Improving uncertainty estimation in urban hydrological modeling by statistically describing bias

D. Del Giudice[1,2], **M. Honti**[1], **A. Scheidegger**[1], **C. Albert**[1], **P. Reichert**[1,2], and **J. Rieckermann**[1]

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

Correspondence to: D. Del Giudice (dario.delgiudice@eawag.ch)

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. 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 logsinh 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.

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, 2009; Korvin 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, 2009; Korvin 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., 2012). 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, the 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., 2012) 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 into a different direction of error separation, Sikorska et al. (2012) 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 input errors from 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 timescales 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 the following.

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 observed 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 data set 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 Methods

2.1 Likelihood function

To statistically estimate the predictive uncertainty of urban drainage models, we need a likelihood function (a.k.a. sampling model), $f\left(\boldsymbol{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x}\right)$, which combines (in this particular case) a deterministic model (a.k.a. simulator), M, with

a probabilistic error term. $f\left(\boldsymbol{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x}\right)$ describes the joint probability density of observed system outcomes, $\boldsymbol{y}_o$, given the simulator and error model parameters ($\boldsymbol{\theta}, \boldsymbol{\psi}$), and external driving forces, $\boldsymbol{x}$, such as precipitation. The probability density, $f\left(\boldsymbol{y}_o|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x}\right)$, 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 $\boldsymbol{\psi}$. A generic likelihood function assuming a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x})$ of output transformed by a function $g()$ can be written as

$$f(\boldsymbol{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x}) = \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{\det(\boldsymbol{\Sigma}(\boldsymbol{\psi}, \boldsymbol{x}))}}$$

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

where n is the number of observations, i.e., the dimension of $\boldsymbol{y}_o$, which could be, for instance, a sewer flow time series. $\boldsymbol{y}_{\mathrm{M}}$ are the corresponding model predictions. The tilde denotes transformed quantities, i.e., $\tilde{\boldsymbol{y}} = g(\boldsymbol{y})$. Note that Eq. (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 about the error distribution (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 variance–covariance matrix of the residuals Σ (Sects. 2.1.1, 2.1.2) and (ii) the transformation function g (Sect. 2.1.3). 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).

2.1.1   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, $\tilde{Y}_o$, is modeled as the sum of a deterministic model output $\tilde{y}_M(x, \theta)$ and an error term representing the measurement noise of the system response E:

$$\tilde{Y}_o(x, \theta, \psi) = \tilde{y}_M(x, \theta) + E(\psi), \tag{2}$$

where variables in capitals represent random variables, whereas those in lowercase are deterministic functions.

Assuming independent identically distributed normal errors in the transformed space, E follows a multivariate normal distribution with mean 0 and a diagonal covariance matrix,

$$\Sigma_E = \sigma_E^2 \, \mathbb{1}. \tag{3}$$

We then have $\psi = \sigma_E$, and the covariance matrix of Eq. (1) is given by $\Sigma = \Sigma_E$. As $E(\psi)$ 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.

2.1.2   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" behavior 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 modeling (Reichert and Schuwirth, 2012).

Technically, model inadequacy (also called bias or discrepancy) is considered by augmenting the independent error model with a bias term:

$$\tilde{Y}_o(x, \theta, \psi) = \tilde{y}_M(x, \theta) + B_M(x, \psi) + E(\psi). \tag{4}$$

On the one hand, this model bias B_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 modeled 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, B_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. (4). 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 favor 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, $B_M(x, \psi) + E(\psi)$, is not testable due to the Bayesian

interpretation of $B_M(x, \psi)$. However, when estimating both $B_M(x, \psi)$ and $E(\psi)$, the assumptions regarding the observation error, $E(\psi)$, 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. In contrast, 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 OU process (Uhlenbeck and Ornstein, 1930), the discretization of which would be a first-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

$$\mathrm{d}B_M(t) = -\frac{B_M(t)}{\tau}\mathrm{d}t + \sqrt{\frac{2}{\tau}}\sigma_{B_{ct}}\mathrm{d}W(t), \qquad (5)$$

where τ is the correlation time and $\sigma_{B_{ct}}$ 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. (1) with covariance matrix $\Sigma = \Sigma_E + \Sigma_{B_M}$ with

$$\Sigma_{B_M, i, j}(\psi) = \sigma_{B_{ct}}^2 \exp\left(-\frac{1}{\tau}|t_i - t_j|\right). \qquad (6)$$

