being independent on the rainfall values. This work illustrates how to overcome the limits of its applicability only to variables with continuous strictly increasing cumulative distributions.

Finally, as concerns the impact of the Gaussian approximation on the final rainfall estimate uncertainty, the quantitative evaluation shows that the use of a transformation can improve the KED estimates, if the transformation is correctly selected.

Acknowledgement: This work was carried out in the framework of the Marie Skłodowska Curie Initial Training Network QUICS. The QUICS project has received funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under grant 607000. M. A. Rico-Ramirez also acknowledges the support of the Engineering and Physical Sciences Research Council (EPSRC) via grant EP/ I012222/1. The authors would like to thank the UK Met Office and the Environment Agency (environmentagency.gov.uk), which provided the radar rainfall data and the rain gauge data to develop this study. The rain gauge data can be requested at national.requests@environmentagency.gov.uk and the radar data can be downloaded through the British Atmospheric Data Centre (badc.nerc. ac.uk; Met Office, 2003). We are also thankful to Andreas Scheidegger and Jörg Rieckermann, from EAWAG, and Antonio M. Moreno Rodenás and Henry Badger, from TU Delft, for providing technical and scientific feedback. We also thank the reviewers for providing insightful comments that helped to improve the manuscript.

Appendix B

Accounting for variation in rainfall intensity and surface slope in wash-off model calibration and prediction within the Bayesian framework

Manoranjan Muthusamy[1,2], Omar Wani[3,4], Alma Schellart[1], Simon Tait[1]

[1] Department of Civil and Structural Engineering, University of Sheffield, Sheffield, UK
[2] School of Water, Energy and Environment, Cranfield University, Cranfield, UK
[3] Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland
[4] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

Journal: Water Research

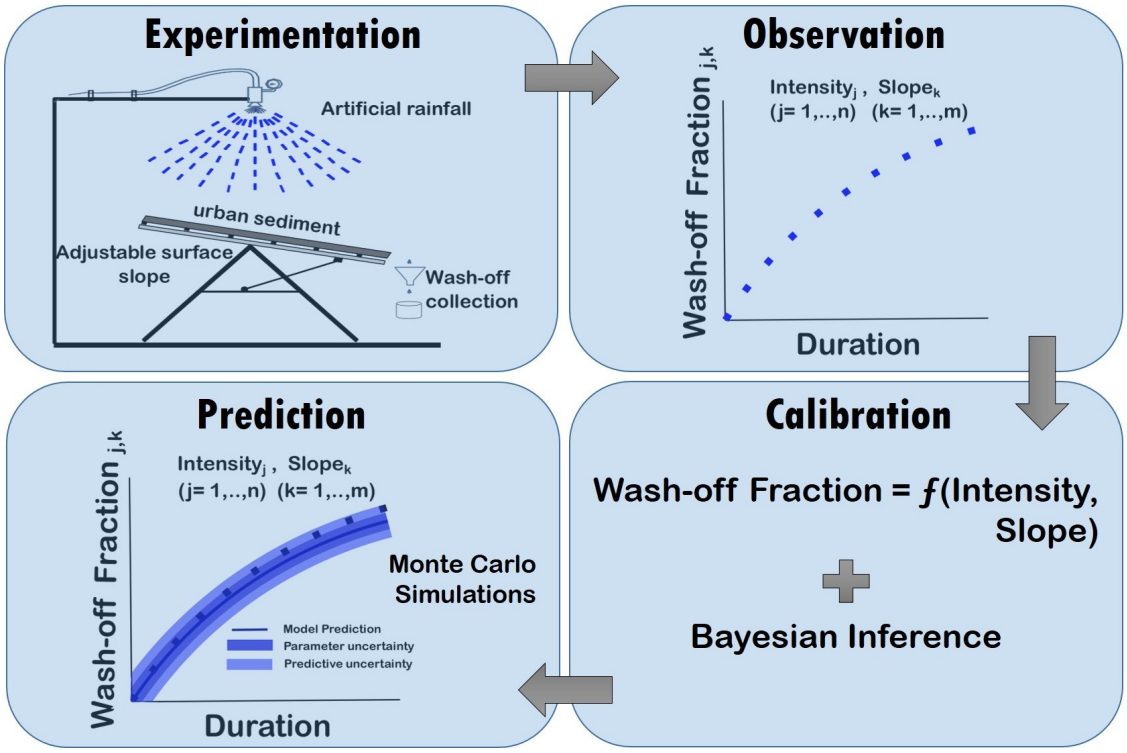
Publication date: 12th June 2018

DOI: https://doi.org/10.1016/j.watres.2018.06.022

Author contributions: MM, with inputs from OW, AS, and ST, designed the new exponential model. OW and MM wrote the code for Bayesian inference and uncertainty analysis. MM conducted the lab and simulation experiments in consultation with ST and AS. MM wrote and revised the manuscript with contributions from OW. AS and ST reviewed and commented on the manuscript and the revision.

Abstract

Exponential wash-off models are the most widely used method to predict sediment wash-off from urban surfaces. In spite of many studies, there is still a lack of knowledge on the effect of external drivers such as rainfall intensity and surface slope on wash-off predictions. In this study, a more physically realistic "structure" is added to the original exponential wash-off model (OEM) by replacing the invariant parameters with functions of rainfall intensity and catchment surface slope, so that the model can better represent catchment and rainfall conditions without the need for lookup tables and interpolation/extrapolation. In the proposed new exponential model (NEM), two such functions are introduced. One function describes the maximum fraction of the initial load that can be washed off by a rainfall event for a given slope and the other function describes the wash-off rate during a rainfall event for a given slope. The parameters of these functions are estimated using data collected from a series of laboratory experiments carried out using an artificial rainfall generator, a 1 m^2 bituminous road surface and a continuous wash-off measuring system. These experimental data contain high temporal resolution measurements of wash-off fractions for combinations of five rainfall intensities ranging from 33 to 155 mm/h and three catchment slopes ranging from 2 to 8%. Bayesian inference, which allows the incorporation of prior knowledge, is implemented to estimate parameter values. Explicitly accounting for model bias and measurement errors, a likelihood function representative of the wash-off process is formulated, and the uncertainty in the prediction of the NEM is quantified. The results of this study show: 1) even when the OEM is calibrated for every experimental condition, the NEM's performance, with parameter values defined by functions, is comparable to the OEM. 2) Verification indices for estimates of uncertainty associated with the NEM suggest that the error model used in this study is able to capture the uncertainty well.



1 Introduction

Urban surface sediment's ability to act as a transport medium to many contaminants makes it one of the major source of pollutants in an urban environment (Collins and Ridgeway, 1980; Guy, 1970; Lawler et al., 2006; Mitchell et al., 2001). Hence there is an increasing interest in being able to better predict the sediment wash-off from urban surfaces. But, modelling sediment wash-off is not a straightforward exercise as it requires the understanding of complex interactions between external drivers with a highly variable nature such as rainfall, catchment surfaces and particle characteristics (Deletic et al., 1997; Egodawatta and Goonetilleke, 2008; Sartor and Boyd, 1972). Currently, the most widely used wash-off models are originally developed using laboratory experiments and consequently include empirical parameters without clear physical interpretations. The exponential wash-off equation (Eq. (1)) proposed by Sartor and Boyd (1972) is one such model whose performance is highly dependent on the accurate estimation of parameter k:

$$w_t = w_0\left(1 - e^{-kR_t}\right) \tag{1}$$

Where w_t is the total transported sediment load up to time t; w_0 is initial load of sediment on the catchment surface; R_t is cumulative rainfall depth at time t, i.e. $i_t t$ where i_t is average rainfall intensity over time t; and k is an empirical wash-off coefficient.

Equation (1) has widely been used in several software packages (e.g. SWMM) with or without modifications (e.g. Zug et al., 1999; Huber and Dickinson, 1992). Since rainfall is the main driver of the wash-off process (Deletic et al., 1997; Egodawatta et al., 2007; Sartor and Boyd, 1972; Shaw et al., 2010), understandably most of these modifications are focused on the effect of rainfall. Recently, Egodawatta et al. (2007) suggested an introduction of a 'capacity factor' which gives a more physical interpretation to the empirically calibrated original model shown in Eq. (1). According to Eq. (1), if the rainfall continues for long enough, it can wash off all the sediment available at the beginning of the event regardless of the rainfall intensity of the event. In other words, the maximum washoff fraction (w_t/w_0) is always one. But Egodawatta et al. (2007) showed that a storm event has the capacity to wash-off only a fraction of sediments available and once this maximum fraction is reached the wash-off becomes almost zero, even though a significant fraction of sediment is still available on the surface. They suggested the introduction of an additional term referred to as the 'capacity factor' (C_F) to replicate this finding in the model equation as shown is Eq. (2)

$$\frac{w_t}{w_0} = C_F\left(1 - e^{-kR_t}\right) \tag{2}$$

Although the above modification was shown to be a meaningful refinement, C_F was investigated against rainfall intensity in isolation in Egodawatta et al. (2007). Muthusamy et al. (2018) further showed that C_F also varies with catchment surface slope in addition to rainfall intensity. Despite surface slope's direct impact on main underlying processes of sediment wash-off which are impact energy from rainfall drops (Coleman, 1993) and shear stress from overland flow (Akan, 1987; Deletic et al., 1997), there is a clear lack of attention given to surface slope in the literature. Results from Muthusamy et al. (2018) showed that the surface slope has a significant effect on the wash-off load and this effect should not be neglected in the prediction of wash-off.

In spite of the modifications suggested by various studies including Egodawatta et al. (2007) and Muthusamy et al. (2018), the calibration parameters k and the newly introduced C_F still need to be calibrated for the conditions of each catchment. In general, this is achieved by using a combination of look up tables/charts and interpolation/extrapolation of existing data. However, with the absence of such commonly accepted look up tables/charts, the modellers are forced to use a constant value for parameters regardless of catchment conditions. This calls for an alternative and a more transparent way of estimating the calibration parameters.

Furthermore, none of the abovementioned studies include any information on the uncertainty in the estimation of the calibration parameters and their dependency structure which needs to be accounted in the prediction of wash-off using these parameters. Although adequate treatment of propagation of uncertainties in model prediction is a currently heavily researched area in hydrology, there are only a few studies on uncertainty related to wash-off modelling (e.g. Sage et al., 2016; Dotto et al., 2012). In this regard, Dotto et al. (2012) compared a number of uncertainty techniques applied in urban water stormwater quality modelling and found that a Bayesian approach, although computationally demanding, to be one of the preferable uncertainty assessment technique. A Bayesian approach helps to identify different sources of uncertainty such as parameter uncertainty, model bias and measurement noise and consequently, helps to separately analyse them, though this requires knowledge about the error process (Dotto et al., 2012). In this regard, Sage et al. (2016) discussed the consequences of using a wrong error model in the prediction of uncertainty in wash-off modelling and called for more attention to be paid for the selection of error model.

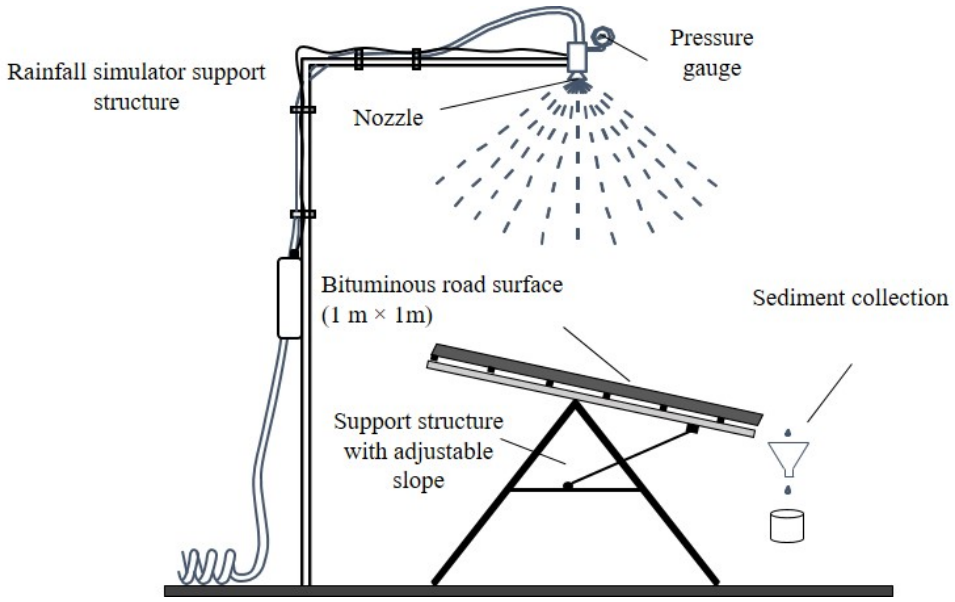


Figure 1: Sketch of the experimental setup.

