We are grateful for the constructive comments of the two reviewers.

`Reviever comments are typeset in bold, our replies from the author comment in plain, and our report about the changes in italic.`

REVIEWER #1

R1: The authors conclude that the estimated uncertainty depends strongly on the uncertainty assessment method. This is quite an expected conclusion. We are talking about "an estimate" of uncertainty, therefore the estimation method does influence.

Yes, we agree that in case of estimation procedures the method will influence the outcome, and that the triviality of such a finding should be mentioned along with some relevant references. However, there is a qualitative difference between formal and informal statistical uncertainty assessment methods with regard to this statement. Several studies have shown that many popular informal likelihood functions in GLUE yield statistically inconsistent uncertainty intervals for hydrological time-series (Mantovan & Todini, 2006; Stedinger et al. 2008) and thus the free choice among the diverse informal likelihoods guarantees that the outcome may be statistically wrong in similarly diverse ways. In contrast to this freedom of choice in GLUE, formal methods – including most Bayesian approaches – are based on the principle that statistically valid likelihood functions should be properly representing the (potentially complex) statistical properties of residuals and hence there exists an "absolute" optimal likelihood formulation for each application that achieves the above mentioned goal with minimum statistical complexity. Several attempts have been made to find such formal likelihood functions for hydrological forecasting (Kuczera et al. 2006; Renard et al. 2011; Schoups and Vrugt 2010; Honti et al. 2013), but so far none of them was able to completely fulfill all requirements. This is in line with Beven's critique on formal methods that full statistical coherence may be actually impossible to reach and therefore the drawbacks of an informal approach are much less important in practice (Beven et al. 2007).

Here we applied a formal likelihood function for the time-series approach and a formal yet approximate likelihood for the quantile approaches. The surprising part of our (otherwise trivial) finding was that one can still get significantly different uncertainty from two formal approaches that each promise to capture the "true" uncertainty. The reasons behind this can be two-fold: First of all it is possible that – similarly to informal methods – the applied formal methods are still statistically

inconsistent estimators and that's why they delivered so different estimates for the very same total predictive uncertainty. Besides this, a second explanation could be that in the study we actually dealt with two different uncertainties. Based on the demonstrated interaction between stochastic model bias and flow quantiles we believe that time-series and quantile uncertainties represent fundamentally different uncertainties – regardless of the assessment methodology – and therefore it is application-specific which can and should be used.

We will add a short paragraph to the discussion mentioning that the difference between the results of different uncertainty assessments is not unprecedented, but still surprising for formal statistical methods.

Added a new subsection (now 4.1) to the discussion about the relativeness of uncertainty.

R1: My main concern however, is that the manuscript is too long. I do not think that it needs 53 pages of HESSD to illustrate the methods/results and convey the message. In the present form, it is difficult to get the essence of the paper quickly and in some places it may confuse the readers. Therefore, I recommend to significantly reduce the manuscript length. Some parts may be presented as supplemental materials if the authors find it necessary. I have a number of specific comments and suggestions for reducing the manuscript length

The length of the manuscript will be significantly reduced according to the specific suggestions below.

The length of the manuscript was reduced from 53 to 44 pages by shortening text, merging figures and moving several tables to Supplementary Material.

R1: P503, L22 - P504, L3: No need to explain this; should be deleted.

OK

Done

R1: P505, L29: full name for i.i.d. should be given.

OK

Done

R1: P506, L12: 'complicated statistical properties' e.g.?

We mean heteroscedasticity, autocorrelation and non-normality. These will be mentioned.

These properties are now explicitly named in the sentence.

R1: P509, L20: In Fig. 2 caption, you have 'logSPM'. In Fig. 3, you have 'CRRM'. But the section heading 2.3 is 'Hydrological model'. This is somewhat confusing.

LogSPM is the name of the conceptual rainfall-runoff model (CRRM) developed by Kuczera et al (2006). The caption of Fig. 2 will be rephrased to 'Schematic structure of the applied hydrological model'.

Done

R1: P513, L6-9: If 1, 2, . . . here refer to the stages (Fig. 3), say Stage 1, Stage 2, etc here.