In contrast to the formulation given by Eq. (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 exponent 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., 2012). 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

$$\mathrm{d}B_M(t) = -\frac{B_M(t)}{\tau}\mathrm{d}t + \sqrt{\frac{2}{\tau}\left(\sigma_{B_{ct}}^2 + (\kappa x(t-d))^2\right)}\mathrm{d}W(t), \quad (7)$$

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. (7) the recursion formula for the variance:

$$E\left[B_M^2(t_i)\right] = E\left[B_M^2(t_{i-1})\right]\exp\left(-\frac{2}{\tau}\Delta t\right)$$
$$+ \left[\sigma_{B_{ct}}^2 + (\kappa x_{i-\delta})^2\right]\left(1 - \exp\left(-\frac{2}{\tau}\Delta t\right)\right). \quad (8)$$

The parameters ψ of this bias are given by

$$\psi = (\sigma_{B_{ct}}, \tau, \kappa, \delta)^T. \qquad (9)$$

The resulting bias covariance matrix, Σ_{B_M}, is given by

$$\Sigma_{B_M, i, j}(\psi, x) = E\left[B_M^2\left(\min(t_i, t_j)\right)\right]\exp\left(-\frac{1}{\tau}|t_j - t_i|\right). \quad (10)$$

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 min. 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 time steps. In our experience, this component did not have a significant effect at this short timescale and its elimination led to a greater simplicity and robustness of the error model.

2.1.3 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

$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0, \end{cases} \quad (11)$$

$$g^{-1}(z) = \begin{cases} (\lambda z + 1)^{1/\lambda} & \text{if } \lambda \neq 0 \\ \exp(z) & \text{if } \lambda = 0, \end{cases} \quad (12)$$

$$\frac{dg}{dy} = y^{\lambda - 1}, \quad (13)$$

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

$$g(y) = \beta \log\left(\sinh\left(\frac{\alpha + y}{\beta}\right)\right), \quad (14)$$

$$g^{-1}(z) = \left(\text{arcsinh}\left(\exp\left(\frac{z}{\beta}\right)\right) - \frac{\alpha}{\beta}\right)\beta, \quad (15)$$

$$\frac{dg}{dy} = \coth\left(\frac{\alpha + y}{\beta}\right), \quad (16)$$

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 $\tilde{y}$,
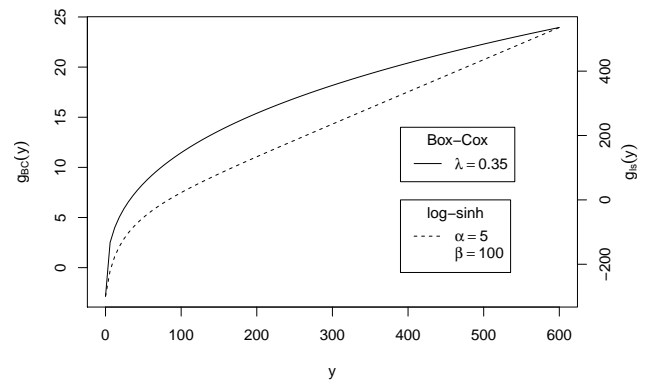

Fig. 1. Behavior of the Box–Cox (solid line) and log-sinh (dashed line) transformation as a function of the output variable (e.g., discharge in $L\,s^{-1}$) with parameters used in this study.

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 $\alpha = 5\,L\,s^{-1}$ and $\beta = 100\,L\,s^{-1}$. 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 flowmeters 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 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, and (iv) probabilistic predictions for the extrapolation period. In these 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.

2.2.1 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. 2.1.2, 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 $\boldsymbol{\psi}$ one should also lack knowledge about $\boldsymbol{\psi}^2$, but no distribution exists that is uniform on both $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^2$ (Christensen et al., 2010).

