Interactive comment on "Informal uncertainty analysis (GLUE) of continuous flow simulation in a hybrid sewer system with infiltration inflow – consistency of containment ratios in calibration and validation?" by A. Breinholt et al.

Response to K.J. Beven:

Thank you very much for your review, comments and suggestions.

1. **I think the presentation could be usefully revised to make more of the evident epistemic uncertainties in the modelling process for this study**

We agree that epistemic uncertainties play a significant role in the rather poor consistency between calibration and validation periods. We also provide several examples of this, e.g. in Figure 10 (left): a small rain event was recorded (<5mm recorded at each rain gauge) but unexpected large flow rates were observed (probably due to flow gauge error or rainfall not caught by the rain gauges). In Figure 10 (right), a presumed malfunctioning of one rain gauge (P316) causes the flow predictions to be underestimated (consequently the flow observations are consistently close to, or above the upper prediction limit for all the illustrated rain events). In Figure 11 (left) a presumed change in the mean dry weather flow level in the validation year 2009 (or a possible flow meter bias) meant that the observed flow observations in dry weather were very low and consistently close to the predicted lower bound. Figure 11 (right) shows a large validation event that is actually more extreme (in terms of accumulated mm) than any of the rainfall events from the calibration period. It appears that the upper prediction limit underestimate the peak of the hydrograph, the predicted timing of the peak is biased and the measured hydrograph has a slower recession at the tail of the hydrograph than predicted. Why is this? Well the model was not calibrated for such a large event however it may also be that the flow meter is wrong, or that the rainfall was heterogeneously distributed. We cannot know for sure. Hence we suggest including a Section 4.7 ("Epistemic uncertainties") in which this can be more thoroughly discussed. Section 4.7 would then end up concluding that the evidence from the changing nature of the errors in this study between and within periods (epistemic uncertainties) suggests that it might be very difficult to test GLUE's ability to provide uncertainty bounds that bracket observations in both calibration and validation, which would also be the case if a formal likelihood approach had been applied. Instead we should try to learn from the significant discrepancies between model and observations.

2. **Interestingly there is rather consistent coverage of the observations between different periods for different behavioural thresholds –does this suggest that some form of non-stationary error correction might be worth investigating?**

We agree when considering Figure 6, it seems that overall there is rather consistent coverage of observations for different behavioural thresholds, however this is not the case when focusing on dry and wet weather periods separately which is due to many epistemic events as discussed above. We cannot really see how a non-stationary error correction can fix this?

3. **Detailed comments:**

a) **P5: Add paragraph split**

We will add before “The scope of this paper..”

b) **P6,1950s**

Thank you. This will be Corrected.

c) **P10, L5 is this residual error variance assuming a zero mean bias (as normally used in NSE, but worth saying explicitly)?**

Yes, we are assuming a zero mean bias, and we will make that clear in the text.

d) **Equations 1 are not themselves likelihoods – should use proportionality signs not equals signs**

Equals signs will be replaced by proportionality signs.

e) **Worth noting that Equation 2 is effectively a Bayesian updating of likelihoods – but again should be proportionality not equals**

Yes we will correct it and include a few lines about the updating.

f) **P11 This could be necessary if the dotty plots show high likelihood values at the lower or upper end of any of the prior parameter ranges- but this is quite common, and ranges might be limited by physical considerations not just a fall in likelihood???**

We agree and will include a remark about this.

g) **We instead took a statistical approach to the acceptability criterion requiring a given prediction interval to bracket the proportion of the observations consistent with the chosen interval- should this really be described as statistical when your weights are not based on a formal statistical model and you are leaving error series implicit?**

You are right. This is confusing. We have deleted “statistical approach” and propose to rephrase the sentence: “We retain a sufficient number of parameter sets to bracket a desired proportion of observations”.

h) **- Should also mention here that this “third” option has been used in the past – e.g. by Xiong and Connor paper cited. Might be better to say that there have been three methods used in the past (actually 4 because of the pre-defined limits of acceptability approach suggested in Beven manifesto paper for which there have now been a number of applications, and some previous applications can be interpreted in this way)**

