Accepted Version of Article: Hutton, C., Kapelan, Z., Vamvakeridou-Lyroudia, L., and Savić, D. (2014). "Application of Formal and Informal Bayesian Methods for Water Distribution Hydraulic Model Calibration." J. Water Resour. Plann. Manage., 140(11), 04014030.

The Application of Formal and Informal Bayesian Methods for Water Distribution Hydraulic Model Calibration

Hutton, C.J.[1], Kapelan, Z.[2], Vamvakeridou-Lyroudia, L.[3], and Savić, D.[4]

Abstract

Water Distribution System model parameter calibration is an important step to obtain a representative system model, such that it may be applied to understand system operational performance, often in real-time. However, few approaches have attempted to quantify uncertainty in calibrated parameters, model predictions, and consider the sensitivity of model predictions to uncertain parameters. A probabilistic Bayesian approach is here applied to calibrate - and quantify uncertainty in - the pipe roughness groups of an Epanet2 hydraulic model of a real-life water distribution network. Within the applied Bayesian framework, the relative performance of formal and informal Bayesian likelihoods in implicitly quantifying parameter and predictive uncertainty is considered. Both approaches quantify posterior parameter uncertainty with similar posterior distributions for parameter values (mean and standard deviation). However, the predictive uncertainty intervals identified with the informal likelihood are too narrow, regardless of the behavioural threshold applied to derive these bounds. In contrast, the formal Bayesian approach produces more realistic 95% prediction intervals based on their statistical coverage of the observations. This results as the error model standard deviation is jointly inferred during calibration, which also helps to avoid potential over-conditioning of the posterior parameter distribution. However, posterior diagnostic checks reveal that the prediction intervals are not valid at percentiles other than the 95% interval as the assumptions of normality,

---

[1]Associate Research Fellow, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, U.K. Email: c.j.hutton@ex.ac.uk
[2]Professor, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, U.K.
[3]Senior Research Fellow, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, U.K.
[4]Professor, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, U.K.

residual homoscedasticity, and non-correlation, often assumed in hydraulic model calibration, do no hold. More robust calibration requires the development of error models better suited to the nature of residual errors found in Water Distribution System models.

Subject Headings: Water Distribution Systems; Uncertainty Principles; Numerical Models; Calibration

Introduction

Water distribution system hydraulic models (e.g. EPANET2; Rossman 2000) are widely applied to aid Water Distribution System (WDS) analysis, planning (Kapelan et al. 2005), and to derive better system operational performance in real-time (Jamieson et al. 2007; Preis et al. 2010; Romano et al. 2012). Offline calibration prior to model application is a necessary step to derive a representative model of the system to be simulated (Savic et al. 2009). Optimisation based approaches have been widely applied for WDS model calibration, whereby model parameters (e.g. pipe roughness) are adjusted to minimise the difference between observed and predicted model states (e.g. nodal pressures and/or pipe flow rates). Methodological development has focussed primarily on developing more efficient means to identify optimal model parameters (for a review see Savic et al. 2009). Despite the fact that there are multiple sources of system uncertainty that affect the quality of model predictions, including model structural, input (e.g. demand), parameter, and measurement uncertainty (Hutton et al. 2012b) relatively few approaches have attempted to quantify model parameter uncertainty (Kang and Lansey 2011; Kapelan et al. 2007), and in turn, the uncertainty in subsequently derived predictions.

In WDS models model parameter uncertainty has been quantified, post calibration, using the First Order Second Moment (FOSM) method (Bush and Uber 1998; Lansey et al. 2001). The method makes potentially restrictive assumptions, including model linearity and normality and independence of calibration parameter values and measurement errors. The parameter response surface can differ significantly from the multi-normal distribution assumed in the first order approximation (Vrugt et al. 2003). In light of these potential difficulties, Kapelan et al. (2007) applied the SCEM-UA optimisation algorithm (Vrugt et al. 2003) within a formal Bayesian framework to calibrate a WDS model, explicitly explore the posterior parameter distributions, and in doing so, quantify uncertainty in the posterior parameter and predictive distributions. The calibration problem effectively reduced to a least squares problem, which also makes potentially restrictive assumptions, including Gaussianity of model residuals. The validity of these assumptions, and in turn the validity of the predictive distributions, requires further evaluation in the context of WDS models. Often such assumptions are not fully evaluated (Huang and McBean 2007), yet may lead to the case where the parameter response surface is over-conditioned; that is, parameters are identified that appear well constrained, but are in fact wrong because of the influence of different forms of model uncertainty (Beven et al. 2008; Hutton et al. 2012b). Calibration parameter errors may cascade and introduce uncertainty into model predictions, subsequently derived planning and/or control optimisation decisions (Sumer and Lansey 2009), and into water quality model predictions (Savic et al. 2009). Thus, it is important that the robustness of the calibration procedure to methodological assumptions is adequately considered.

In Hutton et al (2012b) a framework was presented that considers the uncertainty cascade within the context of Water Distribution Systems. With a view towards strengthening understanding of uncertainties within this framework, and the tools by which they may be robustly quantified, the aim of this paper is to compare both formal and informal Bayesian approaches for WDS model calibration, which to the authors' knowledge have not been applied comparatively in this context. Following a consideration of the calibration problem from a probabilistic, Bayesian perspective, the alternative formal and informal Bayesian likelihoods are introduced, alongside the objectives to be addressed to evaluate the methodologies in the context of WDS model calibration. The calibration case study and Bayesian likelihoods are then described, followed by a presentation of the results, discussion and conclusions.

Probabilistic Calibration - Formal and Informal Likelihood Functions

In light of uncertainty in the parameter values of real-life system models (e.g. of natural and manmade systems), model calibration is widely set within a probabilistic Bayesian framework (Beven and Freer 2001; Draper 1995; Freni et al. 2009b; Vrugt et al. 2003), where the uncertainty in calibrated parameter values is represented probabilistically, $\mathrm{P}(\theta)$. Bayes' equation provides a means to revise the probability distribution of the model parameter values, in light of new data (Y), to derive the probability of the parameters, conditional on the available data:

$$P(\theta \mid \mathrm{Y}) \propto P(\mathrm{Y} \mid \theta) P(\theta) \quad (1)$$

The second right hand term is the prior distribution of model parameters, representing the prior knowledge of the parameter value distribution before obtaining the new data. This prior is combined with the likelihood function, which is the probability of the observed data, given the model parameters. The likelihood function, alongside the parameter sampling procedure, represents a key decision in the calibration procedure.

