Response to Referee #2

General Comments

R2_1: This paper is a case study. There is no new methodological development presented. The paper is well written, with the case study thoroughly presented, and the application of the AIC criterion is well discussed.

The description of the study site was necessary to give an insight into the investigated field site, the type and quality of data available, and the complexity of the numerical model set-up. The description of the study site was also useful to present an approach to convert rough and random distributed sedimentological borehole data and lithological information into hydraulic information, e.g., hydrostratigraphic layer and hydraulic parameter, suitable for numerical investigations.

The study site gives an example of an aquifer system that is described by an extensive data set of piezometric pressure heads collected over 20 years, however only spare information about boundary and initial conditions are available. Such a data base available for the model calibration could suggest using a complex numerical approach. We could demonstrate the usefulness of using the AIC, BIC, and KIC to select the optimal and true model concept for a groundwater model that could be set-up in a similar kind within many regions of the world.

It was not the aim of our research to develop new methodologies, but to apply the AIC to a field-generated and sophisticated data set used for groundwater management strategies. Similar field conditions and data types are not investigated in the current available literature (e.g, *Foglia et al., 2007; Hill, 2006; Hill and Tiedeman, 2007; Katumba et al., 2008; Parker et al., 2010; Poeter and Anderson, 2005; Singh et al., 2010; and Ye et al., 2010).*

We clarified the aim of our investigations in the introduction:

> *"The application of the AIC is relatively new in groundwater modeling and still not standard, although it has been applied in several studies (e.g., Foglia et al., 2007; Hill, 2006; Hill and Tiedeman, 2007; Katumba et al., 2008; Parker et al., 2010; Poeter and Anderson, 2005; Singh et al., 2010; and Ye et al., 2010). Foglia et al. (2007) uses piezometric pressure heads and stream flow gauges for a groundwater model with a huge area of that were monitored over some month and calibrates the hydraulic conductivity. Poeter and Anderson (2005) analyzed synthetic data sets,. Katumba et al. (2008) investigates the likelihood of models of tank experiments, and Parker et al. (2010) analyzes two impeller flow loggings. Singh et al. (2010) and Ye et al. (2010) compared the model uncertainty with respect to the estimated recharge for the Yucca Mountain nuclear waste repository that is well documented over decades of years. In this study, a typical field-generated data set, as often available for numerical investigations for groundwater management issues was investigated. The data set suffers from a lack of information on boundary and initial conditions, however, observation data were collected in great quantities and over a long-term. Information criteria, such as the AIC, might be helpful to define the best model concept with respect to the model performance and uncertainty."*

Foglia, L., Mehl, S.W., Hill, M.C., Perona, P., Burlando, P. 2007. Testing Alternative Ground Water Models Using Cross-Validation and Other Methods, Ground Water, 45(5): 627-641.

Hill, M. C. 2006. The practical use of simplicity in developing ground water models. Ground Water, 44(6):775-81.

Hill, M.C., Tiedeman, C.R. 2007. Effective groundwater model calibration - with analysis of data, sensitivities, predictions, and uncertainty. Hoboken, USA: John Wiley & Sons. ISBN: 978-0-471-77636-9.

Kashyap, R. L. 1982. Optimal choice of AR and MA parts in autoregressive moving average models, IEEE Trans. Pattern Anal. Machine Intell., 4(2), 99–104.

Parker, A. H., West, L. J., Odling, N. E. and Bown, R. T. 2010. A Forward Modeling Approach for Interpreting Impeller Flow Logs. Ground Water, 48: 79–91.

Poeter, E.P., Anderson, D.R. 2005. Multimodel ranking and inference in ground water modeling. Ground Water, 43(4):597-605.

Singh, A., Mishra, S., Ruskauff, R. 2010. Model averaging techniques for quantifying conceptual model uncertainty. Ground Water, 48 (5): 701-715.

Ye, M., Pohlmann, K. F., Chapman, J. B., Pohll, G. M. and Reeves, D. M. 2010. A Model-Averaging Method for Assessing Groundwater Conceptual Model Uncertainty. Ground Water, 48: 716–728.

R2_2: However, the paper would benefit from an English language edit, see a few selected examples below.

We improved the English language, rephrased many parts, and followed your examples to change the sentences.