OK

Added a leading 'Stage' to each number

R1: P513, L21-24: Not clear, what do you want to suggest here?

We wanted to say that 30 years of daily data supplies sufficient samples for the Q5 and Q95 flow quantiles from a single realization of generated weather time-series. The original sentence will be rephrased.

Done

R1: P514, L21-22: Why do you merge this only for the quantile approach? If I understood correctly, you could do the same for the time series approach too. I think the important thing is whether the generated (by the weather generator) precipitation (P) is a better representation for the catchment and/or if the rainfall-runoff model simulates catchment runoff better with this P. If yes, I would use only this (i.e.

Stage 2 and 3) in both cases, otherwise only the observed P (i.e. Stage 1 and 3). Or use both in both approaches: in this case the results should be presented/discussed in pairs, otherwise it becomes more confusing.

We could not merge stages 1-2 for the time-series approach because in stage 1 we calibrated the hydrological model using observed precipitation and observed discharge. In stage 2 we already use generated precipitation but we do not have a matching discharge observation set. We could not use observed precipitation in stage 3 even if it was better performing with the model because stage 3 is the future climate. Merging stages 1 and 2 is only meaningful in the K2 approach because there we calibrate the model using the observed discharge data and the generated precipitation.

R1: P522, L24 - P523, L5: The assumptions should be part of the method section. + R1: P523, L6-24: These are known, no need to repeat this way, but it may be useful to discuss if these known limitations would have influenced the conclusion drawn on the hypotheses.

These assumptions are not specific to our study; they rather belong to the generally accepted methodology for assessing the hydrologic impacts of climate change. Accordingly, the limitations arising from these assumptions have been discussed before, but we think that they are still important to put our results in context (that is there is a huge amount of uncertainty that doesn?t show up despite the detailed analysis of certain uncertainty components).

If these issues don't need to be discussed then the entire 4.3 section can be removed with an optional move of the assumptions to the methods section.

R1: P524, L14: '. . .typical'

Will be changed to 'common'.

We ended up at 'already well known'

R1: Tables 1-4: I do not think that these tables need to be in the main body of the manuscript. I suppose the model is not used for the first time. So the previous publications can be referred to for the details of the parameters and distributions. The relations of these parameters or parameter values were not particularly discussed in the results and discussions, therefore these details do not necessarily help the readers.

Some of these details may be presented as supplemental materials.

These tables will be moved to supplementary materials.

Tables 1 – 4, and 6 were moved to Supplementary Material

R1: Table 5: Separate table is not needed. This can be easily described within the text.

OK, data will be moved to text.

Done

R1: Table 7: Are these numbers same as in the lower panel of Figs. 6-9? If yes, the lower panel of these figures should be removed. Then these figures (6-9) can be combined into one with four panels. + R1: Figs. 6-9: Please see my comment on Table 7 above.

OK, Reviewer #2 has proposed a similar change as well.

We removed the lower panels and merged the upper panels into a new Fig. 6

R1: Fig. 3: What is 'climate' (second column) here? Temperature, Radiation . . .? Better to specify them directly.

Climate means here the theoretical distribution of weather parameters. Our weather generator creates time-series of temperature, precipitation, relative humidity and short-wave radiation. Potential evapotranspiration is calculated from these parameters and serves as an input to the hydrological model besides precipitation and air temperature.

As we did not consider transient climate change we used two discrete weather generator parameter sets for present and future climate. Present means the distribution characteristic to the calibration period (1981-2010) while future is the distribution between 2036 and 2065. The observed climatic variables in the 1981-2010 period can be considered as a single realization of the 'present' climate.

This will be explained in the text in more detail.

We changed the table header in Fig. 3: The 'Climate' column specifies the dis-

tribution ('present' / 'future') and the 'Weather' column tells whether the actual weather data are observed or generated. This is now mentioned in the caption.

