DISS. ETH NO. 22875

Improving output and input statistical error descriptions in urban hydrological modeling

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

Presented by

Dario Del Giudice

Ing. env. dipl. EPF
born on 18.05.1988
citizen of Italy

accepted on the recommendation of

Prof. Dr. Eberhard Morgenroth, examiner
Dr. Jörg Rieckermann, co-examiner
Prof. Dr. Peter Reichert, co-examiner
Dr. Carlo Albert, co-examiner
Prof. Dr. Wolfgang Nowak, co-examiner

2015

Exploring the unknown requires tolerating uncertainty.
(Brian Greene)

It's much more interesting to live not knowing
than to have answers which might be wrong.
(Richard P. Feynman)

Abstract

Hydrological or drainage models can be valuable tools to support water management in urban environments. They can help to assess the characteristics of a catchment and to make predictions about its response (e.g. discharge). Unfortunately, parameters and results of these models are not perfect. These inaccuracies manifest themselves as model discrepancy, the so-called bias. Model bias represents the systematic deviation between model output and the real system response. There are two main causes of bias. First, there are errors in the input estimates, e.g. due to the insufficient coverage of pluviometers. Second, there are deficits in the model structure, e.g. due to the use of oversimplified empirical equations. Besides model bias, there are output measurement errors, e.g. due to imprecise flowmeters.

Until now, a large proportion of hydrological studies implicitly neglected input and structural errors. Neglecting these can lead to: i) misrepresented output measurement errors, ii) incorrect parameter estimates which compensate for model discrepancy, and iii) predictions which underestimate the uncertainty. Furthermore, these approaches do not provide guidance for finding the reasons for bias. Therefore, they cannot support its reduction. This inappropriate error consideration can finally lead to faulty risk assessment, flawed evaluation of decision alternatives, and, consequently, impaired urban water management. More advanced error assessment techniques are therefore called for.

The main goal of this thesis is to better represent input, structural, and output measurement errors. This can meliorate parameter estimation and prediction generation. The main thesis contributions are two. The first is a realistic description of the uncertainties in stormwater, wastewater, and sediment transport predictions. The second is a sound representation of errors in the rainfall inputs. Additionally, we discuss and compare different statistical methods used for hydrological inference. Finally, we also present a new tool to improve sewer flow monitoring.

In the first part of this work, we adapt a bias description, originally coming from statistics, to runoff modeling. We represent model bias as an autocorrelated Gaussian process and propose different parameterizations of this process. This "upgraded" bias description is tested using a parsimonious model of a small stormwater catchment. Results show that accounting for bias makes runoff predictions more reliable (i.e. realistically more uncertain) than before.

In part two, the bias description is further examined in a different case study consisting of a large wastewater catchment. We compare this error model with an alternative one. This second approach includes the bias within the model equations rather than in the output. This analysis corroborates our previous findings: the bias description can appropriately

quantify runoff uncertainties over several days ahead, even when the underlying model is overly simplistic. Furthermore, the method can even improve the accuracy and precision of short-term predictions. Finally, we discuss several theoretical and practical issues for optimally describing the bias of urban hydrological models.

In part three, we analyze the relation between model bias and model complexity. We here compare the bias of multiple stormwater models with a gradient of complexity. Results show that analyzing the bias reduction from the simplest to the most complex and accurate model can effectively quantify the decrease of structural error. Studying the bias of the least biased model can also approximately indicate the contribution of input error to output uncertainty.

The fourth part of this works, finally explores in depth the reasons for bias. Here, we use an autocorrelated Gaussian process to describe and reduce input uncertainty. This novel stochastic input process (SIP) represents the catchment-averaged precipitation. SIP is examined using an accurate sewer model forced with inaccurate rainfall data. Results demonstrate that the method, when compared to previous error models, can help to preserve the physical meaning of model parameters and generate reliable predictions. Furthermore, SIP can estimate the precipitation more realistically than before. These enhanced performances, however, come at a higher computational cost.

The findings from this research show that we have two options to improve the uncertainty quantification in urban hydrology. The simpler alternative is to correct modeling errors at the output. In this thesis we have proposed a statistical bias description to do that. By making predictions more reliable, this approach might be favored in engineering-oriented studies, focusing for instance on decision support. The second alternative is to describe and reduce the sources of errors. In this thesis we developed a method to assess input errors. This more advanced contribution might be particularly relevant for studies aiming at developing new methods and/or understanding more in depth the drainage system.

Zusammenfassung

In der Hydrologie oder Stadtentwässerung sind Modelle wertvolle Hilfsmittel um wasserwirtschaftliche Probleme zu lösen. Sie können helfen die hydrologischen Eigenschaften eines Einzugsgebiets abzuschätzen und dessen Niederschlags-Abfluss-Verhältnisse vorherzusagen. Allerdings sind weder die Parameter noch die Vorhersagen solcher Modelle perfekt. Es resultieren systematische Abweichungen zwischen der realen Reaktion des Systems und den Ausgangsgrössen des Modells, sogenannte Modellfehler. Modellfehler haben hauptsächlich zwei Ursachen: Erstens gibt es Fehler in der Schätzung der Eingangsgrössen des Niederschlag-Abfluss-Modells, z.B. aufgrund ungenügender räumlicher Abdeckung eines Einzugsgebiets mit Niederschlagsmessungen. Zweitens gibt es Modellstrukturfehler, z.B. aufgrund der Verwendung zu stark vereinfachte empirische Gleichungen. Weiterhin stellen Messfehler in den Ausgangsgrössen, z.B. aufgrund unpräziser Durchflussmessungen eine Herausforderung bei der Systemanalyse dar.

Leider vernachlässigen praktisch alle Arbeiten in der (städtischen) Hydrologie immer noch die angesprochenen Fehler in den Eingangsdaten und der Modellstruktur. Das führt dazu, dass Messfehler in den Ausgangsdaten ungenau abgebildet werden, verzerrte Parameterwerte geschätzt und infolgedessen unzuverlässige Prognosen gemacht werden. Somit werden weder spezifische Ursachen von Modellfehlern festgestellt noch wird angestrebt diese zu reduzieren. Zusammenfassend ist diese Art der Berücksichtigung von Ungewissheiten unbefriedigend, da sie zu fehlerhaften Risikoabschätzungen führen kann. Für die Bewertung von Entscheidungsalternativen ist sie also mangelhaft, und beeinträchtigt die Siedlungswasserwirtschaft. Um diese Unzulänglichkeiten zu überwinden müssen wir deshalb belastbarere Methoden zur Fehlerschätzung entwickeln.

Das Hauptziel dieser Dissertation ist, Fehler in den Eingangsgrössen unserer Modelle, in deren Struktur und in den Ausgangsgrössen besser abzubilden. Insbesondere leistet diese Arbeit zwei wesentliche wissenschaftliche Beiträge: Erstens, eine realistischere Beschreibung der Unsicherheit von Niederschlags-Abfluss-Prognosen, für Regenabwasser, Mischabwasser und Feststofftransport, und eine bessere Beschreibung der Fehler in Regendaten, als wichtige Eingangsinformation für die Modelle. Ausserdem werden i) statistische Methoden für die Kalibrierung von hydrologischen Modellen verglichen und diskutiert und ii) eine neue Methode vorgeschlagen, die natürliche Tracer im Abwasser benutzt, um die Qualität von Durchfluss-Messungen zu überprüfen.

Im ersten Teil dieser Arbeit wurde eine Methode entwickelt um Modellfehler explizit zu berücksichtigen. Dazu wurde eine Beschreibung des Modellfehlers aus der angewandten

Statistik angepasst, bei der der Modellfehler als additiver autokorrelierter Gauss-Prozess berücksichtigt wird. Insbesondere haben wir neue Parametrisierungen vorgeschlagen um die Unsicherheiten in der Niederschlag-Abfluss-Prognose realistischer abzubilden. Anhand eines Fallbeispiels für ein Regenabwassersystem wurde gezeigt, dass die Abflussprognosen deutlich zuverlässiger sind, wenn der Modellfehler statistisch beschrieben wird. Typischerweise werden auf diese Weise Prognosen mit grösseren Unsicherheiten erzeugt im Vergleich zu solchen, die mit herkömmlichen Methoden gemacht werden. Allerdings spiegelt dies besser den aktuellen Kenntnisstand wieder und führt letztlich zu besseren siedlungswasserwirtsachaftlichen Entscheiden.

Im zweiten Teil untersuchen wir die Beschreibung des Modellfehlers anhand eines weiteren Fallbeispiels, diesmal für ein grösseres Teileinzugsgebiet, das im Mischsystem entwässert wird. Hier werden die Niederschlags-Abfluss-Prognosen mit dem additiven stochastischen Modellfehler erstmals mit einem alternativen Fehlermodell verglichen. Dieses Fehlermodell ist interessant, weil es den Modellfehler als Unsicherheit in Bezug auf den System-Zustand innerhalb der Modellgleichungen beschreibt, anstatt über einen additiven Fehler in den Ausgangsgrössen. Die Resultate der Untersuchungen bestätigen im Wesentlichen unsere vorherigen Erkenntnisse. Erstens kann ein additiver stochastischer Modellfehler die Unsicherheiten in Niederschlags-Abfluss-Prognosen angemessen über mehrere Tage in die Zukunft beschreiben, selbst wenn ein einfaches hydrologisches Modell verwendet wird. Zweitens kann unsere Methode die Genauigkeit und Präzision von kurzfristigen Prognosen verbessern. Darüber hinaus werden in diesem Methodenvergleich mehrere theoretische und praktische Themen diskutiert um Modellfehler in der Siedlungshydrologie optimal zu beschreiben.

Im dritten Teil der vorliegenden Arbeit wird die Beziehung zwischen Modellfehler und Modellkomplexität untersucht. Insbesondere wird analysiert, wie der additive stochastische Modellfehler von der Komplexität des Niederschlag-Abfluss-Modells abhängt. Die Ergebnisse zeigen erwartungsgemäss, dass der Modellfehler vom einfachsten bis zum komplexesten Modell abnimmt. Das kann ein effektives Mittel sein, um die Verminderung der strukturellen Fehler zu quantifizieren. Ausserdem kann über eine solche Untersuchung des Modellfehlers abgeschätzt werden, was der Beitrag der Fehler in den Eingangsdaten an der Prognoseunsicherheit ist.

Im vierten Teil dieser Arbeit wird schliesslich eine neue Methode vorgeschlagen um Fehler in den Eingangsdaten verlässlicher zu beschreiben als bisher. Diese zielt darauf ab, den Modellfehler nicht nur phänomenologisch zu beschreiben, sondern Hinweise auf mögliche Ursachen zu bekommen. Diese Information könnte in einem späteren Schritt genutzt werden um Fehler durch eine verbesserte experimentelle Ausgestaltung zu reduzieren, beispielsweise durch verbesserte Erhebung von Niederschlag. Hier haben wir uns auf eine verbesserte Beschreibung der Fehler in Regendaten fokussiert und verwenden dazu einen autokorrelierten stochastischen Gauss-Prozess. Dieser neue stochastische Fehlerprozess für Eingangsdaten (engl., ,,stochastic input prozess": SIP) beschreibt den flächengemittelten Niederschlag auf das Einzugsgebiet, wie er von den meisten Modellen verwendet wird. In einem Fallbeispiel benutzen wir die SIP-Methode um das Verhalten eines kleinen Entwässerungsnetzes zu beschreiben, wobei wir zwar eine gute Modellstruktur, jedoch fehlerhafte Regendaten verwenden. Die Resultate zeigen, dass die Methode dazu beiträgt, die physikalische Bedeutung der Modellparameter zu erhalten und so zuverlässige Prognosen zu generieren. Die bisher vorgeschlagenen Ansätze führen zu

verzerrten Parameterwerten. Allerdings erfordert SIP auch einen ein bis zwei Grössenordnungen grösseren Rechenaufwand.

Schlussfolgernd lässt sich sagen, dass es grundsätzlich zwei attraktive Möglichkeiten gibt, um die Quantifizierung von Unsicherheiten in der Siedlungsentwässerung zu verbessern. Die einfachste Alternative ist, Modellfehler in den Ausgangsdaten über einen additiven stochastischen Prozess zu korrigieren. In dieser Dissertation wird dazu eine statistische Beschreibung des Modellfehlers vorgeschlagen und erfolgreich in mehreren Fallbeispielen angewedet. Da dieser Ansatz die Prognose zuverlässiger macht und relative einfach anzuwenden ist, kann er für praktische Ingenieurprojekte angewendet werden. Die zweite Alternative ist, die massgeblichen Fehlerquellen möglichst realistisch zu beschreiben und, wenn möglich, in einem nachfolgenden Schritt zu reduzieren. In dieser Dissertation wird dazu eine Methode vorgeschlagen, um den Fehler in Regendaten so verlässlich wie möglich zu schätzen. Dieser komplexere Ansatz kann für wissenschaftliche Untersuchungen verwendet werden, die sich auf die Entwicklung von neuen Methoden konzentrieren, und die versuchen den Niederschlags-Abfluss-Prozess besser zu verstehen.

Acknowledgments

I want to sincerely express my gratitude towards all the fantastic people that accompanied and supported me along this exciting journey. Primarily, I want to say a big thank you to Jörg Rieckermann, my direct supervisor, for his instructive feedback, for striving to find an optimal balance between giving me guidance and freedom, and, especially, for caring about my development as a complete person. I also want to deeply thank my other supervisors, Peter Reichert, Eberhard Morgenroth, and Carlo Albert for the willingness to share their knowledge and for the faith in my work. Each of them fostered my development in a unique way, providing me with invaluable tools to navigate the scientific world. I am also very grateful to my co-examiner Wolfgang Nowak who, not only contributed with several helpful advices about my work, but also acted as a trusted mentor when discussing my career development. I would additionally like to thank Max Maurer for having backed my education with his advices and the sponsoring of several courses.

At Eawag, and in particular within SWW/Eng, I experienced a congenial and helpful atmosphere. For that, I want to thank the whole Eawag team and particularly the reception, aQa, and Lib4ri staff, the IT, HR, Finances, and Cluster supporters. A special thanks goes to my doctoral fellows Omar Wani, David Machac, Christoph Egger, Sven Eggiman, Dorothee Spuhler, Michele Laureni, Lena Mutzner, Ann-Kathrin McCall, Hanspeter Zöllig, Alexandra Fumasoli, Johny Habermacher, Basil Thalmann, Anna Chomiak, Christian Thürlimann, Pascal Wunderlin, Andrea Portmann, Peter Desmond, Mariane Schneider, and other colleagues like Tobias Doppler, Christoph Ort, Kai Udert, João Leitão, Frank Blumensaat, Philipp Beutler, Alma Masic, Anne Dietzel, Lisa Scholten, Nele Schuwirt, Raoul Schaffner, Canan Aglamaz, Fabrizio Fenicia, Simone Ulzega, James Irving, Gregoire Mariethoz, Alberto Montanari, Ariane Eberhardt, Claire Wedema, Martin Fencl, and Nicolas Derlon. Each of them made my journey better in a special way.

I want to gratefully acknowledge my coauthors Ania Sikorska, Mark Honti, David Dürrenmatt, Vojtěch Bareš, Andreas Scheidegger, Roland Löwe, Peter S. Mikkelsen, Kris Villez, Marc Neumann, and Henrik Madsen. I have learned several valuable lessons from our fruitful collaboration. I also express my gratitude to Cristina Rachelly, Rao Fu, Patrick Kornberger, Daniel Eilertz, Theresa Rossboth, and the Adliswil wastewater operators, for their support in our field study.

Finally, my parents from the sunny shores of Italy deserve my profound appreciation for their sincere love. Grazie.

Contents

Chapter 1

Introduction

1.1 Preamble: challenges in model-based urban water management

The relationship between aquatic environments and human activities is an ambivalent one. On the one hand, we need clean water for drinking, sanitation, irrigation, and leisure. On the other hand, as a result of these activities, we release toxic waste into water bodies, which negatively impacts water quality, and, consequently, future usage of aquatic resources along with human health. Additionally, to facilitate these exchanges of water, human settlements are usually located around or within freshwater systems. This proximity, however, can become dangerous when the system does not respond as predicted. An example of that are flood events in urban areas caused by precipitations higher than the designed capacity of the drainage network.

To face these challenges, water quantity and quality models have commonly been used (Clarke, 1973; Reckhow, 1994b; Beven, 2011). Computer codes can help to estimate the properties of the underlying environmental system and to predict its behavior under future conditions (Omlin and Reichert, 1999). In urban systems, which are in a delicate situation due to their high population density, models are particularly useful to deal with hydrologic challenges (Dotto et al., 2012). These urban drainage models (UDMs) are essential to produce runoff predictions and storm water quality estimates (Vezzaro et al., 2013b), to assess the effectiveness of proposals of sewer system redesign or upgrade (Breinholt et al., 2013), to test whether various management strategies can meet desired water quality standards (Zoppou, 2001), to support real-time control of storage basins via optimal regulation of gates and pumps, and to estimate the risk of sewage emissions into natural waters (Löwe et al., 2013). In other words, models strive to represent a drainage systems and its response to different conditions in order to answer questions about it (Butler and Davies, 2010). Examples of those questions are "what will the discharge at the sewer outlet be during a storm?" and "what are the hydrologic consequences of city expansion and changes in rainfall extremes?".

Environmental models such UDMs, however, are neither able to represent the underlying system exactly nor to predict its dynamics in a perfect way (Box, 1976; Beck and Young, 1976; Schilling and Fuchs, 1986; Deletic et al., 2011). Instead, UDMs are affected by several types of errors which can considerably impair their predictive ability and, consequently, the decision

making process based on these models (Reckhow, 1994a; Freni and Mannina, 2010).
First, computer codes can only crudely approximate the reality of a complex system such as an urban watershed, with a variety of land uses, drainage pipe materials, overland and subterraneous flow paths, and wetting conditions. These oversimplifications are known as model structural deficits (Gupta et al., 2012).
Second, the rainfall forcing a hydrologic response of the watershed can be difficult to estimate (Schilling, 1991; Berne et al., 2004; Sikorska et al., 2012b). Indeed, rainfall inputs are temporally and spatially varying precipitation fields which can be sampled only in specific locations. These "sampling errors" linked to the spatial interpolation of the measurements, together with "measurement errors" associated to the recording devices, can lead to inaccurate and imprecise input data (McMillan et al., 2011). Although neglected by most studies which use measured input (Beven and Young, 2013), in real-time runoff forecasts, the uncertainty associated with rainfall extrapolations in the future additionally plays a role (Löwe et al., 2014b).
Third, model equations include constant values, parameters, which might represent some aggregated properties of the real catchment to be estimated by inference from output data (Beck, 1994). Model parameters, however, can never be determined with certainty, mainly due to inadequate input data and model formulation and due to incomplete output data needed for their estimation (McLean and McAuley, 2012). Furthermore, environmental output data used for parameter inference are usually randomly and systematically erroneous (McMillan et al., 2012; Reichert and Schuwirth, 2012). This has been extensively acknowledged when dealing with flow observations in open channels (Montanari and Di Baldassarre, 2013; Sikorska et al., 2013) and wastewater pipes (Dürrenmatt et al., 2013) and when measuring suspended solids in rivers (Rode and Suhr, 2007).

Due to the aforementioned error sources, UDM predictions are uncertain (Freni et al., 2009b). In environmental and urban hydrological modeling, this uncertainty is usually considered by "disturbing" the model output with a white noise (Kleidorfer, 2009a; Dotto et al., 2011). This procedure is equivalent to assuming that all error sources finally produce a simple output error which is uncorrelated, normally distributed, and constant (Reichert and Mieleitner, 2009). Starting from the 1970s, researchers acknowledged that this simplistic error representation produces biased model results, mainly manifesting as biased parameter estimates and overconfident (i.e. falsely too narrow) predictions (Clarke, 1973; Beck and Young, 1976; Sorooshian and Dracup, 1980; Kuczera, 1983). Additionally, this lumped approach cannot support the understanding and reduction of the error causes, since it precludes error separation into its individual contributions (Yang et al., 2007b; Honti et al., 2013). In recent years, these pitfalls have been repeatedly recognized in hydrology, including urban drainage modeling (Muleta et al., 2013). So far, however, uncorrelated error representations are still the prevalent ones (Freni and Mannina, 2010; Dotto et al., 2012). Yet, in order to draw appropriate conclusions about model parameters and predictions, a realistic description of output errors, or even better, an explicit consideration of the error sources is required (Vrugt et al., 2008; Dietzel and Reichert, 2012).

1.2 Promising statistical solutions

1.2.1 Correcting the symptoms of model errors

From the branch of statistics dealing with error models, Bayesian inference, and Gaussian processes, a promising error description recently appeared (Kennedy and O'Hagan, 2001; Craig et al., 2001; Higdon et al., 2005; Bayarri et al., 2007). It consists in representing the errors as a sum of an uncorrelated term, accounting for random measurement noise, and an autocorrelated one. The latter, also called "bias process", incorporates the systematic deviations of model results from output data and accounts for the effects of input and structural errors. Similar approaches trying to more realistically represent the autocorrelated behavior of modeling errors, however, already appeared few decades earlier in the hydrological modeling literature (Clarke, 1973; Sorooshian and Dracup, 1980; Kuczera, 1983). The main advantage of the statistical bias description over these pioneering efforts is that it is formulated as a continuous process which helps to discriminate between i) the combined effects of input and structural errors, ii) parametric uncertainty, and iii) white observation "noise". From initial experiments, this error description appears to appropriately describe the uncertainties of simple inaccurate models involving the exponential decay of a substance (Bayarri et al., 2007) or the microbial growth in a mixed tank reactor (Reichert and Schuwirth, 2012). It is not clear, however, whether this error representation is applicable to the hydrological and hydraulic models used in urban catchments. These UDMs have to deal with systems that quickly and dramatically respond to precipitation (Coutu et al., 2012b) and that are more complex than stationary and idealized chemical reactors. Urban catchments are mostly ephemeral and thus range among those hydrologic systems for which an appropriate error description is particularly challenging (Evin et al., 2014). Consequently, this promising statistical technique is going to require a substantial adaptation from the simple representation used in the initially-tested toy models. Indeed, the errors characterizing sewer runoff simulations are not only autocorrelated but also display a strong variability in variance which results from the irregular succession of wet and dry-weather periods. Besides these open conceptual aspects, it is also not granted that basic inference methods using classic Markov chain Monte Carlo (MCMC) algorithms (Metropolis et al., 1953), will work here. Instead, in non-trivial circumstances such as in parameter inference for hydrological models, more sophisticated strategies might be required (Vrugt et al., 2009b). Furthermore, besides adopting more advanced inference methods, the use of fast surrogate models (statistical emulators) may be required to accelerate the computations of slow models such as complex UDMs (Reichert et al., 2011; Albert, 2012). Finally, once adapted to UDM, the theoretical and practical advantages and limitations of the bias description will need to be elucidated i) by investigating it in dissimilar systems and ii) by comparing its performances in inference and predictions with other methods for uncertainty assessment (e.g. Dotto et al. (2011); Breinholt et al. (2012); Kavetski et al. (2006)).

1.2.2 Towards understanding the causes of model errors

While the statistical description of model bias displays the potential to improve prediction reliability (related to a sufficient coverage of the validation data), it is not straightforward how much it can help improve model parameter estimation and quantify the causes of bias (Bayarri et al., 2007; Higdon et al., 2005; Reichert and Mieleitner, 2009; Dietzel and Reichert, 2012).

Instead, to go a step further and describe the errors where they occur can have additional advantages of paramount importance (Renard et al., 2010). In the case of rainfall input forcings, for instance, it can help to understand what the contribution of that source in different weather conditions is, and most importantly, how to correct the rainfall estimation to ensure a more realistic parameter calibration (Renard et al., 2011). Furthermore, separately treating input uncertainty makes it possible to disentangle its contribution to the output uncertainty. This is useful to assess in how far prediction uncertainty can be reduced by reducing a particular error source and therefore to guide our efforts to minimize the uncertainties (Sikorska et al., 2012b).

A description of the sources of errors in a rigorous statistical framework, although highly valuable, is currently missing in UDM (Del Giudice et al., 2013). In the neighboring natural hydrology, however, statistical error models have been proposed to simultaneously quantify the errors at their sources and their final effects downstream of the modeling chain (Honti et al., 2013). Attempting to quantify input errors, for instance, scaling the precipitation measurements with event-specific parameters (rainfall multipliers) has become popular in runoff modeling (Kavetski et al., 2006; Vrugt et al., 2008). These procedures assume that, for every precipitation event, the true rainfall over the catchment is proportional to the measured one via a factor to infer (Sun and Bertrand-Krajewski, 2013). Although this method works in simple situations of random input measurement noise or of bias proportional to the measured values, it can run into problems when facing realistic conditions. For instance, in case of a recorded event with temporal dynamics substantially different from the "true" one, this method will fail to appropriately quantify input uncertainty. This non-trivial input bias can occur when a pluviometric station, being located far from the catchment centroid, misses a storm completely but the flowmeter still records an increased runoff at the outlet.

The open problem of dynamic biases in rainfall estimates representative for the whole catchment requires a different, more realistic solution. One promising possibility from the field of statistics consists in describing a priori the catchment input as a Gaussian process in an appropriate space (Sigrist et al., 2012). Via the use of available rainfall data, a hydrological model, and flow (i.e. output) data, such a process could be updated to reflect to true probabilistic precipitation over the catchment. This idea is appealing because, contrary to the multipliers, it does not assume linear dependence between the recorded and the true precipitation. Therefore, it has the potential to provide a more realistic rainfall and parameter estimation in cases of complex input measurement biases. However, it is still unclear how to exactly parameterize the rainfall process and how to estimate it numerically in a computationally feasible way. Additionally, it remains to be explored how this prototype will perform in a real urban hydrological case and if its practical advantages reflect the theoretical benefits.

1.2.3 Summary of remaining modeling challenges and possible solutions

In summary, environmental sciences and urban hydrology are facing a difficult situation: drainage model predictions are necessary for better risk assessment, decision making, and water management, but model results cannot be completely relied upon. In this context, two particularly pressing needs are i) to obtain a realistic representation of the uncertainties associated with model results and ii) to quantify the reasons for these uncertainties, especially related to

rainfall inputs, in order to effectively meliorate the quality of model predictions. Current approaches used in UDM, being overly simplistic, do not provide an adequate solution to those challenges. The need for a more realistic and informative uncertainty quantification could be satisfied instead by transferring promising errors from statistics. In particular, a bias description has the potential to produce more reliable UDM predictions (need i)), whereas a stochastic input process seems to be the key to improve predictive accuracy by suitably quantifying and reducing modeling errors induced by rainfall inputs (need ii)). The transferral of these techniques from statistics to UDM, however, is not going to be an easy task, since it implies substantial field-specific conceptual and practical adaptations together with tests in real case studies.

1.3 Goals, novelty, and contribution

As mentioned above, several problems related to the uncertainty inherent in UDMs can be mitigated by the adaptation and use of advanced statistical techniques. The general goal of this thesis is to contribute to a reliable application of environmental models used in urban water management by improving the quantification of their uncertainties. In particular, the aim is to adapt and further develop innovative methods to realistically assess and possibly reduce the uncertainties of environmental predictions, with a focus on UDMs.

1.3.1 Objectives and research questions

The specific objectives of this work are:

I. Investigate how to transfer and adapt a description of model output bias from applied statistics to ensure reliable predictions of urban runoff.

II. Explore how an appropriately-parametrized bias description performs in several systems and in comparison to statistical techniques currently applied in urban hydrology.

III. Understand how informative can a bias process be with respect to its causes, namely input and structural errors.

IV. Develop and test a probabilistic input description which, even in presence of severe rainfall errors, allows for improved inference of rainfall and parameters, and permits to separate input uncertainty from other uncertainty contributions.

Associated to these goals, several research questions will be addressed:

Q1. Is a statistical bias description a conceptually and practically effective tool to improve the uncertainty assessment in urban hydrology with respect to traditional calibration methods?

Q2. What are the benefits and limitations of the statistical bias description compared to existing inference methods?

Q3. What is the most appropriate bias descriptions for robust and reliable sewer flow predictions?

Q4. What numerical scheme is suitable for a Bayesian inference with consideration of model bias?

Q5. Can the bias description preserve the physical meaning of model parameters and therefore "protect" the inference from the corrupting effect of input and structural errors?

Q6. How can we maximize the understanding of the bias and how much can we learn about its causes without describing the error sources?

Q7. What is an appropriate way to stochastically describe and infer rainfall uncertainty in hydrological model calibration from a conceptual and numerical perspective?

Q8. What are the advantages and disadvantages of stochastic rainfall description compared to a bias description and to the previous method employing rainfall multipliers?

Q9. How can a stochastic rainfall description be applied when the underlying model is structurally deficient and the goal is to disentangle the contribution of input errors from structural deficits?

1.3.2 Novelties and contributions

The main innovative contributions of this research can be summarized as follows.

- A statistically sound formulation and investigation of model bias in urban hydrology. This makes it possible to obtain reliable runoff predictions in a variety of case studies even in presence of input and structural errors.
- A rigorous and realistic consideration of input uncertainty in hydrological inference. This should help assess and reduce rainfall errors, quantify their effects on model predictions and improve model parameter estimation.
- An advancement towards the comprehension and disentanglement of the uncertainty components by presenting several complementary tools for uncertainty assessment. This is useful to better understand the causes of the uncertainties in order to more effectively improve model predictions.
- A research-based formulation of recommendations to guide future studies in urban hydrology involving parameter estimation and uncertainty assessment.

1.4 Outline of the thesis

This work proceeds from a description of output uncertainty to a description and minimization of input error sources. This involves using increasingly complex techniques. This is necessary to move from a "end-of-pipe" symptom correction of the errors to an in-depth understanding of the causes of errors able to support their eventual reduction (Yang et al., 2008; Reichert and Mieleitner, 2009; Salamon and Feyen, 2010; Renard et al., 2010; Honti et al., 2013).

Chapter 2 shows how to reliably describe hydrological model bias, the systematic output deviations resulting from input and structural errors. Since the bias magnitude and temporal variability can substantially differ from one catchment to another, we propose different parameterizations of the bias which should be of help in other case studies. After analyzing stormwater predictions of a small catchment, we discuss the advantages of the bias description over traditional simplified methods.

In order to better understand the suitability of the bias description and to put its performances into perspective, in Chapter 3 we compare the technique with a similar one currently applied

in urban hydrology. This alternative approach to consider input and structural errors comes from control theory rather than statistical inference. We test the methods in a large combined sewer. Given the results of the case study and theoretical considerations, we give suggestions on the optimal field of application for the bias description.

In Chapter 4 we then take a step forward, not only describing but also trying to gain an understanding about the behavior of the bias with drainage models of different complexities. By combining the bias description with a multimodel comparison, we focus on interpreting the bias evolution when model complexity increases. This is useful to comprehend how far can an output error model be informative about the error sources. We discuss the maximal information extractable by the output analysis and the limitations that can be overcome only by describing the error sources. Additionally, we show that the bias description can be successfully applied with models having multiple outputs.

We finally move to a more in-depth uncertainty analysis in Chapter 5, this time directly modeling the error sources with a statistical approach. Here we focus on input errors for which we present a novel representation as a continuous stochastic process. By using a small combined sewer as a test case, we show how this approach outperforms others in terms of parameter and input estimation in situations of erroneous rainfall measurements. Besides discussing the conceptual appeal of such an inference method we also point out practical considerations linked to its computational cost and underlying assumptions.

Overall conclusions are drawn in Chapter 6 which includes a discussion of the main lessons learned, the answers to the initially-raised questions, and an overview of recommended future research directions.

In Appendix A we use, for the first time, a modern data-mining technique called “dynamic time warping” (DTW) to estimate runoff dynamics from inexpensive temperature measurements. Results show that DTW accurately quantifies the flow velocity in a wastewater catchment under a variety of flow conditions. These findings suggest that DTW can improve the assessment and the reductions of output measurement errors in urban hydrology.

Appendix B demonstrates that a statistical bias description can be a valuable tool, not only to improve sewer flow predictions, but also in other environmental modeling endeavors. In particular, here we analyze the evolution of suspended particles in an urbanized river basin. The outcomes of this investigation show that describing the bias can simultaneously improve the reliability of water quality and quantity predictions.

Being structured as cumulative dissertation, Chapters 2, 3, 4, 5, and Appendices A, and B of this thesis correspond to peer-reviewed, in revision or ready-to-submit papers. In particular, Chapters 2, 3, 4, and Appendix A have been published, Appendix B is under review, and Chapter 5 is ready for submission in July 2015. A statement of authors’ contributions is provided at the beginning of each paper.

Chapter 2

Improving uncertainty estimation in urban hydrological modeling by statistically describing bias

D. Del Giudice [a,b], M. Honti [a], A. Scheidegger[a], C. Albert[a], P. Reichert[a,b], J. Rieckermann[a].

[a]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b]ETHZ: Swiss Federal Institute of Technology Zürich, 8093 Zürich, Switzerland

Hydrology and Earth System Sciences (2013), 17, 4209-4225, doi:10.5194/hess-17-4209-2013.

Author contributions

D.D.G. designed the experiments, built the model, further developed and applied the statistical methods, performed the analyses, wrote the paper; M.H., A.S., C.A., and P.R. supported the development and application of the statistical methods; D.D.G., P.R., and J.R. conceived the experiments; all coauthors gave advices, supported result interpretation and paper revision.

Abstract

Hydrodynamic models are useful tools for urban water management. Unfortunately, it is still challenging to obtain accurate results and plausible uncertainty estimates when using these models. In particular, with the currently applied statistical techniques, flow predictions are usually overconfident and biased. In this study, we present a flexible and relatively efficient methodology (i) to obtain more reliable hydrological simulations in terms of coverage of validation data by the uncertainty bands and (ii) to separate prediction uncertainty into its components. Our approach acknowledges that urban drainage predictions are biased. This is mostly due to input errors and structural deficits of the model. We address this issue by describing model bias in a Bayesian framework. The bias becomes an autoregressive term additional to white measurement noise, the only error type accounted for in traditional uncertainty analysis in urban hydrology. To allow for bigger discrepancies during wet weather, we make the variance of bias dependent on the input (rainfall) or/and output (runoff) of the system. Specifically, we present a structured approach to select, among five variants, the optimal bias description for a given urban or natural case study. We tested the methodology in a small monitored stormwater system described with a parsimonious model. Our results clearly show that flow simulations are much more reliable when bias is accounted for than when it is neglected. Furthermore, our probabilistic predictions can discriminate between three uncertainty contributions: parametric uncertainty, bias, and measurement errors. In our case study, the best performing bias description is the output-dependent bias using a log-sinh transformation of data and model results. The limitations of the framework presented are some ambiguity due to the subjective choice of priors for bias parameters and its inability to address the causes of model discrepancies. Further research should focus on quantifying and reducing the causes of bias by improving the model structure and propagating input uncertainty.

2.1 Introduction

Mathematical simulation models play an important role in the design and assessment of urban drainage systems. On the one hand, they are used to investigate the current system, for example regarding the capacity for and likelihood of flooding. On the other hand, engineers use them to predict the consequences of future changes of boundary conditions or control strategies (Gujer, 2008; Kleidorfer, 2009b; Korving and Clemens, 2005). Traditionally, according to standards of good engineering practice, such models were calibrated by adjusting parameters to allow predicted flows to closely reflect field data. In recent years, it has been suggested that predictions of urban drainage models are not of much practical use without an estimate of their uncertainty (Dotto et al., 2011; Kleidorfer, 2009b; Korving and Clemens, 2005; Reichert and Borsuk, 2005). Unfortunately, there are so far no established methods available to assess prediction uncertainty in sewer hydrology in a statistically satisfactory way (Freni et al., 2009b; Breinholt et al., 2012). In the context of design, operation and assessment of urban hydrosystems, it is important to obtain reliable predictions from a calibrated model (Sikorska et al., 2012b). This means that random draws from the model should have similar statistical properties (such as variance or autocorrelation) as the data. Additionally, for reliable predictions, the observed coverage of the simulated uncertainty bounds should match or exceed the nominal coverage. Ideally, this can be achieved by representing the dominant sources of uncertainty explicitly in the model. This could be done by considering uncertainty in (i) model parameters, (ii) measured outputs, (iii) measured

inputs and (iv) the model structure and by propagating these uncertainties to the model output.

While there have been some attempts to formulate a sound "total error analysis framework" in natural hydrology (Kavetski et al., 2006; Vrugt et al., 2008; Reichert and Mieleitner, 2009; Montanari and Koutsoyiannis, 2012), applications in urban hydrology are lacking, probably due to the complexity of these approaches. Instead, it is usually (often implicitly) assumed, first, that the model is correct and, second, that residuals, i.e. the differences between model output and data, are only due to white measurement noise (Breinholt et al., 2012; Dotto et al., 2011). Furthermore, these observation errors are considered to be identically (usually normally) and independently distributed (iid) around zero (Willems, 2012). Unfortunately, these are very strong assumptions in urban hydrology, where processes are faster than in natural watersheds, spatial heterogeneity of precipitation may have a bigger effect (Willems et al., 2012), and rainfall-runoff can increase by several orders of magnitude within a few minutes. This "flashy" reaction can be challenging to reproduce correctly in time and magnitude with current computer models and precipitation measurements (Schellart et al., 2012). In addition, sewer flow data have a high resolution of a few minutes and are usually more precise than those of natural channels. Having temporally dense and precise measurements exacerbate the effects of systematic discrepancies between model outputs and data (Reichert and Mieleitner, 2009). If such model bias, mainly induced by input and structural errors, is not properly accounted for, autocorrelated and heteroskedastic residual errors and overconfident (i.e. too narrow) uncertainty intervals are generated (Neumann and Gujer, 2008).

To better fulfill the statistical assumptions of homoskedasticity and normality of calibration residuals, and so obtain more reliable predictions, a commonly applied technique in hydrology is to transform simulation results and output data. The Box-Cox transformation (Box and Cox, 1964) has indeed been successfully used in several case studies, both rural (e.g., Kuczera, 1983; Bates and Campbell, 2001; Yang et al., 2007b,a; Frey et al., 2011; Sikorska et al., 2012b) and urban (e.g., Freni et al., 2009b; Dotto et al., 2011; Breinholt et al., 2012). Admittedly, transformation stabilizes the variance of the residual errors in the transformed space. Unfortunately, it has almost no effect on the serial autocorrelation of residuals and thus cannot capture model bias.

To account for systematic deviations of model results from field data, it seems promising to apply autoregressive error models that lump all uncertainty components into a single process (Kuczera, 1983; Bates and Campbell, 2001; Yang et al., 2007b; Evin et al., 2013). Such models are not only relatively straightforward to apply, but also often help to meet the underlying statistical assumptions. However, a disadvantage of such lumped error models is that only parameter uncertainty can be separated from the total predictive uncertainty. By not distinguishing among error components, they do not help to reduce predictive uncertainty. To additionally separate bias from random measurement errors, Kennedy and O'Hagan (2001), Higdon et al. (2005), Bayarri et al. (2007) and others suggested using a Gaussian stochastic process to describe the knowledge about the bias, plus an independent error term for observation error. This approach has been applied to environmental modeling and linked to multi-objective model calibration by Reichert and Schuwirth (2012). Recently, a more complex input-dependent description of bias has been applied successfully by Honti et al. (2013). In their study, this solved the problem that model bias was greater during rainy periods than during dry weather, a common situation

in hydrology (Breinholt et al., 2012). Going in a different direction of error separation, Sikorska et al. (2012b) combined the lumped autoregressive error model with rainfall multipliers to separate the effect of input uncertainty from (lumped, remaining) bias and flow measurement errors.

In summary, there are three major interrelated needs in (urban) hydrological modeling: (i) to obtain reliable predictions, (ii) to disentangle prediction uncertainty into its components, and (iii) to fulfill the statistical assumptions behind model calibration. In particular, need (iii) is necessary to fulfill requirements (i) and (ii) in a satisfying way.

To address these issues, here we adapt the framework of Kennedy and O'Hagan (2001), as formulated by Reichert and Schuwirth (2012), to assess model bias along with other uncertainty components. This makes it possible to provide reliable predictions of (urban) hydrological models while improving the fulfillment of the underlying statistical assumptions. At the same time, this approach considers three different uncertainty components, namely output measurement errors, parametric uncertainty and the effect of structural deficits and input measurement errors on model output. With this approach all uncertainties are described in the output. This does not allow separating among input errors and structural deficits. However, a statistical bias description is simpler and less computationally intensive than addressing the causes of bias via mechanistic propagation of rainfall uncertainty (Renard et al., 2011), stochastic time-dependent parameters (Reichert and Mieleitner, 2009), or by combining filtering and data augmentation (Bulygina and Gupta, 2011).

Although focused on urban settings, our methodology is also suitable in other contexts like natural watersheds, where generally processes occur on longer time scales and output measurements are more uncertain. In this paper, we do not advocate an ideal error model that fits every situation. In our view, although very desirable, such a model might be unrealistic because watershed behaviors, measurement strategies and hydrodynamic models differ from case to case. Instead, we suggest a structured approach to find the most suitable description of model bias for a given hydrosystem and a given deterministic model.

Specifically, we investigate different strategies to parameterize the bias description, making it i) input-dependent and ii) output-dependent by applying two different transformations. The innovations of our study are:

i. A formal investigation of model bias in urban hydrology. This makes it possible to obtain reliable uncertainty intervals of sewer flows, also because the underlying statistical assumptions are better fulfilled.

ii. An assessment of the importance of model bias by separating prediction uncertainty into the individual contributions of bias, effect of model parameter uncertainty and measurement errors.

iii. A systematic comparison of different bias formulations and transformations. This is highly relevant for both natural and urban hydrology because we can acquire knowledge for potential future studies.

iv. An assessment of predictive uncertainties of flows for past (calibration) and future (extrapolation) system states. We find that considering bias not only produces reliable prediction intervals. It also accounts for increasing uncertainty when flow predictions move from ob-

served past into unknown future conditions. Furthermore, we discuss how the exploratory analysis of bias and monitoring data can be used to improve the hydrodynamic model.

The remainder of this article is structured as follows: first, we present the statistical description of model bias and compare it to the classical approach. Second, we introduce two different bias formulations and two transformations, and describe how we evaluate the performance of the resulting runoff predictions. Third, we test our approach on a high-quality dataset from a real-world stormwater system in Prague, Czech Republic, and present the results obtained with the different error models. Fourth, we discuss these results as well as advantages and limitations of our approach based on theoretical reflections and our practical experience. In addition, we suggest how to select the most appropriate error formulations for urban and also natural hydrological studies and outline future research needs.

2.2 Methods

2.2.1 Likelihood function

To statistically estimate the predictive uncertainty of urban drainage models, we need a likelihood function (a.k.a. sampling model), $f\left(\mathbf{y}_o|\boldsymbol{\theta},\boldsymbol{\psi},\mathbf{x}\right)$, which combines (in this particular case) a deterministic model (a.k.a. simulator), M, with a probabilistic error term. $f\left(\mathbf{y}_o|\boldsymbol{\theta},\boldsymbol{\psi},\mathbf{x}\right)$ describes the joint probability density of observed system outcomes, $\mathbf{y}_o$, given the model parameters, ξ, and external driving forces, $\mathbf{x}$, such as precipitation. The probability density, $f\left(\mathbf{y}_o|\boldsymbol{\theta},\boldsymbol{\psi},\mathbf{x}\right)$, may have a frequentist or a Bayesian interpretation. While the former considers probabilities as the limiting distribution of a large number of observations, the latter uses probabilities to describe knowledge or belief about a quantity, e.g. output variable. Only frequentist elements in a likelihood function can be empirically tested. To formulate such a likelihood function, we need (i) a simulator of the system with parameters $\boldsymbol{\theta}$, and (ii) a stochastic model of the errors with parameters $(\boldsymbol{\theta},\boldsymbol{\psi})$. A generic likelihood function assuming a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta},\boldsymbol{\psi},\mathbf{x})$ of output transformed by a function $g()$ can be written as:

$$f(\mathbf{y}_o \mid \boldsymbol{\theta},\boldsymbol{\psi},\mathbf{x}) = \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{\det\left(\boldsymbol{\Sigma}(\boldsymbol{\psi},\mathbf{x})\right)}} \exp\left(-\frac{1}{2}\left[\tilde{\mathbf{y}}_o - \tilde{\mathbf{y}}_{\mathrm{M}}(\boldsymbol{\theta},\mathbf{x})\right]^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\psi},\mathbf{x})^{-1}\left[\tilde{\mathbf{y}}_o - \tilde{\mathbf{y}}_{\mathrm{M}}(\boldsymbol{\theta},\mathbf{x})\right]\right) \prod_{i=1}^{n} \frac{\mathrm{d}g}{\mathrm{d}y}\left(y_{o,i},\boldsymbol{\psi}\right), \quad (2.1)$$

where n is the number of observations, i.e. the dimension of $\mathbf{y}_o$, which could be, for instance, a sewer flow time series. $\mathbf{y}_{\mathrm{M}}$ are the corresponding model predictions. The tilde denotes transformed quantities, i.e. $\tilde{\mathbf{y}} = g(\mathbf{y})$. Note that Eq. (2.1) assumes the residual errors to have 0 as expected value.

Uncertainty analysis for predictions is usually preceded by model calibration which requires that the statistical assumptions underlying the likelihood function are approximately fulfilled. This means that the Bayesian part of the likelihood function should correctly represent (conditional) knowledge/belief of the analyst (given the model parameters). This assumption is not testable by frequentist techniques. Instead, the appropriateness of the priors can be checked by carefully eliciting the knowledge of the experts (O'Hagan et al., 2006). Additionally, frequentist assumptions can be tested by comparing empirical distributions with model assumptions. In our error

models, we will have a frequentist interpretation of the observation error that can be tested, whereas the distributional assumption of the bias cannot be tested. If frequentist assumptions are violated, options are (i) to improve the structure of the deterministic model, (ii) to modify the distributional assumptions, or (iii) to improve the error model, e.g., by using a statistical (Bayesian) bias description.

i. Regarding improving the model structure, e.g., by a more detailed description of relevant processes or by increasing the spatial resolution, bias can be reduced, but not completely eliminated for environmental models. Natural systems are so complex that models will always be a simplified abstraction of the physical reality, unable to describe natural phenomena without bias. In addition, increasing model complexity will increase parametric uncertainty and computation time. Thus, adequate model complexity must balance between bias and parametric uncertainty. Input errors are relevant when dealing with highly variable forcing fields, as it is the case for precipitation. Having a denser point measurement network or combining different types of input observations (e.g. from pluviometers, radar and microwave links) can reduce this uncertainty. However, for practical reasons, input errors cannot be completely eliminated (Berne et al., 2004).

ii. Regarding improving the distributional assumptions, a simple way is to transform data and model results and applying the convenient distributional assumptions to the transformed values. This technique is commonly applied in hydrology to reduce heteroscedasticity and skewness of (random observation) errors, while simultaneously accounting for increasing uncertainty during high flow periods (Wang et al., 2012; Breinholt et al., 2012). Alternatively, a similar effect can be achieved through heteroscedastic error models with error variance dependent on external forcings (Honti et al., 2013) or simulated outputs (Schoups and Vrugt, 2010).

iii. Regarding accounting for difficult-to-reduce input and structural errors responsible for autocorrelated residuals, it has been suggested to describe prior knowledge of model bias by means of a stochastic process and to update this knowledge through conditioning with the data (Craig et al., 2001; Kennedy and O'Hagan, 2001; Higdon et al., 2005; Bayarri et al., 2007).

To increase the reliability of our probabilistic predictions and the fulfillment of the underlying assumptions for a given model, we suggest to combine strategies (ii) and (iii).
In the following paragraphs, we consider nine different likelihood functions by systematically modifying (i) the variancecovariance matrix of the residuals $\mathbf{\Sigma}$ (Sects. 2.2.1 and 2.2.1) and (ii) the transformation function g (Sect. 2.2.1). Specifically, we take into account three forms of parameterization of the bias process: neglection of bias (traditional error model with independent observation errors only), a stationary bias process, and an input-dependent bias process. Regarding output transformation, we compare the identity (no transformation), the Box-Cox transformation, and a recently suggested log-sinh transformation (references are given below).

Independent error model

In urban hydrology, the most commonly used statistical technique to estimate predictive uncertainty assumes an independent error model (Dotto et al., 2011; Freni et al., 2009a; Breinholt et al., 2012). Besides the absence of serial correlation, this requires residual errors identically distributed around zero.

The transformed observed system output, $\tilde{\mathbf{Y}}_o$, is modelled as the sum of a deterministic model output $\tilde{\mathbf{y}}_{\mathrm{M}}(\mathbf{x}, \boldsymbol{\theta})$ and an error term representing the measurement noise of the system response $\mathbf{E}$

$$\tilde{\mathbf{Y}}_o(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \tilde{\mathbf{y}}_{\mathrm{M}}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{E}(\boldsymbol{\psi}) \quad , \tag{2.2}$$

where variables in capitals represent random variables, whereas those in lowercase are deterministic functions.
Assuming independent identically distributed normal errors in the transformed space, $\mathbf{E}$ follows a multivariate normal distribution with mean 0 and a diagonal covariance matrix,

$$\boldsymbol{\Sigma}_E = \sigma_E^2 \mathbb{1} \, . \tag{2.3}$$

We then have $\boldsymbol{\psi} = \sigma_E$, and the covariance matrix of Eq. (2.1) is given by $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_E$. As $\mathbf{E}(\boldsymbol{\psi})$ is interpreted to represent observation error, the distributional assumptions are testable through residual analysis. Note that while in hydrology the observation and measurement errors are used as synonyms, in other environmental contexts observation errors can contain additional sampling errors.

Autoregressive bias error model

In contrast to the independent error model, the autoregressive bias error model explicitly acknowledges the fact that simulators cannot describe the "true" behaviour of a system. This has been originally suggested in the statistical literature (Craig et al., 2001; Kennedy and O'Hagan, 2001; Higdon et al., 2005; Bayarri et al., 2007) and later adapted to environmental modelling (Reichert and Schuwirth, 2012).
Technically, model inadequacy (also called bias or discrepancy) is considered by augmenting the independent error model with a bias term:

$$\tilde{\mathbf{Y}}_o(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \tilde{\mathbf{y}}_{\mathrm{M}}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{B}_{\mathrm{M}}(\mathbf{x}, \boldsymbol{\psi}) + \mathbf{E}(\boldsymbol{\psi}) \quad . \tag{2.4}$$

On the one hand, this model bias $\mathbf{B}_{\mathrm{M}}$ can capture the effect of errors in input measurements and structural limitations. On the other hand, it can also describe systematic output measurement errors, e.g. from sensor failure, incorrectly calibrated devices or erroneously estimated rating curves. In its simplest form, the bias is modelled as an autocorrelated stationary random process (Reichert and Schuwirth, 2012). However, it can also have a more complex structure and, for instance, be input-dependent (Honti et al., 2013). Strictly speaking, $\mathbf{B}_{\mathrm{M}}$ represents a bias-correction whereas the bias itself is its negative.
Conceptually, one difficulty is the identifiability problem between model and bias, which is apparent in Eq. (2.1). As both cannot be observed separately, this issue can only be solved by considering prior knowledge on the bias in parameter estimation. This requires a Bayesian framework for inference and prior distributions that favour the smallest possible bias. Indeed, we want output dynamics to be described as accurately as possible by the simulator and only the remaining deviations by the bias. The distribution of the residuals, $\mathbf{B}_{\mathrm{M}}(\mathbf{x}, \boldsymbol{\psi}) + \mathbf{E}(\boldsymbol{\psi})$, is not testable due to the Bayesian interpretation of $\mathbf{B}_{\mathrm{M}}(\mathbf{x}, \boldsymbol{\psi})$. However, when estimating both $\mathbf{B}_{\mathrm{M}}(\mathbf{x}, \boldsymbol{\psi})$ and $\mathbf{E}(\boldsymbol{\psi})$, the assumptions regarding the observation error, $\mathbf{E}(\boldsymbol{\psi})$, can be tested by frequentist tests.
Practically, the choice of an adequate bias formulation is challenging. On the one hand, examples from urban hydrological applications are currently lacking. On the other hand, the bias results

from the complex interplay between the drainage system, the simulator and the monitoring data. This is not straightforward to assess a priori. Notice that the autocorrelated bias and the random observation errors are usually well distinguishable due to their distinct statistical properties.
In the following, we investigate four different bias formulations: (i) constant (i.e., input and output-independent), (ii) output-dependent, (iii) input-dependent, (iv) input and output-dependent. The constant bias is modeled via a standard Ornstein-Uhlenbeck (OU) process. The input-dependence uses a modified OU process, which is perturbed by rainfall. The output-dependence is considered through transformation of measured and simulated data.

Constant bias

The simplest bias formulation is a mean-reverting Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930), the discretization of which would be a fist-order autoregressive process (AR(1)) with Gaussian iid noise. The OU process is a stationary Gauss-Markov process with a long-term equilibrium value of zero (in our application) and a constant variance, either in the original or transformed space. The one-dimensional OU process B_{M} is described by the stochastic differential equation

$$\mathrm{d}B_{\mathrm{M}}(t) = -\frac{B_{\mathrm{M}}(t)}{\tau}\mathrm{d}t + \sqrt{\frac{2}{\tau}}\sigma_{B_{\mathrm{ct}}}\mathrm{d}W(t) \quad , \tag{2.5}$$

where τ is the correlation time and $\sigma_{B_{\mathrm{ct}}}$ is the asymptotic standard deviation of the random fluctuations around the equilibrium. $\mathrm{d}W(t)$ is a Wiener process which is the same as standard Brownian motion (random walk with independent Gaussian increments). For an introduction to stochastic processes see, e.g., Henderson and Plaschko (2006); Iacus (2008); Kessler et al. (2012).
This stationary bias results in the likelihood function of Eq. (2.1) with covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{E}} + \boldsymbol{\Sigma}_{\mathbf{B}_{\mathrm{M}}}$ with

$$\boldsymbol{\Sigma}_{\mathbf{B}_{\mathrm{M}},i,j}(\boldsymbol{\psi}) = \sigma^2_{B_{ct}} \exp\left(-\frac{1}{\tau}|t_i - t_j|\right) \quad . \tag{2.6}$$

In contrast to the formulation given by Eq. (2.6), the covariance in the original formulation by Kennedy and O'Hagan (2001) had an exponent α for the term $|t_i - t_j|$. To guarantee differentiability, this expondent is often chosen to be equal to two. For hydrological applications we prefer an exponent of unity to be compatible with the OU process, which can be assumed to be a simple description of underlying mechanisms leading to a decay of correlation (Yang et al., 2007a; Sikorska et al., 2012b). Indeed, such a covariance structure makes it possible to transfer the autoregressive error models (Kuczera, 1983; Bates and Campbell, 2001; Yang et al., 2007b) to the bias description framework (Honti et al., 2013).

Input-dependent bias

A more complex bias description considers input-dependency to mechanistically increase the uncertainty of flow predictions during rainy periods. Following Honti et al. (2013), we suggest an OU process whose variance grows quadratically with the precipitation intensity, x, shifted in time by a lag d. The equation for the rate of change of the input-dependent bias is then given by:

$$\mathrm{d}B_{\mathrm{M}}(t) = -\frac{B_{\mathrm{M}}(t)}{\tau}\mathrm{d}t + \sqrt{\frac{2}{\tau}\left(\sigma^2_{B_{\mathrm{ct}}} + (\kappa x(t-d))^2\right)}\mathrm{d}W(t) \quad , \tag{2.7}$$

where κ is a scaling factor and d denotes the response time of the system to rainfall. For an equidistant time discretization, with $t_{i+1} - t_i = \Delta t$, assuming that the lag is a constant multiple of Δt, $d = \delta \Delta t$, and the input is constant between time-points, we derive from Eq. (3.4) the recursion formula for the variance

$$E\left[B_{\mathrm{M}}^2(t_i)\right] = E\left[B_{\mathrm{M}}^2(t_{i-1})\right] \exp\left(-\frac{2}{\tau}\Delta t\right) + \left[\sigma_{B_{\mathrm{ct}}}^2 + (\kappa x_{i-\delta})^2\right]\left(1 - \exp\left(-\frac{2}{\tau}\Delta t\right)\right) \quad . \tag{2.8}$$

The parameters $\boldsymbol{\psi}$ of this bias are given by

$$\boldsymbol{\psi} = (\sigma_{B_{\mathrm{ct}}}, \tau, \kappa, d)^T \, . \tag{2.9}$$

The resulting bias covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{B}_{\mathrm{M}}}$, is given by

$$\boldsymbol{\Sigma}_{\mathbf{B}_{\mathrm{M}},i,j}(\boldsymbol{\psi}, \mathbf{x}) = E\left[B_{\mathrm{M}}^2\left(\min(t_i, t_j)\right)\right] \exp\left(-\frac{1}{\tau}|t_j - t_i|\right) \quad . \tag{2.10}$$

In comparison to the original bias formulation by Honti et al. (2013), we modified two aspects. First, we consider the response time of the system by introducing a time lag of the input, which was necessary due to the high-frequent monitoring data with a temporal resolution of 2 minutes. Indeed, instead of working with daily discharge data used in Honti et al. (2013), here we had output observations every two minutes. Second, we omitted the fast bias component, which accounts for additional noise coming into action during the rainy timesteps. In our experience, this component did not have a significant effect at this short time scale and its elimination led to a greater simplicity and robustness of the error model.

Output transformation

In hydrological modeling, it is common practice to apply a transformation to account for increasing variance with increasing discharge. The two variance stabilization techniques which are, in our view, most promising for urban drainage applications are: the Box-Cox transformation (Box and Cox, 1964) and the log-sinh transformation (Wang et al., 2012).

Box-Cox

The Box-Cox transformation has been successfully used in many hydrological studies to reduce the output-dependence of the residual variance in the transformed space (e.g., Kuczera, 1983; Bates and Campbell, 2001; Yang et al., 2007a; Reichert and Mieleitner, 2009; Dotto et al., 2011; Sikorska et al., 2013).
The one-parameter Box-Cox transformation can be written as:

$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \tag{2.11}$$

$$g^{-1}(z) = \begin{cases} (\lambda z + 1)^{1/\lambda} & \text{if } \lambda \neq 0 \\ \exp(z) & \text{if } \lambda = 0 \end{cases} \tag{2.12}$$

$$\frac{dg}{dy} = y^{\lambda - 1} \tag{2.13}$$

where g indicates the forward and g^{-1} the backward transformation, whereas $\frac{dg}{dy}$ is the trans-

formation derivative. $\tilde{y}$, $g(y)$, and z represent the transformed output. λ is a parameter that determines how strong the transformation is. It is chosen from the interval [0,1], with the extreme cases of 1 leading to the (shifted) identity transformation, 0 to a log transformation. We choose a $\lambda = 0.35$, which has lead to satisfactory results in many similar investigations (Willems, 2012; Honti et al., 2013; Yang et al., 2007b,a; Wang et al., 2012; Frey et al., 2011). Assuming a constant variance in the transformed space, this value yields a moderate increase of variance in non-transformed output. This accounts for an observed increase in residual variance while keeping the weight of high discharge observations sufficiently high for calibration. In other words, this moderate λ assures a good compromise between the performances of the error model and the fit of the simulator. The behavior of the Box-Cox transformation and its derivative for the stormwater runoff in our study are shown in Figs. 1 and S1.

Log-sinh

The log-sinh transformation has recently shown very promising results for hydrological applications (Wang et al., 2012). In contrast to the original notation, we prefer a reparameterized form with parameters that have a more intuitive meaning:

$$g(y) = \beta \log\Big(\sinh\big(\frac{\alpha + y}{\beta}\big)\Big), \tag{2.14}$$

$$g^{-1}(z) = \Big(\operatorname{arcsinh}\big(\exp(\frac{z}{\beta})\big) - \frac{\alpha}{\beta}\Big)\beta \quad , \tag{2.15}$$

$$\frac{dg}{dy} = \coth\Big(\frac{\alpha + y}{\beta}\Big) \tag{2.16}$$

where α (originally a/b) and β (originally $1/b$) are lower and upper reference outputs, respectively. α controls how the relative error increases for low flows. For outputs larger than β, instead, the absolute error gradually stops increasing and the scaling of the error (derivative of g) becomes approximately equal to unity. In our study, we chose α to be a runoff in the range of the smallest measured flow and β to be an intermediately high discharge above which uncertainty was assumed not to significantly increase. These considerations are also in agreement with the transformation parameter values determined by Wang et al. (2012). Given the characteristics of our catchment and model we set α=5 l/s and β=100 l/s. The graphs of the transformation function and its derivative with these parameter values are provided in Figs. 1 and S1.

Both transformations are able to reduce the heteroscedasticity of residuals, which represents the fact that flow meters and rating curves are more inaccurate during high flows and systematic errors lead to a higher uncertainty during high flows. Another positive characteristic is that these transformations make error distributions asymmetric, substantially reducing the proportion of negative flow predictions, which can otherwise occur during error propagation.

2.2.2 Inference and predictions

The following steps are needed to calibrate a deterministic model M with a statistical bias description and observation error and to analyze the resulting prediction uncertainties: (i) definition of the prior distribution of the parameters, (ii) obtaining the posterior distribution with Bayesian inference, (iii) probabilistic predictions for the temporal points (in the following called layout) used in calibration, (iv) probabilistic predictions for the extrapolation period. In these

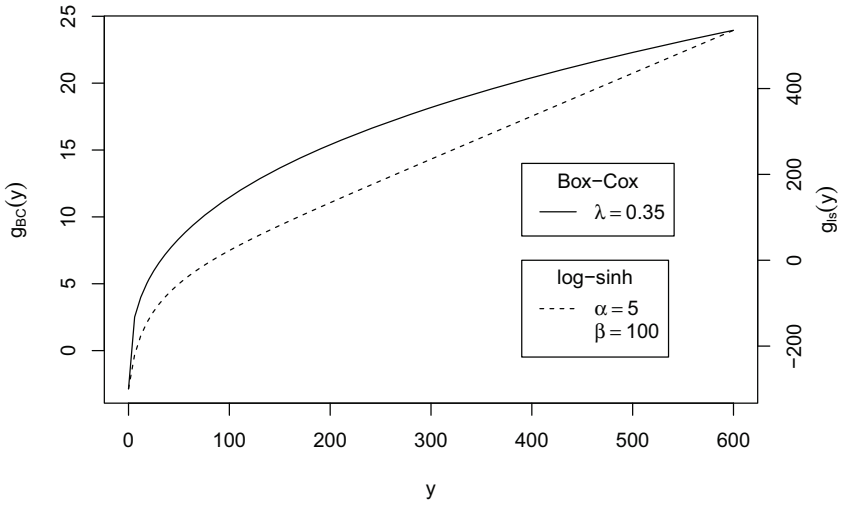

Figure 1: Behavior of the Box-Cox (solid line) and log-sinh (dashed line) transformation as a function of the output variable (e.g. discharge in l/s) with parameters used in this study.

last two phases credible intervals are estimated by uncertainty propagation via Monte Carlo simulations. Finally, one has to assess the quality of the predictions and verify the statistical assumptions. We highly recommended an exploratory analysis of the bias, which can help to improve the structure of the simulator.

Prior definition

First, one has to define the joint prior distribution of the parameters of the hydrological model, $\boldsymbol{\theta}$, and of the error model, $\boldsymbol{\psi}$. In particular, this requires an informative prior of the covariance matrix of the flow measurements. Although a first guess can be obtained from manufacturer's specifications, it is recommended to assess it separately with redundant measurements (see Dürrenmatt et al., 2013). As stated in Sect. 2.2.1, it is important that the prior of the bias reflects the desire to avoid model inadequacy as much as possible. This is obtained by a probability density decreasing with increasing values of $\sigma_{B_{\mathrm{ct}}}$ and κ (e.g. an exponential distribution). This helps to reduce the identifiability problem between the deterministic model and the bias. For the prior for $\sigma_{B_{\mathrm{ct}}}$, one could take into account that the maximum bias scatter is unlikely to be higher than the observed discharge variability. On the other hand, the maximum value of κ is in the same order of magnitude as the maximum discharge divided by the corresponding maximum precipitation of a previously monitored storm event. Additionally, τ should represent the characteristic correlation length of the residuals and could be approximately set to 1/3 of the hydrograph recession time. More prior information may be available from previous model applications to the same or a similar hydrological system. Finally, the parameters of the chosen transformation have to be specified. These parameters influence the priors of σ_E, $\sigma_{B_{\mathrm{ct}}}$, and κ which are defined in the transformed space.

We recognize that assigning priors for bias parameters might be challenging. Therefore we suggest testing a posteriori the sensitivity of the updated parameter distributions to the priors. We advise against using uninformative uniform priors for two reasons. First, as discussed above, our ignorance about bias parameters is not total. Second if one lacks knowledge about ψ one

should also lack knowledge about ψ^2, but no distribution exists that is uniform on both ψ and ψ^2 (Christensen et al., 2010).

Bayesian inference

Second, the posterior distribution of the simulator and the error model parameters $f(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_o, \mathbf{x})$ is calculated using the prior distribution, $f(\boldsymbol{\theta}, \boldsymbol{\psi})$, the likelihood function, $f(\mathbf{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})$, and the observed data, $\mathbf{y}_o$, according to Bayes' theorem:

$$f(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_o, \mathbf{x}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{\psi}) f(\mathbf{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})}{\iint f(\boldsymbol{\theta}', \boldsymbol{\psi}') f(\mathbf{y}_o \mid \boldsymbol{\theta}', \boldsymbol{\psi}', \mathbf{x}) \mathrm{d}\boldsymbol{\theta}' \mathrm{d}\boldsymbol{\psi}'} \quad . \tag{2.17}$$

In other words, during a Bayesian calibration, the joint probability density of parameter and model results, the product of the prior of the parameters and the likelihood, is conditioned on the data.
In order to cope with analytically intractable multi-dimensional integrals, such as the ones in the denominator of Eq. 3.10 or those raising when marginalizing the joint posterior, numerical techniques have to be applied. In this context, Markov Chain Monte Carlo (MCMC) simulations are useful for approximating properties of the posterior distribution based on a sample, even if the normalization constant in Eq. 3.10 is unknown. Details are given in Sect. 2.3.2.

