Response to Referees
(Note: this response is given for both papers: hessd-2009-249 and hessd-2009-250)

Referee #1

Comment: *However, still the study suffers from one of the generic problems of machine learning techniques: The single approaches can be parameterized in very different ways, having much of an influence on the model performance. For example, results from different artificial neural networks might differ substantially depending on the chosen initializations, the learning rate, the number of hidden nodes, the type of activity function, and the learning algorithm used. Although this issue is addressed occasionally in the paper (e.g., with respect to SVM), it is not considered in a systematic way. Consequently, readers would like not only to have a comparison between single realizations of different techniques, but to get some information if these differences are significant or not. It has been often argued that differences between different machine learning techniques might be small compared to the variability encountered within different realizations of models of the same type. Only when that is considered, the paper could really become a benchmark paper*

Response: We fully agree with the reviewer that "The single approaches can be parameterized in very different ways, having much of an influence on the model performance". In this paper an attempt is made to try many parameterizations of the considered models and in comparisons with other models to use (in a certain sense) the best one. This does not concern SVM only. As we stated in Part II, section 2.1, the ANNs were executed 200 times with 200 different random weight initializations. The best model of the 200 runs was identified as the best ANN model. The 200 runs were repeated with each trial of number of hidden nodes (ranging from 3-13). Then the best out of the 200 and optimum number of hidden nodes was selected. In case of GP, many operators were tested with number of generations up to 300, and program size ranged from 80-512 bits. In case of EPR, all possible model structures provided by the EPR software were tested. All of these, with all techniques, done with 12 different realizations of the datasets, which naturally trigger different output possibilities of each technique. Really, from a practical point of view, we cannot see the need of much else to make this a benchmark work for others to build on. We never intended or stated that this work is all-inclusive, nor did we say, figuratively speaking, that all pens shall stop writing and all ink shall dry out!

Comment: *I do not feel happy with the study being split into two papers. On the one hand, the papers are very comprehensive and cannot be presented as two stand-alone papers (methods described in the 1st papers, results presented in the 2nd paper, references nearly exclusively given in the 1st paper). On the other hand, they are partly redundant to each other (e.g., section 2 of 2nd paper and section 4 of the 1st paper). Thus, I would suggest the following:*

1. Skip sections 3 and 5 in the 1st paper, and make it a pure review paper. It could be published nearly as it is (but see specific comments below) and would be a valuable source of information.

Response: We definitely understand the complications with two-part papers. Whenever two-part papers are submitted, one of the classical comments of some reviewers is to integrate them into one paper. Our viewpoint on this issue is:

There is a lot of material beyond what can be included in one paper. Having the two papers separated in two independent ones is an option, given that they are published in the same issue. There are review papers published already. The review presented in the first paper is given as preparation and background for justifying the need for the proposed experiment. There is no redundancy between section 4 of Part I and section 2 of Part II. In part I, we talk about the techniques and how they are different from each other. In part II we show how the techniques were implemented in this particular experiment.

What we suggest, which we think will accommodate this comment, is to keep section 4 in the first paper, and make it almost stand alone review paper that highlight the issues and the need for such a comprehensive modeling experiment. Then, take section 5 to the second paper and make it also almost stand-alone paper. This way each paper can be read separately, however, both are still twp-part papers that together give the entire picture off the issues and the first experimental step on the long road of addressing them.

2. Restrict the model comparison to a single paper. That would require condensing the description of models, of the data set (refer more to relevant papers), and of the results. I recommend focusing more on generic features rather than to the performance of single models on single data sets (see comments below).

Response: Yes, according to our proposal mentioned earlier, the second paper will be slightly modified.

3. All the necessary details that cannot be presented in a HESS paper should be compiled in a technical report. See, e.g., the report by Maier and Dandy (1995) that complemented their paper in Water Resources Research 1996 on artificial neural networks.

Response: Indeed, it is a possibility in case of large amount of material; however, in this particular case we do not see a need for this. With our proposal to restructure the papers for the better balance to have almost-stand-alone papers, this would be automatically addressed.