R2_3: The paper would be strengthened by comparison with the use of other criterion used to weigh up relative model worth, e.g. the similarly constructed BIC, KIC, AICc. Similarly, methods which explore the mathematical limits of parameterisation parsimony could be used to provide a contrast in the analysis, e.g. using methods such as singular value decomposition of the model normal matrix, or as encapsulated in the predictive uncertainty analysis predunc/predvar methods outlined in Doherty (2012). Alternatively, Bayesian model averaging would provide a comparison.

We compared the results of those further criterions, AICc, BIC and KIC that are similarly constructed as the AIC to assess the model worth. For a highly parameterized inversions linked with an uncertainty analysis methods as the “predvar” analysis (Doherty, 2012) might give reasonable results as the AIC, AICc, BIC, KIC ignore the fact that the world is a complex place and that the “maximum likelihood” cannot be computed without recognition of this complexity. The PREDVAR methodology accounts for the loss of system details incurred by using too few parameters, but also for the observation noise that occurs by using too many. Such methodologies consider that the calibration process can normally capture very little of the true complexity of real-world systems. However, in our uncertainty analysis the optimal model obtained using AIC is a “medium complex model” linked with an optimal data fit. High uncertainty was introduced into our model by the boundary and initial conditions and model parameter that are all based on estimates and not on hydraulic field investigations. Therefore, applying the PREDVAR methodology and thus selecting a more complex model than it was chosen by AIC (with more than 15 adjustable model parameters) is not useful to assess the optimal groundwater model at our study site. Thus, we followed researcher (e.g., Poeter and Anderson, 2005; Burnham and Anderson, 2004) that recommend using Kullback-Leibler (K-L) information loss based criteria, such as AIC, AICc, but compared the obtained results with model selections using the BIC and KIC concept. Additionally, with respect to the results obtained by the sensitivity analysis increasing adjustable parameters will result in a sensitivity loss of the observation data which cannot be recommended.

Burnham, K.P., Anderson, D.R. 2004. Multi-model inference: Understanding AIC and BIC model selection. Sociological Methods and Research 33, no. 2: 261–304.

Doherty, J. 2012. Addendum to the PEST manual. Watermark Numerical Computing: Brisbane, Australia.
Poeter, E.P., Anderson, D.R. 2005. Multimodel ranking and inference in ground water modeling. Ground Water, 43(4):597-605.

We added the comparison of AIC with the alternative model selection criteria AICc, BIC and KIC into the abstract:

> *"Residuals, sensitivities, the Akaike Information Criterion (AIC and AICc), Bayesian Information Criterion (BIC), and Kashyap's Information Criterion (KIC) were calculated for a set of seven inverse calibrated models with increasing complexity by gradually rising the number of adjustable model parameters."*

and also:

> *"BIC and KIC selected a simpler model than the model chosen by AIC as optimal. Computing of AIC, BIC, and KIC yielded the most important information to assess the model likelihood."*

We rephrased the title of the methods exploring AIC, AICc, BIC, and KIC into:

> **"Principles to Weigh and Rank Models using AIC, AICc, BIC, and KIC"**

We added a description of the other information criteria (AICc, BIC, KIC) into the methods section:

> *"Several modifications of AIC have been developed. For the case of having a small sample, n/K<40, Burnham and Anderson (2002) suggest using AIC_c:*
>
> $$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \qquad (7)$$
>
> *where AIC is the Akaike Information Criterion as defined by Eq. 4, and K is the number of estimable parameters.*
>
> *AICc tends to AIC when the number of observations is high relative to the number of calibrated parameters as in our study, where n/K equals 5,081/30 giving 169.*
>
> *Further modifications of the AIC were also computed to provide a contrast analysis to the results obtained by the AIC. The BIC (Bayesian Information Criterion) gives a response to the concern that AIC sometimes promotes the use of more parameters than required (Hill and Tidemann, 2007). The BIC is calculated with (Doherty, 2012):*
>
> $$BIC = n\ \ln\left(\hat{\sigma}^2\right) + p\ln(n) \qquad (8)$$
>
> *The KIC (Kashyap's Information Criterion) additionally considers the likelihood of the parameter estimates in light of their prior values and contains a Fisher information matrix term that imbues it with model selection properties not used by AIC, AICc, or BIC. KIC weights and ranks alternative models in the light of the models' predictive performance under cross validation with real hydrologic data (Ye et al., 2008). KIC was derived in the Bayesian context by Kashyap (1982) and is calculated with (Doherty, 2012):*
>
> $$KIC = \left(n - \left(p-1\right)\right)\ln\left(\hat{\sigma}^2\right) - \left(k-1\right)\ln\left(2\pi\right) + \ln\left|\mathbf{J}^t\mathbf{Q}\mathbf{J}\right| \qquad (9)$$
>
> *All models were calibrated to the same data set of piezometric pressure heads, and the model with the smallest information criterion is regarded as the optimal one of all proposed models as selected by AIC, AICc, BIC, and KIC, respectively."*

Burnham, K.P., Anderson, D.R. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer-Verlag. ISBN 0-387-95364-7

Doherty, J. 2012. Addendum to the PEST manual. Watermark Numerical Computing: Brisbane, Australia.