Formal likelihood functions, such as the Gaussian distribution, have been widely chosen within the probabilistic framework, whereby a model of the residuals between observed and predicted model states is used to derive a posterior probability (Dotto et al. 2012; Freni and Mannina 2010). These are the same assumptions typically made using the Least Squares approaches more often applied in WDS calibration (Savic et al. 2009). When the parameters of the error model are jointly inferred alongside those of the model parameters, the method may implicitly account for the effect of other sources of uncertainty on model parameter and predictions estimates, by deriving a correct simulation of the total residual errors. This is in contrast to explicit formal Bayesian approaches that attempt to separate out various sources of model error (Thyer et al. 2009). The problem with the formal Bayesian approach is that the likelihood chosen can strongly condition the shape of the parameter probability distribution (Beven et al. 2008). If the model residuals do not conform to such a distribution, parameter space may become over-conditioned, leading to miss-placed confidence in parameter estimates and model prediction intervals (Beven et al. 2008).

The Generalised Likelihood Uncertainty Estimation (GLUE) procedure was developed following dissatisfaction with the formal Bayesian approach, the potential for over-conditioning the posterior parameter distribution, and the observation that many different models (and parameter sets), produced equally good model predictions (Beven and Binley 1992; Beven and Brazier 2011); a form of equifinality. Recently referred to as the informal, or pseudo Bayesian approach (Freni et al. 2009b; Vrugt et al. 2009), the method employs an informal likelihood function (Smith et al. 2008) in equation (1), the choice of which is largely subjective, and typically based on commonly applied measures of error, such as the sum of square errors. Following parameter sampling, once an informal likelihood is obtained for each parameter, a user defined threshold (t_b) is chosen to determine between the best performing parameter sets – the behavioural models – and the worst performing parameter sets – the non-behavioural models. The informal likelihoods associated with each behavioural parameter set (and associated predictions) are then normalised to unity to derive a probabilistic representation of model parameter and predictive uncertainty. The likelihood information may also be used to conduct a model parameter sensitivity analysis (Beven and Freer 2001), which can reveal important information regarding model structure and parameter dependency (Hutton et al. 2012a), and potentially guide further data collection. The choice of behavioural threshold, however, is alongside the choice of the informal likelihood, a subjective choice that cannot be evaluated a posteriori (Freni et al. 2008; Freni et al. 2009a). A related method to GLUE is the Approximate Bayesian Computation (ABC) approach, which instead of a likelihood function, applies a distance metric to summarise the match between observed and predicted time series, and a threshold to determine whether the simulation should be included to derive the approximate posterior distribution (Wilkinson, 2013).

Whilst existing studies have compared informal and formal Bayesian methods/software (Dotto et al. 2009; Vrugt et al. 2009; Hall et al. 2011), such comparisons have not been made in the context of Water Distribution Systems models. Furthermore, different methods compared in the literature, such as GLUE, may be applied with different combinations of likelihood function and parameter sampling procedure (Dotto et al. 2012; McMillan and Clark 2009; Romanowicz et al. 1994) – the two critical choices in probabilistic Bayesian calibration. Thus, rather than make comparisons at a software level, here we investigate different choices for the likelihood function in equation (1); a critical choice that underpins Bayesian calibration more generally. To address the research aim, the objectives of this paper are to:

1. Apply and compare the formal and informal Bayesian probabilistic approaches to solve a real WDS system calibration problem.
2. Use the posterior parameter distributions to evaluate model parameter sensitivity.
3. Interrogate the posterior predictive distributions to evaluate the relative performance and assumptions made in the formal and informal Bayesian approaches.

Case Study

Description

The two Bayesian calibration procedures are applied to calibrate a hydraulic, demand driven model (EPANET2; Rossman 2000) of a WDS located in the UK (Figure 1), studied previously by Kapelan et al. (2007). The WDS covers an area of approximately 6 km^2, has a ground elevation range of 54-200m above datum, and serves a population of approximately 4,500. The system is supplied by gravity from a service reservoir, and has two pressure reducing valves in the south. The EPANET2 model consists of 451 nodes, 497 pipes and two PRVs (Figure 1).

Calibration data were collected from a normal water use field test conducted in June 1994, with an estimated average demand of 14.4 litres/s. Hourly data were collected for a period of 24 hours from 28 pressure loggers, and the model therefore calibrated for 24 steady-state loading conditions.

Analysis

The network is calibrated for 10 grouped Hazen-Williams pipe roughness coefficients, grouped by pipe material/lining and diameter (Table 1). A uniform prior PDF is assumed for each parameter, which is widely applied in the absence of any prior information (Vrugt et al. 2009; Beven and Freer, 2001), and assigns equal probability across the prior range. Engineering judgment, based on pipe material, lining and diameter is used to set the prior range for each group. Monte Carlo (random) sampling was applied to generate parameter sets from the prior ranges, which given the computational efficiency of the network model, was run overnight to ensure sufficient posterior samples, resulting in a total of 2.4×10^6 samples. More efficient parameter sampling procedures have been applied with both formal and informal likelihood functions (Blasone et al. 2008; Kapelan et al. 2007; McMillan and Clark 2009), which may be better suited to larger distribution networks. It should be noted that the performance of more targeted sampling procedures can be dependent on the chosen likelihood function.

For the Formal Bayesian approach, a Gaussian Likelihood is applied within Equation (1) for model calibration. For computational ease, the log-likelihood is used (Vrugt et al. 2009):

$$L(\theta \mid Y) = -\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n}\frac{1}{2}\ln\left(\sigma_i^{2}\right) - \sum_{i=1}^{n}\frac{(p_i - o_i)^2}{2\sigma_i^{2}} \qquad (2)$$

where n is the number of observations (672), p_i and o_i are the i th model prediction and observation, respectively, and σ_i is the error standard deviation for each observation. The standard deviation is assumed the same for each observation, and is also jointly calibrated as an error model parameter, alongside the 10 pipe roughness groups, sampling from a uniform prior on the interval [0.01, 2]. A total of 20,000 parameter sets were retained for posterior analysis.

The informal likelihood function applied in the study is based on the commonly applied Nash-Sutcliffe Efficiency Statistic (Dotto et al. 2012; Smith et al. 2008):

$$L(Y \mid \theta) = \mathrm{MAX}\left[1 - \frac{\sum_{i=1}^{n}|p_i - o_i|^2}{\sum_{i=1}^{n}|o_i - \bar{o}_i|^2}, 0\right] \qquad (3)$$

Where n is the number of observations (o) and predictions (p). To derive probabilistic information from the informal likelihood, a proportion of total number of model runs needs to be retained; the associated Likelihoods are then normalised to unity to derive probabilistic information. Given that the, so called, behavioural threshold is subjectively derived, a total of 5 thresholds were used between

5000 and 50,000 of the best performing parameter sets – e.g. between the top 0.2% and 2% of simulations.