Specific comments

1. The term "data driven modeling techniques" in the title is too generic. I suggest using the term "machine learning techniques" instead. At least, the latter term should be introduced in the text.
2. P. 7050: Section 2 summarizes various studies that compare different machine learning approaches. However, the approaches discussed here and some of the acronyms used are only introduced in section 4. It should be done the other way round.
3. The 1st paper, especially the introduction of the different machine learning techniques (section 4) provides too many details and acronyms that are not explained, e.g., "full

method, grow method, and ramped half-and-half-methods" (p. 7070, l. 7) or "SRM" and "ERM" (p. 7073, l. 1-2).
4. A list of acronyms should be provided.

Response: On item 1: "data-driven modeling" is a widely accepted term and we see no reason to change it to "machine learning" in the title. DDM is wider since it follows the logic of generally accepted procedures of modeling (which may use the methods of ML but conceptually is different from "learning"). As suggested, the term is introduced in the text of the revised manuscript.
We agree with items 2-3. The revised manuscripts address these editorial points.

Item 4: typically in HESS list of acronyms is not required, so we simply define all variables and acronyms explicitly in the text.

Technical corrections
1. P. 7057, l. 16: What does "less than the ones created" mean?
2. P. 7058, l. 13: Please define "naïve models".
3. P. 7059, l. 11: What does "domain knowledge" mean?
4. P. 7060, l. 23: Be more specific than "has its distinct capabilities and advantages".
5. P. 7060, l. 24: Why was that a "preliminary" study?
6. P. 7064, variables "O", "P", "r": use upper or lower case letters consistently.
7. P. 7064, l. 21 and subsequent lines: Instead of merging the residuals of different models, the uncertainties of the fitted probability distributions should be given

Response:
1. We meant "less than the ones introduced by the ways such techniques were implemented". This is fixed/rewritten.
2. Commonly used terminology, but is briefly defined in the revised manuscript.
3. A term commonly used in machine learning and artificial intelligence (meaning "knowledge about the area (domain, class of systems) in (for) which models are built"). Often may mean "expert knowledge". In the revised manuscript is briefly explained.
4. Is specified in the revised manuscript.
5. "Preliminary" will be removed.
6. R is correlation coefficients but r is residual values. In the revised manuscript the use of small r is avoided.
7. These are 12 realizations of the dataset, but all residuals belong to the same technique (e.g., ANNs). This way, the probability distribution really represents all possible model outcomes. We do not see the need or benefit of fitting a separate distribution for each of the 12 realizations for each technique.

Specific comments on the second paper:
1. The term "data driven modeling techniques" in the title is too generic. I suggest using the term "machine learning techniques" instead. At least, the latter term should be introduced in the text.
2. The paper comes along with only 12 references.
3. It is well known that different measures of performance in fact measure only single

aspects of performance (e,g,, see Janssen & Heuberger (1995), Ecol. Mod. 83: 55-66). On the other hand, most of the different measures are not independent from each other. In the climate modeling community, the Taylor diagram has been developed to visualize comprehensive information about model performance (Taylor, K.E. (2001): Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106: 7183-7192). I suggest using that diagram.
4. What fraction of the total variance is explained by the respective models? That could be given, e.g., by the Nash-Sutcliffe model efficiency.

Response:
1. On item 1: "data-driven modelling" is a widely accepted term and we see no reason to change it to "machine learning". DDM is wider since it follows the logic of generally accepted procedures of modelling (which may use the methods of ML but conceptually is different from "learning")..
2. This is because of the nature of this second part, but this will be slightly modified based on the new proposal of stand alone separated papers.
3. Taylor diagram can be of course used; however, we are limited in space and have chosen to adopt measures that are familiar to hydrologists to allow possible future comparison with the results of other authors.
4. There are various ways to calculate model performance, and we have chosen RMSE (that has close relationship with likelihood – a measure adopted in statistical learning), mean absolute relative error, mean bias, and correlation coefficient. To add yet another performance measure, in our opinion, would add much to the analysis and would not change conclusions.