Considering the above research gaps in the current modelling approach of sediment wash-off, this study aims:
a) To add a more physically realistic "structure" to Eq. (2) by replacing the calibration parameters with functions of external drivers associated with catchment surface and rainfall characteristics and compare its performance with the original model.
b) To identify different sources of uncertainty associated with the new wash-off model developed in (a) and estimate reliable prediction intervals using a suitable error model.

2 Material and methods

2.1 Wash-off data

Data used in this study were collected from a series of laboratory experiments carried out using an artificial rainfall generator, a $1m^2$ bituminous road surface and a continuous wash-off measuring system (Fig. 1). This data contain sediment wash-off data measured against different combinations of rainfall intensity, catchment surface slope and initial sediment load. The road surface was prepared using bituminous asphalt concrete and had a mean texture depth index of 0.4 mm.

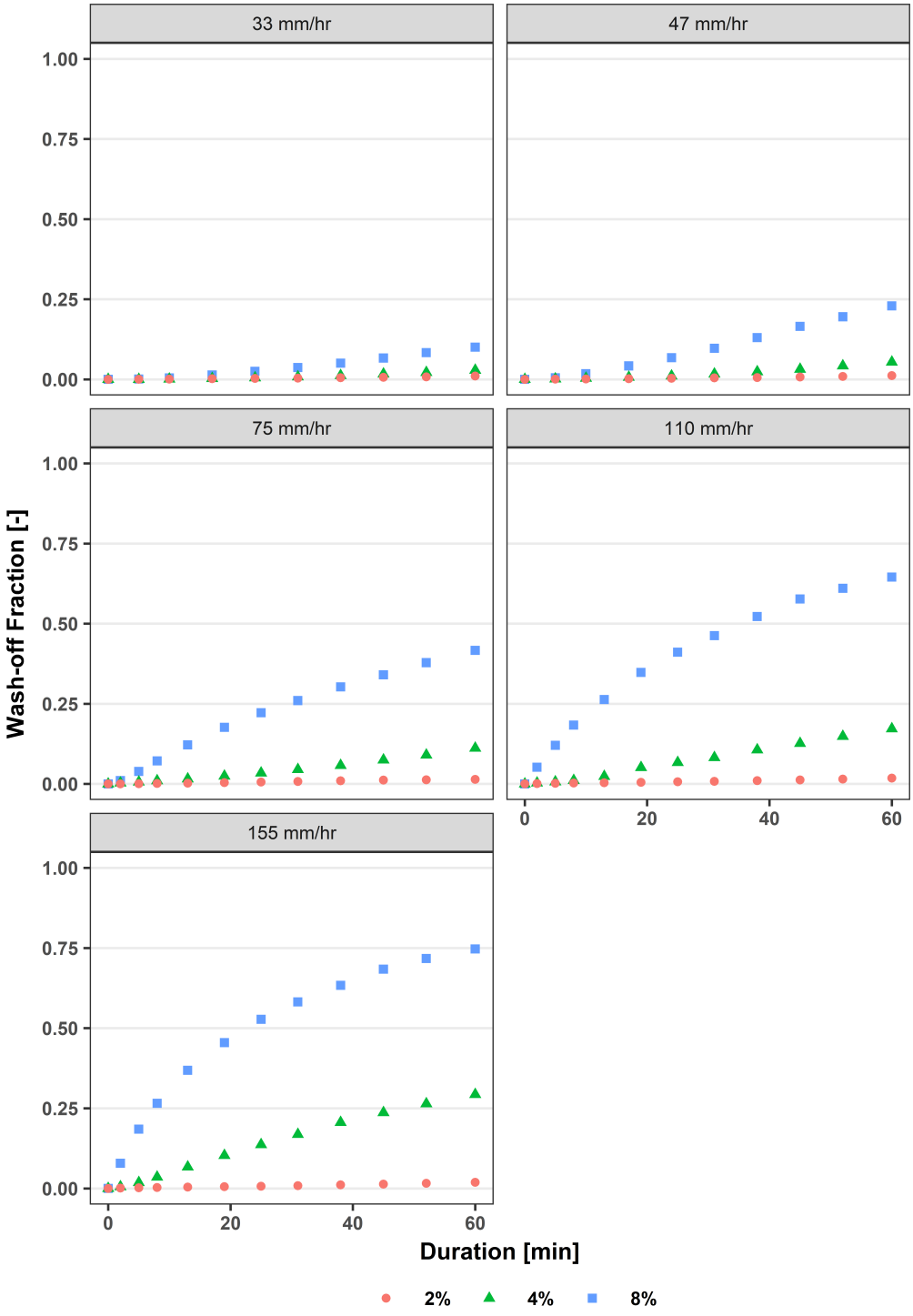


Figure 2: Selected results from Muthusamy et al. (2018): Variation of wash-off fraction for different combinations of rainfall intensity and surface slope.

The D_{10}, D_{50} and D_{90} of the sand used in the experiment are 300 μm, 450 μm and 600 μm respectively. Five intensities ranging from 33 to 155 mm/h, four slopes ranging from 2 to 16% and three initial loads ranging from 50 to 200 g/m^2 were tested in these experiments. For more details on the experimental setup, selection of experimental conditions and data collection the readers are

referred to Muthusamy et al. (2018). As reported in Muthusamy et al. (2018) the effect of initial load on wash-off process was found to be negligible. Hence in this study, experimental results from a constant initial load of 200 g/m^2 as presented in Fig. 2 were used. This figure shows the variation of cumulative wash-off fraction ($F_w = w_t/w_0$) against rainfall intensity and surface slope.

Note that the 16% slope was eliminated from the data, given that such slopes on road surfaces are extreme scenarios and exist only in rare locations. For example, the Department of Transport in the UK suggests a maximum gradient of 10% for roads other than in exceptional circumstances (Manual for Streets, 2009). Since one of the aims of the study is to develop a single model with a fixed set of parameters, the inclusion of results from such an extreme scenario in the calibration may affect the performance of the model for more general cases.

2.2 The modified wash-off model structure and its rationale

The main objective is to replace the calibration parameters in Eq. (2) with functions of surface slope and rainfall intensity, consequently adding a more physically realistic structure to the model. This should make the model robust to new combinations of rainfall intensity and surface slope. To do so, the properties of the model that are sensitive to such parameters need to be identified and understood. From Eq. (2) there are two parameters which define the characteristics of a wash-off curve. The first parameter, C_F, defines the highest wash-off fraction for a given combination of rainfall intensity and a slope. The second parameter, k, defines how fast the wash-off curve reaches the maximum fraction for a given surface slope and rainfall intensity, and hence reflects the erosion rate from the catchment surface. Hence, C_F and k were proposed to be replaced with functions of surface slope and rainfall intensity, as shown in Eq. (3) and (4).

$$C_F = c_1 i_m^{c_2} s^{c_3} \tag{3}$$

$$k = c_4 s \tag{4}$$

Where $c_1, .., c_4$ are constants,[1] i_m is the representative rainfall intensity of a rainfall event (e.g. in this case the constant rainfall intensity set during the experiment, please refer to section 3.4 for discussion on the use of representative rainfall intensities), s is the catchment surface slope. The following criteria were considered when defining Eq. (3) and (4), while also trying to keep these functions as simple as possible to reduce the number of constants:

- C_F - as explained before C_F is a capacity factor which defines the maximum fraction from the initially available sediment that can ever be washed off from a rainfall event for a given slope. Hence, C_F ranges from 0 to 1 and increases with both surface slope and (representative) rainfall intensity of the event. When either of the representative rainfall intensity or slope is zero C_F is zero.
- k - k defines the wash-off rate and it also increases with rainfall intensity and surface slope. But it should be noted that R_t in the exponential term is cumulative rainfall depth at time t, i.e. $i_t t$ which is already a function of average rainfall intensity over time t, i_t . Hence k was taken as a (linear) function of slope only. The complete exponential term reads as $c_4 s i_t t$ which is function of both rainfall intensity and surface slope.

Hereafter this new exponential model will be referred to as NEM and the original exponential model as shown in Eq. (1) will be referred to as OEM.

2.3 Estimation of model parameters and associated uncertainty

Bayesian inference was used to estimate the parameter probability distribution, which allows prior knowledge on the parameters to be incorporated in the estimation and also formally quantifies

uncertainty in the estimation (Dotto et al., 2012; Freni and Mannina, 2010; Del Giudice et al., 2013). In addition, it also helps to capture the dependence structure between parameters (Dotto et al., 2012). Bayesian inference requires the definition of the likelihood function and the prior distribution of the parameters.

The likelihood function

In addition to finding the best estimate of the parameters, we are also interested in the uncertainty associated with the parameter estimation and consequently the uncertainty in the prediction of the wash-off fraction. One way of doing this is to include the error terms which represent the dominant sources of uncertainty explicitly in the likelihood function.We used an error model which accounts for errors due to the model structural deficit (model bias, $\boldsymbol{B_M}$) and measurement noise ($\boldsymbol{E}$). $\boldsymbol{B_M}$ is modelled as an autoregressive stationary random process and $\boldsymbol{E}$ modelled as an independent identically distributed (IID) normal noise. Hence, an observed output, $\boldsymbol{Y_o}$ can be formulated as

$$Y_o(x, \theta, \psi) = y_m(x, \theta) + B_M(x, \psi) + E(\psi) \tag{5}$$

Where x is the external drivers, $\boldsymbol{\theta}$ is deterministic model parameters, $\boldsymbol{\psi}$ error model parameters and $\boldsymbol{y_m}(\boldsymbol{x}, \boldsymbol{\theta})$ is deterministic model output. In this case, $\boldsymbol{Y_o}$ is observed wash-off fractions (F_w) and $\boldsymbol{y_m}$ is the deterministic model output predicted from the NEM (f_w). $\boldsymbol{x}$ represents rainfall intensity and surface slope. $\boldsymbol{\theta}$ represents parameters $c_1, ..., c_4$. $\boldsymbol{\psi}$ represents error model parameters $sd.B$, $sd.E$ and l in which $sd.B$ and l are standard deviation and the correlation length respectively that characterise the autoregressive stationary random process and, $sd.E$ is the standard deviation of the measurement noise. Given the error description of Eq. (5), we define $\boldsymbol{B_M}(\boldsymbol{x}, \boldsymbol{\psi})$ as a multivariate Gaussian distribution with covariance matrix $\Sigma(\boldsymbol{x}, \boldsymbol{\psi})$ and $\boldsymbol{E}(\boldsymbol{\psi})$ as independent, identical normal noise. Therefore, the analytic formulation of the likelihood function with n number of observations can be formulated as

$$\boldsymbol{p(y_o|x, \theta, \psi) = \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{\det(\sum(x, \psi))}} \exp\left(-\frac{1}{2}[y_o - y_m(x, \theta)]^T \sum(x, \psi)^{-1}[y_o - y_m(x, \theta)]\right)} \tag{6}$$

The covariance matrix $\Sigma(\boldsymbol{x}, \boldsymbol{\psi})$ was formulated using an OU process (Uhlenbeck and Ornstein, 1930). For hydrological applications, OU process is a simple description of the underlying mechanisms with an exponential decay of correlation in time (Del Giudice et al., 2013; Sikorska et al., 2012; Yang et al., 2007). A detailed description of the formulation of covariance matrix using OU process can be found in Del Giudice et al. (2013). An autoregressive error model represents model structural deficit better than IID as it accounts for the "memory" in the error time series (Del Giudice et al., 2013). This autoregressive bias error model was originally suggested in other generic statistical applications (Bayarri et al., 2007; Craig et al., 2001; Higdon et al., 2004; Kennedy and O'Hagan, 2001) and later adapted for environmental engineering applications (Reichert and Schuwirth, 2012).

Table 1: k values from the literature.