The sensitivity of model performance to each parameter was evaluated using the results from the formal likelihood, and also from the informal likelihood for each behavioural threshold, following the method applied in Hutton et al. (2012a). First order sensitivity was calculated based on aerial deviation between the prior (uniform) Cumulative Distribution Function (CDF) and posterior CDF obtained for each parameter, which for comparison across parameters, is normalised by the range of the prior for each parameter. The greater the aerial difference, the more concentrated probability mass is in certain areas of parameter space, and therefore the more sensitive the model is to the given parameter. Coefficients of determination between parameter values were also calculated for each likelihood function, and using the informal likelihood, for each behavioural threshold.

Using the Formal Bayesian approach, the 95% confidence intervals are calculated by combining the model predictions at each observation point with their associated probabilities. The 95% prediction intervals are obtained by combining the probability of each parameter set with 50 independent samples taken from the Gaussian distribution (Stedinger et al. 2008). These are then assigned to the prediction at each observation point associated with the parameter set, from which the 95% prediction intervals are derived. The Informal Bayesian uncertainty bounds are derived by assigning the probability of a parameter set to its associated prediction at each observation point. The 95% uncertainty intervals are then derived from the computed cumulative density function across the prediction range.

Results and Discussion

The identified means of the posterior parameter distributions are similar for both the Informal and Formal Bayesian approaches (Table 2), and also to those derived in Kapelan et al. (2007). The exception however is with parameter groups P2 and P3. Whilst similar values are obtained using the formal and informal Bayesian approach applied, H-W roughness coefficients are higher for these roughness groups than those identified in Kapelan et al. (2007), whose standard deviations are also narrower for all parameter groups. Such a result suggests the posterior distribution has been over-conditioned in the least squares approach applied, leading to potential over-confidence in the identified parameters values. In the informal Bayesian approach applied in this study, mean parameter values do not show much sensitivity to the choice of behavioural threshold, whilst there is an increase in the standard deviation of P1-P4.

Figure 2 presents the CDF difference between uniform prior and posterior distributions for the parameter groups to which model performance was most sensitive. In both formal and informal Bayesian calibration, model performance was most sensitive to P1, followed by P4, P2, and to a lesser extent P3 and P5. Model performance is most sensitive to P1 as it is the largest pipe group in the network. Furthermore, the pipes are well distributed relative to the observation locations for this parameter group. P2, P3 and P4, the next most influential pipe groups on model performance, as shown in Figure 1 also contain large numbers of pipes. Parameter groups P5-P10 are less well constrained, in part as a number of these groups have relatively fewer pipes. In the case of P8, which has 43 pipes, these pipes are not distributed in a way to affect the state predictions at the observation locations. In the Formal Bayesian analysis, the error model standard deviation (σ_i) has the second smallest posterior standard deviation, showing that the choice of error model standard deviation is important in calculating the likelihood of a given parameter set.

The four parameters to which the model results are most sensitive, as measured by the difference between prior and posterior distributions, also produce the strongest interactions, as measured by the coefficient of determination between parameter values for the best performing parameter sets, which is shown in Figure 3. Interactions between P3-P4 and P1-P4 are the strongest, when using both the formal and informal Bayesian likelihoods. The pipes in P4 have the largest diameter, as this group represents the main pipe delivering water to the network. Pipes in group P3 and P1 are then connected to the P4 pipes. Thus, pressure predictions at many of the observation locations therefore reflect a trade-off in the roughness values between P3-P4 and P1-P4, a form of equifinality where similar head loss predictions are produced through different combinations of roughness. Interaction between P2-P3 and P1-P3 produce R^2 values of 0.21 and 0.09, respectively, at a behavioural threshold of 5000 in the informal approach. The strength of all interactions reduces with an increase in behavioural threshold. In the formal approach the R^2 value for P2-P3 and P1-P3 is less than 0.05, whilst there is an R^2 value of 0.15 between P2 and P4. This suggests that despite the behavioural threshold not influencing the mean estimates of each parameter (Table 2), interaction between roughness values is important in achieving optimal predictions (e.g. the best performing models).

Figure 4 compares the 95% confidence and prediction intervals to the observations at the four observation locations (A-D) shown in Figure 1. The 95% uncertainty intervals derived from applying the informal Bayesian approach are narrow in comparison to the observations, and do not provide an appropriate statistical coverage of the observations, even when derived using the largest behavioural threshold. In contrast, the 95% prediction intervals derived from the formal Bayesian approach provide a better coverage of the observations, where 6.5% of the observations fall outside of these bounds, which is close to the expected 5%. Informal likelihoods have, in previous applications, been considered suitable to prevent the so-called over-conditioning of the parameter distribution – that is, to prevent inadequate treatment of the errors and therefore overconfidence in tightly constrained parameters (Beven et al. 2008). Whilst this does appear to have occurred in comparison to the other approaches in Table 2, the uncertainty intervals are too narrow. The reason is that unlike other systems, such as catchment systems where the method has been more widely applied (Brazier et al. 2000; Hutton et al. 2012a), the roughness parameters do not produce enough variability in the model response to produce uncertainty bounds that bracket the observations. Thus, other forms of error are not therefore 'mapped' adequately onto the parameter space (Blasone et al. 2008). The relatively small variability on model response as a function of roughness is often the case when measurement data have been collected during normal operating conditions. It would be preferable to calibrate the model using data obtained under hydrant opening, where the observations would be more sensitive to pipe roughness. However, such data are often not available, as in the case study presented here. The results identified here emphasise that when calibrating using data obtained under normal operating conditions, an appropriate consideration of other sources of uncertainty is required.

In contrast to the informal Bayesian approach, the formal approach produces more realistic uncertainty bounds as $\pm_i$ is jointly inferred during calibration alongside the parameter roughness groups. This approach is in contrast to the Bayesian approach applied in Kapelan et al. (2007) where the standard deviation was integrated out, reducing to a least squares problem. By retaining and jointly inferring $\pm_i$, more realistic uncertainty bounds are produced, and the potential for over-conditioning the posterior parameter distributions is reduced.

