Response to Referee #1

General Comments

This paper applies the Akaike Information Criterion (AIC) to rank alternative conceptual models of transient groundwater flow in graben-filling Quaternary sediments near Frankfurt. The alternative models differ in terms of how many hydraulic conductivity and storage parameters are independently estimated during model calibration using the inverse modeling software PEST. Calibration observations included 5081 hydraulic heads over a 20 year period. AIC is one of several criteria that have been developed for ranking alternative models.

These criteria favor models with a better fit to the calibration observations, and at the same time penalize models with a larger number of parameters. They tend to highly rank models that have enough complexity to fit the observations well, yet do not overparameterize the representation of model attributes such as system properties and boundary conditions. In addition to model ranking, these criteria have been used in the past decade or so for model averaging, in which they are the basis for calculating the likelihood or probability of each model under consideration.

This paper is a case study. It does not include development of new methods or concepts. It is a straightforward application of AIC to the alternative models of the Quaternary aquifer near Frankfurt. The paper first describes the field site, the flow model development, and the calibration approach. Then, the AIC criterion and calculation of AIC weights are presented and alternative models 1 through 7 are described, which increase in complexity with model number. Lastly, the results are presented, including sensitivity analyses, AIC model ranks, residuals, and parameter estimates. The results show that model number 4 is the optimal model according to AIC; of the seven models, model 4 is exactly in the middle in terms of the degree of complexity. There are several recent papers in the literature that apply the AIC criterion to models of field groundwater systems (e.g. Folia et al. 2007, Kazumba et al. 2008, Parker et al. 2010, Singh et al. 2010, Ye et al. 2010). Most of these papers apply many different criteria for multi-model comparison, including BIC, AICc, KIC, and others, and evaluate which performs the best in terms of selecting a preferable model. Given the content of these previous papers, I do not believe that the paper under review is a significant original contribution to the hydrologic literature that will be of great interest. The previous papers explore the application of AIC more thoroughly in terms of comparing it to several other model ranking criteria.

This description of the study site was necessary to give an insight into the investigated field site and the numerical model set-up. Compared to the literature cited above our investigations apply the AIC to a field site that contains an extensive data set of piezometric pressure heads collected over 20 years at 47 observation wells. These data base would suggest using a more complex numerical approach than it was later selected by the information criteria.

Important difference exist between the available literature and our research aim. Foglia et al. (2007) uses 32 hydraulic heads and six streamflow gauges for a groundwater model of a huge area of 26 km^2 that were monitored over a short time. This research includes in the parameter estimation process one parameter type, the hydraulic conductivity. With respect to the model set-up, and already without computing information criteria, it is obvious that a complex model is not justified in this case. Poeter and Anderson (2005) analyzed synthetic data sets that are totally different to observation data and boundary conditions obtained from field investigations with respect to the data density and certainty. Katumba et al. (2008) investigated the likelihood of models simulating two tank experiments with an area of 1,000m^2 and 800m^2, which is totally different than developing a groundwater model. Parker et al. (2010) analyzes two impeller flow loggings and thus calibration data and information about boundary and initial conditions are very different to a numerical model set up of a multi-layer aquifer system. Singh et al. (2010) and Yeh et al. (2010) compare model uncertainty with respect to estimated recharge for the Yucca Mountain nuclear waste repository. The Yucca Mountain nuclear waste repository is well investigated over decades of years and boundary conditions as well as observations data sets are clearly documented.

In contrast to the available literature, our study area suffers from a lack of information about boundary and initial conditions, but observation data were collected in great quantities. Our study investigates a groundwater model of i) a complex multi-layer aquifer system (in contrast to Katumba et al. (2008) and Parker et al. (2010)), ii) within a region containing great amounts of field data available for model calibration that could suggest the choice of a quite more complex model (in contrast to Foglia et al. (2007) and Poeter and Anderson (2005)), and thus, iii) several model parameters were included into the uncertainty analysis (in contrast to Foglia et al. 2007), however iv) boundary and initial conditions are only based on rough estimates introducing a high uncertainty into the model concept (in contrast to Singh et al. (2010) and Ye et al. (2010)).