Reference	Land use/catchment type	Value k (mm^{-1})
Alley (1981)	Urban catchment	0.036-0.43
Nakamura (1984)	Various	0.05-10
Nakamura (1984)	Various	0.05-10
Huber and Dickenson (1992)	General	0.04-0.4
Miller (1999)	Residential	0.21
Egodawatta (2007)	Concrete and asphalt roads	5.6×10^{-4}-8.0×10^{-4}

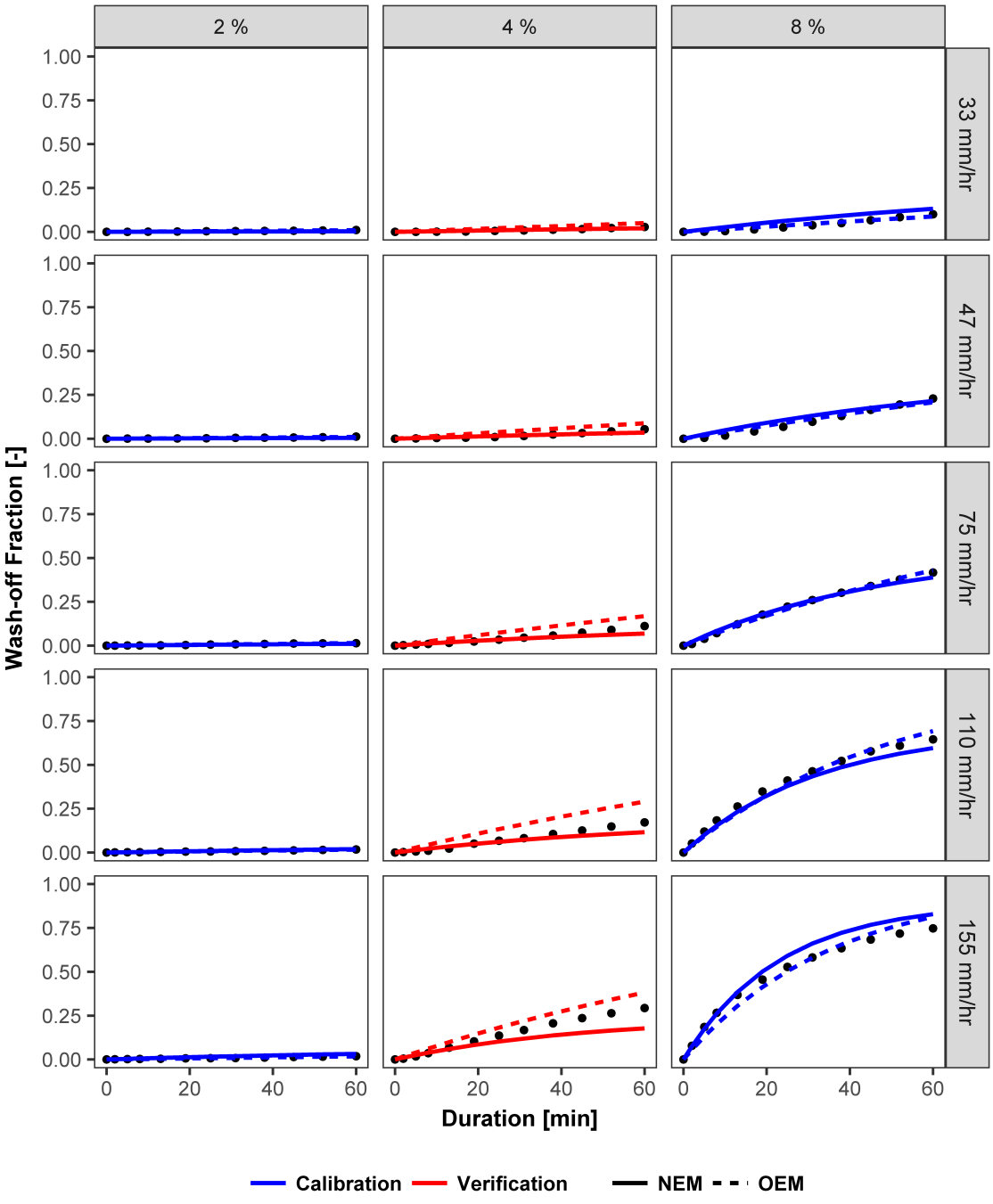


Figure 3: Comparison of the model performance.

Prior distribution of parameters and constraints

Since the introduced parameters $c_1, ..., c_4$ are all new, there is no previous estimation of the exact parameters, but a possible range for some parameters can be derived using our knowledge of the wash-off process, observational data, and the prior belief about values of C_F and k.

Values of c_4 were derived from previous estimations of k as c_4 equals to k/s. The list of k values derived from previous studies is given in Table 1. From the table, the range of 0-10 were selected

for k. In the absence of any information on slope in most of these studies, same range for c_4 was used considering a minimum slope of 1%. Hence a uniform prior with the range 0-1000 was used as a prior distribution for c_4. A uniform prior distribution of model parameters can be used when there is not enough evidence available to choose a different type of distribution (Dotto et al., 2012; Freni and Mannina, 2010).

As discussed previously, the range of C_F is 0-1 because wash off fraction cannot be more than 1. This leads to the constraint $0 \leq c_1 i_m s^{c_3} \leq 1$. The implication of this constraint in the definition of prior probability is not straightforward as it involves three parameters, hence this constraint was used during the estimation of likelihood probability.

It is challenging to define prior distributions for the error model parameters ($sd.B, sd.E$ and l) especially in the case of wash-off modelling as examples from such applications in literature are currently lacking. Out of the three parameters, some information on the measurement noise represented by $sd.E$ can be obtained by frequent tests, i.e. repeating the experiments sufficiently large number of times. However, this is not always possible given the limitation in allocated resources and time. In the absence of much information on any of the error parameters, a uniform prior with the range from 0 to 1(= maximum wash-off fraction) was used for both $sd.B$, $sd.E$ and a uniform prior with the range of 0-200 min was used for correlation length. This range was selected as error correlation is expected to be insignificant beyond such time length.

Bayesian inference

Once the prior distributions (the probability of deterministic and error model parameter, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ without considering the observed output, $\boldsymbol{y_o}$), $p(\boldsymbol{\theta}, \boldsymbol{\psi})$ and the likelihood function (the probability of seeing the observed output, $\boldsymbol{y_o}$, as generated by a model with deterministic and error model parameter $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$), $p(\boldsymbol{y_o}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\psi})$, are defined, the posterior distribution of the deterministic and error model parameters (the conditional probability of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ once the observed output, $\boldsymbol{y_o}$ has been taken into account) can be formulated as,

$$\boldsymbol{p(\theta, \psi|y_o, x)} = \frac{\boldsymbol{p(y_o|x, \theta, \psi)p(\theta, \psi)}}{\int \boldsymbol{p(y_o|x, \theta, \psi)p(\theta, \psi)d\theta d\psi}} \tag{7}$$

Since the direct analytical calculation of $p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y_o}, \boldsymbol{x})$ is generally not possible, numerical techniques such as Markov Chain Monte Carlo (MCMC) simulations have to be applied to generate samples for this distribution. MCMC techniques generate a random walk through the parameter space which will converge to the posterior distribution. In this study, we used robust adaptive Metropolis MCMC sampler presented in Vihola (2012) which is implemented in an R package, *adaptMCMC* (Scheidegger, 2017).
Experimental data with 2% and 8% slopes (two-thirds of the total data) were used for calibration of NEM and the data from the 4% slope (one-third of the total data) were used for verification. The optimal value of each parameter $c_1 ... c_4$ obtained during the calibration stage was then used for validation. Furthermore, the performance of the NEM was compared against the OEM during both calibration and validation stages. In the case of the OEM, the k value was calibrated for each and every combination of surface slope and rainfall intensity during the calibration stage. Linear interpolation of these calibrated k values was then used to obtain new k values during the validation stage for a new surface slope condition.

In addition to deterministic prediction, prediction uncertainty of the NEM was also obtained during both calibration and validation stages. Parameter uncertainty and total predictive uncertainty (parameter uncertainty + model bias + measurement noise) were predicted by sampling from posterior multivariate distributions of parameters $c_1, ... c_4$. Parameter uncertainty was estimated by using deterministic model ($\boldsymbol{y_m}(x, \theta)$) runs and predictive uncertainty was estimated by using the deterministic model together with error model components.

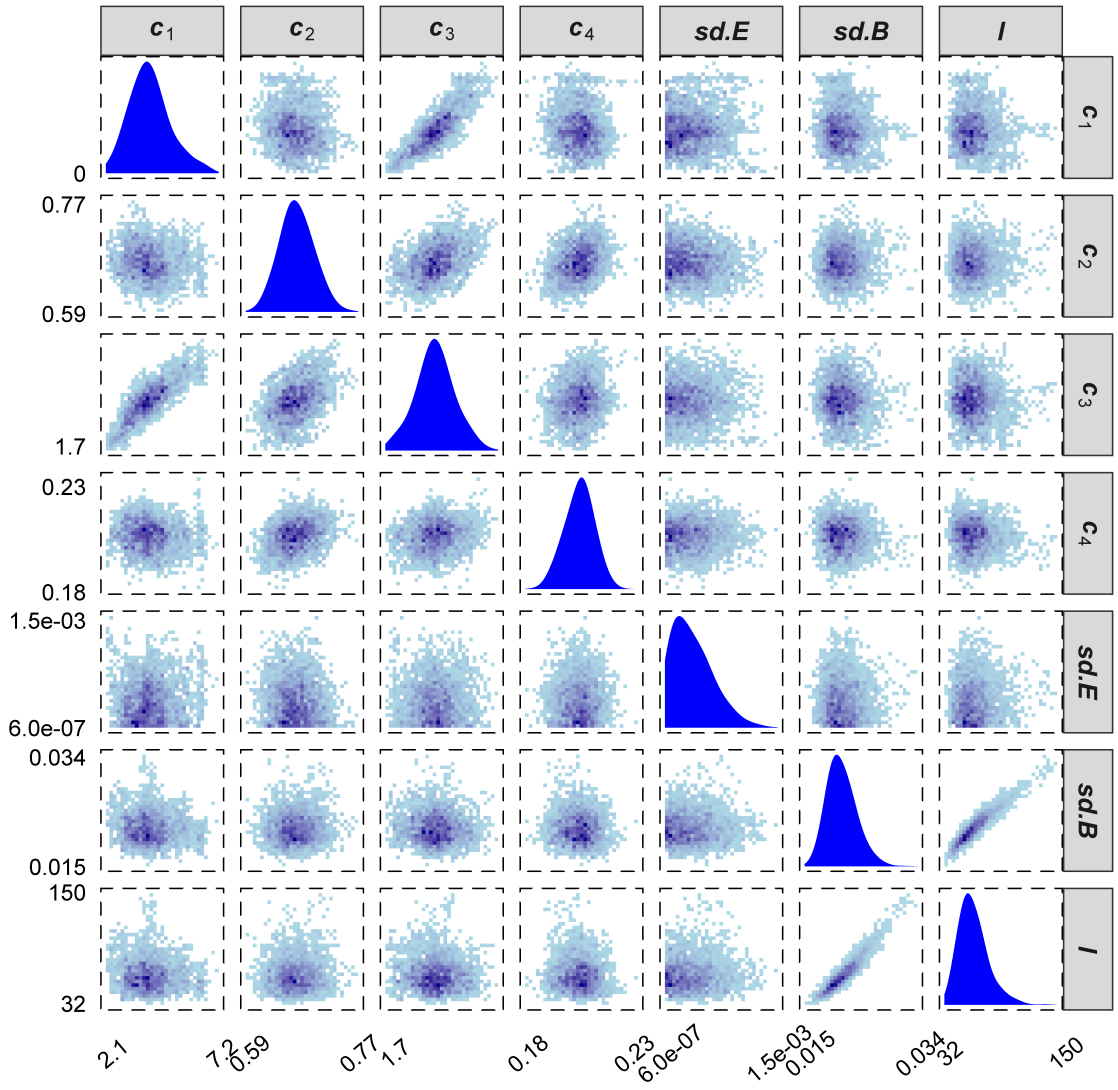


Figure 4: Parameter distribution and bivariate correlation.

3 Results and discussion

3.1 Model performance

Figure 3 shows the model output with the optimal values for $c_1, ..., c_4$ (Table 2) with maximum posterior probability density, i.e. the most probable values given the prior and the observed data. It can be seen from Fig. 3 that with calibration data, the NEM with fixed values of parameters $c_1, ..., c_4$, corresponding to the maximum posterior probability density, performs as well as the OEM which was calibrated for each and every combination of surface slope and rainfall intensity separately. From Table 3, it can be seen that the difference in sum of root mean square error ($\text{RMSE}_{\text{OEM}} - \text{RMSE}_{\text{NEM}}$) from the ten calibrated set of data is -0.07 (Wash-off fraction). However, the robustness of the NEM over the OEM can be seen during the verification stage where the NEM performs better than the OEM in several cases. The difference in sum of root mean square error ($\text{RMSE}_{\text{OEM}} - \text{RMSE}_{\text{NEM}}$) from 5 sets of data during verification stage is 0.09 (Wash-off fraction). The drawback with the OEM is that for a set of new catchment conditions where the OEM has

not been calibrated before, k value needs to be calculated using interpolation/ extrapolation. This might lead to the underperformance of the OEM during validation stage as shown in Fig. 3. Considering the overall performance, the NEM with only 4 parameters $(c_1, ..., c_4)$ performs better than the OEM with 15 parameters $(k_1, ..., k_15)$. Hence, the NEM does not only avoid the need of interpolation to predict the calibration parameter values, it also performs as well as the calibrated OEM.

Table 2: Optimal values of constants of Eq. (3) and (4).

c_1	c_2	c_3	c_4
3.99	0.672	1.99	0.208

Table 3: Performance of OEM and NEM.

