Dear Prof. Efstratiadis,

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:

General comments

1. The paper presents a framework for model calibration, in which data depth measures are used within a new multiobjective optimization algorithm, to identify robust non- dominated solutions. The effectiveness and efficiency of the algorithm are tested on the basis of two typical benchmark problems, while the entire framework is employed in a real-world case study, involving the calibration of a hydrological model (WaSiM) against a number of flood events, in a small experimental catchment in Switzerland.

-

2. The topic of the paper is interesting and the manuscript is well-organized and well-written. However, its originality and novelty are questionable. For, there are two papers that have been recently submitted to HESSD dealing with a very similar subject, where the same algorithms, the same model and the same study area seem to be recycled (Kraußе and Cullmann, 2011a, b). Parts of the text are verbatim reproduction, while some of the tables and figures are repeated. In order to be suitable for publication, a substantial review is essential, to remove the already published components of the paper and provide really original material.

It is true that we submitted two further papers dealing with single-objective robust parameter estimation. According to editor comments we merged these two manuscripts into one paper dealing with single-objective robust parameter estimation. This paper however deals with multi-objective calibration problems. The presentation of the algorithm and all test problems are substantially different than in our other submitted paper. The used hydrologic model and the case study area are the same. This however makes the different approaches compareable and is in our opinion better than to apply the concepts to absolutely different case study areas. The material presented within this paper is substantially new and no verbatim reproduction. For the introduction of the WaSiM model we use the similar Figure as in another submission. Besides that we mark all used references and the material fom other papers.

3. Despite the very promising title (“chances for improving flood forecasting”) and some important hints that are discussed mainly in the first two sections (and have been also discussed in the two aforementioned papers), my final impression was rather about “another calibration exercise”. Specifically, the very challenging task of identifying “robust” (realistic? behavioural?) parameters, which is of major interest in hydrological modelling, is addressed just as an algorithmic issue that is handled through the so-called “Robust Parameter Estimation” (ROPE) approaches. Naming a computational procedure, even the most sophisticated one, as “robust” is, to my point-of-view, not useful, impossible to understand and even misleading. One can find a large number of alternative calibration methods and strategies in the hydrological literature – which of them are robust and under which premises? Are the SCE and GLUE methods, with thousands of applications (and citations) worldwide, robust or not? Who is able to identify the most robust solution, an expert hydrologist or an “expert” algorithm?

We entitled the developed calibration approach based on the data depth technique „Multi-Objective Robust Parameter Estimation Algorithm (MO-ROPE)“ according to the first publication of Bardossy and Singh (2008) who invented the term „ Robust Parameter Estimation Method (ROPE)“. Of course the pure application of an advanced computational prcoedure is not sufficient for the achievement of robustness. According to your comments we added a discussion on this issue in the paper (see page 2 and page 24/25).

4. In general, the parameter estimation problem is satisfactory posed, although some of its aspects may required more development (e.g. the concepts of uncertainty and parsimony are rather poorly explained). The authors, quoting Bardossy and Singh (2008), rightly state that a key goal of model calibration is to find parameters that perform well both in calibration and validation, and at the same time ensure “hydrologically reasonable representation of the corresponding processes” (p. 3699, line 22). The data depth technique, which was initially proposed by Bardossy and Singh (2008) for single-objective functions and now generalized for multiobjective calibration, is next introduced as “a possible approach to achieve this goal” (p. 3696, line 14). However, there is until now little experience with this strategy, to justify such an imperative state- ment. In addition, it is very difficult to trust any automatic method not accounting for the

role of knowledge, in terms of hydrological experience and understanding (cf. Boyle et al., 2000). There are also some practical disadvantages, which are revealed in the case study. Why implementing a computationally expensive technique with negligible physical interpretation, just for rejecting part of the non-dominated solutions that lie in the extremes of the Pareto front? As the authors claim “the tails of the Pareto front estimated in the calibration are not required” (p. 3711, line 20). However, this is not a surprising conclusion: even an elementary approach, based on subjective yet realistic cut-off thresholds, could easily distinguish such “non-behavioural” solutions with negligible effort (cf. Efstratiadis and Koutsoyiannis, 2010).

We weakened our statement and introduce the ROPE method now as „Recent studies using further developed versions of this methodology (e.g. Krausse, 2011a) showed the potentials of the depth based parameter sampling for the estimation of robust parameter vectors.“ .

The depth based sampling is done in the parameter space using the Pareto set. The deep parameter vectors are usually denser distributed in the central part of the Pareto front. There is however no one-to-one mapping of the center region of the Pareto set and the central part of the Pareto front. Furthermore the data depth technique opens the possibility to relate the data depth to the likelihood of the parameter vectors. Nonetheless we compared our approach with the approach using cut-off thresholds and discuss the results.

Besides this consider that the depth based smapling is usually no computationally expensive technique. When the focus is on distributed process-oriented and partly physically-based models, the computing time for the model runs required for the estimation of the Pareto set is very much higher than the computational effort required for the application of the depth based sampling technique.

5. Regarding the presentation of the methodologies, I am afraid that the authors deal with too many issues, thus failing to adequately develop their ideas and highlight their effort. This mainly involves the MO-PSO-GA algorithm, which is presented in a too synoptic manner (section 3.1) that makes difficult to understand the procedures and, especially, the innovations (the search scheme is not fully original, but it is based on an effective combination of various techniques). The same problem exists with the GenDeep function (section 3.2), which was very hard to understand, without referring to the literature. Since this is a relatively new method, I would suggest spending sufficient effort on explaining the details, and, at the same time, drastically eliminating (or even removing) section 2.1, since the concepts and definitions of Pareto optimization are rather trivial.