Figure 5 shows posterior diagnostic checks to evaluate the assumptions made when applying the formal Bayesian error model. Despite the relative success of the formal Bayesian approach in providing 95% uncertainty bounds with a plausible level of accuracy, an evaluation of the assumptions made when applying the formal error model reveal that the Gaussian model does not fully represent the true nature of model errors (Figure 5). Though the mean of $\pm_i$ (Table 2) is adjusted to that of the actual residual distribution, which shows only slight skew (Figure 5a), the model errors exhibit greater kurtosis and heavier tails than can be represented with a Gaussian distribution (Figure 5a and 5c). So, despite the 95% uncertainty intervals providing what appears to be an appropriate statistical coverage, this is not the same for other percentiles (Figure 5c). Furthermore, there is some heteroscedasticity, as the largest residual errors occur at the higher pressure measurements (Figure 5b).

The residual errors also reveal temporal autocorrelation at each observation point (Figure 4), most notably in Figure 4D. Thus, although the Gaussian assumption does not fully hold, the bounds help identify what appears to be a systematic error at observation location D, which contributes 3.5% to the total of 6.5% of the observations that fall outside of the prediction bounds. In general, the temporal correlation in residuals, and the width of the uncertainty bounds at a given observation point reflects the trade-off in calibrating the model to a number of observation locations, which are notable during the night time when demand (and therefore demand uncertainty) is low. These errors may also reflect misspecification of network topology and node elevation. The residuals also show spatially auto-correlation, as shown by the variograms produced in Figure 5d for specific times of the day. At smaller spatial lags (notably less than 500m), the variance in the residuals is smaller, suggesting nearby residuals result from the same error. Residual error variance for a given spatial lag is largest at peak demand hours in the morning (hour 8) and evening (hour 19), which is perhaps where errors in specified system demand are largest.

Summary and Conclusions

Within a probabilistic framework, Formal and Informal Bayesian methods were applied to a Water Distribution System hydraulic model calibration problem. Both methods identify similar posterior parameter distributions for the pipe roughness groups, identifying similar calibrated values (posterior PDF means) and also similar distributions. In comparison to the results derived using a least squares approach within a Bayesian framework (Kapelan et al. 2007), both Formal and Informal Bayesian methods applied here avoid the over-conditioning of the posterior parameter distribution.

The Informal Bayesian approach, however, produced uncertainty bounds that did not adequately bracket the observations. The Formal Bayesian approach produced more realistic 95% uncertainty bounds based on their statistical coverage of the observations. The approach therefore appears more appropriate for pipe roughness calibration of WDS models, as error model parameters are jointly inferred during calibration. An additional benefit of such an approach is that the quantification of uncertainty in future predictions is more robust; information that may then be used in planning decisions (Sumer and Lansey 2009), and also propagated into the application of water quality models (Fisher et al. 2011).

The assumption of Gaussian residuals, however, as applied here in the Formal Bayesian approach, and implicitly assumed in WDS calibration problems based on least squares methods (Kapelan et al. 2007; Savic et al. 2009) was revealed by posterior diagnostics to not fully represent the true nature of model residual errors in the Water Distribution System model. Thus the paper has demonstrated the need for, and methods by which such assumptions should be evaluated in further WDS model calibration. Further work is required to investigate the appropriate of other forms of error model that attempt to deal implicitly with residual errors (Schoups and Vrugt 2010), that in the context of WDS models, for example, may originate from miss-specification of Demand. Furthermore, Formal and Informal Bayesian approaches have been developed in an attempt to deal explicitly with different sources of error, by using multipliers of system input variables (McMillan et al. 2011), and also the limits of acceptability approach (Liu et al. 2009). Investigation is required to evaluate their adaptability to the joint inference problem of pipe roughness and demand (Kang and Lansey 2011), as potential problems can arise when attempting to deal with multiple error sources explicitly (Thyer et al. 2009). As set out in Hutton et al. (2012b) development of appropriate models to deal with errors and uncertainty (alongside development of models in general) should be an iterative process, where assumptions made during model calibration are checked as a means to improve on both the methods for dealing with model errors, as well as the structures of the models themselves (Gelman and Shalazi, 2012). This paper has demonstrated the need for such an approach in the context of WDS model calibration.

Acknowledgments

The work presented in this paper was partially supported by 'PREPARED, Enabling Change', an ongoing European Commission Seventh Framework funded large scale integrating interdisciplinary project (Grant agreement no.: 244232, 2010-2014).

References

Beven, K., and Binley, A. (1992). "The Future of Distributed Models - Model Calibration and Uncertainty Prediction." *Hydrological Processes*, 6(3), 279-298.

Beven, K., and Freer, J. (2001). "Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology." *Journal of Hydrology*, 249(1-4), 11-29.

Beven, K. J., and Brazier, R. E. (2011). "Dealing with Uncertainty in Erosion Model Predictions." Handbook of Erosion Modelling, R. P. C. Morgan and M. A. Nearing, eds., Blackwell, 52-79.

Beven, K. J., Smith, P. J., and Freer, J. E. (2008). "So just why would a modeller choose to be incoherent?" *Journal of Hydrology*, 354(1-4), 15-32.

Blasone, R. S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., and Zyvoloski, G. A. (2008). "Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling." *Advances in Water Resources*, 31(4), 630-648.

Brazier, R. E., Beven, K. J., Freer, J., and Rowan, J. S. (2000). "Equifinality and uncertainty in physically based soil erosion models: Application of the glue methodology to WEPP-the water erosion prediction project-for sites in the UK and USA." *Earth Surface Processes and Landforms*, 25(8), 825-845.

Bush, C. A., and Uber, J. G. (1998). "Sampling design methods for water distribution model calibration." *Journal of Water Resources Planning and Management-Asce*, 124(6), 334-344.

Dotto, C. B. S., Deletic, A., and Fletcher, T. D. (2009). "Analysis of parameter uncertainty of a flow and quality stormwater model." *Water Science and Technology*, 60(3), 717-725.

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

Draper, D. (1995). "Assessment and Propagation of Model Uncertainty." *Journal of the Royal Statistical Society Series B-Methodological*, 57(1), 45-97.

Fisher, I., Kastl, G., and Sathasivan, A. (2011). "Evaluation of suitable chlorine bulk-decay models for water distribution systems." *Water Research*, 45(16), 4896-4908.

Freni, G., and Mannina, G. (2010). "Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution." *Journal of Hydrology*, 392(1-2), 31-39.

Freni, G., Mannina, G., and Viviani, G. (2008). "Uncertainty in urban stormwater quality modelling: The effect of acceptability threshold in the GLUE methodology." *Water Research*, 42(8-9), 2061-2072.