REVIEWER #2

R2: A wide range of uncertainty analysis methods exists involving different levels of mathematical complexity and data requirements. The appropriate method to be used depends upon the nature of the problem at hand including the availability of information, model complexity, and type and accuracy of the results desired. As expected, the uncertainty results varies with the method used to estimate them, this is nothing new (see, e.g., Pappenberger et al., 2006). Different uncertainty analysis methods have their limitations and assumptions. Hence it is important is to test whether these assumptions are valid or realistic under the given modelling situation and problem at a hand and whether the results are realistic and consistent with the observation. However in climate change studies, it is impossible to get the ?truth? observation of the future. This raised the important question of what method is the best for the uncertainty assessment in climate change studies, which is not properly addressed in the paper. I think one should be able to choose an uncertainty analysis method under given assumptions and limitations; and to test whether these assumptions are valid or realistic under the available information he
she had.

We do not fully agree with the statement that we do not address the question of which uncertainty assessment method should be used in climate change studies. What is true is that we did not explicitly recommend any of the tested uncertainty assessment varieties for all kinds of climate change impact studies in general, as we see no way to decide which method will be better suited under the unknown future conditions for which purpose. However, what we found was that the classical time-series uncertainty assessment approach had several problems when flow quantiles were the predicted parameters. From this the most precise conclusion we could draw was that if the flow quantiles are the aims of prediction then a direct quantile uncertainty approach can eliminate most of the problems associated to methods designed to estimate time-series uncertainty. In this sense, we address the issue of what approach should be considered for a specific need.

In contrast to informal methods, formal statistical approaches are based on explicit

assumptions, where mathematics defines unambiguous rules how a quantity of interest is calculated. The assumptions of the applied likelihood function should conform to the statistical properties of the true error process. For the implications on the difference among uncertainty predictions please refer to the answer to the comment of Reviewer #1. Even if one comes up with the perfect error model for the observation period, it would be impossible to prove in advance that the same good performance and thus the meaningful uncertainty predictions would prevail in the future.

We added a section to the Discussion about the relativeness of uncertainty (now 4.1)

R2: The author claimed that the source, structure and composition of uncertainty depended strongly on the uncertainty analysis methods. But what I found missing is that these aspects of uncertainty are not explicitly defined or mentioned in the paper. The authors have used different terms for different sources of uncertainty inconsistently throughout the paper, for example, hydrological, climatic, predictive, future climate, meteorological, total, final, future weather, weather generation, future climate uncertainty etc. Describe how these sources of uncertainty are represented in each stage of their experiments. In the current form all these sources are mixed up and confusing. Does uncertainty in each stage represent single source or accumulated source, for example does uncertainty in stage 2 represent uncertainty in stage 1 + uncertainty in stage 2 or does uncertainty in final stage 4 represent all sources of uncertainty? In this paper this is very important to understand what these terms refer to and how they are estimated. Note that it is often difficult to disaggregate total uncertainty into their source components because of their interaction and non-linearity of the model. Not to confuse, I think it is better to use only three terminologies, e.g. climatic, hydrologic and total uncertainty. Climate uncertainty is represented by 10 GCM-RCM scenarios, hydrologic uncertainty by Bayesian methods and total uncertainty is the combination of both. I strongly suggest to use the consistent terms throughout the paper including in the tables and figures.

Yes, we failed to give a complete overview of the different uncertainty terms we used. We used accumulated uncertainty in each stage because we did not attempt to disaggregate the individual sources. In retrospect this seems to be a good choice as we saw several examples where uncertainty components accidentally cancelled

out each other from one stage to the other.

We agree that climatic, hydrologic and total uncertainty are the most important categories. However, it is sometimes necessary to use sub-categories, such as the uncertainty of downscaling. The weather generator (WG) is also a model and as such it is an imperfect projection of reality, which results in errors on its own right. We think that the uncertainty of realizations of future climate is the climatic uncertainty and it includes weather variability, GCM uncertainty, RCM uncertainty and WG uncertainty as well.

We will add a short paragraph to the introduction defining all uncertainty terms and their relations. In the text each term having alternative names will be unified ('future weather' / 'future climate').

The first part of section 2.4 now presents an overview of the uncertainty terms we use and their relations. Naming was unified as well.