It was not the aim of our research to develop new equations to assess the likelihood of numerical models, but to apply different information criteria to a typical aquifer system as it is developed in several regions of the world and described by observation data in great amount, but suffering from a lack of information about boundary and initial conditions which could lead easily to the wrong model concept. To clarify this we rephrased the goal of our research 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) compare 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.

Specific Comments

R1_1: p. 9689. lines 17-20. The statement beginning "It identifies. . ." seems to state that calculating the AIC for a single model enables one to assess whether that model is satisfactory or, in contrast, needs more complexity introduced. The AIC for a single model cannot do this. The AIC is useful only when it can be calculated for a set of calibrated models with varying levels of complexity represented by varying numbers of model parameters.

We clarified the application of the AIC:

> *"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"*

R1_2: p. 9689, lines 21-23. AIC has been applied in many more groundwater modeling studies than those listed here. The General Comment lists some of these studies, additional studies can be found by searching for AIC on the web sites for Water Resources Research, Ground Water, and the Journal of Hydrology, for example.

We added into the introduction much more information about the aim of our research as given in the response to the general comments of reviewer #1 and added some more literature reference of further studies that also investigate the AIC.

R1_3: p. 9691, line 19: It is not clear what "Mio" means.

We changed the numbering to:

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

R1_4: p. 9694, line 6: Regarding the statement that leakage was adjusted manually, do you mean that the hydraulic conductivity or conductance of the river bed and of Jacobi Pond bottom sediments were adjusted manually?

We clarified which parameters were prescribed, and which were adjusted manually in advance for the simulation of the leakage:

> *"Leakage between groundwater and surface water is driven by the gradient between the surface water stage and the groundwater, and the conductivity of the river bed and Jacobi Pond bottom sediments. The stage of the surface water was prescribed during the simulations, while the hydraulic conductivities of the river bed and Jacobi Pond sediments were adjusted in an initial manual "pre-calibration"."*

R1_5: p. 9694, Section 2.2.4. Please state the total number of observations. It is given later (on p. 9699 line 5), but needs to be stated here. There is an alternative AIC measure, called AICc, that is recommended to be used when (n/p)<40 (Poeter and Anderson 2005). For your most complex model, (n/p)=5081/30 =169. Stating the value of n will make it clear why you use AIC instead of AICc. Also, please briefly state how the observations were weighted.

We stated the number of observation points in section 2.2.4:

> *"Piezometric heads collected at 41 observation wells between 1990 and 2009 were used for the model calibration giving a total amount of 5,081 observation points (Fig. 5)."*

We introduced AIC_c as an alternative measure if the number of observation points is low:

> "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:
>
> $$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \tag{7}$$
>
> *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."*

We added the results obtained by AICc also into the section describing the model ranking and included values of AICc into Figure 7:

> *"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."*

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

We added some information how the observations can be weighted if the relevant information are available:

> *"The two middle terms are constants for a specific data set, and are not affected if parameters are added or removed from the models (Cavanaugh, 1997). Weights were set to one since no information about data uncertainty and measurement error was available. However, when additional information about confidence of the data is available the weight matrix of Eq. 4 allows comparing models based on a weighted data set of observations. This reflects the confidence to specific measurements, or simply, provides the flexibility to scale observations according to additional information or normalization procedures (Hill and Tiedeman, 2002)."*

Cavanaugh, J. E. 1997. Unifying the derivations of the Akaike and corrected Akaike information criteria. Statistics & Probability Letters, 33: 201-208.

Hill, M. C., Tiedeman, C. R. 2002. Weighting observations in the context of calibrating groundwater models. IAHS-AISH Publication; 277: 196-203.

R1_6: p. 9695, lines 7-9: The concept of whether a model can be validated is a controversial subject among groundwater modelers. I suggest that the authors consider the article by Bredehoeft and Konikow (2012) before using the term validation to describe the exercise of assessing model fit to data excluded from the calibration.