2.2.2 Bayesian inference

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

$$f(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{y}_o, \boldsymbol{x}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{y}_o \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{x})}{\iint f(\boldsymbol{\theta}', \boldsymbol{\psi}') f(\boldsymbol{y}_o \mid \boldsymbol{\theta}', \boldsymbol{\psi}', \boldsymbol{x}) \mathrm{d}\boldsymbol{\theta}' \mathrm{d}\boldsymbol{\psi}'}. \quad (17)$$

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 multidimensional integrals, such as the ones in the denominator of Eq. (17) 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. (17) is unknown. Details are given in Sect. 3.2.

2.2.3 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 $\boldsymbol{y}^{L_1} = (y^s_{t_1}, \ldots, y^s_{t_{n_1}})$, 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 $\boldsymbol{y}^{L_1}_{\mathrm{M}}$, which only contains uncertainty from hydrological model parameters; (ii) our best knowledge about the system response $g^{-1}(\tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{B}^{L_1}_{\mathrm{M}})$, which comprehends additional uncertainty from input errors and structural deficits; and (iii) observations of the system response, $g^{-1}(\tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{B}^{L_1}_{\mathrm{M}} + \boldsymbol{E}^{L_1})$, 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

$$E\left[\tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{B}^{L_1}_{\mathrm{M}} \mid \tilde{\boldsymbol{Y}}^{L_1}_o, \boldsymbol{\theta}, \boldsymbol{\psi}\right] =$$

$$\tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{\Sigma}_{B^{L_1}_M} \left(\boldsymbol{\Sigma}_{E^{L_1}} + \boldsymbol{\Sigma}_{B^{L_1}_{\mathrm{M}}}\right)^{-1} \cdot \left(\tilde{\boldsymbol{y}}^{L_1}_o - \tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}}\right) \quad (18)$$

and their covariance matrix by

$$\mathrm{Var}\left[\tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{B}^{L_1}_{\mathrm{M}} \mid \tilde{\boldsymbol{Y}}^{L_1}_o, \boldsymbol{\theta}, \boldsymbol{\psi}\right] = \boldsymbol{\Sigma}_{B^{L_1}_{\mathrm{M}}} \left(\boldsymbol{\Sigma}_{E^{L_1}} + \boldsymbol{\Sigma}_{B^{L_1}_{\mathrm{M}}}\right)^{-1} \boldsymbol{\Sigma}_{E^{L_1}}. \quad (19)$$

To obtain the posterior predictive distribution of the bias-corrected output, $\tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{B}^{L_1}_{\mathrm{M}}$, first, we have to propagate a large sample from the posterior distribution through the simulator, $\boldsymbol{y}^{L_1}_{\mathrm{M}}$, and draw realizations of $\tilde{\boldsymbol{y}}^{L_1}_{\mathrm{M}} + \boldsymbol{B}^{L_1}_{\mathrm{M}}$ by using Eqs. (18) and (19). Then, 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, and 0.975 quantiles). A similar procedure has to be applied to derive the predictive distributions of $\boldsymbol{y}_M^{L_1}$ and $\tilde{\boldsymbol{y}}_M^{L_1} + \boldsymbol{B}_M^{L_1} + \boldsymbol{E}^{L_1}$.

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 $\tilde{\boldsymbol{y}}_M^{L_1} + \boldsymbol{B}_M^{L_1}$ in

$$\boldsymbol{E}^{L_1} = \tilde{\boldsymbol{y}}_o^{L_1} - \left(\tilde{\boldsymbol{y}}_M^{L_1}(\boldsymbol{x}, \boldsymbol{\theta}) + \boldsymbol{B}^{L_1} \right). \tag{20}$$

In this equation $\tilde{\boldsymbol{y}}_o^{L_1}$ refers to the field data. The medians of the components of this sample represent our best point estimates of the observation errors that we will use to test the statistical assumptions as described in Sect. 2.2.5.