R2: P502, L23-L24: This sentence is completely wrong. The uncertainty assessment method depends on the source of uncertainty to be analysed, but not vice versa.

Yes, the sentence is wrong. What we wanted to say was that the choice of the uncertainty assessment method actually determines what sources of uncertainty can be identified at all. It will be corrected to reflect the true message.

The sentence was corrected.

R2: P503, L16: .... in the near future.

OK

Done

R2: P503, L22-P5043: Do not describe the methods in the Introduction section; can be moved to method section.

This part will be removed according to the suggestion of Reviewer #1.

Done

R2: P505, L20: What is predictive uncertainty of hydrological models? If I understood correctly in this paper it is uncertainly of hydrological models given that forecasts are perfect (no uncertainty), but need to define explicitly (see general comments).

We define the predictive uncertainty of the hydrological model as the error between the predictions of the deterministic model and the observations of the true modeled quantity under observed climatic boundary conditions. For the future the hydrological uncertainty would be the total predictive uncertainty if the weather forecasts for the future would be accurate within the present weather observation uncertainty.

This is now defined in the first part of section 2.4

R2: P505, L28: the most popular formal and informal likelihood calculation methods in uncertainty analysis ....

OK

Done

R2: P506, L1: hydrological predictive uncertainty due to invalid statistical assumptions...

OK

Done

R2: P506, L17-L19: I do not agree with this statement. In the time series approach quantile Q5, Q50 and Q95 have already computed from the model realizations.

This is a misunderstanding. In the time-series approach the flow quantiles are not calculated from the outputs of the hydrological model directly, because that would not reflect the true discharge as the model has systematic errors. Instead, quantiles are calculated from the model output plus the bias. As the bias is a stochastic process, there are different realizations of this model+bias composite even for a single parameter set. However, the addition of a stochastic process to the model output results in strongly biased flow quantiles, instead of increased variance. That?s what we meant here as a statistical difficulty.

R2: P507, L20: Show rainfall stations in Fig. 1.

It is problematic because both stations are well outside the figure boundaries. The relatively long distance to the rain gauges is mentioned in the text.

R2: P508, L5: It is not clear why stochastic weather generator is used? I suggest to have a small paragraph after section 2.2 to explain the context of three climate data used in this study.

OK, it will be done. The rationale for using a weather generator was that we wanted to have a realistic variability in the future weather series. For the study regions raw GCM output was very strongly biased for annual total precipitation and underestimated climatic variability.

Added a paragraph to 2.2 to explain why we needed a weather generator.

R2: P509, L5: Number tables or figures based on their first citation in the text, i.e. the first mention of Table should be Table 1.

OK

Done

R2: P511, L24 - L26: What period was used to compute these flow quantiles?

We used the calibration period (2000-2009) to compute these quantiles.

R2: P512, L10: Is total uncertainty the combination of hydrological and climate uncertainty? Define these terms explicitly before they are used (see general comments). What does mean by a hybrid approach?

OK, we will define each term in advance (see answer to general comment #2). We use a mixture of model ensembles (for GCM output) and an explicit statistical error model (for hydrology). Normally these are not mixed and that?s why we called our approach a hybrid.

Uncertainty terms and the fact that we use accumulated uncertainty are now described in the first part of section 2.4

R2: P513, L2-L9: This sentence is not complete, rephrase the sentence e.g. then the flow quantiles were selected for each ....

OK

Done

R2: P513, L2-L9: Use same terminology throughout the text and figure. For example, weather data in the text vs climate in the Fig. 3. From the Fig. 3 it seems climate data refers to all weather data except precipitation.

The 'climate' column in Fig. 3 must have been poorly phrased because it caused a misunderstanding for both reviewers. As mentioned above, climate would refer to the distribution of climatic parameters (incl. precipitation). The figure caption will be changed to deliver a clear message.

Done

R2: P513, L15-L16: What are flow indices? Are these referring to flow quantiles?

Yes, will be changed to flow quantiles.