Freni, G., Mannina, G., and Viviani, G. (2009a). "Uncertainty in urban stormwater quality modelling: The influence of likelihood measure formulation in the GLUE methodology." *Science of the Total Environment*, 408(1), 138-145.

Freni, G., Mannina, G., and Viviani, G. (2009b). "Urban runoff modelling uncertainty: Comparison among Bayesian and pseudo-Bayesian methods." *Environmental Modelling & Software*, 24(9), 1100-1111.

Gelman, A., and Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics (with discussion). British Journal of Mathematical and Statistical Psychology (forthcoming): 1-31.

Hall, J. W., L. J. Manning, and R. K. S. Hankin (2011), Bayesian calibration of a flood inundation model using spatial data, Water Resources. Research, 47, W05529, doi:10.1029/2009WR008541.

Huang, J. J., and McBean, E. A. (2007). "Using Bayesian statistics to estimate the coefficients of a two-component second-order chlorine bulk decay model for a water distribution system." *Water Research*, 41(2), 287-294.

Hutton, C. J., Brazier, R. E., Nicholas, A. P., and Nearing, M. (2012a). "On the effects of improved cross-section representation in one-dimensional flow routing models applied to ephemeral rivers." *Water Resources Research*, 48.

Hutton, C. J., Kapelan, Z., Vamakeridou-Lyroudia, L. S., and Savic, D. (2012b). "Dealing With Uncertainty in Water Distribution Systems' Models: A Framework for Real-time Modelling and Data Assimilation." *Journal of Water Resources Planning and Management-Asce*. 10.1061/(ASCE)WR.1943-5452.0000325

Jamieson, D. G., Shamir, U., Martinez, F., and Franchini, M. (2007). "Conceptual design of a generic, real-time, near-optimal control system for water-distribution networks." *Journal of Hydroinformatics*, 9(1), 3-14.

Kang, D. S., and Lansey, K. (2011). "Demand and Roughness Estimation in Water Distribution Systems." *Journal of Water Resources Planning and Management-Asce*, 137(1), 20-30.

Kapelan, Z. S., Savic, D. A., and Walters, G. A. (2005). "Multiobjective design of water distribution systems under uncertainty." *Water Resources Research*, 41(11).

Kapelan, Z. S., Savic, D. A., and Walters, G. A. (2007). "Calibration of water distribution hydraulic models using a Bayesian-Type procedure." *Journal of Hydraulic Engineering-Asce*, 133(8), 927-936.

Lansey, K. E., El-Shorbagy, W., Ahmed, I., Araujo, J., and Haan, C. T. (2001). "Calibration assessment and data collection for water distribution networks." *Journal of Hydraulic Engineering-Asce*, 127(4), 270-279.

Liu, Y. L., Freer, J., Beven, K., and Matgen, P. (2009). "Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error." *Journal of Hydrology*, 367(1-2), 93-103.

McMillan, H., and Clark, M. (2009). "Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme." *Water Resources Research*, 45.

McMillan, H., Jackson, B., Clark, M., Kavetski, D., and Woods, R. (2011). "Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models." *Journal of Hydrology*, 400(1-2), 83-94.

Preis, A., Whittle, A. J., Ostfeld, A., and Perelman, L. (2010). "On-line hydraulic state estimation in urban water networks using reduced models." *Integrating Water Systems*, 319-324.

Romano, M., Kapelan, Z., and Savic, A. D. (2012). "Automated Detection of Pipe Bursts and other Events in Water Distribution Systems." *ASCE Journal of Water Resources Planning and Management*, 10.1061/(ASCE)WR.1943-5452.0000339

Romanowicz, R., Beven, K. J., and Tawn, J. (1994). "Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach." Statistics for the Environment II: Water Related Issues, V. Barnett and K. F. Turkman, eds., John Wiley and Sons, Chichester, England, 297-317.

Rossman, L. A. (2000). *EPANET2 Users Manual* National Risk Management Research Laboratory, U.S. Environmental Protection Agency, http://www.image.unipd.it/salandin/IngAmbientale/Progetto_2/EPANET/EN2manual.pdf, Cincinnati, OH 45268, USA.

Savic, D. A., Kapelan, Z. S., and Jonkergouw, P. M. R. (2009). "Quo vadis water distribution model calibration?" *Urban Water Journal*, 6(1), 3-22.

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

Smith, P., Beven, K. J., and Tawn, J. A. (2008). "Informal likelihood measures in model assessment: Theoretic development and investigation." *Advances in Water Resources*, 31(8), 1087-1100.

Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, Water Resour. Res., 44, W00B06, doi:10.1029/2008WR006822.

Sumer, D., and Lansey, K. (2009). "Effect of Uncertainty on Water Distribution System Model Design Decisions." *Journal of Water Resources Planning and Management-Asce*, 135(1), 38-47.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and 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, -.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S. (2003). "A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters." *Water Resources Research*, 39(8), -.

Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V., and Robinson, B. A. (2009). "Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?" *Stochastic Environmental Research and Risk Assessment*, 23(7), 1011-1026.

Wilkinson, R.D. (2013). Approximate Bayesian Computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2): 129-141

Table 1. Parameter information for each pipe group; PG = parameter group; D = pipe Diameter; N = number of pipes.

PG	Material/Lining	D (mm)	N	Min	Max
P1	Cast Iron/None	76	50	20	100
P2	Cast Iron/None	102	34	20	100
P3	Cast Iron/None	152	45	20	100
P4	Cast Iron/None	254	37	20	100
P5	Ductile Iron/Cement	100	22	80	130
P6	Ductile Iron/Cement	150	15	80	130
P7	Ductile Iron/Cement	250	1	80	130
P8	Cast Iron/Epoxy	76	43	90	130
P9	MDPE/None	145	7	110	150
P10	PVC/None	152	2	110	150

Table 2. Calibration results comparing the mean and standard deviation (in brackets) of each parameter PDF for the informal likelihood method (for different behavioural thresholds), the formal likelihood method and a least squares approach. PG = parameter group.