2.2.4 Predictions for the validation layout L_2

Fourth, one computes posterior predictive distributions for the extrapolation layout, L_2, where data are not available or not used for calibration. If data are available, they can be used for validation. In our study, L_2 denotes the location and the time points of the extrapolation range, and the associated model output is given by $\boldsymbol{y}^{L_2} = (y_{t_{n_1+1}}^s, \ldots, y_{t_{n_2}}^s)$. 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 Eqs. (35) and (36) of Reichert and Schuwirth (2012) instead of Eqs. (18) and (19). However, the specific form of our bias formulation as an Ornstein–Uhlenbeck process (Eqs. 5, 7) 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 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

$$\mathrm{E}\left[B_{M,j}^{L_2} \mid B_{M,j-1}^{L_2} = b_{j-1}, \boldsymbol{\theta}, \boldsymbol{\psi} \right] = b_{j-1} \cdot \exp\left(-\frac{\Delta t}{\tau} \right), \tag{21}$$

$$\mathrm{Var}\left[B_{M,j}^{L_2} \mid B_{M,j-1}^{L_2}, \boldsymbol{\theta}, \boldsymbol{\psi} \right] =$$
$$\left(\sigma_{B_{\mathrm{ct}}}^2 + \left(\kappa x_{j-d} \right)^2 \right) \cdot \left(1 - \exp\left(-2\frac{\Delta t}{\tau} \right) \right). \tag{22}$$

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 $\boldsymbol{y}_M^{L_2}$, $g^{-1}(\tilde{\boldsymbol{y}}_M^{L_2} + \boldsymbol{B}_M^{L_2})$, and $g^{-1}(\tilde{\boldsymbol{y}}_M^{L_2} + \boldsymbol{B}_M^{L_2} + \boldsymbol{E}^{L_2})$ can be constructed similarly as for layout L_1.

2.2.5 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 bandwidth" (Breinholt et al., 2012). The reliability measures what percentage of the validation data are included in the 95 % credibility intervals of $g^{-1}(\tilde{\boldsymbol{y}}_M + \boldsymbol{B}_M + \boldsymbol{E})$. 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 bandwidth (ABW) measures the average span 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 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. 4), 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. 2.2.3, we can use the median of the posterior of the observation error at layout L_1 to do such a frequentist test. One should test whether these observation errors 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 imprecise prior knowledge of parameter values. In our view, this distorts the interpretation of frequentist tests.

2.2.6 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.

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.

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 %.

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 flowmeter 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
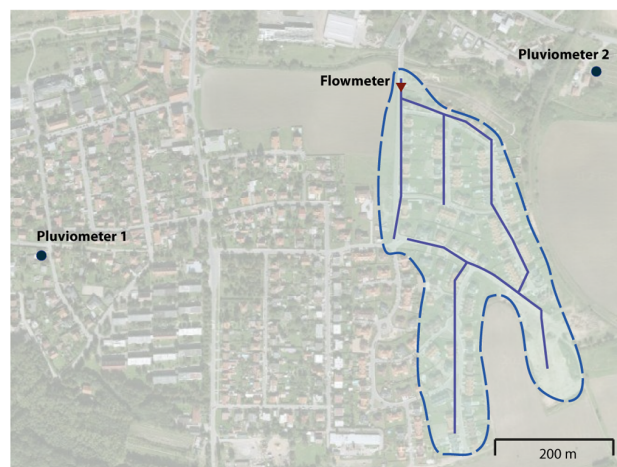

Fig. 2. Aerial photo of our Sadová case study catchment. The map shows the layout of the main stormwater conduits and the location of the rain gauges and the flowmeter.

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 data sets have been aggregated to 2 min time steps using inverse distance weighting.

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 and validation storms had a maximum rain rate ranging from 13 to 65 mm h^{-1} and from 8 to 34 mm h^{-1}, respectively. The monitored rainstorms had a duration of 0.5–4 h with a cumulative height varying from 2.3 to 33 mm. The calibration and validation data of July 2010 are illustrated in Fig. 4.