Model	Sum of root mean square error (RMSE)	
	Calibration	Verification
OEM	0.11	0.20
NEM	0.18	0.11

3.2 Parameter distribution and correlation

This section discusses the posterior distribution of parameters and their multivariate behaviour. Figure 4 shows posterior distributions and a bivariate matrix of the deterministic and error model parameters. The most likely value of $sd.B$ and $sd.E$ are 0.02 (2%) and 0.002 (0.2%) respectively, showing that most of the uncertainty in the wash-off estimation can be explained by the model bias and that uncertainty due to measurement noise is negligible. Although these are approximate representations of the actual system and corresponding uncertainty, we believe that the experiments were conducted with as high a quality as possible. This is one of the reason why a road surface as small as 1 sq.m was selected as it gives a better control over the experimental set-up. For example, the smaller surface area keeps the spatial variability of the rainfall to the minimum. Furthermore, it also keeps the sediment loss during the experiment to insignificant. The maximum sediment loss observed during an experiment was less than 2% which is an indication of the good quality control.

3.3 Performance assessment

Looking at the bivariate plots, there is a strong positive correlation between parameters c_1 and c_3 which indicates that these two parameters compensate each other in order to maximise the posterior probability. This can also be seen between parameters c_2 and c_3, but to a lesser extent. Similarly, the strong positive correlation between $sd.B$ and l means that these parameters compensate each other in order to fit the autoregressive error model B_M. Bayesian inference helps resolve such identifiability issues by allowing for informative priors. Therefore, for real cases, where we have reasons to believe that one of the two parameters should be more constrained, the other parameter value will automatically come out to be constrained after joint inference.

3.4 Estimation of parameter and predictive uncertainty

Figure 5 shows the uncertainty associated with the estimation of the wash-off fraction. Parameter uncertainty was estimated by using deterministic model $(\boldsymbol{y_m}(x, \theta))$ runs and predictive uncertainty was estimated by using the deterministic model together with error model components. Since the latter also includes the uncertainty due to model bias and measurement noise these bands are wider than the parameter uncertainty. The total predictive uncertainty which accounts for parameter

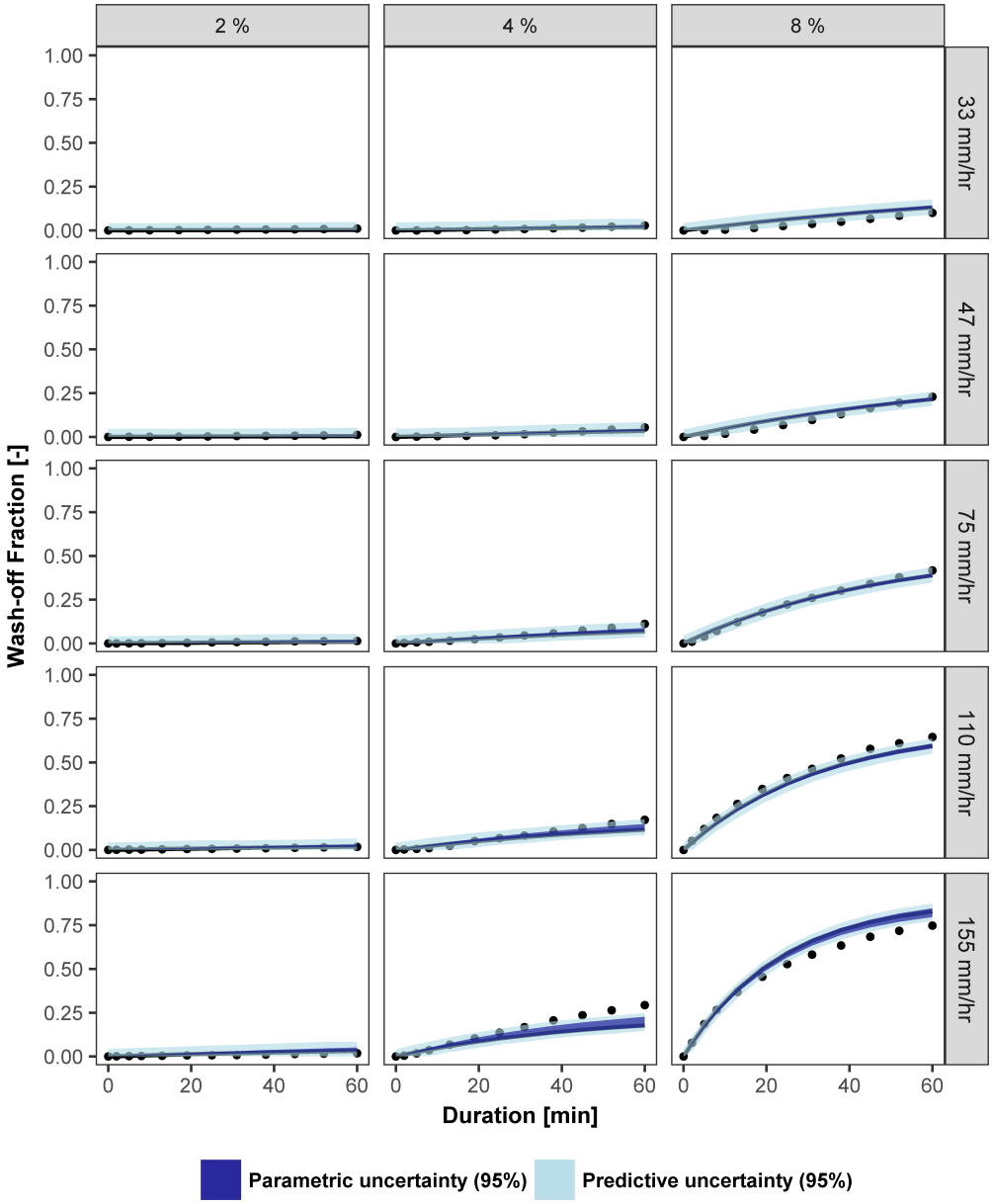


Figure 5: Uncertainty associated with the prediction of wash-off fraction using the NEM.

uncertainty, model bias and measurement noise accounts for ∼0.1 (10%) uncertainty in the wash-off fraction. This constant trend of predictive uncertainty is a reflection of the fact that the error model used here is not explicitly input-dependent bias model, but rather it is a constant variance model. On the other hand, parameter uncertainty increases with increasing wash-off fraction as the variance of parameter uncertainty proportionally increases with mean prediction. The parameter uncertainty accounts for a maximum of 0.06 (6%) wash-off fraction when 95% predictive interval is considered.

To check the reliability of the uncertainty estimation, prediction interval coverage probability (PICP, Ref Eq. (8)) which measures the probability that the observed values lie within the esti-

mated prediction intervals (Shrestha and Solomatine, 2006) was used.

$$PICP = \frac{1}{n}\sum_{i=1}^{n} R \cdot 100\% \quad where\, R \begin{cases} 1, & PL_t^u \leq O_t \leq PL_t^l \\ 0, & otherwise \end{cases} \tag{8}$$

Where, PL_t^u and PL_t^l are upper and lower boundary of the considered prediction interval at time t for a given slope and rainfall intensity, O_t is corresponding measured wash-off fraction at time t. For a better performance, PICP should be close to the considered prediction interval, which is 95% in this case. The calculated PICP during validation stage is 82%, so the corresponding accuracy of the uncertainty estimation is around ∼85% which essentially means that the error model is able to predict the uncertainty reasonably well.

3.5 General discussion

IID is the most commonly used form of error model in urban hydrology (Breinholt et al., 2012; Dotto et al., 2011; Freni et al., 2009; Sage et al., 2015) mainly because of its simplicity. However, it requires the absence of a serial correlation in the error distribution, which can lead to underestimation of uncertainty and biased parameter estimates (Del Giudice et al., 2013). Error process of hydrological phenomena, such as sediment wash-off, are shown to be temporally autocorrelated and assumption of independence is not satisfied (Schoups and Vrugt, 2010; Sage et al., 2016). A likelihood function based on an uncorrelated error model generally leads to narrower posterior probability densities, which results in overconfident parameter estimates and unreliable uncertainty intervals. An autoregressive model helps in preventing such biases both during inference and prediction. Unlike IID, where each data point is a sample of the error distribution, an autoregressive process takes the whole time series of errors as one sample realisation of the process, (in an n-dimensional space), therefore avoiding overconfidence in parameter estimation. For example, Sage et al. (2015) acknowledged that their assumption to model error process associated with wash-off modelling as IID was found to be invalid. Further, Sage et al. (2016) showed that the use of IID to represent the structural deficit of sediment wash-off models violates the statistical properties of the structural deficit and it may result in unreliable estimation of model parameters and total predictive uncertainty. An autoregressive model accounts for this autocorrelation of the error. Hence, it represents the structural deficit better.

This error model can be further improved by accounting for non-normality of the structural bias. However, this adds more complexity. Such added complexity could be an acceptable compromise when there is a very large number of data points to learn about the error model parameters. In our case, the current description of errors seems adequate as suggested by verification measures that show around 85% accuracy of the error models in capturing the uncertainty. Further, we assumed a constant bias to keep this autoregressive error model simple. Nevertheless, it is also possible to describe it as an input - dependent bias (Del Giudice et al., 2013) where bias can be a function of both slope and intensity. The advantage of such bias description still needs to be investigated in the uncertainty analysis of wash-off modelling.

Note that in addition to rainfall intensity and surface slope, other parameters such as sediment size and surface texture will also affect the sediment wash-off, but due to the limitations in the data used in this study, the NEM does not include the effect of these parameters. With smaller sediment sizes and smoother surfaces, the wash-off is expected to be higher. For example, Egodawatta et al. (2007) in a similar experimental study used a larger range (0-1000 μm) sediment resulting in a relatively higher wash-off fraction. Further, Hong et al. (2016) in their studies used a sediment range of 0-400 μm and showed that most (>90%) of the finest particles are removed at the beginning of a rainfall event, with about 10%-20% of medium-size particles are removed over the later part of the even. These studies show that selection of sediment size affects the sediment wash-off process significantly. Hence, the application of the NEM needs to be checked against different sediment sizes and also against different surface textures. It is expected that the values of

$c1, ...c4$ will be different for different particle size distribution of the road sediment and/or different surface roughness. The inclusion of the effect of these parameters explicitly might introduce more complexity in the equation. Nevertheless, such an equation can be applied globally regardless of individual catchment conditions. This is one of the research areas in sediment wash-off modelling that requires to be investigated in detail.

While experimental set-ups like the one used in this study give great flexibility to replicate the real hydrological processes such as sediment wash-off, there are still some limitations which need to be taken into account. The exponential wash-off model was improved based on experimental results which were obtained from rainfall events with constant rainfall intensity throughout the duration of an event. Keeping the rainfall intensity constant makes it easier to understand the physical wash-off process and to consequently modify the wash-off model. In fact, most of the previous studies used a constant intensity rainfall event to understand the wash-off process and consequently apply the results to develop and improve the wash-off equations. These studies include Sartor and Boyd (1972) where the exponential model was originally proposed and Egodawatta et al. (2007) where the capacity factor was first introduced in the exponential wash-off. However, constant intensity rainfall events are never the case in reality. Nevertheless, the equation proposed by Sartor and Boyd (1972) and the consequent refined version (e.g. Egodawatta et al., 2007) were all shown to be applicable for real case studies too. For example, Brodie and Egodawatta (2011) on a follow-up study on Egodawatta et al. (2007) showed that the use of mean rainfall intensity of a real rainfall event as a representative intensity to derive C_F produces reliable predictions. In this regard, application of the NEM also needs to be checked against wash-off events that resulted from real rainfall events. Such validation also needs information about surface slope.

It can also be noted that the rainfall intensities used in this experiments are generally high compared to rainfall intensities observed in the real world. However, the minimum intensity of ∼30 mm/h was selected based on the trial experiments to produce measurable sediment wash-off amounts from the surface. For example, at 2% slope, even the rainfall intensity of 155 mm/h produced only 6 g wash-off total wash-off at the end of 60 min. In addition to selected sediment size and surface roughness, surface size also a deciding factor in the amount of washed off sediment as the larger surface will have a proportionally higher initial sediment load. On the other hand, unlike sediment size and surface roughness, surface size does not affect the underlying physical process and as a result, the wash-off fraction (= washed off load/initial load) will remain same. This provides the flexibility in choosing the surface size for similar wash-off experiments. The small surface size such as the one used in this study ($1\times1\ m^2$) provides a degree of flexibility to change the experiment conditions (e.g. surface slope, initial load) and makes it possible to run such a large number of experiments. Also, it helps to keep the rainfall intensity fairly uniform over the surface. Similar sized experimental surfaces have been used in recent studies to take advantage of the abovementioned points (Egodawatta et al., 2007; Al Ali et al., 2017). However, the trade-off is the physically lesser amount of washed off sediment from the surface and consequently the limitation in testing very mild rainfall conditions in these experiments. Hence, an optimal surface size needs to be selected in future studies which take into account the flexibilities in the experimental setup and the minimum rainfall intensity that can produce a physically measurable sediment wash-off with limited measurement error. However, rainfall intensities used in these experiments are comparable to rainfall intensities used in similar previous wash-off studies. For example, Egodawatta et al. (2007) used a rainfall intensity range of 40 mm/h - 133 mm/h and 20 mm/h - 133 mm/h in their experiments to study the wash-off behaviour. Recently Al Ali et al. (2017) used a constant rainfall intensity of 120 mm/h in similar experimental settings to study the wash-off behaviour from different surfaces. Due to the practical difficulty in covering a large range of rainfall intensity in an experimental set-up, extrapolation of the equation/model outside the experimental conditions is often used. Even the most widely used exponential model was originally developed for much narrower intensity range of 8 mm/h - 20 mm/h (Sartor and Boyd, 1972) and has been used widely for rainfall intensities that are well outside this range. One of the reasons why this is an accepted practice could be that the pattern of observations from previous studies indicate that the underlying physical transport process of wash-off are quite similar, even outside the experimental

conditions that are tested. For instance, the inclusion of a capacity factor as a function of rainfall intensity and slope would be valid for smaller rainfall intensities as even higher intensities have a maximum capacity in wash-off load as seen from the experimental results. Hence, although the NEM has not been calibrated against smaller rainfall intensities, we believe the model structure of the NEM would still be applicable to smaller rainfall intensities. Nevertheless, this should be verified in future studies.