PG	Informal Likelihood 5000	10000	20000	35000	50000	Formal Likelihood	Kapelan et al. (2007)
P1	24(1.9)	24(2.4)	25(3.0)	26(3.9)	27(4.8)	24(1.6)	25(0.8)
P2	69(8.5)	71(10)	72(11.7)	73(12.9)	74(13.4)	69(7.2)	48(1.4)
P3	56(20.8)	56(21.2)	57(21.6)	58(22)	59(22.1)	52(18.9)	42(4.5)
P4	66(6.6)	66(7.5)	65(8.8)	65(9.71)	65(10.2)	66.7(5.3)	66(1.8)
P5	107(14)	106(14)	106(14.4)	105(14.4)	105(14.4)	110(13.5)	113(8.9)
P6	105(14.4)	105(14)	105(14.4)	105(14.5)	105(14.5)	105(14.3)	100(12.9)
P7	104(14.4)	104(14)	104(14.4)	105(14.4)	105(14.4)	104.(14.7)	104(13.5)
P8	109(11.6)	109(11)	110(11.5)	110(11.6)	110(11.5)	107(11.6)	112(10.4)
P9	130(11.6)	130(11)	130(11.6)	130(11.6)	130(11.6)	129(11.6)	130(9.8)
P10	130(11.5)	130(11)	130(11.5)	130(11.5)	130(11.5)	130(10.9)	130(10.9)
σ_i	-	-	-	-	-	1.29(0.04)	-

Table 3. First order parameter sensitivity calculated as the aerial difference between the prior and posterior distribution. PG = parameter group

	Informal Likelihood					Formal Likelihood
PG	5000	10000	20000	35000	50000	