3.2 Model implementation

We modeled 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., 2012). 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 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.,
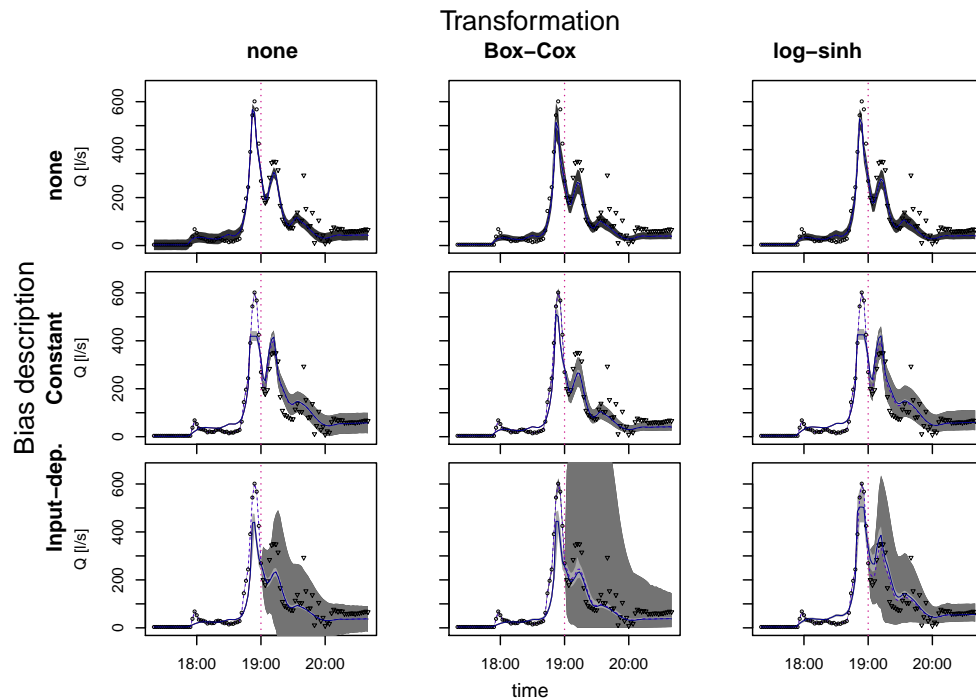
Fig. 3. The 95 % credible intervals for flow predictions for the transition phase obtained with different assumptions on error distribution. The vertical dotted line divides the calibration layout (past) from the validation layout (future). The solid line is the deterministic model output with the optimized 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 and triangles. Triangles were only used to validate the model. 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 E to total uncertainty becomes very small ($\lesssim 1\,\mathrm{L\,s^{-1}}$) and therefore is not visible at this scale. Consequently, the credibility intervals for the system output ($g^{-1}(\tilde{y}_\mathrm{M} + B_\mathrm{M})$) and the observations ($g^{-1}(\tilde{y}_\mathrm{M} + B_\mathrm{M} + E)$) are almost identical and overlap.

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 MCMC algorithm (Hastings, 1970). Before sampling from $f(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{y}_o, \boldsymbol{x})$, 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.

4 Results

In general, accounting for model bias produced substantially wider prediction uncertainty bands in the extrapolation domain 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.

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, 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 a 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 32 % of the maximum runoff modeled in the extrapolation phase.

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, 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 that 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. 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 $g^{-1}(\tilde{y}_{\mathrm{M}} + B_{\mathrm{M}})$, 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 extrapolation 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 when bias is considered, the deterministic model performs best when residual
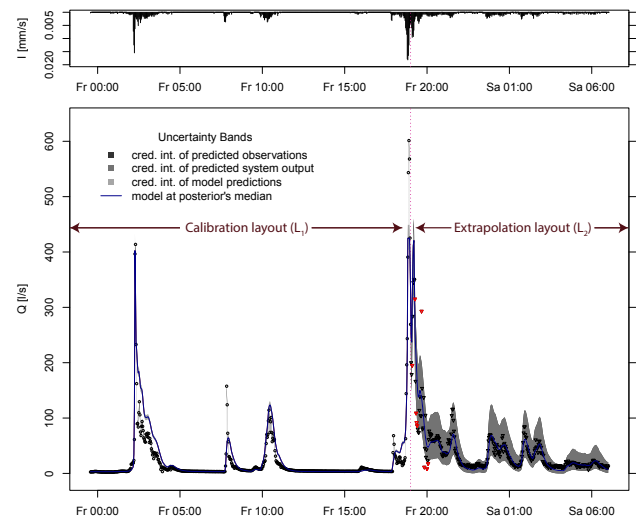

Fig. 4. Probabilistic runoff predictions for part of the calibration (left) and the validation period (right) with the constant bias model and log-sinh transformation. The input time series (hyetograph) is shown on the top. The observed hydrograph is represented by dots, with the triangular data points being used only for validation. The 95 % credible intervals are interpreted as follows: parametric uncertainty due to y_{M} (light gray), parametric plus input and structural uncertainty due to $g^{-1}(\tilde{y}_{\mathrm{M}} + B_{\mathrm{M}})$ (intermediate), total uncertainty due to $g^{-1}(\tilde{y}_{\mathrm{M}} + B_{\mathrm{M}} + E)$ (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.