Technical corrections
1. P. 7098, l. 22: Time needed for the GP depends on the used computer.
2. P. 7101, l. 5-6: The ANN does not force highly non-linear structures. To the contrary, it starts with a very smooth and nearly linear structure that becomes highly nonlinear only during later stages of the learning procedure.
3. P. 7103, l. 17 (and others): How do you measure the nonlinearity of the datasets? Could you give numbers for that?
4. P. 7110, l. 10: Why is the R statistic a key indicator in that study?
5. P. 7111, l. 26: Why do you scale to 0.5 standard deviation, not to 1.0 standard deviation, as is usually done?
6. P. 7130, table 15: What does the grey shading and underlining denote?
7. P. 7132, fig. 1: Please use barplots instead of lines.
8. P. 7122-7128, tab. 7-9, 11, 13: Confidence intervals would be more useful than the range. In addition, I suggest to summarize these results in a matrix of barplots with error bars.
9. In general: Fonts are much too small in most of the graphs.

Response:
1. We agree
2. It all depends on the initialization of weights, and, since these are random, and ANN uses combination of non-linear (transfer) functions, it is safe to say that ANN "force non-

linear structures". We agree that "highly non-linear" was too strong, and changed it to "non-linear".
3. To test (non)linearity, we used the linear regression model as a baseline, and we referred to this when we applied the Gamma test. We calculated the error variance (represented by MSE of the regression model) and compared this to the noise variance on the measured output. If linear regression MSE is not higher than the noise variance, the linear model captured the signal, and this would be good measure of the linearity.
4. This was explained in section 3.2, and the reference was also provided.
5. It is possible, we just took the default of the WinGamma Program, which does not allow for changing this. Since this was applied consistently to all datasets, we do not see any significant impact for the value; whether it is 1.0 or 0.5. This is now explained in the manuscript.
6. In the new version it is removed, underlining just highlights the high values of deterioration.
7. In the new version of the manuscript bar plots will be used.
8. This range is for the 12 datasets, so we do not see the confidence interval to be meaningful based on 12 values. These are for the overall error statistics not for all residuals. But we will consider representing them in graphs, if space allows and in the view of the recommendations to reduce the size.

Referee #2

Comment: *However, I suggest to combine both manuscripts in a single, more condensed paper (for example section 4 of the 1st paper can be considerably shortened).*

Response: We addressed this in our answer to Referee #1

Specific comments:
1st paper
1st section: 1) Page 7057 – I think it is not necessary to introduce the term "soft computing techniques" if data driven modelling is used as a general term.
2nd section: 2) When studies about ANNs are discussed, it should be mentioned which architecture is used. Generally (in both papers), the reader may get the impression that the term ANN is synonym with feed forward neural networks (FFNN) (also see below).
3rd section: 3) Why those six DDM techniques were chosen and not others? In addition, I understand ANNs as a general class of architectures and training strategies which represents many data driven techniques – among them FFNN as a global approximation technique. How about local neural approximation techniques?
4) Page C3515 7063 – Please provide also information about the error function which was used when applying each of the six DDM techniques. Also in the 2nd paper I found sparse information about it.
4th section: 5) The description of the various DDM techniques can be considerably shortened: e.g. ANN from Page 7066 – line 26 to the end of the paragraph, GP from Page 7068 – line 18 to the end of the paragraph, EPR from Page 7072 – line 14 to the end of the paragraph, SVM from Page 7073 – line 10 to Page 7075 – line 1, MT from Page 7075 – line 18 to the end of the paragraph. 6) A discussion of dynamic versus

static regression strategies for EPR (Page 7072 – line 18 -21) would also be interesting in the context of the results of the conducted DDM experiments and can be included in the 2nd paper.

Response:
1) Only "data driven" and "machine learning" terms will be used in the revised manuscript.
2) Yes, FFNN were used indeed. It is mentioned in the revised manuscript.
3) Definitely, there are so many other possible techniques. We chose the ones that are more common than others in hydrology, and we are looking forward to other researchers use other techniques and compare them with ours.
4) We always used the mean squared error as an error function. We agree with the comment and this is mentioned explicitly in the revised manuscript.
5) In the revised version of Part I, we will cut portions of what the referee suggested; but perhaps not all, just to keep the paper informative and stand alone.
6) Dynamic regression is one of the techniques that we did not use. We used static type of techniques throughout the experiment and wanted to remain consistent. Note however that in the rainfall-runoff case studies we used the delayed inputs (rainfall) and this makes the models "dynamic".