Predictions for the calibration layout L_1

Third, one has to compute posterior predictive distributions for the observations that have been used for parameter estimation. The experimental layout of this data set (here: calibration layout), L_1, specifies which output variables are observed, where and when. Here, the model output at calibration layout L_1 is given by the vector $\mathbf{y}^{L_1} = (y^s_{t_1}, \ldots, y^s_{t_{n_1}})$, where y^s denotes the discharge at the location, s, of the measurements and t_i, for $i = 1, \ldots, n_1$, the time points of the measurements.
In order to separate different uncertainty components, we compute predictions from (i) the simulator $\mathbf{y}^{L_1}_{\mathrm{M}}$, which only contains uncertainty from hydrological model parameters, (ii) our best knowledge about the system response $g^{-1}(\tilde{\mathbf{y}}^{L_1}_{\mathrm{M}} + \mathbf{B}^{L_1}_{\mathrm{M}})$, which comprehends additional uncertainty from input errors and structural deficits, and (iii) observations of the system response, $g^{-1}(\tilde{\mathbf{y}}^{L_1}_{\mathrm{M}} + \mathbf{B}^{L_1}_{\mathrm{M}} + \mathbf{E}^{L_1})$ which, in addition, includes random flow measurement errors (note that we mean here the application of the scalar function g^{-1} to all components of the vector specified as its argument). Usually, hydrological "predictions" describe simulation results for time points or locations where we do not have measurements. Here, consistent with Higdon et al. (2005) and Reichert and Schuwirth (2012), "predictions" designate the generation of model outputs (with uncertainty bounds) in general.
To obtain probabilistic predictions for multivariate normal distributions involved in the evaluation of these random variables, the reader is referred to Kendall et al. (1994) and Kollo and von Rosen (2005). Taking as an example the posterior knowledge of the true system output without observation error conditional on model parameters, the expected transformed values are given by

$$E\left[\tilde{\mathbf{y}}^{L_1}_{\mathrm{M}} + \mathbf{B}^{L_1}_{\mathrm{M}} \mid \tilde{\mathbf{Y}}^{L_1}_o, \boldsymbol{\theta}, \boldsymbol{\psi}\right] = \tilde{\mathbf{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{\Sigma}_{\mathbf{B}^{L_1}_M} \left(\boldsymbol{\Sigma}_{\mathbf{E}^{L_1}} + \boldsymbol{\Sigma}_{\mathbf{B}^{L_1}_{\mathrm{M}}}\right)^{-1} \cdot \left(\tilde{\mathbf{y}}^{L_1}_o - \tilde{\mathbf{y}}^{L_1}_{\mathrm{M}}\right) \tag{2.18}$$

and their covariance matrix by

$$\mathrm{Var}\left[\tilde{\mathbf{y}}^{L_1}_{\mathrm{M}} + \mathbf{B}^{L_1}_{\mathrm{M}} \mid \tilde{\mathbf{Y}}^{L_1}_o, \boldsymbol{\theta}, \boldsymbol{\psi}\right] = \boldsymbol{\Sigma}_{\mathbf{B}^{L_1}_{\mathrm{M}}} \left(\boldsymbol{\Sigma}_{\mathbf{E}^{L_1}} + \boldsymbol{\Sigma}_{\mathbf{B}^{L_1}_{\mathrm{M}}}\right)^{-1} \boldsymbol{\Sigma}_{\mathbf{E}^{L_1}} \quad . \tag{2.19}$$

To obtain the posterior predictive distribution of the bias-corrected output, $\tilde{\mathbf{y}}_{\mathrm{M}}^{L_1} + \mathbf{B}_{\mathrm{M}}^{L_1}$, first, we have to propagate a large sample from the posterior distribution through the simulator, $\mathbf{y}_{\mathrm{M}}^{L_1}$, and draw realizations of $\tilde{\mathbf{y}}_{\mathrm{M}}^{L_1} + \mathbf{B}_{\mathrm{M}}^{L_1}$ by using Eqs. 2.18 and 2.19. Than, we transform these results back to the original observation scale by applying the inverse transformation g^{-1}. Finally, to visualize the best knowledge and uncertainty intervals of this distribution, we compute the sample quantile intervals (e.g., 0.025, 0.5, 0.975 quantiles). A similar procedure is to apply to approximate the predictive distributions of $\mathbf{y}_{\mathrm{M}}^{L_1}$ and $\tilde{\mathbf{y}}_{\mathrm{M}}^{L_1} + \mathbf{B}_{\mathrm{M}}^{L_1} + \mathbf{E}^{L_1}$.

Besides calculating the posterior predictive distribution, it is important to check the assumptions of the likelihood function. As our posterior represents our knowledge of system outcomes, bias and observation errors, of which not all have a frequentist interpretation, we cannot apply a frequentist test to the residuals of the deterministic model at the best guess of the model parameters. However, we can perform a frequentist test based on our knowledge of the observation errors. This makes it necessary to split the residuals into bias and observation errors and to derive the posterior of the observation errors alone. A numerical sample of this posterior can be gained by substituting the sample for the random variable $\tilde{\mathbf{y}}_{\mathrm{M}}^{L_1} + \mathbf{B}_{\mathrm{M}}^{L_1}$ in

$$\mathbf{E}^{L_1} = \tilde{\mathbf{y}}_{o}^{L_1} - \left(\tilde{\mathbf{y}}_{M}^{L_1}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{B}^{L_1}\right). \tag{2.20}$$

In this equation $\tilde{\mathbf{y}}_{o}^{L_1}$ refers to the field data. The medians of the components of this sample represent our best point estimates of observation errors that we will use to test the statistical assumptions as described in Sect. 2.2.2.

Predictions for the validation layout L_2

Fourth, one computes posterior predictive distributions for the validation (or extrapolation) layout, L_2, where data are not available or not used for calibration. In our study, L_2 denotes the location and the time points of the extrapolation range, and the associated model output is given by $\mathbf{y}^{L_2} = (y^s_{t_{n_1+1}}, \ldots, y^s_{t_{n_2}})$. This layout, however, could also contain interpolation points between calibration data.

A sample for layout L_2 could be calculated similarly to the one for L_1 by using the Eqs. (35) and (36) of Reichert and Schuwirth (2012) instead of the Eqs. (2.18) and (2.19). However, the specific form of our bias formulation as an Ornstein-Uhlenbeck process (Eqs. 2.4 and 3.4) offers a potentially more efficient alternative. As the OU process is a Gauss-Markov process, we can draw a realization for the entire period by iteratively by drawing the realization for the next time step at time t_j from that of the previous time step at time t_{j-1} from a normal distribution. The expected value and variance of the normal distribution of the bias given the model parameters is given by

$$\mathrm{E}\left[B^{L_2}_{M,j} \mid B^{L_2}_{M,j-1} = b_{j-1}, \boldsymbol{\theta}, \boldsymbol{\psi}\right] = b_{j-1} \cdot \exp\left(-\frac{\Delta t}{\tau}\right), \tag{2.21}$$

$$\mathrm{Var}\left[B^{L_2}_{M,j} \mid B^{L_2}_{M,j-1}, \boldsymbol{\theta}, \boldsymbol{\psi}\right] = \left(\sigma^2_{B_{\mathrm{ct}}} + \left(\kappa x_{j-d}\right)^2\right) \cdot \left(1 - \exp\left(-2\frac{\Delta t}{\tau}\right)\right). \tag{2.22}$$

The sample of the bias for L_2 can be generated by drawing iteratively from these distributions for all required values of j starting from the last result of each sample point from layout L_1. By calculating the results of the deterministic model and drawing from the observation error distribution, samples for $\mathbf{y}_{\mathrm{M}}^{L_2}$, $g^{-1}(\tilde{\mathbf{y}}_{\mathrm{M}}^{L_2} + \mathbf{B}_{\mathrm{M}}^{L_2})$, and $g^{-1}(\tilde{\mathbf{y}}_{\mathrm{M}}^{L_2} + \mathbf{B}_{\mathrm{M}}^{L_2} + \mathbf{E}^{L_2})$ can be constructed similarly as for layout L_1.

Performance analysis

Fifth, the quality of the predictions is evaluated by assessing (i) the coverage of prediction for the validation layout and (ii) whether the statistical assumptions underlying the error model are met for the calibration layout.

Checking the predictive capabilities

The predictive capability of the model can be assessed by two metrics, the "reliability" and the "average band width " (Breinholt et al., 2012). The reliability measures what percentage of the validation data are included in the 95 % credibility intervals of $g^{-1}(\tilde{\mathbf{y}}_{\mathrm{M}} + \mathbf{B}_{\mathrm{M}} + \mathbf{E})$. When this percentage is larger than or equal to 95 %, the predictions are reliable. In general we expect this percentage to be larger than 95 % as our uncertainty bands describe our (lack of) knowledge about future predictions. This combines Bayesian parametric and bias uncertainty with the uncertainty due to the observation error. These three components of predictive intervals are thus systematically more uncertain than the observation error alone. The limiting case of an exact coverage is only expected to occur if parameter uncertainty and bias is small compared to the observation error. In contrast, the average band width (ABW) measures the average breadth of the 95 % credibility intervals. Ideally, we seek the narrowest reliable bands. Besides these two criteria, the Nash–Sutcliffe efficiency index (Nash and Sutcliffe, 1970), a metric often used in hydrology, is applied to evaluate goodness of fit of the deterministic model to the data. As a side note, it has been suggested to check the prediction performance of a model by only examining the number of data points included in the prediction uncertainty intervals resulting only from parameter uncertainty (Dotto et al., 2011). Unfortunately, this is not conclusive because the field observations are not realizations of the deterministic model but of the model plus the errors.

Checking the underlying statistical assumptions

The underlying statistical assumptions of the error model are usually verified by residual analysis (Reichert, 2012). This, however, is only meaningful for frequentist quantities. In a Bayesian framework, probabilities express beliefs, which can differ from one data analyst to another and thus cannot be tested in a frequentist way. In our error model (Eq. 2.1), the observation error is the frequentist part of the likelihood function and frequentist tests can thus only be applied to this term. As outlined in Sect. 2.2.2, we can use the median of the posterior of the observation error at layout L_1 to do such a frequentist test. These posterior observation errors should be tested whether they are (i) normally distributed, (ii) have constant variance and (iii) are not autocorrelated. As the observation errors may only represent a small share of the residuals of the deterministic model, posterior predictive analysis based on independent data, as outlined in the previous paragraph, remains an important performance measure.
As a side note, it is conceptually incorrect to check frequentist assumptions by using the full (Bayesian) posterior distribution (e.g., Renard et al., 2010). Using the full posterior instead of the best point estimate of the observation errors adds additional uncertainty from the incomplete prior knowledge of parameter values. In our view, this distorts the interpretation of frequentist tests.

Improving the simulator

Finally, after performance checking, one should evaluate the opportunity to improve the simulator and/or the measurement design for the model's input. Hints for improvement can be

obtained by exploratory analysis of the bias, for example by investigating the relation between its median and output variables or input. On the one hand, systematic patterns in the relation of the bias to model input or output would suggest the presence of model structural deficits that could be corrected. On the other hand, increasing variance of the bias with increasing discharge could be a sign of excessive uncertainty in rainfall data. This could be improved by more reliable rainfall information.

2.3 Material

To demonstrate the applicability and usefulness of our approach, we evaluate the performance of nine different error models in a real-world urban drainage modeling study. In the following we will briefly describe the case study and details on the numerical implementation of the bias framework.

2.3.1 Case study

We tested the uncertainty analysis techniques on a small urban catchment in Sadová, Hostivice in the vicinity of Prague (CZ). The system has an area of 11.2 ha and is drained by a separate sewer system. It is a green residential area with an average slope of circa 2 %.

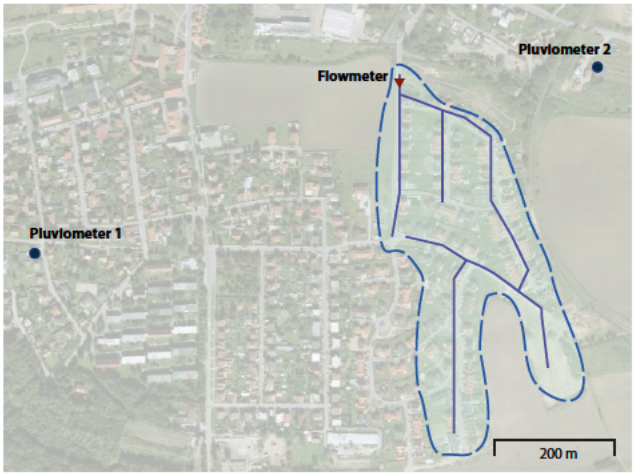

Figure 2: Aerial photo of our Sadová case study catchment. The map shows the layout of the main stormwater conduits and the location of the rain gauges and the flow meter.

The monitoring data of rainfall and runoff were collected in summer 2010 (Bareš et al., 2010). Flow was measured at the outlet of the stormwater system in a circular PVC pipe with a diameter of 0.6 m. A PCM Nivus area-velocity flow meter was used to record water level and mean velocity every 2 min. These output data show that the hydrosystem is extremely dynamic, with a response ranging approximately from 2 $\mathrm{L\,s^{-1}}$ during dry weather to 600 $\mathrm{L\,s^{-1}}$ during strong rainfall.

Rainfall intensities were measured with two tipping bucket rain gauges that were installed only a few hundred meters from the catchment (Fig. 2). These two input temporal datasets have been aggregated to 2 min time steps based on the weighted average distance from the watershed centroid.

For model calibration, we selected two periods with 6 major rainfall events. One on 27 August between 01:52 LT and 12:58, and the second in July between 22 July at 23:32 and 23 July at 19:00. For validation, a single period from 23 July at 19:02 to the next day at 07:00 was selected.

Calibration storms had a peak intensity ranging from 13 to 65 mm/hr, whereas validation events had a maximum rain rate spanning from 8 to 34 mm/hr. The monitored rainstorms had a duration of 0.5-4 hr with a cumulative height varying from 2.3 mm to 33 mm. The calibration and validation data of July 2010 are illustrated in Fig. 4.

2.3.2 Model implementation

We modelled runoff in the stormwater system using the SWMM software (Rossman and Supply, 2010). The model was set to a simple configuration, namely a nonlinear reservoir representing the catchment connected to a pipe with a constant groundwater inflow. Lumped modeling is particularly appropriate when a study focuses on outlet discharge and computation can be a limiting factor (Coutu et al., 2012b). The parameters that we inferred during calibration were the imperviousness, the width, the dry weather inflow, the length of the conduit and the slope of the catchment.

The procedure outlined in Sect. 2.2.2 to compute the posterior predictive distributions was implemented in `R` (R Core Team, 2013). For a simulation, an input file with parameters and rainfall series was read. The input file was iteratively modified to update the parameters by using `awk` (Aho et al., 1987). `awk` was also used to extract the runoff time series from the output file.

From a numerical viewpoint, we solved the "inverse problem" described in Sect. 2.2.2 by using a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm (Hastings, 1970). Before sampling from $f(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_o, \mathbf{x})$, we obtained a suitable jump distribution (a.k.a. transition function or proposal density) by using a stochastic adaptive technique to draw from the posterior (Haario et al., 2001). For better performance we added a size-scaling step, which depends on the target acceptance rate. For our inference problem, this algorithm proved to be more robust than others, such as Vihola (2012). However, research on efficient techniques for posterior sampling is evolving rapidly and other approaches could also be used. See Liang et al. (2011) and Laloy and Vrugt (2012) for recent developments in Bayesian computation.

2.4 Results

In general, accounting for model bias produced substantially wider prediction uncertainty bands and separated them in three components. The bias error models also substantially reduced the magnitude of the identified independent observation errors and decreased their autocorrelation. The different formulations of model inadequacy, however, show a considerable variability in terms of predictive distributions and behavior of the identified observation errors.

2.4.1 Evaluating the performance of probabilistic sewer flow predictions

As expected, the different assumptions underlying the nine error models lead to different credibility intervals for stormwater runoff at the monitoring point (Fig. 3 and Table 1). Predictions did not exhibit considerable sensitivity to the prior for the bias (results not shown).

For our case study, the best error model clearly was the constant bias model with log-sinh transformation (Fig. 4). It leads to high reliability, Nash-Sutcliffe index and sharp total uncertainty intervals.

Although the deterministic model reproduced the measured discharge dynamics well, the total uncertainty during strong rain events in the validation period is still rather large. Indeed, as illustrated in Fig. 4, even the best performing error model has a 95 % interquantile range up to $\sim 140\,\mathrm{L\,s^{-1}}$, which is about 20 % of the maximum runoff modeled in the extrapolation phase.

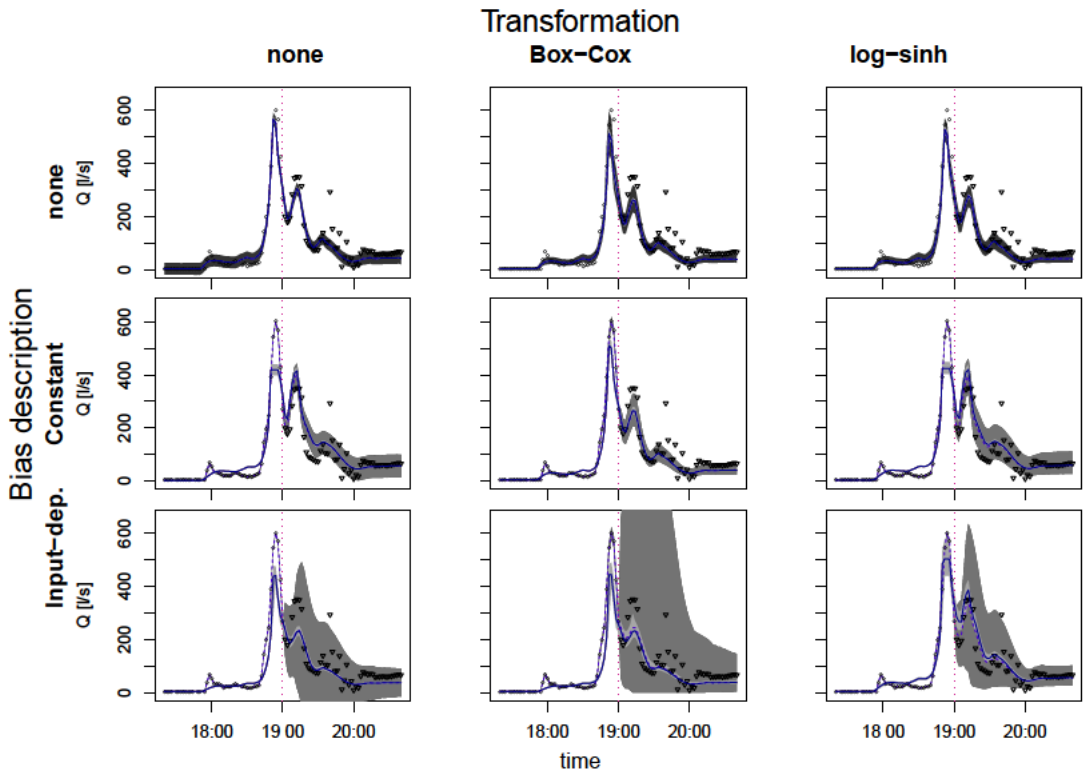

Figure 3: 95 % credible intervals for flow predictions for transition phase obtained with different assumptions on error distribution. The vertical dotted line divides the calibration layout (past) from the validation layout (future). The solid line is the deterministic model output with the optimized parameter set, whereas the dashed line is the bias-corrected output representing our best estimation of the true system response. Observed output of the system is represented by circles, with the triangular ones not being used for calibration. Colors of the credibility intervals: deterministic model predictions (light gray), predictions of the real system output (intermediate gray), predictions of new observations (dark gray). When considering bias, the contribution of uncorrelated observation errors $\mathbf{E}$ to total uncertainty becomes very small ($\lesssim 1$ l/s) and therefore is not visible at this scale. Consequently, the credibility intervals for the system output ($g^{-1}(\tilde{\mathbf{y}}_\mathrm{M} + \mathbf{B}_\mathrm{M})$) and the observations ($g^{-1}(\tilde{\mathbf{y}}_\mathrm{M} + \mathbf{B}_\mathrm{M} + \mathbf{E})$) are almost identical and overlap.

In general, we found that most of the error formulations with model bias produced reliable predictions and around 95 % or more of validation data fell within the 95 % prediction interval range for new observations (Table 1). In addition, the bias framework separated the total uncertainty into parametric uncertainty, effect of input plus structural deficits, and observation errors (Figs. 3 and 4). All the autoregressive error models indicated that most of the predictive uncertainty is due to model bias. Interestingly, uncertainty due to random measurement noise is generally so small that it is not visible in the plots.

In contrast, all error models which ignore model bias, with or without transformation, generated overconfident predictions with too narrow uncertainty bands. As previously stated, they also could not separate the total uncertainty into the individual error contributions.

Besides providing reliable estimates of the total predictive uncertainty, a second advantage of the bias framework is that it takes into account the different knowledge within the calibration and validation layouts. As shown in Fig. 3, the predictions obtained with bias description for the calibration layout, to the left of the dotted line, included most of the observations while being, at the same time, very narrow. This takes into account that in the calibration range, where data are available, our knowledge on stormwater runoff is rather accurate and precise.

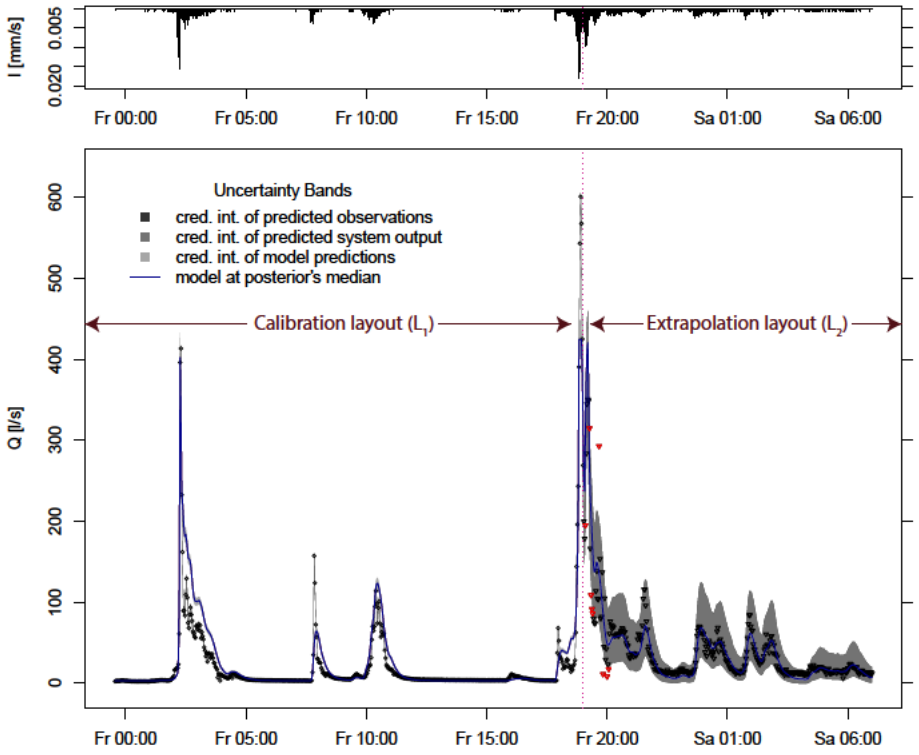

Figure 4: Probabilistic runoff predictions for part of the calibration (left) and the validation period (right) with the constant bias model and log-sinh transformation. The input time series (hyetograph) is shown on the top. The observed hydrograph is represented by circles, with the triangular data points being used only for validation. The 95% credible intervals are interpreted as follows: parametric uncertainty due to $\mathbf{y}_\mathrm{M}$ (light gray), parametric plus input and structural uncertainty due to $g^{-1}(\tilde{\mathbf{y}}_\mathrm{M}+\mathbf{B}_\mathrm{M})$ (intermediate), total uncertainty due to $g^{-1}(\tilde{\mathbf{y}}_\mathrm{M}+\mathbf{B}_\mathrm{M}+\mathbf{E})$ (dark gray). Validation data not included in this dark gray region are marked in red. The prediction intervals for the system output and the observations are almost indistinguishable and therefore only the intermediate gray band is visible at this scale.

In contrast, for the extrapolation domain where no observations are available, the uncertainty intervals are much larger.

In addition, we found that the conditioning on the monitoring data became increasingly weak the further the model predicts into the future. This gradually increases the uncertainty in the transition phase as the prediction horizon moves from the past into the future. Again, this is not possible with the traditional error models. Indeed, models with uncorrelated error terms cannot describe the propagation of information obtained from calibration data to nearby time points. Therefore, their prediction intervals are equally wide for both the calibration and validation layouts.

A third advantage of bias description is that it provides an estimate of the most probable system response $g^{-1}(\tilde{\mathbf{y}}_\mathrm{M}+\mathbf{B}_\mathrm{M})$, which is depicted by the dashed line in Fig. 3. In the calibration layout, it closely follows the observations, which are comparably precise and therefore contain the best information on the state of the system. For the validation layout, this information is lacking. However, instead of abruptly reverting to the simulator, the transition is gradual because the autocorrelated bias carries the information from the last monitoring data into the future. This "de-correlation" typically takes a few correlation lengths (here circa 30–50 min).

As can be seen in Table 1 and Fig. 3, even though the uncertainty intervals are more reliable

Table 1: Prediction performance metrics for the different error models: iid untransformed error model (iidE), Box-Cox transformed (iidE.BC), and log-sinh transformed (iidE.ls), constant untransformed bias (CtB), Box-Cox transformed (CtB.BC), and log-sinh transformed (CtB.ls), input-dependent untransformed bias (IDB), Box-Cox transformed (IDB.BC), and log-sinh transformed (IDB.ls). The criteria on the left represent the Nash-Sutcliffe index in calibration (NS.cal) and validation (NS.val) phases in the non-transformed space, the percentage of validation data points falling into 95 % prediction interval (Cover.val), and the average bound width [$\mathrm{L\,s^{-1}}$] in the extrapolation domain (ABW.val).

	iidE	iidE.BC	iidE.ls	CtB	CtB.BC	CtB.ls	IDB	IDB.BC	IDB.ls
NS.cal	0.948	0.924	0.936	0.885	0.921	0.876	0.847	0.839	0.904
NS.val	0.839	0.782	0.806	0.821	0.796	0.817	0.729	0.731	0.827
Cover.val	89.2	58.1	66.1	95.3	74.7	97.5	95	88.3	90
ABW.val	44.4	21.2	22.9	85	25.8	53.2	81.5	134	55.6

when bias is considered, the deterministic model performs best when residual autocorrelation and heteroscedasticity are not taken into account. This is not surprising since maximizing the posterior with the simple iid error model with no transformation corresponds to minimizing the sum of the squares of the errors and therefore produces the best fit.

Comparing the input-independent and dependent bias formulations, two important points are observed. First, the constant bias description produced on average narrower uncertainty bands than the input-dependent version. The latter, in particular, produced huge uncertainties during rain events and very narrow intervals during dry weather. Second, as expressed in Table 1, the constant bias almost produced the same simulator fit as the simple error model, whereas the input-dependent bias formulation performed on average less satisfactorily.

The transformation created skewed predictive distributions and, as expected, increased the wet weather uncertainty in the "real" space. This substantially reduced occurrence of negative predicted flows with the Box-Cox transformation, and avoided them altogether with log-sinh. The most noticeable observation about transformation is that combining the input-dependent bias with the Box-Cox transformation we obtained the largest uncertainty bands and among the poorest deterministic model performances.

2.4.2 Analysis of estimated observation errors

In general, the analysis of the measurement errors is consistent with the predictive performance analysis. Again, the error model with a constant bias and the log-sinh transformation is among the ones which best fulfills the statistical assumptions (Fig. 5). The estimate of the observation errors has almost no autocorrelation and relatively low heteroskedasticity. Diagnostic plots on $\mathbf{E}$ are only shown as "Supplement" since the usefulness of formal statistical tests can be questioned when the testable errors are much smaller than the bias.

In contrast, the residuals of the iid error models are heteroskedastic and heavily autocorrelated and thus strongly violate the statistical assumptions. They are also several orders of magnitude larger than observation errors estimated with a bias description. Such huge residuals clearly lose their meaning as random measurement errors of the flow.

Finally, besides frequentist analyses of the white measurement noise, one should check what can be learned from an exploratory analysis of the model bias. Plotting the model bias against flow data (Fig. 6) shows an almost constant scatter with only weak trends. In general, we observed a negative bias in the intermediate flow range and a positive bias during severe storms. While the first systematic deviation is caused by slightly overestimating the runoff in the decreasing limb of the hydrograph, the second reveals that the model systematically underestimates the highest

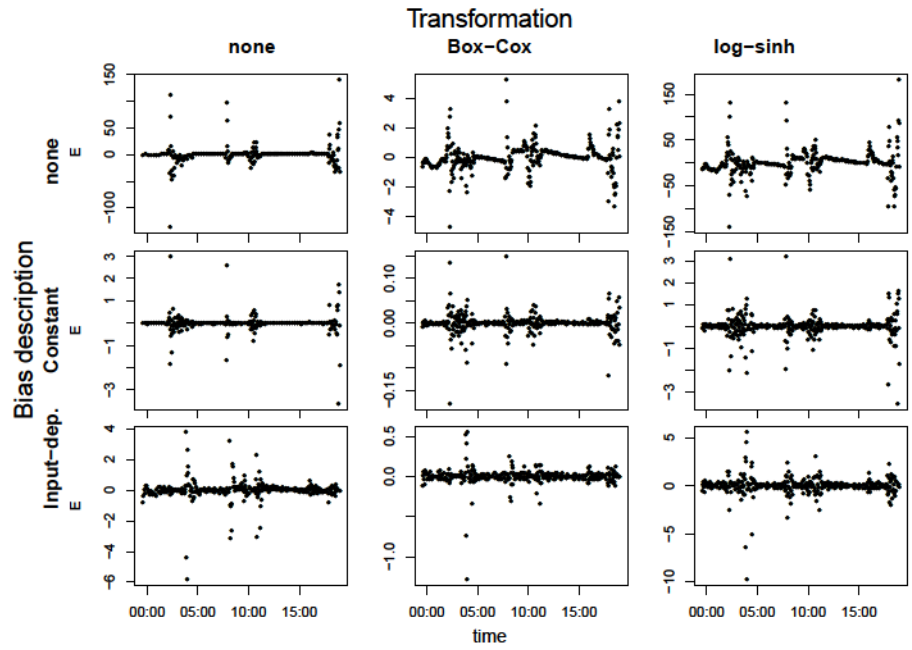

Figure 5: Time series of median of **E**, the part of residuals assumed to come from random observation errors. The range illustrated is the same part of the calibration period as shown in Fig. 4. The ordinate axis is in transformed flow units.

peak discharges (see Fig. 4).

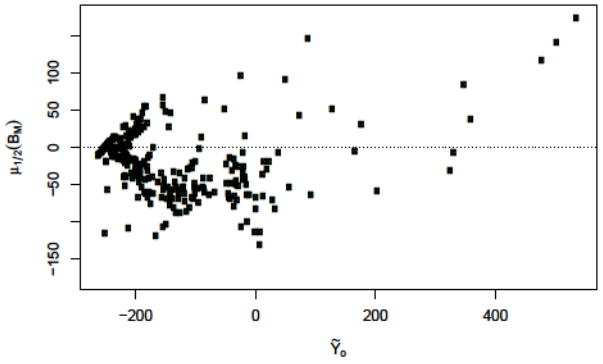

Figure 6: Median of model inadequacy versus transformed observed runoff for the calibration period shown in Fig. 4. Results are shown for the best solution: the constant bias log-sinh transformed error model.

2.5 Discussion

The outcomes of this study, in agreement with theoretical considerations, confirmed that describing bias by means of a stochastic process produces much more reliable and interpretable hydrological predictions than the overly-simplistic traditional error model. Additionally, the bias error model naturally describes the increase in uncertainty about the system response when passing from the calibration to the extrapolation range. In the following, we will interpret the results obtained for our system, assess the differences among the proposed bias description for-

mulations, analyze the dissimilarities between natural and urban hydrology, and finally provide guidelines on how to describe model discrepancies in future studies.

2.5.1 Bias analysis in the case study

The analysis of the uncertainties in our stormwater system demonstrated that our parsimonious deterministic model captures most of the hydrograph dynamics. Nevertheless, it produced partially biased simulations. By accounting for these systematic deviations with the most plausible error model, we observed almost a constant variance of the estimated observation errors in the transformed space. Furthermore, the bias tended to be negative during intermediate rain events and positive bias at very high discharges. These findings indicate that a part of the predictive uncertainty stems from structural deficits due to oversimplification of the simulator, which produce systematic trends in the residuals. Another part of prediction uncertainty, instead, stems from imprecise precipitation measurements, which increases the scatter of the residuals.
If the study's goal is to reduce prediction uncertainty, one should, after detecting structural deficiencies, improve the model. This can be done by modifying process formulations or increasing the model complexity. In our case, analyzing how model discrepancies depend on the measured discharge gave us the necessary information to improve the simulator. Increasing the number of calibration parameters from preliminary simulations reduced a strong positive bias (results not shown) to some mild remaining systematic trends (Fig. 6).

2.5.2 Comparison of different bias descriptions

In this paper we proposed 5 different descriptions of model inadequacy. The bias description where the variance quadratically increases with precipitation is the conceptually most appealing form since it mechanistically accounts for higher uncertainty during rainy periods. Furthermore, in contrast to the empirical output-dependence via data transformation, input-dependence acknowledges that the rising limb of the hydrograph is more uncertain than the recessive limb.
Notwithstanding its theoretical appeal, the input-dependent bias has several drawbacks. First, it has two parameters more than the constant bias, which potentially reduces the robustness of this approach. In particular, during estimation, the proportionality constant κ tends to reach very high posterior values (see Supplement for priors and posteriors) and, in this way, leads to inflated variances during rainfall and too small variances during dry weather. Second, since we always assume a normal distribution of the bias, the input-dependent description frequently requires a transformation anyway in order to avoid negative predictions which are physically meaningless. Third, linked to the two previous considerations, the input and output dependent error model has the tendency to include too many mechanisms to describe model inadequacy. This complex representation reproduces data dynamics "excessively well" and therefore motivates the deterministic model less to fit the observations.
It is interesting to notice that the input-dependent error model never reverted back to a constant bias (i.e. κ never calibrated to 0), even in cases where all performance indicators favored the simpler error description. This can be explained considering that the input-dependent bias has a basic variance plus a variance induced by precipitation. In our case, the posterior basic variance for the input-dependent bias was, irrespective of the transformation, smaller than for the constant bias. This is caused by three combined phenomena: first, an error distribution with smaller variance has generally higher likelihood, second, our simulator could match the baseflow almost exactly, and third, a large part of the calibration period had an output equal to the baseflow. Since the input-dependent error model still had to account for big errors during wet

weather, it did so by increasing the precipitation-induced error variance, producing sometimes too wide uncertainty intervals for the future storm events.

2.5.3 Bias assessment in urban and natural hydrology

Interestingly, for Honti et al. (2013) the input-dependent Box-Cox transformed error model produced the best predictions. This different outcome, however, is not necessarily in contrast with our findings. First, as mentioned in the introduction, different case studies can display extremely dissimilar error properties. Our urban catchment had a much stronger difference in the hydrologic response between dry and wet periods than Honti et al.'s natural watershed, which additionally presented fewer points of constant minimal discharge. Second, their formulation presented an additional precipitation-dependent bias component, while neglecting the delay between precipitation occurrence and uncertainty increase. Third, in Honti et al. (2013) the log-sinh transformation was not implemented.

Comparing urban to natural catchments, an important aspect is, first, that urban hydrosystems react much more rapidly and strongly and therefore require much more frequent measurements of the hydrologic response (typically at minute scale instead of daily scale). Furthermore, the discharge in sewers can be ascertained with much higher precision and accuracy than in the case of rivers. Indeed, in drainage conduits the area-velocity sensors can measure the velocity directly, without requiring a rating curve, which is an additional source of uncertainty in streamflow observations (Sikorska et al., 2013; Montanari and Di Baldassarre, 2013). The elevated temporal density and precision of the measurements, as discussed by Reichert and Mieleitner (2009), leads to an even higher need to address model discrepancies explicitly.

Second, in natural watersheds the high flow can be associated with floodplain inundations which dramatically increase the uncertainty of flow measurements. In sewer systems, by contrast, the well-defined geometry of the pipes and the reliable flow measurement devices lead to a much smaller increase in observation uncertainty with increasing discharge. This situation underlines the advantage of a log-sinh transformation in urban hydrology instead of the traditional Box-Cox. Indeed, the former assumes that residual scatter in high streamflow ranges is limited.

Regarding uncertainty estimation, it seems that a formulation where the standard deviation of the input-dependent bias component linearly increases with precipitation is suboptimal for urban systems. Indeed, most urban water basins, especially those only draining stormwater and having low groundwater infiltration, exhibit extremely high contrasts between low and high flows. Such strong dynamics and linear input dependence, can result in unnecessarily wide uncertainty intervals.

2.5.4 Recommendations

As our results demonstrate, hydrological predictions are more reliable when model deficiencies are considered explicitly, especially for urbanized areas. This is in agreement with many other studies (Breinholt et al., 2012; Yang et al., 2007b; Schoups and Vrugt, 2010; Reichert and Mieleitner, 2009). If the modeller is not interested in separating predictive uncertainty into its contributing sources because data collection or model building processes are fixed, we suggest using a lumped autoregressive error model (Bates and Campbell, 2001; Yang et al., 2007b; Evin et al., 2013). Such formulations are usually sufficient to reliably estimate total output uncertainty. However, it is often useful to assess how far the prediction uncertainty can be reduced by minimizing a particular error source (Sikorska et al., 2012b). In these cases, we recommend applying our five-variant bias description in order to disentangle the effects of model discrepan-

cies and random measurement errors. In particular, we suggest starting with a constant log-sinh transformed bias and setting priors of the error model parameters using the recommendations given in Sect. 2.2.2. This likelihood formulation is simple and robust and proved to perform extremely well in our case study. Then, if the efficiency of this error description is unsatisfactory, the Box-Cox transformation and eventually the input-dependent bias description in its three variants can be applied. Finally, we propose selecting among the error description providing the best validation coverage with the narrowest bands and the most iid transformed observation errors, and applying this likelihood formulation to subsequent predictions.
If the input-dependence is of particular interest though providing dubious predictions, considering what we discussed above, we suggest adapting this dependence on precipitation as a function of the systems dynamics. One possibility could be to modify Eq. (3.4) and, instead of a linear increase of uncertainty with the precipitation, one could adopt a power relationship with the exponent as calibration parameter.

2.6 Conclusions

In this study, we proposed different strategies for obtaining reliable flow predictions and quantifying different error contributions. We adapted a Bayesian description of model discrepancy to urban hydrology, making the bias variance increase during wet weather in five different ways. From the experience gained in this modeling study and theoretical considerations, we conclude that:

i. Due to input uncertainty, structural deficits, and (possibly) systematic errors in flow measurements, urban hydrological simulations are biased. When using precise and high-frequency output measurements to calibrate and analyze the uncertainties of these simulators by means of traditional iid error models, we obtain implausible predictions.

ii. We can obtain much sounder predictions and significantly improve the fulfillment of the assumptions by adding a model discrepancy function to the classical error model. Such a bias term should have a variance that increases during storm events as a function of rainfall and/or runoff. Finally, the results demonstrate that random output observation errors are much smaller than uncertainty due to bias.

iii. In our study, a rather simple constant autoregressive bias with a log-sinh transformation outperformed an input-dependent bias description. Although the latter is conceptually superior, the simpler formulation appeared more suitable for systems with alternating long low flow periods and short high discharge pulses, such as urban watersheds. Indeed, it is apparently less susceptible to producing excessively wide error bands and suboptimal fit. Therefore, we suggest to test this first and then try the other bias descriptions, if necessary.

iv. Although it generally outperforms the traditional error assumptions, we are aware of the limitations of our approach. The presence of bias, to which we have to assign a weight in the form of priors, inevitably introduces subjectivity in the uncertainty analysis. Moreover, this statistical method can 'only' describe the different types of output uncertainties, but cannot address directly the causes of bias nor can it straightforwardly lead to an uncertainty reduction.

v. Despite remaining challenges, our approach has further advantages besides providing separated and plausible prediction intervals. First, a bias description can be associated with a fully-fledged framework which propagates the uncertainty sources and supports bias reduction, to describe "remnant" errors. So far, this remaining bias due to imperfect description of the uncertainty sources has been modeled as white noise. Second, the exploratory analysis of model bias, for example investigating its dependence on the output, can provide valuable insight into whether model discrepancy is dominated by input uncertainty or inadequate model formulation. Third, this statistically sound approach is computationally not more intensive than the classical methods. Indeed, it is cheaper than the mechanistic error propagation frameworks.

vi. Open questions which require further research include how the bias description can be applied to quantify the structural errors of complex hydrodynamic models with multiple outputs. Moreover, it is unclear how input errors can be separated from structural deficits as this requires a probabilistic formulation and propagation through the simulator. A further challenge is that complex sewer models are prohibitively slow for many iterative simulations. This can be possibly overcome by statistical approximations, such as emulators.

Supplementary material related to this article is available online at http://www.hydrol-earth-syst-sci.net/17/4209/2013/hess-17-4209-2013-supplement.pdf.

Acknowledgements

We would like to thank Vojtěch Bareš and Martin Fencl for providing the monitoring data and information about the case study. We are also grateful to David Machac for the computational support and Anna Sikorska, Dmitri Kavetski, Saket Pande, Eberhard Morgenroth, and two anonymous referees for helpful comments. The Swiss National Science Foundation is acknowledged for financing this research within the scope of the COMCORDE project (grant No. CR2212_135551).

Chapter 3

Comparison of two stochastic techniques for reliable urban runoff prediction by modeling systematic errors

D. Del Giudice[a,b], R. Löwe[c], H. Madsen[c], P. S. Mikkelsen[d], J. Rieckermann[a].

[a]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b]ETHZ: Swiss Federal Institute of Technology Zürich, 8093 Zürich, Switzerland
[c]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark.
[d]Department of Environmental Engineering, Technical University of Denmark, Lyngby, Denmark.

Water Resources Research (2015), 51, doi:10.1002/2014WR016678.

Author contributions

D.D.G. and R.L. designed the experiments, performed the analyses, wrote the paper; D.D.G., R.L., P.S.M., and J.R. conceived the experiments; all coauthors gave advices, supported result interpretation and paper revision.

Abstract

In urban rainfall-runoff, commonly applied statistical techniques for uncertainty quantification mostly ignore systematic output errors originating from simplified models and erroneous inputs. Consequently, the resulting predictive uncertainty is often unreliable. Our objective is to present two approaches which use stochastic processes to describe systematic deviations and to discuss their advantages and drawbacks for urban drainage modeling. The two methodologies are an external bias description (EBD) and an internal noise description (IND, also known as stochastic grey-box modeling). They emerge from different fields and have not yet been compared in environmental modeling. To compare the two approaches we develop a unifying terminology, evaluate them theoretically, and apply them to conceptual rainfall-runoff modeling in the same drainage system. Our results show that both approaches can provide probabilistic predictions of wastewater discharge in a similarly reliable way, both for periods ranging from a few hours up to more than one week ahead of time. The EBD produces more accurate predictions on long horizons but relies on computationally heavy MCMC routines for parameter inferences. These properties make it more suitable for off-line applications. The IND can help in diagnosing the causes of output errors and is computationally inexpensive. It produces best results on short forecast horizons that are typical for on-line applications.

3.1 Introduction

Any model in urban hydrology usually delivers results that substantially differ from observations of water level, flow, or water quality (Dotto et al., 2012). These mismatches between modeled and observed output are caused by errors in the input estimation and by simplifications of the system description (Del Giudice et al., 2013). These systematic output deviations can affect the operation of urban drainage and wastewater systems as well as design decisions, which are usually based on model predictions (Vezzaro and Grum, 2014). Consequently, an appropriate description of these systematic deviations can meliorate forecasting and control (Löwe et al., 2014a). Significant efforts have therefore been made in past and recent hydrological research to quantify the uncertainties of model results (Jonsdottir et al., 2007; Yang et al., 2007a; Salamon and Feyen, 2010; Breinholt et al., 2012; Freni and Mannina, 2012; Sikorska et al., 2012b; Evin et al., 2013; Honti et al., 2013).

Runoff modeling in urban hydrology distinguishes itself from its counterpart in natural catchment hydrology by the usually smaller temporal and spatial scales involved in peak discharge generation. Typical time steps for peak discharge simulations are 6 (Kleidorfer et al., 2009) to 15 minutes (Breinholt et al., 2012), but seconds (Freni et al., 2009a) to days (Mejía et al., 2014) have been reported. Typical study areas of sewer watersheds range from dozens (Del Giudice et al., 2015b) to more than one thousand hectares (Breinholt et al., 2011). Furthermore, the majority of sewer peak flow comes from sealed surfaces which dominate urban landscapes (Coutu et al., 2012b). As a result, concentration times of one hour or less are common, which makes model predictions highly sensitive to variations of rainfall input on small scales. This sensitivity to input uncertainty was underlined by previous investigations which suggested that forecasting errors are mainly due to discrepancies in the rainfall input, in particular an insufficient quantification of the spatial rainfall distribution on a scale of a few kilometers or less (Schilling and Fuchs, 1986; Sikorska et al., 2012b; Borup et al., 2013).

The systematic rainfall errors, their routing through a possibly non-linear model, and deficits

in the model structure usually lead to an autocorrelated and heteroscedastic behavior of the residuals of runoff simulations (see Reichert and Mieleitner (2009) or Evin et al. (2013)). Most of the techniques applied for uncertainty quantification in urban hydrology do not explicitly account for this dynamic nature of model errors. Typically, only parametric uncertainty and output measurement noise are considered. This usually leads to biased parameter estimates and to suboptimal forecasting (Thyer et al., 2009; Schoups and Vrugt, 2010; Willems, 2012; Del Giudice et al., 2013).
Recent developments have focused on the attempt to account for systematic behavior of runoff model residuals (by some authors referred to as model bias or discrepancy). The present work aims at comparing two such approaches that have recently been applied in urban hydrology (Bechmann et al., 2000; Breinholt et al., 2012; Del Giudice et al., 2013; Löwe et al., 2014a). In the following, we will denote them as "external bias description" (EBD) and "internal noise description" (IND). Both approaches aim at describing and compensating for the dynamic variations of model residuals. However, they are implemented in different mathematical frameworks, originate from different scientific fields, utilize a distinct terminology, and to date focus on dissimilar applications.
The EBD, on the one hand, was developed against the background of statistical inference in a regression-type framework (see Craig et al. (2001); Kennedy and O'Hagan (2001); Higdon et al. (2005); Bayarri et al. (2007); Reichert and Schuwirth (2012), for example) and has a strong focus on the estimation of parameters and system output, as well as their related uncertainties. The IND, on the other hand, originated from research related to stochastic processes and time series analysis, and was originally applied to forecasting and control of engineered systems such as chemical reactors or heating systems (see Bechmann et al. (2000); Kristensen et al. (2004, 2005) and Friling et al. (2009), for example).
Based on the existing literature, it is difficult to identify the relative advantages and disadvantages of the approaches and to make recommendations on their overall applicability which depends on forecasting horizon and model type. Therefore, the main objectives and innovations of this work are to:

Q1. Present in commensurate terms two advanced approaches for probabilistic model calibration and predictions. Because of their different origins, the EBD and IND have been presented with dissimilar "idioms", which has hindered the collaboration between their respective communities.

Q2. Explore new aspects of the two approaches. For the EBD, this implies testing its performance in short-term predictions, in combined sewer flow modeling, and in the presence of substantial and non-stationary model deficiencies. For the IND, this means testing its performances in discrete short and long-term predictions, observing the uncertainty expansion from the last observation point, and discussing its likelihood function in more detail.

Q3. Discuss the lessons learned from the two approaches and their respective strengths and weaknesses. To do so, we consider both theoretical aspects and the performances of the EBD and IND when applied to a common and complex system and an oversimplified model.

The discussions and results of this investigation will help the modeler to make a more conscious choice about which method to adopt. This choice will depend on the study resources (e.g. black-

box/modifiable model, sufficient/limited computational power) and goals (e.g. predicting over long/short horizons). Furthermore, the reciprocal understanding of the EBD and IND ensuing from this study will help direct future developments of both approaches.

3.2 Brief review of methods applied for uncertainty quantification in conceptual rainfall-runoff modeling

This section provides a brief overview of the techniques applied for quantifying uncertainties, with a focus on conceptual rainfall-runoff modeling. We classify the techniques as shown in Table 1 according to their main characteristics: model formulation (rows) and representation of the errors (columns).

Table 1: Probabilistic approaches for runoff predictions. We included examples from the urban drainage (marked with an asterisk*) and natural hydrology literature. Note that it is not possible to assume the residual errors to be independent and identically distributed (iid) when the system equations contain a noise term.

	Errors iid	**Systematic dynamic deviations described**	**Error sources represented**
Output error modeling (deterministic model + stochastic errors)	Dotto et al. (2012)* Freni and Mannina (2012)* Kleidorfer et al. (2009)* Vezzaro et al. (2013a)*	Del Giudice et al. (2013)* Kuczera (1983) Schoups and Vrugt (2010) Wilkinson et al. (2011)	Kavetski et al. (2006) Renard et al. (2010) Sikorska et al. (2012b)* Sun and Bertrand-Krajewski (2013)*
Internal error modeling (stochastic model + stochastic errors)	- - -	Breinholt et al. (2011, 2012)* Löwe et al. (2014a)* Moradkhani et al. (2012)	Beck and Young (1976) Vrugt et al. (2005) Bulygina and Gupta (2009) Reichert and Mieleitner (2009) Salamon and Feyen (2010)

A natural distinction of the different approaches derives from the way the model is formulated (Renard et al., 2010). In hydrology, we traditionally model the output of a system by using a deterministic model (or simulator). The model output can then be combined with one or more probabilistic error terms. This approach is shown in the first row of Table 1 and we denote it as "output error modeling". Alternatively, the model itself can be stochastic. This is usually done by considering the model states (e.g., in Vrugt et al. (2005); Breinholt et al. (2012); Moradkhani et al. (2012)) or parameters (e.g., in Beck and Young (1976); Reichert and Mieleitner (2009)) as time-varying, random variables. Such approaches are usually implemented in a state-space form, which is common in system theory and statistical filtering (Lin and Beck, 2007; Bulygina and Gupta, 2009; Quinn and Abarbanel, 2010). The model output, a function of these stochastic states, is additionally affected by an observation error term, and the approach is usually combined with data assimilation methods. We denote these approaches as "internal error modeling" and summarize them in the second row of Table 1.
Complementary to how they formulate the model, methods for uncertainty analysis of runoff predictions can be classified by how they characterize modeling errors (columns in Table 1). We suggest distinguishing between three cases:

- Approaches that do not explicitly account for dynamic model discrepancies. These may be Bayesian approaches which assume uncorrelated model residuals or pseudo-Bayesian approaches (such as GLUE (Beven, 1993)). Common to these frameworks is that input and structural uncertainties are assigned to the (constant) model parameters (see discussions in Yang et al. (2008) and Reichert and Mieleitner (2009)). As a result, parameter estimates can become difficult to interpret and the resulting output prediction intervals may be unreliable (Renard et al., 2010; Reichert and Schuwirth, 2012).

- Approaches that explicitly account for dynamic model discrepancies in their formulation. In the case of output error modeling, this can be done by adding a time-varying error term to the model output (for example ARMA models as already suggested by Kuczera (1983) or stochastic differential equations (SDEs) as in Yang et al. (2008) and Del Giudice et al. (2013)). In the case of internal error modeling, a random noise is added to the states to reflect that the states can rarely be predicted exactly (see Breinholt et al. (2012), for example). The state noise provides a quantification of forecast uncertainties. In both methods, structural and input uncertainty are aggregated into one term.

- Approaches that, instead of just describing the output errors, focus on identifying the causes of model inadequacies. To quantify input uncertainty, rainfall multipliers have been proposed (Kuczera et al., 2006; Sun and Bertrand-Krajewski, 2013). Structural uncertainty, instead, has been dealt with by inferring the model equations (Bulygina and Gupta, 2009), the behavior of dynamic parameters (Reichert and Mieleitner, 2009), or the value of model parameters and states (Vrugt et al., 2005).

From the literature (e.g., Dotto et al. (2012); Sikorska et al. (2012b); Del Giudice et al. (2013)), it is clear that the majority of uncertainty modeling studies in urban hydrology do not account for time-dependent systematic model errors. In contrast, the two approaches considered in this article explicitly account for systematic dynamic output errors (second column of Table 1). However, they are generally less conceptually complex and computationally demanding than those presented in the third column of Table 1. The EBD is an output error modeling approach (first row of Table 1), while the IND is an internal error modeling approach (second row of Table 1). The works of Breinholt et al. (2012) and Del Giudice et al. (2013) in urban hydrology and multiple works in natural catchment hydrology (see Table 1) have demonstrated that such approaches are generally capable of producing reliable predictions in conceptual rainfall-runoff modeling.

3.3 Methods

3.3.1 Terminology

We here provide a brief unifying nomenclature to describe our analyses with the two methodologies. We also mention alternative terminology used in hydrology, statistics, and control theory. An illustrative description of this terminology is given in Figure 1.

Parameter estimation consists in identifying parameter values by comparing the model and the output observations. This learning process is also known as parameter inference (Reichert and Schuwirth, 2012), calibration (O'Hagan, 2006), or inverse modeling.

Smoothing refers to identifying system states and/or outputs in a past time, e.g. the calibration period, using the available data before and after that point (Bulygina and Gupta, 2009; Law and Stuart, 2011).

Forecasting denotes the generation of model outputs (and states) starting from the last observation up to an arbitrary number of time steps in the future. This process is also loosely described as making predictions (in the validation period) (Dietzel and Reichert, 2012; Renard et al., 2010; Law and Stuart, 2011; Einicke, 2012), simulations (Platen and Bruti-Liberati, 2010)

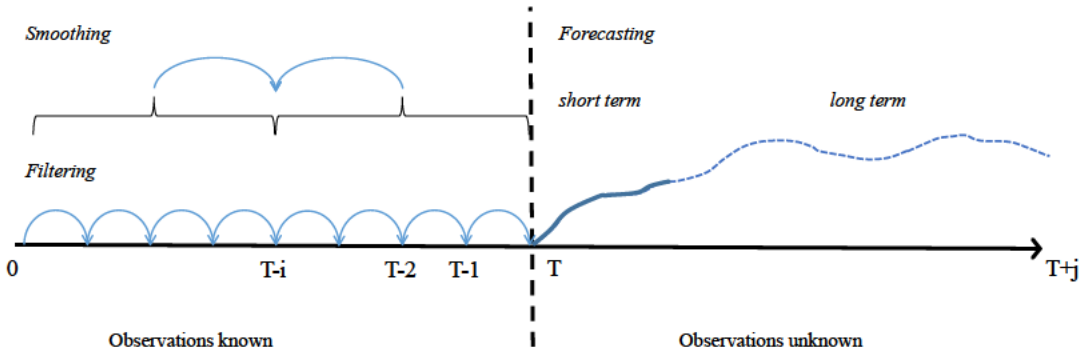

Figure 1: Illustration of the different types of predictions according to the conditioning on output observations.

or, more precisely, ex-post hindcasting (when the input is assumed to be known) (Beven and Young, 2013).

Filtering consists in characterizing the system state at the current time given inputs and observations up to the current time point (Bulygina and Gupta, 2009; Platen and Bruti-Liberati, 2010; Law and Stuart, 2011). Data assimilation is also used to define this process of learning about the current state (O'Hagan, 2006).

3.3.2 Two approaches to explicitly account for dynamic systematic errors in rainfall-runoff modeling

We here explain the external bias description (EBD) and the internal noise description (IND). While the first adds a stochastic process to the system output, the second adds a stochastic process to the states *and* to the output.

Output error modeling and external bias description (EBD)

In deterministic conceptual modeling, differential equations are applied to describe the variation ds of a set of model states s (e.g., water level in an unobserved combined sewer overflow tank, hydraulic heads in specific points of an aquifer, soil moisture content in a catchment) depending on a vector of driving forces (e.g., a rainfall time series) $\mathbf{x}$ and parameters $\boldsymbol{\theta}$ in a function f_{M} (equation 3.1). Bold minuscule denote deterministic vectors while bold majuscule denote stochastic vectors.

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} = f_{\mathrm{M}}(\mathbf{s}, \mathbf{x}, t, \boldsymbol{\theta}). \tag{3.1}$$

The model output $\mathbf{y}_{\mathrm{M}}$ relates to the model states, input, and parameters through a function h:

$$\mathbf{y}_{\mathrm{M}} = h(\mathbf{s}, \mathbf{x}, t, \boldsymbol{\theta}). \tag{3.2}$$

So far, no modeling error has been considered. In order to account for the fact that no system description is perfect and that output observations are affected by errors, two strategies are possible: external or internal error modeling. In external (or output) error modeling, the observed system output $\mathbf{Y}_o$ (e.g., measured discharge just before the entrance of a sewage treatment plant) can be represented as the sum of $\mathbf{y}_{\mathrm{M}}$ plus a stochastic error term. This term aggregates modeling and observation errors and can be independently and identically distributed (iid) (e.g.,

in Kleidorfer et al. (2009); Freni and Mannina (2012)), or autocorrelated in time (e.g., in Kuczera (1983); Bates and Campbell (2001); Frey et al. (2011); Evin et al. (2013)). Several studies (e.g., Yang et al. (2007a); Sikorska et al. (2012b); Honti et al. (2013)) have demonstrated that describing the autocorrelated behavior of the errors produces more reliable predictions. Instead of adding only one autocorrelated error term, recent statistical literature has suggested considering observation noise in addition to input, structural, and parameter uncertainty (equation 2.1) (Craig et al., 2001; Kennedy and O'Hagan, 2001; Higdon et al., 2005; Bayarri et al., 2007). Following the notation of Reichert and Schuwirth (2012), who transferred this approach to environmental modeling, we model the observable system output as:

$$\mathbf{Y}_o = \mathbf{y}_\mathrm{M}(\mathbf{s}, \mathbf{x}, t, \boldsymbol{\theta}) + \mathbf{B}_\mathrm{M}(\mathbf{x}, t, \boldsymbol{\psi}) + \mathbf{E}(\boldsymbol{\psi}), \tag{3.3}$$

where $\mathbf{B}_\mathrm{M}$ is a random process that mimics systematic deviation of model results from the true system output, $\mathbf{E}$ represents uncorrelated observation errors, and $(\boldsymbol{\theta}, \boldsymbol{\psi})$ are the parameters of the simulator and error model. Simplified iid approaches only consider $\mathbf{E}$ while neglecting $\mathbf{B}_\mathrm{M}$. To further improve the error description, modeled and observed outputs could be transformed by a function. This can be useful in hydrology, where the error variance increases during peak discharge. This effect can, however, also be reproduced by a heteroschedastic error model (Evin et al., 2013; Del Giudice et al., 2013). In this study, we achieve satisfactory results with an input-dependent bias description. The specific formulation we use assumes that the bias follows an Ornstein-Uhlenbeck process with input-dependent variance (see Honti et al. (2013) for derivation). In other words, $\mathbf{B}_\mathrm{M}$ is modeled as a continuous version of a first-order autoregressive process with normal independent noise whose variance grows with the rain rate, x, shifted in time by a lag d. The evolution of $\mathbf{B}_\mathrm{M}$ and $\mathbf{E}$ for the scalar case are described by equations 3.4 and 3.5:

$$\mathrm{d}B_\mathrm{M}(t) = -\frac{B_\mathrm{M}(t)}{\tau}\mathrm{d}t + \sqrt{\frac{2}{\tau}\left(\sigma_{B_\mathrm{ct}}^2 + (\kappa x(t-d))^2\right)}\mathrm{d}W(t), \tag{3.4}$$

$$E(t) = \sigma_E E_\mathcal{N}, \tag{3.5}$$

where κ is a scaling factor, d denotes the response time of the system to rainfall, τ is the correlation time of the error process, and σ_{B_ct} is the asymptotic standard deviation of the random fluctuations around the equilibrium. $\mathrm{d}W(t)$ represents increments of a standard Wiener process and therefore has a normal distribution (Kloeden and Platen, 1999; Iacus, 2008), while $E_\mathcal{N}$ is a standard normal random variable.

Internal error modeling and internal noise description (IND)

An alternative way to account for uncertainties when modeling the behavior of a hydrosystem with equations 3.1 and 3.2 is via internal error modeling, which is usually applied in combination with state updating (Kristensen et al., 2004; Moradkhani et al., 2012). Instead of adding stochasticity only to the system output, this approach (also known as state-space modeling or stochastic grey-box modeling) describes the internal evolution of the system as:

$$\mathrm{d}\mathbf{S} = f_\mathrm{M}(\mathbf{S}, \mathbf{x}, t, \boldsymbol{\theta})\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{S}, \mathbf{x}, t, \boldsymbol{\psi})\mathrm{d}\mathbf{W}(t). \tag{3.6}$$

This so-called "state" (or "transition", or "system") equation describes the continuous evolution

of some "hidden" (or "latent") states $\mathbf{S}$ which, being now stochastic, directly account for modeling errors. This vector of usually-unmeasurable variables can be estimated from the measured outputs (Einicke, 2012). $\boldsymbol{\sigma}$ is called "diffusion term", "state noise", or "level disturbance" and accounts for modeling errors by making the states uncertain or random. $f_{\mathrm{M}}(\cdot)$ is called "drift term" and corresponds to the functions constituting the deterministic (part of the) model M. Adding noise to the state equations reflects that the states cannot be predicted exactly, such that any statement about future values of the states must be probabilistic. Hence the model itself is stochastic. This is an important distinction from the EBD, where randomness is only added to the model output. The IND is instead more similar to approaches making model parameters stochastic and time-varying (Reichert and Mieleitner, 2009).
The dynamics of the observed output $\mathbf{Y}_o$ are related to the state equations via an observation equation:

$$\mathbf{Y}_o = h(\mathbf{S}, \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}, t) + \mathbf{E}(\boldsymbol{\psi}), \tag{3.7}$$

which is a potentially non-linear function of states $\mathbf{S}$ and parameters $(\boldsymbol{\theta}, \boldsymbol{\psi})$. The modeled observation process $\mathbf{Y}_o$ is assumed to be subject to independent random normal observation errors $\mathbf{E}$. Similarly to the EBD, transformations can be applied to the observed and modeled output (Breinholt et al., 2012).
We here parametrized the diffusion term as linearly increasing with the model states

$$\boldsymbol{\sigma}(\mathbf{S}, \mathbf{x}, t, \boldsymbol{\psi}) = diag\big(\boldsymbol{\sigma}_s \circ \mathbf{S}\big) \tag{3.8}$$

where $\circ$ is the Hadamard (entrywise) product between the vector of diffusion parameters $\boldsymbol{\sigma}_s$ and the vector of states $\mathbf{S}$, and *diag* indicates that the matrix is diagonal. This formulation produced satisfactory results in previous urban hydrological studies (Breinholt et al., 2011; Löwe et al., 2014a). The assumption of state-dependent noise in the IND is another relevant distinction from the EBD, where the additive noise terms can depend on the input or output, but not on a (hidden) state variable.
The linear state-dependent diffusion imposes a log-normal distribution on the model outputs. We thus use the logarithmic transformation of the modeled and observed outputs for parameter inference. We then back-transform $h(\mathbf{S}, \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}, t) + \mathbf{E}(\boldsymbol{\psi})$ into the real space for forecasting.
As the numerical solution of equation 3.6 with stochastic state-dependent diffusion can be challenging, a Lamperti transformation is commonly applied (Kloeden and Platen, 1999; Iacus, 2008; Moeller, 2010; Breinholt et al., 2011).

3.3.3 Inference and generation of model outputs

To describe how the EBD and IND differ regarding parameter estimation and forecasting, we first discuss the approaches on a conceptual level before addressing their numerical implementation.

Parameter estimation

In a probabilistic framework, the inverse problem of parameter estimation requires assumptions about the error distribution. These assumptions are usually formalized by a likelihood function $\mathcal{L}_M\big(\mathbf{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x}\big)$ that describes the conditional probability density of producing the observed output data given a certain model structure M, inputs $\mathbf{x}$, and parameters $(\boldsymbol{\theta}, \boldsymbol{\psi})$. Calibration parameters of the hydrological model $(\boldsymbol{\theta})$ and of the error description $(\boldsymbol{\psi})$ are presented in Table 2.

Table 2: Conceptual model and error model calibration parameters ($\boldsymbol{\theta}, \boldsymbol{\psi}$). The notation for prior distributions is: LN(μ, σ): lognormal, N(μ, σ): normal, TN(μ, σ, a_1, a_2): truncated normal, Exp(λ^{-1}): exponential. The symbols are: μ: expected value, σ: standard deviation, a_1: lower limit, a_2: upper limit, λ: rate.

Name	Description and alternative name	Units	Prior (for EBD)	Prior (for IND)
Deterministic model parameters ($\boldsymbol{\theta}$)				
$ln(A_{imp})$	$\log_e$ of the impervious catchment (A)	$ln(ha)$	N(4.31,0.86)	N(4.31,0.86)
k	mean reservoir residence time	h	TN(4.5,0.9,0,∞)	TN(4.5,0.9,0,∞)
$s_{1,0}$	initial condition of reservoir 1 ($s1\_ini$)	m^3	LN(675, 135)	-
$s_{2,0}$	initial condition of reservoir 2 ($s2\_ini$)	m^3	LN(675, 135)	-
$ln(s_{1,0})$	initial condition of reservoir 1	$ln(m^3)$	-	N(6.5, 0.19)
$ln(s_{2,0})$	initial condition of reservoir 2	$ln(m^3)$	-	N(6.5, 0.19)
Error model parameters ($\boldsymbol{\psi}$)				
τ	correlation length of $\mathbf{B}$ ($corrlen$)	h	LN(10,3)	-
$\sigma_{B_{ct}}$	standard deviation of $\mathbf{B}$ ($sd.B\_Q$)	m^3/h	TN(0,40,0,∞)	-
κ	proportionality constant between input and uncertainty increase (ks_Q)	m^2	TN(0,57965,0,∞)	-
d	lag (in timesteps) between input and uncertainty increase ($Delta$)	10 min	Exp(6)	-
σ_{s_1}	diffusion scaling for $ln(s_1)$	[-]	-	N(-10,1000)
σ_{s_2}	diffusion scaling for $ln(s_2)$	[-]	-	N(-10,1000)
σ_E	standard deviation of $\mathbf{E}$ ($sd.Eps\_Q$)	m^3/h	LN(20,2)	-
$ln(\sigma_E)$	standard deviation of $\mathbf{E}$	$ln(m^3/h)$	-	N(-2.55,0.255)

Parameter estimation in the EBD approach In the current state of the EBD approach, we assume that the data generating process follows a multivariate normal distribution with mean $\mathbf{y}_M$ and covariance $\boldsymbol{\Sigma}$:

$$\mathcal{L}_M(\mathbf{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x}) = \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{\det\big(\boldsymbol{\Sigma}(\boldsymbol{\psi}, \mathbf{x})\big)}} \exp\left(-\frac{1}{2}\Big[\mathbf{y}_o - \mathbf{y}_M(\boldsymbol{\theta}, \mathbf{x})\Big]^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\psi}, \mathbf{x})^{-1} \Big[\mathbf{y}_o - \mathbf{y}_M(\boldsymbol{\theta}, \mathbf{x})\Big]\right) \tag{3.9}$$

where n is the number of observations i.e. the length of the vector $\mathbf{y}_o$ (e.g., a measured discharge time series at the outlet of a catchment). $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{B}_M} + \boldsymbol{\Sigma}_{\mathbf{E}}$ is the total error covariance matrix accounting for the autocorrelated and heteroskedastic bias process arising from input and structural errors and for iid observation errors.