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.

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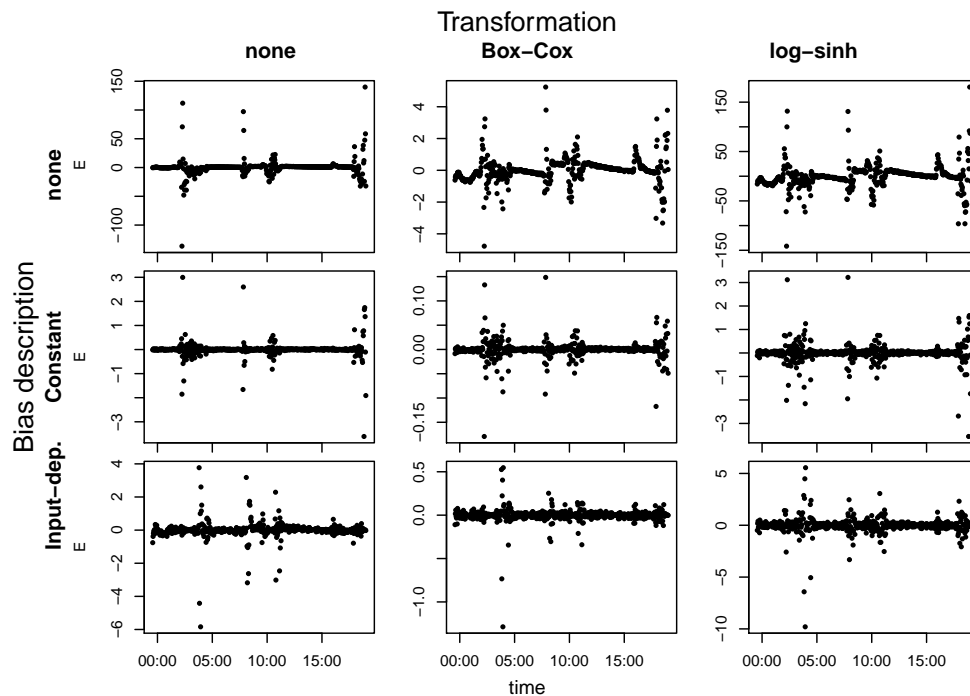
Fig. 5. Time series of median of **E**, the part of residuals assumed to come from random observation errors. The range illustrated is the same part of the calibration period as shown in Fig. 4. The ordinate axis is in transformed flow units.

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 that best fulfills the statistical assumptions (Fig. 5). The estimate of the observation errors has almost no autocorrelation and relatively low heteroskedasticity. Diagnostic plots on **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 peak discharges (see Fig. 4).
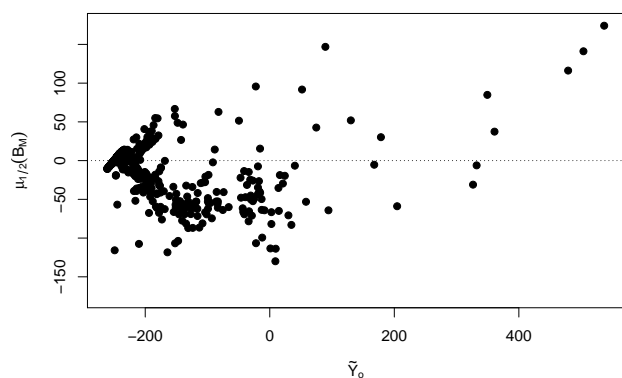
Fig. 6. Median of model inadequacy versus transformed observed runoff for the calibration period shown in Fig. 4. Results are shown for the best solution: the constant bias log-sinh transformed error model.

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

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 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 (Frey et al., 2011). Another part of prediction uncertainty, instead, stems from imprecise precipitation measurements, which increase 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).

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 that 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.

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.