4 Conclusions

In this study, we proposed an improved exponential wash-off model where a more physically realistic structure was added to the original exponential model by replacing the calibration parameters with functions of external drivers associated with catchment surface and rainfall characteristics. This improvement avoids the need for empirical look-up table/charts and interpolation/ extrapolation and introduces some transparency in the parameter estimation which is otherwise a "black box" approach. Further, replacing the invariant calibration parameters with functions of external drivers (i.e. rainfall intensity and surface slope) makes it easier to investigate the propagation of errors in the external drivers (e.g. rainfall intensity) as these external drivers are now explicitly defined in the new equation. This new exponential model (NEM) was calibrated and verified using the experimental data collected for different combinations of surface slopes and rainfall intensities. Bayesian inference, which allows the incorporation of prior knowledge, was implemented to estimate the distribution of the parameters of the newly introduced functions. In addition, by statistically describing model bias and measurement noise, different sources of uncertainty in the prediction of the NEM were separately estimated.

During calibration, the NEM with a fixed set of parameter values performs as well as the OEM which is calibrated for each and every experimental condition separately. At validation, the NEM's performance improves over the OEM, reflecting the ability of the NEM to perform better under new catchment conditions. Verification measures show the uncertainty estimates associated with the NEM predictions are plausible, indicating that the use of two error terms, autoregressive error and independently identically distributed error, to represent model bias and measurement noise respectively was a reasonable representation of the error process associated with sediment wash-off modelling. The total uncertainty in the prediction of wash-off fraction - which accounts for both model bias and measurement noise - was found to be $\sim$0.1 (10%) when 95% predictive interval is considered, out of which a maximum of 0.06(6%) was due to the parameter uncertainty.
It should be noted that the optimal values of $c_1, ...c_4$ in the NEM needs to be checked against different sediment sizes and different surface roughness as these are two other major external drivers which would affect the sediment wash-off. Nevertheless, the model structure of the NEM would be applicable for any sediment size and surface texture as the underlying physical processes will be the same as those on which the model structure of the NEM was developed.

Acknowledgement: The authors thank Jörg Rieckermann for his engagement in profitable discussions. This research was done as part of the Marie Curie ITN - Quantifying Uncertainty in Integrated Catchment Studies project (QUICS). This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement no. 607000.

Bibliography

[1] Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P., and Rasmussen. An introduction to the European hydrological system - systeme hydrologique europeen ,SHE, 2 : Modelling system the systme hydrologique europen , or european hydrological system (SHE), is an advanced , physically-based , distri. *Journal of Hydrology 87* (1986), 61–77.

[2] Abell, R., Thieme, M. L., Revenga, C., Bryer, M., Kottelat, M., Bogutskaya, N., Coad, B., Mandrak, N., Balderas, S. C., Bussing, W., Stiassny, M. L. J., Skelton, P., Allen, G. R., Unmack, P., Naseka, A., Ng, R., Sindorf, N., Robertson, J., Armijo, E., Higgins, J. V., Heibel, T. J., Wikramanayake, E., Olson, D., López, H. L., Reis, R. E., Lundberg, J. G., Sabaj Pérez, M. H., and Petry, P. Freshwater Ecoregions of the World: A New Map of Biogeographic Units for Freshwater Biodiversity Conservation. *BioScience 58*, 5 (2008), 403.

[3] Ajami, N. K., Duan, Q., and Sorooshian, S. An integrated hydrologic bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research 43*, 1.

[4] Ajami, N. K., Hornberger, G. M., and Sunding, D. L. Sustainable water resource management under hydrological uncertainty. *Water Resources Research 44*, 11 (2008), 1–10.

[5] Akan, A. O. Pollutant Washoff by Overland Flow. *Journal of Environmental Engineering 113*, 4 (aug 1987), 811–823.

[6] Akbari, M., and Afshar, A. Similarity-based error prediction approach for real-time inflow forecasting. *Hydrology Research 45*, 4-5 (2014), 589.

[7] Al, S. E. T., and Elson, P. E. A. N. Extraordinary flood response of a small urban watershed. *Bulletin of the American Meteorological Society 86*, 12 (2005), 1730–1732.

[8] Al Ali, S., Bonhomme, C., Dubois, P., and Chebbo, G. Investigation of the wash-off process using an innovative portable rainfall simulator allowing continuous monitoring of flow and turbidity at the urban surface outlet. *Science of the Total Environment 609* (2017), 17–26.

[9] Alley, W. M. Estimation of Impervious-Area Washoff Parameters. *Water Resources Research 17*, 4 (1981), 1161–1166.

[10] Ammann, L., Reichert, P., and Fenicia, F. A framework for likelihood functions of deterministic hydrological models. *Hydrology and Earth System Sciences Discussions 2018* (2018), 1–39.

[11] Arnal, L., Ramos, M. H., De Perez, E. C., Cloke, H. L., Stephens, E., Wetterhall, F., Van Andel, S. J., and Pappenberger, F. Willingness-to-pay for a probabilistic flood forecast: A risk-based decision-making game. *Hydrology and Earth System Sciences 20*, 8 (2016), 3109–3128.

[12] Aronica, G., Bates, P. D., and Horritt, M. S. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. *Hydrological Processes 16*, 10 (2002), 2001–2016.

[13] Azimi-Zonooz, A., Krajewski, W. F., Bowles, D. S., and Seo, D. J. Spatial rainfall estimation by linear and non-linear co-kriging of radar-rainfall and raingage data. *Stochastic Hydrology and Hydraulics 3*, 1 (1989), 51–67.

[14] Bailey, R., and Dobson, C. Forecasting for floods in the severn catchment. *Journal of the Institution of Water Engineers and Scientists* (1981).

[15] Barancourt, C., Creutin, J. D., and Rivoirard, J. A method for delineating and estimating rainfall fields. *Water Resources Research 28*, 4 (1992), 1133–1144.

[16] Bárdossy, A., and Hörning, S. Gaussian and non-Gaussian inverse modeling of groundwater flow using copulas and random mixing. *Water Resources Research 52*, 6 (2016), 4504–4526.

[17] Bates, B. C., and Campbell, E. P. A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. *Water Resources Research 37*, 4 (2001), 937–947.

[18] Beck, M. B. Water quality modeling: A review of the analysis of uncertainty. *Water Resources Research 23*, 8 (1987), 1393–1442.

[19] Beckers, J. V., Weerts, A. H., Tijdeman, E., and Welles, E. ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction. *Hydrology and Earth System Sciences 20*, 8 (2016), 3277–3287.

[20] Beersma, J. J., and Adri Buishand, T. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. *Climate Research* (2003).

[21] Beersma, J. J., and Buishand, T. A. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation . *Climate Research 25*, 2 (2003), 121–133.

[22] Benke, K. K., Lowell, K. E., and Hamilton, A. J. Parameter uncertainty, sensitivity analysis and prediction error in a water-balance hydrological model. *Mathematical and Computer Modelling 47*, 11-12 (2008), 1134–1149.

[23] Bergström, S. Development and Application of a Conceptual Runoff Model for Scandinavian Catchments. *Smhi* (1976).

[24] Berndt, C., Rabiei, E., and Haberlandt, U. Geostatistical merging of rain gauge and radar data for high temporal resolutions and various station density scenarios. *Journal of Hydrology 508* (2014), 88–101.

[25] Betterle, A., Schirmer, M., and Botter, G. Characterizing the spatial correlation of daily streamflows. *Water Resources Research 53*, 2 (2017), 1646–1663.

[26] Beven, K. Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. *Hydrological Processes 16*, 2 (2002), 189–206.

[27] Beven, K. Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal 61*, 9 (2016), 1652–1665.

[28] Beven, K., and Binley, A. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes 6*, 3 (1992), 279–298.

[29] Bircher, S., Skou, N., Jensen, K. H., Walker, J. P., and Rasmussen, L. A soil moisture and temperature network for SMOS validation in Western Denmark. *Hydrology and Earth System Sciences 16*, 5 (2012), 1445–1463.

[30] Bogner, K., and Pappenberger, F. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research 47*, 7 (2011).

[31] Bogner, K., Pappenberger, F., and Cloke, H. L. Technical Note: The normal quantile transformation and its application in a flood forecasting system. *Hydrology and Earth System Sciences 16*, 4 (2012), 1085–1094.

[32] Borgomeo, E., Pflug, G., Hall, J. W., and Hochrainer-Stigler, S. Assessing water resource system vulnerability to unprecedented hydrological drought using copulas to characterize drought duration and deficit. *Water Resources Research 51*, 11 (2015), 8927–8948.

[33] Borup, M., Grum, M., Madsen, H., and Mikkelsen, P. S. A partial ensemble Kalman filtering approach to enable use of range limited observations. *Stochastic Environmental Research and Risk Assessment 29*, 1 (2015), 119–129.

[34] Boyle, D. P., Gupta, H. V., and Sorooshian, S. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research* (2000).

[35] Breinholt, A., Møller, J. K., Madsen, H., and Mikkelsen, P. S. A formal statistical approach to representing uncertainty in rainfall-runoff modelling with focus on residual analysis and probabilistic output evaluation - Distinguishing simulation and prediction. *Journal of Hydrology 472*, Supplement C (2012), 36–52.

[36] Bringi, V. N., Keenan, T. D., and Chandrasekar, V. Correcting C-band radar reflectivity and differential reflectivity data for rain attenuation: A self-consistent method with constraints. *IEEE Transactions on Geoscience and Remote Sensing 39*, 9 (2001), 1906–1915.

[37] Butler, D., and Clark, P. Sediment management in urban drainage catchments. *Report, Construction Industry Research and Information Association (CIRIA)* (1995).

[38] Butts, M., and Graham, D. Flexible Integrated Watershed Modeling with MIKE SHE. In *Watershed Models*. 2005.

[39] Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology 298*, 1-4 (2004), 242–266.

[40] Buytaert, W., Dewulf, A., De Bièvre, B., Clark, J., and Hannah, D. M. Citizen Science for Water Resources Management: Toward Polycentric Monitoring and Governance? *Journal of Water Resources Planning and Management 142*, 4 (2016), 01816002.

[41] Carbajal, J. P., Leitão, J. P., Albert, C., and Rieckermann, J. Appraisal of data-driven and mechanistic emulators of nonlinear simulators: The case of hydrodynamic urban drainage models. *Environmental Modelling & Software 92* (jun 2017), 17–27.

[42] Chech, T. V. Principles of water resources: history, development, management, and policy.

[43] Cheng, Q., Agterberg, F. P., and Ballantyne, S. B. The separation of geochemical anomalies from background by fractal methods. *Journal of Geochemical Exploration 51*, 2 (1994), 109–130.

[44] Chilès, J., and Delfiner, P. *Geostatistics. Modeling spatial uncertainty*. 2014.

[45] Coccia, G., and Todini, E. Recent developments in predictive uncertainty assessment based on the model conditional processor approach. *Hydrology and Earth System Sciences 15*, 10 (2011), 3253–3274.

[46] Coleman, T. J. A comparison of the modelling of suspended solids using SWMM3 quality prediction algorithms with a model based on sediment transport theory. In *6th Int. Conf. on Urban Storm Drainage* (Reston, 1993), ASCE.

[47] Cover, T. M., and Thomas, J. A. *Elements of Information Theory*. 2005.

[48] Cox, R. T. Probability, Frequency and Reasonable Expectation. *American Journal of Physics 14*, 1 (1946), 1–13.

[49] Craig, P. S., Goldstein, M., Rougier, J. C., and Seheult, A. H. Bayesian Forecasting for Complex Systems Using Computer Simulators. *Journal of the American Statistical Association 96*, 454 (2001), 717–729.

[50] CRAN. *www.r-project.org*, 2015.

[51] Cressie, N. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1993.

[52] Cressie, N., and Hawkins, D. M. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology 12*, 2 (1980), 115–125.