P1	0.51	0.54	0.55	0.55	0.54	0.62
P2	0.16	0.15	0.16	0.17	0.18	0.22
P3	0.07	0.05	0.04	0.03	0.02	0.12
P4	0.19	0.17	0.15	0.14	0.14	0.23
P5	0.05	0.02	0.01	0.00	0.00	0.10
P11	-	-	-	-	-	0.36

Table 4. Coefficient of determination (R^2) calculated between the best performing parameter sets in the formal and informal Bayesian calibration.

Parameter Interaction	Informal Likelihood 5000	10000	20000	35000	50000	Formal Likelihood
P3-P4	0.47	0.43	0.38	0.36	0.33	0.27
P1-P4	0.28	0.26	0.22	0.21	0.20	0.22
P2-P3	0.21	0.17	0.10	0.04	0.03	0.02
P1-P3	0.09	0.07	0.04	0.03	0.02	-
P2-P4	-	-	-	-	-	0.15
P1-P2	-	-	-	-	-	0.03

Figure 1. U.K. water distribution network used in this study showing the main calibraton pipe groups, and also the location of sensors (S) within the network. Note: pipe thickness is to help differentiate between groups, and is not related to pipe diameter. The letters A to D indicate the observation location pressure measurements shown in Fig.4

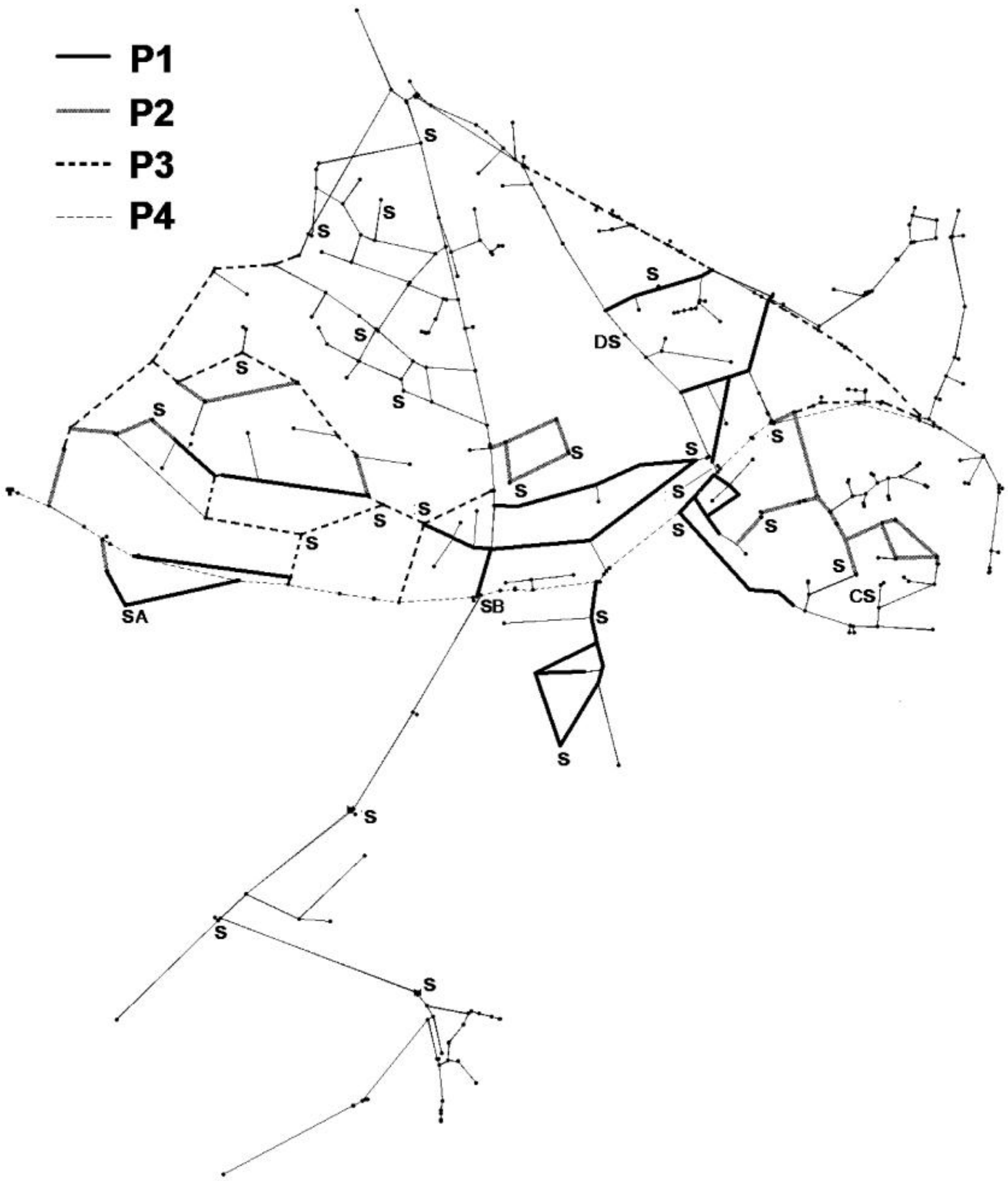
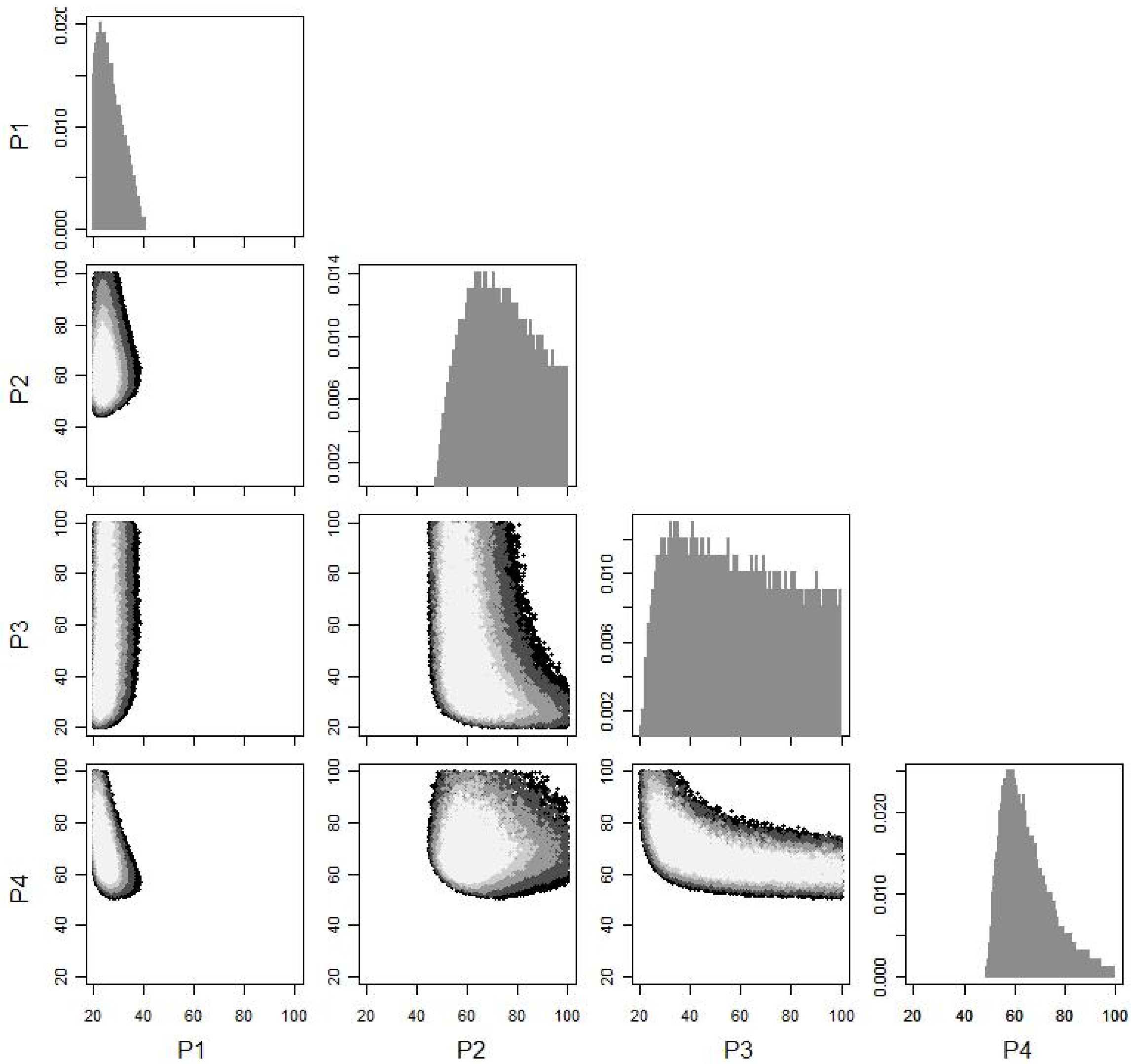