Since equation 2.1 has three terms to identify given one observation vector, a Bayesian approach involving the use of prior information is necessary (Craig et al., 2001; Bayarri et al., 2007; Reichert and Schuwirth, 2012). For statistical inference, the likelihood function is combined with the prior information on parameters to infer their posterior distribution according to Bayes' law:

$$f_{post}(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_o, \mathbf{x}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{\psi}) \mathcal{L}_M(\mathbf{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})}{\iint f(\boldsymbol{\theta}, \boldsymbol{\psi}) \mathcal{L}_M(\mathbf{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}} \quad . \tag{3.10}$$

Numerically, we approximated this distribution by a Markov chain Monte Carlo (MCMC) algorithm (Honti et al., 2013; Del Giudice et al., 2013).

Parameter estimation in the IND approach Considering the focus of the IND on online (i.e. real time) applications, computationally efficient routines for parameter inference are important. For time series data, the likelihood function is given as a product of one-step-ahead conditional densities (Box et al., 2008; Madsen, 2007). This approach is more efficient and easier to implement than sampling from the multivariate likelihood function when accounting for all the observations at a time. This likelihood would be a path integral, i.e. an infinite-dimensional integral over all possible realizations of the model states (e.g., Restrepo (2008), Balaji (2009) and Quinn and Abarbanel (2010)). We define:

$$\mathcal{L}_M(\mathbf{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x}) = \left(\prod_{i=2}^{n} p(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})\right) p(\mathrm{y}_{o_1}|\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x}) = \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{\det\big(\boldsymbol{\Sigma}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})\big)}} \cdot$$

$$\exp\left(\sum_{i=2}^{n}\Big(-\frac{1}{2}\Big[\mathrm{y}_{o_i} - \mathrm{E}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})\Big]^{\mathrm{T}} \boldsymbol{\Sigma}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})^{-1}\Big[\mathrm{y}_{o_i} - \mathrm{E}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x})\Big]\Big)\right) \tag{3.11}$$

$$\cdot p(\mathrm{y}_{o_1}|\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x}),$$

where $\mathrm{E}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}, \boldsymbol{\psi})$ is the mean and $\boldsymbol{\Sigma}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}, \boldsymbol{\psi})$ the covariance of the one-step-ahead predictions generated using an extended Kalman Filter. This product of conditional densities assumes independence and normality of the one-step-ahead forecast errors ("innovations") at each time step given the observations up to time $i-1$. These innovations are the results of input and structural errors. It is implicitly assumed that the transformed states given all observations up to $i-1$ are also normally distributed (Law and Stuart, 2011) and that they follow a Markov process (Bulygina and Gupta, 2009; Moradkhani et al., 2012). To gain insight into whether the conditional densities of the states can be considered Gaussian, we can analyze the empirical distribution of the one-step-ahead errors.

In the IND, inference is usually performed on a frequentist basis (Breinholt et al., 2012), but a Bayesian framework has also been adopted (Melgaard, 1994; Sadegh et al., 1994). For comparability with the EBD, we will use a Bayesian calibration and therefore also make use of equation 3.10. Traditionally, in the IND, Bayesian estimation has consisted in maximizing the posterior $f(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_o, \mathbf{x})$ rather than characterizing its full distribution (Melgaard, 1994; Sadegh et al., 1994; Walter and Pronzato, 1997). Numerically, the so-called maximum a posteriori (MAP) estimation is here performed with an extended Kalman filter (EKF) (Law and Stuart, 2011). The EKF provides a consistent first-order approximation to the estimate of a nonlinear model at the observation time, as well as the errors of this estimate (Kao et al., 2004). Details on the EKF equations can be found in Appendix 3.A. Quinn and Abarbanel (2010), Balaji (2009), and Law and Stuart (2011) provide further discussions on the assumptions behind approximate Gaussian filters (as the EKF).

Smoothing

It can be useful to predict system output and/or states for points in time where flow data have been employed for parameter inference, in the so-called "calibration period" (or calibration layout). This retrospective analysis, called smoothing, consists in identifying system states (or output) from all available (noisy) output data (Einicke, 2012; Bulygina and Gupta, 2009).

Smoothing with EBD Here, we condition the Gaussian bias process on the observations and updated parameters, and propagate the parametric uncertainty of the simulator and the error models via Monte Carlo simulations (Reichert and Schuwirth, 2012; Del Giudice et al., 2013). To predict the observed system response in the calibration layout, we approximate the distributions of $\mathbf{y}_\mathrm{M} + \mathbf{B}_\mathrm{M} + \mathbf{E}$ for every temporal point i of the dataset, i.e. for $i = 1, \ldots, n$.

Smoothing with IND Commonly, the IND is applied in combination with extended Kalman filtering to update the model states considering one data point at a time (Kristensen and Madsen, 2003). For comparability with the EBD, we here generate smoothened estimates of the model states and outputs in the calibration period. In this setting, conditioning on data can be performed by combining a filter moving forward in time with one going backwards (i.e. from the future to the present) (Einicke, 2012). The smoothed model states are assumed to be normally distributed and related to the output through equation 3.7.

Forecast of future output

Forecast with EBD The posterior predictive distribution of runoff in the extrapolation layout (also called validation period) is computed via Monte Carlo simulations. To approximate the distribution of $\mathbf{y}_\mathrm{M} + \mathbf{B}_\mathrm{M} + \mathbf{E}$, we first obtain realizations of $\mathbf{y}_\mathrm{M}$ by propagating a sample of $\boldsymbol{\theta}_{post}$ through f_M. Second, we compute trajectories of $\mathbf{B}(\boldsymbol{\psi}_{post})$ and $\mathbf{E}(\boldsymbol{\psi}_{post})$ and add them to the results of the simulator (Reichert and Schuwirth, 2012; Del Giudice et al., 2013). In this procedure, the bias-corrected model is not conditioned on data and therefore its predictive uncertainty becomes larger than in the calibration period. However, as the autocorrelated bias has a "memory", observed output still influences these predictions if the analyzed time is close to the last calibration point. An explanation on how to produce EDB forecasts in this (initial) extrapolation phase is given in Appendix 3.C.

Forecast with IND Unconditional output can be generated from stochastic grey-box models by performing "scenario (or ensemble) simulations" (Platen and Bruti-Liberati, 2010) from equation 3.6. To compute trajectories from the stochastic differential equations describing the state-space model, we use discrete-time approximations. For each solution of equation 3.6, the predictions for $\mathbf{Y}_o$ are derived by inserting the simulated paths of the states into equation 3.7. In this setting, normality is assumed only for the model states at the forecast starting point j, conditional on the previous time steps observations $\mathbf{Y}_{o,j-1}$.

3.3.4 Design of computer experiments

To compare the performances of the two approaches, we performed three numerical experiments. First, we analyzed the parameter estimates we obtained after calibration. Second, we compared the quality of long-term predictions over 14 days (5328 time steps) and, third, short-term forecasts over 200 minutes (20 time steps). Although this is longer than the usual 1-5 time steps of on-line applications, we selected this forecasting horizon for illustrative purposes. Since future rainfall was assumed known, both types of predictions were, strictly speaking, ex-post hindcasts.

3.3.5 Performance metrics

To evaluate the performances of the EBD and IND, we used 4 performance metrics, together with a visual inspection of model predictions and quantile-quantile plots. To assess the quality of the underlying deterministic model, we considered the median of the probabilistic simulations. We used i) the Nash-Sutcliffe efficiency index (NS, optimally approaching 1 from below) and ii)

the normalized (or relative) bias (NB, optimally approaching 0). Both statistics are commonly used in hydrology to assess the accuracy in fitting the peaks of the hydrographs and preserve water balance, respectively (Bennett et al., 2013; Bulygina and Gupta, 2009; Coutu et al., 2012b).

To assess the quality of ex-post forecasts, we focused on 95% prediction intervals, while also analyzing the other quantiles via QQ plots (Supporting Information). Specifically, we evaluated the iii) "coverage", which measures the percentage of validation measurements falling into the 95% prediction intervals and iv) the interval (skill) score ($\mathsf{S}^{int}_{0.05}$, optimally approaching 0 from above), which provides a simultaneous assessment of the precision and reliability of the prediction intervals (Gneiting and Raftery, 2007):

$$\mathsf{S}^{int}_{\alpha} = (u - l) + \frac{2}{\alpha}(l - \mathrm{y}_{o_j})H\{l - \mathrm{y}_{o_j}\} + \frac{2}{\alpha}(\mathrm{y}_{o_j} - u)H\{\mathrm{y}_{o_j} - u\} \tag{3.12}$$

where $\alpha = 0.05$ corresponds to the confidence level, u and l to the 97.5 and 2.5 quantiles of the predictive distribution of $\mathbf{Y}_o$ at the time point j, and y_{o_j} to the data in the extrapolation layout. H denotes the unit step function, which takes the value of 1 if its argument is greater than 0 and 0 otherwise. We averaged $\mathsf{S}^{int}_{0.05}$ over all time steps considered.

3.4 Hydrological Application

In the following, we describe the analyzed watershed, the deterministic model used, the available hydrological measurements, the chosen priors, and the computer implementation of our study.

3.4.1 Case study

For our application, we chose the sewer system located in the Ballerup area close to Copenhagen (Denmark) (Figure 2). The catchment has a total surface area of approximately 1300 ha and is mainly laid out as a separate system, although it does have a small combined section. The runoff in this area is strongly influenced by rainfall-dependent infiltration, and the catchment contains several basins and pumping stations. Several previous modeling studies were undertaken using this catchment (Breinholt et al., 2011, 2012; Löwe et al., 2014a). Tipping bucket rain gauge measurements were available from the Danish Water Pollution Committee's (SVK) network (Jørgensen et al., 1998). 1-minute observations from the two pluviometers located near the catchment were averaged and used as input for the runoff model. Flow measurements were available with a temporal resolution of 5 minutes. The time of concentration of the catchment is approximately 60 minutes. As Schilling (1991) recommends a temporal resolution of rainfall measurements of at least 0.2 to 0.33 times the concentration time of an urban watershed, we adopted a modeling time step of 10 min and averaged flow and pluviometric data to this time discretization.

3.4.2 A parsimonious hydrological model

The sewer flow at the monitoring point, $\mathbf{y}_{\mathrm{M}}$, is modeled as a superposition of wastewater flow and rainfall-runoff. While the stormwater runoff (equations 3.13 and 3.14) is described by a cascade of two virtual reservoirs, the wastewater hydrograph (equation 3.15) is represented as a superposition of 4 harmonic functions (Figure 3). The model dynamics is defined by the following deterministic equations:

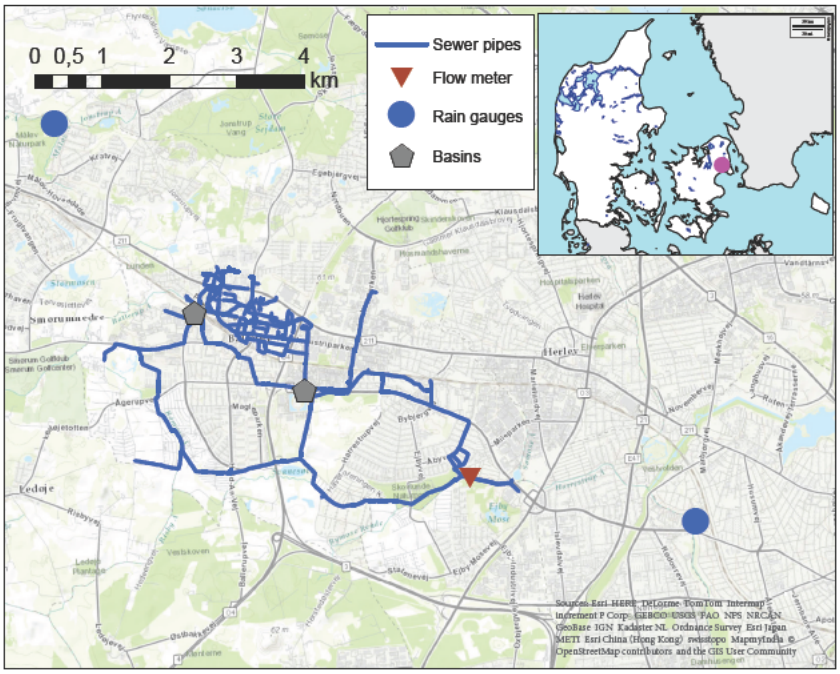

Figure 2: The studied Ballerup sewer network with the rain gauges used for deriving the model input and the flow meter used for measuring the system output.

$$f_{\mathrm{M}}(\mathbf{s},\mathbf{x},t,\boldsymbol{\theta})\mathrm{d}t = \mathrm{d}\begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix} = \begin{bmatrix} A_{imp}\cdot x(t) + a_0 - \frac{1}{k}s_1(t) \\ \frac{1}{k}s_1(t) - \frac{1}{k}s_2(t) \end{bmatrix}\mathrm{d}t \tag{3.13}$$

with output

$$\mathrm{y}_M(\mathbf{x},t,\boldsymbol{\theta}) = \frac{1}{k}s_2(t) + w_{dw}(t), \tag{3.14}$$

where $w_{dw}(t)$ describes the diurnal variation of dry weather wastewater flow

$$w_{dw}(t,\boldsymbol{\theta}) = \sum_i^2 (\varsigma_i sin\frac{i2\pi t}{24} + \chi_i cos\frac{i2\pi t}{24}) \tag{3.15}$$

s_1 and s_2 correspond to the states of the system, i.e. the levels in the virtual storage tanks, and vary as a function of time (in hours). The vector $\boldsymbol{\theta}$ of physical model parameters includes the impervious catchment area A_{imp}, the mean dry weather flow at the catchment outlet a_0, the mean travel time (or reservoir residence time) k, and parameters ς_1, ς_2, χ_1, and χ_2. These last 4 variables describe the dry weather variation of the catchment outflow as a harmonic function. The vector $\mathbf{x}$ of model inputs includes the rainfall measurements averaged from the two pluviometers.

This simplified model disregards infiltration and does not include losses from sewer overflows. However, as a so-called "grey-box model", it captures the major processes with components that have a physical meaning. As such, its major advantage is that its equations are suitable to be incorporated into the IND framework (Appendix 3.B) and it is computationally fast enough to be applied in a forecast setting with data assimilation (Breinholt et al., 2012; Löwe et al., 2014a). Simple models have often proven useful and sufficient in off-line and on-line applications

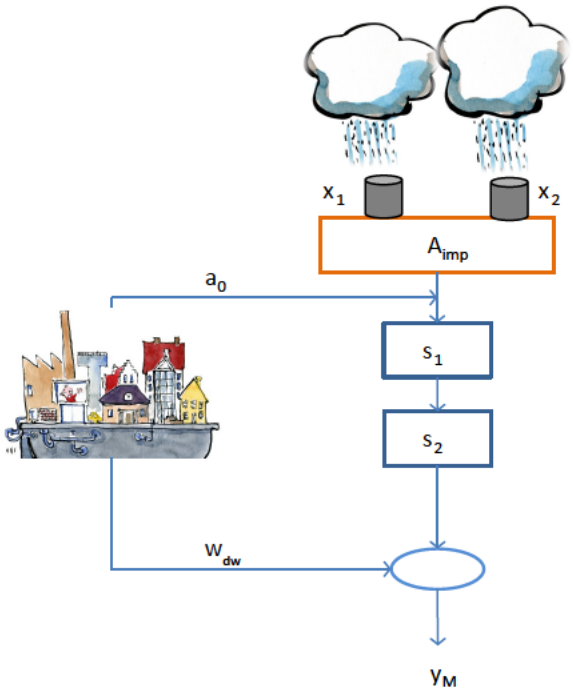

Figure 3: The linear reservoir cascade model considered for hydrological modeling. On the left the wastewater generation is illustrated, while on the right the rainfall-runoff process is shown. Symbols' description is given in Section 3.4.2. Drawings by F. Ahlefeldt.

(Coutu et al., 2012b; Wolfs et al., 2013; Mejía et al., 2014) and when modeling the integrated urban drainage system (Freni et al., 2009a).

3.4.3 Prior knowledge of model parameters

We selected prior distributions for the EBD based on the experience gained during previous studies in the same and similar catchments (Breinholt et al., 2012; Löwe et al., 2014a). Prior knowledge on simulator parameters was described by lognormal or normal distributions with a coefficient of variation of 0.2.

For the bias, we defined a probability density decreasing with increasing values of $\sigma_{B_{\mathrm{ct}}}$ and κ (here a truncated normal distribution) (Reichert and Schuwirth, 2012; Del Giudice et al., 2013). This helps to reduce the identifiability problem between the deterministic model and the bias term, and avoids model bias as much as possible. Regarding the correlation time of the bias, τ, we chose a prior value of 10h, close to 1/3 of the recession time of a consequential flood event not used for calibration.

For the IND approach, all parameters, except k, are defined in a logarithmic space to avoid negative values for the parameters. With respect to the standard deviation of the observation error σ_ϵ, we specified a prior as consistent as possible with the one of the bias description. Regarding the initial model states, we analytically calculated the filling of the reservoirs for no rain condition (see Supporting Information). The results obtained were similar to the system states in dry weather calculated in previous studies (Breinholt et al., 2011).

The parameters of the dry-weather-flow compartment were not inferred simultaneously with the other parameters due to numerical difficulties encountered in the IND routine. Instead,

we independently estimated them with a least squares method. For that, we selected data (not shown) from a period with no rain ranging from 07/18/2010 until 07/28/2010. The resulting dry-weather parameters were: $a_0 = 281.5\frac{m^3}{h}$, $\chi_1 = -47.4$, $\chi_2 = 21.3$, $s_1 = -43.4$, and $s_2 = -84.2$. The prior distributions of simulator and error model parameters are summarized in Table 2.

3.4.4 Computer implementation

The conceptual hydrological model and the EBD routine for uncertainty analysis were implemented in `R` (R Core Team, 2013). During inference (equation 3.10), we first obtained an optimal jump distribution and chain starting point by sequentially using the stochastic techniques described by Haario et al. (2001) and Vihola (2012), and then sampled from the target distribution by using a Metropolis-Hastings algorithm (Hastings, 1970). Finally, we approximated the predictive distribution of $\mathbf{Y}_o$ by propagating a posterior parameter sample through the simulator and the error model.
The IND routine was implemented in the open source software `CTSM` (Juhl et al., 2013), which is available as a package for `R`. Posterior maximization was performed using the PORT algorithm through the `R` function `nlminb` (Gay, 1990). To generate forecasts with the SDEs, we applied an Euler-Maruyama scheme (see, e.g., Kloeden and Platen (1999); Iacus (2008)), which involved 5000 realizations of the process $\mathbf{S}$.

3.5 Results

Predicting sewage flow with the EBD and IND approaches we found that: i) both methodologies provided forecast coverage of the validation data close to the nominal 95%; ii) reproducing the observations during heavy storm events (where the model has high discrepancies from data) was challenging for both methods. Even so, the uncertainty estimates of the two approaches dramatically outperformed those of a simplified approach using an iid error model (see Supporting Information).

3.5.1 Experiment 1: Parameter estimation

The data used for inference include two separate periods, as presented in Figure S1. The parameters inferred for the different modeling approaches are shown in Figure 4. The calibration with the IND was approximately two orders of magnitude faster than with the EBD. In the EBD the inference produced approximately bell-shaped marginals. The only distribution with a complex shape is that of d, which represents the time steps after which the rainfall influences runoff uncertainty. The posterior initial model states s_1 and s_2 remained close to their prior estimates and were similar for the EBD and IND. For the effective area A_{imp}, we observed bigger values of approximately 39 ha for the EBD approach, while the IND estimated an optimum of 33 ha. For the time constant k, approximately the same value was obtained with both frameworks (2.5 h). In both approaches, the inferred observation noise was considerably smaller than the bias or diffusion term (Figure 4). Due to the different ways of considering errors in the two methods, the other stochastic process parameters $\boldsymbol{\psi}$ cannot be compared directly.

3.5.2 Experiment 2: long-term forecasting

Long-term predictions for the two approaches were similar in terms of interquantile width and reliability (Figure 5). Credible intervals for IND predictions, however, were slightly wider than

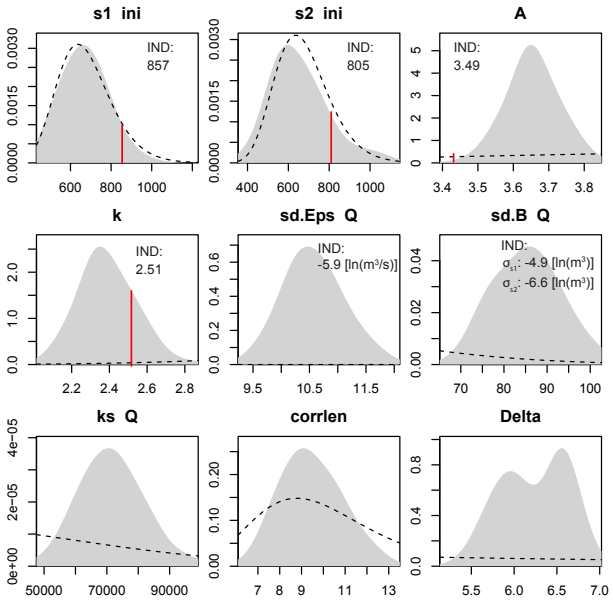

Figure 4: Prior (black, dashed) and posterior (gray area) marginal distributions from Bayesian inference in the EBD framework. The corresponding maximum a posteriori estimates from the IND framework are also displayed. Meaning and units of the parameters are given in Table 2.

those for the EBD and therefore covered the validation data better. Higher data coverage also resulted in a $\approx$ 50 % better average interval score $\mathsf{S}_{0.05}^{int}$ than for the EBD. The median of the probabilistic predictions was closer to the observations for the EBD than for the IND approach. The model calibrated with the EBD fitted validation peak discharge data better and obtained a better NS than the IND (32% higher). In general, with both error descriptions, the model consistently underestimated wet weather flows. This underprediction is confirmed by the QQ plot analysis (Figure S2). Here, the EBD-calibrated simulator performed slightly better than the IND. The latter had a $\mathsf{NB} \approx 40$ % larger and quantiles more distant from the 1:1 line. As expected, the EDB and IND outperformed the forecasts where model bias was neglected, both in terms of data coverage (i.e. reliability) and interval scores (Figure S4).

3.5.3 Experiment 3: short-term forecasting

As shown in Figure 6, the percentage of data points covered by the 95% credible interval of the short-term predictions was close to the nominal coverage. This means that the predictions were approximately reliable, although the underlying simulator appears to systematically deviate from reality. This is particularly interesting during the flood event on the right side of Figure 6, where the underlying model heavily underestimated the receding section of the hydrograph, yet the probabilistic predictions, after data assimilation, still encompassed most of the validation data. Indeed, with the simplified analysis that uses an iid error model, we obtained much poorer prediction intervals than with the two proposed methodologies (Figure S5).

During storm events, interval scores $\mathsf{S}_{0.05}^{int}$, which penalize too wide and unreliable uncertainty bands, were moderately higher (i.e. worse) for the EBD, especially in the decreasing limb of the flood hydrograph. Visual inspection shows that this is related to the slightly overconfident predictions of the EBD in this period. In contrast, during dry weather the EBD and IND produced similar predictions.

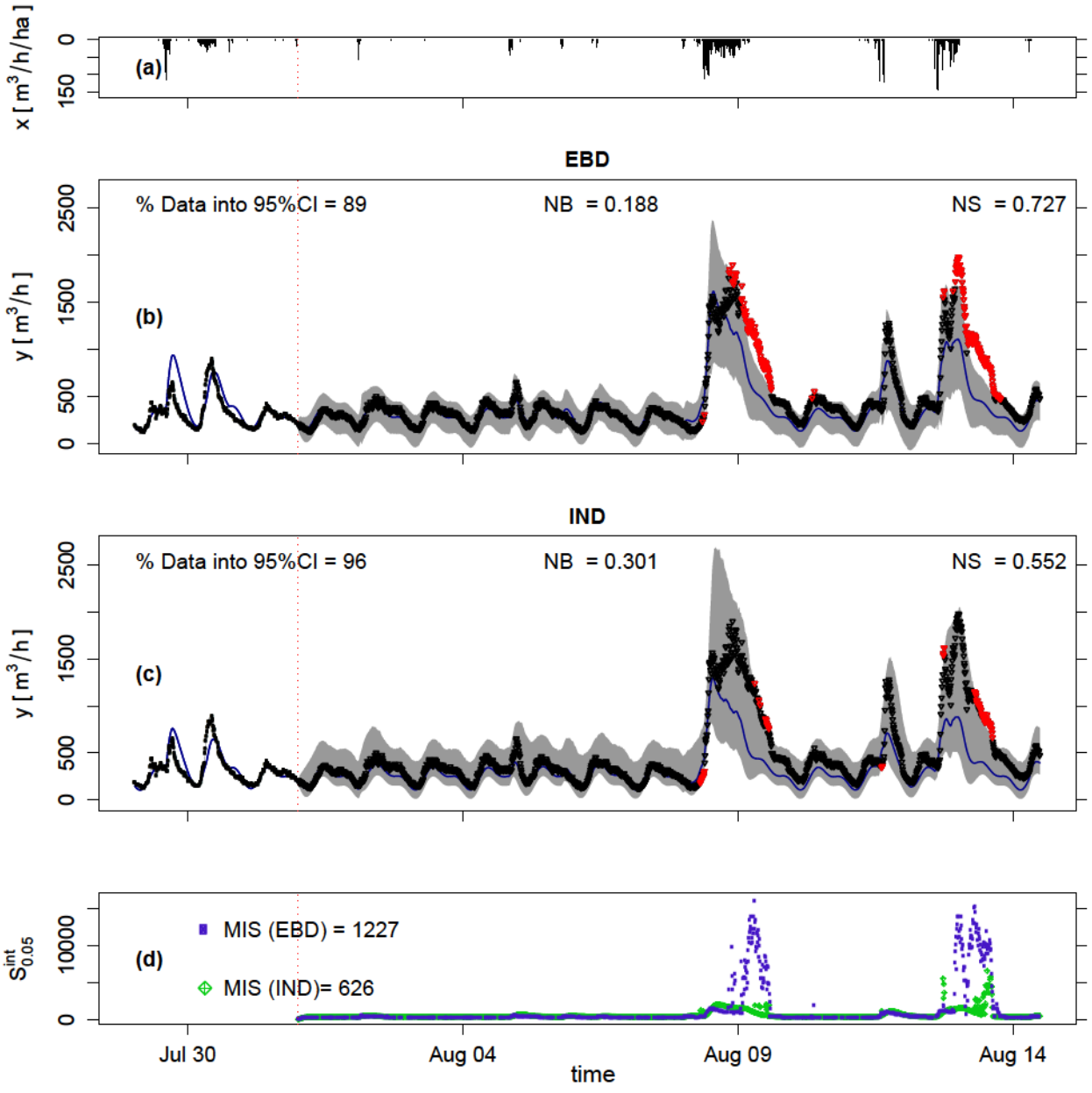

Figure 5: Smoothing (left) and long-term forecasting (right) results with the two methods. (a) Average rain intensity over the catchment (input data); (b) 95% credible intervals (gray) using the EBD approach, output data (dots, red when outside the intervals), median of the deterministic model (blue line); (c) 95% credible interval using the IND approach; (d) interval skill scores $S_{0.05}^{int}$ for the validation period together with its mean value MIS. The Nash-Sutcliffe coefficient (NS), the Normalized Bias index (NB), and the other performance indicators refer to the extrapolation period on the right of the dotted line.

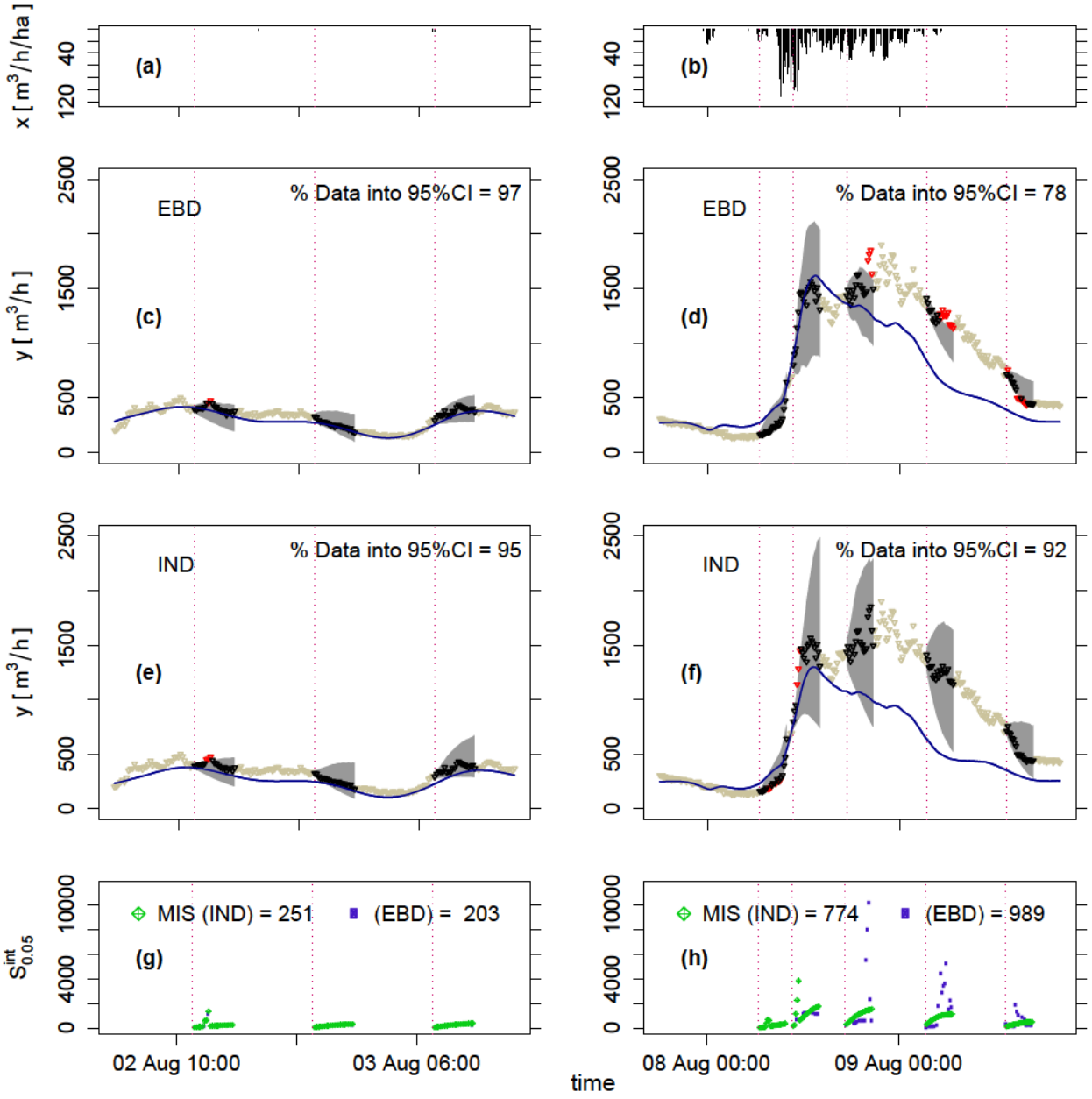

Figure 6: Short-term forecasts for illustrative points during a dry (a), (c), (e), (g) and a wet (b), (d), (f), (h) period. (a), (b) Rain intensity (input data); (c), (d) 20-step flow forecasts for the EBD approach (95% credible intervals (gray), output data (dots), median for the previously calibrated deterministic model (blue line); (e), (f) 20-step flow forecasts for the IND approach; (g), (h) interval skill scores $S_{0.05}^{int}$ for the different forecast horizons together with their mean value MIS.

3.6 Discussion

3.6.1 Prediction analysis

As shown in the case study application, both methodologies were able to provide both short-term and long-term reliable predictions. This is remarkable for two reasons. First, the underlying lumped reservoir model was a simplified representation of reality and therefore unable to consider all mechanisms occurring in the catchment (e.g. spatially varying soil water content, infiltration). Second, the validation conditions were consistently different from the calibration circumstances (more substantial peak discharges and infiltration-inflow). These considerations suggest that the methods are relatively robust against non-stationary inputs and boundary conditions, and structural errors of the model. Furthermore, both the EBD and IND could account for increased uncertainty during more dynamic wet periods, the first thanks to the input-dependence of the bias and the second due to the state-dependency of the noise. This is consistent with the conclusions of previous studies (Breinholt et al., 2012; Dietzel and Reichert, 2012; Honti et al., 2013; Del Giudice et al., 2015b). Furthermore, for both methods, conditioning on data generated generally reliable and precise short-term forecasts in all flow conditions, even when the calibrated simulator heavily deviated from the measurements (Figure 6).

Large deviations between model predictions and observations on long forecast horizons are mostly caused by the very simple model structure and system non-stationarities, but are also influenced by the error description. As discussed in Bayarri et al. (2007) and Del Giudice et al. (2013), the bias description might produce model performances which are slightly inferior to simplified approaches based on an iid error assumption. This can be explained by the fact that the inference with the EBD does not force the simulator to reproduce the observations with biased (i.e. over-tuned) parameters. The reverse, however, can also be true, and in this experiment the model fitted the data better with the bias than without it. Reduced model fit can be even more pronounced in the IND where parameter inference is performed in a one-step-ahead prediction setting. Breinholt et al. (2012) demonstrated a very satisfactory forecast performance of the approach on short horizons, which diminishes on longer horizons until becoming inferior to simplified approaches. In the present study, we also observe the highest forecast accuracy on the shortest horizons (Figure 6).

Parameter estimation in the IND relies on the assumption of normality and independence of the one-step-ahead prediction errors (innovations) and of Gaussianity of the transformed system states. By inspecting the innovations (Figures S13-S15), this assumption appears to be valid in our study.

In agreement with previous studies (Honti et al., 2013; Del Giudice et al., 2013; Breinholt et al., 2012), we generally found that both the EBD and the IND (Figures 5, 6) produced much less overconfident and therefore more reliable uncertainty bands than simplified approaches (Figures S4, S5).

3.6.2 Commonalities and differences of the methods

Theoretical considerations

The main difference between the two approaches considered is that the IND describes model inadequacies as part of the model states, while the EBD adds them to the model output. In other words, the IND propagates the input and structural errors identified during calibration through the model, while the EBD treats the model as a “perfect” black box to which these

errors are added. In addition, the EBD was developed with a focus on statistical inference and long-term prediction, while on-line applications were the focus for the IND. This background defines how the methods were implemented and what advantages and disadvantages they have. Flexible model structures for describing the time-dependent behavior of systematic errors can be implemented in both approaches. Input- and output-dependence of systematic errors can be considered in the EBD (Del Giudice et al., 2013). In the IND, state- and input-dependent diffusion terms can be implemented, but only the former were documented in previous applications (Breinholt et al., 2012; Löwe et al., 2014a), while the latter is the subject of ongoing research.

Practical aspects

The most suitable error characterization needs to be identified depending on the specific case study with both approaches. Adding linear state-dependent noise in the model equations, as in the IND, has the advantage that it guarantees positive values of the model output. When modeling the errors in the output equations, as in the EBD, output transformation might be required to ensure non-negative predictions (e.g. in Frey et al. (2011) and Sikorska et al. (2012b)).
On the one hand, the implementation of the noise term as part of the model states in the IND seems intuitively more appropriate, because the systematic error description becomes a part of the model and the noise is routed through the model. In combination with data assimilation routines, the IND also allows for the identification of hidden states from data, which is a useful feature in process monitoring and system control, for example. On the other hand, the solution of stochastic differential equations is more complex than that of ordinary differential equations and this limits how complex the model can be.
The IND, being an "intrusive" method, cannot easily be applied to existing hydrological software packages such as SWMM. Instead, this is easily done with the "non-intrusive" EBD, on condition that the model is fast enough to be applied in MCMC.
Parameter inference in the two approaches is largely driven by their focus areas, and that applies to both conceptual formulation and the numerical techniques. The EBD applies a Bayesian approach using MCMC which is slow but allows for the identification of the whole distribution of the parameters. The IND commonly applies Maximum a Posteriori (or Likelihood) estimation for parameter inference. Currently, only the mode of the parameter distribution is considered and parametric uncertainty is neglected during forecasting. This approach is computationally very efficient and identifies model parameters which are optimal for on-line predictions.
An updating of the model states is readily implemented in the IND framework, but it leads to a violation of the water balance (see e.g., Salamon and Feyen (2010); Reichert and Mieleitner (2009)). It is therefore not particularly suitable for design studies, while it can be very useful in on-line applications where only the correspondence between forecasted and observed output is of interest.

3.7 Conclusions

In this study we, for the first time, compared and discussed two probabilistic techniques to reliably quantify predictive uncertainty in rainfall-runoff modeling in urban catchments. The first approach was an external bias description (EBD), representing model discrepancies in the output space. The second was an internal noise description (IND), considering model inadequacies in the system equations. Based on theoretical considerations and the results of the case study,

we conclude that:

Q1. Both approaches describe systematic model errors in a way suitable for hydrological modeling. Both can produce reliable forecasts in the short-term, which is useful, e.g., for real-time model predictive control of sewer networks and wastewater treatment plants, as well as for long-term analyses. As demonstrated in our case study, this seems to be the case even for very simple rainfall-runoff models applied to a complex sewer system with non-stationary behavior.

Q2. Both methods also have some limitations. First, although they explicitly account for the effects of model inadequacies, neither of them provides comprehensive information on underlying causes of bias. The IND, through an analysis of model states, can, however, give some hint on which model compartment is most uncertain. The EBD can be rather demanding on a computational level during parameter inference because it requires tens of thousands of MCMC simulations. Furthermore, it does not provide a data assimilation routine in its current implementation. In contrast to the IND, the EBD can readily be applied to any existing engineering software. Additionally, in its current implementation, the IND makes simplifying assumptions on the distribution of the states and outputs. These guarantee a very high computational efficiency but need to be tested via residual analysis.

Q3. Although both techniques generally outperform those that do not account for systematic model errors, especially in quantifying predictive uncertainties, each has its optimal field of application. The EBD is usually able to provide accurate and precise long-term forecasts with various kind of models, provided that the model reasonably describes the system studied. The IND, on the other hand, is especially suitable for short-term forecasts where new output measurements are continuously available for updating. Additionally, it appears able to provide reliable prediction even in cases where the underlying model is highly simplified. Finally, it allows for the identification of hidden model states, which is useful to identify the behavior of a variable when only indirect measurements are available.

Q4. Expected developments of the EBD involve the investigation of the reasons for bias. Current research in the IND is focusing on reducing the likelihood approximations and producing an ensemble-based version that would make it applicable to existing models.

Acknowledgements

The data and codes used are available upon request from the first author (Dario.DelGiudice@eawag.ch). The authors are grateful to Carlo Albert for the interesting discussions about inference in state space modeling and his useful comments on the manuscript. Fabrizio Fenicia, Anna Sikorska, Wolfgang Nowak, Nataliya Le Vine, and Eberhard Morgenroth are also acknowledged for their feedback. This work was partially supported by the Swiss National Science Foundation (grant No. CR2212_135551) and by the Danish Council for Strategic Research (DSF) as part of the Storm- and Wastewater Informatics (SWI) project.

3.A Equations for state updating with the IND using the EKF

Posterior maximization with the IND likelihood (equation 3.11) adopts an extended Kalman filter (EKF). The filtering procedure is briefly synthesized from Kristensen et al. (2004) and

Kao et al. (2004). For each candidate parameter set $(\boldsymbol{\theta}_c, \boldsymbol{\psi}_c)$ generated during optimization, the innovations $\mathrm{y}_{o_i} - \mathrm{E}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}_c, \boldsymbol{\psi}_c)$ and their covariances $\boldsymbol{\Sigma}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}, \boldsymbol{\theta}_c, \boldsymbol{\psi}_c)$ are continuously updated following this assimilation scheme:

Step i:. Project the state ahead for the next timestep i solving the state prediction equation representing the deterministic model:

$$\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} = f_{\mathrm{M}}(\mathbf{S}, \mathbf{x}, t, \boldsymbol{\theta}). \tag{1.1}$$

for time interval $[t_{i-1}, t_i[$. The so-obtained state $\mathbf{S}_{i|i-1}$ is used to predict the (a priori) output at time i:

$$\mathrm{E}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}) = h(\mathbf{S}_{i|i-1}, \mathbf{x}_i, \boldsymbol{\theta}_c, \boldsymbol{\psi}_c, t_i). \tag{1.2}$$

Step ii:. Project the (a priori) error-covariance matrix ahead:

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = \mathbf{M}\mathbf{P} + \mathbf{P}\mathbf{M}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\sigma}} \tag{1.3}$$

where the resulting covariance matrix is defined as $\mathbf{P}_{i|i-1} \equiv \mathrm{E}\left[(\mathbf{S}_{i|i-1} - \mathbf{S}_*)(\mathbf{S}_{i|i-1} - \mathbf{S}_{*,i})^{\mathrm{T}}\right]$ with $\mathbf{S}_{*,i}$ representing the true state. In equation 1.3, $\mathbf{M}$ is the Jacobian matrix of the deterministic model f_{M}, and $\boldsymbol{\Sigma}_{\boldsymbol{\sigma}}$ is the estimated system noise covariance for the prediction of $\mathbf{P}$.

Step iii:. When the next output measurement y_{o_i} becomes available (or assimilable) the states are updated (or corrected):

$$\mathbf{S}_{i|i} = \mathbf{S}_{i|i-1} + \mathbf{K}_i\big(\mathrm{y}_{o_i} - \mathrm{E}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}})\big) \tag{1.4}$$

where $\mathbf{K}_i$ is the Kalman gain defined as $\mathbf{K}_i \equiv \mathbf{P}_{i|i-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}})$, with $\mathbf{H}$ being the Jacobian matrix of the stochastic model h, and $\boldsymbol{\Sigma}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}}) \equiv \mathbf{H}\mathbf{P}_{i|i-1}\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Sigma}_E$ being the innovation covariance matrix.

Step iv:. Finally, the updated (a posteriori) error-covariance matrix is computed as:

$$\mathbf{P}_{i|i} = \mathbf{P}_{i|i-1} - \mathbf{K}_i\boldsymbol{\Sigma}(\mathrm{y}_{o_i}|\mathrm{y}_{o_{i-1}})\mathbf{K}_i^{\mathrm{T}} \tag{1.5}$$

This procedure of sequential state update is repeated for every timestep i of the calibration period.

3.B Specific model equations with the IND

Combining the simulator equations (equation 3.13 - equation 3.15) with the state noise (equation 3.8) and the Lamperti transformation we obtain the following state-space description of the system studied:

$$\mathrm{d}\begin{bmatrix} ln(s_1(t)) \\ ln(s_2(t)) \end{bmatrix} = \begin{bmatrix} (exp(ln(A_{imp})) \cdot x(t) + a_0) \cdot exp(-ln(s_1(t))) - \frac{1}{k} - \frac{1}{2}\sigma_{s_1}^2 \\ \left(\frac{1}{k} exp(ln(s_1(t)))\right) \cdot exp(-ln(s_2(t))) - \frac{1}{k} - \frac{1}{2}\sigma_{s_1}^2 \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_{s_1} & 0 \\ 0 & \sigma_{s_2} \end{bmatrix} \mathrm{d}W_t, \tag{2.6}$$

$$ln(Y_0) = ln(\frac{1}{k} s_2(t) + df(t)) + E. \tag{2.7}$$

3.C Short-term forecasts with the EBD

In its current implementation, the on-line predictions with the bias correction are calculated by following these steps:

Step i:. Select the current time point j (e.g. the last element of a data time series) and its corresponding output observation y_{o_j}.

Step ii:. Condition the Gauss-Markov process on y_{o_j}. This involves computing the mean and variance of the bias according to equation 27 and 28 of Reichert and Schuwirth (2012), which in turn requires calculating $\mathbf{\Sigma_E}(\boldsymbol{\psi}_{post})$ and $\mathbf{\Sigma}_{\mathbf{B}_M}(\boldsymbol{\psi}_{post})$ according to equation 3 and 10 of Del Giudice et al. (2013).

Step iii:. Draw $\sim 10^3$ samples of the bias process in this past period.

Step iv:. Use each last element (i.e. the one at time j) of the bias sample as starting point for simulating trajectories of $\mathbf{B}_\mathrm{M}$ over the desired number of time steps in the future. These realizations are based on equations 21 and 22 of Del Giudice et al. (2013).

Step v:. As in equation 2.1, add to the bias realizations sample paths of the white noise (see equation 3.5) and an equal number of runs of the model $\mathbf{y}_\mathrm{M}(\boldsymbol{\Theta}_{post})$.

Step vi:. Finally, produce the desired sample quantiles $\mathbf{y}_\mathrm{M} + \mathbf{B}_\mathrm{M} + \mathbf{E}$ to plot the total uncertainty bands, usually corresponding to the region between the 95% credible intervals.

Step vii:. Repeat for each time j of interest.

Chapter 4

Model bias and complexity - understanding the effects of structural deficits and input errors on runoff predictions

D. Del Giudice [a,b], P. Reichert[a,b], V. Bareš[c], C. Albert[a], J. Rieckermann[a].

[a]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b]ETHZ: Swiss Federal Institute of Technology Zürich, 8093 Zürich, Switzerland
[c]Czech Technical University in Prague, 166 29 Prague, Czech Republic

Environmental Modelling and Software (2015), 64, 205 - 214, doi:10.1016/j.envsoft.2014.11.006.

Author contributions

D.D.G. designed the experiments, built the models, performed the analyses, wrote the paper; V.B. collected the data and built the models; D.D.G. and P.R. conceived the experiments; all coauthors gave advices, supported result interpretation and paper revision.

Abstract

Oversimplified models and erroneous inputs play a significant role in impairing environmental predictions. To assess the contribution of these errors to model uncertainties is still challenging. Our objective is to understand the effect of model complexity on systematic modeling errors. Our method consists of formulating alternative models with increasing detail and flexibility and describing their systematic deviations by an autoregressive bias process. We test the approach in an urban catchment with five drainage models. Our results show that a single bias description produces reliable predictions for all models. The bias decreases with increasing model complexity and then stabilizes. The bias decline can be associated with reduced structural deficits, while the remaining bias is probably dominated by input errors. Combining a bias description with a multimodel comparison is an effective way to assess the influence of structural and rainfall errors on flow forecasts.

4.1 Introduction

Models are important to predict future behavior of environmental systems. They help, for instance, decision makers to choose among different management policies by enabling them to compare the consequences of several alternatives with the status quo.
Environmental models used in natural and urban hydrology can be anything between extremely simple or overly complex. Among the simplest continuous-time models are lumped reservoir models (e.g. Breinholt et al., 2012; Sikorska et al., 2012b). These parsimonious simulators (i.e. deterministic models) usually can be very rapidly calibrated, have easily identifiable parameters and can reproduce well the hydrologic response of a simple system (Coutu et al., 2012b). However, they cannot describe the effects of complex flow processes. This is the task of physically-based and spatially-distributed simulators, which can model runoff characteristics in different points of the drainage networks and even predict surface inundation processes (Butts et al., 2004; Leitao et al., 2010; Butler and Davies, 2010).

Simulators which are too simple or too complex are both affected by uncertainties. The first category is principally affected by model structural deficits, while the second is strongly influenced by parametric uncertainty (Reichert, 2012; Jackson et al., 2010).
In hydrology, model structural deficits arise from process misspecifications (e.g., neglecting infiltration or oversimplifying the functioning of hydraulic structures), insufficient spatial resolution (e.g., modeling the whole heterogeneous catchment as one storage), oversimplified empirical equations (e.g., assuming a linear response of the catchment or neglecting back-water effects), numerical errors (e.g., using a poor solver), etc. (Gupta et al., 2012). Parametric uncertainty, expressing the incomplete knowledge of the "correct" parameter values, arises from theoretical and practical non-identifiability. The first type of non-identifiability is linked to a model structure capable of producing the same output with different parameters, while the second is due to insufficient or imprecise output observations (McLean and McAuley, 2012).
Besides (in)adequate system description and parameter identifiability, another predominant contributor to the uncertainty of hydrologic predictions is input uncertainty. This is associated with the use of inaccurate data to force the simulator due to inappropriate sampling of the rainfall field and/or measurement errors (McMillan et al., 2011).

There are at least three main motivations to explicitly assess the effects of structural deficits and input inaccuracies in urban drainage, natural hydrology, and in environmental modeling in general. First, these error sources can be the dominant cause of predictive uncertainty (Renard et al., 2011). Second, neglecting structural and input errors leads to autocorrelated residuals and therefore biased parameter estimates, as well as underestimation of uncertainty (Neumann and Gujer, 2008; Sikorska et al., 2012b). Third, we are interested in understanding how much a better model structure can improve the precision and accuracy of our predictions.

To our knowledge, very few studies have rigorously tried to quantify the combined results of structural deficits and input errors on urban flow simulations. Indeed, the effect of structural deficits on modeling errors and predictions has widely been neglected by the urban drainage community, apparently due to the lack of simple techniques to deal with it (Dotto et al., 2011). Instead, parametric uncertainty appears to be the only type of model uncertainty to have been analyzed (e.g., Freni et al., 2009b). Few recent exceptions have been the investigations of Sun and Bertrand-Krajewski (2013), Sikorska et al. (2012b), and Del Giudice et al. (2013). While the first study estimated rainfall errors but assumed no structural deficits, the others implicitly considered the combined impact of structural and input inadequacies by using autoregressive output error models. These error formulations can indeed account for the systematic deviations of model results from output data, the so-called model bias (or discrepancies). Theses studies, however, have not focused on the individual contribution of oversimplified process description on predictive uncertainty. Understanding the reasons for systematic errors is essential to assess, for example, if model prediction will most benefit from a more adequate model formulation or more representative input information.

Recently, a few statistical approaches have appeared in the hydrological literature to explicitly assess, and in some cases reduce, the effects of structural deficits on model residuals (viz. structural calibration errors) and on probabilistic predictions (viz. structural uncertainty). Among the three most promising and rigorous methodologies which also account for imprecise input information are: those using stochastic, time-dependent parameters (Reichert and Mieleitner, 2009; Renard et al., 2010; Lin and Beck, 2012), structural multipliers (Salamon and Feyen, 2010), and Bayesian data assimilation (Bulygina and Gupta, 2009). The drawback of these frameworks is that they are conceptually and practically demanding. Such complexity, although necessary to optimally quantify the time-dependent propagation of structural errors, can hinder the application of these techniques, especially when dealing with computationally expensive models (Dietzel and Reichert, 2012).

In this paper we therefore propose a statistical alternative to quantify the influence of structural errors on environmental predictions, with a particular focus on hydrology. We simultaneously consider the impact of rainfall errors which are intricately entangled with that of inappropriate model structure. Our formal methodology combines the strengths of model comparison (Butts et al., 2004; Zhang et al., 2011; Jackson et al., 2010) with those of a Bayesian description of the systematic model deviations from data (Craig et al., 2001; Kennedy and O'Hagan, 2001; Higdon et al., 2005; Bayarri et al., 2007). In recent studies, we have shown how this relatively simple technique can efficiently account for model structure and input errors. In this way it produces reliable flow predictions which also distinguish uncertainty due to input plus structural errors,

parameters, and measurement noise (Reichert and Schuwirth, 2012; Dietzel and Reichert, 2012; Honti et al., 2013; Del Giudice et al., 2013).
Our objective here is to use a Bayesian description of output bias to discriminate as best as possible the effects of structural deficits from those of errors in input (e.g. rainfall) measurements. To reach this goal, we quantify the combined effect of structural and input errors over a class of alternative deterministic models calibrated on the same flow data. In particular, we investigate how a stochastic process representing model inadequacies behaves as a function of increasing simulator complexity. This analysis, together with our mechanistic understanding of the system, will facilitate the acquisition of knowledge about the bias and its interpretation. This is relevant because the bias description can be a useful tool for modelers to produce more reliable predictive intervals and reduce overtuning of calibration parameters (Bayarri et al., 2007).

4.2 Methodology

We briefly clarify our terminology of structure-related problems and review the Bayesian description of model bias used to quantify the effects of structural deficits and input errors. We then show how to use this statistical technique of modeling output errors in combination with model comparison to associate the bias to the effects of structural deficits and input errors.

4.2.1 Definition of structural deficits

We define "model structural deficits" or errors as the inadequate selection of model variables and processes, inadequate process formulation and the inadequate choice of the spatial and temporal resolution of the model. These deficits can also comprise suboptimal computer implementation, including poor numerical approximations, and software bugs. In essence, all erroneous oversimplifications of the actual system contribute to structural deficits. Model structural deficits are analogous to what Gupta et al. (2012) call "model structural inadequacy" and to "model structure uncertainty" plus "model technical uncertainty" in Refsgaard et al. (2007).

4.2.2 Brief description of inference and predictions with bias

Model bias is defined as the systematic deviation of simulator results from observed outputs. This phenomenon is also known as model inadequacy or discrepancy and affects most environmental models. To describe model discrepancy with an additive stochastic term is one of the simplest yet formal ways to account for output bias and so derive reliable probabilistic predictions. In the following points we summarize the core concepts of the bias description for calibration and uncertainty analysis. For a mathematical derivation the reader is referred to Craig et al. (2001); Kennedy and O'Hagan (2001); Higdon et al. (2005); Bayarri et al. (2007); Reichert and Schuwirth (2012). Details specific to hydrology are described in Honti et al. (2013) and Del Giudice et al. (2013).

Step i:. **System's output representation**: in the presence of bias, the observations $\mathbf{Y}_o$ of the system response (e.g., runoff at the outlet) can be modeled as a sum of a deterministic model output (possibly multivariate) $\mathbf{y}_M$ plus a Gaussian process $\mathbf{B}_M(\mathbf{x}, \boldsymbol{\psi})$ (an autocorrelated bias correction) and a Gaussian white noise $\mathbf{E}(\boldsymbol{\psi})$:

$$\tilde{\mathbf{Y}}_o(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \tilde{\mathbf{y}}_M(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{B}_M(\mathbf{x}, \boldsymbol{\psi}) + \mathbf{E}(\boldsymbol{\psi}) \quad . \tag{2.1}$$

Here, random variables are represented in capitals, whereas those in lowercase are deterministic. While $\mathbf{B}_M(\boldsymbol{\psi})$ mimics the combined effect of input errors, structural deficits and (possibly) measurement biases, $\mathbf{E}(\boldsymbol{\psi})$ represents the random errors of the output measurements. The equation describing the evolution of the bias process is given in Appendix B. The tilde indicates transformed quantities, i.e. $\tilde{\mathbf{y}} = g(\mathbf{y})$. Output transformation is used to improve the realism of the error model (Sect. 4.2.3). The input of the system, e.g. rainfall, is given by $\mathbf{x}$. The variables $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are the simulator and error model parameters, respectively. These parameters have prior distributions which formulate (incomplete) knowledge about their value as probabilities. These distributions are updated during inference. We adopt a Bayesian framework to overcome identifiability problems between the parameters of the models and those of the autocorrelated bias process. Indeed, the assignment of appropriate priors supports the distinction between model and bias and facilitates the inference of sensible parameters.

Step ii:. **Parameter inference**: from the system representation in Eq. 2.1, a hierarchical probability model for the observation distribution, with $\mathbf{B}_M(\boldsymbol{\psi})$ and $\mathbf{E}(\boldsymbol{\psi})$ at the intermediate level, can be formulated. The likelihood function $f\left(\mathbf{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{x}\right)$ is here normally distributed in the $g-$transformed space (see Appendix C). By combining prior knowledge about the parameters with the likelihood function evaluated at the actual observations, we can derive the joint posterior distribution of parameters $(\boldsymbol{\theta}, \boldsymbol{\psi})$. This process is also called Bayesian calibration, parameter estimation, updating, or inverse modeling.

Step iii:. **Predictions in the calibration domain**: To analyze how our model was deviating from the true system response, we can predict the system response given the actual calibration data, $g^{-1}(\tilde{\mathbf{y}}_M^c + \mathbf{B}_M^c|\mathbf{y}_o^c)$. This bias-corrected model output is our best approximation of true output in the calibration phase, denoted with c. The act of predicting "in the past" is also referred to as smoothing (Bulygina and Gupta, 2009). Typically, the predictions of the system response in this period have low uncertainty and closely follow the measured output. This is due to the conditioning of the bias on calibration data.

Step iv:. **Predictions in the extrapolation domain**: To forecast system behavior for a period, denoted with e, without output data, we extrapolate our mechanistic and posterior knowledge about it. If data are available but not used for calibration, they can be used for conditional validation. This phase involves a propagation of $\boldsymbol{\Theta}_{post}$, the random variable following the posterior distribution of simulator parameters, and $\boldsymbol{\Psi}_{post}$, the random variable following the posterior distribution of error model parameters. This enables us to calculate the distributions of the model results $\mathbf{y}_M^e(\boldsymbol{\Theta}_{post})$, the true system output $g^{-1}(\tilde{\mathbf{y}}_M^e(\boldsymbol{\Theta}_{post}) + \mathbf{B}_M^e(\boldsymbol{\Psi}_{post})|\mathbf{y}_o^c)$ and the true output with observation error $g^{-1}(\tilde{\mathbf{y}}_M^e(\boldsymbol{\Theta}_{post}) + \mathbf{B}_M^e(\boldsymbol{\Psi}_{post}) + \mathbf{E}(\boldsymbol{\Psi}_{post})|\mathbf{y}_o^c)$. When the extrapolation domain is not adjacent to the calibration period, as in this study, the conditioning of the initial condition on $\mathbf{y}_o^c$ becomes negligible (see discussions on correlation length in Del Giudice et al. (2013) and Reichert and Schuwirth (2012)).

4.2.3 How to connect structural errors with the bias of the alternative models

The idea of analyzing the performances of alternative model structures for the same case study is known in hydrology (Butts et al., 2004; Schoups et al., 2008; Fenicia et al., 2013). These

studies have demonstrated that multimodel comparison is an effective way to detect model structural deficits and to select the optimal model parameterization for the purpose of the investigation. Our approach combines the strengths of analyzing multiple simulator variants and those of statistically describing model bias.

We argue that the changes of the bias as a function of the alternative model complexities can be used to quantify the effects of structural deficits. By analyzing the reduction of the standard deviation of the bias, σ_B, after inference with the alternative structures we have quantitative, although indirect, information about the reduction of structural calibration errors. By comparing the width of the predictive distributions for our best estimates of the true system response, $g^{-1}(\tilde{\mathbf{y}}_M + \mathbf{B}_M)$, we can quantify the reduction of structural and input uncertainty due to a decline of structural uncertainty.

Data transformation $g()$

In order to account for the heteroschedasticity of the uncertainty of model predictions we transform the modeled and observed response of the system (Yang et al., 2007b; Frey et al., 2011; Breinholt et al., 2012; Dietzel and Reichert, 2012). As these and several other studies show, output transformation can stabilize the variance of the calibration residuals while accounting for the increase of uncertainty with higher values of the predictand (e.g., during high flow situations). Several strategies are possible to make the error model heteroscedastic (see e.g., Del Giudice et al., 2013). We here selected for $g()$ a two-parameter Box-Cox transformation (Box and Cox, 1964) to ensure i) an appropriate consideration of residual heteroscedasticity and ii) the same error parameterization for all model structures. Output transformation also has the following effect: while in the transformed space the error terms $\mathbf{B}_M$ and $\mathbf{E}$ in Eq. 1 are normal and homoscedastic, the inverse transformation makes the predictive distributions $g^{-1}(\tilde{\mathbf{y}}_M + \mathbf{B}_M)$ and $g^{-1}(\tilde{\mathbf{y}}_M + \mathbf{B}_M + \mathbf{E})$ asymmetric and with a spread increasing with the output. The Box-Cox transformation functions are given in the Appendix. We here kept the transformation parameters fixed during calibration.

Link to multi-objective calibration

As recently presented by Reichert and Schuwirth (2012), we connected the Bayesian bias formulation with multi-objective calibration. Multi-objective calibration can refer to several interrelated concepts: a compromise in optimally adjusting to different sections of an output time series, different objective functions, or different output variables (see Spaaks and Bouten (2013) and references therein). In our application, multi-objective calibration refers to the simultaneous use of 2 prediction variables to infer model parameters. Using multiple outputs per timepoint can not only provide further insights into the reasons for model structural errors, but also potentially increases the amount of information available to identify calibration parameters. The multi-objective approach is incorporated in the bias description by setting different (hyper)priors of the amount of model inadequacy for the different outputs. This expresses how much bias the analyst is willing to accept in each predictand (Dietzel and Reichert, 2012). The choice of how to weight the different outputs remains subjective. This cannot be avoided in the presence of model bias. With the suggested methodology, however, the assumptions behind the weights at least become transparent.

Evaluating model performance

To assess the performance of each model structure, we used several metrics along with a visual inspection of the predictions.

In the calibration phase we analyzed two statistics connected to the (mis)fit of model results to data. First, we observed σ_B, the posterior standard deviation of the Gauss-Markov process representing model discrepancy. This indicates the amount of bias identified for each model structure. Second, we computed the Nash-Sutcliffe efficiency, NS (Nash and Sutcliffe, 1970). This coefficient is often used in hydrology to evaluate the match of the deterministic model to the output data.

In the extrapolation (or validation) phase, instead, we focused on how the calibration bias translates to future forecast and how predictive uncertainty changes with different model parameterizations. Besides monitoring again the NS, we also computed Cover[%], the percentage of validation data included within the 95 % credibility intervals of $g^{-1}(\tilde{\mathbf{y}}^e_M(\boldsymbol{\Theta}_{post}) + \mathbf{B}^e_M + \mathbf{E}^e)$. As suggested by recent statistical hydrological studies (see Sikorska et al. (2013), Breinholt et al. (2012) and references therein), when Cover[%] was approximately equal to or larger than 95, we considered the predictions to be reliable. Given the autocorrelation of $\mathbf{B}_M$, this statement holds only for validation windows which are substantially larger than τ, the correlation length of the bias term.

The third statistic for predictive performance analysis was $\langle \gamma_{y+B} \rangle$. This value represents the average band width of the 95 % interquantile range, IQR_{95}, of the predictive distributions of $g^{-1}(\tilde{\mathbf{y}}^e_M + \mathbf{B}^e_M \mid \boldsymbol{\theta}, \boldsymbol{\psi})$ as a function of the parameters. The symbol γ_{y+B} denotes the time-dependent IQR_{95} of the mentioned distribution. The operator $\langle(\cdot)\rangle$ denotes averaging with respect to both time and posterior parameter distribution. The formula to compute this index is:

$$\begin{aligned}\langle \gamma_{y+B} \rangle \equiv \left\langle \gamma_{\mathbf{y}_M(\boldsymbol{\theta})+\mathbf{B}_M(\boldsymbol{\psi})} \right\rangle_{\mathbf{t},\boldsymbol{\Theta}_{post},\boldsymbol{\Psi}_{post}} &= \left\langle \mathrm{IQR}_{95}\left(g^{-1}(\tilde{y}_{M_t} + B_{M_t} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \right) \right\rangle_{\mathbf{t},\boldsymbol{\Theta}_{post},\boldsymbol{\Psi}_{post}} \\ &\approx \frac{1}{n}\sum_{t=1}^{n}\left[\frac{1}{m}\sum_{k=1}^{m}\left[g^{-1}\Big(\tilde{y}_{M_t,k} + 1.96\sigma_{B,k}\Big) - g^{-1}\Big(\tilde{y}_{M_t,k} - 1.96\sigma_{B,k}\Big) \right] \right] \end{aligned} \quad (2.2)$$

where $\mathbf{t}$ is the vector of n timepoints constituting the extrapolation period e, $\frac{1}{m}\sum_{k=1}^{m}$ computes the expected value over the posterior parameter sample of length m, and $\frac{1}{n}\sum_{t=1}^{n}$ computes the expected value over e. For each model parameterization, $\langle \gamma_{y+B} \rangle$ was calculated as follows:

i) Propagate a large posterior sample ($m \sim 10^3$ elements) of $\boldsymbol{\Theta}_{post}$ through the simulator forced with the input (e.g. precipitation) of the extrapolation period, $\mathbf{x}^e$, ii) transform each realization of the deterministic model and add $\pm 1.96\sigma_B$ to obtain a large sample of the 2.5% and 97.5% quantiles of $\tilde{\mathbf{y}}^e_M + \mathbf{B}^e_M$, iii) transform the quantiles back to the real space, iv) compute for each point in time an average value of the two quantiles over the length of the posterior sample, v) calculate the average interquantile distance over time for each output.

Finally, we computed the $\langle \gamma_y \rangle$ of each model alternative. This score, representing the average band width of the 95 % credible intervals of $\mathbf{y}^e_M(\boldsymbol{\Theta}_{post})$, indicates the effect of parameter

uncertainty on model output. The formula to compute this index is:

$$\langle \gamma_y \rangle \equiv \left\langle \gamma_{\mathbf{y}_M(\mathbf{\Theta}_{post})} \right\rangle_{\mathbf{t}} = \left\langle \mathrm{IQR}_{95}\big(y_{M_t}(\mathbf{\Theta}_{post})\big) \right\rangle_{\mathbf{t}} \approx \frac{1}{n} \sum_{t=1}^{n} \mathrm{IQR}_{95}\big(y_{M_t}(\mathbf{\Theta}_{post})\big) \tag{2.3}$$

A summary of the chosen statistics of model error magnitude and predictive powers is given in Table 1.

Index	Description	Interpretation	Usage
σ_B	Posterior distribution of the s.d. of the bias for each output	Magnitude of the systematic deviations	Calibration
NS	Nash-Sutcliffe accuracy coefficient	Model's ability to predict the data	Calibration Extrapolation
Cover[%]	Pct. of data within the 95% total uncertainty intervals	if $\geq 95\%$, predictions are reliable	Extrapolation
$\langle \gamma_{y+B} \rangle$	IQR_{95} of the true system output with no parameter uncertainty	Effect of input and structural errors	Extrapolation
$\langle \gamma_y \rangle$	IQR_{95} of the model output	Effect of parameter uncertainty	Extrapolation

Table 1: Statistics to evaluate model performances during calibration and predictions.

4.3 Material

We here present the specific experimental setup used to test our methodology's ability to quantify the effects of structural deficits. This section includes the description of the urban watershed studied with its measured time series, the hydrodynamic models used to simulate the system behavior, the error model configuration, and information on the computer implementation.