[53] Daniell, J., Wenzel, F., and Schaefer, A. The economic costs of natural disasters globally from 1900-2015: historical and normalised floods, storms, earthquakes, volcanoes, bushfires, drought and other disasters. In *EGU General Assembly Conference Abstracts* (apr 2016), vol. 18 of *EGU General Assembly Conference Abstracts*, pp. EPSC2016–1899.

[54] de Wit, A. J. W., de Bruin, S., and Torfs, P. J. J. F. Representing Uncertainty in Continental-Scale Gridded Precipitation Fields for Agrometeorological Modeling. *Journal of Hydrometeorology* (2008).

[55] Del Giudice, D., Albert, C., Rieckermann, J., and Reichert, P. Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation. *Water Resources Research 52*, 4 (2016), 3162–3186.

[56] Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. *Hydrology and Earth System Sciences 17*, 10 (2013), 4209–4225.

[57] Del Giudice, D., Reichert, P., Bareš, V., Albert, C., and Rieckermann, J. Model bias and complexity - Understanding the effects of structural deficits and input errors on runoff predictions. *Environmental Modelling and Software 64*, 1 (2015), 205–214.

[58] Deletic, A., Dotto, C. B. S., McCarthy, D. T., Kleidorfer, M., Freni, G., Mannina, G., Uhl, M., Henrichs, M., Fletcher, T. D., Rauch, W., Bertrand-Krajewski, J. L., and Tait, S. Assessing uncertainties in urban drainage models. *Physics and Chemistry of the Earth 42-44*, January 2016 (2012), 3–10.

[59] Deletic, A., Maksimovic, C., and Ivetic, M. Modelling of storm wash-off of suspended solids from impervious surfaces. *Journal of Hydraulic Research 35*, 1 (1997), 99–118.

[60] Delrieu, G., Wijbrans, A., Boudevillain, B., Faure, D., Bonnifait, L., and Kirstetter, P. E. Geostatistical radar-raingauge merging: A novel method for the quantification of rain estimation accuracy. *Advances in Water Resources* (2014).

[61] Diggle, P. J., and Ribeiro, P. J. *Model-based Geostatistics (Springer Series in Statistics)*. 2007.

[62] Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., and Shrestha, D. L. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. *Hydrology and Earth System Sciences 19*, 7 (2015), 3181–3201.

[63] Dotto, C. B. S., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D. T., and Fletcher, T. D. Performance and sensitivity analysis of stormwater models using a Bayesian approach and long-term high resolution data. *Environmental Modelling & Software 26*, 10 (2011), 1225–1239.

[64] Dotto, C. B. S., Mannina, G., Kleidorfer, M., Vezzaro, L., Henrichs, M., McCarthy, D. T., Freni, G., Rauch, W., and Deletic, A. Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling. *Water Research 46*, 8 (2012), 2545–2558.

[65] Duan, Q., Sorooshian, S., and Gupta, V. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* (1992).

[66] Ed, M. F. B., and Hons, B. E. Characterising Urban Pollutant Loads. *Civil Engineering*, February (2010), 1–248.

[67] Eggimann, S., Mutzner, L., Wani, O., Schneider, M., Spuhler, D., Moy De Vitry, M., Beutler, P., and Maurer, M. The Potential of Knowing More: A Review of Data-Driven Urban Water Management. *Environmental Science and Technology 51*, 5 (2017).

[68] Egodawatta, P., and Goonetilleke, A. Understanding road surface pollutant wash-off and underlying physical processes using simulated rainfall. *Water Science and Technology 57*, 8 (2008), 1241–1246.

[69] Egodawatta, P., Thomas, E., and Goonetilleke, A. Mathematical interpretation of pollutant wash-off from urban road surfaces using simulated rainfall. *Water Research 41*, 13 (2007), 3025–3031.

[70] Emmanuel, I., Andrieu, H., Leblois, E., and Flahaut, B. Temporal and spatial variability of rainfall at the urban hydrological scale. *Journal of Hydrology 430-431* (2012), 162–172.

[71] EPA. *Environmental Protection Agency, (https://www.epa.gov/water-research/storm-water-management-model-swmm)*, 2015.

[72] Erdin, R., Frei, C., and Künsch, H. R. Data Transformation and Uncertainty in Geostatistical Combination of Radar and Rain Gauges. *Journal of Hydrometeorology 13*, 1987 (2012), 1332–1346.

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

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

[75] Ferriman, A. {BMJ} readers choose the "sanitary revolution" as greatest medical advance since 1840. *BMJ 334*, 7585 (jan 2007), 111.

[76] Freer, J., Beven, K., and Ambroise, B. Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach. *Water Resources Research 32*, 7 (1996), 2161–2173.

[77] Freni, G., and Mannina, G. Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution. *Journal of Hydrology 392*, 1 (2010), 31–39.

[78] Freni, G., Mannina, G., and Viviani, G. Uncertainty assessment of an integrated urban drainage model. *Journal of Hydrology 373*, 3 (2009), 392–404.

[79] Freni, G., Mannina, G., and Viviani, G. The influence of rainfall time resolution for urban water quality modelling. *Water Science and Technology 61*, 9 (2010), 2381–2390.

[80] Frey, M. P., Stamm, C., Schneider, M. K., and Reichert, P. Using discharge data to reduce structural deficits in a hydrological model with a Bayesian inference approach and the implications for the prediction of critical source areas. *Water Resources Research 47*, 12 (2011), 1–18.

[81] Furquim, G., Pessin, G., Faiçal, B. S., Mendiondo, E. M., and Ueyama, J. Improving the accuracy of a flood forecasting model by means of machine learning and chaos theory. *Neural Computing and Applications 27*, 5 (2016), 1129–1141.

[82] G. E. P. Box & Cox, D. R. An Analysis of Transformations. *Journal of the Royal Statistical Society 26*, ii (1964), 211–252.

[83] Gamerith, V., Neumann, M. B., and Muschalla, D. Applying global sensitivity analysis to the modelling of flow and water quality in sewers. *Water Research 47*, 13 (2013), 4600–4611.

[84] Ganti, V., Chadwick, A. J., Hassenruck-Gudipati, H. J., Fuller, B. M., and Lamb, M. P. Experimental river delta size set by multiple floods and backwater hydrodynamics. *Science Advances 2*, 5 (2016), 1–11.

[85] Genz, A., and Carlo, M. Numerical Computation of Multivariate Normal Probabilities available. *Journal of Computational and Graphical Statistics 1*, 2 (1992), 141–149.

[86] Germann, U., Berenguer, M., Sempere-Torres, D., and Zappa, M. REAL - Ensemble radar precipitation estimation for hydrology in a mountainous region. *Quarterly Journal of the Royal Meteorological Society* (2009).

[87] Gupta, H. V., Sorooshian, S., and Yapo, P. Y. Toward improved calibration of hydrologic models: Multiple and noncommensurable measure of information. *Water Resources Research 34*, 4 (1998), 751–763.

[88] Guy, H. P. Sediment Problems in Urban Areas. *Geological survey circular 601-E, U.S. Geological Survey* (1970).

[89] Hall, J. Handling uncertainty in the hydroinformatic process. *Journal of Hydroinformatics* (2003).

[90] Hall, J., and Solomatine, D. A framework for uncertainty analysis in flood risk management decisions. *International Journal of River Basin Management 6*, 2 (2008), 85–98.

[91] Hall, J. W., Manning, L. J., and Hankin, R. K. Bayesian calibration of a flood inundation model using spatial data. *Water Resources Research 47*, 5 (2011), 1–14.

[92] Halls-Moore, M. Bayesian Statistics: A Beginner's Guide.

[93] Han, F., and Zheng, Y. Joint analysis of input and parametric uncertainties in watershed water quality modeling: A formal Bayesian approach. *Advances in Water Resources 116*, October 2017 (2018), 77–94.

[94] Hardy, G. H. *A Mathematician's Apology*. Cambridge University Press, mar 2012.

[95] Harrison, D., Driscoll, S. J., and Kitchen, M. Improving precipitation estimates from weather radar using quality control and correction techniques. *Meteorol. Appl.* (2000).

[96] Herr, H. D., and Krzysztofowicz, R. Generic probability distribution of rainfall in space: The bivariate model. *Journal of Hydrology* (2005).

[97] Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., and Ryne, R. Combining Field Data and Computer Simulations for Calibration and Prediction. *SIAM Journal on Scientific Computing 26*, 2 (jan 2004), 448–466.

[98] Hoeting, J. J. A., Madigan, D., Raftery, A. E. A., and Volinsky, C. T. Bayesian model averaging: A tutorial. *Statistical science 14*, 4 (1999), 382–401.

[99] Hofer, T., Gruber, G., Gamerith, V., Montserrat, A., Corominas, L., and Muschalla, D. Using Temperature Sensors to Detect Occurrence and Duration of Combined Sewer Overflows. *13th International Conference on Urban Drainage 2012*, September (2014), 1–8.

[100] Hojberg, A. L., Troldborg, L., Stisen, S., Christensen, B. B., and Henriksen, H. J. Stakeholder driven update and improvement of a national water resources model. *Environmental Modelling and Software 40* (2013), 202–213.

[101] Hong, Y., Bonhomme, C., Le, M. H., and Chebbo, G. New insights into the urban washoff process with detailed physical modelling. *Science of the Total Environment 573* (2016), 924–936.

[102] Honti, M., Stamm, C., and Reichert, P. Integrated uncertainty assessment of discharge predictions with a statistical error model. *Water Resources Research 49*, 8 (2013), 4866–4884.

[103] Hoss, F., and Fischbeck, P. S. Performance and robustness of probabilistic river forecasts computed with quantile regression based on multiple independent variables. *Hydrology and Earth System Sciences 19*, 9 (2015), 3969–3990.

[104] Huber, W., and Dickenson, r. *Storm Water management Model version 4: User's Manual*, 1992.

[105] Hyvärinen, A., and Oja, E. Independent Component Analysis: Algorithms and Applications. *Neural Networks 13*, 45 (2000), 411–430.

[106] Ilja Van Meerveld, H. J., Vis, M. J., and Seibert, J. Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences 21*, 9 (2017), 4895–4905.

[107] IPCC. Climate Change 2014 Synthesis Report Summary Chapter for Policymakers. *Ipcc* (2014), 31.

[108] Jaynes, E. T. Probability Theory: The Logic of Science. *The Mathematical Intelligencer* (2003).

[109] Jensen, K. H., and Illangasekare, T. H. HOBE: A Hydrological Observatory. *Vadose Zone Journal 10*, 1 (2011), 1.

[110] Jensen, N. E., and Pedersen, L. Spatial variability of rainfall: Variations within a single radar pixel. *Atmospheric research 77*, 1-4 (2005), 269–277.

[111] Jeon, J. H., Park, C. G., and Engel, B. A. Comparison of performance between genetic algorithm and SCE-UA for calibration of SCS-CN surface runoff simulation. *Water (Switzerland) 6*, 11 (2014), 3433–3456.

[112] Joanes, D. N., and Gill, C. A. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society Series D: The Statistician 47*, 1 (1998), 183–189.

[113] Juston, J. M., Kauffeldt, A., Montano, B. Q., Seibert, J., Beven, K. J., and Westerberg, I. K. Smiling in the rain: Seven reasons to be positive about uncertainty in hydrological modelling. *Hydrological Processes 27*, 7 (2013), 1117–1122.

[114] Karlsson, M., and Yakowitz, S. Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research* (1987).

[115] Kavetski, D. *Parameter Estimation and Predictive Uncertainty Quantification in Hydrological Modelling*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018, pp. 1–42.

[116] Kavetski, D., Kuczera, G., and Franks, S. W. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research 42*, 3 (2006), 1–9.

[117] Kavetski, D., Kuczera, G., and Franks, S. W. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research 42*, 3 (2006), 1–10.

[118] Kelly, K. S., and Krzysztofowicz, R. Probability distributions for flood warning systems. *Water Resources Research* (1994).

[119] Kennedy, M. C., and O'Hagan, A. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*, 3 (2001), 425–464.

[120] Kerkez, B., Gruden, C., Lewis, M., Montestruque, L., Quigley, M., Wong, B., Bedig, A., Kertesz, R., Braun, T., Cadwalader, O., Poresky, A., and Pak, C. Smarter stormwater systems. *Environmental Science and Technology 50*, 14 (2016), 7267–7273.

[121] Kirstetter, P. E., Andrieu, H., Boudevillain, B., and Delrieu, G. A Physically based identification of vertical profiles of reflectivity from volume scan radar data. *Journal of Applied Meteorology and Climatology 52*, 7 (2013), 1645–1663.

[122] Klein, B., Meissner, D., Kobialka, H. U., and Reggiani, P. Predictive uncertainty estimation of hydrological multi-model ensembles using pair-copula construction. *Water (Switzerland) 8*, 4 (2016).

[123] Klein, J. P., and Mosechberger, M. L. *Survival Analysis: techniques for censored and truncated data*. 2003.