We added a link to the publication of Bredehoeft and Konikow (2012) in the section were the validation is discussed:

> *"This procedure was chosen due to the analysis of Bredehoeft and Konikow (2012). They emphasize that a professional judgment of the model is only possible using historical data, while the validation of the model against future response remains challenging. However, errors resulting from conceptual errors will neither be addressed by using historical nor future data in the validation (Bredehoeft and Konikow, 2012)"*

Bredehoeft, J. D., Konikow, L. F. 2012. Ground-Water Models: Validate or Invalidate. Ground Water, 50 (4): 493–495.

R1_7: p. 9695, equation (2): The commonly used expression for AIC applied to models calibrated by least squares regression omits the middle two terms of equation 2 (e.g., Poeter and Hill 2007, Singh et al. 2010). Burnham and Anderson (2002, p. 63), which is cited for equation 2, also present the equation for AIC without the middle two terms for the least squares case. Please explain why you use the form of AIC in equation 2, and please cite a reference for this form of AIC.

We corrected Eq. 2 and added the term for considering the weights of the observation data as given in Ye et al. (2008).

> "AIC is defined as follows (Ye et al., 2008):
>
> $$AIC = n\ln(\hat{\sigma}_{ML}{}^{2}) + n\ln(2\pi) + n + \ln|Q^{-1}| + 2p \qquad (4)$$
>
> *where p equals the number of estimated model parameters plus one, n is the number of observations, Q is the weight matrix, and $\hat{\sigma}^{2}_{ML}$ represents an estimate of the variance of weighted residuals, …."*

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

R1_8: p. 9695, equation (3). This expression needs to include the observation weights. Using the definition of weights in PEST, the squared term in the summation needs to be the product of the weight and the residual, not just the residual.

We added the observation weight to Equation 3.

> *"…and $\hat{\sigma}^{2}_{ML}$ represents an estimate of the variance of weighted residuals, which is given by:*
>
> $$\hat{\sigma}^{2}_{ML} = \frac{\sum_{j=1}^{n} (\varepsilon\, q)_i^{\,2}}{n} \qquad (5)$$
>
> *where ε stands for the residuals (observed minus calculated values), and q is the weight of the j^th observation, respectively, which is always one for the present study."*

R1_9: p. 9697, Section 3.1., General: Please provide the equation for the sensitivity coefficients plotted in Fig. 6. The measure plotted appears to be an overall measure of the sensitivity of all observations in a group to all parameters of the model. In addition, the measure appears to be normalized to the largest sensitivity in the model with the fewest parameters. For readers to evaluate the results in Fig. 6, the equation is needed.

We added into the methods (chapter 2.2) the necessary equations for computing the composite scaled sensitivities:

> *"In PEST the composite sensitivity s_j of a parameter i is computed with (Doherty, 2010):*
>
> $$s_i = \left(\mathbf{J}^{\mathbf{t}}\mathbf{QJ}\right)_{ii}^{1/2} / m \qquad (2)$$
>
> *where **J** is the Jacobian matrix**,** **Q** is the weight matrix, **J**$^{\mathbf{t}}$**QJ** is the normal matrix, and m is the number of observations with non-zero weights.*
>
> *The composite observation sensitivity s_j of observation j is computed in PEST with (Doherty, 2010):*
>
> $$s_j = \left\{\mathbf{Q}\left(\mathbf{J}^{\mathbf{t}}\mathbf{J}\right)\right\}_{jj}^{1/2} / n \qquad (3)$$
>
> *where **J**^{t}**J** is the Hessian matrix, j is the counter of the observations, and n is the number of adjustable parameters."*

Doherty, J. 2010. PEST - Model-Independent Parameter Estimation. User's Manual, 5th ed.; Watermark Numerical Computing: Brisbane, Australia.

We also clarified in this section the approach illustrated in figure 6:

> *"Composite observation sensitivity s_j were computed for each observation point to be an overall measure of the sensitivity of all 5,081 observation points to all adjustable parameter in the model, respectively.*"

Sensitivities were not normalized to the largest sensitivity in the model with the fewest parameters.

R1_10: p. 9697, line 11: What is the "colmation" layer? Does this refer to the river and pond bed sediments?