4.3.1 Case study and data

The area we studied was a small stormwater network in Hostivice, in proximity to Prague (CZ). The catchment has a surface of 11.2 ha, which is mainly occupied by streets and medium density houses surrounded by gardens (Fig. 1). The dominant process contributing to the stormwater discharge is precipitation-induced runoff from the impervious surfaces. Groundwater infiltration into the sewer pipes is a minor phenomenon.

The precipitation and stormwater data of the catchment have been collected in a dedicated monitoring campaign (Bareš et al., 2010) already used in a previous study on uncertainty quantification (Del Giudice et al., 2013). The watershed was instrumented with two tipping bucket pluviometers (Model SR03 from Fiedler) located on its sides, an ultrasonic flowmeter at the system outlet (Model PCM4 from Nivus) and an ultrasonic level gauge (Model US1200 from Fiedler) located further upstream (Fig. 1). The two rainfall signals were aggregated into one input timeseries with a time step of two minutes, the same time resolution as the field data (Fig. S1). The two analyzed outputs were water level in an upstream manhole, H.up, and outlet streamflow, Q. We selected 22^{nd}-23^{rd} July 2010 as calibration period (Fig. S9) and 27^{th} August 2010 for validation (Fig. 4).

4.3.2 The deterministic models

We constructed 5 model structures to simulate the sewer flow (Fig. 2). These alternative hypotheses with increasing detail aim to represent the typical levels of complexity used in urban hydrological modeling (Coutu et al., 2012b; Leitao et al., 2010). The model structures differ in the number of elements included to describe the topology and rainfall-runoff processes of the drainage system and/or the number of calibration parameters, the latter being a classical measure of model complexity (Spiegelhalter et al., 2002). The characteristics of each model parameterization, in increasing order of complexity, are:

M1: lumped structure with a single linear reservoir. The water height in the reservoir reproduces the behavior of the water level, H.up, whereas its output simulates the catchment streamflow, Q. An additional constant flux represents the baseflow. 2 parameters are calibrated.

M2: includes 2 non-linear reservoirs (subcatchments) in series which simulate the runoff generation, and 4 conduits in series which route the overland flow. An additional constant flux captures the baseflow. 7 parameters are calibrated.

M3: comprises 16 reservoirs (subcatchments) and 25 conduits in series. Infiltration/inflow is defined by a series of constant fluxes and unit hydrographs. 11 parameters are calibrated.

M4: distributed combination of 47 reservoirs (subcatchments) and 58 branched conduits. Infiltration/inflow is defined by a series of constant fluxes and rainfall-derived unit hydro-

graphs. 15 parameters are calibrated.

M5: architecturally identical to the previous one. Its flexibility, however, is increased since 20 parameters are now calibrated.

All model parameterizations, except for the simplest one, which is programmed in R (R Core Team, 2013), are implemented in the open source software EPA-SWMM (Rossman and Supply, 2010). More information about the model parameters is given in the Supporting Material.

4.3.3 Formulation of prior knowledge

The prior distributions of simulator parameters represent the formalized existing knowledge about the watershed's characteristics and behavior. Having a more or less direct physical meaning, the parameters of the hydrodynamic model can be relatively easily elicited from experts (Scholten et al., 2013). Here, prior simulator parameters were estimated via an engineering analysis of the catchment. For instance, the prior mean of the parameters related to soil imperviousness were estimated by comparing the land-use categories to their typical imperviousness (Butler and Davies, 2010). The parameters connected to the roughness coefficient were instead assigned by analyzing the typical Manning's n values for given soil categories, e.g. asphalt or concrete pipe (Rossman and Supply, 2010). Finally, parameters associated to geometrical characteristics of the system (width, surface) were ascertained by a spatial analysis. Details about the distribution of the deterministic model parameters are given in the Supporting Material (Tab S1 and S2).
The prior (hyper)parameters of the observation errors **E** can be formulated by quantifying the random fluctuations of the sensor in a period not used in the analysis, by integrating the manufacturer specifications about the measurement accuracy, or by performing independent tracer tests. In this case we used all three sources of information. Note that the prefix "hyper" refers to parameters of a parameter distribution. For σ_{E_Q} and $\sigma_{E_{H.up}}$, the random standard deviations of the measurements of our output variables, we selected lognormal distributions centered in very low values. We defined the first moment of σ_{E_Q} as 0.1 % of a high flowrate (300 $\mathrm{l\,s^{-1}}$) and the first moment of $\sigma_{E_{H.up}}$ as 0.1% of the approximate measurement range of the level sensor (1 m). In applying a data transformation, we assume that measurement errors increase with increasing output.
While eliciting priors for the parameters of **E** can usually be straightforwardly done by analyzing the measurement process, defining priors for the parameters of **B** is less obvious. Indeed, **B** expresses the combined effect of model structure and input errors, which are challenging to determine before observing the data. To quantify the prior distribution of the bias parameters we considered two objectives (Dietzel and Reichert, 2012). First, we wanted our simulator to explain the output observations as much as possible. For this reason, for σ_B, the standard deviation of the bias process, we selected 2 normal truncated distributions (one for each output) centered in 0 and solely defined in the real positive domain. In this way we favor a priori small magnitudes of the bias. Second, through the priors of the bias, we can influence how to "distribute" the bias between different model variables. For instance, specifying a very small (hyper)prior of $\sigma_{B_{H.up}}$ and a very wide (hyper)prior of σ_{B_Q}, we could force the simulator to adjust better the time series of water level than that of discharge. In this case, the prior information did not favor any prior. Therefore, we selected the standard deviation of σ_{B_Q} to be 50% of a high flowrate (300 $\mathrm{l\,s^{-1}}$) and $\sigma_{B_{H.up}}$ to be 50% of a high water level (10 cm).

Regarding τ, the other bias parameter, we assigned it a prior mean of slightly less than 1/3 of the hydrograph recession time. Our goal here was to capture the average correlation structure of the residuals. We also decided to assign the same prior bias to all model structures in order not to influence the comparison. Detailed information about the distribution of error model (hyper)priors is given in Tab. 2.

Table 2: Error model calibration parameters ($\boldsymbol{\psi}$). The notation for prior distributions is: LN(μ, σ): lognormal, TN(μ, σ, a_1, a_2): truncated normal, Exp(λ^{-1}): exponential. The symbol meaning is: μ: expected value, σ: standard deviation, a_1: lower limit, a_2: upper limit, λ: rate.

Name	Description	Units	Prior
corrlen (τ)	Correlation Length of $\mathbf{B}_Q$ and $\mathbf{B}_{H.up}$	min	LN(3, 3)
sd.Eps_Q (σ_{E_Q})	Standard Deviation of $\mathbf{E}_Q$	g(l/s)	LN(0.3·$\frac{\mathrm{d}g}{\mathrm{d}y}\Big\vert_{300}$, 0.06·$\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)_{300}$)
sd.Eps_H.up ($\sigma_{E_{H.up}}$)	Standard Deviation of $\mathbf{E}_{H.up}$	g(m)	LN($10^{-3}\cdot\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)_1$, $2\cdot10^{-4}\cdot\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)_1$)
sd.B_Q (σ_{B_Q})	Standard Deviation of $\mathbf{B}_Q$	g(l/s)	TN(0, 150·$\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)_{300}$, 0, ∞)
sd.B_H.up ($\sigma_{B_{H.up}}$)	Standard Deviation of $\mathbf{B}_{H.up}$	g(m)	TN(0, 0.05·$\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)_1$, 0, ∞)

4.3.4 Specific error model definition

We modeled the inadequacies observed in our case study with an output-dependent bias description (Del Giudice et al., 2013). Output dependence was achieved via a Box-Cox transformation (see Sect. 4.2.3). This means that the Gauss-Markov process $\mathbf{B}_M(\boldsymbol{\psi})$ described in Appendix B, has constant variance in the transformed space, but, when non-linearly transformed back to a real space, its variances increases with higher output values.
For the variable Q, we selected $\lambda_1 = 0.5$ [-] and $\lambda_2 = 0$ [$\mathrm{l\,s^{-1}}$], whereas for H.up we selected $\lambda_1 = 0.5$ [-] and $\lambda_2 = 0.01$ [m]. These values ensured a realistic representation of the uncertainties during high and low flows with all model structures (see Results). They are very close to those used in similar studies (Sikorska et al., 2013; Honti et al., 2013; Dietzel and Reichert, 2012).
We decided to use the same error model parameterization for each prediction variable to ensure comparability of the bias parameters for all model variants.

4.3.5 Computer implementation

The inference and prediction routines were programmed in R (R Core Team, 2013). We first maximized the posterior with the "Multi-Level Single-Linkage" algorithm for global optimization (Johnson, 2014). Then, we used this optimum as a starting point for the posterior characterization via Metropolis Markov chain Monte Carlo (MCMC) sampling (Metropolis et al., 1953). We first tuned the jump distribution sequentially using two stochastic algorithms (Vihola, 2012; Haario et al., 2001) and then sampled keeping the jump distribution fixed. This posterior exploration involved circa $5{\cdot}10^4$ simulator runs. Finally, we selected a representative MCMC sample of 1900 elements to numerically approximate the predictive distributions. Plots of the posterior distributions are given in the Supporting Information.

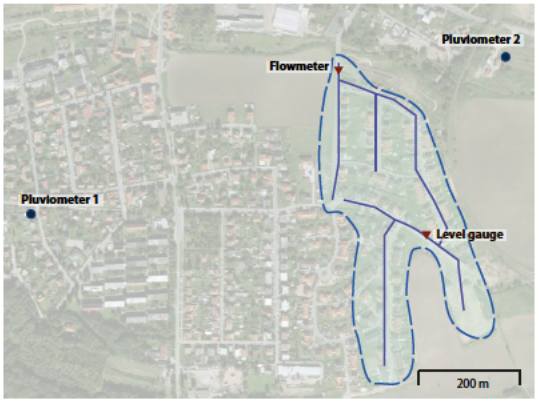

Figure 1: Hostivice watershed, Prague, Czech Republic. The main stormwater pipes and the measurement stations are indicated.

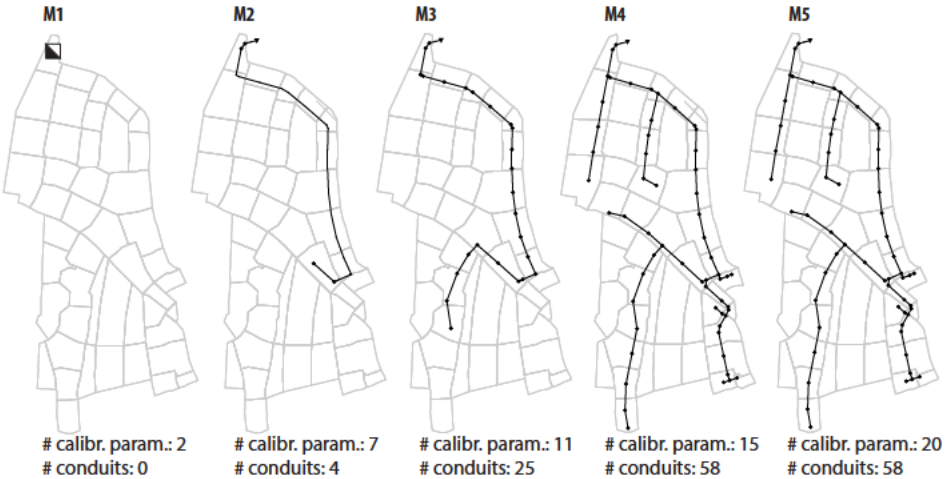

Figure 2: Increasingly complex model structures considered in this study. The black lines represent the conduits whereas the light gray lines illustrate the subcatchment boundaries and are only given as spatial reference. The points represent the nodes of the drainage system. For the first structure the square symbolizes the reservoir used to model the whole catchment.

4.4 Results

In general, our results showed a pronounced decrease of bias with increasing model complexity up to the intermediate model structure (M3). With more realistic and flexible structures beyond M3, indeed, bias decreased less and less. Additionally, parameter uncertainty reached a minimum in M3 and increased for simpler and more complex models. Predictive uncertainty intervals of all models contain circa 95 % or more of the validation data. The models behaved similarly for both outputs, outlet flow and water level in the network.

4.4.1 Calibration

As expected, the inferred magnitude of the bias, represented by the posterior $\sigma_{B_{H.up}}$ and σ_{B_Q}, decreased with increasing model complexity (Fig. 3, top row). After an initially pronounced decline, the bias gradually stabilized to its minimum. The Nash-Sutcliffe efficiency (NS) mirrored this behavior by rapidly rising at first as a function of model complexity, and then tapering off (Fig. 3, bottom row). Even the best fitting model structure showed a significant remaining bias (circa half of the initial bias of M1). The bias reduction and the corresponding matching improvement appeared slightly more pronounced for the discharge downstream than for the upstream water level.

Most parameters were well identified during the inference (Fig. S3-S7). Some posterior marginals (as, for instance, those of σ_B and τ), however, showed a relevant correlation. Observation of the spread of posteriors revealed that parametric uncertainty decreased until model structure M3 and then increased again (Fig. S9).

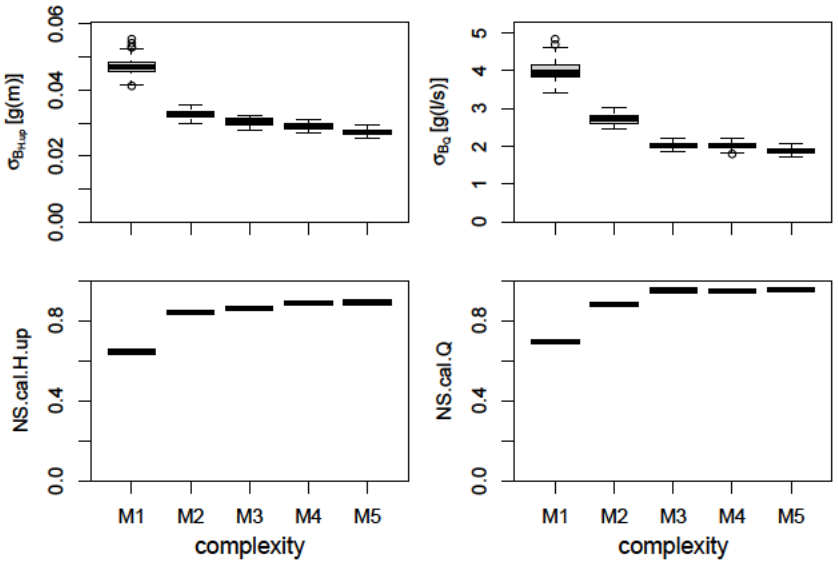

Figure 3: Indices demonstrating the simulator performances in the calibration period for both model outputs (water level on the left, discharge on the right) with increasingly complex models. Top row: the box plots illustrate the posterior distributions of σ_B and indicate the amount of bias identified. Bottom row: the Nash-Sutcliffe criterion is shown, which expresses how well the median of the simulator, $\mathbf{y}_M^c$, matches the calibration time series, $\mathbf{y}_o^c$.

4.4.2 Predictions in the extrapolation domain

After propagating the uncertainties in the validation period with the same error model, we observed a prominent prevalence of uncertainty coming from bias (gray in Fig. 4). There, it can also be seen that the nominal coverage of the uncertainty intervals (95%) is always approximately met or exceeded, which means that our predictions are reliable (see Sect. 4.2.3). At the same time, the predictive quantile-quantile plots (Fig. S2) showed a slight overestimation of total uncertainty for some model structures. In general, however, the bands appeared to be plausibly wide.

As noticed during inference (Fig. 3) and also throughout validation, the bias (and the total uncertainty) decreased when making the structure more complex. A similar improvement was observable in model fit, represented by an increasing NS (first and third row of Fig. 5). Although the most complex model structures fitted validation data very well (NS close to 1), the remaining bias was still about half of the original bias of M1. Indeed, as visible in Fig. 4, the bias was the main uncertainty component (gray) for every model followed by parameter (light gray) and observational uncertainty (dark gray).

In the second row of Fig. 5 it can be seen that the effect of parametric uncertainty is substantially less than that of bias. The uncertainty coming from $\mathbf{y}_{\mathrm{M}}(\boldsymbol{\Theta}_{post})$, in contrast to the bias behavior, did not exhibit a constant trend as a function of complexity. Instead, the output parameter uncertainty at first showed a decrease and then a rise in function of model complexity.

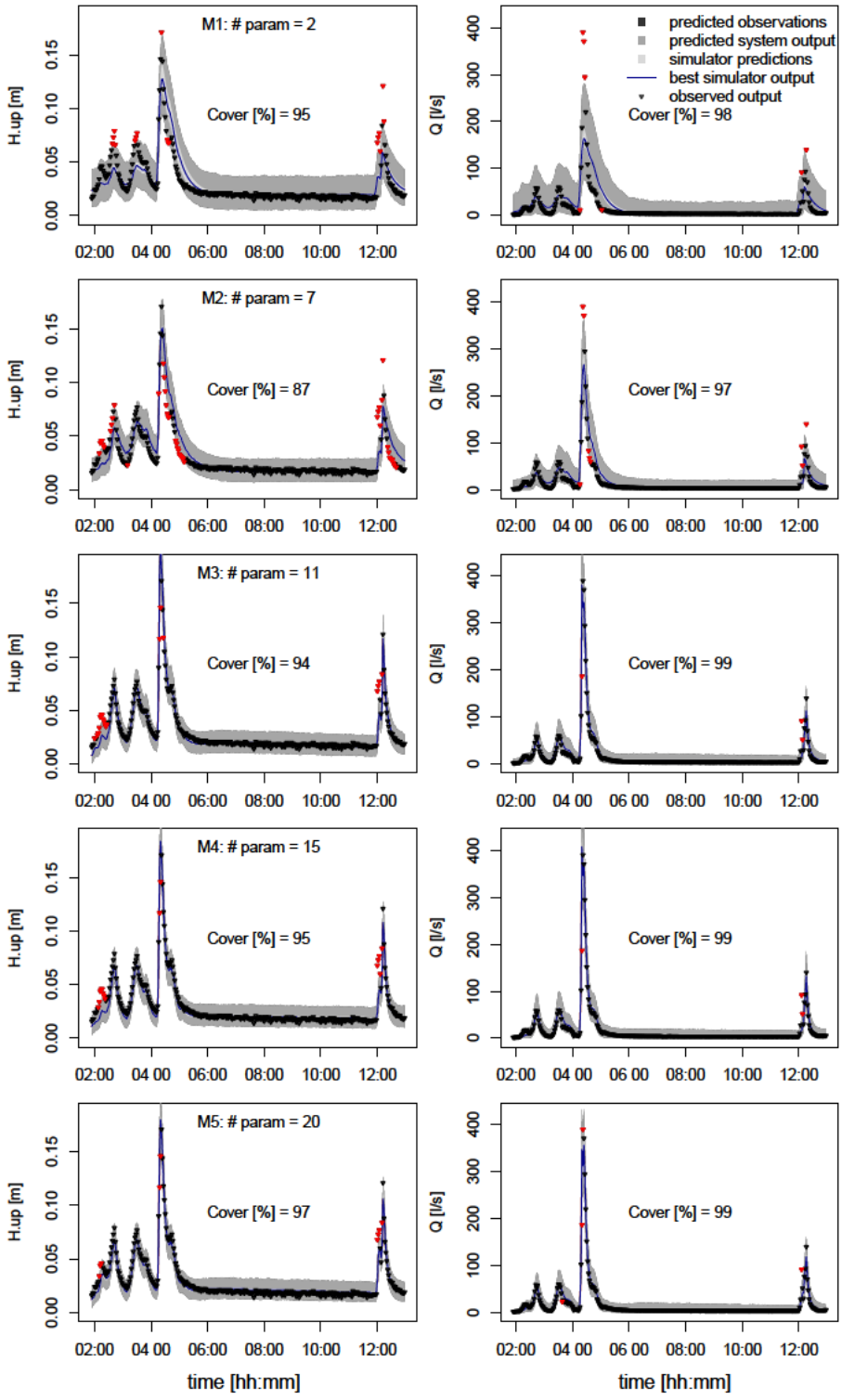

Figure 4: Predictions of water level upstream (H.up) and discharge (Q) in the extrapolation period for the alternative model structures. The validation data are represented by triangular points. The 95% credible intervals are interpreted as follows: parametric uncertainty due to $\mathbf{y}_M^e(\boldsymbol{\Theta}_{post})$ (light gray), parametric plus input and structural uncertainty due to $g^{-1}(\tilde{\mathbf{y}}_\mathrm{M} + \mathbf{B}_\mathrm{M})$ (gray), total uncertainty due to $g^{-1}(\tilde{\mathbf{y}}_\mathrm{M} + \mathbf{B}_\mathrm{M} + \mathbf{E})$ (dark gray, not distinguishable at this scale [see Support for a magnification]). Validation measurements not included in this dark gray region are marked in red. The dark blue line is the median of $\mathbf{y}_\mathrm{M}$ and represents our best knowledge of the future system response. The number of the simulator parameters which we inferred is also shown.

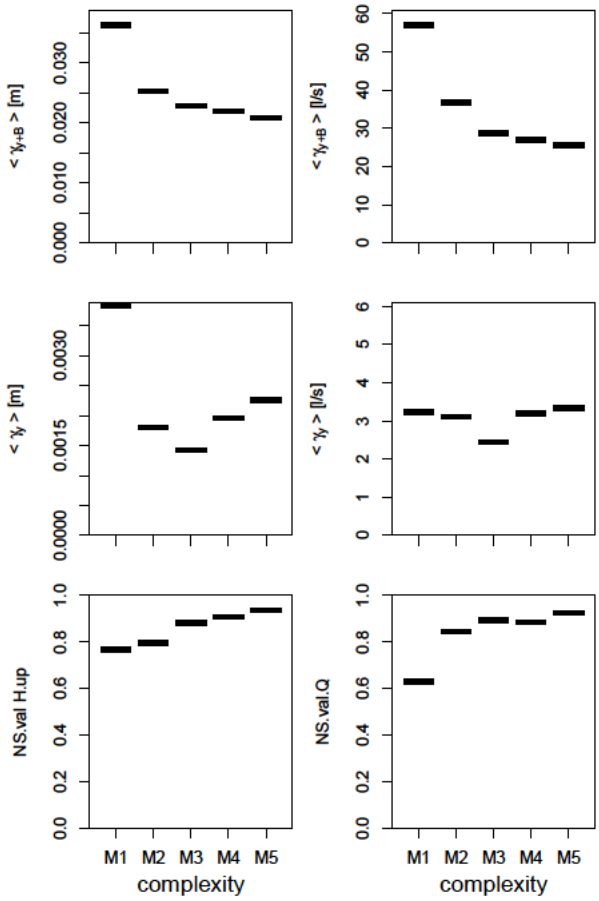

Figure 5: Predictive performances of the different models for water level (left) and flow (right). $\langle \gamma_{y+B} \rangle$ represents the average width of the 95 % predictive intervals due to model bias and given the parameters. Similarly, $\langle \gamma_y \rangle$ quantifies the effects of posterior parametric uncertainty on model output. On the bottom, NS evaluates how closely the simulator fits the validation data. While the first two indices would be ideally 0, the NS would be 1 in the case of a perfect match.

4.5 Discussion

In this paper we combined a stochastic bias description with a multimodel comparison to learn about the behavior of the Gaussian bias process. Results of our case study show that consideration of systematic deviations leads to reliable runoff predictions for all models and outputs. Furthermore, we observed a progressive decrease of bias with increasing model parameterization. In the following we will i) link the gradual decrease of the bias to structural and input errors, ii) analyze the behavior of parametric uncertainty, iii) discuss our approach with respect to optimal model selection, iv) assess the benefits and v) the limitations of this study in relation to others, and vi) recommend further research tasks.

4.5.1 Connecting the bias behavior to structural and input errors

As expected, the bias generally decreased with increasing model complexity. Since the input errors remain the same, we can associate the reduction of bias with a reduction of structural deficits. Additionally, since from a certain model complexity performance did not improve any more and given our knowledge of the system, we can think of the remaining bias here as being primarily due to input errors. This is very plausible because the input for our models was measured by pluviometers located a few hundreds of meters outside the catchment area (Fig. 1). Furthermore, output measurements were conducted with greatest care, in order to reduce systematic errors. An analysis of the uncertainty of the predictions suggests that by reducing the structural deficits we can reduce the initial bias uncertainty by about 50%.

As we think that the model M5 sufficiently parameterizes the main hydrological process, its remaining bias (gray region of the bottom row of Fig. 4) can be interpreted as largely caused by input uncertainty. Note that the uncertainties are not additive. Actually, due to partial dependence among the error types, it is probable that the total uncertainty, in terms of predictive variance, is less than the sum of the individual variances.

4.5.2 Interpreting parametric uncertainty

In agreement with Schoups et al. (2008), in our example, total predictive uncertainty did not increase with mechanistic model complexity (Fig. 4). This is, however, connected to the relatively small parametric uncertainty affecting even the most flexible model structure M5. Presumably, by further increasing the number of calibration parameters we would observe a rise in total uncertainty because the information content in the two output time series would not be enough to reduce the prior parametric uncertainty. This would especially affect prediction variables which have not been used for calibration (e.g. the water level at some point of the network without observations).

Predictive uncertainty due to simulator parameters is minimal for the intermediate structure and rises for simpler and more complex ones (Fig. 5). This behavior can be interpreted as follows. Assuming reasonably identifiable model parameters, output parametric uncertainty usually rises with the standard deviation of calibration residual errors. Simultaneously, for similarly fitting models, this uncertainty generally increases with the number of parameters due to reduced identifiability. For the simplest model structures we foresee high variance due to large bias. For the intermediate models, with small residuals and high parsimony, we anticipate low variance. For the most complex and overparametrized models we expect again high parameter spread. This pattern is evident in Fig. 5. It might, however, become less clear when the architecture of the model changes.

4.5.3 Relations to model selection

In this study we tried to represent the typical complexity levels used in urban runoff forecasting. We selected a completely lumped approach as the simplest model, M1 (Coutu et al., 2012b), a fully-detailed network modeling as the two most complex structures, M4 and M5, and two simplified network structures in between (Leitao et al., 2010). The methodology we presented can also be useful to support optimal selection among these commonly applied parameterizations. Indeed, if the goal of a study is to select the best model for predicting a certain output (e.g. stormwater flow or water level), our approach can help by providing its predictive uncertainty. Based on the considerations in Sect. 4.5.1, total uncertainty is a good indicator of model complexity control since it simultaneously accounts for bias and parameter uncertainty, the factors that a modeler might want to minimize. In our case, structure M5 has these optimal characteristics. However, other criteria might also be relevant for model selection and other complexity control methods could complement our analysis of the uncertainty types (Schoups et al., 2008; Leube et al., 2013).

4.5.4 Advantages of the methodology

Based on our current experience, the proposed approach has several benefits for environmental engineering in general and urban hydrology in particular: i) it provides reliable, sharp, and robust probabilistic predictions for all analyzed model structures and output variables. This is an important strength with respect to the oversimplified approaches currently applied in urban drainage (Freni et al., 2009b; Dotto et al., 2011) and environmental analysis as a whole (Liu et al., 2010). ii) It can integrate available information from different measurement types and quality levels. This can improve identifiability of model parameters and thus reduce predictive uncertainty. It can also provide insights to better understand where and how the model is deficient. iii) It can help quantifying how much of the total predictive uncertainty comes from parameter uncertainty, input errors, and random observation errors, especially when the system is well-known and the output observations are accurate. This was not possible in previous studies analyzing the structural uncertainty by means of multimodel comparison (Butts et al., 2004; Zhang et al., 2011). iv) It can point to the causes for bias by only analyzing the output of the system and of the model. This output analysis is usually much simpler and faster than other stochastic techniques developed to quantify the causes of bias (Lin and Beck, 2012; Reichert and Mieleitner, 2009; Renard et al., 2010; Bulygina and Gupta, 2009). v) It can easily accommodate different kinds of models and commercial software since we do not need to modify the simulator equations (as for instance in Vrugt and Robinson (2007); Reichert and Mieleitner (2009); Bulygina and Gupta (2009); Breinholt et al. (2012)). In contrast, we just have to add an external stochastic process.

4.5.5 Limitations

The downside of its simplicity is that the methodology can only partially quantify the effects of the different error sources. For instance, we can affirm that a significant portion of the remaining bias is due to input errors, yet some remaining structural deficits and systematic measurement errors might possibly play a role. Additionally, we cannot separately assess the individual effects of structural or input error, nor precisely understand which misspecified process or faulty rainfall sensor causes output errors. This implies, for instance, that we cannot know in advance how much the total uncertainty of a given model structure would decrease if the input uncertainty

was eliminated. Moreover, when using an output error, we can only predict the uncertainty of output components that we can measure.
Although simpler and faster than similar approaches, it still requires the construction of increasingly realistic and flexible models. Furthermore, it necessitates an appropriate parameterization of the bias and tens of thousands of simulator runs for posterior characterization, which might be computationally very expensive for highly complex models. On the other hand, CPU intensive calculations are a drawback of most Bayesian techniques for uncertainty analysis (Yang et al., 2008; Dietzel and Reichert, 2012).

4.5.6 Outlook to future research

In describing model deficiencies as an autoregressive Gaussian process and calibrating different model structures, we increased the information extractable from the measured time series. If we only use a lumped normal bias in the output space instead of describing the errors where they arise, we are indeed limited in the amount of information that allows input and structural uncertainty to be separately qualified. In order to even more precisely understand where to focus our efforts to reduce predictive uncertainty, techniques which support the identification of the causes of bias should be investigated. This could mean propagating input uncertainty (Renard et al., 2011), making parameters stochastic and time-varying (Reichert and Mieleitner, 2009), or adding a dynamic noise to the model states (Vrugt and Robinson, 2007). These approaches, being computationally demanding, could make use of quick statistical surrogate models like mechanistic emulators (Kennedy and O'Hagan, 2001; Bayarri et al., 2007; Reichert et al., 2011; Albert, 2012), which would facilitate their applicability.

4.6 Conclusions

The goal of this study was to present a relatively simple yet statistically sound method to understand how model inadequacies depend on structural complexity. For this purpose, we used a Bayesian description of model bias, which we demonstrated in previous studies to improve the reliability of environmental forecasts. Here, we combined the strengths of this approach with those of multimodel comparison, to learn more about the causes of model inadequacies from the output data. We have demonstrated the usefulness of our method by analyzing the behavior of several stormwater models, although the technique is general and applicable to many other environmental case studies. Based on our results, theoretical considerations and previous experiences in the literature we conclude that:

I. Analyzing a system with increasingly complex model structures while describing their bias as an autoregressive process is an effective way to better interpret output uncertainty. This allows us to associate the reduction of bias of hydrodynamic models with the effect of structural errors, an uncertainty component that is challenging to quantify. With this method we can also approximately quantify the effects of input errors. If the most complex model has sufficient flexibility to adequately describe the dominant hydrological processes and if output measurements are accurate, the remaining bias can be assumed to be mainly caused by input uncertainty.

II. Our technique provides statistically-sound predictions of several output variables in a relatively simple way. This is highly relevant for environmental predictions, particularly in

hydraulic or hydrological modeling where reliable river or sewer flow predictions are still challenging.

Acknowledgements

We are grateful to Anna Sikorska and three anonymous referees for the helpful comments on the manuscript. The Swiss National Science Foundation is acknowledged for financing this research within the scope of the COMCORDE project (grant No. CR2212_135551). Additional support was provided by the Czech Science Foundation, project No. 14-22978S. The city of Hostivice deserves our acknowledgments for their support.

Supplementary data related to this article can be found at http:// dx.doi.org/10.1016/j.envsoft.2014.11.006.

Chapter 5

Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation

D. Del Giudice[a,b], C. Albert[a], J. Rieckermann[a], P. Reichert[a,b].

[a]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b]ETHZ: Swiss Federal Institute of Technology Zürich, 8093 Zürich, Switzerland

Water Resources Research (under review)

Author contributions

D.D.G. designed the experiments, built the model, developed the method, performed the analyses, wrote the paper; P.R. developed the method, conceived and designed the experiments; all coauthors gave advices, supported result interpretation, and paper revision.

Abstract

Rainfall input uncertainty is a major concern in hydrological modeling. Unfortunately, during inference, input errors are usually neglected, which can lead to biased parameters and implausible predictions. Rainfall multipliers can reduce this problem but still fail when the observed input (precipitation) has a different temporal pattern from the true one or if the true non-zero input was not detected. In this study we propose an improved input error model, which is able to overcome these challenges and to assess and reduce input uncertainty. We formulate the average precipitation over the watershed as a stochastic input process (SIP) and, together with a model of the hydrosystem, we include it in the likelihood function. During statistical inference, we use input (rainfall) and output (runoff) data to learn about the "true" rainfall, model parameters, and runoff. We test the methodology with the rainfall-discharge dynamics of a small urban catchment. To assess its advantages, we compare SIP with simpler techniques: i) standard least squares (LS), ii) Bayesian bias description (BD), and iii) rainfall multipliers (RM). We also compare two scenarios: accurate versus inaccurate forcing data. Results show that, when inferring the input with SIP, physical parameters are "protected" from the corrupting impact of input errors. This is not the case with LS and RM. When using imperfect rain data, SIP infers the most realistic parameter values, together with BD. In addition, it infers time series of whole-catchment precipitation and its associated uncertainty. During validation, SIP also delivers realistic uncertainty intervals for both rainfall and runoff. Thus, we recommend this technique in all cases where the input of a system is the predominant source of uncertainty and millions of model runs can be performed in reasonable time. Furthermore, the high-resolution rainfall intensities obtained with SIP can help validate areal rainfall estimates from other methods and constitute an important contribution toward the disentanglement of predictive uncertainties.

5.1 Introduction

One of the main sources of uncertainty in hydrological modeling are input errors. These are predominantly associated to errors in the estimation of the true precipitation over a watershed (Kuczera et al., 2006; Vrugt et al., 2008). Hydrological systems are indeed heavily input-driven and inaccuracies in rainfall characterization can dramatically impair the quality of calibration results and model output.

Rainfall input errors affecting model calibration arise from a variety of reasons: inadequate areal coverage of point-scale pluviometers, inexact spatial interpolation, mechanical limitation of the gauge, wind effects etc. (McMillan et al., 2011; Renard et al., 2011). Furthermore, precipitation provided at an insufficient temporal resolution can substantially impair the model ability to represent runoff, especially for small and fast-reacting catchments (Ochoa-Rodriguez et al., 2015).

Traditionally, input uncertainty has generally been neglected in the inference process because of the mathematical complexity of including it in a likelihood function. The likelihood is the parametric model describing the probability distribution of observations given the values of model parameters and the inputs. This probability function is needed to extract information about model parameters and input from observed data. To make correct inference, the likelihood should consider all relevant mechanisms and error contributions. However, as discussed, e.g., by Yang et al. (2008), Sikorska et al. (2012b), and Reichert and Schuwirth (2012), likelihood functions formulated as uncorrelated normal distributions of observations, that are centered at

the outputs of a deterministic model, are still frequently used. This means that all discrepancies between deterministic model results and output data only stem from measurement errors (Tomassini et al., 2009). Due to input errors and/or structural deficits this assumption is usually unrealistic (Yang et al., 2007a). Normal iid (independent and identically distributed) likelihoods have repeatedly shown to produce biased estimates of model parameters and unreliable predictions (Renard et al., 2011; Honti et al., 2013; Del Giudice et al., 2015a).

One alternative to the iid uncertainty description is the use of autoregressive error models (Kuczera, 1983; Yang et al., 2007a). Although these likelihoods are still simple, they implicitly acknowledge the existence of errors besides the random output measurement noise (including inaccuracies in the input estimation). The effects of these errors on model output has been described by autocorrelated stochastic processes added to the model output (Frey et al., 2011; Evin et al., 2013). In recent studies focusing on reliable output predictions, (iid) observation errors have been explicitly considered in addition to the process describing correlated deviations (Reichert and Schuwirth, 2012; Del Giudice et al., 2013; Dietzel and Reichert, 2014). This formulation of (autocorrelated) model deficiencies, in the following called "bias description" (BD), makes it possible to learn about systematic discrepancies of model output from calibration data and to more realistically assess the associated uncertainties.

While likelihoods describing bias are more plausible than iid ones, they still have some limitations: i) they can only provide limited information about the causes of model bias and, therefore, do not help much to disentangle input from structural errors; ii) they can only partially buffer the corruption of model parameter estimates; iii) they do not contribute to quantifying the uncertainty of unobserved variables (such as water level in an arbitrary point of the drainage network) (Reichert and Mieleitner, 2009; Del Giudice et al., 2015b).

A more satisfying approach for considering input errors is to make the input uncertain and propagate it through the model (Honti et al., 2013). A simple way of doing so, which has become popular in hydrology, is the use of so-called rainfall multipliers (RM) (Kuczera et al., 2006; Sun and Bertrand-Krajewski, 2013). These are event-specific random variables multiplied with the observed rain to provide the input to the model. These multipliers (and their uncertainty) is then estimated jointly with the other model parameters to correct for (possible) rainfall input errors during the calibration period. While using rainfall multipliers is relatively straightforward, they have important drawbacks: multipliers do not provide a realistic assessment of input uncertainty if, for example the temporal dynamics, i.e. the "shape", of a recorded storm event is significantly different from the true precipitation dynamics or if a storm bypasses the pluviometric stations so that they do not record any precipitation although the catchment shows a runoff response (Renard et al., 2011; Vrugt et al., 2008). While the first disadvantage can be reduced with multipliers varying within the storm event (Reichert and Mieleitner, 2009), the second one cannot be solved within this framework and requires a fresh approach.

In this study, we therefore suggest a novel input uncertainty model that describes the input of a hydrosystem as a continuous stochastic process. This makes it possible to formulate a more realistic likelihood function than those discussed above. This also allows us to learn about and reduce input as well as output uncertainties. Via Bayesian inference, we show how to update our prior beliefs about the parameters and rainfall patterns from the simultaneous use of input data (here: from pluviometers), output data (from a flowmeter at the outlet of the catchment), the runoff model (a lumped linear reservoir), a rainfall model (a Gauss-Markov process), and models of the input and output observation errors (both normal distributions). We name this

method Stochastic Input Process, SIP.
SIP, which can have much broader applications, has the following benefits:

I. It can probabilistically estimate the true input to a system in cases of sparse, inaccurate, or imprecise input measurements, if the output measurements are comparably accurate. This can be valuable to reconstruct past precipitation records from flow data or to spatially upscale point measurements. Reliable precipitation estimates can also be very useful to test hydrologic theories and benchmark recordings from other sensors like radars (Kirchner, 2009).

II. It can reduce the bias in inferred parameters of hydrological models and therefore in runoff predictions. This can substantially support regionalization studies, which try to establish relations between hydrological model parameters calibrated in gauged catchments and properties of these catchments (Kuczera et al., 2006).

III. It can produce not only a reliable assessment of total output uncertainty, but also quantify the contributions due to parameter and input uncertainty. Supporting uncertainty separation, SIP can help assess in how far prediction uncertainty can be reduced by providing better rainfall data and can therefore guide our efforts to minimize the uncertainty sources (Sikorska et al., 2012b).

We examine the ability of our Bayesian approach to produce realistic posterior parameter estimates and reliable predictions, and compare it with the three methods mentioned above: the simple least squares (LS) formulation assuming iid errors, an autoregressive bias description (BD), and the event-dependent rainfall multiplier (RM) error model. As illustrative example, we perform inference and prediction for a monitored urbanized watershed which we model with a parsimonious hydrological model of combined waste- and rainwater discharge.

5.2 Method

We briefly describe the LS, BD, and RM approaches, three commonly used techniques to calibrate and predict with environmental and, specifically, rainfall-runoff models. All techniques are implemented in a Bayesian framework, meaning that the likelihood function is combined with a prior distribution of the parameters in order to obtain posterior parameter estimates from observations. This allows us to make use of our existing knowledge about physical and error model parameters. Prior knowledge can reduce the identifiability problem between the process-based model and the statistical error model (Bayarri et al., 2007).
After this review, we explain more in depth the concepts and numerics of the method developed in this paper that is based on describing rainfall as a Stochastic Input Process (SIP). We finally discuss the numerical experiment we performed to demonstrate the usefulness of the SIP calibration scheme in presence of important input errors. A graphical comparison of the methods is provided in Figure 1.

5.2.1 Alternative methods used for comparison

Standard least squares - LS method

The standard non-linear least squares is the simplest of the 4 approaches. This regression method describes the residual errors, the differences between the output of a deterministic model, $\mathbf{y}_M$, and observations, $\mathbf{y}_o$, as normally distributed and independent (Figure 1, panel a). The implicit

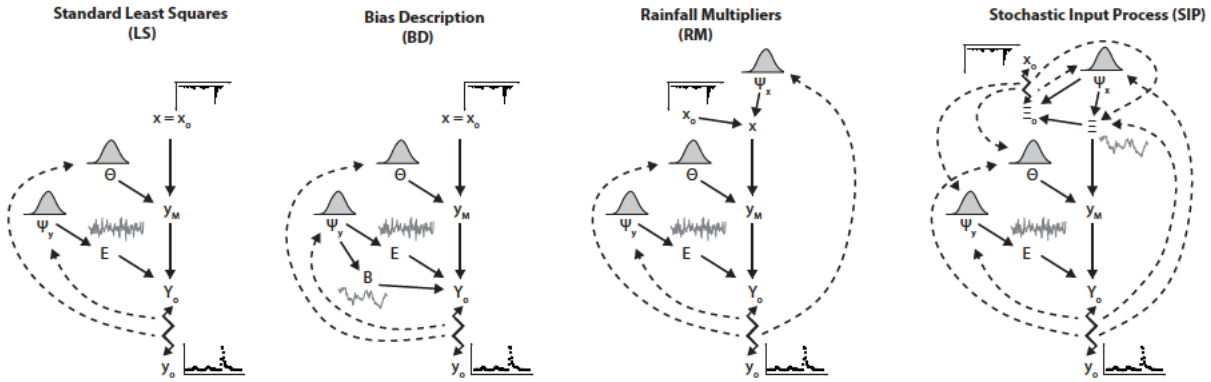

Figure 1: Representation of the 4 methods we use to describe the uncertainties during inference and predictions. The random processes $\mathbf{E}$, $\mathbf{B}$, and $\boldsymbol{\Xi}$ have different illustrations depending on their consideration of autocorrelation. The dashed lines exemplify the learning process of inference while the solid lines illustrate the information flow during predictions. The zigzags typify the comparison between modeled and measured time series taking place during inference. While the RM implicitly accounts for input uncertainty by inferring additional parameters, SIP explicitly considers it in the likelihood function.

assumption here is that model results deviate from data only because of random observation errors, $\mathbf{E}$. The other error sources, like input and structural errors, are neglected or somehow considered to have the same effect as white observation noise (Vrugt et al., 2008; McMillan et al., 2011). Relaxing the assumptions of constant variance and normality of the residual errors via output transformation can make the LS approach slightly more appropriate for hydrological applications (Wang et al., 2012). The probabilistic model used for inference and prediction by this approach can be written as:

$$f(\mathbf{y}_o \mid \boldsymbol{\theta}, \psi_y, \mathbf{x}) = \frac{(2\pi)^{-\frac{q}{2}}}{\sqrt{\det\big(\boldsymbol{\Sigma}(\psi_y)\big)}} \cdot \exp\left(-\frac{1}{2}\Big[g(\mathbf{y}_o) - g\big(\mathbf{y}_M(\boldsymbol{\theta}, \mathbf{x})\big)\Big]^{\mathrm{T}} \boldsymbol{\Sigma}(\psi_y)^{-1}\Big[g(\mathbf{y}_o) - g\big(\mathbf{y}_M(\boldsymbol{\theta}, \mathbf{x})\big)\Big]\right) \prod_{i=1}^{q} \frac{\mathrm{d}g}{\mathrm{d}y}(y_{o,i}) \;, \quad (2.1)$$

where $\boldsymbol{\theta}$ is the vector of hydrological model parameters, ψ_y is the vector of parameters of the error term, $\mathbf{x}$ is the time-varying model input (in our case precipitation), and the scalar transformation function, g, when applied to a vector, returns the vector of the function applied to all components of its argument. For the LS method, the input used in equation (2.1) is given by the observed input

$$\mathbf{x} = \mathbf{x}_o \quad (2.2)$$

(ignoring input uncertainty), and the covariance matrix $\boldsymbol{\Sigma}(\psi_y)$ is given by the diagonal matrix

$$\boldsymbol{\Sigma}(\psi_y)_{i,j} = \delta_{ij}\sigma_E^2 \;, \quad (2.3)$$

where i and j are subscripts running over the time domain of length q, and δ represents the Kronecker delta. The results of the deterministic model are represented by $\mathbf{y}_M(\boldsymbol{\theta}, \mathbf{x})$, whereas the corresponding observed values are denoted by $\mathbf{y}_o$. We here consider the heteroscedasticity of

the errors (i.e. the dependence of the error variance on the corresponding model output) via an output transformation, g, whose functional form is given in Appendix 5.A. The only parameter $\boldsymbol{\psi}_y$ of this error model is σ_E, the (constant) standard deviation of the output measurement noise in transformed units. Several studies have shown that this error description is too simple for complex environmental systems where input and structural errors play an important role (e.g., Yang et al., 2007a,a; Vrugt et al., 2008; Renard et al., 2011; Honti et al., 2013). However, we chose the LS method as a base case for comparison with the better methods because it is still widely applied in environmental modeling.

Statistical bias description - BD method

A way to consider the effects of input errors, and possibly structural deficits, on model output is to mimic the systematic deviations of model results from data with an autocorrelated stochastic process. This approach comes from the branch of statistics dealing with error models, Bayesian inference, and Gaussian processes (Kennedy and O'Hagan, 2001; Bayarri et al., 2007; Brynjarsdóttir and O'Hagan, 2014), and has recently been adapted to improve hydrological predictions (Reichert and Schuwirth, 2012; Honti et al., 2013; Del Giudice et al., 2015a).
The idea behind the Bias Description (BD) is to add an error process additional to **E** to the deterministic model output. This model discrepancy term, **B**, describes what we know about the bias correction needed to fit the output data in presence of incorrect inputs and structural deficits. **B** here follows an OrnsteinUhlenbeck (OU) dynamics (see Platen and Bruti-Liberati, 2010; Andersen et al., 2009, and references therein) for g-transformed output and data

$$\mathrm{d}B(t) = -\frac{B(t)}{\tau}\mathrm{d}t + \sqrt{\frac{2}{\tau}}\sigma_B \mathrm{d}W(t) \quad , \tag{2.4}$$

where τ is the correlation time and σ_B is the asymptotic standard deviation of the statistical fluctuations around the average value of **B**, here 0. $W(t)$ is a Wiener process, also called standard Brownian motion, or random walk with independent Gaussian increments. The first part of the (Langevin) equation (2.4) describes a deterministic dampening, central-restoring force, or pull towards the long-run mean of zero. The second term counterbalances this tendency by adding stochastic white noise. This leads to random oscillations of realizations of this process around the equilibrium state with standard deviation σ_B and correlation time τ. We chose an OU process, because it is time-continuous, linear, Markovian, it has a finite stationary variance, and it can be integrated analytically (Ibe, 2013; Wolfgang and Baschnagel, 2013).
The BD technique is similar to the autoregressive output error models of Yang et al. (2007a) and Evin et al. (2013), except that besides the autocorrelated error term **B**, representing a correction of model deficiencies, we now also separately represent the measurement noise, **E**. As adequate information on measurement precision is usually available and **B** and **E** have different properties, their identifiability is typically high (Reichert and Schuwirth, 2012). In this approach, the likelihood function has the same basic form as in equation (2.1), but the covariance matrix is non-diagonal:

$$\boldsymbol{\Sigma}(\boldsymbol{\psi}_y)_{i,j} = \sigma_B^2 e^{-\tau^{-1}|t_i - t_j|} + \delta_{ij}\sigma_E^2 \quad . \tag{2.5}$$

In this equation, i and j are subscripts spanning over the time domain, and τ and σ_B are the (hyper)parameters, $\boldsymbol{\psi}_y$, of the Gaussian bias process. As the effect of input errors is corrected at the output, also this technique is based on using the observed input (2.2) when applying the likelihood function (2.1) with (2.5).

Multiplicative rainfall error model - RM method

The RM (rainfall multiplier) approach explicitly considers input (in our case rainfall) uncertainty by perturbing the observed precipitation time series with independent random factors for all storm events (Kuczera et al., 2006; McMillan et al., 2011). To make the inference tractable, these (latent) factors are kept constant during an event (Vrugt et al., 2008). We then apply the likelihood function (2.1) with (2.3) and with the perturbed input

$$x_i = \beta_{j(i)} x_{o,i} \quad , \tag{2.6}$$

where the index i runs through all elements of the rainfall time series, whereas the index j remains constant for all values of i within any given storm event. The parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_{n_s})$ represent the linear rainfall bias corrections for all (n_s) storm events. The priors for the elements of $\boldsymbol{\beta}$ are formulated hierarchically with lognormal distributions centered at 1 and with a joint standard deviation σ^β. Centering these distributions at 1 implies a preference for the observed input. While some applications of RM kept σ^β fixed (Kuczera et al., 2006; Sun and Bertrand-Krajewski, 2013), making the error model non-hierarchical, we prefer to infer this hyperparameter to learn about the overall input variance detected during calibration (Li et al., 2012; Sikorska et al., 2012b). Despite using the same likelihood function as for the LS method, the replacement of the input description (2.2) by (2.6) leads to the consideration of input uncertainty at the level of whole storm events and augments the parameter vector by the parameters $\boldsymbol{\Psi}_x = \{\boldsymbol{\beta}, \sigma^\beta\}$ of the rainfall error model. The basic difference between LS and RM is that, while the former assumes the observed rainfall to be the true input, the latter considers sections of the true input to be unknown multiples of the recorded precipitation during that time period (Figure 1). While the RM technique provides a simple approximation for the uncertainty of the rainfall volumes, its limited ability to deal with strongly dynamic input errors has been widely acknowledged (Kuczera et al., 2006; Vrugt et al., 2008; Sikorska et al., 2012b). Consequently, a more realistic statistical representation of the catchment-averaged precipitation is needed (Salamon and Feyen, 2010; Renard et al., 2011).

5.2.2 Joint inference of input, hydrological model and output error parameters - SIP method

Overall concept

The framework we propose to quantify, reduce, and propagate input uncertainty is based on the inference of a latent Gaussian stochastic process, the "rainfall potential", $\boldsymbol{\xi}$. This time series can be transformed to the areal averaged precipitation over the watershed. Together with $\boldsymbol{\xi}$, the parameters of the hydrological model, $\boldsymbol{\Theta}$, those of the input, $\boldsymbol{\Psi}_x$, and those of the output error model, $\boldsymbol{\Psi}_y$, are also inferred. Furthermore, similarly to Sigrist et al. (2012), we simultaneously estimate $\boldsymbol{\xi}_o$, the "rainfall potential" at the pluviometric station. Here, we define "rainfall potential" as a quantity that describes the potential for having rainfall in a given catchment or at a given site and is not meant in a physical sense. These "rainfall potentials" can be transformed to the corresponding rainfall by the scalar function h:

$$\mathbf{x} = h(\boldsymbol{\xi}) \quad , \quad \mathbf{x}_o = h(\boldsymbol{\xi}_o) \tag{2.7}$$

(again, application of the scalar function h to the vector $\boldsymbol{\xi}$, representing the time series of rainfall potentials, returns the vector with elements that result from the application of h to the elements of $\boldsymbol{\xi}$). As multiple values of the "rainfall potentials" are mapped to zero precipitation, this function is not invertible (Figure 2). Therefore, we avoid denoting the elements of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_o$ the "transformed rainfall". However, whenever the rainfall intensities are not zero and thus, h^{-1} exists, they are transformed rainfall intensities.
The 2 main differences of the suggested technique to RM are:

I. The SIP technique does not assume (pieces of) the true precipitation to be proportional to the observed time series. Instead, our knowledge on true input is inferred from prior knowledge, input observations, and output observations. This makes it possible to deal with time-varying observation errors of the rainrate and with unrecorded storms (that bypassed the observation site but led to runoff increase), a situation intractable with rainfall multipliers. These features make the suggested technique conceptually more satisfying than the techniques described in Section 5.2.1.

II. The joint input and output likelihood function of SIP does not have a simple explicit form as for the RM, but it is instead given in a discretized form of a high dimensional integral over all possible realizations of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_o$. Unfortunately, this makes the suggested technique computationally more demanding than all three techniques used to compare with.

The SIP likelihood function can be written as:

$$f(\mathbf{y}_o, \mathbf{x}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x) = \int f\big(\mathbf{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}_y, \mathbf{x} = h(\boldsymbol{\xi})\big) f(\mathbf{x}_o|\boldsymbol{\xi}_o) f(\boldsymbol{\xi}_o|\boldsymbol{\xi}, \boldsymbol{\psi}_x) f(\boldsymbol{\xi}|\boldsymbol{\psi}_x) \mathrm{d}\boldsymbol{\xi} \mathrm{d}\boldsymbol{\xi}_o \,, \tag{2.8}$$

where integration is over all possible discretized time series of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_o$. This formulation is the discretized version of what in physics is called path integral (for an application in the environmental sciences see, e.g., Quinn and Abarbanel (2010)).
In the following, we describe the elements of this likelihood function. $f(\mathbf{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}_y, \mathbf{x})$ is the likelihood of observed output given the parameters of the hydrological model, $\boldsymbol{\theta}$, the parameters of the output error model, $\boldsymbol{\psi}_y$, and the rainfall input, $\mathbf{x}$; $f(\mathbf{x}_o|\boldsymbol{\xi}_o)$ is the model of observed input, $\mathbf{x}_o$, given the input potential at the observation site, $\boldsymbol{\xi}_o$; $f(\boldsymbol{\xi}_o|\boldsymbol{\xi}, \boldsymbol{\psi}_x)$ is the model for the rainfall potential at the observation site, $\boldsymbol{\xi}_o$, given the rainfall potential for the whole catchment, $\boldsymbol{\xi}$, and the input model parameters, $\boldsymbol{\psi}_x$; and $f(\boldsymbol{\xi}|\boldsymbol{\psi}_x)$ is the a priori model for the rainfall potential in the catchment, $\boldsymbol{\xi}$, given the input model parameters, $\boldsymbol{\psi}_x$. Subsequently, we describe the numerical method to implement inference of parameters and time series of rainfall potential, and, finally, we elucidate how to make predictions with SIP.

Output probabilistic model

The term $f(\mathbf{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}_y, \mathbf{x})$ represents the probabilistic hydrological model for the observed discharge, $\mathbf{y}_o$, as a function of the rainfall, $\mathbf{x}$, the parameters of the hydrological model, $\boldsymbol{\theta}$, and those of the output error model, $\boldsymbol{\psi}_y$. This probability density function is assumed to be given by equation (2.1) with the covariance matrix given by equation (2.3) as for the LS and RM approaches. The difference is, again, the representation of the input. The LS approach assumes the observed input to be error-free (equation 2.2). The RM approach assumes the true input

to be a linearly-scaled and random version of the observed one (equation 2.6). In our new approach, we use our best knowledge of the true input by inferring the rainfall potential, $\boldsymbol{\xi}$, and the corresponding input, $\mathbf{x}$, jointly with the parameters of the hydrological model and the error models. In some sense, we use the output of the catchment as an additional rain gauge, to gain spatially-integrated information about the rainfall in the catchment.
Similarly to the RM approach, the use of the probability density function (2.1) with the covariance matrix given by equation (2.3) assumes that model structural deficits are negligible and can be "absorbed" by parameter uncertainty. This means that we assume that the systematic deviations of model output from observations are dominated by problems in acquiring the input with sufficient accuracy. For the simple hydrosystem under study, this assumption is very plausible (see Section 5.4). However, it would be possible to use an output model that accounts for the effect of structural errors (see Section 5.5.4).

Prior rainfall model

We base the description of the prior rainfall model on a "rainfall potential", $\boldsymbol{\xi}$, that follows an Ornstein-Uhlenbeck process with mean zero and asymptotic standard deviation unity, from which we get the distribution of rainfall intensity by a transformation, h: $\mathbf{x} = h(\boldsymbol{\xi})$. In continuous-time formulation, the rainfall potential then follows the stochastic differential equation

$$\mathrm{d}\xi(t) = -\frac{\xi(t)}{\tau}\mathrm{d}t + \sqrt{\frac{2}{\tau}}\mathrm{d}W(t) \quad , \tag{2.9}$$

which is solved by a Gaussian process with conditional expectation and variance given by

$$\mathrm{E}[\xi(t) \mid \xi(t_0)] = \xi(t_0)\exp\left(\frac{t_0 - t}{\tau}\right) \quad , \quad \mathrm{Var}[\xi(t) \mid \xi(t_0)] = 1 - \exp\left(-2\frac{t - t_0}{\tau}\right) \quad . \tag{2.10}$$

The discrete time series, $\boldsymbol{\xi}$, used for our model, consists of an evaluation of this process for a discrete set of time points. The resulting probability density, $f(\boldsymbol{\xi}|\boldsymbol{\psi}_x)$, describes our prior knowledge of the rainfall potential for spatially average precipitation time series in the catchment during rainy periods.

In our case study, we estimated the parameterization of the transformation h and the choice of the parameter values from a long precipitation time series with few zeros (Figure 2). Similarly to Sigrist et al. (2012), we chose a power function, but we differentiated its coefficients for three rain intensity intervals: no rain, light rain, and heavy precipitation. The coefficients were constrained to guarantee differentiability of h over the full range of its argument. This led to the following choice:

$$x = h(\xi) = a(\xi - b)^{\alpha} + c \quad ,$$

$$\xi = h^{-1}(x) = b + \left(\frac{x - c}{a}\right)^{1/\alpha} \text{ if } a \neq 0 \quad , \tag{2.11}$$

$$\frac{\mathrm{d}h}{\mathrm{d}\xi} = a\alpha(\xi - b)^{\alpha - 1} \quad ,$$

with 3 sets of parameters for the intervals $]-\infty, \xi_1]$, $[\xi_1, \xi_2]$, and $[\xi_2, \infty[$ (see Figure 2). Note

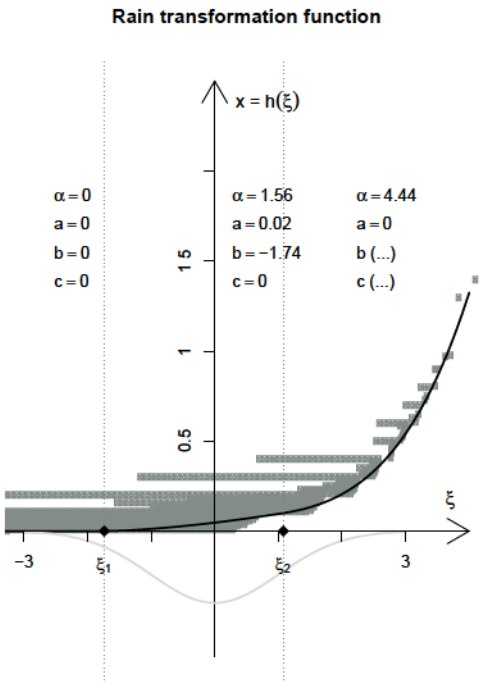

Figure 2: Empirical function to transform a normally distributed value into a rain value (see equation 2.11). The data used to parameterize the function are displayed. The values of b and c for $\xi > \xi_2$ were derived from continuity constraints for h and $\frac{dg_x}{d\xi}$ at ξ_2. The dotted vertical lines define three domains of the transformation whose parameters are indicated above.

that for $\xi \leq \xi_1$, $a = c = 0$, so that h is only invertible for $\xi > \xi_1$.
As will be shown later, the properties of this Ornstein-Uhlenbeck process are convenient for efficient posterior sampling (Sect. 5.2.2). Alternative stochastic models for the precipitation (e.g., Paschalis et al., 2013) could also be used, but they would need a different numerical approach.

Model for rainfall observations

As mentioned in the introduction, the true input to the catchment, $\mathbf{x}$, differs from the observed input, $\mathbf{x}_o$, mainly due to sampling errors caused by insufficient gauge coverage and/or the imperfect spatial interpolation scheme between gauges (McMillan et al., 2011). Similarly to Sigrist et al. (2012), our error model for input observation errors related to the distance and intrinsic inaccuracy of pluviometers is multivariate normal in the space of the "rainfall potential":

$$f(\boldsymbol{\xi}_o \mid \boldsymbol{\xi}, \psi_x) = \frac{1}{(2\pi)^{n_{\mathbf{x}_o}/2}} \frac{1}{\sqrt{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\xi}_o}(\psi_x)\right)}} \exp\left(-\frac{1}{2}(\boldsymbol{\xi}_o - \boldsymbol{\xi})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\xi}_o}(\psi_x)^{-1}(\boldsymbol{\xi}_o - \boldsymbol{\xi})\right) . \quad (2.12)$$

Due to insignificant correlations found in long time series of reference data sets (Figure S1), we parametrized the covariance matrix in equation 2.12 as:

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}_o}(\psi_x)_{i,j} = \delta_{i,j}\sigma_{\xi}^2 \quad , \quad (2.13)$$

where $\sigma_{\boldsymbol{\xi}}^2$ is the variance characterizing the deviation in rainfall potential between the observation site and the input to the catchment (Figure S2). The larger $\sigma_{\boldsymbol{\xi}}$, the less accurately we assume the observed rainfall to represent the true rainfall over the catchment.
Finally the probability distribution of rainfall given the rainfall potential is given by

$$f(\mathbf{x}_o \mid \boldsymbol{\xi}_o) = \delta\big(\mathbf{x}_o - h(\boldsymbol{\xi}_o)\big) \quad . \tag{2.14}$$

This Dirac function represents the transformation from rainfall potential into actual rainrate at the measurement station.

Numerical implementation of inference with SIP

Bayesian updating of the prior process $\boldsymbol{\Xi}$ and distributions $\boldsymbol{\Theta}$, $\boldsymbol{\Psi}_y$, and $\boldsymbol{\Psi}_x$ was based on the Markov chain Monte Carlo (MCMC) scheme proposed by Tomassini et al. (2009). In particular, we adopted a Metropolis-within-Gibbs algorithm which sequentially samples different conditional distributions while keeping the other parameters or process realizations constant. Using the index k for the elements of these Markov chains, we sequentially generate the elements $k+1$ of the different chains as outlined below. The starting point for the iterations could be obtained, for instance, by drawing a vector of parameters and a realization of the input processes from the prior distribution. The pseudo-code of this algorithm is as follows:

I. Sample a new point of the Markov chain of the hydrological model and output error parameters, ($\boldsymbol{\theta}^{k+1}$, $\boldsymbol{\psi}_y^{k+1}$), for the conditional distribution

$$f(\boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1} | \mathbf{y}_o, \mathbf{x}_o, \boldsymbol{\psi}_x^k, \boldsymbol{\xi}^k, \boldsymbol{\xi}_o^k) = f(\boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1} | \mathbf{y}_o, \boldsymbol{\xi}^k) \propto f\big(\mathbf{y}_o | \boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, h(\boldsymbol{\xi}^k)\big) \cdot f(\boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}) \tag{2.15}$$

by Metropolis sampling: draw a candidate point for ($\boldsymbol{\theta}^{k+1}$, $\boldsymbol{\psi}_y^{k+1}$) from $\mathcal{N}\big((\boldsymbol{\theta}^k, \boldsymbol{\psi}_y^k), \boldsymbol{\Sigma}_y\big)$ as the proposal (or jump) distribution with covariance matrix $\boldsymbol{\Sigma}_y$ and accept or reject this candidate by the Metropolis rule using the density (2.15). This step requires running the deterministic model.

II. Sample a new point of the Markov chain of the input error model parameters, $\boldsymbol{\psi}_x^{k+1}$, for the conditional distribution

$$f(\boldsymbol{\psi}_x^{k+1} | \mathbf{y}_o, \mathbf{x}_o, \boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, \boldsymbol{\xi}^k, \boldsymbol{\xi}_o^k) = f(\boldsymbol{\psi}_x^{k+1} | \boldsymbol{\xi}^k, \boldsymbol{\xi}_o^k) \propto f(\boldsymbol{\xi}_o^k | \boldsymbol{\xi}^k, \boldsymbol{\psi}_x^{k+1}) \cdot f(\boldsymbol{\xi}^k | \boldsymbol{\psi}_x^{k+1}) \cdot f(\boldsymbol{\psi}_x^{k+1}) \tag{2.16}$$

by Metropolis sampling: draw a candidate point for $\boldsymbol{\psi}_x^{k+1}$ from $\mathcal{N}\big(\boldsymbol{\psi}_x^k, \boldsymbol{\Sigma}_x\big)$ as the proposal distribution with covariance matrix $\boldsymbol{\Sigma}_x$ and accept or reject this candidate by the Metropolis rule using the density (2.16). The parameters of the Ornstein-Uhlenbeck process for $\boldsymbol{\xi}$ are mean zero and standard deviation unity. Furthermore, the correlation time (here: τ_ξ =10.6 min) was estimated with a long precipitation time series. Therefore, in the current application, the density $f(\boldsymbol{\xi}|\boldsymbol{\psi}_x)$ does not depend on $\boldsymbol{\psi}_x$ and cancels for the rejection rate calculation. This step does not require any hydrological model run.

III. Sample a new element of the Markov chain of the rainfall potential time series at the input observation site, $\boldsymbol{\xi}_o^{k+1}$, for the conditional distribution

$$f(\boldsymbol{\xi}_o^{k+1}|\mathbf{y}_o, \mathbf{x}_o, \boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, \boldsymbol{\psi}_x^{k+1}, \boldsymbol{\xi}^k) = f(\boldsymbol{\xi}_o^{k+1}|\mathbf{x}_o, \boldsymbol{\psi}_x^{k+1}, \boldsymbol{\xi}^k) \propto f(\mathbf{x}_o|\boldsymbol{\xi}_o^{k+1}) f(\boldsymbol{\xi}_o^{k+1}|\boldsymbol{\xi}^k, \boldsymbol{\psi}_x^{k+1}) \,. \quad (2.17)$$

For time indices, i, at which $x_{o,i} > 0$, we can directly calculate $\xi_{o,i}^{k+1} = h^{-1}(x_{o,i})$ since h^{-1} exists for arguments that are larger than zero. For time indices, i, at which $x_{o,i} = 0$, we sample $\xi_{o,i}^{k+1}$ from a normal distribution with mean zero and standard deviation unity that is truncated to values $\xi < \xi_1$.