[124] Klemeš, V. COMMON SENSE AND OTHER HERESIES: Selected Papers on Hydrology and Water Resources Engineering. *Canadian Water Resources Journal / Revue canadienne des ressources hydriques 25*, 1 (jan 2000), 108.

[125] Koch, J., Siemann, A., Stisen, S., and Sheffield, J. Spatial validation of large-scale land surface models against monthly land surface temperature patterns using innovative performance metrics. *Journal of Geophysical Research: Atmospheres 121* (2016), 5430–5452.

[126] Koenker, R., and Bassett, G. Regression quantiles. *Econometrica 46*, 1 (1978), 33–50.

[127] Krzysztofowicz, R. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research* (1999).

[128] Krzysztofowicz, R. The case for probabilistic forecasting in hydrology. *Journal of Hydrology 249*, 1-4 (aug 2001), 2–9.

[129] Kuczera, G., Kavetski, D., Franks, S., and Thyer, M. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology 331*, 1-2 (2006), 161–177.

[130] Laio, F., and Tamea, S. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci. 11*, 4 (may 2007), 1267–1277.

[131] Lall, U., and Sharma, A. A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research* (1996).

[132] Lang, T. J., Nesbitt, S. W., and Carey, L. D. On the correction of partial beam blockage in polarimetric radar data. *Journal of Atmospheric and Oceanic Technology 26*, 5 (2009), 943–957.

[133] Lawler, D. M., Petts, G. E., Foster, I. D. L., and Harper, S. Turbidity dynamics during spring storm events in an urban headwater river system: The Upper Tame, West Midlands, UK. *Science of the Total Environment 360*, 1-3 (2006), 109–126.

[134] Le Coz, J., Patalano, A., Collins, D., Guillén, N. F., García, C. M., Smart, G. M., Bind, J., Chiaverini, A., Le Boursicaud, R., Dramais, G., and Braud, I. Crowdsourced data for flood hydrology: Feedback from recent citizen science projects in Argentina, France and New Zealand. *Journal of Hydrology 541* (2016), 766–777.

[135] Li, H., Luo, L., Wood, E. F., and Schaake, J. The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *Journal of Geophysical Research Atmospheres 114*, 4 (2009), 1–10.

[136] Lien, G. Y., Kalnay, E., and Miyoshi, T. Effective assimilation of global precipitation: Simulation experiments. *Tellus, Series A: Dynamic Meteorology and Oceanography 65* (2013), 1–16.

[137] Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology 201*, 1-4 (1997), 272–288.

[138] Lo, S. W., Wu, J. H., Lin, F. P., and Hsu, C. H. Visual sensing for urban flood monitoring. *Sensors (Switzerland) 15*, 8 (2015), 20006–20029.

[139] Lopez Lopez, P., Verkade, J. S., Weerts, A. H., and Solomatine, D. P. Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: A comparison. *Hydrology and Earth System Sciences 18*, 9 (2014), 3411–3428.

[140] Madsen, H. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources 26*, 2 (feb 2003), 205–216.

[141] Maheepala, U. K., Takyi, A. K., and Perera, B. J. C. Hydrological data monitoring for urban stormwater drainage systems. *Journal of Hydrology 245*, 1-4 (may 2001), 32–47.

[142] Maisels, C. K. *Early civilizations of the Old World: the formative histories of Egypt, the Levant, Mesopotamia, India and China*. Routledge, 2003.

[143] Mantovan, P., and Todini, E. Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology 330*, 1-2 (2006), 368–381.

[144] Marsh, T., and hannaford, J. Uk hydrometric register, hydrological data uk series. *Center for Ecology and Hydrology, Wallingford, UK* (2008).

[145] Mazzoleni, M., Juliette Cortes Arevalo, V., Wehn, U., Alfonso, L., Norbiato, D., Monego, M., Ferri, M., and Solomatine, D. P. Exploring the influence of citizen involvement on the assimilation of crowdsourced observations: A modelling study based on the 2013 flood event in the Bacchiglione catchment (Italy). *Hydrology and Earth System Sciences 22*, 1 (2018), 391–416.

[146] McInerney, D., Thyer, M., Kavetski, D., Bennett, B., Lerat, J., Gibbs, M., and Kuczera, G. A simplified approach to produce probabilistic hydrological model predictions. *Environmental Modelling & Software* (2018).

[147] MetOffice UK. Online, 2016.

[148] Millar, R. G. Analytical determination of pollutant wash-off parameters. *Journal of Environmental Engineering Vol. 125,*, No. 10 (Technical Note) (1999), 989–992.

[149] Miller, K. L., Berg, S. J., Davison, J. H., Sudicky, E. A., and Forsyth, P. A. Efficient uncertainty quantification in fully-integrated surface and subsurface hydrologic simulations. *Advances in Water Resources 111*, June 2017 (2018), 381–394.

[150] Milly, A. P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Zbigniew, W., Lettenmaier, D. P., Stouffer, R. J., and Milly, P. C. D. Stationarity Is Dead : Stationarity Whither Water Management ? *Science 319*, 5863 (2008), 573–574.

[151] Mitchell, G., Lockyer, J., and McDonald, A. Pollution Hazard from Urban Nonpoint Sources: A GIS-model to Support Strategic Environmental Planning in the UK. *Technical Report, School of Geography, University of Leeds 1,2* (2001), 240pp.

[152] Mlodinow Leonard. *The Drunkard's walk : how randomness rules our lives / Leonard Mlodinow.* Pantheon Books New York, 2008.

[153] Montanari, A., and Brath, A. A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research 40*, 1 (2004), 1–11.

[154] Montanari, A., and Koutsoyiannis, D. A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research 48*, 9 (2012), 1–15.

[155] Montanari, A., & Brath, A. A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research 40*, 1 (2004), 1–11.

[156] Montserrat, A., Gutierrez, O., Poch, M., and Corominas, L. Field validation of a new low-cost method for determining occurrence and duration of combined sewer overflows. *Science of the Total Environment 463-464* (2013), 904–912.

[157] Montserrat, A., Hofer, T., Poch, M., Muschalla, D., and Corominas, L. Using the duration of combined sewer overflow events for the calibration of sewer hydrodynamic models. *Urban Water Journal 14*, 8 (2017), 782–788.

[158] Moy De Vitry, M., Dicht, S., and Leitão, J. P. FloodX: Urban flash flood experiments monitored with conventional and alternative sensors. *Earth System Science Data* (2017).

[159] Müller, H. Bayesian Transgaussian Kriging. *15th European Young Statisticians Meeting 2007*, 1 (2007), 1–5.

[160] Muthusamy, M., Tait, S., Schellart, A., Beg, M. N. A., Carvalho, R. F., and de Lima, J. L. M. P. Improving understanding of the underlying physical process of sediment wash-off from urban road surfaces. *Journal of Hydrology 557* (feb 2018), 426–433.

[161] Mutzner, L., Staufer, P., and Ort, C. Model-based screening for critical wet-weather discharges related to micropollutants from urban areas. *Water Research 104* (2016), 547–557.

[162] Nace, R. L. Pierre Perrault: the Man and His Contribution To Modern Hydrology. *JAWRA Journal of the American Water Resources Association 10*, 4 (1974), 633–647.

[163] Nakamura, E. Factors affecting stormwater quality decay coefficient. In *Proceedings of the Third International Conference on Urban Storm Drainage* (Goteborg, 1984), A. S. P. Balmer, P. Malmquist, Ed., pp. 979–988.

[164] Nanding, N., Rico-Ramirez, M., and Han, D. Comparison of different radar-gauge merging techniques. *Journal of Hydroinformatics* (2015).

[165] Nash, J. E., and Sutcliffe, J. V. River Flow Forecasting Through Conceptual Models Part I-a Discussion of Principles*. *Journal of Hydrology 10* (1970), 282–290.

[166] Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., and Weijs, S. V. A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal 61*, 9 (2016), 1666–1678.

[167] O'Hagan, A. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety 91*, 10-11 (2006), 1290–1300.

[168] Overgaard, J. Energy-based land-surface modelling: new opportunities in integrated hydrological modelling. 128.

[169] Pappenberger, F., and Beven, K. J. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research 42*, 5 (2006), 1–8.

[170] Peleg, N., Blumensaat, F., Molnar, P., Fatichi, S., and Burlando, P. Partitioning the impacts of spatial and climatological rainfall variability in urban drainage modeling. *Hydrol. Earth Syst. Sci. 21*, 3 (mar 2017), 1559–1572.

[171] Pianosi, F., and Raso, L. Dynamic modeling of predictive uncertainty by regression on absolute errors. *Water Resources Research 48*, 3 (2012), 1–11.

[172] Pokhrel, P., Robertson, D. E., and Wang, Q. J. A Bayesian joint probability post-processor for reducing errors and quantifying uncertainty in monthly streamflow predictions. *Hydrology and Earth System Sciences 17*, 2 (2013), 795–804.

[173] R. *R: a language and environment for statistical computing.(R development core team - https://cran.r-project.org/)*, 2018.

[174] Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review 133*, 5 (2005), 1155–1174.

[175] Rajagopalan, B., and Lall, U. A k-nearest-neighbor Simulator for Daily Precipitation and Other Variables. *Water Resources Research 35*, 10 (1999), 3089–3101.

[176] Rasmussen, M. A low cost calibration method for urban drainage models. *Proceedings of 11th Internation Conference on Urban Drainage* (2008), 1–6.

[177] Refsgaard, J. C. Parameterisation, calibration and validation of distributed hydrological models. *Journal of Hydrology 198*, 1-4 (1997), 69–97.

[178] Refsgaard, J. C., and Henriksen, H. J. Modelling guidelines - Terminology and guiding principles. *Advances in Water Resources 27*, 1 (2004), 71–82.

[179] Refsgaard, J. C., van der Sluijs, J. P., Brown, J., and van der Keur, P. A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources 29*, 11 (2006), 1586–1597.

[180] Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., and Vanrolleghem, P. A. Uncertainty in the environmental modelling process - A framework and guidance. *Environmental Modelling and Software 22*, 11 (2007), 1543–1556.

[181] Reggiani, P., Renner, M., Weerts, A. H., and Van Gelder, P. A. H. J. M. Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research* (2009).

[182] Reichert, P., Borsuk, M., Hostmann, M., Schweizer, S., Spörri, C., Tockner, K., and Truffer, B. Concepts of decision support for river rehabilitation. *Environmental Modelling and Software 22*, 2 (2007), 188–201.

[183] Reichert, P., and Schuwirth, N. Linking statistical bias description to multiobjective model calibration. *Water Resources Research 48*, 9 (sep 2012).

[184] Reichert, P., White, G., Bayarri, M. J., and Pitman, E. B. Mechanism-based emulation of dynamic simulation models: Concept and application in hydrology. *Computational Statistics and Data Analysis 55*, 4 (2011), 1638–1655.

[185] Remy, N., Boucher, A., and Wu, J. *Applied Geostatistics with SGeMS: A user's guide*. 2009.

[186] Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research 46*, 5 (2010), 1–22.

[187] Rico-Ramirez, M. A., and Cluckie, I. D. Classification of ground clutter and anomalous propagation using dual-polarization weather radar. *IEEE Transactions on Geoscience and Remote Sensing* (2008).

[188] Riggelsen, C. Learning parameters of Bayesian networks from incomplete data via importance sampling. *International Journal of Approximate Reasoning 42*, 1 (2006), 69–83.

[189] Rinderknecht, S. L., Albert, C., Borsuk, M. E., Schuwirth, N., Künsch, H. R., and Reichert, P. The effect of ambiguous prior knowledge on Bayesian model parameter inference and prediction. *Environmental Modelling and Software 62* (2014), 300–315.

[190] Ringgaard, R., Herbst, M., Friborg, T., Schelde, K., Thomsen, A. G., and Soegaard, H. Energy Fluxes above Three Disparate Surfaces in a Temperate Mesoscale Coastal Catchment. *Vadose Zone Journal 10*, 1 (2011), 54.

[191] Roscoe, K., Weerts, a. H., and Schroevers, M. Estimation of the uncertainty in water level forecasts at ungauged river locations using Quantile Regression. *International Journal of River Basin Management 14*, November (2012), 1–31.

[192] Sabourin, A., and Renard, B. Combining regional estimation and historical floods: A multivariate semiparametric peaks-over-threshold model with censored data. *Water Resources Research 51*, 12 (2015), 9646–9664.

[193] Sadegh, M., Ragno, E., and Aghakouchak, A. Water Resources Research. 5166–5183.

[194] Sage, J., Bonhomme, C., Al Ali, S., and Gromaire, M. C. Performance assessment of a commonly used "accumulation and wash-off" model from long-term continuous road runoff turbidity measurements. *Water Research 78* (2015), 47–59.

[195] Sage, J., Bonhomme, C., Berthier, E., and Gromaire, M.-C. Assessing the Effect of Uncertainties in Pollutant Wash-Off Dynamics in Stormwater Source-Control Systems Modeling: Consequences of Using an Inappropriate Error Model. *Journal of Environmental Engineering 143*, August (2016), 1–9.