Done

R2: P513, L15-L16: In general 10% of the observed flow data should be outside the Q95 and Q5 interval. In stage 3, the observed flow data in not known. However in stage 1 and in stage 2, it is possible to compute the actual percentage of observed flow data inside the Q95 and Q5 interval (percentage coverage). Did you compute the percentage coverage in these stages?

It is only possible in stage 1 as there are no observations for stage 2 (which would be a different realization of the past). This was done in the study presented in Honti et al. (2013) for the very same observed data from the Mnchaltorfer Aa catchment. The actual coverage was a bit higher than 10%.

R2: P514, L7: The details of the approximate quantile likelihood function are described in Appendix B and C. (Appendix C is not referred

to the text).

OK

Done

R2: P515, L20-L22: Stage 3 GCM-RCM chain is better than Stage 2 weather generation. I do not see significant uncertainty in the figure 4, indeed quantiles are matching very well.

This is actually by accident. As uncertainty accumulates along the stages, the stepping back to the right QQ curve from stage 2 to stage 3 means that the high bias of stage 2 was accidentally cancelled out by a similarly large bias introduced in stage 3. As one cannot guarantee such a lucky elimination of uncertainty effects the overall conclusion is that the uncertainty of stage 3 is significant, despite that the final QQ curve happened to fall on the proper location.

R2: P517, L15: Given that the paper is a bit long, section 3.2 and section 3.1 can be merged.

We think that merging sections 3.1 and 3.2 would not easily reduce the length of the manuscript without getting rid of a significant portion of findings describing the differences the test catchments. Given that other parts will be strongly reduced, we would retain from this option unless even further shortening is necessary

R2: Fig 3: I understand what WG, CC means, but it is never defined in the text. This is confusing text - White stars indicate stages of calibration. Does it mean that calibration of hydrological parameters is done only in stage 1?

OK, all abbreviations will be defined. Yes, the star means that we only calibrate in stage 1. In other stages it would not be possible in the absence of measured flow data.

Abbreviations are defined and figure caption is changed

R2: Fig 4: To be consistent with Fig 3, use stage 0, 1, 2, 3 and 4 in the legends if these refer to observation, hydrological model, weather generator and GCM-RCM chain.

OK

Done

R2: Fig 4:9: What does Box and whisker diagram represent, how these are calculated?

From Fig 5 to Fig 9 the whiskers represent the 95% uncertainty interval of the flow quantile. These are calculated from the different realizations.

R2: Fig 5: I think figure 5 contains too much information.

Background signs ('WG' and 'FUTURE') will be removed to improve clarity.

Done

R2: Fig6 – Fig10: Use Stages in either y axis or x axis in both panels. Use the consistent notations (e.g., Stage 0, not St 0). Table 7 also presents the relative change, so no need to repeat in figure, I strongly suggest to remove the bottom panels. Then Fig 6, 7 and 8 can be combined together to compare TS, K1 and K2 approach in a single plot.

OK, figures 6-9 will be combined.

Done

R2: Table 1: Define all affected storage symbols.

OK

Done

R2: Table 2-4: Endnote symbols a and b need to be swapped. Table 2: Table 2 does not present parameters for paves area.

OK, will be corrected

Done

R2: I think Table1-4 can be removed or moved to the appendix.

OK

Done

References

Beven K., Paul Smith, Jim Freer, Comment on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology" by Pietro Mantovan and Ezio Todini, Journal of Hydrology, Volume 338, Issues 3?4, 30 May 2007, Pages 315-318, ISSN 0022-1694, http://dx.doi.org/10.1016/j.jhydrol.2007.02.023.

Kuczera G., Dmitri Kavetski, Stewart Franks, Mark Thyer, Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, Journal of Hydrology, Volume 331, Issues 1?2, 30 November 2006, Pages 161-177, http://dx.doi.org/10.1016/j.jhydrol.2006.05.010.

Mantovan P., E. Todini. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. J. Hydrol., 330 (1?2) (2006), pp. 368?381. http://dx.doi.org/10.1016/j.jhydrol.2006.04.046

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, Water Resour. Res., 47, W11516, doi:10.1029/2011WR010643.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46, W10531, doi:10.1029/2009WR008933.

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.