Figure 2. Posterior parameter distributions (diagonal plots) for P1-P4, the most influential pipe group parameters from informal Bayesian calibration (x-axis represents the prior roughness range; y-axis the posterior probability). The off-diagonal plots reveal parameter interaction across the prior ranges of

566 each parameter. The behavioural thresholds for each parameter set are shown from white (threshold =
567 5000) to black (threshold = 35000).



568

569

570

571

572

573

574

575

576

Figure 3. Posterior parameter distributions (diagonal plots) for P1-P4, the most influential pipe group parameters from formal Bayesian calibration (x-axis represents the prior roughness range; y-axis the posterior probability). The off-diagonal plots reveal parameter interaction across the prior ranges of each parameter. The behavioural thresholds for each parameter set are shown from white (threshold = 5000) to black (threshold = 35000).

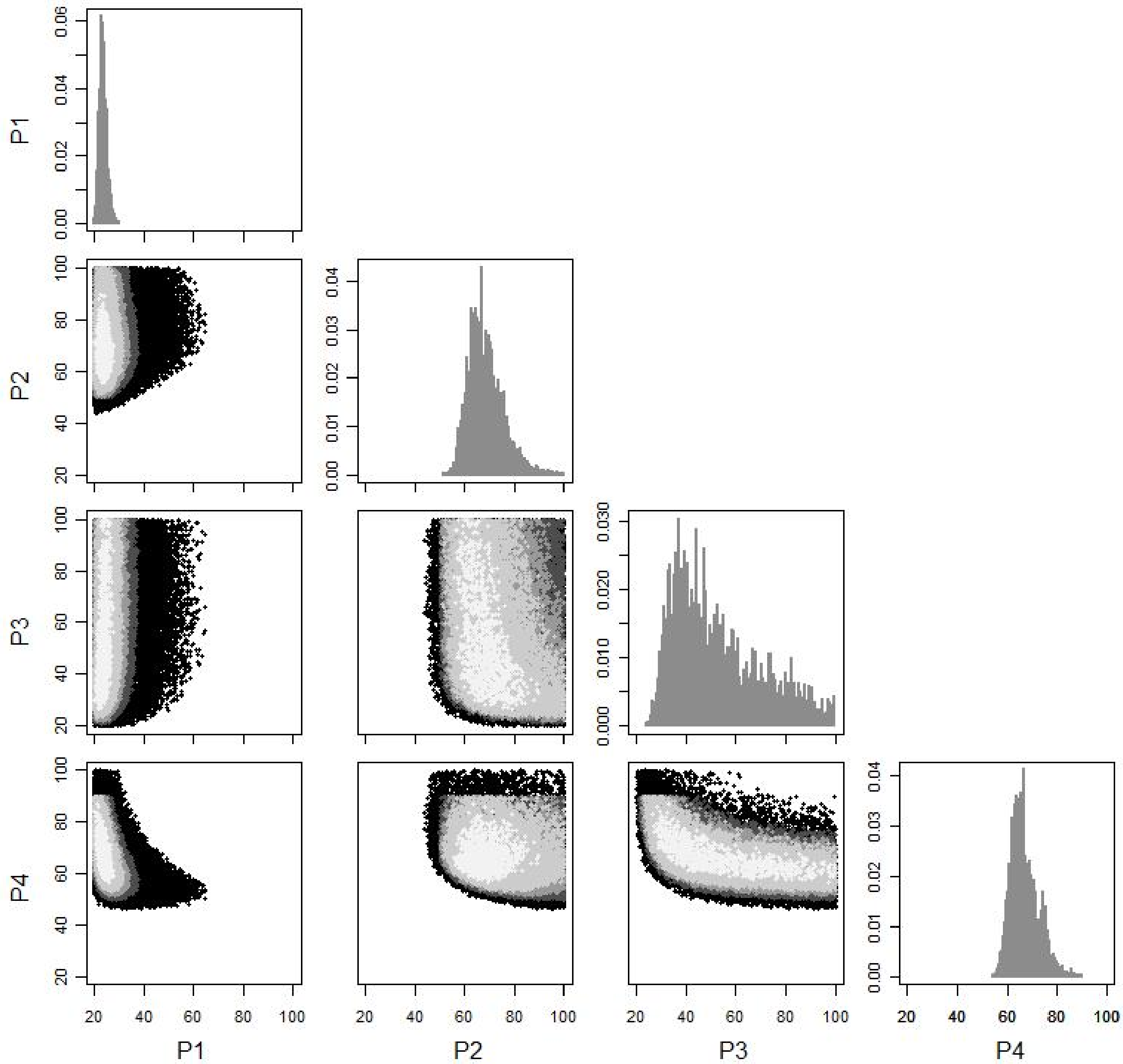
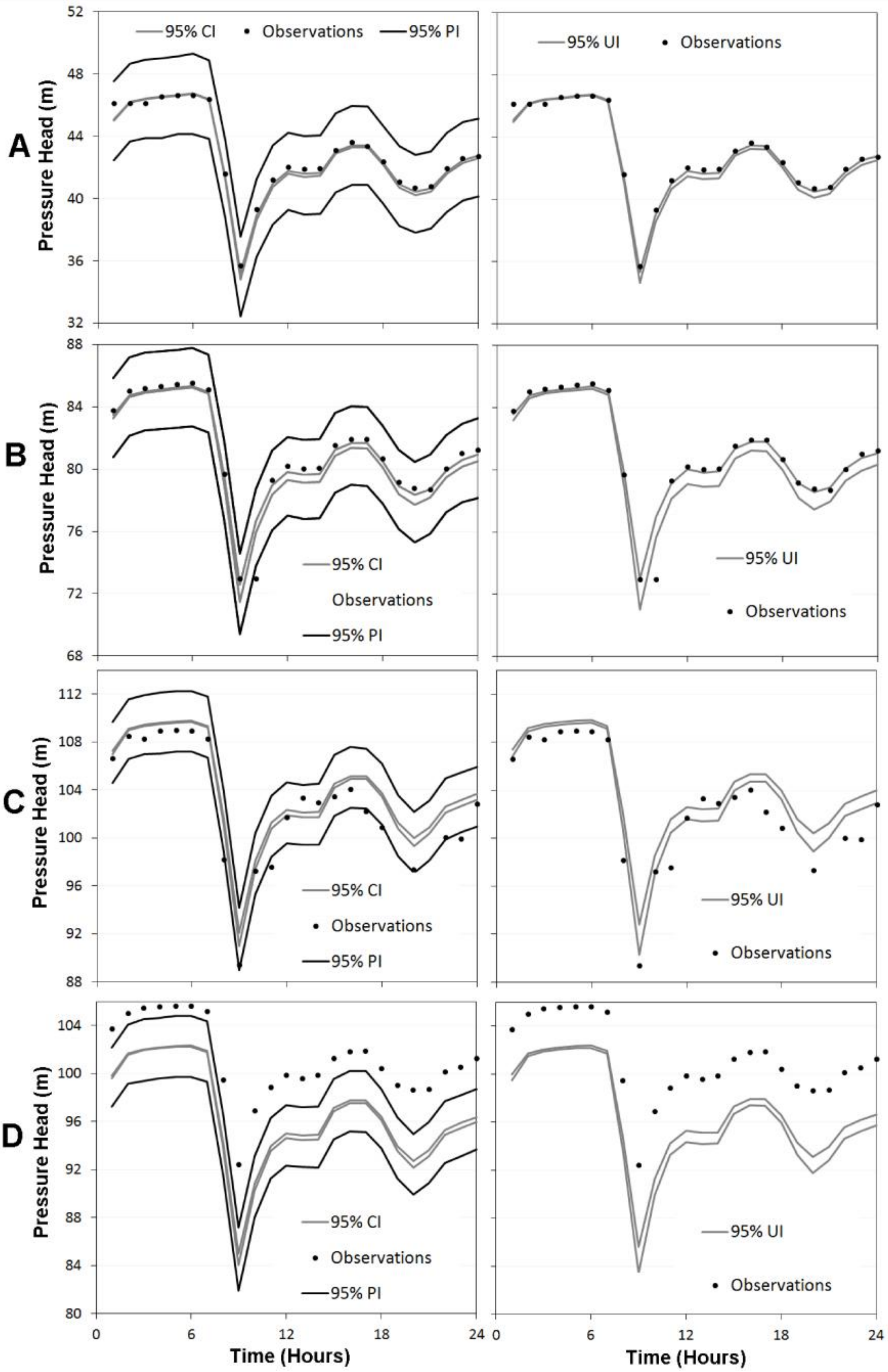


Figure 4. Formal Bayesian confidence intervals (CI) and prediction intervals (PI; left-hand figures), and informal Bayesian uncertainty intervals derived when using a behavioural threshold of 50000

592 samples (right-hand figures) for four selected sensors (A-D) as shown in Figure 1.



593

Figure 5. (a) Posterior checks of residual error model assumptions: comparison of theoretical (fitted) distribution (smooth curve) and actual residual distribution (bullets); (b) plot of residual errors as a function of pressure head; (c) quantile-quantile plot comparing the fitted distribution to the observed distribution; (d) spatial variograms of residual error variance plotted for four representative hours during the simulation.