IV. Sample a new element of the Markov chain of the rainfall potential time series for the catchment, $\boldsymbol{\xi}^{k+1}$, for the conditional distribution

$$f(\boldsymbol{\xi}^{k+1}|\mathbf{y}_o, \mathbf{x}_o, \boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, \boldsymbol{\psi}_x^{k+1}, \boldsymbol{\xi}_o^{k+1}) = f(\boldsymbol{\xi}^{k+1}|\mathbf{y}_o, \boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, \boldsymbol{\psi}_x^{k+1}, \boldsymbol{\xi}_o^{k+1}) \propto f\big(\mathbf{y}_o|\boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, h(\boldsymbol{\xi}^{k+1})\big) f(\boldsymbol{\xi}_o^{k+1}|\boldsymbol{\xi}^{k+1}, \boldsymbol{\psi}_x^{k+1}) f(\boldsymbol{\xi}^{k+1}|\boldsymbol{\psi}_x^{k+1}) \,. \quad (2.18)$$

This is similar to drawing a new element of the Markov chain of a time-dependent parameter (Tomassini et al., 2009; Reichert and Mieleitner, 2009), with the difference that we condition not only on the parameters, $(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x)$, and the observed output, $\mathbf{y}_o$, but in addition on the rainfall potential of the observed input, $\boldsymbol{\xi}_o$.

In principle, we could draw a candidate realization from the Ornstein-Uhlenbeck process $f(\boldsymbol{\xi}^{k+1}|\boldsymbol{\psi}_x^{k+1})$, and use the two other factors in equation (2.18), $f\big(\mathbf{y}_o|\boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, h(\boldsymbol{\xi}^{k+1})\big)$ and $f(\boldsymbol{\xi}_o^{k+1}|\boldsymbol{\xi}^{k+1}, \boldsymbol{\psi}_x^{k+1})$, for the calculation of the rejection ratio of Metropolis sampling. However, in that case we would not profit from what we learned in the past (up to step k) about the posterior of $\boldsymbol{\xi}$. This would lead to a very low acceptance rate. To better profit from what we already learned up to step k, we do this only for pieces of the time series of $\boldsymbol{\xi}^{k+1}$, keeping the remainder of the time series at their previous values, $\boldsymbol{\xi}^k$, for the intervals that were not yet updated, or their new values, $\boldsymbol{\xi}^{k+1}$, for the intervals that were already updated (Figure S3). This leads to the following algorithm for the construction of $\boldsymbol{\xi}^{k+1}$ from $\boldsymbol{\xi}^k$:

(a) Divide the calibration period and the previous realization of the rainfall potential, $\boldsymbol{\xi}^k$, into m subintervals of similar length (using random disturbances to avoid that the interval boundaries are the same in successive steps). Denote with $\boldsymbol{\xi}_l^k = \boldsymbol{\xi}^k\Big|_{I_l}$ the restrictions of $\boldsymbol{\xi}^k$ to the subinterval I_l.

(b) Repeat the following 3 substeps $\forall\ l = 1, ..., m$ subintervals I_l in order to draw $\boldsymbol{\xi}^{k+1}$, a sample of the updated $\boldsymbol{\Xi}$. The more substeps are considered, the more hydrological model runs will be required for one iteration.

   i. Draw a candidate sample $\boldsymbol{\xi}_l^{k+1'}$ over the subinterval I_l from an Ornstein-Uhlenbeck process conditional on the values of $\boldsymbol{\xi}^k$ at its start time point, s,

and its end point, u. This conditioning guarantees that replacing the current sample over the subinterval I_l by the candidate still guarantees continuity of the process over the full time domain. Conditional mean and variance of this process is given by

$$\begin{aligned}\mathrm{E}\big[\Xi(t)|\xi(s),\xi(u)\big] = &\frac{\exp(-(t-s)/\tau_\xi)\big[1-\exp(-2(u-t)/\tau_\xi)\big]}{1-\exp(-2(u-s)/\tau_\xi)}\,\xi(s)\\ &+\frac{\exp(-(u-t)/\tau_\xi)\big[1-\exp(-2(t-s)/\tau_\xi)\big]}{1-\exp(-2(u-s)/\tau_\xi)}\,\xi(u)\end{aligned} \tag{2.19}$$

$$\mathrm{Var}\big[\Xi(t)|\xi(s),\xi(u)\big] = \frac{\big[1-\exp(-2(u-t)/\tau_\xi)\big]\big[1-\exp(-2(t-s)/\tau_\xi)\big]}{1-\exp(-2(u-s)/\tau_\xi)}\,. \tag{2.20}$$

Then, replace the current sample (that may have already been modified on previous intervals) by the candidate, $\boldsymbol{\xi}_l^{k+1'}$, in the interval I_l. We denote this candidate sample over the full time domain by $\boldsymbol{\xi}_{<l}^{k+1} \cup \boldsymbol{\xi}_l^{k+1'} \cup \boldsymbol{\xi}_{>l}^{k}$.

ii. Compute the acceptance probability, r, of this candidate sample according to:

$$r = \min\left[1, \frac{f(\boldsymbol{\xi}_o|\boldsymbol{\xi}_{<l}^{k+1} \cup \boldsymbol{\xi}_l^{k+1'} \cup \boldsymbol{\xi}_{>l}^{k}, \boldsymbol{\psi}_x^{k+1}) f\big(\mathbf{y_o}|\boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, h(\boldsymbol{\xi}_{<l}^{k+1} \cup \boldsymbol{\xi}_l^{k+1'} \cup \boldsymbol{\xi}_{>l}^{k})\big)}{f(\boldsymbol{\xi}_o|\boldsymbol{\xi}_{<l}^{k+1} \cup \boldsymbol{\xi}_l^{k} \cup \boldsymbol{\xi}_{>l}^{k}, \boldsymbol{\psi}_x^{k+1}) f\big(\mathbf{y}_o|\boldsymbol{\theta}^{k+1}, \boldsymbol{\psi}_y^{k+1}, h(\boldsymbol{\xi}_{<l}^{k+1} \cup \boldsymbol{\xi}_l^{k} \cup \boldsymbol{\xi}_{>l}^{k})\big)}\right]. \tag{2.21}$$

This part requires running the hydrological model which might be time consuming. Differently from Tomassini et al. (2009), in this acceptance ratio, the rainfall observations are considered in the form of probability density of the rainfall potential at the observation site, $f(\boldsymbol{\xi}_o|\boldsymbol{\xi}, \boldsymbol{\psi}_x)$, in addition to the probability density for the observed output, $f(\mathbf{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}_y, h(\boldsymbol{\xi}))$.

iii. Set $\boldsymbol{\xi}_l^{k+1} = \boldsymbol{\xi}_l^{k+1'}$, i.e. accept $\boldsymbol{\xi}_l^{k+1'}$, with probability r, otherwise set $\boldsymbol{\xi}_l^{k+1} = \boldsymbol{\xi}_l^{k}$, i.e. reject $\boldsymbol{\xi}_l^{k+1'}$. This piecewise updating of $\boldsymbol{\Xi}$ ensures a much higher efficiency than drawing $\boldsymbol{\xi}^{k+1}$ over the full time domain and comparing it with $\boldsymbol{\xi}^{k}$.

(c) After having completed these m substeps, set $\boldsymbol{\xi}^{k+1} = \boldsymbol{\xi}_1^{k+1} \cup ... \cup \boldsymbol{\xi}_m^{k+1}$ and move to the next iteration (step 1 above).

After having repeated these steps 1 - 4 of the MCMC algorithm to convergence, we obtain a sample of the joint posterior distribution and input processes $f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \boldsymbol{\xi}, \boldsymbol{\xi}_o|\mathbf{y}_o, \mathbf{x}_o)$. The posterior of the parameters only can be gained through marginalization: $f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x|\mathbf{y}_o, \mathbf{x}_o) = \int f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \boldsymbol{\xi}, \boldsymbol{\xi}_o|\mathbf{y}_o, \mathbf{x}_o)\mathrm{d}\boldsymbol{\xi}\mathrm{d}\boldsymbol{\xi}_o$. A sample from this distribtution is obtained from the sample of $f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \boldsymbol{\xi}, \boldsymbol{\xi}_o|\mathbf{y}_o, \mathbf{x}_o)$ by disregarding the information on $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_o$.

5.2.3 Predictions in the calibration and validation periods

Once having a statistically calibrated model, we are usually interested to quantify our knowledge of the true system output, $\mathbf{y}$. This is done by calculating

$$f(\mathbf{y}^{L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) = \int f(\mathbf{y}^{L_2} \mid \boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}_y \mathrm{d}\boldsymbol{\psi}_x \ , \quad (2.22)$$

where the superscripts L_1 and L_2 indicate that we may be interested in predictions for another time period (here: "layout" (L)), L_2, than we have observations, L_1. As in most studies on inference and uncertainty analysis, we still assume input data to be available also in L_2 and thus operate in "prediction" or "hindcasting" mode (Renard et al., 2011; Del Giudice et al., 2015a). When evaluating the quality of the models, we want to compare observations of the system by predicted observations. This requires us to predict our knowledge of observations, $\mathbf{y}_o^{L_2}$, rather than our knowledge of the true output, $\mathbf{y}^{L_2}$:

$$f(\mathbf{y}_o^{L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) = \int f(\mathbf{y}_o^{L_2} \mid \boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}_y \mathrm{d}\boldsymbol{\psi}_x \ . \quad (2.23)$$

The essential difference between the equations (2.22) and (2.23) is that the latter also considers output observation error, usually in the form of iid Gaussian noise.
In the following subsections we describe the specifics of predicting $\mathbf{y}^{L_2}$ and $\mathbf{y}_o^{L_2}$ with the 4 different methods described here. In addition, for the technique RM and SIP which infer the rainfall input also, we will discuss the formulation of our posterior knowledge of the rainfall.

Predictions with alternative methods

Predictions with LS As the traditional least squares approach assumes that the uncertainty in the system output predictions only arises from incomplete knowledge about model parameters, its predictive distribution can thus be obtained by propagating the posterior of the model parameters:

$$\mathbf{Y}^{L_2} = \mathbf{y}_M^{L_2}(\boldsymbol{\Theta}_{\text{post}}^{L_1}, \mathbf{x}) \quad . \quad (2.24)$$

For the prediction of observations, the observation error, $\mathbf{E}$, must be added on the g-transformed scale:

$$\mathbf{Y}_o^{L_2} = g^{-1}\Big(g\big(\mathbf{y}_M^{L_2}(\boldsymbol{\Theta}_{\text{post}}^{L_1}, \mathbf{x})\big) + \mathbf{E}^{L_2}(\boldsymbol{\Psi}_{y,\text{post}}^{L_1})\Big) \quad (2.25)$$

for time points in L_2 that are not identical to time points in L_1, where we know the observed output. Note that according to the model assumption (2.2), the observed input is used for $\mathbf{x}$ in these equations.
Numerically, a sample of $\mathbf{Y}^{L_2}$ is generated by propagating the parameter sample through the deterministic model, $\mathbf{y}_M^{L_2}$. To generate a sample of $\mathbf{Y}_o^{L_2}$, sample points of the normal distribution of $\mathbf{E}^{L_2}$ with the corresponding sample points of $\boldsymbol{\Psi}_{y,\text{post}}^{L_1}$ must be added on the transformed scale and the sum transformed back to the original scale as indicated in equation (2.25).

Predictions with BD The bias description approach assumes that the uncertainty in the system output predictions arises from incomplete knowledge about model parameters and from

input and structural errors. Thus, our best knowledge of the true system output needs consideration of the bias (and the transformation g):

$$\mathbf{Y}^{L_2} = g^{-1}\Big(g\big(\mathbf{y}_M^{L_2}(\mathbf{\Theta}_{\text{post}}^{L_1}, \mathbf{x}) + \mathbf{B}_{\text{post}}^{L_2}(\mathbf{\Psi}_{y,\text{post}}^{L_1})\Big) \quad . \tag{2.26}$$

For the prediction of observations, the observation error, $\mathbf{E}$, must be added in addition to the bias (on the transformed scale also):

$$\mathbf{Y}_o^{L_2} = g^{-1}\Big(g\big(\mathbf{y}_M^{L_2}(\mathbf{\Theta}_{\text{post}}^{L_1}, \mathbf{x})\big) + \mathbf{B}_{\text{post}}^{L_2}(\mathbf{\Psi}_{y,\text{post}}^{L_1}) + \mathbf{E}^{L_2}(\mathbf{\Psi}_{y,\text{post}}^{L_1})\Big) \quad . \tag{2.27}$$

Again, the observed input (2.2) is used in these equations, as the effect of input errors to the output is corrected in the output by the additive term $\mathbf{B}$. As the distributions of $\mathbf{B}_{\text{post}}$ and $\mathbf{E}$ conditional on their parameters are normal (with expectation and variance given by equations 35 - 38 in Reichert and Schuwirth (2012)), we can again propagate the posterior sample of the model parameters through these equations and sample from the corresponding normal distributions to get a posterior sample of (2.26) and (2.27), respectively.

Predictions with RM The rainfall multipliers approach assumes that the uncertainty in the system output predictions arises from incomplete knowledge about model parameters and from input imprecision. The output of the system is assumed to be equal to the model output forced with uncertain input $\mathbf{y}_M(\mathbf{\Theta}, \mathbf{X})$. As for each storm event, j, the uncertain input is equal to the observed input times a factor β_j. Compared to the LS approach, this leads to the expansion of the parameter vector. The predictions are thus still given by the equations (2.24) and (2.25) with the exception that the use of the observed input (2.2) is replaced by the input given by equation (2.6). If the time points to quantify the posterior knowledge extend to storm events that have not been used for calibration, our knowledge of the rainfall multiplier is described by a hierarchical model based on a conditional lognormal distribution with mean unity, the standard deviation of which is distributed according to the posterior of the parameter $\sigma_{\text{post}}^{\beta}$.

In addition to the posterior of the model output, the RM technique provides a posterior estimate of the rainfall given by (see equation 2.6)

$$X_i = \beta_{j(i)} x_{o,i} \tag{2.28}$$

with $\beta_{j(i)}$ as defined above.

The numerical implementation is again similar to the LS approach. For storm events included into the calibration phase, the rainfall multiplier β_j is part of the parameter sample, for rainfalls not included, it is drawn from a lognormal distribution with mean unity and standard deviation $\sigma_{\text{post}}^{\beta}$.

Predictions with SIP

Our approach of using a stochastic input process assumes that the uncertainty in the system output predictions arises from incomplete knowledge about model parameters and from input imprecision. The distributions representing our knowledge of true and observed output are given by considering an additional integration over the rainfall potentials, $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_o$, in the equations (2.22) and (2.23), and eliminating arguments that are not relevant. This leads to

$$f(\mathbf{y}^{L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) = \int f\big(\mathbf{y}^{L_2} \mid \boldsymbol{\theta}, \boldsymbol{\psi}_y, h(\boldsymbol{\xi}^{L_1 \cup L_2})\big) \cdot f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \boldsymbol{\xi}^{L_1 \cup L_2}, \boldsymbol{\xi}_o^{L_1 \cup L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}_y \mathrm{d}\boldsymbol{\psi}_x \mathrm{d}\boldsymbol{\xi}^{L_1 \cup L_2} \mathrm{d}\boldsymbol{\xi}_o^{L_1 \cup L_2} \quad (2.29)$$

and

$$f(\mathbf{y}_o^{L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) = \int f\big(\mathbf{y}_o^{L_2} \mid \boldsymbol{\theta}, \boldsymbol{\psi}_y, h(\boldsymbol{\xi}^{L_1 \cup L_2})\big) \cdot f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \boldsymbol{\xi}^{L_1 \cup L_2}, \boldsymbol{\xi}_o^{L_1 \cup L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}_y \mathrm{d}\boldsymbol{\psi}_x \mathrm{d}\boldsymbol{\xi}^{L_1 \cup L_2} \mathrm{d}\boldsymbol{\xi}_o^{L_1 \cup L_2} \,. \quad (2.30)$$

The posterior of our knowledge of the true rainfall over the catchment is given by

$$f(\mathbf{x}^{L_1 \cup L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) = \int f(\mathbf{x}^{L_1 \cup L_2} \mid \boldsymbol{\xi}^{L_1 \cup L_2}) \cdot f(\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x, \boldsymbol{\xi}^{L_1 \cup L_2}, \boldsymbol{\xi}_o^{L_1 \cup L_2} \mid \mathbf{y}_o^{L_1}, \mathbf{x}_o^{L_1 \cup L_2}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\psi}_y \mathrm{d}\boldsymbol{\psi}_x \mathrm{d}\boldsymbol{\xi}^{L_1 \cup L_2} \mathrm{d}\boldsymbol{\xi}_o^{L_1 \cup L_2} \,. \quad (2.31)$$

Numerically, this requires us to do the inference over the combined time frame, $L_1 \cup L_2$, but only using output observations from L_1. A sample for the true output (2.29) is then obtained by propagating the posterior sample components corresponding to $\boldsymbol{\theta}$, $\boldsymbol{\psi}_y$ and $\boldsymbol{\xi}$ through the model $\mathbf{y}_M\big(\boldsymbol{\theta}, \boldsymbol{\psi}_y, h(\boldsymbol{\xi})\big)$. For the sample for (2.30), we have to add sample points of the normal distribution of observation errors with the corresponding parameters $\boldsymbol{\psi}_x$ and considering output transformation as in equation (2.25). Finally, a sample of the smoothing distribution of the true input, $\mathbf{x}$, is obtained by applying the transformation $\mathbf{x} = h(\boldsymbol{\xi})$ to the sample of $\boldsymbol{\xi}$.

5.2.4 Rainfall scenarios

To test the performances of the SIP method compared to the other three error descriptions we considered 2 typical scenarios of rainfall data availability. Scenario Sc1 uses as input the rainfall recorded by 2 of our own pluviometers located in the direct vicinity of the catchment (Section 5.3.3). Using the highly-representative high-resolution data from these gauges is an illustrative example of the best case scenario of input data availability. Scenario Sc2 uses as input the rainfall recorded by a pluviometer managed by the Swiss meteorological office (Section 5.3.3). Using data from this less-representative gauge is a typical example of (suboptimal) input data availability. In this study we focus on point-scale pluviometers, since they still are the most common source of rainfall measurements (McMillan et al., 2011).

5.2.5 Performance assessment

An optimal error description should produce posterior model parameters which are highly representative of the average conditions of the physical system (Vrugt et al., 2008; Brynjarsdóttir and O'Hagan, 2014). Furthermore, it should ensure reliable (i.e. with high coverage of data), accurate (i.e. on average close to the data or unbiased), and precise (i.e. sharp or with low dispersion) predictions, especially in the extrapolation domain. For this reason we inspected the following factors:

I. **Correctness of the updated parameters $\boldsymbol{\theta}, \boldsymbol{\psi_y}, \boldsymbol{\psi_x}$.** To assess the physical realism of

the estimated parameters, we compare the posterior marginals with the priors and with the posterior marginals of the other scenario and likelihood functions.

II. **Prediction accuracy**. As a measure of model adequacy, we calculate the Nash-Sutcliffe efficiency at the maximum of the posterior, NS. The closer this coefficient is to 1, the better the model fits the data, especially during high-flow periods (Reichert and Mieleitner, 2009; Coutu et al., 2012b).

III. **Prediction reliability**. We analyze the data coverage of the 95% interquantile intervals. If the percentage of data points falling into these total uncertainty bands is larger or equal to 95, we considered the predictions to be reliable (Wang et al., 2012; Li et al., 2012).

IV. **Integrated predictive performance**. We adopt two metrics to quantify how reliable, accurate, and precise predictions are. The first is the interval (skill) score (Gneiting and Raftery, 2007), S_α^{int}:

$$S_\alpha^{int} = \sum_{i=1}^{n} \left[\left(q_i^{1-\alpha/2} - q_i^{\alpha/2}\right) + \frac{2}{\alpha}\left(q_i^{\alpha/2} - \mathrm{y}_{o,i}\right) H\left(q_i^{\alpha/2} - \mathrm{y}_{o,i}\right) + \frac{2}{\alpha}\left(\mathrm{y}_{o,i} - q_i^{1-\alpha/2}\right) H\left(\mathrm{y}_{o,i} - q_i^{1-\alpha/2}\right) \right] \quad (2.32)$$

where α corresponds to the confidence level (set to 0.05), n is the number of timesteps within the considered period, q_i^α is the α-quantile of the predictive distribution at time point i, and $\mathrm{y}_{o,i}$ is the observation at time i. H denotes the unit step function which takes the value of 1 if its argument is greater than 0 and 0 otherwise. The better the quality of the predictions, the closer to 0 this statistics is. Second, we also create predictive quantile-quantile plots, which analyze the probability of the observations being distributed as the model output (including all uncertainties). The more reliable and precise the predictive distribution is, the closer to the identity line the observed p-values are (Renard et al., 2011).

5.3 Materials

To demonstrate the relevance of our method in hydrology, we tested it in a urban catchment with real observations. We here describe the analyzed sewershed, the conceptual rainfall-runoff hydrological model adopted, the pluviometric and discharge data used, and the prior distribution of the parameters.

5.3.1 System

The test case hydrosystem is a small partially-combined sewer network located in Adliswil in the proximity of Zurich, Switzerland (Figure 3). The watershed surface is about 28.6 hectares, only a fraction of which contributes to the stormwater outflow. The area is characterized by medium density of housing and a slope of about 8.7 %.

5.3.2 Hydrological model

The hydrological model of the hydrosystem consists of two components, one for the stormwater runoff and the other for the produced wastewater. This parsimonious modeling concept is akin to the one adopted by Del Giudice et al. (2015a).

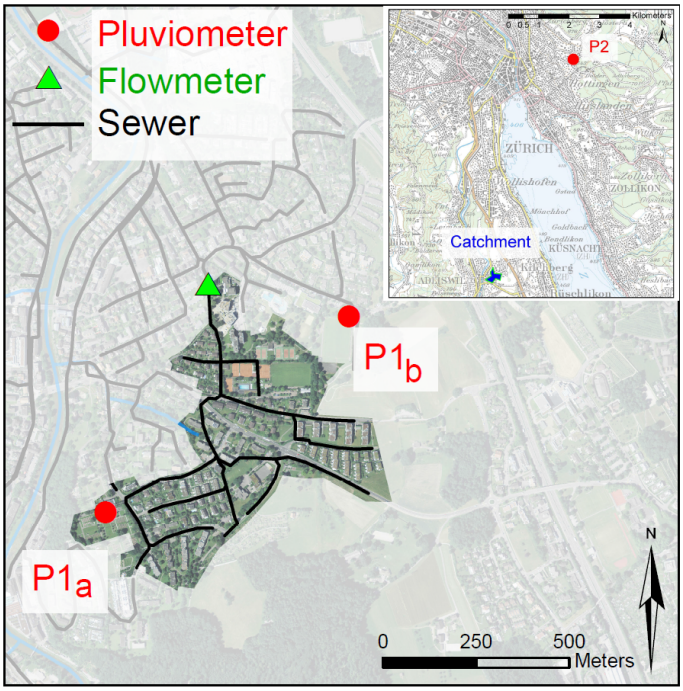

Figure 3: Map of the study area. The urban catchment and monitoring sites for input and output measurements are represented. Data from $P1_a$ and $P1_b$ are combined to provide the input of Sc1, while data from P2 are used as input in Sc2.

The stormwater runoff is modelled by a linear reservoir which is alimented by a time-varying precipitation input, $\mathbf{x}$, and a constant infiltration rate from groundwater, x_{gw}. The dynamics of this compartment is described by the following ODE which can be solved analytically:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = A \cdot x(t) + x_{gw} - \frac{s(t)}{k} \quad , \tag{3.33}$$

where A is area contributing to the rainfall-runoff and k is the mean residence time in a virtual reservoir.

The produced wastewater during dry weather is described by the harmonic function:

$$w(t) = \sum_{i}^{2} (\varsigma_i sin\frac{i2\pi t}{24} + \chi_i cos\frac{i2\pi t}{24}) \quad , \tag{3.34}$$

with ς_1, ς_2, χ_1, and χ_2 representing the coefficients of the trigonometrical series. The combined discharge at the outlet of the system is modeled by the superposition of the storm- and wastewater:

$$y_M(t) = \frac{s(t)}{k} + w(t) \quad . \tag{3.35}$$

The model, as well as the analyses, have been coded in the statistical programming language `R` (R Core Team, 2013).

5.3.3 Dataset

The measurements of precipitation and discharge were performed from June to November 2013. We recorded the rainfall data with two weighing gauges ($P1_a$ and $P1_b$ in Figure 3). Averaging the

recordings ($\Delta t = 1$ min) from those gauge (OTT Pluvio^2), we derived the input to the system in scenario Sc1 (Section 5.2.4). As input in scenario Sc2 we instead used rainfall observed ($\Delta t = 10$ min) at the pluviometric station of Zurich-Fluntern (P2 in Figure 3), which belongs to the network of the Swiss meteorological office (`www.hw.zh.ch/hochwasser/foto/DB%20SMA.pdf`). Precipitation data from these 3 and other 4 stations located around the catchment (not shown) were analyzed to parameterize the function transforming the standard OU-process into precipitation (see Figure 2) and the prior of the "rainfall potential" (Figures S1 and S2).
Wastewater flow was measured ($\Delta t = 4$ min) at the outlet of the catchment by a radar-based contact-free sensor (Flo-Dar 4000 SR). From the recorded data we selected 3 events for calibration and 2 for validation (see Results). These storms were characterized by a significant response of the system (i.e. maximal flowrate one order of magnitude larger than during dry-weather) and negligible infiltration from groundwater. Being separated by several days in between, the individual flood events can be considered as independent.

5.3.4 Prior distributions

The marginal prior distributions of the hydrological model and error model parameters are given in Table 1. We estimated the prior marginals of A, k, x_{gw}, ς_1, ς_2, χ_1, χ_2, and σ_E based on a least-squares calibration employing measurements not used in the final analysis (data not shown). Having an interpretation related to the hydrological system and measurement device, from now on we refer to these constants as "physical parameters". Regarding the priors of the bias error model (BD), we followed the guidelines provided in previous works (Reichert and Schuwirth, 2012; Brynjarsdóttir and O'Hagan, 2014; Del Giudice et al., 2015a). For σ_B, the magnitude of the bias, we determined its prior standard deviation by analyzing the model discrepancy between the model forced with accurate rainfall and the measured discharge. The prior expected value of τ, the bias autocorrelation time scale, was set approximately equal to 1/3 the duration of the falling limb of a storm hydrograph. As for the other parameters, the priors of the bias were based on analyses of events independent from those used for calibration and validation. In the multiplicative error model (RM), similarly to Sikorska et al. (2012b), we assumed a priori no bias in the rainfall measurements and estimated the mean standard deviation of the input uncertainty, σ^β, to 10%.

5.4 Results

In general, all error models performed similarly well in the best rainfall scenario (Sc1). However, when using less accurate precipitation measurements (Sc2), SIP and BD provided the most realistic parameter estimates and reliable output predictions for both the calibration and the extrapolation phase. SIP additionally corrected the precipitation and quantified its uncertainty. The rainfall multipliers, instead were not able to represent missing peaks accurately and, in most cases, underestimated input uncertainty. Calibration with SIP, however, took 10-100 times longer than with other inference schemes (Figure S12). The outcomes of the case study application are discussed more extensively in the following paragraphs.

5.4.1 Estimated parameters during calibration

In the scenario with optimal rainfall data (Sc1), inferred hydrological model parameters had a similar distribution for all error representations (upper row in Figure 4). Very little model bias was identified (low posterior σ_B), a condition confirmed by the RM error model which did not display an increase in σ^β. Scenario Sc2 with lower input data quality induced different

Table 1: Hydrological model and error model calibration parameters ($\boldsymbol{\theta}, \boldsymbol{\psi}_y, \boldsymbol{\psi}_x$). The notation for prior distributions is: LN(μ, σ): lognormal, TN(μ, σ, a_1, a_2): truncated normal. The symbol meaning is: μ: expected value, σ: standard deviation, a_1: lower limit, a_2: upper limit.

Symbol	Description	Units	Prior
A	Area contributing to outflow	m^2	LN(11815.8, 1181.6)
k	water residence time	hr	LN(0.079, 0.016)
x_{gw}	groundwater infiltration	l/s	LN(2.05, 0.013)
$-\varsigma_1$	trigonometric coefficient of the sewage flow	-	LN(0.25, 0.094)
$-\varsigma_2$	trigonometric coefficient of the sewage flow	-	LN(0.84, 0.019)
$-\chi_1$	trigonometric coefficient of the sewage flow	-	LN(0.68, 0.019)
χ_2	trigonometric coefficient of the sewage flow	-	LN(0.077, 0.01)
σ_E	Standard deviation of $\mathbf{E}$	g(l/s)	LN(4.1 $\left.\frac{\mathrm{d}g}{\mathrm{d}y}\right\|_{50}$, 0.41 $\left.\frac{\mathrm{d}g}{\mathrm{d}y}\right\|_{50}$)
σ_B	Standard deviation of $\mathbf{B}$	g(l/s)	TN(0, 3.77$\left.\frac{\mathrm{d}g}{\mathrm{d}y}\right\|_{50}$, 0, ∞)
τ	Correlation length of $\mathbf{B}$	hr	LN(0.47, 0.047)
β_j	Rainfall multiplier for the calibration event j	-	LN(1, σ^β)
σ^β	Standard deviation of the multipliers	-	LN(0.1, 0.02)
$\sigma^2_{\boldsymbol{\xi}}$	Variance between rainfall potentials	-	LN(0.4, 0.2)

performances of the error models. While with BD and SIP posterior physical parameters were similar among them and to the other scenario, with LS and RM those posteriors differed from the scenario Sc1 (second row in Figure 4). The most affected parameters were those related to the hydrologic response time of the catchment (k) and to the output measurement errors (σ_E), both of which increased dramatically. The other 2 parameters common to all error models were less affected by the inaccurate rain. In particular, x_{gw}, representing infiltration from groundwater, was not affected by inference scheme or rainfall data quality. The effective impervious area, A, was also only mildly altered by the bad input data. This appears to be connected to the rainfall characteristics of both scenarios (Figure 5) which, despite showing different temporal behavior, have similar volumes. This is in agreement also with the only slight deviations of the multipliers β_1, β_2, β_3 from unity, even with the worse rain (Sc2).

An analysis of the error-model-specific parameters (bottom two rows in Figure 4) shows that, as expected, all parameters related to the amount of input uncertainty increase when comparing Sc1 with Sc2 . In particular, with less accurate rain, more output bias is detected (higher σ_B) and more uncertainty is identified with SIP (higher $\sigma^2_{\boldsymbol{\xi}}$). With RM, instead, the increased input uncertainty is barely recognized (σ^β hardly increases).

5.4.2 Estimated input and output during calibration

Forced with accurate rain (Sc1), the chosen hydrological model fitted the calibration data accurately (Nash-Sutcliffe efficiency around 0.9 with all error models, Figure 5, first row). Predictions were also reliable for all error descriptions (data coverage around 95%) and sharp (i.e. precise). Estimated input uncertainty with RM and SIP was also very low. In the more realistic scenario Sc2, however, a substantial distinction among error models becomes evident (Figure 5, second row, and Figures S13-S16). Both LS and RM significantly increase output uncertainty producing unrealistically wide (i.e. imprecise) prediction intervals while still missing the mis-recorded

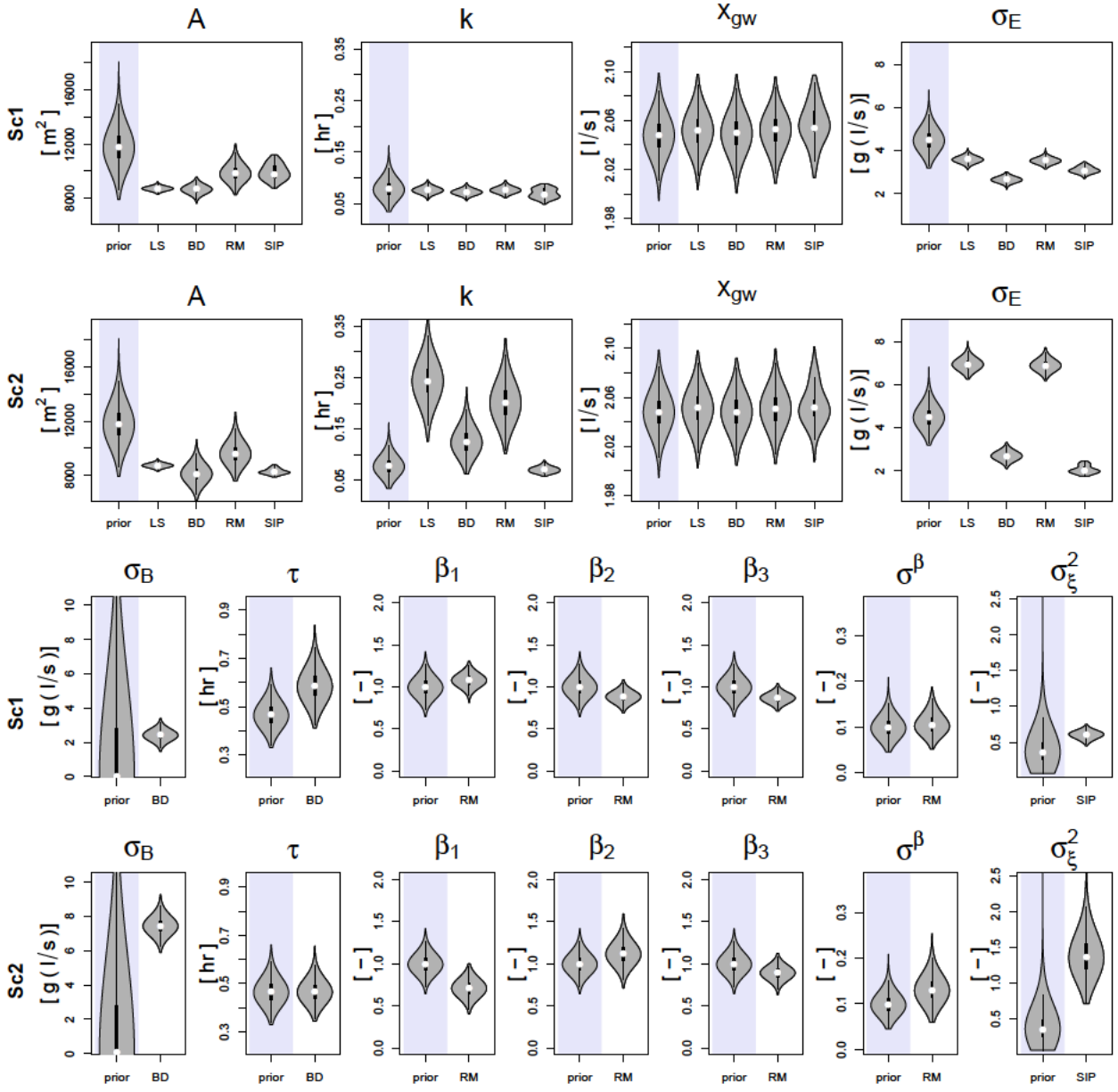

Figure 4: Marginal prior and posterior distributions of physical parameters common to all error models (first 2 rows) and of parameters typical of each error model (bottom 2 rows). Comparing results between Sc1 (accurate rain) and Sc2 (faulty rain) shows the corrupting effects of input uncertainties on the physical parameters and should be detected by the error model parameters. See Table 1 for an explanation of the symbols.

rainfall peaks. Instead, BD was able to effectively assimilate the deviating flow data and in this way correct model output in a reliable and precise way. The most interesting result, however, was produced by SIP. Not only were the output predictions the most accurate and precise among the 4 cases but the rainfall was also realistically estimated for Sc2. As shown in Figures 6 and S21, even when using highly biased input data, SIP produced estimates very close to the optimal data (plotted for comparison). The rainfall multipliers, instead, were unable to deal with these dynamic input biases.

5.4.3 Estimated input and output during extrapolation

In the validation period, when using only input data but not output data, BD and SIP produced the most reliable flow predictions (coverage close to 95%), regardless of data quality (Figure 5, bottom 2 rows, and Figures S17-S20). When using the SIP technique, however, the hydrological model produced slightly less accurate results than with the other methods. Regarding the input, SIP realistically allowed for more uncertainty compared to RM. Contrary to the calibration phase, differences among the error models are visible in both rainfall scenarios. With the most accurate and precise rainfall (Sc1), the bias description performed best, especially in terms of accuracy and precision during low flows. Regarding rainfall, SIP allowed for slightly more uncertainty than RM, especially during the moments of maximum intensity. Using rainfall data from the less representative pluviometer (Sc2) helped to further differentiate the input estimates of the two error models. While none was able to account for missing peaks without considering output information, SIP was substantially less overconfident than RM (Figure S21). Concerning the flow predictions with inaccurate rainfall, SIP appears to have generated realistic uncertainty bands during high flows. It was also more precise than the other methods during low flow events. All-in-all, the BD method dominated in accuracy and reliability, although it slightly overestimated predictive uncertainty (Figure S22).

5.5 Discussion

5.5.1 Interpreting posterior parameters, input, and output

As expected theoretically, and as confirmed by the results of the case study, describing input uncertainty in a realistic way helps to protect model parameters from shifting to unrealistic values (Kuczera et al., 2006; Vrugt et al., 2008; Li et al., 2012). Parameters estimated with SIP using erroneous input data were very close to their most realistic value obtained using the best rain data (Figure 4). This means that SIP avoided the compensation of input errors by shifts in model parameters. Instead, with simpler error models like LS and RM, some parameters were forced to unrealistic posteriors to help the model fit flow data notwithstanding the erroneous forcing. This occurs frequently in hydrological modeling where input data errors can corrupt model parameters away from their original meaning as catchment or measurement characteristics (Renard et al., 2011). Interestingly, the bias description had a similar parameter preserving effect as SIP. This robustness of BD, only speculated in previous studies (Bayarri et al., 2007; Del Giudice et al., 2015a), was confirmed by our study thanks to the comparison of the "unbiased model" (Sc1) with a "biased model" (Sc2). In other words, both SIP and BD helped alleviate the (distorting or overtuning) impact of errors in the regressor, i.e. the areal precipitation. The ability of BD to infer parameters close to their physical value even in the presence of input errors is very promising. This is probably due to a combination of two reasons. First, our chosen bias process and its parameters are very reasonable (see Section

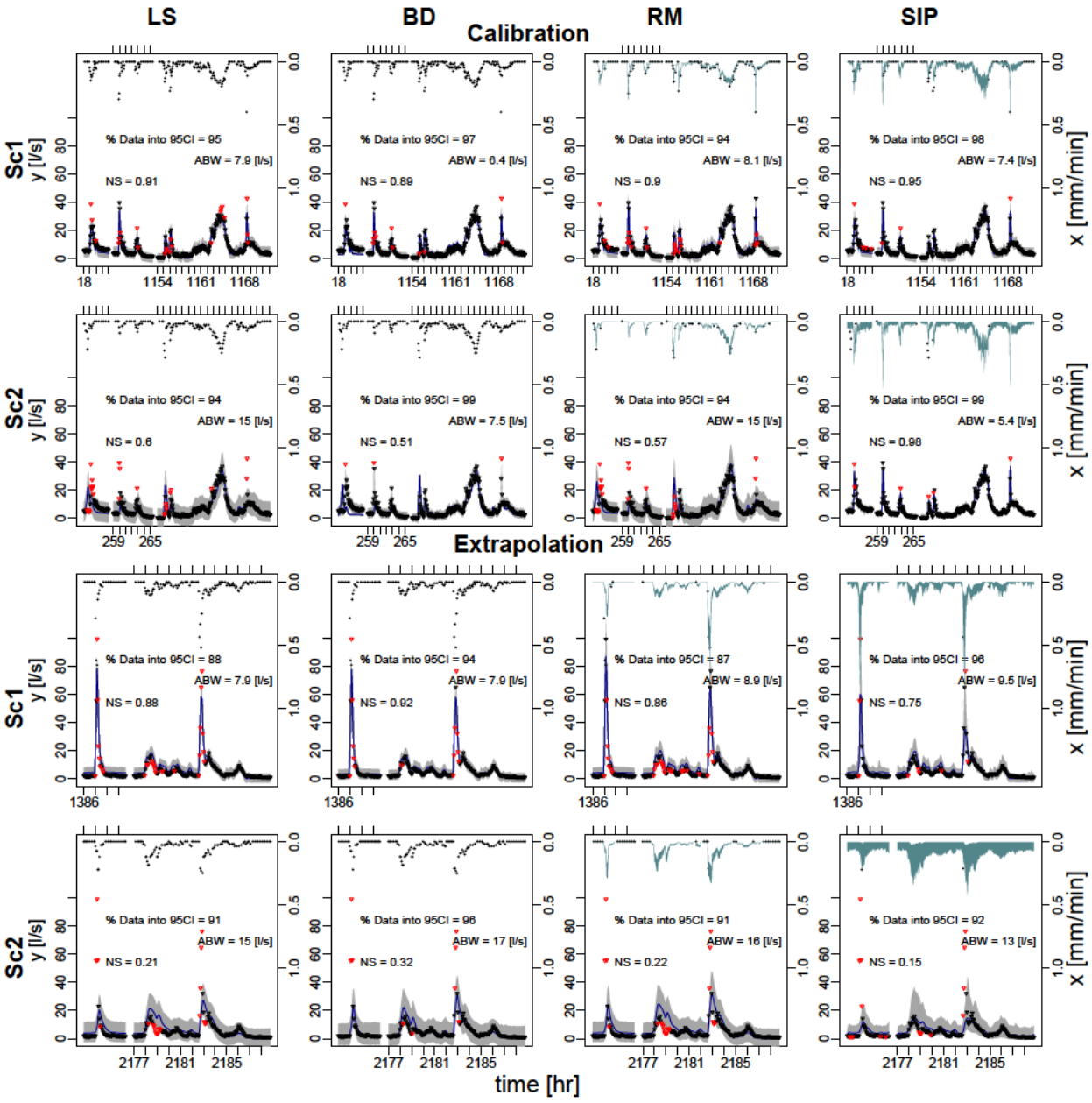

Figure 5: Total output (bottom of each frame) and input uncertainty (top of each frame) in the calibration and validation phases with all error models and rainfall scenarios. The points represent measured data. Red triangular dots indicate runoff data not included within the 95% output credible intervals. Predictions are considered reliable if red dots are $\leqslant 5\%$ of the total. ABW is the average band width. NS expresses the accuracy of the model median (blue line). While LS and BD do not assess input uncertainty, RM and SIP do. Uncertainty intervals are generally wider in extrapolation, since the runoff data there are only used for a posteriori validation.

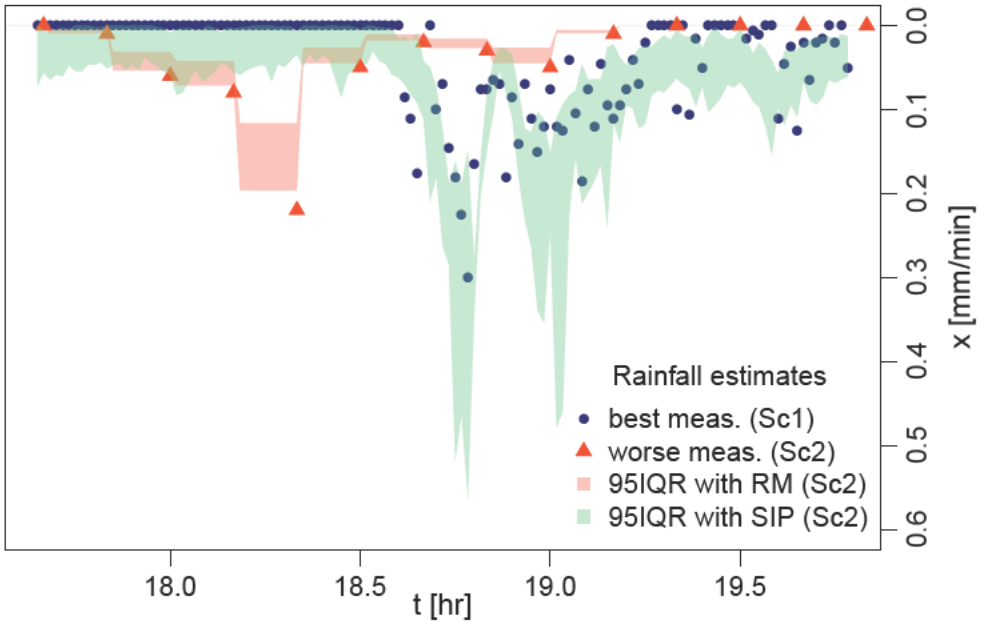

Figure 6: Zoom of the estimated input uncertainty with RM and SIP using the less representative rainfall data for a storm event in the calibration period (Figure 5, second row, fourth column). The accurate data from Sc1 are only used for a posteriori validation. Whole-catchment precipitation inferred with SIP is substantially more realistic than the one with RM. SIP, as a continuous-time stochastic process, is better able to learn from the flow data.

5.3.4 and Brynjarsdóttir and O'Hagan (2014)). Second, our runoff model appeared to represent the hydrological processes sufficiently well (see Figure 5). SIP and BD also provided similarly reliable and precise smoothing estimates, i.e. predictions in the calibration phase, for the output even in Sc2. The mechanisms behind the two methods, however, are different. SIP avoided parameter overtuning by flexibly adapting the input process and therefore adjusting the biased model at its source. BD instead avoided parameter compensation by not forcing the model to fit the data but rather allowing the autocorrelated discrepancy term to bridge the gap between model output and data.

In contrast to our expectations, considering storm-dependent rainfall multipliers did not improve the calibration of physical parameters compared to the simplest least squares approach. This is probably linked to the fact that we are analyzing a rapidly-reacting catchment with more detailed data than what is normally available for natural systems (Beuchat et al., 2011). This can explain why our results with RM differ from those of studies conducted at coarser time resolutions (Vrugt et al., 2008; Li et al., 2012) or with small and non-systematic input errors (Sun and Bertrand-Krajewski, 2013).

As observed in Figures 5 and 6, model input is estimated much more realistically by the SIP method than by RM. Both methods learn from the output about the input dynamics. Discharge data integrate rainfall-runoff processes over the entire catchment (Frey et al., 2011), and, by using the hydrological model "backward", they can be used to reliably learn about precipitation (Kirchner, 2009). SIP, however, is not limited to follow the temporal dynamics of the measured rain and can, therefore, more effectively learn from runoff dynamics. This flexibility is even larger than what was obtained by Reichert and Mieleitner (2009) who let the multiplicative factors vary within the event. Here, indeed, we can additionally handle time periods where

rainfall was completely missed.
In case of contradicting input and output measurements SIP makes a compromise between matching the input to the rainfall data and the output to the discharge data. As expected, the direction of this compromise is mainly dictated by our a priori assumptions on the errors of the input measurements (σ_ξ) and output measurements (σ_E) and by the relative number of input and output data (defined by data resolution). Here, in the best case (Sc1) we had 4 times more input than output data, whereas in the worse case (Sc2) the size of the output data set was 2.5 times the size of the input data set. This last point probably explains why SIP considered the information content of the output time series relatively strongly (Figures 6 and S16).
Interestingly, although SIP is meant to realistically estimate the input mainly during the calibration period, it also appears to perform well during extrapolation. In this phase, where only rainfall data are assimilated, obviously no method can correct an erroneous input. Compared to RM, however, SIP more reliably estimated (wider) uncertainty bands for the input (Figure S21). Such a comparison of the inferred rainfall against independent pluviometric data is a strong test of the robustness of the input estimates (Kirchner, 2009; Vrugt et al., 2008).
Runoff predictions during extrapolation had, in all cases, a similarly reasonable coverage, even in the worst scenario and with the simplest error model (Figure 5, last row, first column). This is probably because rainfall biases, although important, only last a few time steps and quickly vanish. Among all error models, however, BD still provided slightly more accurate and reliable uncertainty intervals than the other methods. This confirms the advantages of the bias description for reliable flow prediction, as discussed in previous studies (Honti et al., 2013; Del Giudice et al., 2015a).

5.5.2 Advantages and limitations of SIP

Based on theoretical reflections and experiences from this case study, this novel formulation of input uncertainty as a stochastic process has the following advantages over previous methods:

I. Compared to LS and BD, SIP provides a more accurate assessment of model input during the calibration phase and a realistic characterization of input uncertainty also when extrapolating to the validation period. When input errors are the main contributor to predictive uncertainty, SIP helps to infer more realistic physical parameters than those obtained with LS. Furthermore, by stochastically describing and propagating the input, SIP can support the decomposition of output uncertainty into its sources, a highly desirable feature (Vrugt et al., 2008; Renard et al., 2011). This was not possible with LS, which implicitly partitions output uncertainty into parameter and output measurement errors. Characterization of output uncertainty with SIP is similar to that of BD (compare Figures S14 with S16 and S18 with S20). However, by describing the uncertainties where they arise instead than at the "end of the pipe", uncertainty separation with SIP is conceptually sounder than with BD.

II. Compared to RM, SIP represents a more appropriate model for forcing errors arising from rainfall measurements of suboptimal quality (e.g. collected by an low-resolution rain gauge or one which is located away from the catchment). SIP provides a better rainfall description especially in two cases. First, when the temporal pattern of whole-catchment precipitation during the storm event is different from the observed dynamics. In fact, contrarily to RM, SIP does not assume the storm to have a certain "shape" dictated by

the input measurements. Therefore, SIP can compensate for time-varying input errors and generate a reasonable rainfall dynamics. Describing the input as a continuous stochastic process can estimate the rainrate fluctuations at very fine scales (Sigrist et al., 2012). This can be very useful in hydrology (Paschalis et al., 2013), for instance by helping to downscale coarse rainfall measurements to a temporal resolution appropriate for urban hydrology (Ochoa-Rodriguez et al., 2015) or flood frequency analysis (Beuchat et al., 2011). Second, SIP can even handle complete rainfall misses (i.e. rainy periods with 0 recorded rain), which cannot be tackled with RM. Because of this superior input characterization, especially when the number of events is not very high, SIP can estimate more meaningful physical parameters than RS.

Notwithstanding its several improvements with respect to existing methods, our approach still has some limitations:

I. The main disadvantage of the suggested technique is its computational requirements. Inferring input requires the propagation of a large number of suggested inputs through the model what makes inference computationally much more demanding. This is a generic problem of any technique that makes the attempt of identifying uncertainties and errors where they originate (Yang et al., 2008). Describing the input stochastic process as an Ornstein-Uhlenbeck process and then sampling from it with an advanced MCMC strategy made the inference tractable. While this is more practical than estimating one multiplier per time point, which is virtually unfeasible, it is still computationally much more expensive than any of the other methods tested here.

II. Inference of the dynamics of rainfall at an arbitrary temporal resolution from output is obviously limited by the retention time of the hydrological system. While this was possible for our urban catchment, it may be more limited in natural catchments. Rather than "doing hydrology backward" (Kirchner, 2009), we thus suggest to add a "backward component" to the statistical inference process based on observed input that can identify dynamics at a higher temporal resolution.

III. To ensure a reasonable parameterization of the prior input process requires rainfall measurements additional to those used during calibration and validation. The collection of a relatively large rainfall data set necessary for SIP might be expensive. This, however, can also be seen as an advantage over the other methods, which cannot use this prior information as effectively. Furthermore, the need for additional data is usually not problematic as routine rainfall measurements, e.g. provided by meteorological offices, could be used to parameterize the prior input process.

IV. Finally, the current implementation of SIP assumes that the main reason for model bias are input errors and therefore uses as output error model a LS likelihood (Equations 2.1 and 2.3). As recognized for multipliers (Li et al., 2012; Sun and Bertrand-Krajewski, 2013), this has the potential of producing rainfall estimates which are unrealistically compensating for structural inadequacies. While this effect might be useful in some situations, e.g. to detect unexpected or difficult-to-measure inputs such as groundwater infiltration, we generally prefer having input estimates as independent as possible from the hydrological model (Kirchner, 2009). This can be accomplished by using a model with minimal structural

errors, as done here. However, for more general situations, in Section 5.5.4 we discuss possible strategies to cope with model structural deficits.

5.5.3 Recommendations

Depending on the available resources and the specific objectives of hydrological modeling in a given study, we provide our perspective on which of the four techniques discussed in this paper to preferably apply (but see also Section 5.5.4 for the development of even better alternatives):

I. If a realistic model is available but rainfall data are limited and of insufficient quality and the study focus is to estimate the physical properties of the catchment, the dynamics of the catchment-averaged rainrate, or the contribution of the error sources to output uncertainty, then we suggest using SIP whenever this is computationally feasible.

II. Under the same conditions as above, but if computational requirements make it impossible to perform millions of simulations, we suggest using RM.

III. If a realistic model is available and input and output data are of high quality and the study focus is to estimate the physical properties of the catchment and to predict its output, then we suggest using LS. Note that these conditions are rarely met.

IV. If the available model is structurally deficient, input is reasonably-known, and the study focus is to reliably predict the system output, then we suggest using BD.

V. If the model is structurally uncertain and the input is poorly observed, we recommend to combine BD with either SIP or RM, depending on the computational possibilities. This is particularly relevant when, besides output predictions, input uncertainty estimation is also of interest. Although this is a slight extension of the four alternatives discussed in this paper, it should be straightforward to apply.

We did not mention all possible situations and we focused on the four techniques discussed in this paper. In the next section, we will provide an outlook to promising developments towards even better techniques.

5.5.4 Outlook

In this first application of SIP, as done by several studies on input uncertainty (Kuczera et al., 2006; McMillan et al., 2011; Sun and Bertrand-Krajewski, 2013), for the sake of simplicity, we deliberately adopted a simple output error model similar to LS. For our system-model combination, assuming minimal structural errors resulted to be appropriate. However, since the long-term goal is to target also more realistic situations, we suggest some future directions of research to deal with input and structural errors in hydrology. Note that our overall concept is to address errors where they are generated, rather than correcting the output for the effects of these errors, as BD does. This explicit assessment of the uncertainty sources is conceptually the most appropriate approach. However, this implies the need to propagate the uncertainties through the (typically nonlinear) model. For this reason, these inference techniques will inherently be computationally demanding and we will still have to rely on simpler approaches, such as BD, for computationally very demanding models.

I. Increasing computational power of future hardware will certainly contribute to making approaches for inferring uncertainty that has to be propagated through a model possible. Nevertheless, further developing the numerics of SIP by more advanced sampling techniques for inference based on likelihoods that are formulated as infinite dimensional integrals, is an important branch of research. One option could be to use so-called "Hamiltonian Monte Carlo" algorithms, a promising class of methods that profit from concepts of molecular dynamics to increase the efficiency of MCMC schemes (Brooks et al., 2011).

II. Relaxing the distributional assumptions of SIP by not relying on a transformed Ornstein-Uhlenbeck process could lead to a better description of our prior knowledge on the rain rate. However, this would require the use of a different numerical approach. Research under point 1 above could contribute to make this feasible.

III. Combine SIP with techniques describing model structural deficits to make it possible to infer the input jointly with model structural deficits. This is a very important research direction as we often have both sources of error and are interested in disentangling their contributions to the overall output error (Salamon and Feyen, 2010; Renard et al., 2011). A promising way of doing this is to combine SIP with stochastic, time-dependent parameters as outlined in Reichert and Mieleitner (2009). This approach, although computationally demanding, would enable us to directly capture the sources of uncertainty. Combining SIP with BD or similar autoregressive output error models (as done for RM by Sikorska et al. (2012b) and Li et al. (2012)) could be also a pragmatic alternative. In both cases, however, to minimize the identifiability problem between model parameters and stochastic processes, the use of realistic priors for input errors (Renard et al., 2011) or model discrepancy (Brynjarsdóttir and O'Hagan, 2014) will be decisive.

IV. Extending the input error model to combine different types of input data such as those from radars or microwave links in addition to rain gauges could reduce input uncertainty. Indeed, these alternatives provide better information on spatially integrated rain rates than rain gauges.

5.6 Conclusions

In this study, we aimed at improving parameter inference, better estimating areal precipitation, and contributing to uncertainty separation in hydrological modeling. The main novelty of this work is the development of a more realistic statistical error model for rainfall input. Our advanced inference strategy can jointly estimate rain intensities and model parameters. In particular, we suggest to describe the catchment-averaged precipitation as a stochastic input process (SIP). This appropriately parameterized and transformed Ornstein-Uhlenbeck process is updated in a Bayesian framework by combining rainfall data (the input), system understanding (the hydrological model), and runoff data (the output). We applied SIP to a parsimonious urban rainfall-runoff model and compared the effects of optimal versus mediocre rainfall data. For a better understanding of SIP performance, we compared its results with those obtained with simpler methods, namely the standard least squares (LS), the rainfall multipliers (RM), and the bias description (BD). By combining conceptual arguments with the results of our case study, we conclude that:

I. SIP can effectively deal with severe input errors such as unrecorded or temporally-shifted rainfall peaks. In such situations, simpler methods assuming multiplicative forcing errors provide inaccurate rainfall estimates and biased model parameter values. Given an accurate hydrological model and high-quality discharge data, SIP can asses input uncertainty more reliably than RM. Being formulated as a time-continuous process, SIP can also accurately infer the average rainfall over the catchment at every desired temporal resolution.

II. In our case study, when forcing the model with highly inaccurate input data, similarly to BD, SIP was able to produce physically-coherent parameters. Simpler methods such as LS and RM instead produced biased parameter estimates. Furthermore, also in prediction mode, SIP estimated input and output uncertainty more reliably than RM.

III. Despite those advantages over previous methods, the increased computational requirements of SIP can be limiting for practical applications. Furthermore, as RM, SIP can unintentionally compensate for model structural deficits by incorrectly adjusting the input.

IV. We recommend SIP to reduce the corrupting effects of input uncertainty on hydrological model parameters and to estimate the input to a catchment in an accurate probabilistic way. Further developments will aim at improving its numerical efficiency and extending its applicability to the consideration of structurally inadequate models.

Acknowledgements

The data and codes used are available upon request from the first author. The authors are very grateful to Tobias Doppler for data collection and compilation, and Hans Rudolf Künsch for stimulating discussions. This work was supported by the Swiss National Science Foundation (grant No. CR2212_135551).

5.A log-sinh transformation

The log-sinh transformation has recently shown very promising results for hydrological applications (Wang et al., 2012; Del Giudice et al., 2013). In contrast to the original notation, we suggest a reparameterized notation with parameters that have a more intuitive meaning:

$$g(y) = \beta \log\Big(\sinh\big(\frac{\alpha + y}{\beta}\big)\Big) \quad , \tag{1.1}$$

$$g^{-1}(z) = \Big(\operatorname{arcsinh}\big(\exp(\frac{z}{\beta})\big) - \frac{\alpha}{\beta}\Big)\beta \quad , \tag{1.2}$$

$$\frac{dg}{dy} = \coth\Big(\frac{\alpha + y}{\beta}\Big) \quad , \tag{1.3}$$

where α (originally a/b) and β (originally $1/b$) are "low" and "high" outputs, relatively to observations. α and β control the degree of heteroscedasticity of the predictions (higher when $\alpha \ll \beta$). As in the aforementioned studies, we chose β to be an intermediately high discharge above which uncertainty was assumed not to significantly increase. To ensure a mild degree of transformation, we set α=25 l/s (Figure S23). This provided a plausible representation of the

output-dependent uncertainties with the best rainfall scenario and all error models and enabled predictive intervals to properly encompass high and low flow data.

Chapter 6

Conclusions and outlook

The objective of this thesis was to improve the uncertainty quantification and reduction in urban drainage modeling (UDM) by using statistically sound approaches. This is important to foresee the evolution of the system (e.g. a sewer network) in its current state or after having implemented new management plans. Probabilistically evaluating the environmental consequences of making a certain decision can help to compare the desirability of different scenarios and thus assist water management. A rigorous assessment of model uncertainties can also increase the understanding of the underlying system by providing more realistic estimates of the physical properties of the catchment. The challenge related to uncertainty assessment in UDM, and environmental modeling in general, is that the models used to mimic the system are only partially accurate. This model inaccuracy, or discrepancy from the observed data, has two main origins. First, we are generally unable to perfectly characterize, with only a few equations, the catchment's behavior (e.g. the discharge dynamics at its outlet). Second, we usually do not have fully-representative data (e.g. rainrates) to force the model. In other words, structural deficits and input errors constitute main obstacles to the proper application of urban drainage models (UDMs).

This thesis contributed to the mitigation of the effects of structural and input errors of the model in two ways. First, a method to more appropriately describe the consequences of these errors on model predictions was adapted from applied statistics. In particular, an autocorrelated stochastic process to represent model discrepancies from data was reparametrized to more suitably characterize the uncertainties of UDMs. This bias description was tested in a variety of drainage systems ranging from 12ha to 1300ha and in different conditions (long and short-term predictions, simple and complex models, single and multiple outputs). Second, a method to address one of the sources of bias was developed. In particular, input uncertainty was described as a stochastic process and reduced by assimilating rainfall and runoff data during calibration. This novel method was applied to a sewer system and compared with previous approaches, including the bias description. The conclusions of these analyses are summed up in the ensuing paragraphs.

6.1 Applicability of the statistical methods

In this thesis, several insights were gained from theoretical developments, conceptual considerations, and applications to real world examples. Many of the initially-raised questions (Sect. 1.3.1) could be answered, while others remain open for further research. The main lessons drawn

from this research about analyzing the uncertainties in urban hydrology can be summarized as follows:

- The bias description is an effective tool for urban and natural hydrology to describe predictive uncertainties arising from structural and input errors (Q1). Specifically, its benefits over traditional methods (e.g. standard least squares approaches) are in most cases increased reliability. This means that predictions become wider to appropriately bracket validation data and inferred parameter values become more resistant to bias and therefore less overtuned (Chapters 2, 3, 4, 5, and Appendix B). Compared to more complex methods which also have these advantages (e.g. approaches describing the errors at their sources), the bias description is typically easy to apply, computationally efficient, and readily applicable to diverse models without being "intrusive" in their formulation (Q2, Chapter 3). Consequently, the bias description appears particularly appropriate for predicting on different temporal horizons with model discrepancies ranging from low to moderate and when calibration and prediction conditions are relatively similar. The main limitations of the bias description are that it provides limited insights into the causes of modeling errors and that its formulation and parameters are challenging to define a priori (see following points).

- There is no unique bias description but rather a series of parameterizations have been explored (Chapter 2). Specifically, the bias can have a constant variance, a variance increasing with the output (indirectly, via output transformation) and/or a variance increasing with the input. This makes it adaptable to disparate case studies involving, for instance, runoff modeling in a large sewer system (Chapter 3) or sediment transport in a small river basin (Appendix B). This flexibility, however, can pose a challenge to a priori define a bias implementation ready-to-use with a given model and dataset. A formulation which works in most catchments and with models of different complexities (Chapters 2, 4, 5 and Appendices B) involves using an Ornstein-Uhlenbeck process in an appropriately transformed space (Q3). Typically, a Box-Cox or log-sinh transformation will ensure a realistic representation of the errors both during high and low flows. In the few situations where a transformation is less suitable, an input-dependent variance is likely to be effective (Chapter 3). Alternatively, output transformation and input dependence can be also combined.

- To define prior distributions for the bias parameters can be challenging since, before having conducted the study, typically not much knowledge is available on how the model will deviate from the data. Usually, however, during hydrologic studies, (approximate) prior information on model performances can be derived, e.g., from a previous monitoring campaign, from experience with a similar case, or from available data not used for calibration or validation. Additionally, it is advisable to define these a priori parameters in a way to favor minimal biases and maximal model fit, because we prefer the model, rather than the bias process, to represent the data.

- From a numerical perspective, the bias description involves an inference procedure of comparable complexity to that of traditionally employed Bayesian approaches with simpler error models (e.g. standard least square approaches). This implies using iterative Markov Chain Monte Carlo (MCMC) methods. Among those, adaptive algorithms are suggested

to optimize sampling performances (Q4, Chapter 2). However, thousands of model runs are still a requirement of this procedure. Depending on the complexity of the underlying model, this might be a limiting factor for applications involving slow models. This problem, however, can be substantially mitigated be using emulators, fast statistical approximations of the original model (Reichert et al., 2011; Albert, 2012). Although the uncertainty induced by using a surrogate model is generally low, more experience is required to assess the impact of emulation on parameter inference.

- It is still an open issue whether the bias description can help to preserve the physical meaning of the model parameters (Q5). Current results suggest that, when describing the bias rather than ignoring it, the mean value of the inferred parameters might be more compatible with the physical properties they are meant to represent (Bayarri et al., 2007). This "protective" effect seems to be especially active when the bias is due to input errors rather than structural deficits (Chapter 5) and/or when the prior knowledge about the bias is sufficient (Brynjarsdóttir and O'Hagan, 2014). In general, however, the bias description mainly increases the parameter uncertainty (Chapters 2, 3, 5, Higdon et al. (2005)). Indeed, when a model incorrectly represents the system and its processes, even an appropriate error model cannot provide fully interpretable and physically meaningful parameters.

- Although the bias description is primarily a technique to describe output systematic deviations rather than explain them, if appropriately analyzed, can also provide useful insights into the reasons of these deviations (Q6). Observing the relation between detected bias and measured output can, for instance, help to uncover and correct model structural deficits (Chapter 2). Comparing the bias of several models of the same system can provide useful hints about the relative importance of input and structural errors (Chapter 4). To disentangle the effects of these two error contributions, however, any method that only describes output errors is by definition limited. To separate the uncertainties and reduce their sources, more complex methods are required.

- Using a stochastic input process (SIP) in hydrological inference appears to be an effective way to address the reasons for bias, especially in situations where rainfall uncertainties play a dominant role (Q7). An appropriately parameterized Ornstein-Uhlenbeck process demonstrated sufficient flexibility in describing the temporal evolution of whole-catchment precipitation. Furthermore, a distinctive MCMC algorithm made the inference with SIP numerically tractable.

- Compared to the bias description, SIP can not only reliably quantify the effect of input errors on model output, but it can also assess input uncertainty itself. Furthermore, SIP is able to reduce the reasons for bias by correcting input biases during the inference phase (Q8). Compared to a previous method to quantify input uncertainty, the rainfall multipliers technique, SIP is much more realistic. SIP is not only able to deal with ideal situations of constant over- or under-estimation of the precipitation. It can also account for more complex biases occurring for instance when one rainfall peak is recorded with delay or completely missed. This means that SIP can estimate model input both more reliably and accurately than before. A better input estimation is particularly effective in counteracting the biasing effects of semi-representative precipitation measurements over

parameter inference. In this way, SIP is a valuable tool to assess realistic values of model parameters related to physical characteristics of the underlying system.

- Despite its several advantages over previous methods, SIP still has two potentially limiting features. First, it is not yet clear how to optimally apply SIP in situations when important structural deficits coexist with input errors. For the moment, the methodology is targeted to cases where rainfall uncertainty dominates over structural deficits. An open question, however, is how SIP can be appropriately used with structurally inadequate models and still disentangle input uncertainty from the contribution of structural deficits (Q9).
  Second, even using a customized MCMC algorithm which allows SIP to be used with parsimonious models, the inference procedure might still be one or two order of magnitude more time consuming than for the bias description. In Section 6.3, I will lay out few ideas to overcome the challenges of confounding structural deficits and computational burden.