We think all four previously applied methods should be stated.

i) **P17 attributed to the inability of the GLUE methodology to fully describe the uncertainty of the system.**
- but one of the advantages of the GLUE method (relative to formal statistical methods) is to detect failures due to either model or data limitations. It is quite clear from some of the plots that – for whatever reason – the model cannot predict the observations but in a way that could also not really be reproduced by a stationary statistical error model (even allowing for heteroscedasticity). So this should not really be described an inability but it is rather informative about something that needs improving in the modelling process. You do, after all, discuss these sources of additional error later in the paper.

We suggest rephrasing to: "Does the observed inconsistency suggest an inability of the GLUE methodology to fully describe the uncertainty of the system?" and later in Section 4.7 we will return to this question and discuss it more thoroughly.

We discuss/conclude later in the paper that the inconsistencies is due to epistemic errors rather than an inability of the GLUE methodology to fully describe the uncertainty.

j) **P20 It is important to recognize that the GLUE methodology as applied here and in many other GLUE studies implies a transfer of all uncertainties to the model parameters**
- No!!!! You have stated in the introduction that you wanted to test the implicit error handling in GLUE, but you are now forgetting that each parameter set carries along with it an implicit(non-stationary) error series. Models that underpredict in calibration are expected to underpredict in similar circumstances in prediction etc. The uncertainties are NOT being transferred to the model parameters, but the error series are (implicitly) weighted along with the simulated outputs from that model. This does not, of course, ensure that the errors reproduced in this way will be similar in the predicted period – especially when the sources of error are epistemic as discussed here, but please do not perpetuate this misinterpretation (also in Conclusions P22 L22).

You do have an important point here and we will rephrase the sentence so that it becomes clearer that each parameter set carries along with it an implicit (non-stationary) error series and therefore uncertainty is not directly transferred to the parameters. Also we will rephrase in the conclusion.

k) **This means e.g. that insufficient rain input will be compensated for by adjusting the size of the paved area, which adds a level of variation in addition to that caused by parameter correlation (see Table 7), and the posterior parameter ranges therefore lack physical interpretation and thus cannot be used for e.g. inference about the relative size of infiltration area versus size of paved area, which otherwise would be desired knowledge.**

- but this has nothing to do with GLUE implying a transfer of uncertainties to model parameters. Such compensations will be apparent in any calibration exercise (they may even be worse in a formal statistical calibration because of the stretching of the likelihood surface) UNLESS you build in prior knowledge about what is acceptable or not acceptable for parameters and their interactions.

We will argue instead that the changing nature of errors across different periods is the cause for the wide posterior parameter ranges.

l) **P21 where the contributing runoff area and pipe network data can be estimated independently**
- but you have already noted that this may not be possible because of lack of knowledge about what is actually connected to the network – another source of epistemic errors in addition to the nonstationarity in the input errors (and perhaps observed flows)

Yes we agree that this suggestion may not provide better agreement between modelled outputs and observations and therefore the sentence will be deleted. Instead we would like to add a comment to the paragraph just above:

"The experiences from this investigation have shown that calibration of much more complex models (physically distributed, hydrodynamic) used in practical urban drainage engineeringin catchments with insufficient rain gauge coverage to questionable flow measurements from shorter measuring campaigns is problematic."

And we should add: "not least because a calibrated model normally implies a reduction in the safety factor used in modelling of urban drainage systems"

m) **P23. however we call for further comparisons between formal and informal approaches in which both calibration and validation periods are included for performance comparison in real-world applications, and we suggest that users of formal approaches demonstrate that their error assumptions are valid.**
- surely you cannot really call for this when you have not made the effort to do so yourselves.

We agree and have deleted the sentence.

n) **- Is it not better to say simply here that the evidence from the changing nature of the errors in this study between and within periods suggests that it might be very difficult to find a valid error model for use in a formal likelihood approach, and that we should therefore try to learn from the significant discrepancies between model and observations.**

We agree and will insert your suggestion in the conclusion.