We refurbished the presentation of the algorithm and tried to introduce the concepts of this paper in more detail in order to provide the developed approach in a more understandable way. Furthermore we eliminated the section 2.1.

6. The testing framework for evaluating the performance of the MO-PSO-GA algorithm, on the basis of two rather simple benchmark problems, is insufficient. To make sense, this test should involve a representative sample of multiobjective functions, including high-dimensional problems (in terms of both the number of control variables and the number of objectives), and different levels of complexity, regarding the geometry of the Pareto front.

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.

7. Although the title focuses on flood forecasting, little attention is given to the specific aspects, challenges and peculiarities of this problem. The authors could also take advantage from related applications (e.g., Pappenberger et al., 2007; Moussa and Chahinian, 2009), thus providing a much more attractive paper.

Tahnk you for this hint. The publication of Moussa and Chahinian (2009) provides some useful ideas. We compared our results considering the comparison of single and multi-objective claibration with some general findings of this paper.

Minor comments and technical corrections:

1. Page 3697, line 2: “The developed approach is tested on synthetical data.” I do not agree characterizing the benchmarks problems as “synthetic”. The term is used when contrasting to actual or historical conditions.

Done.

2. Page 3697, line 20: ". . . where = ( 1, . . ., 2) is a d-dimensional vector" Use bold fonts for and vectors and change 2 by d.

We refurbished the definitions of the test problems. In the current manuscript all parameters in the test problems are now denoted with bold letters (**x, y, …**)

3. Page 3698, line 11: "Often both terms [Pareto set and Pareto front] are used syn- onymously." The authors have right, but they should further emphasize on the negative impacts of this practice, which often leads to misleading conclusions.

We agree to you. This concept is very important for the understanding of this paper. That is why we moved this statement form the eliminated section 2.1. into the introduction (page 2) and added another statement to avoid any misunderstandings or misleading conclusions in the remainder of this paper.

4. Page 3700, lines 5-8: "One starting point which recently attracted rising scientific interest is a more intelligent selection of the calibration data . . ., another one is the development of advanced methods for the identification of parsimonious model param- eters" Parsimony is associated to the model structure, not the parameters. It is a key concept in modelling, asking to represent a model structure with as few parameters as possible, where the essential number of parameters depends on the available information. Please, see the related discussion and the literature provided by Efstratiadis and Koutsoyiannis (2010).

We considered this issue and introduced a more detailed discussion on this issue on page 2 with a reference to your publication Efstratiadis and Koutsoyiannis (2010).

5. Page 3700, line 29: It is preferable using "simple" instead of "small" (example).

Done.

6. Page 3703, line 3: Change to read "population-based".

Obsolete due to a reformatting of the manuscript.

7. Page 3705, eq. 2: It is very hard to understand this equation. What is the vector u? What is the symbol T?

The symbol T indicates that the labeled vector is transposed. This equation counts the number of points in the halfspace defined by the normal vector u through the parameter vector . The minimum number of points for all possible halfspaces (defined by any possible normal vector u) is the halfspace depth hdepth. We included a figure that illustrates this for the 2-dimensional case (Fig. 2).

8. Page 3705, lines 13-17: "The developed solution addresses some of the drawbacks of existing multi-objective and robust single-objective calibration procedures. It pro- vides a good possibility for the identification of robust model parameter vectors with respect to multiple calibration objectives." This statement is not justified and should be removed.

9. Page 3706, section 3.3.1: How are the constraints of test function 1 handled?

We included a discussion on this issue. See page 3 and 4.

„The AMALGAM framework contains a simple handling of boundary constraints. Infeasible solutions that are out of the bounds are just set to the bounds in order to preserve their feasibility. In order to enable the framework to deal with more complex bounds, we added some features of a con- straint handling technique based on adaptive penalty func- tions and a distance measure proposed by Woldesenbet et al. (2007). This method uses the number of feasible individu- als in the population in order to be able to control whether a modified objective function focusses just on the objective values or the constraint violation. As long as all members of the current population are feasible the objectives remain unchanged.

10. Page 3709, line 3: Change to read "Wolpert and Macready (1997)".

Done.

11. Page 3710, lines 15-16: “In a first case study we calibrated WaSiM with the MO-ROPE algorithm in terms of two objective functions: rPD and NS. Additionally we applied the single-objective robust parameter estimation algorithm ROPEPSO to this problem using rPD, NS and their aggregate FloodSkill as objective functions.” The three functions (PD, NS, FloodSkill) are not defined.

The old manuscript already provides a Table with a definition of all used objective functions (Table 3). In the new mansucript we introduce the performance criteria in more detail (page 15/16) and provide the mentioned Table with a reference in the text (Table 8).

12. Page 3715, lines 21-25: “This underlines that robust parameter estimation can identify the most robust solutions within the given constraints. However, a good selection of appropriate calibration objectives and a suitable model structure are as important as a reliable and robust model parametrisation.” This explains why the use of term “robust” should be done more carefully. Robustness is a combination of all the aforementioned aspects, i.e. the model structure, which should be as parsimonious as possible, the data, which should be as representative as possible, and the calibration.

Your comment helped us to improve this paper. We edited this statement and included a discussion on this issue. Once again we conculde it at the end of the last case study on page 24 („Furthermore it underlines that a successful robust modelling does not require just an advanced parameter estimation procedure but also the selection of a as parsimonious as possible model structure, representative calibration data and appropriate calibration objectives. The combination of multi-objective optimisation and depth based parameter sampling can be a good tool to obtain robust model parameters. Alone for itself the depth based sampling is however not sufficient to achieve robustness.“)

Kind regards,

Thomas Kraußе, Johannes Cullmann, Philipp Saile and Gerd Schmitz