Technical corrections:
1st paper
1) Page 7057 – last line "technique"
2nd section: 2) Some acronyms like GP, SVM, EPR are used before defined 3) Page 7060 – line 11 "disaggregates . . . from the . . ."
3rd section: 4) Eq. 5 – "IPE_{ij}= . . ."
4th section: 5) Page 7073 – line 15 "where . . ." 6) Page 7074 – line 4 "Müller . . ." – u-umlaut
5th section: 7) Page 7080 – line 26 "subcatchment . . ." 8) Page 7081 – line 15 "(Berger, 1992) 9) Page 7081 – line 16 "More . . ."
figures section: 10) Page 7091 – add labels to x- and y- axis

Response: All editorial corrections will be taken care of.

Specific comments: 2nd paper
2nd section: 1) Pages 7098/7099 – Why did you test several kernels for the SVM models and not different (also linear) activation functions in the case of ANN.
3rd section: 2) The results are very carefully and comprehensively analysed and discussed!
5th section: Page 7113 – line 26 I would also expect benefits if DDM techniques would be combined in ensembles e.g. using bayesian model averaging, see
Th.Wöhling and J. A. Vrugt: Combining multi-objective optimization and bayesian model averaging to calibrate forecast ensembles of soil hydraulic models. Water Resources Research, 44:W12432, 2008. doi: 10.1029/2008WR007154.

Response:

1) The test of the different kernels was brief just to get an indication which one is more suitable. The rest of the experiment was conducted with one selected kernel. Also, the choice of kernels in SVM seems to have considerable attention in literature compared to activation functions of ANNs. TO make things clear: we used non-linear hidden nodes, and linear output nodes (architectures generally accepted for numeric prediction). However making all nodes linear will make ANN just a linear regression model. But again, so many different things can be tested, and so many experiments conducted, but unfortunately there always time limitations. However, we like to acknowledge our shortcomings in this regard, and the corresponding recommendation is added.
2) Thank you!
3) We agree that this would be an excellent thing to do, and we will add this as a recommendation for future work.

Technical corrections:
C3523 2nd paper
1st section: 1) Page 7097 – acronym AMI is used before defined
5th section: 2) Page 7114 – line 12 "pre-processing . . ."
figures section: Page 7138/7139 – ylabel->closing bracket

Response: All editorial corrections are taken care of.

Referee #3

Comments: *The authors made a comparison of several data driven modeling techniques over several different hydrological datasets. Although much effort was put into this endeavour, the final results were not very satisfactory, and I would recommend major revision.*
1. The division of the work into two papers is not well justified. While it is not uncommon for authors to publish \concepts and methodology" as a separate paper from the results/applications, their \concepts and methodology" part usually presents a new theory or model, followed by the application of that theory/model in a second paper. Here the \concepts and methodology" only reviews existing models/methods, hence it is difficult to justify a separate paper.

Response: We already addressed this issue and how we will tackle this in light of the suggestion of Referee #1.

2. The goal of this paper is to be able to recommend to readers that for a given type of hydrological dataset, modeling method A outperforms method B, etc. Unfortunately this type of recommendation is not really convincing since some of the complicated nonlinear models have many parameters, solution approaches etc., so that scientist X using a different setup of method A than scientist Y may end up concluding method B is better than A. In particular, the single ANN (neural network) method as used in this paper is certainly suboptimal. The nonlinear ANN is known to suffer from multiple minima, so one needs to use an ensemble average of say 30 ANN models (trained using

random initial weights) to overcome this problem. See Bishop's 1995 book (Neural Networks for Pattern Recognition, Sect.9.6). For all the datasets I have worked with, the ensemble model forecasts have always outperformed forecasts from a single ANN, which gives the erratic behavior mentioned on p.7109, line 3. Even for the SVM method (which does not have to perform nonlinear optimization over a large number of weights as in ANN), how one searches for the 3 hyperparameters can lead to different skills, e.g. using a finer grid search for the hyperparameters often improves the SVM model. The poor performance of SVM mentioned on p.7103, line 24 suggests that suboptimal hyperparameters were found. Hence my point is that it is difficult to have confidence in the recommendation of \method A outperforms method B" when it is not clear how well the methods have been set up.

