Dear reviewer,

we greatly appreciate your thoughtful comments that helped improve the manuscript. We trust that all of your comments have been addressed accordingly in a revised manuscript. Thank you very much for your effort. In the following, we give a point- by-point reply to your comments:

>I agree in many points with the first reviewer. The manuscript is well written and clear. The idea to >combine advantages from MO-calibration and depth sampling sounds interesting and should be >further tested. However, there are a number of issues that need to be addressed before publication. >My major concern is that the manuscript is a mixture of two topics that should be addressed >separately and each one in more depth.

In our opinion a the robust multi-objective calibration technqiue requires both a reliable procedure for the estimation of the Pareto-optimal set and an effective post-processing technique. This is why we include a limited discussion on reliable techniqes for the estimation of the Pareto-optimal set.

We showed that the developed MO-PSO-GA (that was already succesfully applied within the frame of previous single-objective ROPE algorithms) can both outperform other single-strategy search strategies and help to improve existing advanced multi-method search approaches. That is why we integrated the developed approach into the existing approved multi-method AMALGAM framework in order to achieve an optimal effectivity and efficiency for the estimation of the Pareto-optimal set.

Nonetheless we agree to your comment and are convinced that the application of the depth based sampling technique as a kind of post-processing technique is the main thrust of this paper. Therefore we refurbished the manuscript by eliminating the section 2.1. (since the concepts and definitions of Pareto optimization are rather trivial, according to the comments of the third reviewer). and added further test cases to illustrate the advantages of the depth-based sampling. The depth based sampling in combination with multi-objective optimisation is the key concept of this paper.

>The first topic is the introduction of the new multi-objective calibration algorithm MO-PSO-GA. In >agreement with the first reviewer, I suggest to test the new algorithm in more depth with other test >cases (more parameters, rougher response surface) and to compare it to the performance of more >recent MO-algorithms (e.g. AMALGAM).

We agree to you. Therefore, we integrated many more complex test problems with up to 30 parameters. The selected set of test problems is oriented to the one used in the first presentation of the AMALGAM approach. We showed that the developed MO-PSO-GA can both outperform other single-strategy search strategies and help to improve existing advanced multi-method search approaches. That is why we integrated the developed approach into the existing AMALGAM framework in order to achieve an optimal effectivity and efficiency for the estimation of the Pareto-optimal set.

>The second topic is the combination of MO-calibration with depth sampling in the pa- rameter space. >This topic should be presented independent of MO-PSO-GA, since this combined approach can be >used independent of the way the Pareto front is determined. Also, I recommend that the authors >design synthetic test examples also for this topic where they demonstrate the clear advantage of the >new approach for improved validation results in addition to the real-world case study. From the results >presented I got the impression that the main effect of depth sampling is to exclude the outer margins >of the Pareto front.

We included synthetic test studies based on the extended set of test problems used for test of the multi-objective optimisation problems. We introduced a simple uncertainty model into these problems and show that the deep parameters provide a good approximation of the set of Pareto-fronts in the uncertain objective space. Furthermore the deep parameter vectors are less sensitive to small changes in order to approximate the set of Pareto-front for the given problems.

The depth based sampling is done in the parameter space and NOT in the objective space. It identifies deep parameter vectors in the central region of the complete Pareto-optimal set and not in the Pareto-front. For many problems, e.g. the presented real-world case study with the model WaSiM-ETH, the central regions of the Pareto-optimal set correspond to model performances that are especially located in the central part of the Pareto-front for the calibration data. Consider however that there is no one-to-one mapping of these regions. Thus, the depth based sampling does not just cut the tails of the Pareto front. Furthermore we compared the depth-based sampling with an cutoff approach based on subjective threshold values for the individual criteria. Notwithstanding the fact that the threshold values are subjective and hard to estimate a priori, the depth based sampling provided better results than the simple cutoff technique. Additionally the data depth provides the possibility to relate the likelihood of the parameters to their depth.

>There is one point in the method that I am a bit puzzled and that should be thoroughly discussed by >the authors: I do not understand why a large effort should be spent to sample the entire Pareto front >(one of the problems to be addressed with MO-algorithms) and in the next step to reduce the set to >the center part of the front.

The data depth can be computed to measure the centrality of a point with respect to a given point set or distribution. It is necessary to estimate the complete distribution or point set in order to compute the data depth of a point with respect to this set. In our case it is thus necessary to identify the complete Pareto-optimal set in order to be able to sample deep parameter vectors. Consider once again that the deep parameter vectors are deep in the parameter space and not the objective space. Thus, the task of the depth-based sampling is not to reduce the set to he center part of the Pareto-front but to the central region of the Pareto-optimal set in the parameter space. There is no one-to-one relationship between these two sets.

>Finally, the case study suffers from a problem, that unfortunately is very common in hydrology. The >authors take an existing model, which they are aware of that it is not well suited (as presented in the >discussion p. 3715, l. 1-5) and make a large effort (high performance computations required to >produce the results) to find a valid parameter set. It is not clear, why not more effort goes into finding a >better description/model for the catchment?

We partly agree to this comment. Let us explain the background of this study. The further development of the data depth technique and its application for the calibration of hydrologic models focussing on flood forecasting is part of a bigger project whose goal is the development of an operational flood forecasting framework for fast responding catchments.

According to our experience and many other studies the WaSiM model is one of the best suitable for such tasks (cf. Cullmann 2006, Marx 2007, Grundmann 2010; the references are given in the manuscript). We chose the Rietholzbach catchment for this study because it has been intensively monitored for a longer time period and a lot of modelling studies have taken place in this catchment, also with the WaSiM model. Furthermore previous studies showed that improvements in the model calibration for small headwater catchments can due to the huge uncertainties tremendously improve the model performance for operational monitored gauging stations in the lower reaches (cf. Cullmann 2006 and Grundmann 2010). We intensively communicated with the research group monitoring the Rietholzbach catchment considering the question of the model decision. According these discussions and our experience the WaSiM model is due to all shortcomings one of the most suitable for the modelling of flood events. A large amount of the uncertainties in small headwater catchments is averaged out in lower reaches that are nonetheless still subject to a very short response time.

Of course one can put lot of effort to find even better descriptions/models for such small catchments focussing on flood events. Such research studies are very sensefull and important. However, this goal is not within the focus of this research paper.

Kind regards,

Thomas Kraußе, Johannes Cullmann, Philipp Saile and Gerd Schmitz