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 modeler 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., 2012). In these cases, we recommend applying our five-variant bias description in order to disentangle the effects of model discrepancies 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. 2.2.1. 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 system's dynamics. One possibility could be to modify Eq. (7) and, instead of a linear increase of uncertainty with the precipitation, one could adopt a power relationship with the exponent as calibration parameter.

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 the following.

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 that 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 (e.g., Renard et al., 2010). 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 that 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 (Reichert et al., 2011).

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.

References

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

Bareš, V., Stránský, D., Kopecká, J., and Fridrich, J.: Monitoring povodí a stokové síto Města Hostivice – lokalita Sadová [Monitoring a sewer watershed in Hostivice municipality – Sadová district], Tech. rep., Czech Technical University in Prague, 2010 (in Czech).

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

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

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

Box, G. and Cox, D.: An analysis of transformations, J. Roy. Stat. Soc. B, 26, 211–252, 1964.

Breinholt, A., Møller, J., Madsen, H., and Mikkelsen, P.: A formal statistical approach to representing uncertainty in rainfall-runoff modelling with focus on residual analysis and probabilistic output evaluation-distinguishing simulation and prediction, J. Hydrol., 472–473, 36–52, doi:10.1016/j.jhydrol.2012.09.014, 2012.

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

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

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

Craig, P., Goldstein, M., Rougier, J., and Seheult, A.: Bayesian forecasting for complex systems using computer simulators, J. Am. Stat. Assoc., 96, 717–729, 2001.

Dotto, C. B. S., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D. T., and Fletcher, T. D.: Performance and sensitivity analysis of stormwater models using a Bayesian approach and long-term high resolution data, Environ. Model. Softw., 26, 1225–1239, 2011.

Dürrenmatt, D., Del Giudice, D., and Rieckermann, J.: Dynamic time warping improves sewer flow monitoring, Water Res., 47, 3803–3816, doi:10.1016/j.watres.2013.03.051, 2013.

Evin, G., Kavetski, D., Thyer, M., and Kuczera, G.: Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resour. Res., 49, 4518–4524, doi:10.1002/wrcr.20284, 2013.

Freni, G., Mannina, G., and Viviani, G.: Uncertainty assessment of an integrated urban drainage model, J. Hydrol., 373, 392–404, doi:10.1016/j.jhydrol.2009.04.037, 2009a.

Freni, G., Mannina, G., and Viviani, G.: Urban runoff modelling uncertainty: Comparison among Bayesian and pseudo-Bayesian methods, Environ. Modell. Softw., 24, 1100–1111, doi:10.1016/j.envsoft.2011.03.013, 2009b.

Frey, M. P., Stamm, C., Schneider, M. K., and Reichert, P.: Using discharge data to reduce structural deficits in a hydrological model with a Bayesian inference approach and the implications for the prediction of critical source areas, Water Resour. Res., 47, W12529, doi:10.1029/2010WR009993, 2011.

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

Haario, H., Saksman, E., and Tamminen, J.: An Adaptive Metropolis Algorithm, Bernoulli, 7, 223–242, available at: http://www.jstor.org/stable/3318737, 2001.

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

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

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D.: Combining Field Data and Computer Simulations for Calibration and Prediction, SIAM J. Sci. Comput., 26, 448–466, doi:10.1137/S1064827503426693, 2005.

Honti, M., Stamm, C., and Reichert, P.: Integrated uncertainty assessment of discharge predictions with a statistical error model, Water Resour. Res., 49, 4866–4884, doi:10.1002/wrcr.20374, 2013.

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

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

Kendall, M., Stuart, A., and Ord, J.: Advanced Theory of Statistics, Distribution Theory (Vol. 1), London, Arnold, 1994.

Kennedy, M. and O'Hagan, A.: Bayesian calibration of computer models, J. Roy. Stat. Soc. B, 63, 425–464, 2001.

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

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

Kollo, T. and von Rosen, D.: Advanced multivariate statistics with matrices, Vol. 579, Springer, 2005.

Korving, H. and Clemens, F.: Impact of dimension uncertainty and model calibration on sewer system assessment, Water Sci. Technol., 52, 35–42, 2005.

Kuczera, G.: Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty, Water Resour. Res., 19, 1151–1162, doi:10.1029/WR019i005p01151, 1983.