Hill, M.C., Tiedeman, C.R. 2007. Effective groundwater model calibration - with analysis of data, sensitivities, predictions, and uncertainty. Hoboken, USA: John Wiley & Sons. ISBN: 978-0-471-77636-9.

Kashyap, R. L. 1982. Optimal choice of AR and MA parts in autoregressive moving average models, IEEE Trans. Pattern Anal. Machine Intell., 4(2), 99–104.

Ye, M., Meyer, P.D., Neuman, S.P. 2008. On model selection criteria in multimodel analysis. Water Resources Research, 44(3): W03428.

We rephrased the title of the results section dealing with the ranking of the models into:

"Comparative Results of the Model Selection Criteria"

We illustrated the results of the four information criteria in Figure 7 and Tab. 2, and compared their results:

"Both, AIC and AICc assess the similar model as optimal. The lowest AIC and AICc value is achieved by Model 4 with 15 adjustable parameters. The selection of AIC and AICc mirrors the trend of the model fit that improved distinctively between Model 1 and Model 4, and stagnated with more than 15 adjustable model parameters. Model 2 (5 adjustable parameters) and Model 7 (30 adjustable parameters) were assessed similarly poor due to a lack of model fit to the data (Model 2) or an unjustified complexity (Model 7)."

"All information criteria (AIC, AICc, BIC, and KIC) selected Model 1 (uncalibrated model based on sedimentological information) as worst model (highest information criteria). However, differences occurred in the selection of the optimal model and model ranking (Fig. 7).

Fig. 7: AIC (diamond), AICc (square), BIC (triangle), KIC (circle) assessment of the calibrated models with respect to complexity and model fit.

Tab 2: Differences Δ_j of the AIC, BIC and KIC values to the optimal model, respectively, and likelihood of the flow models from the Akaike weights (AIC w_j).

The BIC assesses a very simple model, Model 2 with 5 adjustable parameters, as the optimal model and Model 7 (30 adjustable parameters) as unfeasible. BIC values of the different models are varying more pronounced than AIC values differ (Tab. 2). The KIC evaluates Model 3 (10 adjustable parameters) as optimal model and also Model 7 (30 adjustable parameters) as worst model. BIC and KIC choose as best model approaches with fewer adjustable parameters as they assume that in the true model still the prior information exist (Burnham and Anderson, 2004). Thus, they select for greater certainty, which threatens to capture a precise, but less accurate answer than that selected by AIC. Also due to a decreasing sensitivity of the observation data with increasing parameter freedom, Model 3, as selected by KIC, might still provide a valuable model concept with a reasonable precise match of the observation data. Finally, all selection criteria argue against increasing the model complexity to more than 15 adjustable parameters."

We added results of the comparison of the information criteria into the conclusions:

"*Computing the AIC, AICc, BIC, and KIC allowed the evaluation of the benefit adjusting high*

numbers of model parameters. The simplest model based on sedimentological information as well as the complex models were rejected by all information criteria since they are likely to be under- or overparameterized."

"Differences prevail in the choice of the optimal model. AIC selects as best model a model of "medium complexity". It adjusted five of ten storage coefficients and all ten horizontal conductivities, while keeping the vertical conductivities tied by one order of magnitude lower. The results of the optimal model selected by the AIC approximately resemble observed hydraulic piezometric heads, while keeping estimated model parameters at a minimum. The AIC was able to maintain parsimony and makes predictions with a reasonable uncertainty. KIC and BIC give preference to simpler models increasing the model certainty and to maintain prior information. The optimal models selected by BIC and KIC adjusted only five or ten hydraulic conductivities, respectively, while storage coefficients are kept as deduced from the sedimentological investigations. The model fit is unacceptable in the optimal model selected by BIC. The KIC might be able to select the optimal model for an aquifer system that is described by more precise and well-known field data about model parameter than they were available at our study site. However, in situations with poor information about model parameter and boundary conditions the AIC selection should be given preference as it chooses a parsimony model, but with a sufficient freedom to receive an acceptable model fit. The choice made by AIC reflects the data available for calibration better than the optimal models chosen by the KIC and BIC. In our case, where extensive observation data were available, computing the AIC, and eventually the KIC, can improve model confidence, as it avoids an under- or overparameterization of conceptual models for a given data set. However, to decide between the optimal model selected by the AIC and KIC, respectively, the modeler still needs an overview about the data types converted to boundary and initial conditions and model parameters, which is disregarded in the model ranking by all information criteria."