6.2 Broader benefits for science and practice

The technical advancements summarized in the previous section can be expected to have impacts going beyond the branch of urban hydrology dealing with uncertainty assessment. The knowledge acquired and tools developed can indeed positively influence neighbouring research fields and society in general.

- Motivate better error models accounting for error autocorrelation in other fields where generating predictions is important. An example could be considering bias in groundwater pollutant transport modeling (Liu et al., 2010) or in the inference of oxygen concentration in marine environments (Liu et al., 2011), where suboptimal standard least squares approaches are currently applied.

- Foster a probabilistic consideration of the input in other complex systems where the input is important but difficult to determine correctly. This can be useful, for instance, to estimate the temporal dynamics of nutrient loading entering a lake (Couture et al., 2014).

- Support the model-based learning about the physical properties of the system. Developing statistical methods more resilient to model bias can help all physical sciences to learn about quantities not easy to measure directly. These quantities can be as disparate as the average geological properties of an area (Gupta et al., 2012), the working efficiency of a machine (Brynjarsdóttir and O'Hagan, 2014), or the properties of a beam of subatomic particles (Higdon et al., 2005).

- Contributing to correct uncertainty assessment is useful in environmental management as it provides a sound basis for ecological risk assessment and rational decision making (Reckhow, 1994a; Reichert, 2012). In urban drainage, for instance, this can help engineers and policymakers to design more economically and environmentally favorable control strategies to reduce risks of sewer overflows and flooding.

6.3 Suggestions for future research

During this research, progress has been made in the assessment of output uncertainty and in the quantification and reduction of input errors. The application focus has been urban hydrology

where input usually refers to precipitation and output to sewer runoff at the outlet of the catchment. As mentioned above, however, the developed tools can have broader repercussions in every branch of environmental sciences where estimating model parameters and making probabilistic predictions is of interest. Based on the developed knowledge, here are few areas potentially useful to further investigate in urban drainage modeling and, more broadly, in environmental sciences, as it relates to uncertainty quantification:

Transfer the bias description into practice. From a scientific viewpoint, this technique to describe systematic modeling error appears to have reached a mature point. Indeed, in its current form, it has been successfully applied to several case studies (Reichert and Schuwirth (2012); Dietzel and Reichert (2012, 2014); Del Giudice et al. (2013, 2015a); Honti et al. (2013); Sikorska et al. (2015), just to mention a few). Consequently, only few conceptual developments are to be expected, such as increasing the learning ability of the bias process to make it more accurately correct model extrapolations. Further developments of tools for formulating better priors of the bias can also be valuable in cases when improving the model representation of the system is not a viable option (Brynjarsdóttir and O'Hagan, 2014). In my opinion, however, it is preferable not to make the bias description overly complex but rather to focus on meliorating the system understanding and the model. After having reduced the bias as much as possible, the statistical description presented here can be a useful tool to account for remaining errors. Given these considerations, I would suggest that the next developments for the bias description mainly concern more practical aspects. This could involve, for instance, making the inference of the bias numerically more efficient and therefore applicable with more complex models, by combining it with fast emulators, more effective MCMC algorithms, and optimized matrix operations. It this way, the bias description could become a standard tool for reliable probabilistic predictions of water quality and quantity.

Improve the efficiency of Bayesian inference with the stochastic input process. SIP, the developed method to reliably and accurately estimate the time evolution of model input, is, contrarily to the bias description, still at its early stage. Therefore, more theoretical and practical advancements can be foreseen, the first of which relates to its numerical implementation. Currently SIP might be impractical when combined with slow-running models, since it can involve millions of model executions. Hence, the development of more efficient computational schemes for SIP inference would represent a relevant contribution. One option could be the adoption of so-called "Hamiltonian Monte Carlo" algorithms (Neal, 2011), a class of MCMC methods developed to effectively tackle challenging inference problems involving for instance stochastic process estimation.

Combine the stochastic input process with a method to consider model structural errors. SIP currently focuses on rainfall errors. In many cases, however, model structural errors can be another relevant source of model bias. If the goal is to obtain input estimates more representative of the true precipitation over the catchment (or of any other physical input) or to separate the output bias into its components, then a realistic consideration of structural deficits is necessary. Promising options are to combine SIP with a bias description for model output (similarly to what done by Li et al. (2012) and Sikorska et al. (2012b)) or to introduce the bias term inside the model equations (Reichert and Mieleitner, 2009). To minimize the identifiability problem between model parameters and the stochastic processes, representing input and structural inadequacies will require the proper use of prior knowledge.

Gain more experience with the stochastic input process. In parallel with the previously-mentioned developments, it will be useful to test SIP in other catchments and using different sources of input data such as radar measurements. This will help to generalize the conclusions derived in this thesis and to provide further guidance for subsequent applications.

Bibliography

Ahnert, M., Kuehn, V., Krebs, P., 2010. Temperature as an alternative tracer for the determination of the mixing characteristics in wastewater treatment plants. Water Research 44, 1765–1776.

Aho, A., Kernighan, B., Weinberger, P., 1987. The AWK programming language. Addison-Wesley Longman Publishing Co., Inc.

Albert, C., 2012. A mechanistic dynamic emulator. Nonlinear Analysis: Real World Applications 13, 2747 – 2754. doi:http://dx.doi.org/10.1016/j.nonrwa.2012.04.003.

Alex, J., Benedetti, L., Copp, J., Gernaey, K.V., Jeppsson, U., Nopens, I., Pons, M.N., Rieger, L., Rosen, C., Steyer, J.P., Vanrolleghem, P., Winkler, S., 2008. Benchmark Simulation Model no. 1 (BSM1).

Andersen, T.G., Davis, R.A., Kreiss, J.P., Mikosch, T.V., 2009. Handbook of financial time series. Springer Science & Business Media.

Arattano, M., Marchi, L., 2005. Measurements of debris flow velocity through cross-correlation of instrumentation data. Natural Hazards and Earth System Sciences 5, 137–142.

Bailly-Comte, V., Martin, J.B., Screaton, E., 2011. Time variant cross correlation to assess residence time of water and implication for hydraulics of a sink-rise karst system. Water Resources Research 47.

Balaji, B., 2009. Continuous-discrete path integral filtering. Entropy 11, 402–430.

Banasik, K., Walling, D., 1996. Predicting sedimentgraphs for a small agricultural catchment. Nordic hydrology 27, 275–294.

Bareš, V., Stránský, D., Kopecká, J., Fridrich, J., 2010. Monitoring povodi a stokove sito Města Hostivice - lokalita Sadová. [Monitoring a sewer watershed in Hostivice municipality - Sadová district] (In Czech). Technical Report. Czech Technical University in Prague.

Bates, B., Campbell, E., 2001. A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resources Research 37, 937–947.

Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., Tu, J., 2007. A framework for validation of computer models. Technometrics 49, 138–154.

Bechmann, H., Madsen, H., Poulsen, N.K., Nielsen, M.K., 2000. Grey box modeling of first flush and incoming wastewater at a wastewater treatment plant. Environmetrics 11, 1–12.

Beck, M., 1991. Principles of modelling. Water Science & Technology 24, 1–8.

Beck, M.B., 1994. Understanding uncertain environmental systems, in: Predictability and non-linear modelling in natural sciences and economics. Springer, pp. 294–311.

Beck, M.B., Young, P., 1976. Systematic identification of DO-BOD model structure. J. Environ. Eng. Div. Am. Soc. Civ. Eng., 103 5, 902–927.

Beck, M.S., 1983. Correlation in instruments: cross correlation flowmeters. Instrument Science and Technology 2nd Ed. , B.E. Jones (Ed.), Adam Hilger Ltd, Bristol, U.K.

Beck, M.S., Dran, J., Plaskows, A., Wainwright, N., 1969. Particle Velocity and Mass Flow Measurement in Pneumatic Conveyors. Powder Technology 2, 269–277.

Bekele, E.G., Nicklow, J.W., 2007. Multi-objective automatic calibration of swat using nsga-ii. Journal of Hydrology 341, 165–176.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C., et al., 2013. Characterising performance of environmental models. Environmental Modelling & Software 40, 1–20.

Berne, A., Delrieu, G., Creutin, J.D., Obled, C., 2004. Temporal and spatial resolution of rainfall measurements required for urban hydrology. Journal of Hydrology 299, 166 – 179. doi:10.1016/j.jhydrol.2004.08.002.

Berretta, C., Gnecco, I., Lanza, L., La Barbera, P., 2007. Hydrologic influence on stormwater pollution at two urban monitoring sites. Urban Water Journal 4, 107–117.

Bertrand-Krajewski, J., Briat, P., Scrivener, O., 1993. Sewer sediment production and transport modelling: A literature review. Journal of hydraulic research 31, 435–460.

Beuchat, X., Schaefli, B., Soutter, M., Mermoud, A., 2011. Toward a robust method for subdaily rainfall downscaling from daily data. Water Resources Research 47. URL: http://dx.doi.org/10.1029/2010WR010342, doi:10.1029/2010WR010342.

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water resources 16, 41–51.

Beven, K., Young, P., 2013. A guide to good practice in modeling semantics for authors and referees. Water Resources Research 49, 5092–5098. doi:10.1002/wrcr.20393.

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

Borup, M., Grum, M., Mikkelsen, P.S., 2013. Comparing the impact of time displaced and biased precipitation estimates for online updated urban runoff models. Water Science & Technology 68, 109–116.

Box, G., 1976. Science and statistics. Journal of the American Statistical Association 71, 791–799.

Box, G., Cox, D., 1964. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26, 211–252.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 2008. Time series analysis : Forecasting and control. 746 pp, Wiley, 10.1002/9781118619193.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. Water Resources Research 36, 3663–3674.

Branisavljević, N., Prodanović, D., Pavlović, D., 2010. Automatic, semi-automatic and manual validation of urban drainage data. Water Science and Technology 62, 1013. doi:10.2166/wst.2010.350.

Breinholt, A., Grum, M., Madsen, H., Thordarson, F.Ö., Mikkelsen, P.S., 2013. Informal uncertainty analysis (GLUE) of continuous flow simulation in a hybrid sewer system with infiltration inflow. Hydrology and Earth System Sciences 17, 4159.

Breinholt, A., Møller, J., Madsen, H., Mikkelsen, P., 2012. 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 doi:10.1016/j.jhydrol.2012.09.014.

Breinholt, A., Thordarson, F.Ö., Møller, J.K., Grum, M., Mikkelsen, P.S., Madsen, H., 2011. Grey-box modelling of flow in sewer systems with state-dependent diffusion. Environmetrics 22, 946–961.

Brooks, S., Gelman, A., Jones, G., Meng, X.L., 2011. Handbook of Markov Chain Monte Carlo. CRC press, ISBN 9781420079418.

Brynjarsdóttir, J., O'Hagan, A., 2014. Learning about physical parameters: The importance of model discrepancy. Inverse Problems 30, 114007.

Bulygina, N., Gupta, H., 2009. Estimating the uncertain mathematical structure of a water balance model via bayesian data assimilation. Water Resources Research 45.

Bulygina, N., Gupta, H., 2011. Correcting the mathematical structure of a hydrological model via bayesian data assimilation. Water Resour. Res. 47, W05514. doi:10.1029/2010WR009614.

Butler, D., Davies, J., 2010. Urban Drainage. 3rd ed., Spon Press.

Butts, M.B., Payne, J.T., Kristensen, M., Madsen, H., 2004. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. Journal of Hydrology 298, 242 – 266. doi:http://dx.doi.org/10.1016/j.jhydrol.2004.03.042.

Chebbo, G., Bachoc, A., Laplace, D., Le Guennec, B., 1995. The transfer of solids in combined sewer networks. Water science and technology 31, 95–105.

Chivers, C., 2012. General markov chain monte carlo for bayesian inference using adaptive metropolis–hastings sampling. URL: cran.r-project.org/web/packages/MHadaptive/MHadaptive.pdf.

Christensen, R., Johnson, W., Branscum, A., Hanson, T., 2010. Bayesian ideas and data analysis: An introduction for scientists and statisticians. CRC.

Cios, K.J., Swiniarski, R.W., Pedrycz, W., Kurgan, L.A., Cios, K., Swiniarski, R., Kurgan, L., 2007. The Knowledge Discovery Process, in: Data Mining. Springer, New York and NY, pp. 9–24.

Cirpka, O.A., Fienen, M.N., Hofer, M., Hoehn, E., Tessarini, A., Kipfer, R., Kitanidis, P.K., 2007. Analyzing Bank Filtration by Deconvoluting Time Series of Electric Conductivity. Ground Water 45, 318–328.

Clarke, R., 1973. A review of some mathematical models used in hydrology, with observations on their calibration and use. Journal of hydrology 19, 1–20.

Coutu, S., Del Giudice, D., Rossi, L., Barry, D., 2012a. Modeling of facade leaching in urban catchments. Water Resources Research doi:10.1029/2012WR012359.

Coutu, S., Del Giudice, D., Rossi, L., Barry, D., 2012b. Parsimonious hydrological modeling of urban sewer and river catchments. Journal of Hydrology 464 - 465, 477 – 484. doi:10.1016/j.jhydrol.2012.07.039.

Couture, R.M., Tominaga, K., Starrfelt, J., Moe, S.J., Kaste, Ø., Wright, R.F., 2014. Modelling phosphorus loading and algal blooms in a nordic agricultural catchment-lake system under changing land-use and climate. Environmental Science: Processes & Impacts 16, 1588–1599.

Craig, P., Goldstein, M., Rougier, J., Seheult, A., 2001. Bayesian forecasting for complex systems using computer simulators. Journal of the American Statistical Association 96, 717–729.

Das, P., Haimes, Y.Y., 1979. Multiobjective optimization in water quality and land management. Water Resources Research 15, 1313–1322.

Davis, C.M., Fox, J.F., 2009. Sediment fingerprinting: Review of the method and future improvements for allocating nonpoint source pollution. Journal of Environmental Engineering 135, pp. 15. doi:10.1061/(ASCE)0733-9372(2009)135:7(490).

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., Rieckermann, J., 2013. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. Hydrology and Earth System Sciences 17, 4209–4225. doi:10.5194/hess-17-4209-2013.

Del Giudice, D., Lwe, R., Madsen, H., Mikkelsen, P.S., Rieckermann, J., 2015a. Comparison of two stochastic techniques for reliable urban runoff prediction by modeling systematic errors. Water Resources Research doi:10.1002/2014WR016678.

Del Giudice, D., Reichert, P., Albert, C., Bareš, V., Rieckermann, J., 2015b. Model bias and complexity - understanding the effects of structural deficits and input errors on runoff predictions. Environmental Modelling and Software doi:10.1016/j.envsoft.2014.11.006.

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., Tait, S., 2011. Assessing uncertainties in urban drainage models. Physics and Chemistry of the Earth, Parts A/B/C .

Deletic, A., Maksimovic, E., Ivetic, M., 1997. Modelling of storm wash-off of suspended solids from impervious surfaces. Journal of Hydraulic Research 35, 99–118.

Dietzel, A., Mieleitner, J., Kardaetz, S., Reichert, P., 2013. Effects of changes in the driving forces on water quality and plankton dynamics in three swiss lakes–long-term simulations with belamo. Freshwater Biology 58, 10–35.

Dietzel, A., Reichert, P., 2012. Calibration of computationally demanding and structurally uncertain models with an application to a lake water quality model. Environmental Modelling and Software 38, 129–146. doi:http://dx.doi.org/10.1016/j.envsoft.2012.05.007.

Dietzel, A., Reichert, P., 2014. Bayesian inference of a lake water quality model by emulating its posterior density. Water Resources Research 50, 7626–7647.

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

Dotto, C.B.S., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D.T., Fletcher, T.D., 2011. Performance and sensitivity analysis of stormwater models using a Bayesian approach and long-term high resolution data. Environmental Modelling and Software 26, 1225–1239. doi:http://dx.doi.org/10.1016/j.envsoft.2011.03.013.

Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R., et al., 2004. Calibration of watershed models. American Geophysical Union.

Dürrenmatt, D., Del Giudice, D., Rieckermann, J., 2013. Dynamic time warping improves sewer flow monitoring. Water Research doi:10.1016/j.watres.2013.03.051.

Dürrenmatt, D.J., 2011. Data mining and data-driven modeling approaches to support wastewater treatment plant operation. ETH, Zürich.

Dürrenmatt, D.J., Wanner, O., 2008. Simulation of the wastewater temperature in sewers with TEMPEST. Water Science and Technology 57, 1809–1815.

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrological Sciences Journal–Journal des Sciences Hydrologiques 55, 58–78.

Einicke, G.A., 2012. Smoothing, filtering and prediction: Estimating the past, present and future. New York: InTech .

Evin, G., Kavetski, D., M., T., G., K., 2013. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resources Research URL: http://dx.doi.org/10.1002/wrcr.20284, doi:10.1002/wrcr.20284.

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

Fenicia, F., Kavetski, D., Savenije, H.H., Clark, M.P., Schoups, G., Pfister, L., Freer, J., 2013. Catchment properties, function, and conceptual model representation: is there a correspondence? Hydrological Processes doi:10.1002/hyp.9726.

Freni, G., Mannina, G., 2010. Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution. Journal of Hydrology 392, 31–39.

Freni, G., Mannina, G., 2012. Uncertainty estimation of a complex water quality model: The influence of box–cox transformation on bayesian approaches and comparison with a non-bayesian method. Physics and Chemistry of the Earth, Parts A/B/C 42, 31–41.

Freni, G., Mannina, G., Viviani, G., 2009a. Uncertainty assessment of an integrated urban drainage model. Journal of Hydrology 373, 392–404. doi:http://dx.doi.org/10.1016/j.jhydrol.2009.04.037.

Freni, G., Mannina, G., Viviani, G., 2009b. Urban runoff modelling uncertainty: Comparison among Bayesian and pseudo-Bayesian methods. Environmental Modelling and Software 24, 1100–1111. doi:http://dx.doi.org/10.1016/j.envsoft.2011.03.013,.

Frey, M.P., Stamm, C., Schneider, M.K., Reichert, P., 2011. 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. doi:10.1029/2010WR009993.

Friling, N., Jiménez, M.J., Bloem, H., Madsen, H., 2009. Modelling the heat dynamics of building integrated and ventilated photovoltaic modules. Energy and Buildings 41, 1051–1057.

Gay, D.M., 1990. Usage summary for selected optimization routines.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian data analysis, (Chapman & Hall/CRC Texts in Statistical Science) .

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.

Gresch, M., Braun, D., Gujer, W., 2010. The role of the flow pattern in wastewater aeration tanks. Water Science and Technology 61, 407–414.

Gujer, W., 2008. Systems analysis for water technology. Springer.

Gupta, H.V., Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resources Research 48. doi:10.1029/2011WR011044.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resources Research 34, 751–763.

Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive Metropolis algorithm. Bernoulli 7, pp. 223–242.

Hastings, W., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Henderson, D., Plaschko, P., 2006. Stochastic Differential Equations in Science and Engineering. World Scientific, Singapore.

Higdon, D., Kennedy, M., Cavendish, J.C., Cafeo, J.A., Ryne, R.D., 2005. Combining field data and computer simulations for calibration and prediction. SIAM J. Sci. Comput. 26, 448–466. doi:10.1137/S1064827503426693.

Honti, M., Stamm, C., Reichert, P., 2013. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resources Research doi:10.1002/wrcr.20374.

Hoppe, H., Messmann, S., Giga, A., Grüning, H., 2009. Options and limits of quantitative and qualitative online-monitoring of industrial discharges into municipal sewage systems. Water Science and Technology 60, 859–859.

Iacus, S.M., 2008. Simulation and inference for stochastic differential equations: with R examples. Springer.

Ibe, O., 2013. Markov processes for stochastic modeling. Newnes. ISBN: 978-0-12-407795-9.

Iglewicz, B., Hoaglin, D.C., 1993. How to detect and handle outliers. ASQC Quality Press, Milwaukee and Wis.

ISO, 1977. Iso 2975-7:1977 Measurement of water flow in closed conduits – tracer methods – part 7: Transit time method using radioactive tracers .

ISO, 1992. Iso 9555-2:1992, Measurement of liquid flow in open channels – tracer dilution methods for the measurement of steady flow – part 2: Radioactive tracers .

ISO, 1994. Iso 9555-1:1994, Measurement of liquid flow in open channels – tracer dilution methods for the measurement of steady flow – part 1: General .

Jackson, C.H., Sharples, L.D., Thompson, S.G., 2010. Structural and parameter uncertainty in bayesian cost-effectiveness models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 59, 233–253. doi:10.1111/j.1467-9876.2009.00684.x.

Johnson, S.G., 2014. The NLopt nonlinear-optimization package. URL: http://ab-initio.mit.edu/nlopt. accessed and archived at http://www.webcitation.org/6P7wZcS7B on 26 April 2014.

Jonsdottir, H., Nielsen, H.A., Madsen, H., Eliasson, J., Palsson, O.P., Nielsen, M., 2007. Conditional parametric models for storm sewer runoff. Water resources research 43.

Jørgensen, H.K., Rosenørn, S., Madsen, H., Mikkelsen, P.S., 1998. Quality control of rain data used for urban runoff systems. Water science and technology 37, 113–120.

Juhl, R., Kristensen, N.R., Bacher, P., Kloppenborg, J., Madsen, H., 2013. CTSM-R user guide.

Jun, B.H., 2011. Fault detection using dynamic time warping (DTW) algorithm and discriminant analysis for swine wastewater treatment. Journal of Hazardous Materials 185, 262–268.

Kao, J., Flicker, D., Henninger, R., Frey, S., Ghil, M., Ide, K., 2004. Data assimilation with an extended kalman filter for impact-produced shock-wave dynamics. Journal of Computational Physics 196, 705–723.

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. theory. Water Resour. Res. 42, W03407. doi:10.1029/2005WR004368.

Kendall, M., Stuart, A., Ord, J., 1994. Advanced Theory of Statistics, Distribution Theory (Volume 1). London [etc.]: Arnold [etc.].

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

Keogh, E.J., Pazzani, M.J., 1999. Scaling up dynamic time warping to massive dataset. Principles of Data Mining and Knowledge Discovery 1704, 1–11.

Kessler, M., Lindner, A., Sorensen, M., 2012. Statistical methods for stochastic differential equations. volume 124. CRC Press.

Kirchner, J.W., 2009. Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. Water Resources Research 45.

Kişi, Ö., 2004. Daily suspended sediment modelling using a fuzzy differential evolution approach/modélisation journalière des matières en suspension par une approche dévolution différentielle floue. Hydrological sciences journal 49, 183–197.

Kleidorfer, M., 2009a. Uncertain calibration of urban drainage models: A scientific approach to solve practical problems .

Kleidorfer, M., 2009b. Uncertain Calibration of Urban Drainage Models: A Scientific Approach to Solve Practical Problems. Ph.D. thesis.

Kleidorfer, M., Deletic, A., Fletcher, T.D., Rauch, W., 2009. Impact of input data uncertainties on urban stormwater model parameters. WATER SCIENCE AND TECHNOLOGY 60, 1545–1554. doi:10.2166/wst.2009.493.

Kloeden, P.E., Platen, E., 1999. Numerical Solution of Stochastic Differential Equations. volume 3rd. Springer.

Kollo, T., von Rosen, D., 2005. Advanced multivariate statistics with matrices. volume 579. Springer.

Korving, H., Clemens, F., 2005. Impact of dimension uncertainty and model calibration on sewer system assessment. Water Science & Technology 52, 35–42.

Kracht, O., Gresch, M., Gujer, W., 2007. A stable isotope approach for the quantification of sewer infiltration. Environ. Sci. Technol. 41, 5839–5845. doi:10.1021/es062960c.

Kracht, O., Gresch, M., Gujer, W., 2008. Innovative tracer methods for sewer infiltration monitoring. Urban Water Journal 5, 173–185. doi:10.1080/15730620802180802.

Kristensen, N.R., Madsen, H., 2003. Continuous Time Stochastic Modelling CTSM 2.3 - Mathematics Guide. URL: http://www.webcitation.org/6PI9H6pBR.

Kristensen, N.R., Madsen, H., Ingwersen, S.H., 2005. Using stochastic differential equations for pk/pd model development. Journal of pharmacokinetics and pharmacodynamics 32, 109–141.

Kristensen, N.R., Madsen, H., Jørgensen, S.B., 2004. Parameter estimation in stochastic grey-box models. Automatica 40, 225–237.

Kuczera, G., 1983. Improved parameter inference in catchment models: 1. evaluating parameter uncertainty. Water Resources Research 19, 1151–1162. doi:10.1029/WR019i005p01151.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. Journal of Hydrology 331, 161–177.

Laloy, E., Vrugt, J., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. Water Resources Research 48, W01526. doi:10.1029/2011WR010608.

Law, K.J., Stuart, A.M., 2011. Evaluating data assimilation algorithms. Mon. Wea. Rev. doi:http://dx.doi.org/10.1175/MWR-D-11-00257.1.

Leitao, J.P., Simoes, N.E., Maksimovic, C., Ferreira, F., Prodanovic, D., Matos, J.S., Marques, A.S., 2010. Real-time forecasting urban drainage models: full or simplified networks? Water Science and Technology 62, 2106–2114. doi:{10.2166/wst.2010.382}.

Leube, P.C., De Barros, F.P., Nowak, W., Rajagopal, R., 2013. Towards optimal allocation of computer resources: Trade-offs between uncertainty quantification, discretization and model reduction. Environmental Modelling & Software 50, 97–107. doi:http://dx.doi.org/10.1016/j.envsoft.2013.08.008.

Li, M., Yang, D., Chen, J., Hubbard, S.S., 2012. Calibration of a distributed flood forecasting model with input uncertainty using a bayesian framework. Water Resources Research 48. doi:10.1029/2010WR010062.

Liang, F., Liu, C., Carroll, R., 2011. Advanced Markov chain Monte Carlo methods: learning from past samples. volume 714. Wiley.

Lin, Z., Beck, M., 2007. On the identification of model structure in hydrological and environmental systems. Water resources research 43, W02402.

Lin, Z.L., Beck, M.B., 2012. Accounting for structural error and uncertainty in a model: An approach based on model parameters as stochastic processes. Environmental Modelling & Software 27-28, 97–111.

Liu, X., Cardiff, M., Kitanidis, P., 2010. Parameter estimation in nonlinear environmental problems. Stochastic Environmental Research and Risk Assessment 24, 1003–1022. doi:10.1007/s00477-010-0395-y.

Liu, Y., Arhonditsis, G.B., Stow, C.A., Scavia, D., 2011. Predicting the hypoxic-volume in chesapeake bay with the streeter–phelps model: A bayesian approach1. JAWRA Journal of the American Water Resources Association 47, 1348–1363.

Löwe, R., Mikkelsen, P., Madsen, H., 2013. Stochastic rainfall-runoff forecasting: parameter estimation, multi-step prediction, and evaluation of overflow risk. Stochastic Environmental Research and Risk Assessment , 1–12doi:10.1007/s00477-013-0768-0.

Löwe, R., Mikkelsen, P.S., Madsen, H., 2014a. Stochastic rainfall-runoff forecasting: parameter estimation, multi-step prediction, and evaluation of overflow risk. Stochastic Environmental Research and Risk Assessment 28, 505–516.

Löwe, R., Thorndahl, S., Mikkelsen, P.S., Rasmussen, M.R., Madsen, H., 2014b. Probabilistic online runoff forecasting for urban catchments using inputs from rain gauges as well as statically and dynamically adjusted weather radar. Journal of Hydrology 512, 397–407.

MacDonald, D.D., Dipinto, L.M., Field, J., Ingersoll, C.G., Lvong, E.R., Swartz, R.C., 2000. Development and evaluation of consensus-based sediment effect concentrations for polychlorinated biphenyls. Environmental Toxicology and Chemistry 19, 1403–1413.

Madsen, H., 2007. Time series analysis. CRC Press. ISBN: 9781420059670.

Mannina, G., Viviani, G., 2010. An urban drainage stormwater quality model: Model development and uncertainty quantification. Journal of Hydrology 381, 248–265.

McLean, K.A.P., McAuley, K.B., 2012. Mathematical modelling of chemical processes - obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. The Canadian Journal of Chemical Engineering 90, 351–366. doi:10.1002/cjce.20660.

McMillan, H., Jackson, B., Clark, M., Kavetski, D., Woods, R., 2011. Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models. Journal of Hydrology 400, 83–94.

McMillan, H., Krueger, T., Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrological Processes 26, 4078–4111.

Mejía, A., Daly, E., Rossel, F., Jovanovic, T., Gironás, J., 2014. A stochastic model of streamflow for urbanized basins. Water Resources Research 50, 1984–2001.

Melgaard, H., 1994. Identification of physical models. Ph.D. thesis. Technical University of Denmark.

Merz, B., 2006. Hochwasserrisiken. Möglichkeiten und Grenzen der Risikoabschätzung. E. Schweizerbartsche Verlagsbuchhandlung, Stuttgart. ISBN-13: 978-3510652204.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21, 1087–1092.

Moeller, J.K., 2010. Stochastic state space modelling of nonlinear systems. Ph.D. thesis. Technical University of Denmark (DTU).

Montanari, A., Di Baldassarre, G., 2013. Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty. Advances in Water Resources 51, 498–504. doi:10.1016/j.advwatres.2012.09.007.

Montanari, A., Koutsoyiannis, D., 2012. A blueprint for process-based modeling of uncertain hydrological systems. Water Resources Research 48. doi:10.1029/2011WR011412.

Moore, R., 1984. A dynamic model of basin sediment yield. Water Resources Research 20, 89–103.

Moradkhani, H., DeChant, C.M., Sorooshian, S., 2012. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-markov chain monte carlo method. Water Resources Research 48. doi:10.1029/2012WR012144.

Mourad, M., Bertrand-Kralewski, J.L., 2002. A method for automatic validation of long time series of data in urban hydrology. Water Science and Technology 45, 263–270.

Muleta, M., McMillan, J., Amenu, G., Burian, S., 2013. Bayesian approach for uncertainty analysis of an urban storm water model and its application to a heavily urbanized watershed. Journal of Hydrologic Engineering 18, 1360–1371. doi:10.1061/(ASCE)HE.1943-5584.0000705.

Muleta, M.K., Nicklow, J.W., 2005. Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. Journal of Hydrology 306, 127–145.

Müller, M., 2007. Dynamic Time Warping, in: Information Retrieval for Music and Motion. Springer Berlin Heidelberg, pp. 69–84.

Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models part I : A discussion of principles. Journal of Hydrology 10, 282 – 290. doi:10.1016/0022-1694(70)90255-6.

Neal, R.M., 2011. Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC. chapter MCMC Using Hamiltonian Dynamics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. doi:doi:10.1201/b10905-610.1201/b10905-6.

Neumann, M.B., Gujer, W., 2008. Underestimation of uncertainty in statistical regression of environmental models: influence of model structure uncertainty. Environmental science & technology 42, 4037–4043.

Ochoa-Rodriguez, S., Wang, L.P., Gires, A., Pina, R.D., Reinoso-Rondinel, R., Bruni, G., Ichiba, A., Gaitan, S., Cristiano, E., van Assel, J., et al., 2015. Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation. Journal of Hydrology .

O'Hagan, A., 2006. Bayesian analysis of computer code outputs: A tutorial. Reliability Engineering & System Safety 91, 1290–1300.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertain judgements: eliciting experts' probabilities. Wiley. com.

Omlin, M., Reichert, P., 1999. A comparison of techniques for the estimation of model prediction uncertainty. Ecological Modelling 115, 45–59.

Parker, G.T., Droste, R.L., Rennie, C.D., 2013. Coupling model uncertainty for coupled rainfall/runoff and surface water quality models in river problems. Ecohydrology 6, 845–851.

Paschalis, A., Molnar, P., Fatichi, S., Burlando, P., 2013. A stochastic model for high-resolution space-time precipitation simulation. Water Resources Research 49, 8400–8417.

Petrow, T., Merz, B., Lindenschmidt, K.E., Thieken, A., 2007. Aspects of seasonality and flood generating circulation patterns in a mountainous catchment in south-eastern germany. Hydrology and Earth System Sciences Discussions Discussions 4, 589–625.

Piatyszek, E., Joannis, C., Aumond, M., 2002. Using typical daily flow patterns and dry-weather scenarios for screening flow rate measurements in sewers. Water Science and Technology 45, 75–82.

Piatyszek, E., Voignier, P., Graillot, D., 2000. Fault detection on a sewer network by a combination of a Kalman filter and a binary sequential probability ratio test. Journal of Hydrology 230, 258–268.

Platen, E., Bruti-Liberati, N., 2010. Numerical Solution of Stochastic Differential Equations with Jumps in Finance. volume 64. Springer.

Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, J., Saporta, D., Benito, G., Hedo, M., Molina, A., 2010. Validation and reconstruction of flow meter data in the Barcelona water distribution network. Control Engineering Practice 18, 640–651.

Quinn, J.C., Abarbanel, H.D., 2010. State and parameter estimation using monte carlo evaluation of path integrals. Quarterly Journal of the Royal Meteorological Society 136, 1855–1867. URL: http://dx.doi.org/10.1002/qj.690, doi:10.1002/qj.690.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/. accessed and archived at http://www.webcitation.org/6P7vrJNzs on 26 April 2014.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of speech recognition. Prentice Hall signal processing series, Prentice Hall PTR, Upper Saddle River and NJ.

Reckhow, K., 1994a. Importance of scientific uncertainty in decision making. Environmental Management 18, 161–166.

Reckhow, K.H., 1994b. Water quality simulation modeling and uncertainty analysis for risk assessment and decision making. Ecological Modelling 72, 1–20.

Refsgaard, J.C., van der Sluijs, J.P., Hojberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the environmental modelling process - a framework and guidance. Environmental Modelling & Software 22, 1543 – 1556. doi:http://dx.doi.org/10.1016/j.envsoft.2007.02.004.

Reichert, P., 1998. AQUASIM 2.0 - Tutorial: Computer Program for the Identification and Simulation of Aquatic Systems.

Reichert, P., 2012. Conceptual and practical aspects of quantifying uncertainty in environmental modelling and decision support, International Environmental Modelling and Software Society (iEMSs). pp. 1013–1020. Archived at http://www.webcitation.org/6P7w8shw9.

Reichert, P., Borsuk, M., 2005. Does high forecast uncertainty preclude effective decision support? Environmental Modelling & Software 20, 991–1001.

Reichert, P., Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. Water Resour. Res. 45, W10402. doi:10.1029/2009WR007814.

Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resources Research 48, W09543. doi:10.1029/2011WR011391.

Reichert, P., White, G., Bayarri, M.J., Pitman, E.B., 2011. Mechanism-based emulation of dynamic simulation models: Concept and application in hydrology. Comput. Stat. Data Anal. 55, 1638–1655. doi:10.1016/j.csda.2010.10.011.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46, W05521. doi:10.1029/2009WR008328.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resour. Res. 47, W11516. doi:10.1029/2011WR010643.

Restrepo, J.M., 2008. A path integral method for data assimilation. Physica D: Nonlinear Phenomena 237, 14–27.

Rieckermann, J., Borsuk, M., Reichert, P., Gujer, W., 2005a. A novel tracer method for estimating sewer exfiltration. Water Resources Research 41, 11 PP. doi:200510.1029/2004WR003699.

Rieckermann, J., Neumann, M., Ort, C., Huisman, J.L., Gujer, W., 2005b. Dispersion coefficients of sewers from tracer experiments. Water Science and Technology 52, 123–133.

Rode, M., Arhonditsis, G., Balin, D., Kebede, T., Krysanova, V., Van Griensven, A., Van der Zee, S.E., 2010. New challenges in integrated water quality modelling. Hydrological processes 24, 3447–3461.

Rode, M., Suhr, U., 2007. Uncertainties in selected river water quality data. Hydrology and Earth System Sciences Discussions 11, 863–874.

Rode, M., Suhr, U., Wriedt, G., 2007. Multi-objective calibration of a river water quality modelinformation content of calibration data. Ecological Modelling 204, 129–142.

Rossi, L., Chèvre, N., Fankhauser, R., Margot, J., Curdy, R., Babut, M., Barry, D.A., 2013. Sediment contamination assessment in urban areas based on total suspended solids. Water research 47, 339–350.

Rossi, L., Krejci, V., Rauch, W., Kreikenbaum, S., Fankhauser, R., Gujer, W., 2005. Stochastic modeling of total suspended solids (TSS) in urban areas during rain events. Water Research 39, 4188–4196.

Rossman, L., Supply, W., 2010. Storm water management model user's manual, version 5.0. National Risk Management Research Laboratory, Office of Research and Development, US Environmental Protection Agency.

Sadegh, P., Melgaard, H., Madsen, H., Holst, J., 1994. Optimal experiment design for identification of grey-box models, in: American Control Conference, 1994, IEEE. pp. 132–137.

Sakoe, H., 1978. Dynamic-Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics Speech and Signal Processing 26, 43–49.

Salamon, P., Feyen, L., 2010. Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation. Water Resources Research 46, W12501.

Schellart, A., Shepherd, W., Saul, A., 2012. Influence of rainfall estimation error and spatial variability on sewer flow prediction at a small urban scale. Advances in Water Resources 45, 65 – 75. doi:10.1016/j.advwatres.2011.10.012.

Schilling, W., 1991. Rainfall data for urban hydrology: what do we need? Atmos. Res 27, 5–21.

Schilling, W., Fuchs, L., 1986. Errors in stormwater modeling - a quantitative assessment. Journal of Hydraulic Engineering 112, 111–123.

Schmelter, M., Hooten, M., Stevens, D., 2011. Bayesian sediment transport model for unisize bed load. Water Resources Research 47.

Scholten, L., Scheidegger, A., Reichert, P., Maurer, M., 2013. Combining expert knowledge and local data for improved service life modeling of water supply networks. Environmental Modelling & Software 42, 1 – 16. doi:http://dx.doi.org/10.1016/j.envsoft.2012.11.013.

Schoups, G., van de Giesen, N.C., Savenije, H.H.G., 2008. Model complexity control for hydrologic prediction. Water Resources Research 44, n/a–n/a. doi:10.1029/2008WR006836.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-gaussian errors. Water Resour. Res. 46, W10531. doi:10.1029/2007WR006720.

Seibert, J., 1999. Regionalisation of parameters for a conceptual rainfall-runoff model. Agricultural and forest meteorology 98, 279–293.

Sigrist, F., Künsch, H.R., Stahel, W.A., et al., 2012. A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. The Annals of Applied Statistics 6, 1452–1477.

Sikorska, A., Banasik, K., 2010. Parameter identification of a conceptual rainfall-runoff model for a small urban catchment. Annals of Warsaw University of Life Sciences-SGGW. Land Reclamation 42, 279–293.

Sikorska, A., Scheidegger, A., Chiaia-Hernandez, A., Hollender, J., Rieckermann, J., 2012a. Tracing of micropollutants sources in urban receiving waters based on sediment fingerprinting .

Sikorska, A.E., 2013. Uncertainty analysis of rainfall-runoff predictions for a small urbanized basin. Ph.D. thesis.

Sikorska, A.E., Del Giudice, D., Banasik, K., Rieckermann, J., 2015. The value of streamflow data in improving tss predictions - bayesian multi-objective calibration. Journal of Hydrology Under Review.

Sikorska, A.E., Montanari, A., Koutsoyiannis, D., 2014. Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. Journal of Hydrologic Engineering .

Sikorska, A.E., Scheidegger, A., Banasik, K., Rieckermann, J., 2012b. Bayesian uncertainty assessment of flood predictions in ungauged urban basins for conceptual rainfall-runoff models. Hydrology and Earth System Sciences 16, 1221. doi:10.5194/hess-16-1221-2012.

Sikorska, A.E., Scheidegger, A., Banasik, K., Rieckermann, J., 2013. Considering rating curve uncertainty in water level predictions. Hydrology and Earth System Sciences 17, 4415–4427. doi:10.5194/hess-17-4415-2013.

Sil, B.S., Choudhury, P., 2010. Application of multi-objective technique in modeling water and sediment flow in river reaches, in: INTERNATIONAL CONFERENCE ON MODELING, OPTIMIZATION, AND COMPUTING (ICMOS 20110), AIP Publishing. pp. 504–511.

Smits, J., Moens, M., Klootwijk, M., van Vliet, H., 2008. Testing flow-meters using a field laboratory, in: 11th International Conference on Urban Drainage. Edinburgh and Scotland and UK.

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

Spaaks, J.H., Bouten, W., 2013. Resolving structural errors in a spatially distributed hydrologic model using ensemble kalman filter state updates. Hydrology and Earth System Sciences 17, 3455–3472. doi:10.5194/hess-17-3455-2013.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639. doi:10.1111/1467-9868.00353.

Sun, S., Bertrand-Krajewski, J.L., 2013. Separately accounting for uncertainties in rainfall and runoff: Calibration of event-based conceptual hydrological models in small urban catchments using bayesian method. Water Resources Research 49, 5381–5394. doi:10.1002/wrcr.20444.

Taylor, K.G., Owens, P.N., 2009. Sediments in urban river basins: a review of sediment–contaminant dynamics in an environmental system conditioned by human activities. Journal of Soils and Sediments 9, 281–303.

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

Tomassini, L., Reichert, P., Künsch, H.R., Buser, C., Knutti, R., Borsuk, M.E., 2009. A smoothing algorithm for estimating stochastic, continuous time model parameters and its application

to a simple climate model. Journal of the Royal Statistical Society: Series C (Applied Statistics) 58, 679–704.

Uhlenbeck, G., Ornstein, L., 1930. On the theory of the Brownian motion. Physical Review 36, 823841.

Van Griensven, A., Meixner, T., 2007. A global and efficient multi-objective auto-calibration and uncertainty estimation method for water quality catchment models. Journal of Hydroinformatics 9, 277–291.

Vezzaro, L., Grum, M., 2014. A generalized dynamic overflow risk assessment (dora) for urban drainage real time control. Journal of Hydrology 10, 292–303.

Vezzaro, L., Mikkelsen, P., Deletic, A., McCarthy, D., 2013a. Urban drainage models - simplifying uncertainty analysis for practitioners. Water Science and Technology 68, 2136–2143. doi:10.2166/wst.2013.460.

Vezzaro, L., Mikkelsen, P.S., Deletic, A., McCarthy, D., 2013b. Urban drainage models–simplifying uncertainty analysis for practitioners. Water Science & Technology 68, 2136–2143.

Vihola, M., 2012. Robust adaptive metropolis algorithm with coerced acceptance rate. Statistics and Computing 22, 997–1008.

Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with markov chain monte carlo simulation. Water Resour. Res. 44, W00B09. doi:10.1029/2007WR006720.

Vrugt, J.A., Braak, C.J.F.t., Gupta, H.V., Robinson, B.A., 2009a. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic environmental research and risk assessment 23, 1011–1026. Online first DWK KB-01 PE&RC.

Vrugt, J.A., Diks, C.G., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resources Research 41.

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and bayesian model averaging. Water Resour. Res. 43, W01411. doi:10.1029/2007WR006720.

Vrugt, J.A., Ter Braak, C., Diks, C., Robinson, B.A., Hyman, J.M., Higdon, D., 2009b. Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. International Journal of Nonlinear Sciences and Numerical Simulation 10, 273–290.

Wagener, T., McIntyre, N., Lees, M., Wheater, H., Gupta, H., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. Hydrological Processes 17, 455–476.

Walling, D., 2005. Tracing suspended sediment sources in catchments and river systems. Science of the total environment 344, 159–184.

Walling, D., Webb, B., 1996. Erosion and sediment yield: a global overview. IAHS Publications-Series of Proceedings and Reports-Intern Assoc Hydrological Sciences 236, 3–20.

Walter, E., Pronzato, L., 1997. Identification of parametric models from experimental data. 413 pp, Springer.

Wang, Q., Shrestha, D., Robertson, D., Pokhrel, P., 2012. A log-sinh transformation for data normalization and variance stabilization. Water Resources Research 48, W05514. doi:10.1029/2011WR010973.

White, K.L., Chaubey, I., 2005. Sensitivity analysis, calibration, and validations for a multisite and multivariable swat model1.

Wilkinson, R.D., Vrettas, M., Cornford, D., Oakley, J.E., 2011. Quantifying simulator discrepancy in discrete-time dynamical simulators. Journal of Agricultural, Biological, and Environmental Statistics 10, 77–3.

Willems, P., 2012. Model uncertainty analysis by variance decomposition. Physics and Chemistry of the Earth, Parts A/B/C 42, 21–30. doi:http://dx.doi.org/10.1016/j.pce.2011.07.003.

Willems, P., Molnar, P., Einfalt, T., Arnbjerg-Nielsen, K., Onof, C., Nguyen, V.T.V., Burlando, P., 2012. Rainfall in the urban context: Forecasting, risk and climate change. Atmospheric Research 103, 1 – 3. doi:10.1016/j.atmosres.2011.11.004.

Wolfgang, P., Baschnagel, J., 2013. Stochastic Processes: From Physics to Finance. Springer-Verlag, Berlin.

Wolfs, V., Villazon, M., Willems, P., 2013. Development of a semi-automated model identification and calibration tool for conceptual modelling of sewer systems. Water Science and Technology 68, 167–175. doi:10.2166/wst.2013.237.

Yang, J., Reichert, P., Abbaspour, K., 2007a. Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur river basin (Switzerland). Water Resources Research 43, W10401. doi:10.1029/2006WR005497.

Yang, J., Reichert, P., Abbaspour, K.C., Xia, J., Yang, H., 2008. Comparing uncertainty analysis techniques for a SWAT application to the chaohe basin in china. Journal of Hydrology 358, 1–23.

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., 2007b. Hydrological modelling of the Chaohe basin in China: Statistical model formulation and Bayesian inference. Journal of Hydrology 340, 167–182. doi:10.1016/j.jhydrol.2007.04.006.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. Journal of hydrology 204, 83–97.

Zhang, X., Hoermann, G., Gao, J., Fohrer, N., 2011. Structural uncertainty assessment in a discharge simulation model. Hydrological Sciences Journal 56, 854–869. doi:10.1080/02626667.2011.587426.

Zoppou, C., 2001. Review of urban storm water models. Environmental Modelling & Software 16, 195–231.

Appendix A

Dynamic time warping improves sewer flow monitoring

D. J. Dürrenmatt[a,b], D. Del Giudice[a,b], J. Rieckermann[a].

[a]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b]ETHZ: Swiss Federal Institute of Technology Zürich, 8093 Zürich, Switzerland

Water Research (2013), 47, 3803 - 3816, doi: 10.1016/j.watres.2013.03.051.

Author contributions

D.D. conceived and designed the experiments, developed the method, collected the data, performed the analyses, wrote the paper; D.D.G. collected the data, performed the analyses; J.R. conceived the experiments; all coauthors gave advices, supported result interpretation, and paper revision.

Abstract

Successful management and control of wastewater and storm water systems requires accurate sewer flow measurements. Unfortunately, the harsh sewer environment and insufficient flow meter calibration often lead to inaccurate and biased data. In this paper, we improve sewer flow monitoring by creating redundant information on sewer velocity from natural wastewater tracers. Continuous water quality measurements upstream and downstream of a sewer section are used to estimate the travel time based on i) cross-correlation (XCORR) and ii) dynamic time warping (DTW). DTW is a modern data mining technique that warps two measured time series non-linearly in the time domain so that the dissimilarity between the two is minimized. It has not been applied in this context before. From numerical experiments we can show that DTW outperforms XCORR, because it provides more accurate velocity estimates, with an error of about 7% under typical conditions, at a higher temporal resolution. In addition, we can show that pre-processing of the data is important and that tracer reaction in the sewer reach is critical. As dispersion is generally small, the distance between the sensors is less influential if it is known precisely. Considering these findings, we tested the methods on a real-world sewer to check the performance of two different sewer flow meters based on temperature measurements. Here, we were able to detect that one of two flow meters was not performing satisfactorily under a variety of flow conditions. Although theoretical analyses show that XCORR and DTW velocity estimates contain systematic errors due to dispersion and reaction processes, these are usually small and do not limit the applicability of the approach.

A.1 Introduction

Successful management and control of wastewater and storm water systems requires accurate sewer flow measurements during dry and wet weather. However, due to the harsh environment and given current monitoring techniques, accurate measurements are not easy and discharge data are often biased. Hoppe et al. (2009) report flow measurement errors under normal operating conditions from 2 to 20%, with inductive measuring instruments being more accurate than Venturi channels and tracer dilution methods. While level measurements seem to be rather accurate, reported errors of velocity measuring devices range from 4%, for the average velocity computed from multi-point measurements, to 18% from single-point monitoring devices. Similarly, Smits et al. (2008) report deviations of 5-10% for Doppler and Ultrasonic devices from reference measurements under field conditions using a calibrated pump.

Unfortunately, in practice, flow meters are hardly ever checked against such reference measurements on a routine basis. On the one hand, reference measurements during realistic storm conditions are dangerous and thus rarely performed. On the other hand, even reference measurements only provide limited insight at a specific point of time, usually during average flow conditions. Therefore, it is very difficult to assess the data quality for an individual flow meter. Consequently, many attempts have been made to develop methods for the detection of anomalies such as sensor faults, shifts or systematic biases (Piatyszek et al., 2000; Branisavljević et al., 2010). However, while automated data filtering might be a viable option to validate data a posteriori (Mourad and Bertrand-Kralewski, 2002; Piatyszek et al., 2002; Quevedo et al., 2010), it is preferable to detect such anomalies already during a monitoring campaign − if possible even in real time. While the calibration of the level measurement of area-velocity flow meters, for instance, is a straightforward task, the in-situ calibration of the velocity sensor, however, is

difficult.

In this paper, we therefore suggest to retrieve velocity information from the fluctuations of natural wastewater tracers. To this aim we adapt the dynamic time warping (DTW) approach and demonstrate its superiority over an established method based on cross-correlation analysis (XCORR). Theoretical considerations and numerical experiments are used to determine the field of application and current limitations of the approach. The results from a real-world case study, where we assessed the data quality of a flow meter in a small community, are satisfactory and demonstrate the usefulness of the suggested procedure.

This paper is structured as follows: In section A.2, we develop the methodology and briefly describe i) the experimental design, ii) the XCORR and DTW methods and iii) select suitable statistics for performance evaluation. In section A.3 we first describe the benchmark simulation environment for the numerical experiments, which consists of an inflow generator module, a hydrodynamic pipe flow model and a sensor module and then give the details for the case study. Section A.4 describes the obtained results. Finally, we discuss the results and draw conclusions in sections A.5 and A.6.

A.2 Methods

A.2.1 Using natural tracers to improve discharge monitoring

Fluctuations of substances, compounds and physical properties naturally occurring in the system can be seen as natural (reactive) tracers and tracked. Several studies have shown that natural tracers can provide valuable information for system identification and analysis (Cirpka et al., 2007; Kracht et al., 2007, 2008; Davis and Fox, 2009; Ahnert et al., 2010; Gresch et al., 2010). Probably their biggest advantage is that they do not require system manipulation, such as adding Dirac pulses of artificial tracers (Rieckermann et al., 2005a,b).

The methods presented here likewise use the fluctuations naturally occurring within a sewer channel as tracers. In Figure 1 (left), a setup is presented where two sensors are mounted in the sewer to assess an area-velocity flow meter. The estimate $\hat{\theta}$ of the travel time of a water packet between the measuring locations, θ, is acquired by a method that has similarities to a tracer dilution experiment for flow measurements in open channels (ISO, 1992, 1994) and the transit time method (ISO, 1977) for liquid pipe flows. However, these methods require the injection of a known mass of a tracer substance at the influent of an observed system and measurement of its concentration at the effluent. Instead, we suggest to track the characteristic patterns in an upstream signal and assign these patterns to the associated patterns in the downstream signal. Here, the upstream signal is T_A, which is the measurement of T_{in} and the downstream signal is T_B, which measures T_{out}. In contrast to the dilution experiment, however, it is not possible to measure discharge with the given experimental setup.

The underlying principle to estimate velocities is illustrated in Figure 1, right. "Water packet" I is observed in the system influent at time t_0 and two time steps Δt later in the effluent. Hence, the observed travel time is $\hat{\theta}_0 = 2\Delta t$. Similarly, the travel time is $2\Delta t$ for packet II, $3\Delta t$ for packet III and $4\Delta t$ for packet IV.

In fact, the discharge through the reactor varies with time and is not known, so are the flow velocities. Let $v = (v_0, v_1, \ldots, v_k, \ldots, v_L)$ be the flow velocity at time $t = (t_0, t_1, \ldots, t_k, \ldots, t_L)$. So, packet I travels with v_0 at the first and v_1 at the second time step. Because the distance between the observations is L, the equation $L = \Delta t v_0 + \Delta t v_1$ holds. Systematically formulating

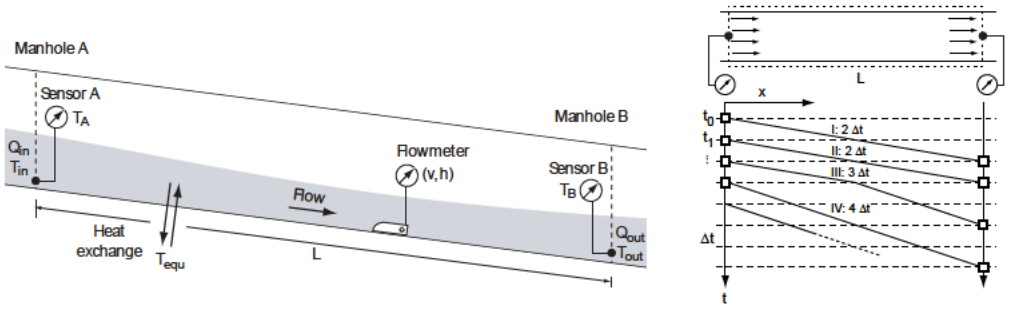

Figure 1: Left: Experimental setup. Two temperature probes are mounted in a sewer section of length L (system boundary indicated as dashed box), one at the inlet measuring T_{in} as T_A and the other at the outlet measuring T_{out} as T_B. These sensors are used to check a flow meter, which often independently measures flow velocity v and water level h. Note that heat exchange processes affect wastewater temperature. Right: Illustration of the time required for water packets I-IV to flow through an ideal plug-flow reactor.

the observations in this form but replacing v_i with $v_i = L/\theta_i$ gives

$$
\begin{aligned}
1 &= \Delta t\theta_0^{-1} + \Delta t\theta_1^{-1} && (1)\\
1 &= \Delta t\theta_1^{-1} + \Delta t\theta_2^{-1} && (2)\\
1 &= \Delta t\theta_2^{-1} + \Delta t\theta_3^{-1} + \Delta t\theta_4^{-1} && (3)\\
1 &= \Delta t\theta_3^{-1} + \Delta t\theta_4^{-1} + \Delta t\theta_5^{-1} + \Delta t\theta_6^{-1} && (4)\\
1 &= \dots && (5)
\end{aligned}
$$

or

$$\frac{1}{\Delta t} e = \mathbf{S}\theta_{inv} \tag{6}$$

with e being a vector of ones, $\mathbf{S}$ a structure matrix and θ_{inv} a vector of the reciprocals of the unknown travel times. Given enough observations, this equation system can theoretically be solved for θ. If the travel times θ are relatively short compared to the variation in $\theta(t)$, it is feasible to approximate the real travel times by the observed travel times, i.e. $\theta_k \approx \hat{\theta}_k$.

Although this will only be exact in dispersion- and reaction-free systems, in A.A it is shown for a sewer modeled by a tanks-in-series system of N continuous stirred-tank reactors with first order degradation reaction, that a systematic relative error of

$$E_{\theta,rel} = 1 - \frac{N}{2\pi f\theta} \arctan\left(\frac{2\pi f\theta}{N + k\theta}\right). \tag{7}$$

results for systems with constant influent discharge Q, a harmonically oscillating influent concentration series with frequency f and reaction coefficient k. It can be shown that the same considerations hold for a sewer reach, where $A = const.$ because $Q = const.$ and for temperature times series using the heat balance instead of mass balance. The error diminishes when $N \to \infty$, which means a dispersion-free system, and $k \to 0$, which means a reaction-free system. For $k > 0$ and $N < \infty$ however, $E_{\theta,rel}$ is positive, θ is under-estimated.

Given the travel time of a packet and assuming that the travel time is shorter in comparison to

fluctuations in the flow velocity, the latter can be approximated by

$$v(t) \approx \frac{L}{\hat{\theta}(t)} \tag{8}$$

where $\hat{\theta}(t)$ denotes the travel time.
In the following, we first describe how we estimate $\hat{\theta}(t)$ based on cross correlation, second based on dynamic time warping and third how we compare their performance.

Cross correlation (XCORR)

The XCORR-technique has a rather long history and has already been suggested for the flow meter design by Beck et al. (1969) and Beck (1983); recent applications are discharge estimation in Karst systems (Bailly-Comte et al., 2011) and the estimation of debris flow velocity (Arattano and Marchi, 2005). In principle, the XCORR method determines the shift in the time axis that maximizes the correlation between two signals (or two synchronized signal windows). Formally, the cross correlation of two time series T_A and T_B with lag τ is

$$R_{T_A,T_B}(\tau) = \int_0^T T_A(t-\tau)T_B(t)dt \tag{9}$$

and one tries to find τ that maximizes $R_{T_A,T_B}(\tau)$, denoted τ_{max} and formally expressed by

$$\tau_{max} = \max_{\tau} R_{T_A,T_B}(\tau). \tag{10}$$

The resulting lag τ is then a measure for the average flow time within the signal window (Beck, 1983). In order to get a high-resolution series of flow times, one wants to minimize the length of the signal window. Small Short signal windows, however, can lead to instabilities (see below).

Dynamic Time Warping (DTW)

DTW is a method for measuring the similarity between data signals varying in time. In contrast to the well-known linear alignments of two signals by cross-correlation, it is used to optimally align two sequences by non-linearly warping the time-axis of the sequences until their dissimilarity is minimized. Specifically, DTW non-linearly expands and compresses signals in time by comparing the distance of each point of the first sequence with every point of the second one (Rabiner and Juang, 1993). The result is a warping path that contains information on how to translate, compress and expand patterns so that similar features are matched (Jun, 2011).
Originally, DTW was applied in the field of speech recognition (Sakoe, 1978), but it is now also used in other fields for sequence alignment and to measure (dis)similarities. Among other applications, it has been successfully used to identify hydraulic residence times in WWTP reactors from water quality measurements Dürrenmatt (2011). In contrast to common applications of DTW, we are not interested in dissimilarity measures or the aligned sequences, but rather in the warping path itself. As it maps all the points of an influent series to the points of an effluent series, it provides an estimate for the hydraulic residence time.
To align the two sequences $X = (x_1, x_2, \ldots, x_n, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_m, \ldots, y_M)$ with DTW, first a distance matrix $\mathbf{D} \in \mathbb{R}^{N \times M}$ with the Euclidian distance between all points of the two series

$$D_{n,m} = \sqrt{(x_n - y_m)^2} \tag{11}$$

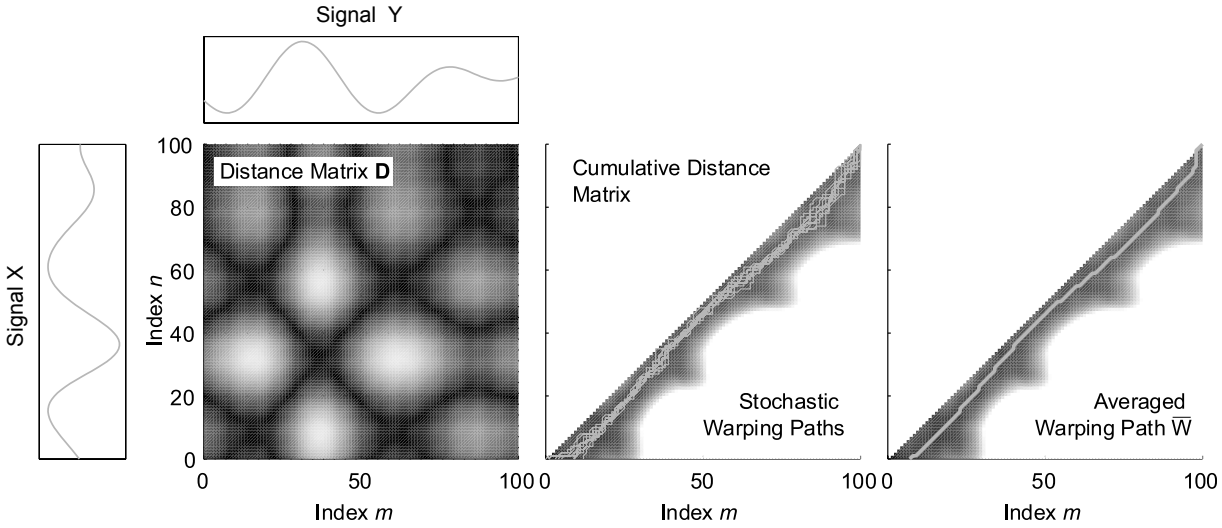

Figure 2: Illustration of the DTW algorithm. Given two time series, X and Y, the distance matrix, $\mathbf{D}$, for $m \geq n$ is first calculated (using Eq. 11; left). The cumulative distance matrix is computed by solving Eq. (13) (middle) and is applied to find the warping path, W, with the smallest cumulative distance. This is repeated Z times and yields Z warping paths (middle, squiggly lines, which are averaged to warping path $\bar{W}$ (right).

is computed. Second, a warping path, $W = (w_1, w_2, \ldots, w_k, \ldots, w_K)$, is computed as a sequence of consecutive matrix elements that define a mapping between A and B with the k-th element being $w_k = (n, m)_k$. The warping path must satisfy the following conditions (Müller, 2007): i) the warping path starts and ends in diagonally opposite corners of the matrix ($w_1 = (1, 1)$, $w_K = (N, M)$), ii) the path is continuous, allowed steps are restricted to $w_k - w_{k-1} \in \{(0, 1), (1, 0), (1, 1)\}$, and, additionally for this application, iv) a pattern appears first in sequence X and then in sequence Y ($w_k = (n, m)$ with $m \geq n$).

As a consequence of the first condition, the algorithm needs some 'burn-in" time to avoid physically impossible travel times and achieve an appropriate alignment. While many warping paths exist that satisfy these conditions, the interest lies in the particular path that minimizes the total distance, d, defined by

$$d(X, Y) = \sum_{w_k} D_{n,m} \tag{12}$$

This path can be efficiently found by evaluating the recurrence

$$p_{n,m} = D_{n,m} + \min\left(p_{n-1,m-1}, p_{n-1,m}, p_{n,m-1}\right), \tag{13}$$

where the cumulative distance, $p_{n,m}$ (Figure 2, middle), is defined as the distance in the cell (n, m) and $p_{n-1,m-1}$, $p_{n-1,m}$, and $p_{n,m-1}$ are the minimal cumulative distances of the neighboring cells obtained through dynamic programming (Keogh and Pazzani, 1999). The algorithm is illustrated in Figure 2.

Given the time series T_A and T_B, each element in the computed warping path W represents the mapping of the i-th point in time series T_A at t_i to the j-th point in series T_B at t_j. As the mapping is the result of tracking an imaginary water packet through a sewer reach, the difference $t_j - t_i$ is the travel time of the packet in the reach. If the travel time is short compared to the variability of v, then $\hat{\theta}_{i,j} \approx t_j - t_i$. Otherwise, a linear equation system, as given in Eq. (6), can be set up and solved.

Special attention must be given to mappings for which $t_j - t_i = 0$. These occur if DTW selects matrix elements for the warping path that lie on the diagonal of the cumulative distance matrix. They are physically not meaningful and must be dropped.

DTW always finds the warping path that minimizes the cumulative distance and considers it as the "correct" time alignment (Rabiner and Juang, 1993). When using noisy and erroneous input signals however, unrealistic warping paths may result when the Euclidian distances between the upstream and downstream measurements are small. To generate smoother warping paths, we therefore repeat the computation of the warping path Z times. In each run, we perturb the data by adding a random term, $\varepsilon = \mathcal{N}(0, \sigma_m)$ to each data point of the influent and effluent series. ε is normally distributed with zero mean and a standard deviation σ_m, for which the accuracy of the quality sensor is a first guess.

The Z individual warping paths are then combined into an averaged warping path $\bar{W}$ by calculating the mode along the diagonal axis $(1,1) - (N,M)$, as shown in Figure 2. The standard deviation along the path (σ_s) indicates how well-defined a specific point of the warping path is. Therefore we used it as a quality measure.

To remove outliers, the travel time estimates are further processed with the modified Z-Score statistical test (Iglewicz and Hoaglin, 1993). The modified Z-Score, M_Z, is calculated with

$$M_Z = \frac{0.6745 \left(\hat{\theta} - median \left(\hat{\theta} \right) \right)}{MAD} \tag{14}$$

where $\hat{\theta}$ is a vector of the travel time estimates and the median absolute deviation, MAD, is defined by

$$MAD = median \left(|\hat{\theta} - median \left(\hat{\theta} \right) | \right). \tag{15}$$

Now, all flow time estimates are kept for which $|M_Z| < p$ with a positive constant p.

Based on our numerical experiments, we obtained best performance by accepting the points corresponding to the lowest 10% quantile of σ_s for velocity estimation and by choosing $p = 1.7$.

A.2.2 Performance assessment

After a visual quality check, performance is quantified based on the coefficient of variation of the root mean squared deviation, $CV(RMSD)$, between measured (v_{meas}) and estimated velocity (v_{est}) time series.

$$CV(RMSD) = \frac{RMSD}{\bar{v}_{meas}} \tag{16}$$

where $\bar{v}_{meas}$ denotes the average of the measured series and

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} \left(v_{meas,i} - v_{est,i} \right)^2}{N}} \tag{17}$$

All time series are of length N.

Plotting v_{est} against v_{meas} helps to identify systematic deviations between the measurements and the estimates. Where such a plot results in a not very intuitive point cloud, plotting the weighted average can provide better insight. The average is weighted by σ_i, which is the standard deviation over all calculated warping paths at point i. This indicates how well defined a point in the warping path is.

The average $\bar{v}_{(m,n)}$ within a window $[v_m, v_n]$ is defined by

$$\bar{v}_{(m,n)} = \frac{\sum_{i=m}^{n} w_i v_i}{\sum_{i=m}^{n} w_i} \tag{18}$$

with

$$w_i = \frac{1}{1+\sigma_i}. \tag{19}$$

To obtain a meaningful comparison to the XCORR method, we also repeated the computation of XCORR velocity estimates Z times by perturbing the data with random errors and accepting the estimates based on the criterion given in Table 1, as described for the DTW method above.

A.2.3 Investigating the field of application using scenario and sensitivity analysis

For real-world applications, some important implications arise from the placement and characteristics of the water quality sensors. Regarding the placement, i) the distance L must be known as exactly as possible, ii) lateral and transversal mixing must be complete, iii) sewer properties should not change and iv) there must not be any lateral inflow. This is, because the natural tracers yield an average velocity v over distance L. To compare this estimate to the values of an existing velocity meter, the velocity in the section between the quality sensors and the original device should ideally be uniform, or at least v should change slower than the travel time of the packet.