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

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

Montanari, A. and Di Baldassarre, G.: Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty, Adv. Water Resour., 51, 498–504, doi:10.1016/j.advwatres.2012.09.007, 2013.

Montanari, A. and Koutsoyiannis, D.: A blueprint for process-based modeling of uncertain hydrological systems, Water Resour. Res., 48, W09555, doi:10.1029/2011WR011412, 2012.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I : A discussion of principles, J. Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Neumann, M. B. and Gujer, W.: Underestimation of uncertainty in statistical regression of environmental models: influence of model structure uncertainty, Environ. Sci. Technol., 42, 4037–4043, 2008.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T.: Uncertain judgments: eliciting experts' probabilities, Wiley. com, 2006.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: http://www.R-project.org/, last access: 1 April 2013.

Reichert, P.: Conceptual and practical aspects of quantifying uncertainty in environmental modelling and decision support, 2012 International Congress on Environmental Modelling and Software, International Environmental Modelling and Software Society (iEMSs), 2012.

Reichert, P. and Borsuk, M.: Does high forecast uncertainty preclude effective decision support?, Environ. Modell. Softw., 20, 991–1001, 2005.

Reichert, P. and Mieleitner, J.: Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, Water Resour. Res., 45, W10402, doi:10.1029/2009WR007814, 2009.

Reichert, P. and Schuwirth, N.: Linking statistical bias description to multiobjective model calibration, Water Resour. Res., 48, W09543, doi:10.1029/2011WR011391, 2012.

Reichert, P., White, G., Bayarri, M. J., and Pitman, E. B.: 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, 2011.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, doi:10.1029/2009WR008328, 2010.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., and Franks, S. W.: 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, 2011.

Rossman, L. and Supply, W.: Storm water management model user's manual, version 5.0, National Risk Management Research Laboratory, Office of Research and Development, US Environmental Protection Agency, 2010.

Schellart, A., Shepherd, W., and Saul, A.: Influence of rainfall estimation error and spatial variability on sewer flow prediction at a small urban scale, Adv. Water Resour., 45, 65–75, doi:10.1016/j.advwatres.2011.10.012, 2012.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46, W10531, doi:10.1029/2007WR006720, 2010.

Sikorska, A. E., Scheidegger, A., Banasik, K., and Rieckermann, J.: Bayesian uncertainty assessment of flood predictions in ungauged urban basins for conceptual rainfall-runoff models, Hydrol. Earth Syst. Sci., 16, 1221–1236, doi:10.5194/hess-16-1221-2012, 2012.

Sikorska, A. E., Scheidegger, A., Banasik, K., and Rieckermann, J.: Considering rating curve uncertainty in water level predictions, Hydrol. Earth Syst. Sci. Discuss., 10, 2955–2986, doi:10.5194/hessd-10-2955-2013, 2013.

Uhlenbeck, G. and Ornstein, L.: On the theory of the Brownian motion, Phys. Rev., 36, 823–841, 1930.

Vihola, M.: Robust adaptive Metropolis algorithm with coerced acceptance rate, Stat. Comput., 22, 997–1008, 2012.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A.: 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, 2008.

Wang, Q., Shrestha, D., Robertson, D., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, Water Resour. Res., 48, W05514, doi:10.1029/2011WR010973, 2012.

Willems, P.: Model uncertainty analysis by variance decomposition, Phys. Chem. Earth, 42, 21–30, doi:10.1016/j.pce.2011.07.003, 2012.

Willems, P., Molnar, P., Einfalt, T., Arnbjerg-Nielsen, K., Onof, C., Nguyen, V.-T.-V., and Burlando, P.: Rainfall in the urban context: Forecasting, risk and climate change, Atmos. Res., 103, 1–3, doi:10.1016/j.atmosres.2011.11.004, 2012.

Yang, J., Reichert, P., and Abbaspour, K.: Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland), Water Resour. Res., 43, W10401, doi:10.1029/2006WR005497, 2007a.

Yang, J., Reichert, P., Abbaspour, K. C., and Yang, H.: Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference, J. Hydrol., 340, 167–182, doi:10.1016/j.jhydrol.2007.04.006, 2007b.