We clarified this:

> *"…. leakage through the colmation layer of the river and pond bed sediments."*

R1_11: p. 9699, Section 3.3.1 and Table 3: Weighted residuals are dimensionless quantities, so the standard error of the weighted residuals is dimensionless rather than having units of meters as indicated in the manuscript.

We deleted the units that were given for the standard error of the weighted residuals:

> *"The smallest standard error of the weighted residuals was obtained with 0.22 to 0.23 near the Jacobi pond.*
>
> *….computed standard errors of the weighted residuals increased to 0.47 to 0.51, which can still be assessed as sufficient…*
>
> *…showed the highest calibration standard error with up to 1.34 that might result…*
>
> *…standard error of weighted residuals improved from 1.18 (Model using only sedimentological information) to 0.74 (Model 7)."*

R1_12: p. 9699, line 26: The model validation data set is mentioned on lines 26-27, and is followed by the general statement about model fit on lines 27-29. That statement refers to Table 3, which summarizes the fit for all observations, those included in the calibration and those used just for validation. If you retain the idea of model validation in the paper, it would be best to show the fit to the calibration and validation observations separately.

General statements about residuals are valid for the inverse model calibration that excluded the observation data used for the model validation. We modified table 3 and added a separate line giving the model fit obtained during the validation. We added also some explanation into the text:

> *"Groundwater levels simulated by the optimal model matched measured values at most locations reasonably well (group 1, 4, 5, and 6) and demonstrated that model parameters were estimated within a reliable range (Tab. 3). However, at two locations (group 2 and 3) the model fit was distinctively poor and similar standard error as obtained with the model based on sedimentological information."*

R1_13: p. 9700, lines 13-15: The Quaternary aquifer was described as consisting of unconsolidated sediments (section 2.1.1, Fig. 3). The explanation of possible highly permeable fractures in the aquifer is quite inconsistent with this aquifer geology.

We changed this into:

> *"These differences may result from the impact of secondary flow pathways or local heterogeneities that were missed by the interpretation of the borehole data."*

Editorial Comments

Generally, the paper is written fairly well, and is well organized. Some parts would benefit from additional careful editing to improve the English usage. Some examples are:

p. 9691, line 18, change "rebuild" to "rebuilt"

we corrected this typo

p. 9693, line 16 "in the" not "in of the"

we corrected this typo

p. 9693, line 24 change to "The main inflow into the groundwater system resulted in. . ."

we changed this into: "*The main inflow into the groundwater system is recharge…"*

p. 9694, line 5 change "prevailed" to "were"

we changed prevailed into were

p. 9695, line 9 change "observations" to "observation"

we corrected this typo

p. 9696, lines 4-5 rephrase "since it gets larger then"

we changed this into: "…. *as this term increases within rising amount of adjustable parameters*."

p. 9696, line 9 change "is denoting" to "denotes"

we changed is denoting into denotes

p. 9696, line 22 omit “amount”

we changed this into: “….*model domain giving 30 adjustable parameters.”*

p. 9697, line 18 change “amount” to “number”

we changed amount into “*number*”

p. 9697, line 20 change to “the highest . . .” Change “amount” to number”

we changed this into: “…*five parameters revealed the highest sensitivity coefficients (Fig. 5). Increasing the number of adjustable parameters…”*

p. 9697, line 22 Change “amount” to number”

we changed amount into *“number”*

p. 9698, line 2 change to “allowed evaluation of the conceptual. . .”

we changed this into: “*Computing the AIC, AICc, BIC, KIC allowed the evaluation of the best conceptual model…”*

p. 9898, line 15 change “as similar worse” to “similarly poor”

we changed this into: “…*were assessed similarly poor*….”

p. 9899, line 5 omit “sum”

we changed this into: “*…. giving a total number of 5,081 piezometric pressure head data…”*

p. 9899, lines 12, change to “and also displayed the impact of. . .”

we changed this into: “*Within the southern part groundwater levels varied up to 2.1 m and also displayed the impact of the water works Goldstein*.”

p. 9700, line 7, change to “Very limited information was available. . .”

we changed this into: “*Very limited information was available from field investigations about hydraulic conductivity and storage*.”