Response: We agree with the referee that with such nonlinear techniques, it is very difficult to say that method A outperforms method B. That is why we never stated this as an objective. What we said on Page 7058 that we want to evaluate and test the predictive abilities of these techniques that all are treated similarly. We never argued regarding individual models vs. ensemble. If ensemble is to be tested, it has to be done for all techniques. Only when many other modeling techniques, other optimization algorithms, other modeling strategies (e.g., individual vs. ensemble), and other input configuration and pre-processing are tested by other researchers, we, collectively as a community, MIGHT be able to say whether method A outperform method B or not. And this is really the goal of our work. By the way, using ensemble overcomes some of the erratic behavior….well, of course, just because you smooth out such erratic behavior but this does not eliminate the fact that the technique may misbehave sometimes.
A recommendation to use ensembles is added to recommendations.

3. The paper uses an excessive number of tables (16 tables in total), many of them poorly designed, and many with trivial information content. Take Table 7 for example: column 3 (RMSE ave) has one significant figure, so the numbers are either 0.04 and 0.05. Maybe the range of values is really 0.0449 - 0.0450 instead of 0.4-0.5, but how can one tell with only one significant figure? Column 9 (MB ave) is even more strange, some numbers are given with 3 decimals while others are with 2 decimals. Look at the bottom two entries { is 0.00 larger or smaller than 0.001? How can one tell when \0.00" could be 0.004 or 0.0001? In Table 9, column 6 (MARE ave) all models tied with exactly the same value \0.04", thanks to the single significant figure used. Then there are tables with trivial information content { see Table 12. Tables 6, 8, 10, 14 also have minimal information content.

Response: Agreed. This is addressed in the revised manuscript. We will make all numbers consistent with regard to significant figures.

4. I am not familiar with the Kolmogorov-Smirnov (KS) test, but the fact that it finds highly significant results while two probability distributions are visually indistinguishable (p.7107, first paragraph) leads me to suspect it was not done correctly. When a statistical test is applied and the results show incredibly highly level of significance, it is usually the result of overestimating the independent degrees of freedom

in the data. The authors did not discuss how the independent degrees of freedom were estimated in their KS test given that the data must have autocorrelation. Also the 12 cases were from resampling from the same dataset.

Response: We understand the concerns of the reviewer. Indeed the manuscript was not fully clear on this issue. What we did is this: the distributions were fitted and it was found they were all Logistic and were inspected visually (Fig 7-9). Then, K-S test was applied to the *raw* data, and in some cases it has shown that distributions are different. We believe that we have adhered to the procedures of the KS test, and indeed its results may differ from what the visual inspection of the fitted distributions (not real data!) may suggest. It was done on the model residuals and those were really highly independent; there was no autocorrelation in the model residuals. This is now better explained in the manuscript.

5. In short, I think the authors tried to cover too many methods, too many datasets, too many statistical tests, tables etc., and ended up spreading themselves too thin (e.g. important details not given, lack of in-depth discussion of tables, figures etc.) The revision should aim at reducing quantity and improving quality.

Response: Indeed, we tried many methods and several datasets – this was a deliberate intention in this study. Hopefully by addressing all previous comments of the three referees, our both papers will read better.

Minor comments
1. p.7097, line 12: Spell out the acronym AMI and give reference.
2. p.7103, line 7: MARE is \standardized" by observed values, hence not affected by the narrow range of the dataset.
p.7103, line 8: The argument that the R statistic is the most important indicator is not convincing.
3. p.7103, line 21: SVM with radial basis function kernel can also handle highly nonlinear data.
4. p.7104, line 21. I don't follow the sentence \... no correlations found among the probability distributions...". How does one calculate the correlation of probability distributions?
5. Table 4. The table caption should define the acronyms MARE, MB and R, etc.
6. Many figures have tiny fonts.
7. Fig.3. Define Tr, Va, Te in legend. Why is the magnitude of y1 greater than that of y2? And that of y3 greater than that of y4? I expect greater difference between Te and Tr than between Tr and Va.
8. Fig.5. Curves have wrong labels? K-nn and M5 are supposed to have overlapping curves, not SVM and M5.

Response:
1: Done.
2: the fact that MARE is not working well on the Soil moisture case, is explained in section 3.2, and the reference was also provided.

3: Indeed, this is true.
4: Indeed, the sentence was unclear and was removed.
5, 6: Agreed, and is addressed. At production process, if asked, we will update the fonts.
7: Agreed, addressed.
8: Fig 5 correctly represents the fitted distributions. K-S test (done on raw data!) has shown slight difference between kNN and M5.