[196] Salvadori, G., and De Michele, C. Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research 40*, 12 (2004), 1–17.

[197] Sartor, J. D., Boyd, G. B., and Agardy, F. J. Water pollution of street surface contaminants. *Journal (Water Pollution Control Federation) 46*, 3 (1974), 458–467.

[198] Scharling, P. B., Rasmussen, E. S., Sonnenborg, T. O., Engesgaard, P., and Hinsby, K. Three-dimensional regional-scale hydrostratigraphic modeling based on sequence stratigraphic methods: A case study of the Miocene succession in Denmark. *Hydrogeology Journal 17*, 8 (2009), 1913–1933.

[199] Scheidegger, A. adaptMCMC: Implementation of a Generic Adaptive Monte Carlo Markov Chain Sampler, 2017.

[200] Scheidegger, A., and Rieckermann, J. Bayesian assimilation of rainfall sensors with fundamentally different integration characteristics. *WRaH Proceedings* (2014).

[201] Schleiss, M., Chamoun, S., and Berne, A. Stochastic simulation of intermittent rainfall using the concept of "dry drift". *Water Resources Research* (2014).

[202] Schoups, G., and Vrugt, J. A. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research 46*, 10 (2010), 1–17.

[203] Schuurmans, J. M., and Bierkens, M. F. P. Effect of spatial distribution of daily rainfall on interior catchment response of a distributed hydrological model. *Hydrology and Earth System Sciences 3*, 4 (2006), 2175–2208.

[204] Sharma, A., and Mehrotra, R. An information theoretic alternative to model a natural system using observational information alone. *Water Resources Research 50*, 1 (2014), 650–660.

[205] Sharma, A., Tarboton, D. G., and Lall, U. Streamflow simulation: A nonparametric approach. *Water Resources Research* (1997).

[206] Shaw, S. B., Stedinger, J. R., and Walter, M. T. Evaluating Urban Pollutant Buildup/Wash-Off Models Using a Madison, Wisconsin Catchment. *Journal of Environmental Engineering* (2010).

[207] Shrestha, D. L., and Solomatine, D. P. Machine learning approaches for estimation of prediction interval for the model output. *Neural networks : the official journal of the International Neural Network Society 19*, 2 (mar 2006), 225–235.

[208] Shrestha, D. L., and Solomatine, D. P. Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. *International Journal of River Basin Management 6*, 2 (2008), 109–122.

[209] Sideris, I. V., Gabella, M., Erdin, R., and Germann, U. Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. *Quarterly Journal of the Royal Meteorological Society* (2014).

[210] Sikorska, A. E., Montanari, A., and Koutsoyiannis, D. Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques. *Journal of Hydrologic Engineering* (2015).

[211] Sikorska, A. E., and Renard, B. Calibrating a hydrological model in stage space to account for rating curve uncertainties: general framework and key challenges. *Advances in Water Resources 105* (2017), 51–66.

[212] Sikorska, A. E., Scheidegger, A., Banasik, K., and Rieckermann, J. Bayesian uncertainty assessment of flood predictions in ungauged urban basins for conceptual rainfall-runoff models. *Hydrology and Earth System Sciences 16*, 4 (2012), 1221–1236.

[213] Sikorska, A. E., Scheidegger, A., Banasik, K., and Rieckermann, J. Bayesian uncertainty assessment of flood predictions in ungauged urban basins for conceptual rainfall-runoff models. *Hydrology and Earth System Sciences 16*, 4 (2012), 1221–1236.

[214] Sivakumar, B. *Chaos in hydrology: Bridging determinism and stochasticity.* 2016.

[215] Solomatine, D. P., Maskey, M., and Shrestha, D. L. Instance-based learning compared to other data-driven methods in hydrological forecasting. *Hydrological Processes* (2008).

[216] Solomatine, D. P., and Shrestha, D. L. A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Research* (2009).

[217] Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., and Xu, C. Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. *Journal of Hydrology 523*, 225 (2015), 739–757.

[218] Sorooshian, S., and Gupta, V. K. Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness. *Water Resources Research 19*, 1 (feb 1983), 260–268.

[219] Souza Filho, F. A., and Lall, U. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. *Water Resources Research 39*, 11 (2003), 1–13.

[220] Stedinger, J. R., Vogel, R. M., Lee, S. U., and Batchelder, R. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research 44*, 12 (2008), 1–17.

[221] Stisen, S., Hojberg, A. L., Troldborg, L., Refsgaard, J. C., Christensen, B. S., Olsen, M., and Henriksen, H. J. On the importance of appropriate precipitation gauge catch correction for hydrological modelling at mid to high latitudes. *Hydrology and Earth System Sciences 16*, 11 (2012), 4157–4176.

[222] Stisen, S., Koch, J., Sonnenborg, T. O., Refsgaard, J. C., Bircher, S., Ringgaard, R., and Jensen, K. H. Moving beyond run-off calibration-Multivariable optimization of a surface-subsurface-atmosphere model. *Hydrological Processes*, May (2018), 2654–2668.

[223] Stisen, S., Sonnenborg, T. O., Højberg, A. L., Troldborg, L., and Refsgaard, J. C. Evaluation of Climate Input Biases and Water Balance Issues Using a Coupled Surface-Subsurface Model. *Vadose Zone Journal 10*, 1 (2011), 37.

[224] Sun, R., Yuan, H., and Liu, X. Effect of heteroscedasticity treatment in residual error models on model calibration and prediction uncertainty. *Journal of Hydrology* (2017).

[225] Systems, C. Collection Systems 2011. *Environment 1*, 60 (2011), 716–727.

[226] Székely, G. J., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Annals of Statistics 35*, 6 (2007), 2769–2794.

[227] Terenin, A., and Draper, D. Cox's Theorem and the Jaynesian Interpretation of Probability. *ArXiv e-prints* (July 2015).

[228] Thorndahl, S., Beven, K. J., Jensen, J. B., and Schaarup-Jensen, K. Event based uncertainty assessment in urban drainage modelling, applying the GLUE methodology. *Journal of Hydrology 357*, 3-4 (2008), 421–437.

[229] Todini, E. A Bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. *Hydrology and Earth System Sciences* (2001).

[230] Todini, E. A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management 6*, 2 (2008), 123–137.

[231] TOMST. *https://tomst.com/web/en/systems/tms/tms-3/*, 2017 (accessed February , 2017).

[232] Trenberth, K. E., Smith, L., Qian, T., Dai, A., and Fasullo, J. Estimates of the Global Water Budget and Its Annual Cycle Using Observational and Model Data. *Journal of Hydrometeorology 8*, 4 (2007), 758–769.

[233] Uhlenbeck, G. E., and Ornstein, L. S. On the theory of the Brownian motion. *Physical Review 36*, 5 (1930), 823–841.

[234] van Andel, S. J., Weerts, A., Schaake, J., and Bogner, K. Post-processing hydrological ensemble predictions intercomparison experiment. *Hydrological Processes 27*, 1 (2013), 158–161.

[235] van de Beek, C. Z., Leijnse, H., Stricker, J. N. M., Uijlenhoet, R., and Russchenberg, H. W. J. Performance of high-resolution X-band radar for rainfall measurement in The Netherlands. *Hydrology and Earth System Sciences 14*, 2 (2010), 205–221.

[236] van der Vaart, a. W. *Asymptotic Statistics*. 1998.

[237] Van der Waerden, B. L. Order tests for two-sample problem and their power I-III. *Indagationes Mathematicae 14,15* (1953), 453—-458 (14), 303—-316(15).

[238] Velasco-Forero, C. A., Sempere-Torres, D., Cassiraga, E. F., and Jaime Gómez-Hernández, J. A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. *Advances in Water Resources* (2009).

[239] Verdin, A., Rajagopalan, B., Kleiber, W., and Funk, C. A Bayesian kriging approach for blending satellite and ground precipitation observations. *Water Resources Research* (2015).

[240] Verkade, J. S., Brown, J. D., Davids, F., Reggiani, P., and Weerts, A. H. Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine. *Journal of Hydrology 555* (2017), 257–277.

[241] Verkade, J. S., and Werner, M. G. Estimating the benefits of single value and probability forecasting for flood warning. *Hydrology and Earth System Sciences* (2011).

[242] Vihola, M. Robust adaptive Metropolis algorithm with coerced acceptance rate, 2012.

[243] Villarini, G., Mandapaka, P. V., Krajewski, W. F., and Moore, R. J. Rainfall and sampling uncertainties: A rain gauge perspective. *Journal of Geophysical Research: Atmospheres 113*, 11 (2008), 1–12.

[244] Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S. A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research 39*, 8 (2003).

[245] Wagener, T., and Gupta, H. V. Model identification for hydrological forecasting under uncertainty. *Stochastic Environmental Research and Risk Assessment 19*, 6 (dec 2005), 378–387.

[246] Wallingford. Hr wallingford isis software. *Hydraulic Unit, WAllingford, UK* (1997).

[247] Wallingford. Wallingford water, a flood forecasting and warning system for the river soar. *Wallingford, UK* (1997).

[248] Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P. A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research 48*, 5 (2012), n/a–n/a.

[249] Wang, X., Gebremichael, M., and Yan, J. Weighted likelihood copula modeling of extreme rainfall events in Connecticut. *Journal of Hydrology 390*, 1-2 (2010), 108–115.

[250] Wani, O., Beckers, J. V., Weerts, A. H., and Solomatine, D. P. Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting. *Hydrology and Earth System Sciences 21*, 8 (2017), 4021–4036.

[251] Ward, P. J., Jongman, B., Weiland, F. S., Bouwman, A., Van Beek, R., Bierkens, M. F., Ligtvoet, W., and Winsemius, H. C. Assessing flood risk at the global scale: Model setup, results, and sensitivity. *Environmental Research Letters 8*, 4 (2013).

[252] Weerts, A. H., Winsemius, H. C., and Verkade, J. S. Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrology and Earth System Sciences 15*, 1 (jan 2011), 255–265.

[253] Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., and Heynert, K. The Delft-FEWS flow forecasting system. *Environmental Modelling and Software 40* (2013), 65–77.

[254] Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C. Y. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences 15*, 7 (2011), 2205–2227.

[255] Wilson, P. S., and Toumi, R. A fundamental probability distribution for heavy rainfall. *Geophysical Research Letters 32*, 14 (2005), 1–4.

[256] Wood, S. J., Jones, D. a., and Moore, R. J. Accuracy of rainfall measurement for scales of hydrological interest, 2000.

[257] Yang, J., Reichert, P., and Abbaspour, K. C. Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland). *Water Resources Research 43*, 10 (2007), n/a—-n/a.

[258] Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resources Research 39*, 7 (2003), 1–15.

[259] Yates, S. R., Warrick, A. W., and Myers, D. E. A disjunctive kriging program for two dimensions. *Computers and Geosciences 12*, 3 (1986), 281–313.

[260] Yates, S. R., Warrick, A. W., and Myers, D. E. Disjunctive Kriging: 1. Overview of Estimation and Conditional Probability. *Water Resources Research 22*, 5 (1986), 615–621.

[261] Yates, S. R., Warrick, A. W., and Myers, D. E. Disjunctive Kriging: 2. Examples. *Water Resources Research 22*, 5 (1986), 623–630.

[262] Zug, M., Phan, L., Bellefleur, D., and Scrivener, O. Pollution wash-off modelling on impervious surfaces: Calibration, validation, transposition. In *Water Science and Technology* (jan 1999), vol. 39, No longer published by Elsevier, pp. 17–24.

Curriculum Vitae

Omar Wani specializes in water science and engineering, with particular interest in the mathematical modelling of hydrologic systems. He has been a part of several collaborative research projects under the broader ambit of earth and environmental sciences. He likes to work at the intersection of science and application. Apart from research, he enjoys stimulating conversations, reading poetry, running along streams, hiking up to glaciers, and travelling in general.

Academic Qualification

2014 - 2018	PhD in Environmental Engineering, ETH Zürich, Switzerland
2012 - 2014	Master of Science with distinction - in Water Science and Engineering, IHE Delft Institute for Water Education, the Netherlands in Hydro Science and Engineering, TU Dresden, Germany
2008 - 2012	Bachelor of Technology in Civil Engineering, (emphasis - Hydraulic Engineering), Indian Institute of Technology, Roorkee, India

Professional Experience

2014 - 2018	PhD Researcher, Swiss Federal Institute of Aquatic Science and Technology, Eawag, Switzerland
2016 - 2018	Teaching Assistant, Summer School in Environmental Systems Analysis, Eawag, Switzerland
Nov 2017 - Jan 2018	Visiting Researcher, Division of Geological and Planetary Sciences, California Institute of Technology, USA
Feb 2016 - Apr 2016	Visiting Researcher, Water and Environmental Engineering Group, University of Bristol, UK
Feb 2014 - Aug 2014	Research Trainee, Deltares, the Netherlands

Awards

2014 - 2017	Marie Curie Fellowship
2012 - 2014	Erasmus Mundus Scholarship