Regarding the sensors characteristics, it is important that i) the wastewater tracer exhibits sufficient variations, ii) the accuracy and sampling frequency are as high as possible, iii) the design and placement prevents clogging, iv) the sensors must be of the same type, to avoid phase shifts. As shown by Beck et al. (1969) phase shifts from different response times not necessarily cancel out.

As such practical aspects influence the field of application of the method, we investigate some of these aspects with numerical experiments on different scenarios using a benchmark simulation environment. Specifically, we investigated i) the effect of dispersion and reaction, ii) the effect of the sensor response time and the measuring error, and iii) the effect of the variation in the influent signal. In addition, the sensitivities of the parameters of the XCORR and the DTW methods are investigated by increasing or decreasing their values while observing the change in performance. From this, we then derive recommendations regarding the applicability of the method as well as the choice of important parameters.

In the following, we investigated the field of application using heat as a tracer. Heat is convenient, because temperature probes are robust and inexpensive sensors that allow to track natural temperature fluctuations in near real-time.

A.3 Material

A.3.1 Benchmark simulation environment (BSE)

For the numerical experiments, our BSE permits to virtually investigate a wide range of different real-world situations, or scenarios. The BSE implements or interfaces data generation algorithms, a hydrodynamic solver of the Saint-Venant equations, a sensor model, the XCORR and DTW methods and modules for performance analysis as described above. The flow scheme of a scenario run is given in Figure 3 and default parameter values for each particular module

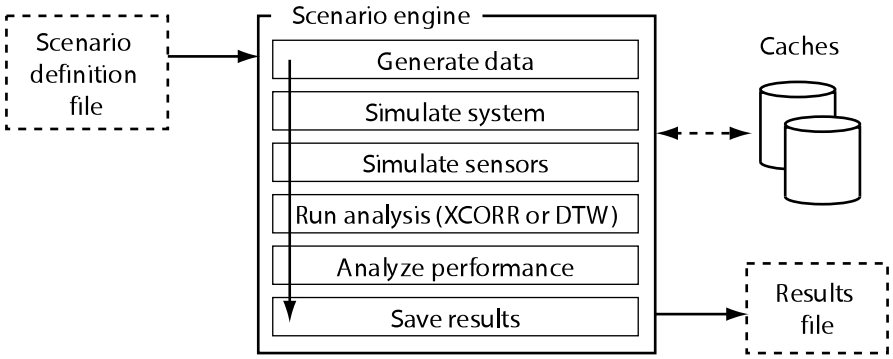

Figure 3: Overview of the benchmark simulation environment. The scenario engine receives a definition file, performs the indicated tasks and saves the outputs to a results file. Computation is sped up by caching recurring results.

are listed in Table 1. In the following subsections, we briefly describe the individual models. Implementation details are given in the Supporting Information.

System inflow generator

Two alternatives are available for the generation of influent series. The first is an autoregressive process of arbitrary order (AR(n)):

$$S_{i+1} = \sum_n \rho_n S_{i-n} + \mathcal{N}(0,1) \tag{20}$$

Because the coefficients for a higher order AR process are not necessarily stable, the Ornstein-Uhlenbeck (OU) process is provided as second alternative. An OU process (Uhlenbeck and Ornstein, 1930) is a mean-reverting random walk in continuous time, or, in other words, the continuous analogue to an AR(1) process. An exact solution of the stochastic differential equation which describes the process is

$$S_{i+1} = S_i e^{-\lambda_{OU}\Delta t_{OU}} + \mu_{OU}\left(1 - e^{-\lambda_{OU}\Delta t_{OU}}\right) + \sigma_{OU}\sqrt{\frac{1 - e^{-2\lambda_{OU}\Delta t_{OU}}}{2\lambda_{OU}}}\mathcal{N}(0,1) \tag{21}$$

where μ_{OU} denotes the mean, λ_{OU} the mean reversion rate and σ_{OU} the volatility. Δt_{OU} is the sampling interval of the OU process. As the OU parameters cannot be intuitively chosen, Table 2 shows OU process parameters conditioned on observations from sewers with low (Wangen), medium (Dübendorf) and high discharge (Zürich). Seasonal effects are not considered. Realizations for each case are shown in Figure 4.

Sewer transport and transformation model

Sewer transport and transformations of wastewater quality parameters were carried out with a numerical heat transfer model. This consists of a hydrodynamic model, in which temperature has been introduced as a parameter. As cooling, particularly heat transfer to the surrounding soil, is the dominant heat exchange process in gravity sewers (Dürrenmatt and Wanner, 2008), this was modeled using a Newtonian cooling process. The model was implemented in AQUASIM (Reichert, 1998), using the diffusive wave approximation, a downstream free-surface boundary condition and the friction approach according to Manning-Strickler. The reaction coefficient was estimated from previous experimental data, where the best fit was found using the simplex

Table 1: List of all parameters of the benchmark simulation environment with their default values. They were calibrated for data from a typical sewer reach in a medium-sized city in Switzerland.

Parameter	Symbol	Default
Influent signal[a]		
Sampling interval	Δt	10 s
Sampling points	N	2000
Lower bound	min	1 ℃
Upper bound	max	30 ℃
Variations	OU-process with $\mu_{OU} = 19.98$ ℃, $\sigma_{OU} = 2.6 \cdot 10^{-2}$ ℃/s$^{1/2}$ and $\lambda_{OU} = 2.8 \cdot 10^{-4}$ 1/s	
Influent discharge[b]		
Sampling interval	Δt	10 s
Sampling points	N	2000
Lower bound	min	0.0005 m^3/s
Upper bound	max	1 m^3/s
Variations	OU-process with $\mu_{OU} = 6.5 \cdot 10^{-2}$ m^3/s, $\sigma_{OU} = 1.8 \cdot 10^{-3}$ m^3/s$^{3/2}$ and $\lambda_{OU} = 1.0 \cdot 10^{-2}$ 1/s	
Hydrodynamic model		
Manhole distance	L	50 m
Pipe diameter	D	1.2 m
Strickler coefficient	k_{st}	72 m$^{1/3}$/s
Slope	S_0	0.001
Heat transfer coefficient	k	$4.4 \cdot 10^{-4}$ s^{-1}
Equilibrium temperature	T_{equ}	12 ℃
Dispersion[c]	D	(Approximated by num. disp.)
Spatial resolution	Δx	1 m
Temporal resolution	Δt	10 s
Sensor model		
Response time	T_{90}	90 s
Measurement error	White noise $\varepsilon_m \sim \mathcal{N}(\mu_m = 0, \sigma_m = 0.01)$	
Cross correlation (XCORR)		
Number of realizations	Z	100
Superimposed noise	σ_m	0.02
Window length	w	900 s
Normalize series	b_{norm}	No
Calculate derivative	b_{deriv}	Yes
Acceptance criterion	In lower 10% percentile of σ_s	
Outlier statistics	p	1.7
Dynamic time warping (DTW)		
Number of realizations	Z	100
Superimposed noise	σ_m	0.02
Normalize series	b_{norm}	No
Calculate derivative	b_{deriv}	Yes
Acceptance criterion	In lower 10% percentile of σ_s	
Outlier statistics	p	1.7

[a]Corresponds to T_{in} in Figure 1

[b]Corresponds to Q_{in} in Figure 1

[c]Dispersion can either be approximated by solving the dispersion-advection-reaction with $D \neq 0$ or by introducing sufficiently high numerical dispersion by adequate choice of Δx (setting $N = L$)

Table 2: Parameterization of Ornstein-Uhlenbeck processes to simulate discharge and temperature profiles in sewers with low (Wangen, ≈ 2'500 population equivalents, PE), medium (Dübendorf, ≈ 19'500 PE) and high discharge (Zürich, ≈ 400'000 PE). For simplicity, seasonal effects are not considered. μ_{OU} denotes the mean, σ_{OU} the volatility and λ_{OU} the mean reversion rate. The unit of σ_{OU} is the unit of μ_{OU} divided by $[\sqrt{s}]$.

Description		μ_{OU}	σ_{OU}	λ_{OU} [1/s]
Discharge (Wangen)	[m^3/s]	0.007	0.00010	0.00042
Discharge (Dübendorf)	[m^3/s]	0.065	0.0018	0.0088
Discharge (Zürich)	[m^3/s]	2.6	0.0088	0.00022
Temperature (Wangen)	[℃]	13.8	0.013	0.00030
Temperature (Dübendorf)	[℃]	20.0	0.0053	0.00018
Temperature (Zürich)	[℃]	13.1	0.0067	0.000062

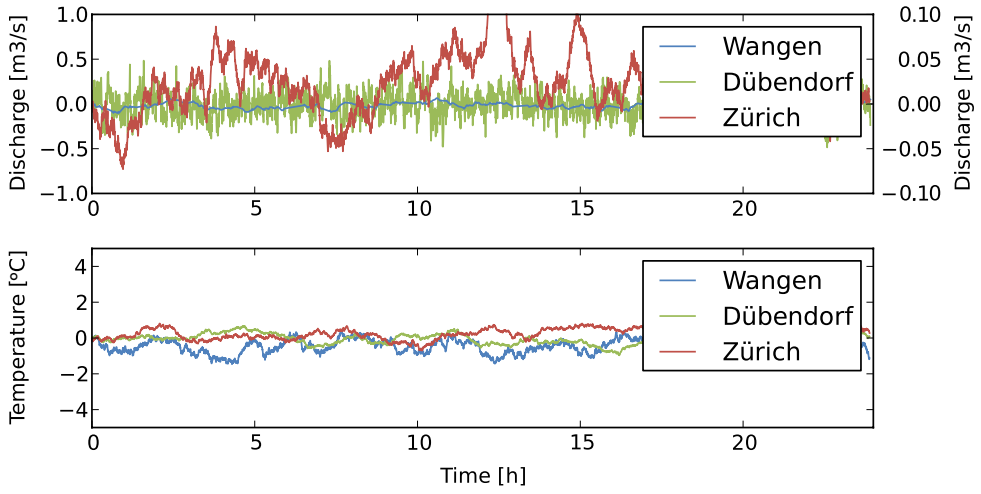

Figure 4: Realizations of OU-processes simulated with parameter values given in Table 2. For better visibility, the mean values were set to zero ($\mu_{OU} = 0$); $\Delta t_{OU} = 10$ s.

method and a least-squares objective function.

Sensor model

For realistic tests of the proposed methods, the dynamic behavior of sensors needs to be considered, in addition to measurement noise. With regard to the dynamic behavior, the response time T_{90} is a measure which expresses the duration, after a step change, until observations reach 90% of the final value of the step response. The response time T_{90} is divided into two parts, i.e. the delay time (T_{10}) and the rise time. The delay time is the time elapsed until 10% of the final value of the step response is reached.

The considered model closely follows the sensor model developed in the framework of the Benchmark Simulation Model No. 1 (Alex et al., 2008) for wastewater treatment plants. The temperature probes in the case study are "Class A" sensors with very low response times. Sensors of "Class A" are modeled by two first-order transfer functions in series of linear time-invariant systems, i.e.

$$H(s) = \frac{1}{1 + Ts} \frac{1}{1 + Ts} \tag{22}$$

with

$$T = \frac{T_{90}}{233.4\text{s}}. \tag{23}$$

where T_{90} is the response time in seconds. For a typical sensor of this group, the delay time is approximately 13% of the response time.

A.3.2 Case study

In a case study, we installed two temperature probes (Onset HOBO TMC20-HD) at a distance of 37.3 m upstream and downstream of two existing flow meters in a circular sewer ($D = 1.0$ m, $Q_{mean} = 85$ Ls^{-1}) in Dübendorf, Switzerland.

Flow meter 1 was a modern area-velocity flow meter. It computes average velocities from multiple velocity estimates in up to 16 vertical cross-sections. The individual velocity estimates are obtained from cross-correlation analysis, according to the manufacturer with a precision of +-1% of the measurand. In addition, it uses two redundant water level measurements by means of ultra-sound and pressure. Flow meter 2 was also an area-velocity flow meter, although with a simpler monitoring principle that computes cross-sectional averages from estimates of maximum flow velocities from a single measurement. Own test in a laboratory flume suggest a precision of +- 10% of the velocity sensor (single standard deviation). It only uses a single bubbler sensor to measure flow depth. The upstream temperature probes was installed in a manhole upstream of the flow meters and the downstream sensor approx. 2 meters downstream of the flow meters.

The data were recorded unsupervised with a temporal resolution of 5 seconds over a period of one week. The data also included rain events, which is particularly valuable since it allows investigating the performance of the flow meters over a wide range of flows. DTW and XCORR were then used to compute average velocities. Default parameter values were used (Table 1), except for Δt, which was set to 5s. Prior to the analysis, temperature series were normalized to zero mean and unit variance (Table 1).

A.4 Results

A.4.1 Benchmark simulation environment

Comparison of XCORR and DTW

The results of flow velocity estimation with the XCORR method based on synthetic data are depicted in Figure 5. The data were generated with the parameter values listed in Table 1.
First, it is visible that only 10 % of the values were accepted, sorted by the standard deviation of the generated realizations as described in Section A.2.1 (cf. filled black circles in Figure 5d). Second, it is interesting that the accepted values are approximately in the same range, which is close to the average flow velocity. This is confirmed by comparing the estimated velocity v_{est} to the true velocity v_{true} (Figure 6, left). It can be seen that for a rather large range of true values, the same flow velocity is estimated. This means, that XCORR has an inherent bias towards the average flow velocity, which makes it difficult to detect rapid changes and sharp peaks.
In comparison, the DTW method shows a much better performance. First, the results displayed in Figure 5e) and f) clearly show that more values are accepted than for XCORR and that DTW estimates cover a much wider range of flows (Figure 6, right) than the previous methods. Second, as expected from Eq. (7), it is visible that DTW over-estimates flow velocities systematically. However, the deviation is small and generally less than 0.05 m/s. Larger deviations are computed for border regions with high and low flow velocities, where there are only very few accepted estimates. This is very promising, because DTW not only clearly outperforms XCORR, but can also detect small deviations.
In addition, these results illustrate how the proposed method can be applied to find systematic deviations of the measuring device. While the time series plots in Figure 5d) and f) are suitable for online signal diagnosis and real-time performance assessment, Figure 6 compares measured to estimated values for a specific period of time. This is best performed off-line using long time series, because a larger data base leads to more accepted data points and, consequently, to a more precise diagnosis.

Assessment of the field of application

The field of application was assessed for both, the DTW and the XCORR method. The main focus, however, was given to the DTW method, as it clearly outperforms XCORR. The findings are comparable to those of the XCORR method and their interpretation is very similar.
First, the effect of dispersion was investigated by varying the length of the sewer reach, and the effect of reaction by varying the first order reaction coefficient. As mentioned above, we expect that approximating the hydraulic residence time of a water packet by its travel time has an increasing systematic error for increasing dispersion and increasing reactions, as shown in Eq. (7). The results are plotted in Figure 7a.

They indicate that, even for long sewer reaches of several hundred meters, the effect of dispersion is negligible, while higher reaction coefficients quickly decrease performance. The elevated CV(RMSD) for very short reaches ($L = 10$ m) is explained by the fact that a sampling interval of 10 seconds is rather long for the short hydraulic residence times. In practice, the distance between measuring locations will likely be less than 200 meters, because lateral inflows must be avoided and constant sewer properties are required. Considering that the real sewer system for which the model was calibrated had small reaction coefficient of $k = 4.4 \cdot 10^{-4}$ s^{-1}, it is unlikely

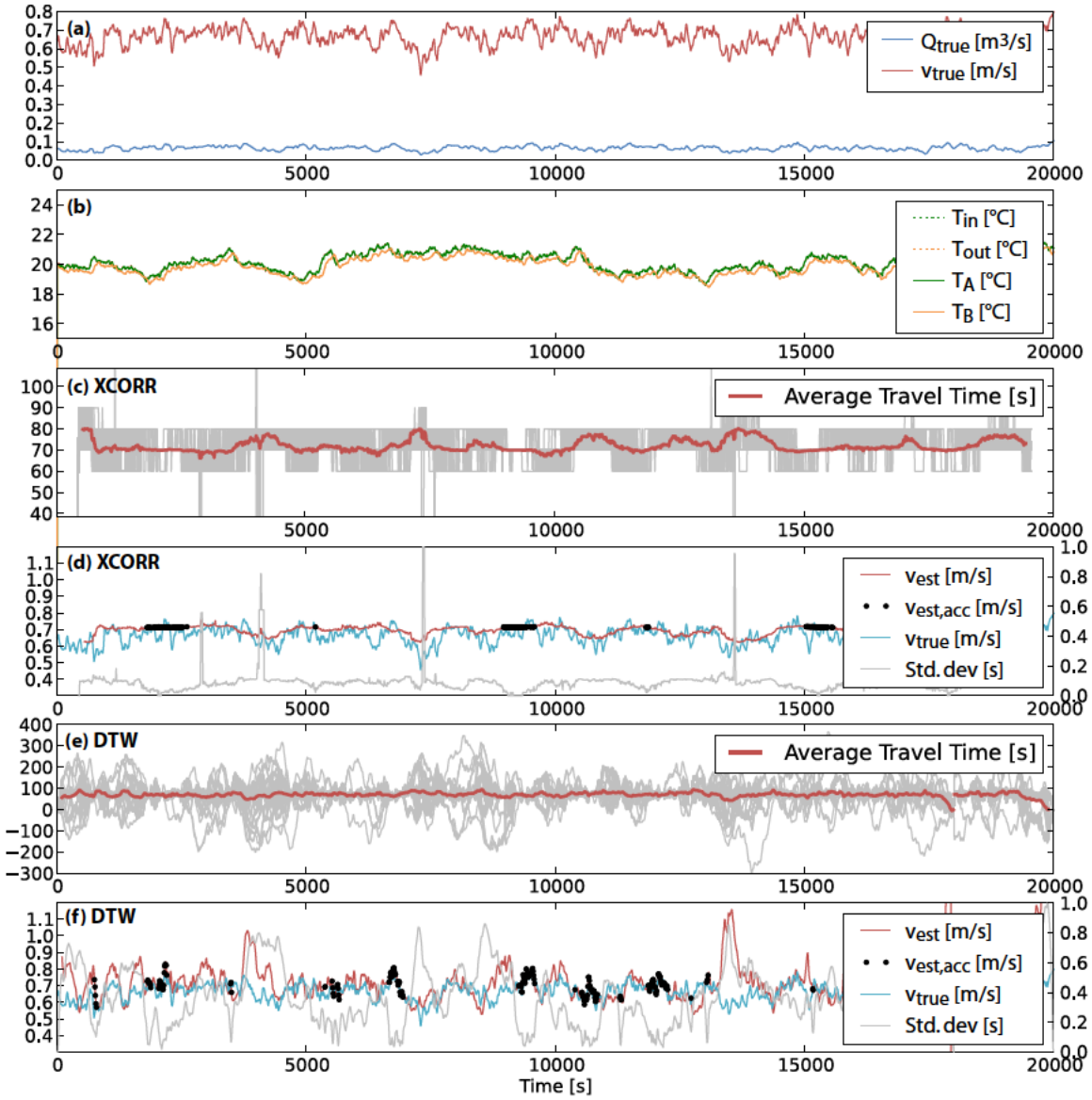

Figure 5: Estimation of the flow velocity with the XCORR and DTW methods. (a) Generated discharge and velocity series; (b) generated temperature series T_{in}, simulated series T_{out} and corresponding outputs of the sensor models, T_A and T_B, respectively. Note that only the sensor response is visible, because on this time scale the difference to the true signals is negligibly small; (c) estimated travel times using XCORR for all realizations as well as the averaged travel time (bold red line); (d) comparison of the estimated velocity using XCORR with the true synthetic velocities, values are accepted (filled black circles) when the performance statistic (green line, right axis) fulfils the given criterion. (e) estimated travel times using DTW, (f) comparison of DTW estimates to true velocities. Parameter values and acceptance criteria are given in Table 1.

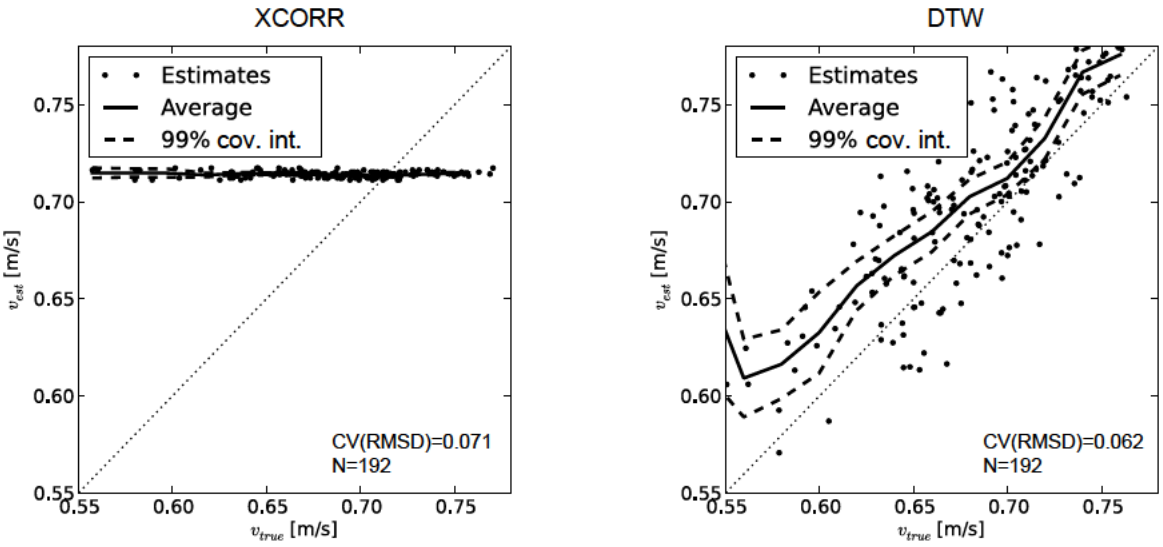

Figure 6: Comparison of the estimated velocity v_{est} with the true velocity v_{true} for the XCORR (left) and DTW method (right). Accepted data points are indicated, as well as the weighted average with the 99% coverage intervals. For this figure, a total of N values within the 10% percentile of the standard deviation of the paths were accepted.

that dispersion or reaction will significantly decrease the performance.

Second, we investigated the effect of sensor response time (T_{90}) and random measuring error (σ_e) on the performance (Figure 7b). Although, in the simulations, both sensors had the same characteristics, the realizations of the random errors were different in each iteration. It can be seen that the measuring errors have a stronger impact on the results than the response time of the sensors.

Third, we found that the volatility σ_{OU} has a larger effect than the mean reversion rate λ_{OU} (results not shown). However, the effect of the variation in the influent signal was roughly the same over the entire parameter range.

In addition, we investigated the sensitivity of the performance to important parameters of the method, such as the chosen pre-processing, the sampling interval and the added noise in the repeated simulations.

Regarding preprocessing, which often plays a critical role in many data mining applications (Cios et al., 2007), we investigated the suitability of: i) normalizing the temperature signals to have zero mean and unit variance and ii) the calculation of the first-order derivative, which is equivalent to high-pass filtering. Since the major aim of the preprocessing technique is the equalization of the magnitude of both signals, we assessed the performance of the estimates conditional on the first-order reaction coefficient. As shown in Figure 7c, the selection of the calculation of the first-order derivative yields best performance for both methods.

For the sampling interval Δt, we only find negative effects where sampling intervals are too long in comparison to the travel time (not shown). If the manhole distance is short (L <100m), the sampling interval should not exceed 15s to avoid CV(RMSD)> 0.1.

Finally, the sensitivity to the value of σ_s for XCORR and DTW is shown in Figure 7d. Firstly, the CV(RMSD) has a minimum for $\sigma_s > 0$ °C, which justifies the use of the iterative simulations. For the DTW method, choosing $\sigma_s > 0.02$ °C leads to unsatisfactory performance. The behavior of XCORR is similar, although the performance decrease only for $\sigma_s > 0.04$ °C. This can be

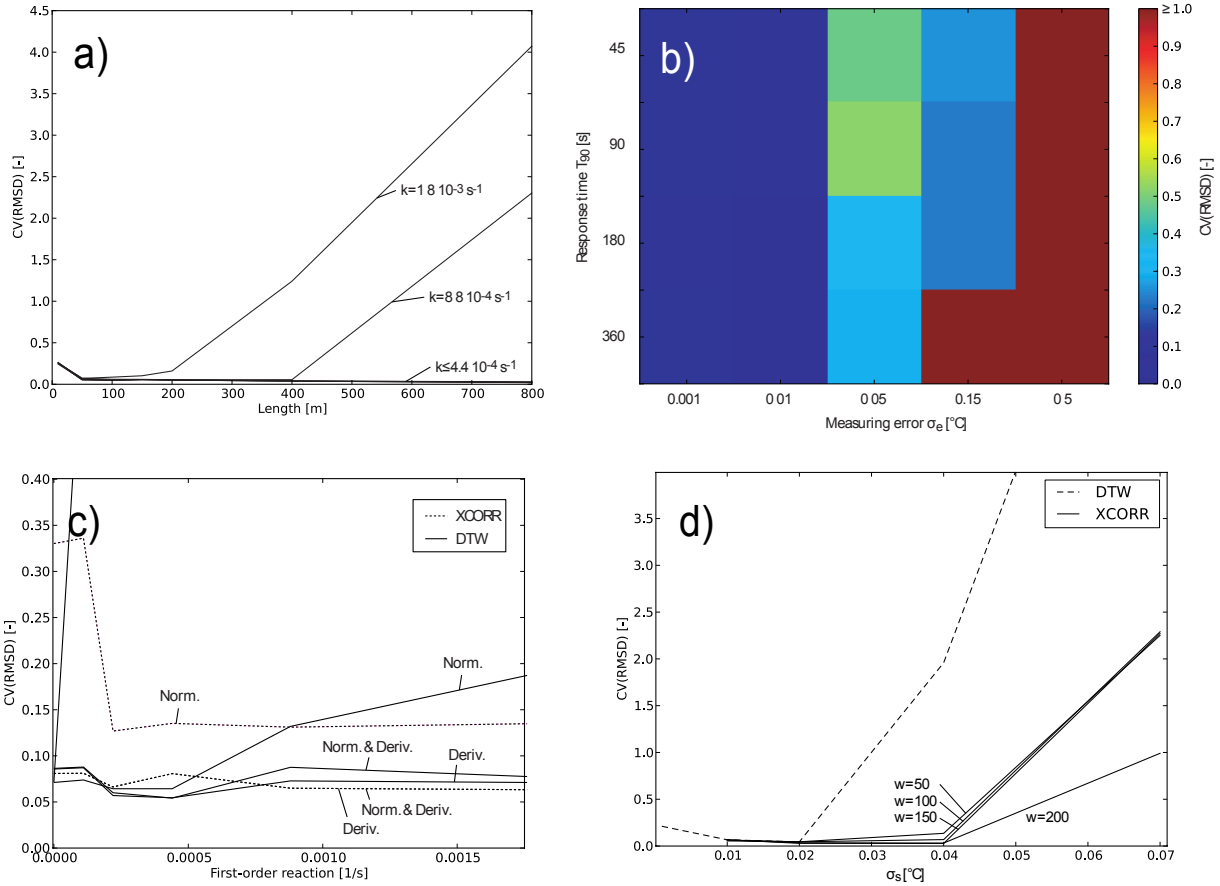

Figure 7: a) Analysis of the effect of dispersion (by varying the length of the sewer section) and reaction (by varying the first-order reaction coefficient) on the accuracy of flow velocity estimation with the DTW method; b) Accuracy of the DTW method for different values of sensor response time, T_{90}, and standard deviation of the measuring error, σ_e. For the analysis it was assumed that both sensors exhibit similar behavior, i.e. have the same values for T_{90} and σ_e; c) Analysis of the influence of the preprocessing for systems that exhibit different first-order reaction coefficients for the XCORR method (dashed lines) and the DTW method (solid lines). As preprocessing techniques, the calculation of the first order derivative ("Deriv.") and the normalization of the signal to have zero mean and unit variance ("Norm.") were considered. Note that the results for the XCORR method with derivation as preprocessing technique overlap with those with derivation and normalization as technique; d) Sensitivity of the DTW method (solid line) and the XCORR method with different window sizes w (dashed lines) on the standard deviation of the noise added during stochastic sampling, σ_s.

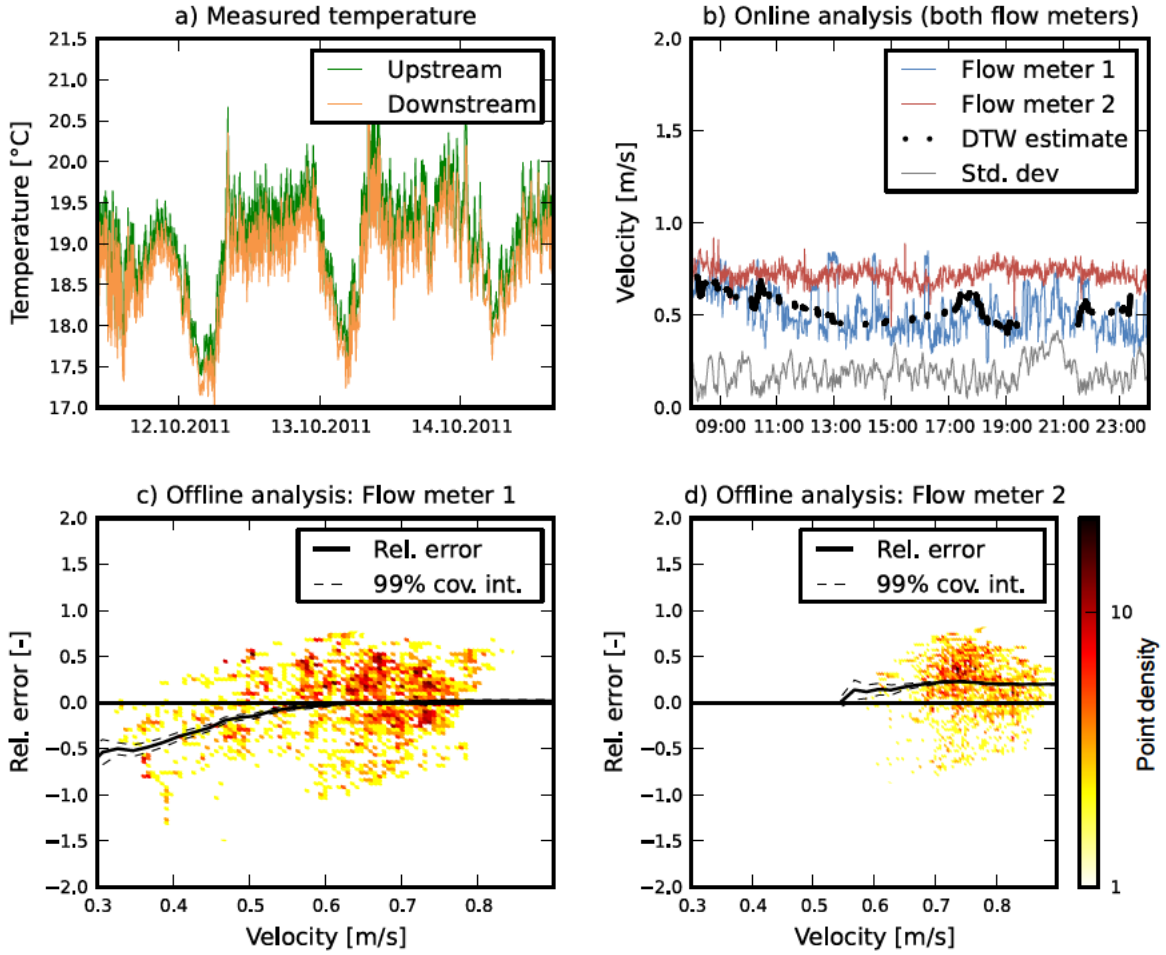

Figure 8: Diagnosis of the ultrasonic-doppler velocity probes of two flow meters. Short-term temperature data (a) are used to estimate the flow velocity as a function of time, which can be considered for online monitoring (b). The results of an off-line analysis, using long-term temperature data (c and d) suggest that Flow meter 2 overestimates average velocities during high flow velocities (point density cloud with logarithmic scale). The solid black line is the average, computed with Eq. (18). The dashed black lines are coverage intervals, as presented in Figure 6

explained by XCORR's requirement of rather large windows, which also means that the increase in CV(RMSD) for higher σ_s is more pronounced for smaller window sizes w. Further results (not shown) suggest that $30 < w < 100$ is an optimal window size for XCORR. At least for the investigated system, the performance is unstable for lower window sizes, and inaccurate for longer sizes due to averaging effects.

A.4.2 Case study

The results of the case study are shown in Figure 8. They indicate that the DTW estimates match the flow velocity measurements of "Flow meter 1" (FM 1) closely and rather deviate from the measurements of "Flow meter 2" (FM 2). In this regard, the results from the offline analysis are most conclusive. For example, they clearly demonstrate that Flow meter 2 not only fails to observe velocities lower than 0.5 m/s, it also exhibits a significant positive relative error.

A.5 Discussion

A.5.1 Numerical experiments and case study

Considering the results of the off-line analysis presented for the XCORR and the DTW method in Figure 6, a CV(RMSD) below 7.5 % was obtained for each optimal configuration. A closer look at the estimates that were actually accepted, however, reveals that XCORR only accepts velocities that are close to the average velocity. The DTW method, which produces more accurate results, does not show this behavior and provides estimates within the full range of values. This lies in the "local" nature of DTW, which, in contrast to XCORR, does not need to compare windows of sufficiently large sizes to provide a stable estimate. However, estimates based on large window sizes, calculated with XCORR in high temporal resolution, are only meaningful if the time lag within each window remains constant. Local warping obviously decreases the cross-correlation, which explains the fact that the range of the XCORR estimates is narrower.

The results of the case study show that a redundant measurement of the flow velocity from water quality measurements can help to assess the accuracy of two different flow meters. More specifically, the DTW estimates correspond better with the measurements of FM 1 than with those of the downstream flow meter FM 2. This corresponds to our expectations and might reflect the superior monitoring principle of the more modern device FM 1. However, these results must nevertheless be taken with care: Firstly, the estimates for the latter only contains accepted points in the range of 0.5 to 0.9 m/s while the estimates for the former range from 0.2 to 0.9 m/s. Secondly, both flow meters were installed at the same manhole, one slightly upstream and one slightly downstream of the shaft. Although the installation was professional and undertaken with the greatest care, it is thus not impossible that the upstream flow meter affects the flow field seen by the downstream meter.

A.5.2 DTW for flow velocity estimation

From a practical point of view, we believe that both methods are suitable to check the performance of flow meters, respectively their velocity measurements. Based on the above results, we would clearly recommend the DTW method over XCORR, although the latter has a slightly better computational performance. This is, because XCORR yields estimates in a limited range of values only with inferior or equal performance. In addition, it requires an additional parameter, the window size w. For safe and robust application, however, it must be ascertained that for all parameters of the DTW method, there are sufficiently accurate estimates available.

For example, a measuring error in the distance between the measuring locations, L, linearly propagates to the flow velocity v because $v = L/\theta$ (with hydraulic residence time θ). In practice, lengths of sewer stretches can typically be extracted from information systems or can be measured on maps or, strongly recommended, in the field. The sensitivity analysis revealed that the high-pass filtering is the best preprocessing method and the sampling intervals Δt of a few seconds guarantee a satisfactory performance. With regard to the stochastic approach, an increasing number of iterations Z stabilizes the results, although Dürrenmatt (2011) demonstrated that $Z = 100$ is often sufficient for wastewater systems. Finding a good value for parameter σ_s, in contrast, is less obvious. As shown in Figure 7d, ideal results can often only be obtained within a narrow range of σ_s. Since both parameters, Z and σ_s, only aim at smoothing and stabilizing the results, visual inspection of the result helps to determine whether parameter choices were adequate. For example, discontinuities pinpoint too few iterations while noisy individual

warping paths suggest a too high σ_s (cf. Figure 5c and e).

To accept or reject a velocity estimate we combined two distinct strategies. First, we used the modified Z-Score as a simple outlier statistics (cf. Section A.2.2). Second, we used the variance in the individual warping paths as a measure for the robustness of the computed results. From our experience, we suggest to accept a pre-defined percentage of the "best" values, which are no outliers and lie in clearly defined regions of the result space. While this choice is reasonable, it does not guarantee that the estimates are accurate, after all. High deviations in sensor measurements, as well as differences between deployed sensors, for example from clogging or fouling, would still provide "best", but false, estimates. Therefore, care has to be taken that none of the sensors clogs, or that the respective sections are eliminated from the data before analysis. Our experience shows that, where long time series are available, an off-line analysis can still be performed. On the one hand, sensors either clog seldomly or self-clean during high flows. On the other hand, a small fraction of heavily biased values disappears in the noise. From a practical point of view, the DTW calculations can be run on common desktop computers. In our case, the calculations for the case study took about 12 min for roughly 56,000 data points and Z=100 realizations.

A.5.3 Flow monitoring in sewers

Considering the results of the off-line analysis presented for the XCORR and the DTW method in Figure 6, a CV(RMSD) below 7.5 % was obtained for each case. These results indicate that the errors of the estimates are significantly lower than typical measuring errors of flow measuring methods as mentioned in the introduction. Because of the presented methods' simple experimental set-up, inexpensive sensors and a rather low amount of maintenance required, it seems to be a highly practicable approach to check velocity measurements by online or off-line diagnosis.

Instead of using temperature, our approach can be applied on any physical property, as long as it has near-conservative behavior. Using several tracers with different diurnal profiles, such as temperature and conductivity, should result in a gain of information and improve the performance during periods where a single natural tracer signal does not exhibit sufficient variations. Therefore, future developments should consider using simultaneous measurements of independent wastewater tracers, such as temperature and conductivity. In addition, using complementary information on water levels might even allow to directly infer sewer discharges based on an advection-dispersion model. Alternatively, a filtering algorithm could be implemented to continuously estimate parameter values of the applied model on the fly (Piatyszek et al., 2000). Also, nonparametric travel-time distributions could be determined by sophisticated deconvolution as suggested by Cirpka et al. (2007).

Our study demonstrates that it is possible to create redundant information on sewer flow velocities, which aids the diagnosis of erroneous sewer flow data. Sewer flow data are notoriously erroneous due to the harsh and hazardous environment as well as the lack of simple and fast methods for in-situ flow meter calibration. Unfortunately, erroneous data are particularly critical for extreme storm and flow events, which are often used in hydrodynamic modeling and the design of expensive urban drainage infrastructure.

A.6 Conclusions

Successful supervision and control of wastewater and storm water systems requires accurate sewer flow measurements. In this study, we investigated the potential of cross-correlation (XCORR) and dynamic time warping (DTW) to retrieve sewer flow velocities from online measurements of natural wastewater tracers and draw the following main conclusions:

- XCORR identifies the shift between two patterns, one measured upstream and one measured at a downstream boundary of a sewer section, by maximizing the cross-correlation. In contrast, DTW extracts travel times from the temporal shift between the two patterns. As it computes a non-linear warping path which minimizes the dissimilarity, it is conceptually superior.

- We comprehensively assessed the field of application of the method based on theoretical considerations and synthetic data. Synthetic data were generated from a new benchmark simulation environment that is able to generate virtual influent time series, numerically simulate sewer flow and transport and model sensor behavior. The results from numerical experiments on dispersion, reaction, reach length, sensor performance and other important influence factors show that pre-processing of the data is important and that tracer reaction in the sewer reach is critical. As dispersion is generally small, the distance between the sensors is less influential if it is known precisely. In general, the method should be very well suited for the conditions found in typical sewer systems.

- This was confirmed in a full-scale case study, where DTW was used to check the performance of two different flow meters. Based on the DTW results from temperature online measurements, we were able to show that one flow meter provided a more reliable velocity measurement. Our results suggest errors of less than 7.5% for DTW velocity estimates, which is in the low range of velocity and flow measuring errors reported in literature.

- Although theoretical analyses show that XCORR and DTW velocity estimates contain systematic errors due to dispersion and reaction processes, these are usually small and do not limit the applicability of the approach. Because of the simple set-up and low experimental costs for sensors and maintenance, we believe that our method is highly suitable to check sewer flow monitoring devices online or off-line.

A.A Derivation of the systematic relative error for a tanks-in-series model

To derive Eq. (7), which expresses the systematic error when approximating the real travel time by the observed travel time, $\theta = \hat{\theta}$, it is assumed that the discharge in the sewer, Q, is constant and that the flow is uniform. It is further assumed that there is a first-order degradation reaction, r, taking place.

Given these assumptions, the flow and mass transfer can be mathematically modeled by a tanks-in-series model which consists of a cascade of N continuous stirred-tank reactors (CSTRs) with equal and constant volumes (total volume V).

The mass balance of compound C over reactor j in a tanks-in-series model with $1 \leq j \leq N$

equal reactors assuming constant reactor volume $V_j = \frac{V}{N}$, is

$$\frac{dC_j}{dt} = \frac{1}{\theta_j}\left(C_{j-1}(t) - C_j(t)\right) + r_j(t) \tag{A.1}$$

where the hydraulic residence time (HRT) in the reactor is denoted by $\theta_j = \frac{V_j}{Q} = \frac{\theta}{N}$. The HRT of the entire cascade is θ, and r_j is the first-order reaction defined by $r_j = -kC_j(t)$ with the reaction constant k.
Let the influent discharge, Q, be constant while the influent concentration, C_0, periodically oscillates according to

$$C_0(t) = a\sin\left(2\pi f t + b\right) + c \tag{A.2}$$

where a is the amplitude, f is the frequency, b is the relative phase shift and c is an offset.
The set of ordinary differential equations $(\frac{dC_1}{dt}, \frac{dC_2}{dt}, ..., \frac{dC_N}{dt})$ that defines the tanks-in-series model has a closed-form solution for the given influent discharge and influent concentration. The asymptotic solution (independent of the initial conditions) for the effluent concentration of reactor N is

$$\begin{aligned} C_N(t) \;=\; & a\left(\frac{1}{\sqrt{(1+k\theta/N)^2+(2\pi f\theta/N)^2}}\right)^N \cdot \\ & \sin\left(2\pi f t + b - N\cdot\arctan\left(\frac{2\pi f\theta}{N+k\theta}\right)\right) + \\ & c\left(\frac{1}{1+k\theta/N}\right)^N \end{aligned} \tag{A.3}$$

Similar to the influent series in Eq. (A.2) (which is, in fact, the special case of $N = 0$), Eq. (A.3) too is a harmonic oscillation. However, if $N > 0$, the amplitude is lower, and when $k > 0$, the offset c decreases (and increases for $k < 0$). In addition, the effluent series exhibits an additional phase shift compared to the influent signal. The difference in the relative phase shift between the influent and the effluent signal divided by $2\pi f$ corresponds to the observed travel time and is given by

$$\hat{\theta} = \frac{N}{2\pi f}\arctan\left(\frac{2\pi f\theta}{N+k\theta}\right) \tag{A.4}$$

It is clear that $\hat{\theta} = \theta$ is only valid as $N \to \infty$, in which case the cascade approximates plug-flow behavior (Gujer, 2008).
If θ is approximated by $\hat{\theta}$ when $N \ll \infty$ and $k \neq 0$, a systematic error is introduced. The relative error $E_{\theta,rel}$ is defined as

$$E_{\theta,rel} = \frac{1}{\theta}\left(\theta - \hat{\theta}\right) \tag{A.5}$$

and, when applying Eq. (A.4) it is given by

$$E_{\theta,rel} = 1 - \frac{N}{2\pi f\theta}\arctan\left(\frac{2\pi f\theta}{N+k\theta}\right). \tag{A.6}$$

Although this theoretical analysis only holds for compounds for which the mass balance applies, the derivation for temperatures T, thus by formulating a heat balance, is straightforward.

Appendix B

The value of streamflow data in improving TSS predictions - Bayesian multi-objective calibration

A.E. Sikorska[a,b,c], D. Del Giudice[a,d], K. Banasik [a], J. Rieckermann[a].

[a]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b]University of Zurich, Dept. of Geography, 8057 Zürich, Switzerland
[c]Warsaw University of Life Sciences, Dept. of Hydraulic Engineering, 02-787 Warsaw, Poland
[d]ETHZ: Swiss Federal Institute of Technology Zürich, 8093 Zürich, Switzerland

Journal of Hydrology (under review).

Author contributions

A.S. conceived and designed the experiments, developed the method, collected the data, performed the analyses, wrote the paper; D.D.G. developed the method, wrote the paper; J.R. conceived the experiments; all coauthors gave advices, supported result interpretation, and paper revision.

Abstract

The concentration of total suspended solids (TSS) in surface waters is a commonly used indicator of water quality impairments. Its accurate prediction remains, however, problematic because: i) TSS build-up, erosion, and wash-off are not easily identifiable; ii) calibrating a TSS model requires observations of sediment loads, which are rare, and streamflow observations, to calculate concentrations; iii) predicted TSS usually deviate systematically from observations, an effect which is commonly neglected. Ignoring systematic errors during calibration can lead to overconfident (i.e. unreliable) uncertainty estimates during predictions. In this paper, we therefore investigate whether a statistical description of systematic model errors makes it possible to generate reliable predictions for TSS. In addition, we explore how the reliability of TSS predictions increases when streamflow data are additionally used in model calibration. A key aspect of our study is that we use a Bayesian multi-output calibration and a novel autoregressive error model, which describes the model predictive error as a sum of independent random noise and autocorrelated bias. Our results show that using a statistical description of model bias provides more reliable uncertainty estimates of TSS than before and including streamflow data into calibration makes TSS predictions more precise. For a case study of a small ungauged catchment, this improvement was as much as 15%. Our approach can be easily implemented for other water quality variables which are dependent on streamflow.

B.1 Introduction

Human-induced changes within a catchment often impair water quality in receiving rivers. To assess possible impacts of these changes on water condition and to propose mitigation strategies, various quality indicators are often assessed. Among these, the amount of total suspended solids (TSS) is one of the most commonly used indicators of surface water quality problems (e.g. MacDonald et al., 2000; Rossi et al., 2013). The reason for this is that TSS have a direct (physical, biological, and ecological) and indirect (toxicological) impact on aquatic ecosystems (e.g. Chebbo et al., 1995; Rossi et al., 2005; Sikorska et al., 2012a; Taylor and Owens, 2009; Walling, 2005). Thus, TSS is considered a good proxy for current water conditions and is useful to assess the risk of water quality hazard (Kişi, 2004; Parker et al., 2013). To assess such hazards, the TSS release to and its transport in rivers is commonly modelled with build-up/wash-off (BW) processes (e.g. Coutu et al., 2012b; Deletic et al., 1997; Moore, 1984; Zoppou, 2001). Such an approach is very promising because it imitates natural sediment processes as a function of dry periods, where sediment accumulates, and wet periods, where sediment is washed off. However, an accurate prediction of TSS in rivers is still problematic, because of three main reasons:

First, TSS build-up, erosion, and wash-off are a complex interplay of numerous real-world processes within the catchment and include many random and non-linear relations which are not well identifiable (Bertrand-Krajewski et al., 1993; Deletic et al., 1997). Therefore, the release of TSS into a stream is even less predictable than surface runoff generation (Schmelter et al.). Moreover, it is difficult to provide observation data with enough spatial resolution to model such a complex process. To do so, observational data at different points during the TSS generation process would be required, whereas only data at end-points (in surface waters) are usually collected.

Second, as for most environmental models, the accurate prediction of TSS requires observed data for calibration of a BW model and for water quality assessment, the concentration of TSS

(C_{TSS}) is often required. These data are usually rare and inaccurate because it is difficult to measure representative values for all conditions occurring in the river (Deletic et al., 1997; Walling and Webb, 1996). This representativeness has two aspects. On the one hand, TSS concentrations indicate a high variability across the river cross section and over time (Chebbo et al., 1995). Thus, observations at a single location may not properly capture current conditions, as opposed to punctual streamflows (Q) that can better describe flow conditions in the river channel. Consequently, observational errors of C_{TSS} may be substantial and much higher than for Q (McMillan et al., 2012). On the other hand, TSS in the river is sensitive to changes in the catchment. Thus, series that are too long may not reflect stationary conditions, which are usually assumed during parameter calibration (e.g. Merz, 2006, Petrow et al., 2007).

Third, predictions of a BW model, as any environmental model, are subject to uncertainty (Yang et al.; 2007b). It is recognized that this uncertainty is caused by the model's inability to reproduce observed patterns (model structural deficits), limited parameter identifiability (parametric uncertainty), errors in input data, e.g. rainfall (input uncertainty), and in output data for model calibration (observation uncertainty) (e.g., McMillan et al., 2012; Reichert and Schuwirth, 2012; Sikorska et al., 2014). Additional errors may occur, when the output of a BW model, the TSS load, must be converted to the concentration. This concentration is usually modelled as a function of Q in the river (Kişi, 2004), which for future conditions must be predicted with a hydrological model (Rode et al., 2010). Thus, inaccuracies in the description of hydrological processes will be mapped to sediment related processes (Banasik and Walling, 1996). This can be avoided only if uncertainty of TSS processes is properly acknowledged.

All these difficulties lead to problems in reliably calibrating a BW model, which turn into simulation errors in the prediction interface. Consequently, BW models are usually less accurate than hydrological models (Berretta et al., 2007; Bertrand-Krajewski et al., 1993). Yet, different uncertainty contributions have different effects on prediction uncertainties. While observation errors are typically well described with random errors, the presence of model structure deficits and input uncertainty lead to systematic errors, here referred to as bias. The first category of model errors can be accounted for with a typical regression approach. The second category requires more elaborate consideration. Finally, parametric uncertainty represents the third error category. It is linked to the model capability of reproducing the same output with different parameter values and to the lack of enough information in the calibration dataset. This error can be represented by defining prior parameter ranges or distributions. These three error categories contribute in different ways to the (total) prediction uncertainty of a BW (and a TSS) model. However, it remains unclear how these different error categories in BW models can be addressed adequately with respect to their properties and how they contribute to the TSS prediction uncertainty. Also, it is not transparent how predictions of a BW model can be improved with currently available calibration techniques given that TSS observational data are not generally available.

Given this, the feature of having more accurate streamflow observations, and as a result of this, more accurate hydrological models, could be potentially used to support predictions of less accurate water quality models such as BW models (Gupta et al., 1998). Some studies have attempted to use both hydrological and water quality models to improve water quality predictions by means of a multi-objective calibration (e.g., Efstratiadis and Koutsoyiannis, 2010; Gupta et al., 1998; Rode et al.,2007; Van Griensven and Meixner, 2007; Yapo et al., 1998). A multi-objective calibration can involve using multiple calibration sets, multiple objective functions, or multiple

outputs. Especially the latter concept might be beneficial in terms of TSS predictions because it gives a possibility of using (usually) more frequently available streamflow data as additional information in a BW model calibration. The advantage of such a multi-output approach is that it helps to provide more reliable estimates, usually represented by wider uncertainty bands, but at the same time also more precise than could be achieved with a single-output calibration only. Despite a poor accuracy of BW models, only a few studies have investigated the benefits of using one of these multi-objective approaches in improving TSS predictions (Bekele and Nicklow, 2007; Das and Haimes, 1979; Muleta and Nicklow, 2005; Rode et al., 2007; Sil and Choudhury, 2010; White and Chaubey, 2005). Most of these studies have optimized some statistical metrics or have used likelihood functions, making unrealistic assumptions on the output error distribution, usually using a traditional Gaussian error model. This type of error model assumes independence and sometimes variance stationarity on errors. However, it has been shown that errors of most environmental models are strongly correlated and heteroscedastic over time (Sikorska et al., 2012b; 2013; Yang et al.; 2007b). These properties of model errors require adequate statistical consideration because model estimates based on an inadequate model error description result in unreliable predictions (Brynjarsdóttir and O'Hagan, 2014). Yet, the application of multi-output calibration to BW models with a statistical description of their systematic model errors is still missing. Additionally, the value of using more frequently available streamflow data as additional information in BW model calibration remains unexplored.
To address these issues, we here propose a methodology to improve predictions of streamflow-dependent variables, such as C_{TSS}, by adapting a Bayesian multi-output calibration concept and by considering both systematic and random output errors. In our previous studies that focused on surface runoff predictions (Sikorska et al., 2012b; 2013) we did not attempt to separate structural deficits from observed errors nor considered multiple outputs together. Although a lumped error model can be justified to describe errors of streamflow and water level, which have shown very small observational uncertainties (e.g. Del Giudice et al., 2015b), it is suboptimal for representing TSS errors. Therefore, the objectives of our present study are three-fold:

I. We improve the fulfillment of assumptions on BW model errors by explicitly representing systematic errors using a statistical description of a bias.

II. Next, we provide more reliable estimates of TSS prediction uncertainty than before by explicitly acknowledging autocorrelation and heteroscedasticity of model errors with a novel error model (autoregressive error model). The advantage of this model over a traditional Gaussian model is that it better describes the present model errors, which are almost always autocorrelated.

III. We investigate whether we can reduce the predictive uncertainty of TSS by using Q data as additional information for calibration of BW models by means of a recently proposed Bayesian multi-objective approach.

The novelty of our work lies in: (1) Exploring the capability of a multi-objective approach to a simultaneous calibration of two dependent variables (multiple outputs) in this regard for the first time to model Q and C_{TSS}. (2) Adapting a Bayesian approach which uses a bias description to account for model systematic deviations and apply it for the first time to model river water quality. This has not been done so far for Q and C_{TSS}. (3) Investigating the value of Q data as additional information in multi-output calibration to provide more accurate TSS predictions.

We illustrate our approach with an example of a small urbanized catchment in Warsaw (Poland) using a conceptual hydrological and a BW model and six weeks of recorded data (precipitation, evapotranspiration, streamflow, TSS concentrations) with a 1-hour resolution. This paper is organized as follows. Next, we describe the methodology of our approach and three numerical experiments. After, we introduce the test catchment, the conceptual models, and available data. Finally, we present and discuss our results and draw our main conclusions and recommendations.

B.2 Methods

B.2.1 Stochastic description of a model and prediction error

Environmental model results systematically deviate from observations (Beck, 1991; 1991; Kennedy and O'Hagan, 2001; Wagener et al., 2003). Traditionally, these deviations are minimized by means of least squares error fitting which makes strong assumptions (independent and identically distributed errors, iid) that are usually violated. While this is not too critical for the best estimates, it plays a major role when interest lies in uncertainty intervals, which are flawed if the error model is inappropriate (see discussions in Dietzel et al., 2013; Neumann and Gujer, 2008; Yang et al., 2007b). Unfortunately, most uncertainty analysis approaches available in water quality modelling are performed with error models which implicitly assume such iid errors because they have convenient mathematical properties (e.g., Freni and Mannina, 2010; Mannina and Viviani, 2010; Parker et al., 2013; Schmelter et al., 2011; Yang et al.).

In this paper, we suggest to improve the representation of model errors by using a statistical description of model bias. Formally, this is modelled as an autocorrelated error-term in addition to the random errors, which are used to describe the uncertainty in the output measurements. Such a description better describes model errors and thus allows for more reliable estimates which are supported by the fulfillment of the underlying assumptions. Both systematic and random errors are modelled as additive to the output of the deterministic model (Del Giudice et al., 2013; Reichert and Schuwirth, 2012) giving a stochastic model output:

$$\tilde{\boldsymbol{Y}}\left(\boldsymbol{X},\boldsymbol{\theta}^{M},\boldsymbol{\theta}^{\epsilon}\right)=\boldsymbol{Y}_{M}\left(\boldsymbol{X},\boldsymbol{\theta}^{\boldsymbol{M}}\right)+\tilde{\boldsymbol{B}}_{M}\left(\boldsymbol{X},\boldsymbol{\theta}^{\epsilon}\right)+\tilde{\boldsymbol{E}}_{Y}\left(\boldsymbol{\theta}^{\epsilon}\right), \tag{1}$$

where $\boldsymbol{Y}_{M}\left(\boldsymbol{X},\boldsymbol{\theta}^{M}\right)$ is the output of the deterministic model M. $\tilde{\boldsymbol{Y}}\left(\boldsymbol{X},\boldsymbol{\theta}^{M},\boldsymbol{\theta}^{\epsilon}\right)$ is the output of the entire model, i.e., the combination of the deterministic model and the error model, and it mimics "true" but unobservable system response ($\boldsymbol{Q}$ or $\boldsymbol{C}_{TSS}$). As such, the model output $\tilde{\boldsymbol{Y}}\left(\boldsymbol{X},\boldsymbol{\theta}^{M},\boldsymbol{\theta}^{\epsilon}\right)$ is a random variable. It depends on external inputs, $\boldsymbol{X}$, parameters of the deterministic model, $\boldsymbol{\theta}^{M}$, and parameters of the error term: $\boldsymbol{\theta}^{\epsilon}$. $\tilde{\boldsymbol{B}}_{M}\left(\boldsymbol{X},\boldsymbol{\theta}^{\epsilon}\right)$ mimics model bias and lumps the effects of input and structural uncertainty, and $\tilde{\boldsymbol{E}}_{Y}\left(\boldsymbol{\theta}^{\epsilon}\right)$ represents random measurement noise, and together they describe the model prediction error $\tilde{\boldsymbol{\epsilon}}(\boldsymbol{\theta}^{\epsilon})$. The variables with the tilde represent random variables, while those with bold font are vectors. As an alternative to the additive description, a formulation of a multiplicative or a combined additive and multiplicative bias would be possible (Reichert and Schuwirth, 2012). As the two error terms have different characteristics, it is relatively easy to identify their parameters (Dietzel and Reichert, 2012; Reichert and Schuwirth, 2012). The additive formulation of the errors and the model might cause an identifiability problem, which can be, however, solved within a Bayesian framework (Gelman et al., 2003) (see section B.2.5).

B.2.2 Prediction and likelihood

We model the "true" system response as a random variable ($\tilde{\boldsymbol{Y}}$ in Eq. 1). For the prediction, given the input $\boldsymbol{X}$, the output of the stochastic model from Eq. 1 can be described by the predictive probability distribution $p(\tilde{\boldsymbol{Y}}|\boldsymbol{X})$. To calculate this distribution, we use a Bayesian approach, in which $p(\tilde{\boldsymbol{Y}}|\boldsymbol{X})$ is estimated by marginalizing the joint distribution of $\tilde{\boldsymbol{Y}}$ and all (model and error) parameters grouped into $\boldsymbol{\theta}$:

$$p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{X}\right) = \int p\left(\boldsymbol{Y}^o|\boldsymbol{\theta},\boldsymbol{X}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}, \tag{2}$$

where $p\left(\boldsymbol{\theta}\right)$ describes the prior knowledge of all parameters $\boldsymbol{\theta}$=$\{\boldsymbol{\theta}^M, \boldsymbol{\theta}^\epsilon\}$. $p\left(\boldsymbol{Y}^o|\boldsymbol{\theta},\boldsymbol{X}\right)$ is the likelihood function of the model M that measures the probability that data $\boldsymbol{Y}^o$ could be generated with the model M given the input $\boldsymbol{X}$ and a candidate parameter set sampled randomly from $p\left(\boldsymbol{\theta}\right)$. The calculation of $p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{X}\right)$ requires the prior and the likelihood for the model M to be specified.

B.2.3 Bayesian updating with calibration data

The distribution $p\left(\boldsymbol{\theta}\right)$ in Eq. 2 represents the knowledge about the parameters before considering any calibration data. If observed data for calibration, i.e., $\{\boldsymbol{X}^o;\boldsymbol{Y}^o\}$ become available, this distribution can be conditioned on the information contained in data, which results in the so-called posterior distribution of model parameters:
$p\left(\boldsymbol{\theta}|\boldsymbol{X}^o,\boldsymbol{Y}^o\right) \propto p\left(\boldsymbol{\theta}\right) \cdot p\left(\boldsymbol{Y}^o|\boldsymbol{\theta},\boldsymbol{X}^o\right)$.

B.2.4 Procedure for TSS model calibration and numerical experiments

To predict C_{TSS}, Q is needed. Thus, we construct a TSS model, which has a build-up/wash-off (BW) and a rainfall-runoff (RR) component. It thus has the same input as the RR model, usually precipitation which is zero during build-up (dry) periods, and two outputs: Q from the RR component and C_{TSS} from the BW component (Fig. 1a). We formulate a generic framework which treats the models as black boxes and thus any build-up/wash-off model and any rainfall-runoff model can be used as BW and RR. To test whether the reliability of the TSS predictions increases when Q data are used in model calibration, we propose three numerical experiments with different configurations of model structure and of calibration data (Tab. 1). These scenarios are built upon the most common procedures used to calibrate a TSS model for predicting the C_{TSS} and require dataset which can usually be gathered in the catchment.

Table 1: Comparison of investigated scenarios.

Scenario	Calibrated submodels	Calibration Input data	Calibration Output data	Input data for prediction
A	RR, BW	$\boldsymbol{P}^{o*}$	$\boldsymbol{C}^o_{TSS}$	$\boldsymbol{P}^{f*}$
B	BW	$\boldsymbol{Q}^o$	$\boldsymbol{C}^o_{TSS}$	$\boldsymbol{Q}^f$
C	RR, BW	$\boldsymbol{P}^{o*}$	$\boldsymbol{Q}^o$, $\boldsymbol{C}^o_{TSS}$	$\boldsymbol{P}^{f*}$
D**	RR	$\boldsymbol{P}^{o*}$	$\boldsymbol{Q}^o$	$\boldsymbol{P}^{f*}$

P is precipitation, Q is streamflow, C_{TSS} is TSS concentration. The superscripts ^{o} and ^{f} refer to the observed or to the future respective variable. The bold font indicates a vector. RR - rainfall-runoff model and BW - build-up/wash-off model. $^{*)}$ We refer here to precipitation as the most common input into a RR model. This input has to be adopted for each specific RR model if additional input variables are required, e.g. evapotranspiration or temperature. For details on the model structures applied in this study see Sect. B.3.2. $^{**)}$ Scenario D is used here only for a comparison with the RR model performance which is possible to obtain in a single-output calibration.

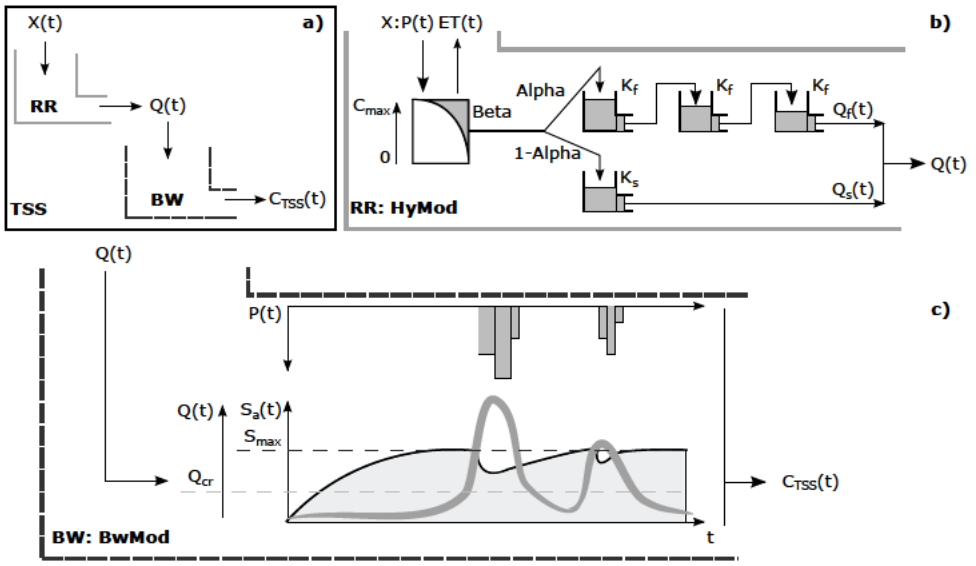

Figure 1: a) Schema of the TSS model with two components: rainfall-runoff (RR) and build-up/wash-off (BW); b) Rainfall-runoff model used in the study: HyMod; c) Applied build-up/wash-off model: BwMod.

In all three scenarios we use observed $\boldsymbol{C}^o_{TSS}$ as output, whereas the input varies in different scenarios. In scenario A we use only observed precipitation $\boldsymbol{P}^o$ as input (Sect. B.2.4). Instead, in scenario B we use only observed streamflow $\boldsymbol{Q}^o$ as input (Sect. B.2.4). Finally, in scenario C we use $\boldsymbol{P}^o$ as an input and then explore the advantage of using both variables i.e. $\boldsymbol{Q}^o$ as an (intermediate) output of the RR component and $\boldsymbol{C}^o_{TSS}$ (Sect. B.2.4). The first two scenarios rely on the single-output calibration, whereas the latter requires the multi-output calibration. For predictions we use precipitation in scenario A and C, and streamflow in scenario B.
In addition, we evaluate the fourth experiment (D), in which we only calibrate the RR model, with $\boldsymbol{P}^o$ as input and $\boldsymbol{Q}^o$ as output. This scenario relies on a single-objective calibration and thus is conceptually similar to scenario B, which models BW processes instead (see Sect. B.2.4 and Supplementary material). Scenario D is used only as reference and it illustrates the prediction performance for the RR model that can be achieved in a single-objective calibration (see also Tab. 1).

Scenario A: single-output calibration with precipitation as input and C_{TSS} as output

TSS model concept

This procedure is used if only C_{TSS} is of interest and Q predictions are not directly required. The TSS model, labelled as M_{TSS}, is calibrated with precipitation ($\boldsymbol{P}^o$) and $\boldsymbol{C}^o_{TSS}$ data only. Thus, Q is an intermediate state which is not inferred during the calibration. However, predicted $\boldsymbol{C}_{TSS}$ is diluted by the streamflow and thus also contains information on flow dynamics. Predictions only require $\boldsymbol{P}$ as an input. The output of the deterministic TSS model conditioned on $\boldsymbol{P}^o$ and model parameters $\boldsymbol{\theta}^{TSS}$, which is $\boldsymbol{C}_{TSS}\left(\boldsymbol{P}^o, \boldsymbol{\theta}^{TSS}\right)$, (Fig. 1a) is described as:

$$\boldsymbol{C}_{TSS}\left(\boldsymbol{P}^o, \boldsymbol{\theta}^{TSS}\right) = \mathrm{M}_{\mathrm{TSS}}\left(\boldsymbol{P}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{TSS}}\right) = \mathrm{M}_{\mathrm{BW}}\left(\mathrm{M}_{\mathrm{RR}}\left(\boldsymbol{P}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{RR}}\right), \boldsymbol{\theta}^{\mathrm{BW}}\right) = \mathrm{M}_{\mathrm{BW}} \circ \mathrm{M}_{\mathrm{RR}}\left(\boldsymbol{P}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{RR}}, \boldsymbol{\theta}^{\mathrm{BW}}\right). \tag{3}$$

M_{RR} and M_{BW} stand for the RR and the BW model while $\boldsymbol{\theta}^{RR}$ and $\boldsymbol{\theta}^{BW}$ stand for their parameters. $\mathrm{M}_{RR}\left(\boldsymbol{P}^o, \boldsymbol{\theta}^{RR}\right)$ is the output of the RR model, which is simply a modelled streamflow, and the input into the BW model. $\boldsymbol{\theta}^{TSS}$ represents the parameters of the TSS model.