R2_4: Finally, there was a brief note of the presence of bias for the very simple sedimentological estimate of parameters, but no exploration of the model bias was provided in the paper. An exploration of the relationship between bias and degrees of parsimony would be an interesting way to strengthen the paper (e.g. using methods discussed in Doherty and Christensen 2012).

We applied the methodology as discussed in Doherty and Christensen (2012) to compare the simple model, based on sedimentological information, with the optimal models selected by the AIC. We introduced the method in section 2.3

"Finally, using the paired model methodology (Doherty and Christensen, 2012) the benefit of a more complex model associated with good calibration results versus a simple model yielding a higher certainty is assessed. Simulation results of both models are given against each other in a scatter plot. Coefficients (intercept and slope) of the regression line allow analyzing the bias of the simple versus the results obtained by the optimal and more complex model with a higher degree of freedom and uncertainty."

We added the results obtained from the paired model methodology into chapter 3.2.

"The model based solely on sedimentological information is assessed by all information criteria as worst model. The bias and worth of this simple model can be explored in detail with the paired model methodology as given in Doherty and Christensen (2012). The model output of the simple uncalibrated model is compared against the results of the optimal model selected by the AIC (Fig. 8). The regression coefficients (intercept and slope) of the line

through the scatter plot allow addressing effects of simplification on the model predictions. The intercept differs distinctively from zero indicating the null space contribution of the parameter matrix to the prediction error and thus that the simple model possess consistent an error into the predictions (Doherty and Christensen, 2012). The slope of the scatter line is near 1. Hence, parameter surrogacy does not affect the uncalibrated model's ability to predict the piezometric pressure heads. The correlation coefficient of 0.99 indicated that the model based on sedimentological information might give already reasonable results. However, due its null space contribution to the prediction error the uncalibrated model based on sedimentological information can be excluded to provide already a true model."

Doherty, J., Christensen, S. 2012: Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resources Research. 47, W12534.

We added a new figure (Figure 8) to illustrate the obtained results from the paired model methodology.

"Fig 8: Paired model analysis: predicted piezometric pressure heads of Model 1 (based on sedimentological information) versus the results of the optimal model selected by AIC (Model 4), regression line equation, and correlation coefficient (R^2)."

We added the results obtained from the paired methodology into the conclusion:

"The simplest model based on sedimentological information as well as the complex models were rejected by all information criteria since they are likely to be under- or overparameterized. The paired model methodology also displays the high bias possessed by the simple model into the model predictions."

Specific comments

R2_5: Page 9691, line 19. What is Mio?

We changed the numbering to:

*"…from 560,000 m^3/a (1995) to 1.4*106 m^3/a in 2000."*

R2_6: Page 9699, line 24, 'calibration errors' change to 'residuals' for consistency in terminology.

In Tab. 3 we gave the standard error obtained by the different models, we clarified this:

"Calibration results obtained for observation wells located near the river Main (group 6) showed the highest standard error of the residual with up to 1.34 that might result from the interpolation of the river stage within the model domain."

R2_6: Page 9689, line 15, 'In addition, AIC allows to rank the models' Change to 'In addition, AIC allows the ranking of models

We changed this into:

"In addition, AIC allows the ranking of models and…"

R2_7: Page 9691, line 18. '...was rebuild' replace with ' change to '...was rebuilt'.

We changed this into:

> *"About 100 years later, the water works was rebuilt and…"*

R2_8: Page 9696, line 11, the less plausible it is to be the best one.' Change to 'the less likely it is to be the best one.'

We changed this into:

> *"The larger the AIC difference of a model, the less likely it is to be the best one."*

R2_9: Page 9700, line 23, 'Computing the AIC allowed to evaluate the benefit of adjusting high numbers of model parameters.' change to 'Computing the AIC allowed the evaluation of '

We changed this into:

> *"Computing the AIC, AICc, BIC, and KIC allowed the evaluation of the benefit adjusting high numbers of model parameters."*