The TSS concentration in surface water, $\boldsymbol{C}_{TSS}\left(\boldsymbol{P}^o, \boldsymbol{\theta}^{TSS}\right)$, in Eq. 3 represents the output of the deterministic model. To obtain a stochastic model output, $\boldsymbol{C}_{TSS}\left(\boldsymbol{P}^o, \boldsymbol{\theta}^{\mathrm{TSS}}\right)$ is combined with an output of the error model and can be described according to Eq. 1 as:

$$\tilde{\boldsymbol{C}}_{TSS}\left(\boldsymbol{P}^o, \boldsymbol{\theta}\right) = \boldsymbol{C}_{TSS}\left(\boldsymbol{P}^o, \boldsymbol{\theta}\right) + \tilde{\boldsymbol{B}}_{M_{BW} \circ M_{RR}}\left(\boldsymbol{P}^o, \boldsymbol{\theta}\right) + \tilde{\boldsymbol{E}}_{C_{TSS}}\left(\boldsymbol{\theta}\right), \tag{4}$$

where $\tilde{\boldsymbol{B}}_{M_{BW} \circ M_{RR}}\left(\boldsymbol{P}^o, \boldsymbol{\theta}\right)$ lumps the systematic prediction errors of C_{TSS} due to both the BW and the RR model components and $\tilde{\boldsymbol{E}}_{C_{TSS}}\left(\boldsymbol{\theta}\right)$ represents random errors of C_{TSS} observation. For simplicity we represent all parameters as $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{RR}, \boldsymbol{\theta}^{BW}, \boldsymbol{\theta}^{\epsilon}\}$.

Likelihood function

The likelihood of the TSS model describes the probability of observing data $\boldsymbol{C}^o_{TSS}$ given the model and its parameters which results in: $p\left(\boldsymbol{C}^o_{TSS} | \boldsymbol{\theta}, \boldsymbol{P}^o\right)$.

Calibration and prediction

The prior knowledge about parameters, $p\left(\boldsymbol{\theta}\right)$, is updated with recorded data $\{\boldsymbol{P}^o, \boldsymbol{C}^o_{TSS}\}$ to the posterior $p\left(\boldsymbol{\theta} | \boldsymbol{P}^o, \boldsymbol{C}^o_{TSS}\right)$ by using the likelihood. According to Eq. 2, the knowledge about the future realization of $\tilde{\boldsymbol{C}}^f_{TSS}$ conditioned on calibration data $\{\boldsymbol{P}^o, \boldsymbol{C}^o_{TSS}\}$ and future assumed input $\boldsymbol{P}^f$ will be described with the following predictive distribution:

$$p\left(\tilde{\boldsymbol{C}}^f_{TSS} | \boldsymbol{P}^o, \boldsymbol{C}^o_{TSS}, \boldsymbol{P}^f\right) = \int p\left(\boldsymbol{C}^f_{TSS} | \boldsymbol{\theta}, \boldsymbol{P}^f\right) p\left(\boldsymbol{\theta} | \boldsymbol{P}^o, \boldsymbol{C}^o_{TSS}\right) d\boldsymbol{\theta}. \tag{5}$$

Scenario B: single-output calibration with Q as input and C_{TSS} as output

TSS model concept

In scenario A, streamflow was only an internal state of the TSS model and not an output. Thus, it was not used for the model calibration. In scenario B, we alternatively use observed streamflow $\boldsymbol{Q}^o$ instead of observed precipitation as an input for the TSS model. As calibration output data we use only $\boldsymbol{C}^o_{TSS}$. Thus, only the BW component is inferred (Fig. 1a) and Eq. 3 can be rewritten and simplified to:

$$\boldsymbol{C}_{TSS}\left(\boldsymbol{Q}^o, \boldsymbol{\theta}^{TSS}\right) = \mathrm{M}_{\mathrm{BW}}\left(\boldsymbol{Q}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{BW}}\right). \tag{6}$$

Using Eq. 6, in which $\boldsymbol{C}_{TSS}\left(\boldsymbol{Q}^o, \boldsymbol{\theta}^{TSS}\right)$ describes only the output of the deterministic TSS model, the predicted stochastic output, $\tilde{\boldsymbol{C}}_{TSS}\left(\boldsymbol{Q}^o, \boldsymbol{\theta}\right)$, can now be described according to Eq. 1 as:

$$\tilde{\boldsymbol{C}}_{TSS}\left(\boldsymbol{Q}^o, \boldsymbol{\theta}\right) = \boldsymbol{C}_{TSS}\left(\boldsymbol{Q}^o, \boldsymbol{\theta}\right) + \tilde{\boldsymbol{B}}_{M_{BW}}\left(\boldsymbol{Q}^o, \boldsymbol{\theta}\right) + \tilde{\boldsymbol{E}}_{C_{TSS}}\left(\boldsymbol{\theta}\right). \tag{7}$$

The systematic prediction error in Eq. 7, $\boldsymbol{B}_{M_{BW}}\left(\boldsymbol{Q}^o, \boldsymbol{\theta}\right)$, now describes only the bias of the BW component. $\boldsymbol{E}_{C_{TSS}}\left(\boldsymbol{\theta}\right)$, in a similar way as in scenario A, represents the observation errors of C_{TSS}. $\boldsymbol{\theta}$ again lumps all parameters, i.e. $\boldsymbol{\theta}^{RR}$, $\boldsymbol{\theta}^{BW}$ and $\boldsymbol{\theta}^{\epsilon}$. This procedure is suitable only if high quality streamflow data are available and thus the observational error of streamflow can be assumed to be negligibly small, which is the case in this study.

Likelihood function

The likelihood of the TSS model is now: $p\left(\boldsymbol{C}^o_{TSS} | \boldsymbol{\theta}, \boldsymbol{Q}^o\right)$.

Calibration and prediction

The prior $p(\boldsymbol{\theta})$ is updated with streamflow and TSS data $\{\boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}\}$ to the posterior: $p(\boldsymbol{\theta}|\boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS})$. Similar to Sect. B.2.4, the knowledge about the future realization of $\tilde{\boldsymbol{C}}^f_{TSS}$, conditioned on data $\{\boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}\}$ and future input $\boldsymbol{Q}^f$ is represented by:

$$p\left(\tilde{\boldsymbol{C}}^f_{TSS}|\boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}, \boldsymbol{Q}^f\right) = \int p\left(\boldsymbol{C}^f_{TSS}|\boldsymbol{\theta}, \boldsymbol{Q}^f\right) p\left(\boldsymbol{\theta}|\boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}\right) d\boldsymbol{\theta}. \tag{8}$$

Scenario C: multi-output calibration with precipitation as input, C_{TSS} and Q as output

TSS model concept

In scenario C we extend the approach from scenario A to two outputs: Q and C_{TSS} using precipitation $\boldsymbol{P}^o$ as the model input (Tab. 1). For making predictions, however, only precipitation is required. The TSS model now uses both components, i.e., the RR model and the BW model (Fig. 1a) giving the deterministic outputs as:

$$\begin{aligned} \boldsymbol{Q}\left(\boldsymbol{P}^o, \boldsymbol{\theta}^{RR}\right) &= \mathrm{M_{RR}}\left(\boldsymbol{P}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{RR}}\right) \quad \text{and} \\ \boldsymbol{C}_{TSS}\left(\boldsymbol{Q}, \boldsymbol{\theta}^{BW}\right) = \mathrm{M_{BW}}\left(\boldsymbol{Q}\left(\boldsymbol{P}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{RR}}\right), \boldsymbol{\theta}^{\mathrm{BW}}\right) &= \mathrm{M_{BW}}\left(\mathrm{M_{RR}}\left(\boldsymbol{P}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{RR}}\right), \boldsymbol{\theta}^{\mathrm{BW}}\right) = \\ &\mathrm{M_{BW}} \circ \mathrm{M_{RR}}\left(\boldsymbol{P}^{\mathrm{o}}, \boldsymbol{\theta}^{\mathrm{RR}}, \boldsymbol{\theta}^{\mathrm{BW}}\right) \end{aligned} \tag{9}$$

The predicted stochastic outputs $\tilde{\boldsymbol{Q}}(\boldsymbol{P}^o, \boldsymbol{\theta})$ and $\tilde{\boldsymbol{C}}_{TSS}(\boldsymbol{Q}, \boldsymbol{\theta})$ may be described according to Eqs. 1 and 9 as:

$$\begin{aligned} \tilde{\boldsymbol{Q}}(\boldsymbol{P}^o, \boldsymbol{\theta}) &= \boldsymbol{Q}(\boldsymbol{P}^o, \boldsymbol{\theta}) + \tilde{\boldsymbol{B}}_{M_{RR}}(\boldsymbol{P}, \boldsymbol{\theta}) + \tilde{\boldsymbol{E}}_Q(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{C}}_{TSS}(\boldsymbol{Q}, \boldsymbol{\theta}) &= \boldsymbol{C}_{TSS}(\boldsymbol{Q}, \boldsymbol{\theta}) + \tilde{\boldsymbol{B}}_{M_{BW}}(\boldsymbol{Q}, \boldsymbol{\theta}) + \tilde{\boldsymbol{E}}_{C_{TSS}}(\boldsymbol{\theta}), \end{aligned} \tag{10}$$

where systematic errors of the RR and the BW model are described explicitly by $\tilde{\boldsymbol{B}}_{M_{RR}}(\boldsymbol{P}, \boldsymbol{\theta})$ and $\tilde{\boldsymbol{B}}_{M_{BW}}(\boldsymbol{Q}, \boldsymbol{\theta})$, respectively. $\tilde{\boldsymbol{E}}_Q(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{E}}_{C_{TSS}}(\boldsymbol{\theta})$ represent random observation errors of Q and C_{TSS}. $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{RR}, \boldsymbol{\theta}^{BW}, \boldsymbol{\theta}^{\epsilon}\}$ and $\boldsymbol{Q}$ is the streamflow predicted with the model RR.

Likelihood function

The likelihood of the TSS model is the product of both likelihoods of the RR and the BW model: $p(\boldsymbol{C}^o_{TSS}, \boldsymbol{Q}^o|\boldsymbol{\theta}, \boldsymbol{P}^o) = p(\boldsymbol{C}^o_{TSS}|\boldsymbol{\theta}, \boldsymbol{P}^o) \cdot p(\boldsymbol{Q}^o|\boldsymbol{\theta}, \boldsymbol{P}^o)$.

Calibration and prediction

Again, the distribution $p(\boldsymbol{\theta})$ describes the prior belief about all parameters. This prior is updated with recorded precipitation, streamflow and TSS data $\{\boldsymbol{P}^o, \boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}\}$ to the posterior: $p(\boldsymbol{\theta}|\boldsymbol{P}^o, \boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS})$. The future realization of $\tilde{\boldsymbol{C}}^f_{TSS}$ is predicted conditioned on future precipitation $\boldsymbol{P}^f$ and calibration data $\{\boldsymbol{P}^o, \boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}\}$ according to Eq. 2 as:

$$p\left(\tilde{\boldsymbol{C}}^f_{TSS}|\boldsymbol{P}^o, \boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}, \boldsymbol{P}^f\right) = \int p\left(\boldsymbol{C}^f_{TSS}, \boldsymbol{Q}^f|\boldsymbol{\theta}, \boldsymbol{P}^f\right) p\left(\boldsymbol{\theta}|\boldsymbol{P}^o, \boldsymbol{Q}^o, \boldsymbol{C}^o_{TSS}\right) d\boldsymbol{\theta}. \tag{11}$$

B.2.5 Bias consideration in multi-output calibration

To account for systematic errors, we apply a statistical bias description as suggested in Reichert and Schuwirth (2012) to one output, C_{TSS}, in scenarios A and B and to multiple, here two,

outputs, i.e., C_{TSS} and Q in scenario C. A bias description for multiple outputs has already been successfully applied in lake water quality modeling (Dietzel et al., 2013) and hydrodynamic simulations (Del Giudice et al., 2015b). By the multi-output calibration we refer in this study to the simultaneous use of two (or more) prediction variables per time point to infer model parameters. Each of these prediction variables is an output of a different model component which may have a different inadequacy (bias) in describing the corresponding "true" variable. In this context, the multi-output concept is linked to a bias description by specifying the bias for each of the modelled outputs independently. Within the Bayesian framework it means that a prior on bias for each of the modelled variable needs to be defined separately. These priors should be defined in a way that expresses the need to avoid bias as much as possible and thus prefers the model over the bias to describe the observed data. Furthermore, this also reduces the potential identifiability problem between the model and the bias process. Optionally, by adding a weighting factor to the bias of multiple variables, one can choose how much bias one is willing to accept among these variables. For simplicity, we assign here the same relative priors on bias parameters for both outputs which means we do not prefer a priori any output to be predicted better than another.

B.2.6 Parametrization of the prediction error (bias+random noise)

Model bias $\tilde{B}_M$

To deal with the correlated systematic error, $\tilde{\boldsymbol{B}}_M$ where M refers to the chosen model, we use a statistical description in which the bias is modelled as an Ornstein-Uhlenbeck (OU) process (Andersen et al., 2009; Platen and Bruti-Liberati, 2010; Uhlenbeck and Ornstein, 1930). By using a correlated bias, we explicitly acknowledge that the model cannot perfectly reproduce the observed variable. The correlation structure of $\tilde{\boldsymbol{B}}_M$ is chosen in such a way that it becomes similar to the autoregressive error (AR) model of Yang et al. (2008; 2007b), which however lacked the term $\tilde{\boldsymbol{E}}_Y$ (random noise). Thus, a bias term is modelled as a stationary Gauss-Markov process. Since we want our model to describe the data as best as possible, we assign a prior mean of 0 to the bias term. The stochastic differential equation describing the evolution of the bias term is

$$d\tilde{B}_M(t) = -\frac{\tilde{B}_M(t)}{\tau_M}dt + \sqrt{\frac{2}{\tau_M}}\sigma_{B_M}dW(t) \tag{12}$$

where bias parameters are defined as: τ_M - a correlation time, σ_{B_M} - a standard deviation, where M refers to the model component, here RR or BW. τ_M defines the time length in which the residual correlation is existant and is expressed in the units of the simulation time step while the σ_{B_M} is represented in the units of the modelled variable. $W(t)$ is a Wiener process (Brownian motion). Other bias descriptions are possible, for instance using a non-Markov Gaussian autocorrelated process or a standard Wiener process (see Del Giudice et al. (2013) for more details).

Random noise $\tilde{E}_Y$

The random noise $\tilde{\boldsymbol{E}}_Y$ from Eq. 1 is modelled as a white noise Gaussian process. Thus, no autocorrelation in time is assumed for this error. As we have no evidence to suggest biased observations, we describe $\tilde{\boldsymbol{E}}_Y$ as a variable normally distributed around a mean 0 with a standard deviation σ_{E_Y} which is expressed in the units of the modelled variable.

Output transformation $\psi()$

It has become a common practice in hydrology to apply a transformation, $\psi()$, on the modelled and on the observed variable, in order to help fulfill the assumptions of model errors. The advantage of the transformation over other approaches, which link the error variance to the value of the output variable (Renard et al., 2010) or of the external input (Del Giudice et al., 2013), is that the transformation makes it possible to obtain asymmetric uncertainty bands, which are more intuitive. In this study, we transform the output variables using the two-parameter Box-Cox transformation, which is the most commonly applied transformation function in hydrology (Box and Cox, 1964). Another recently used transformation function is a log-sinh (see Del Giudice et al. (2013) for comparison of both transformation functions). Generally, the success of applying the transformation function can be verified by an inspection of the assumption fulfillment on model residuals and realism of prediction intervals. For a comparison study of different scenarios, as in our case, the same transformation function must be chosen for all scenarios to allow for their fair comparison. Introducing output transformation implies that bias and the random noise should also be described in the transformed space.

B.2.7 The form of the likelihood function

As a likelihood in each of the implemented scenarios, we used a likelihood function, $p\left(\boldsymbol{Y}^o|\boldsymbol{\theta},\boldsymbol{X}\right)$, described by the Gaussian density centered in the deterministic model output $\boldsymbol{Y_M(\theta,X)}$ and transformed by a function $\psi()$, where M refers to the applied model, i.e. RR or BW.

$$p\left(\boldsymbol{Y}^o|\boldsymbol{\theta},\boldsymbol{X}\right) = \frac{\left(2\pi\right)^{-\frac{n}{2}}}{\sqrt{det\left(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{B}}_M+\tilde{\boldsymbol{E}}_Y}\left(\tilde{\boldsymbol{\epsilon}}\right)\right)}}\cdot$$
$$\cdot\exp\left(-\frac{1}{2}\left[\dot{\boldsymbol{Y}}^{\mathrm{o}}-\dot{\boldsymbol{Y}}_{\mathrm{M}}\left(\boldsymbol{\theta},\boldsymbol{X}\right)\right]^{\mathrm{T}}\cdot\boldsymbol{\Sigma}_{\tilde{\boldsymbol{B}}_{\mathrm{M}}+\tilde{\boldsymbol{E}}_{\mathrm{Y}}}\left(\tilde{\boldsymbol{\epsilon}}\right)^{-1}\cdot\left[\dot{\boldsymbol{Y}}^{\mathrm{o}}-\dot{\boldsymbol{Y}}_{\mathrm{M}}\left(\boldsymbol{\theta},\boldsymbol{X}\right)\right]\right)\cdot\prod_{\mathrm{i}=1}^{\mathrm{n}}\frac{\mathrm{d}\psi}{\mathrm{dy}}\left(\mathrm{y}=\boldsymbol{Y}_{\mathrm{i}}^{\mathrm{o}}\right) \tag{13}$$

In Equation above, $\dot{\boldsymbol{Y}}^o$ and $\dot{\boldsymbol{Y}}_M(\boldsymbol{\theta},\boldsymbol{X})$ stand for observations and the deterministic model output after applying the transformation function $\psi()$. $\boldsymbol{\theta}$ is the parameter vector, $\tilde{\boldsymbol{\epsilon}}$ is the error term which consists of the systematic error (bias) $\tilde{\boldsymbol{B}}_M$ and the random noise $\tilde{\boldsymbol{E}}_Y$. n and i represent the length of and the subscript over the calibration period. $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{B}}_M+\tilde{\boldsymbol{E}}_Y}$ is a covariance matrix and is described for each model and variable independently as:

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{B}}_M+\tilde{\boldsymbol{E}}_Y}\left(\tilde{\boldsymbol{\epsilon}}\right)_{i,j} = \sigma^2_{B_M}\cdot exp\left(-\frac{|t_i-t_j|}{\tau_M}\right)+\sigma^2_{E_Y} \tag{14}$$

Where τ_M and σ_{B_M} are the parameters of the bias in the M model, σ_{E_y} is the parameter of the random noise for the variable Y, i and j denotes subscripts over the calibration period and t is time.

For the multi-output calibration with ω number of outputs, the joint likelihood function has to be evaluated which simply results in a product of all likelihoods for each calibrated output:

$$p\left(\boldsymbol{Y}^o_a,...,\boldsymbol{Y}^o_\omega|\boldsymbol{\theta},\boldsymbol{X}\right) = \prod_{a=1}^{\omega}\left[p\left(\boldsymbol{Y}^o_a|\boldsymbol{\theta},\boldsymbol{X}\right)\right] \tag{15}$$

Where $\boldsymbol{Y}^o_a$ represents the observed variable with the subscript a and $p\left(\boldsymbol{Y}^o_a|\boldsymbol{\theta},\boldsymbol{X}\right)$ is its likelihood

function. In case of multiple outputs, the joint likelihood function can result in values close to zero. To deal with potentially small values of the multiplication, we use a log transformation on each likelihood.

B.2.8 Performance analysis

We assess the performance of the TSS model in the three scenarios based on the computed 95% predictive credibility intervals (95%-PIs) using three performance metrics; i) the fulfillment degree of the underlying statistical assumptions on the model errors, ii) data coverage of the 95%-PIs (DCOV) and iii) 95%-PIs sharpness measured by their average bandwidth, ABW, see also Supplementary material. Regarding i) we test the errors of the correlated error model for normality in the calibration, whereas for ii) and iii) we assess the PIs in the validation with the remaining available data which were not used for the inference. It is only meaningful to test the underlying statistical assumptions for the best (biased-corrected) model prediction and not for the whole Bayesian distribution, which additionally cumulates the uncertainty from the imprecise knowledge on model parameters (Del Giudice et al., 2013). Because the assumptions on model errors are made for the errors in calibration, the fulfillment of these assumptions is usually tested in this period. In contrast, it is usual to test PIs in validation because if the inference was successful, the coverage in calibration has to be equal or higher than a theoretical value. The model predictions are considered as reliable when the assumptions are not severely violated, and PIs are sharp and their data coverage is close to 95% (or higher). To assess the specific impact of the observation errors on the predictive uncertainty, we compute two kinds of 95%-PIs; first, only accounting for the parametric uncertainty of the TSS model and its bias, and second, also including the random error associated with the measurement error.

B.3 Material: case study

B.3.1 Research catchment and measured data

We test our approach on a small catchment of Sluzew Creek located in the Warsaw suburbs, Poland. Sluzew Creek is a third-degree watercourse and a tributary of the Wilanowka river, which flows into the Vistula River (see Fig. 2). The area of the analysed catchment is 28.7 km^2 with 41% agricultural and 59% urban land use. Due to progressing urbanization and following land use changes (see Supplementary material), Sluzew Creek experiences high sediment concentrations coming from surface wash-off during heavy summer rainfalls (see Sikorska et al. (2012b; 2013) or Sikorska and Banasik (2010) for more information).

As in most small catchments in Poland, there is no routine monitoring program in Sluzew Creek. Therefore, we performed our own monitoring in the summer of 2012 that consisted of 6 pluviometric stations, 1 evapotranspiration station and 1 streamflow gauge additionally equipped with continuous measurement of TSS concentration at the catchment outlet (Fig. 2). For the proof-of-concept, we used the selected meteorological and hydrological data over 6 weeks with 1 hour resolution. We used the first 3.5 weeks of observations to calibrate the model and validated the results on the remaining 2.5 weeks.

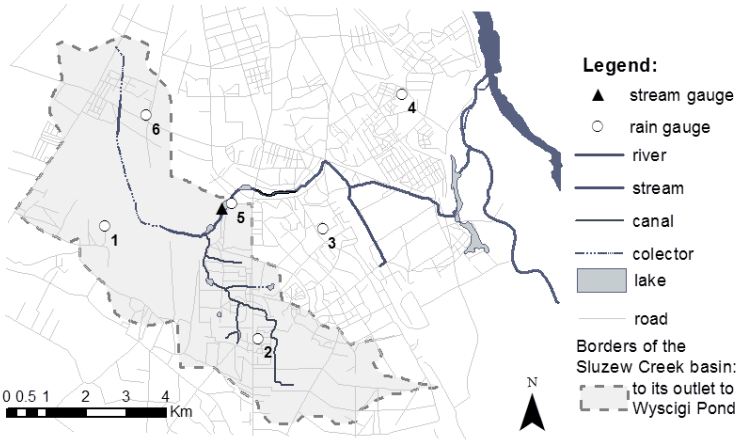

Figure 2: Location of the Sluzew Creek catchment, Warsaw.

B.3.2 TSS model

Rainfall-runoff (RR) model: HyMod

To model hydrological processes within the catchment, we chose the conceptual HyMod model (Boyle et al., 2000; 2004). Due to its simplicity and often satisfactory performance, HyMod is commonly applied in hydrological studies and uncertainty analysis (e.g. Bulygina and Gupta, 2011; Montanari and Koutsoyiannis, 2012; Vrugt et al., 2009a). HyMod is especially suitable to model lowland catchments, in which rainfall dominates runoff generation, as in our case. For catchments with a significant contribution of snow processes, other more suitable models can be applied, e.g. the HBV model (Seibert, 1999).

HyMod consists of a simple rainfall excess model, which is connected to two series of linear reservoirs to route surface and subsurface flow (Fig. 1b). The surface flow is routed through three fast flow reservoirs and the subsurface flow through a single slow flow reservoir. HyMod has five parameters which are: C_{max} $[mm]$ - the maximum storage capacity in the catchment; $Beta$ $[-]$ - the shape factor of the main soil water storage tank that represents the degree of spatial variability of the soil moisture capacity within the catchment; $Alpha$ $[-]$ - the factor distributing flow between two series of reservoirs; and K_f $[h]$ and K_s $[h]$ which represent the residence time of linear reservoirs in the fast and the slow flow series, respectively. The inputs into the HyMod are mean areal precipitation (P) and evapotranspiration (ET).

Build-up/wash-off (BW) model: BwMod

To model TSS transport in the creek, we applied a conceptual BwMod (Sikorska, 2013) model which is based on a common concept of sediment build-up/wash-off (BW) (Coutu et al., 2012a; Deletic et al., 1997; Moore, 1984; Zoppou, 2001). Within this concept we assume that sediments (TSS) are accumulated mostly during dry periods (build-up) on the catchment (mostly impervious) surface and next washed-off during rain events with the stormwater runoff (Fig. 1c). This is reasonable in a lowland partly urbanized catchment with a dominant rainfall contribution to the surface generation.

The BwMod has six parameters: $S_a(t)$ is the amount of sediment available to be washed-off $[kg]$ at time t; $Kappa$ $[-]$ is the sediment accumulation rate; S_{max} - the maximal amount of sediment that can accumulate within the catchment on impervious surfaces $[kg]$; Q_{cr} is the critical streamflow which has to be exceeded to trigger sediment wash-off; a [-] and b [-] are em-

pirical parameters of the wash-off function. For further details see the Supplementary material. Sediment build-up is modelled as an exponential function of time, whereas the wash-off process is described with a power law function that depends on the temporal streamflow in the creak. BwMod takes streamflow as an input to produce C_{TSS} at the catchment outlet. Dependent on the scenario, modelled or observed streamflow are used as an input.

B.3.3 Formulation of prior knowledge about the model parameters

Generally, formulating a prior is always a sensitive issue. From the assumption, the prior should reflect the best of our knowledge. Because in most cases, some knowledge about model parameters and model error is available, we therefore advise using an informative rather than non-informative prior. The benefit of using an informative prior is that it allows for a faster convergence towards the posterior and, in case of model bias, it helps to avoid the identifiability problem between the model and its error. Yet, if no prior information is available, we recommend to use conservative uniform distributions. In our study, we possessed enough information to formulate an informative prior. The priors for the HyMod and BwMod parameters were taken from previous studies on this catchment or were elicited from Polish experts who are familiar with the rainfall-runoff and erosion processes in Sluzew Creek (see Supplementary material). For the measurement uncertainty, given the knowledge on our measurement quality, we assume the zero-mean and a standard deviation equal to 10% of the maximal value observed for each modelled variable (C_{TSS} and Q). For the bias we assume rather wide priors bounded towards 0. Because bias is unlikely to be higher than the maximal variability noticed within the observed period, we took the maximal observed value as a bias standard deviation for each variable, which was also recommended by Del Giudice et al. (2013). Uncertainty in ET and P are considered implicitly by the model bias. For the output transformation $\psi()$, we use the two-parameter Box-Cox transformation with parameters $\lambda_1 = 0.5$ and $\lambda_2 = 0.001$ which proved to be efficient for this catchment (Sikorska et al., 2012b; 2013).

B.3.4 Implementation details

Both the HyMod and the BwMod model as well as the Bayesian inference were implemented in R (R Core Team, 2013). Both models were run with a 1-hour time step, which is reasonable for a partly urbanized catchment with the area of about 28 km^2. The HyMod and the BwMod are both based on a catchment memory scheme (discharge reservoir and sediment build-up and wash-off). Because we possessed the information on precipitation and streamflow for a period preceding the observed TSS concentration, we chose a starting point for the model calibration as the period directly after the last rainfall when its effect was no longer visible. Thus, we assumed that the catchment memory was reset.

The posterior probability distribution $p(\boldsymbol{\theta}|\boldsymbol{X}^o, \boldsymbol{Y}^o)$ was sampled with the adaptive Monte Carlo Markov Chain (MCMC) algorithm proposed by Haario et al. (2001) and implemented by Chivers (2012). This MCMC algorithm sequentially adapts the jump distribution of parameters and thus speeds up the convergence because less model runs are required compared to a traditional non-adaptive algorithm. To improve the identification process, we first searched for a global optimum by maximizing the posterior, as proposed by Del Giudice et al. (2015b). This optimum was next used as starting values for the adaptive MCMC algorithm. The convergence was achieved after about 50000 model runs. Next, from the obtained posterior, we chose a representative sample with 1000 random parameter sets, which were used to draw realizations of the stochastic output $\tilde{\boldsymbol{Y}}$. Based on those realizations, we computed the 95% predictive intervals.

B.4 Results

In general, the TSS model matched the observations well in all three scenarios (for the test on model errors see Supplementary material). We also found that using more information improves the prediction capacity of the TSS model and reduces its predictive uncertainty.

B.4.1 Posterior analysis

The learning process was informative for most of the BwMod parameters in all three scenarios (Fig. 3, second row), because their inferred distributions were shifted from the priors. Yet, inferred values slightly vary. When comparing different calibration datasets, the most informative for sediment processes were those used in scenario B which uses $\boldsymbol{Q}^o$ as an input and $\boldsymbol{C}^o_{TSS}$ as an output and scenario C with $\boldsymbol{P}^o$ used as an input and two outputs $\boldsymbol{C}^o_{TSS}$ and $\boldsymbol{Q}^o$. In both scenarios, their posteriors are slightly more strongly shifted than those in scenario A which used only $\boldsymbol{P}^o$ and $\boldsymbol{C}^o_{TSS}$ during the calibration.
Also, we were indeed able to learn about some HyMod parameters in scenario A (Fig. 3, top row) even without directly using streamflow data during the inference. However, this was much less than for scenario C in which $\boldsymbol{P}^o$ is explicitly used. The posterior of HyMod parameters in scenario B remains the same as the prior because the calibration data used in this scenario, i.e. $\boldsymbol{Q}^o$ and $\boldsymbol{C}^o_{TSS}$, does not contain any information for the hydrological model.
The inference was the most informative for the error models in all scenarios because among all parameters the posteriors of the error parameters showed a different mean compared to their priors. Interestingly, although posteriors on the HyMod and the BwMod parameters varied between different scenarios, the posteriors on error parameters of the sediment component ($\sigma_{B_{BW}}$, τ_{BW} and $\sigma_{C_{TSS}}$) are much more similar in all three scenarios (Fig. 3, third and fourth rows). Especially little has been learnt from the observations on the parameters of the observation error on C_{TSS}, i.e. $\sigma_{C_{TSS}}$. In contrast, the posterior on the systematic error ($\sigma_{B_{BW}}$, τ_{BW}) was slightly shifted in different scenarios. However, the comparison of these parameters in different scenarios in not straightforward and the reader is referred to Sect. B.5 for further discussion.
With respect to the hydrological component, its errors were identified only with the calibration data of scenario C, in which we explicitly acknowledged this error. In contrast, errors of the hydrological component for scenarios A and B were not explicitly modelled. As the error parameters (σ_{E_Q}, $\sigma_{B_{RR}}$ and τ_{RR}) were not considered during the inference, only the posteriors for scenario C are plotted in Figure 3.

B.4.2 Predictive uncertainty in three scenarios

The most reliable C_{TSS} predictions were obtained in scenario C (Fig. 6) because i) the data coverage (DCOV) is closest to the theoretical 95% (91.8% of validation data) and ii) the uncertainty bands were 15% narrower than in the second best scenario - A, see also Table 2. While the 95%-uncertainty bands in scenario A (Fig. 4) are wider, they cover less validation points (89.7%). The TSS model performed the worst in scenario B (Fig. 5) because the 95%-PIs were the narrowest but with the lowest data coverage which was equal to 89.1%. Outliers occur mostly during rain events when C_{TSS} is over-predicted. Also, the BwMod systematically overestimates C_{TSS} in scenario A for both dry and wet conditions, and in scenario B for wet conditions. The BwMod performs best in scenario C because it captures TSS dynamics fairly well. Although, peaks are usually slightly overestimated. As an example, during the biggest recorded event on the 9^{th} of July 2012, the peak of C_{TSS} is overestimated by almost two times more in scenario

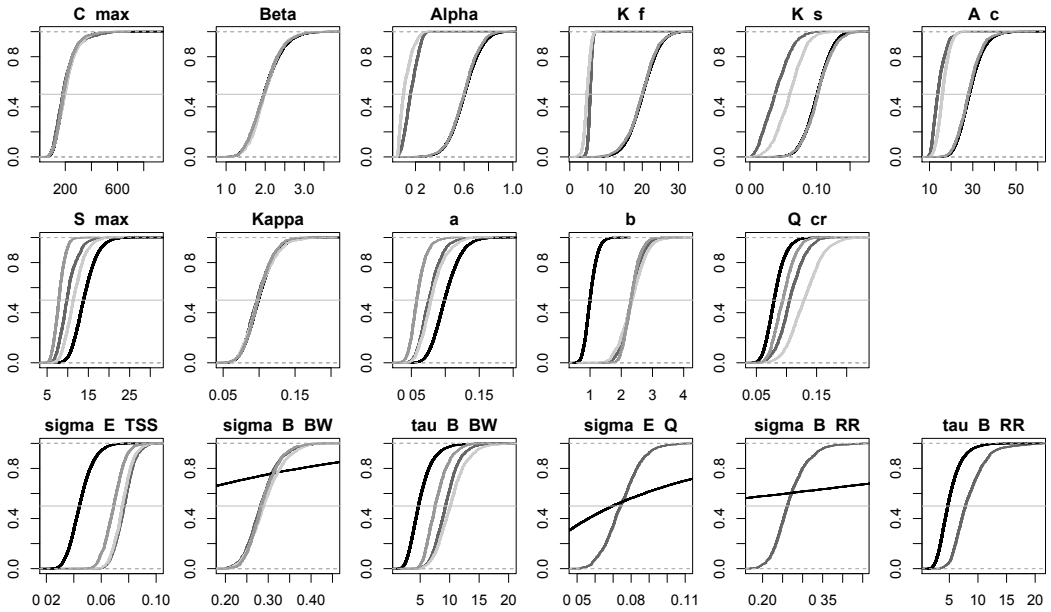

Figure 3: Parameter prior (black line) and posterior cumulative distribution functions (cdf) in scenario A (light grey), scenario B (moderate grey) and scenario C (dark grey). X-axes – parameter values, y-axes – cdfs. Top row - parameters of the HyMod; middle row - parameters of the BwMod; bottom row - parameters of the error model, where σ_{E_Y} refers to the Gaussian error model of the variable Y (C_{TSS} or Q) and parameters σ_{B_M} and τ_{B_M} describe the autoregressive error model of the model M (RR or BW). Note that the parameters of the error model for the RR and Q were only inferred in scenario C and thus are presented only for this scenario.

B (factor=5.4) than in scenarios A (factor=2.4) and C (factor=2.9). The comparison of three scenarios allows us to state that the multi-output calibration together with the bias description results in more reliable prediction uncertainty of C_{TSS} and in a better coverage of the validation data than it is possible to obtain with the single-output calibration.

Table 2: Comparison of prediction performance in investigated scenarios

Predicted output	TSS concentration	$(\tilde{\boldsymbol{C}}_{\boldsymbol{TSS}})$	Streamflow	$(\tilde{\boldsymbol{Q}})$
Index performance in scenario	DCOV [%]	ABW $[g{\cdot}l^{-1}]$	DCOV [%]	ABW $[m^3{\cdot}s^{-1}]$
A	89.7	0.0390		
B	89.1	0.0276		
C	90.7	0.0332	91.2	0.145
D			93.6	0.127

DCOV - data coverage; ABW - average bandwidth. The bold font indicates a vector and the tilde - a random variable.

The results of applying the autoregressive bias model in all scenarios indicated that the most uncertainty is due to the model bias and less due to the measurement error because the uncertainty bands in the calibration period are very narrow and hardly visible. Yet, the PIs included most of the observation data points. This takes into consideration the fact that during the period when observations are available, our knowledge about the system response is very precise. This is seen in Figs. 4–6 as the best biased-corrected model response which closely matches the observed data in the calibration (compare with the model output without bias-correction). In contrast, in the validation no data are used for learning about the model error. This results in wider uncertainty bands and a worse model performance. Interestingly, the effect of the bias

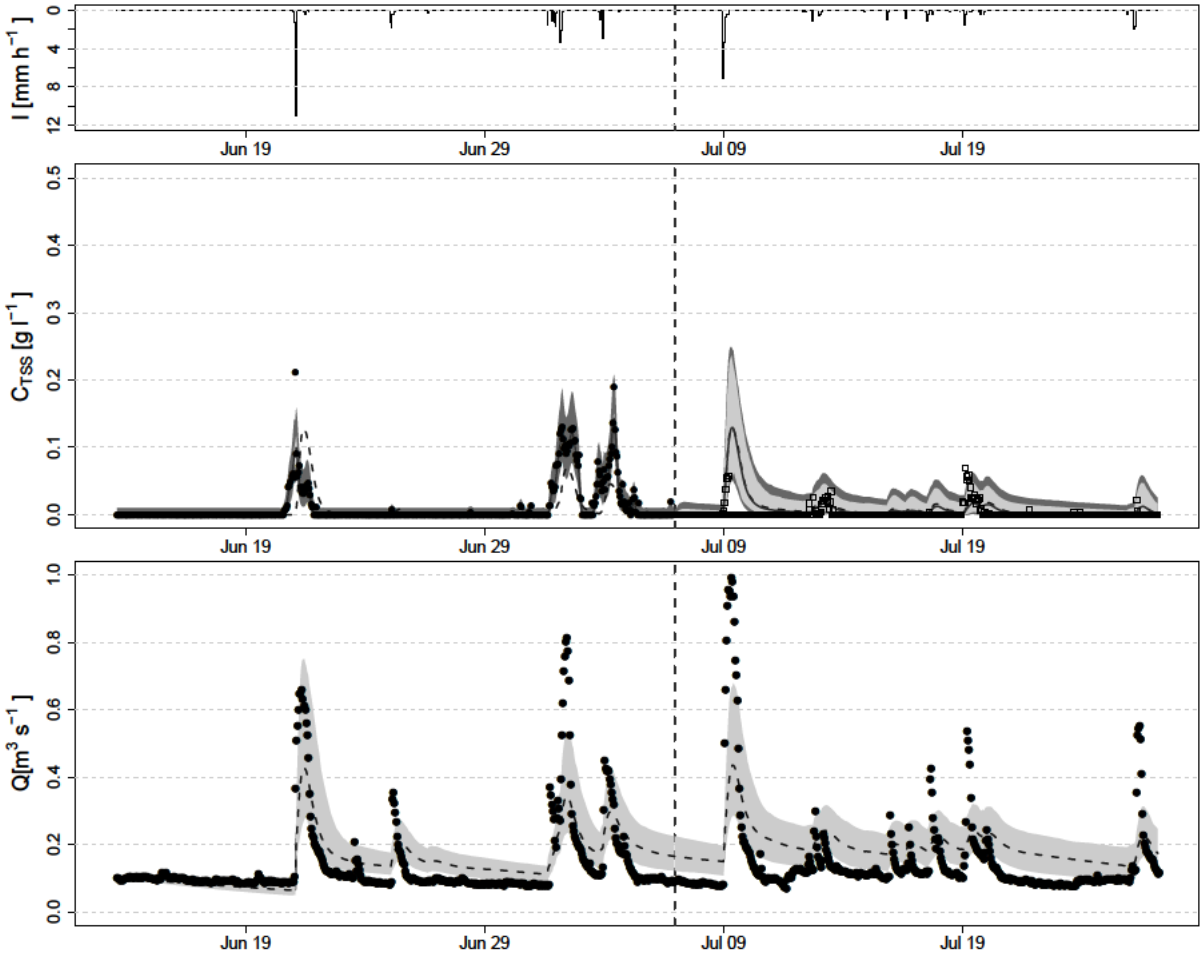

Figure 4: Scenario A: single-output calibration with precipitation as input and TSS concentration, C_{TSS}, as output. Top panel: rainfall intensity $[mm \cdot h^{-1}]$. Middle panel: prediction uncertainty of C_{TSS} $[g \cdot l^{-1}]$ in Sluzew Creek. Bottom panel: approximated streamflow Q $[m^3 \cdot s^{-1}]$. Notation: Gray polygons of C_{TSS} illustrate 95%-PIs due to parametric uncertainty and systematic errors (light) and also due to random noise (dark). Dashed line corresponds to the best simulation of the deterministic output computed with the optimized model parameter set, whereas the solid line represents the most probable bias-corrected model output which is interpreted as our best estimation. Black points depict observation and open dots validation points. Dashed vertical line cuts the validation from the observation period. In contrast, the uncertainty approximates of Q represent only the parametric uncertainty of the HyMod. The computation of predictive uncertainty of Q (with model bias) is not possible within this scenario.

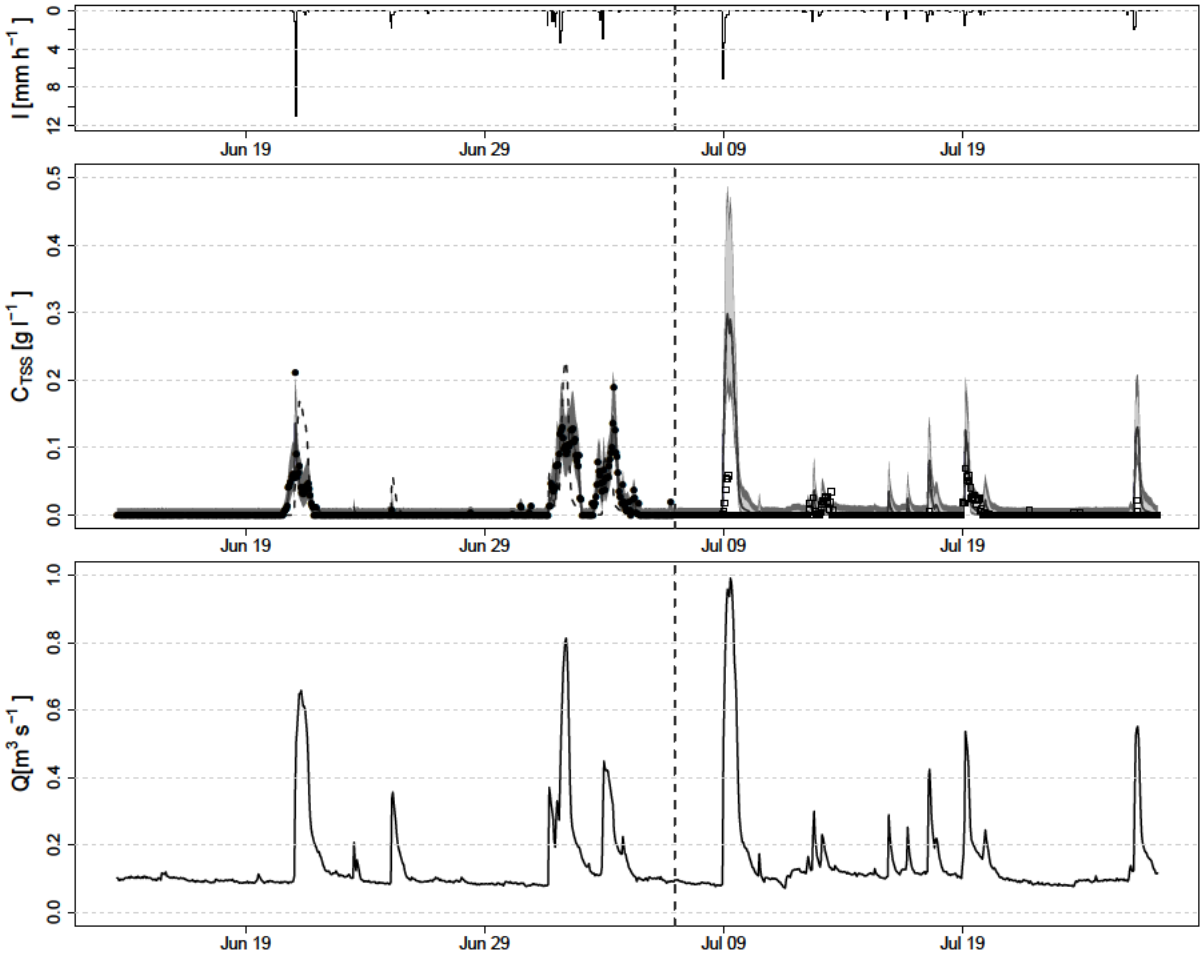

Figure 5: Scenario B: single-output calibration with streamflow as input and TSS concentration, C_{TSS}, as output. Top panel: rainfall intensity $[mm \cdot h^{-1}]$. Middle panel: prediction uncertainty of C_{TSS} $[g \cdot l^{-1}]$ in Sluzew Creek. Bottom panel: observed streamflow $[m^3 \cdot s^{-1}]$ used as an input for the TSS model.
Notation: Gray polygons of C_{TSS} illustrate 95%-PIs due to parametric uncertainty and systematic errors (light) and also due to random noise (dark). Dashed line corresponds to the best simulation of the deterministic output computed with the optimized model parameter set, whereas the solid line represents the most probable bias-corrected model output which is interpreted as our best estimation. Black points depict observation and open dots validation points. Dashed vertical line cuts the validation from the observation period.

correction is still noticeable at the beginning of the validation period.

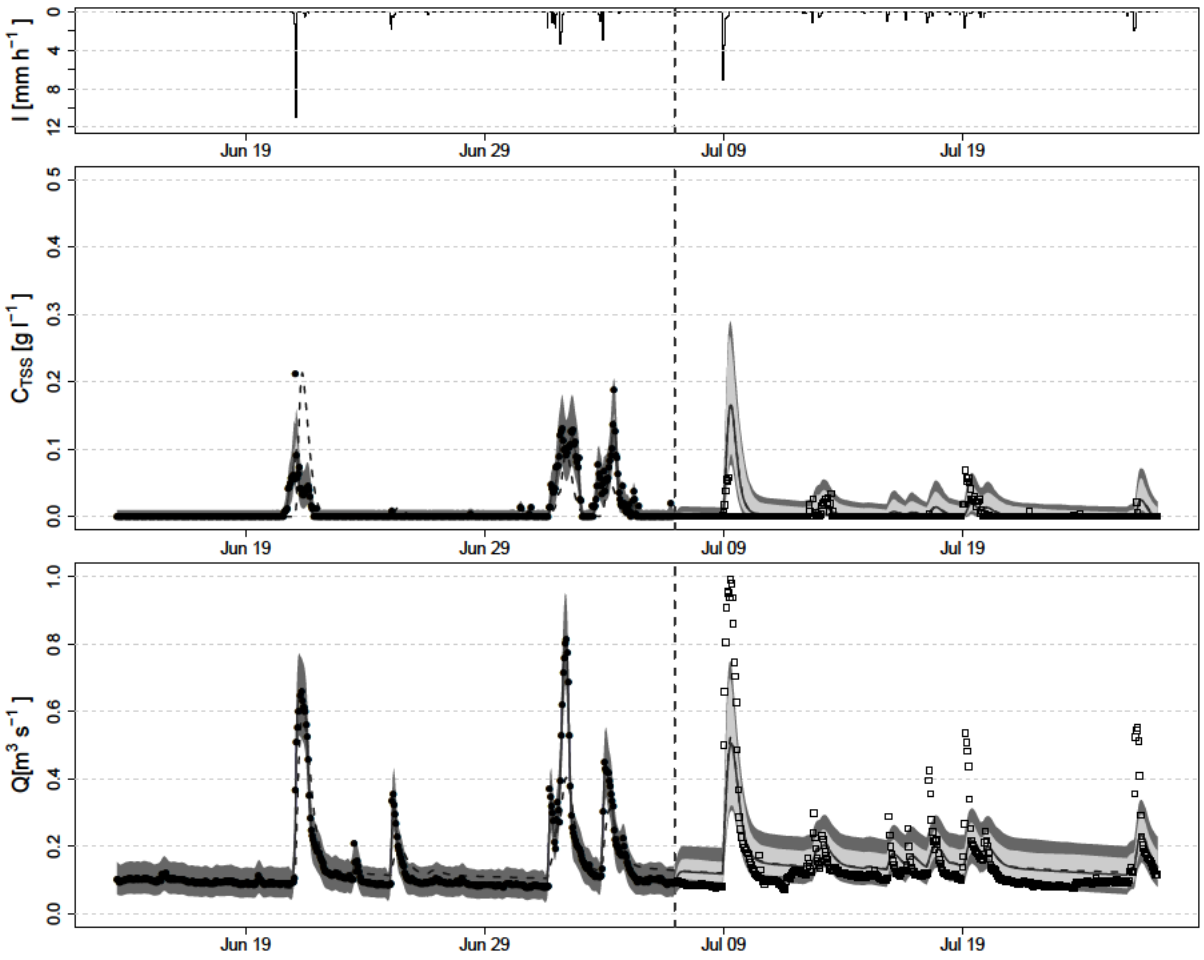

Figure 6: Scenario C: multi-output calibration with precipitation as input, TSS concentration, C_{TSS}, and streamflow, Q, as outputs. Top panel: rainfall intensity $[mm \cdot h^{-1}]$ in Sluzew Creek. Middle panel: prediction uncertainty of C_{TSS} $[g \cdot l^{-1}]$. Bottom panel: prediction uncertainty of Q $[m^3 \cdot s^{-1}]$.
Notation: Gray polygons illustrate 95%-PIs due to parametric uncertainty and systematic errors (light) and also due to random noise (dark). Dashed line corresponds to the best simulation of the deterministic output computed with the optimized model parameter set, whereas the solid line represents the most probable bias-corrected model output which is interpreted as our best estimation. Black points depict observation points, open dots - validation points. Dashed vertical line cuts the validation from the observation period.

A simultaneous multi-output calibration of the TSS model with two variables, C_{TSS} and Q, in scenario C gives the possibility to make explicit statistical statements on error terms of both hydrological and sediment processes. This was not possible in other two scenarios which rely on a single-output calibration and therefore represent errors in both processes with a single term. Thus, it is not surprising that the scenario C provides the best streamflow predictions (Fig. 6). The obtained performance of the HyMod model in this scenario was slightly worse than in the case when it is calibrated only against Q^o using a single-output calibration (scenario D in Table 2), see also Supplementary material. This is a very important finding for prediction purpose since the C_{TSS} is modelled as a function of Q. For the hydrological process, we identified substantial systematic errors and random noise with the help of the autoregressive bias model, similar to sediment erosion and wash-off processes. The findings regarding Q cannot be compared

quantitatively to those of other scenarios because streamflow was not predicted there.

B.5 Discussion

In this study we investigated, first, whether a statistical description of systematic model errors makes it possible to generate reliable predictions for C_{TSS} in receiving waters and, second, if the simultaneous multi-output calibration with additional streamflow data improves TSS estimates given limited C_{TSS} data for a single-output calibration. The results of our case study showed that using the autoregressive bias model leads to more reliable TSS predictions in all three scenarios investigated. Also, results from our comparative study suggest that precipitation data contain more information on the C_{TSS} dynamics than streamflow data, which, however, improves the identification of the TSS processes when used as additional information. In the following, we discuss: i) the informativeness of the different calibration datasets; ii) interpretation of the model bias in three scenarios; iii) benefits and limitations of the multi-objective calibration with bias description; iv) further challenges in predicting TSS.

i) Informativeness of diverse calibration datasets

All calibration scenarios provided rather reliable uncertainty bands as proved in Sect. B.4. Among all, the prediction uncertainty intervals (95%-PIs) of scenario C, which relies on the multi-output calibration approach, were the most reliable for Sluzew Creek. These PIs had data coverage (DCOV) which was closest to the desired 95% and they were 15% sharper than in the second best scenario, A. Moreover, this scenario also provided reliable streamflow predictions. This result showed that including $\boldsymbol{Q}^o$ data as additional information for calibrating the TSS model improves the identification of both hydrological and sediment processes. Thus, we recommend to use this scenario in other catchments, if $\boldsymbol{P}^o$, $\boldsymbol{Q}^o$ and $\boldsymbol{C}^o_{TSS}$ data are available for calibration of the TSS model. The second best results were obtained for scenario A which only calibrates the model on $\boldsymbol{P}^o$ and $\boldsymbol{C}^o_{TSS}$ data, while scenario B which uses $\boldsymbol{Q}^o$ as input performed the worst. This finding, together with the information on relatively small measurement errors of C_{TSS} and Q identified in all three scenarios, suggests that, on the one hand, the dataset including $\boldsymbol{P}^o$ and $\boldsymbol{C}^o_{TSS}$ is slightly more informative than the dataset including $\boldsymbol{Q}^o$ and $\boldsymbol{C}^o_{TSS}$. On first sight, this is a bit surprising because one would expect that streamflow (in which TSS is diluted) would have rather improved predictions. However, it has to be considered that the observed Q^o is usually computed from observed water levels with the use of another model, mostly in the form of a rating curve (Sikorska et al., 2013). The uncertainty associated with this transition is considered here not as the measurement uncertainty of Q but as input uncertainty lumped into the model bias. A worse TSS model performance obtained in scenario B would suggest that this input uncertainty of Q may be higher than the input uncertainty of observed precipitation. Another reason could be that, in this catchment, sediment dynamics are more closely related to rainfall dynamics and less to bottom shear stress or bank erosion by increased streamflow.

On the other hand, the better TSS model performance obtained in scenario A could be a result of a possible error compensation. The TSS model used in scenario B consists only of the BwMod while parameters of the HyMod are kept constant. In contrast, the TSS model in scenario A has more parameters as it consists of the HyMod and the BwMod, in which outputs are correlated. Because the HyMod parameters are adjusted without streamflow data, they increase the flexibility of the TSS model and thus partly compensate for its error. This resulted in a

better TSS model fit to C_{TSS} data. To reduce this effect of an error compensation, we advise against using non-informative priors on parameters of the hydrological model. Informative priors together with correlated outputs allow some information on both variables to be known, as it was in our case. Yet, this may lead to a poor model performance if calibration data contain only little information to inform both variables.
Finally, it should also be noted that the model performance in each dataset strongly depends on the data quality, data representativeness as well as the case study. In our case, to preserve the comparativeness of all three scenarios, we ensured that all calibration data were of high-quality.

ii) Interpretation of the model bias in three scenarios

As our results showed, the explicit consideration of the model bias helps to capture all systematic errors in modelling such as model structural errors, input uncertainty and all other remaining sources which are not explicitly acknowledged. Nevertheless, the interpretation of the bias is different for each scenario and thus its estimates are not comparable between scenarios. In scenario A, the bias parameters describe structural deficits of the combined TSS model, which are the lumped deficits of both the HyMod and the BwMod model and model input which is observed precipitation. With such bias description it is not possible to distinguish between the systematic errors of the HyMod and of the BwMod. In scenario B, the bias represents only the structural deficits of the BwMod model and model input, which was observed streamflow. In scenario C, as opposed to scenario A, the individual biases of the HyMod and the BwMod model are parametrized separately. In both cases bias parameters lump model structure deficits together with model input which is precipitation for the HyMod and streamflow for the BwMod model. Independent bias description for the HyMod and the BwMod makes it possible to identify the errors of both models, as opposed to scenario A. This information on individual systematic errors of each model can be useful in assessing which model could be improved. Yet, because in each scenario the bias represents the aggregated error term of different systematic error sources, assessment of their individual contributions is not possible in any of the scenarios. Thus, although the bias helps to obtain reliable predictions with phenomenological description of errors, it delivers only little insight into causes of the systematic model errors (Del Giudice et al., 2015b; Reichert and Schuwirth, 2012).

iii) Benefits and limitations of multi-objective calibration with bias description

We showed that the multi-objective calibration improves the predictive efficiency of the TSS model and that the bias description leads to more reliable uncertainty estimates. In our study we investigated four of the most common scenarios for predicting TSS with calibration of two outputs at the same time, i.e. Q and C_{TSS}. However, the methodology is directly transferable to model other pollutants which are diluted in streamflow. Also, it can be extended to model more than two outputs at the same time, using Eq. 15 directly.
The autoregressive model allowed for a bias correction in the calibration and in the validation, however, the observed effect of a bias correction differs in both periods. In calibration, correcting for bias resulted in very narrow uncertainty bands because errors (and bias parameters) can be estimated from the observed data at each time point. For the prediction, however, we do not know the observed values of the variable and thus we can only rely on the bias that we estimated during the calibration period while the "real" bias is not known. Because the autoregressive bias model carries the information on estimated model errors from the last observations into the extrapolation period, the effect of the bias correction can still be noticeable at the beginning of

the validation period, usually over a few correlation lengths. This effect vanishes when moving further from the calibration period since our knowledge on the bias becomes more uncertain. This also results in wider uncertainty bands.

Another important issue is related to a possible identifiability problem between the TSS model and the bias parameters. Because of the additive form of the bias, there is a danger of its overestimation if the model cannot sufficiently explain the data. In such a case, most uncertainty would be given to the bias, while the estimated parametric uncertainty of a TSS model is very low and thus, most likely underestimated. To avoid overestimation of the bias, it is important to specify an informative prior on the bias and its probability distribution in which it is preferable to attribute small (or zero) values for the model bias. This can be interpreted as that we give the priority to the TSS model over the bias to explain the observed data. In this context, a careful formulation of the prior on bias is the key factor to obtain both realistic calibration and prediction.

Yet, formulating informative prior on model bias remains the main challenge of the approach. The reason for this is that the bias usually compensates for different error sources, as in our case, and thus does not have a direct physical interpretation. To avoid miss-informative prior, we advise using information from previous studies on the same catchment. Another way of formulating an informative bias is by using expert knowledge. Because the bias is related to the model structure, a modeller usually has some notion on the possible errors and deficits of the model that he is using. However, it may still be difficult to represent such qualitative knowledge on model deficits, e.g. due to omitting some processes, in a quantitative way and even harder to express it as a probability distribution.

In terms of the multi-objective calibration, priors on bias parameters need to be defined for each of the modelled variable independently. An independent bias description gives the modeler the possibility to represent his knowledge on errors of each model component separately, which is more intuitive than formalizing the lumped bias for all submodels. This is an advantage over the single-output calibration where bias compensates for errors of both components and thus becomes less interpretable. As a special case, by alternating priors on both biases, the modeller can specify which variable he or she would prefer to calibrate better based on his/her experience, prior knowledge and prediction goals. As suggested in Reichert and Schuwirth (2012), this trade-off among objectives is transparent here, which was not the case in the traditional multi-objective calibration. Moreover, a multi-output bias approach allows for reliable prediction estimates and for calculation of uncertainty intervals for both variables i.e. C_{TSS} and Q. This is not possible with the single-output calibration, which allows for calculating uncertainty estimates only on C_{TSS}. This is important for risk assessment when both variables have to be precisely predicted.

An explicit bias consideration also allows for identification of systematic and random errors independently. Knowing these contributors, one can better plan strategy for uncertainty reduction in terms of improving process description (if bias dominates) or gathering more or of better quality calibration data (if measurement error dominates). Improving model description is, however, not straightforward because the bias here represents aggregated error of the model structure and input uncertainty. These contributors cannot be separated with the bias description that we used here. To do so, an explicit input uncertainty model should be implemented. For continuous modelling of TSS it is not straightforward because the input uncertainty is assumed to alternate over time, which requires a time-dependent approach.

A multi-objective calibration is, however, computationally more expensive than a single-

objective calibration. Thus, computation time may be limiting in practical studies when numerous or long-term datasets need to be analyzed because many model runs must be evaluated for both the HyMod and the BwMod model. This increase in the computation time is related to the additional parameters which also have to be calibrated. In our case, seven additional parameters of the HyMod and three parameters of the error model had to be inferred, which increased double the computation time compared to the single-output calibration. Moreover, both data have to be of high-quality, which may generate additional costs of obtaining the streamflow data for acquiring and calibrating rating curves, additional lab and field equipment, such as on-line probes, and their maintenance.

iv) Further challenges in predicting TSS

As our case study showed, the information content in availability of TSS data is rather low and the TSS prediction is very uncertain. Although including Q data generally improves TSS prediction, their accuracy is still rather low when compared to streamflow predictions. Thus, on the one hand, more frequent and more precise observations on C_{TSS} are needed to better calibrate a TSS model. In this context, a Bayesian calibration is very promising because it allows for a subsequent model updating when new data become available. On the other hand, there is still room for improvement of a description and modelling of sediment processes. For instance, the BW model that we applied is limited in modelling sediments only due to the wash-off induced by rainfall. However, sediments observed in the stream may also occur due to other processes, e.g. such as channel and bank erosion. These processes were not explicitly represented here. Yet, it must be considered that increasing the complexity of the model structure is usually associated with increased uncertainty due to the identifiability problem of increased number of model parameters.
An alternative solution to improve the description of the sediment processes are stochastic models. Such models rely on statistical information and thus may better imitate stochastic patterns of the catchment. However, they still require a formulation of the predictive uncertainty components, such as input, parameter or output uncertainty. For instance, Rossi et al. (2005) has proposed to incorporate statistical information on sediment processes from existing related studies to model TSS dynamics in an urban catchment as a probability function. Although this approach is promising, it is based on numerous field measurements, which are difficult to obtain for scarce data catchments.

B.6 Conclusions

In this study we investigated whether an explicit statistical description of bias improves predictions for TSS and how the simultaneous multi-output calibration with streamflow as additional data improves TSS estimates. To this end, we adapted a Bayesian multi-objective calibration with an autoregressive bias model and applied it for the first time to model TSS concentration and streamflow. To assess the additional value of this bias description, we performed three numerical experiments by using different datasets and TSS calibration strategies. From our comparative analysis, we derive the following conclusions and recommendations.

- Introducing the autoregressive bias model for representing the systematic model errors helped us to better match the underlying assumptions on the error models. This systematic error is introduced in addition to incomplete knowledge about the model parameters and output observation errors. Thus, in all three scenarios, we were able to provide more

reliable TSS predictions and uncertainty estimates than it was previously possible.

- Among three tested experiments, the multi-objective calibration with two variables, TSS concentration and streamflow, provided the most reliable TSS predictions with predictive uncertainty bands 15% sharper than in the next best scenario. Thus, we recommend this dataset with precipitation as input, and TSS concentration and streamflow as outputs for calibration in other catchments. The next best estimates were obtained for the dataset that used precipitation as input and TSS as output. Thus, in our case, precipitation data contained more information on TSS dynamics than streamflow data.

- We also found that, in our case, most uncertainty was due to the systematic errors of the TSS model, i.e. hydrological and build-up/wash-off model, and less due to the random measurement errors. This finding proves that TSS release to and transport in rivers is a complex stochastic process which is difficult to model.

- In this study, we focused on investigating the value of streamflow information to improve TSS predictions. However, our multi-output approach can be applied to improve predictions of other water quality indicators which are linked to streamflow dynamics or to other interdependent variables.

- The main challenge arises from the need to specify a reliable joint prior on the bias for both sediment and hydrological processes. This is a key factor to obtain reliable predictions. However, the formulation of an informative prior on bias is not a trivial task because it is not straightforward to interpret. Also, computation time in may be a limiting factor for complex models with numerous parameters because many simulations of both models are required to estimate the posterior distribution. Yet, with fast developments in statistical computing, this should not pose a problem.

Acknowledgments

Rectors' Conference of the Swiss Universities (CRUS) is acknowledged for financing this research within the scope of the Swiss-Polish Scientific Exchange Program NMS-CH (Grant No. 11.165). The authors are grateful to Andreas Scheidegger (EAWAG), Jan Seibert (University of Zurich), and three anonymous reviewers for their useful comments, which greatly improved the manuscript. The authors would also like to thank Peter Molnar (ETH Zurich) for consulting the build-up/wash-off (BwMod) model, Wiesława Przewoźniczuk (WULS-SGGW) for preparing the evapotranspiration data, and Tracy Ewen (University of Zurich) for proofreading the manuscript.

Curriculum Vitae

Academic Education

2011–2015	PhD in Environmental Engineering at Swiss Federal Institute of Aquatic Science and Technology (Eawag) and Swiss Federal Institute of Technology (ETH) Zurich, CH.
2009–2011	MSc in Environmental Sciences and Engineering at the Swiss Federal Institute of Technology (EPF) Lausanne, CH.
2007–2009	BSc in Environmental Sciences at the Università di Bologna, IT.
2008–2009	Exchange student in Sciences, Universidad de Granada, ES.
2006–2007	Student in Environmental Sciences at Università Parthenope, IT.

Professional Experience

2012–2015	Teaching assistant for Eawag Summer School in Environmental System Analysis, CH.
2011	Research assistant at Ecological Engineering Laboratory of EPF Lausanne, CH.
2010–2011	Teaching assistant for Water Quality Modeling, Quantitative Methods II, Informatics, and Air Pollution Modeling courses of EPF Lausanne, CH.
2010	Internship in hydrodynamic modeling at e-dric.ch, engineering firm, CH.
2009	Internship in air quality modeling at MED Ingegneria, research center, IT.

Peer-reviewed Publications

Under review	**D. Del Giudice**, C. Albert, J. Rieckermann, and P. Reichert. Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation.
Under review	A.E. Sikorska, **D. Del Giudice**, K. Banasik, and J. Rieckermann. The value of streamow data in improving TSS predictions - multi-objective Bayesian calibration.
2015	**D. Del Giudice**, R. Löwe, H. Madsen, P. S. Mikkelsen, and J. Rieckermann. Comparing two stochastic approaches to predict urban rainfall-runoff with explicit consideration of systematic errors, Water Res. Res., doi:10.1002/2014WR016678.
2015	**D. Del Giudice**, V. Bares, C. Albert, P. Reichert, and J. Rieckermann. Model bias and complexity - understanding the effects of structural deficits and input errors on runoff predictions, Env. Mod. & Soft, doi:10.1016/j.envsoft.2014.11.006.
2013	**D. Del Giudice**, M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, Hydrol. Earth Syst. Sci., doi:10.5194/hess-17-4209-2013.
2013	D. Dürrenmatt, **D. Del Giudice**, J. Rieckermann. Dynamic time warping improves sewer flow monitoring. Water Res., doi: 10.1016/j.watres.2013.03.051.
2012	S. Coutu, **D. Del Giudice**, L. Rossi, D. A. Barry. Parsimonious hydrological modeling of urban sewer and river catchments. J. Hydrol., doi:10.1016/j.jhydrol.2012.07.039.
2012	S. Coutu, **D. Del Giudice**, L. Rossi, D. A. Barry. Modeling of facade leaching in urban catchments. Water Res. Res., doi:10.1029/2012WR012359.