We thank reviewer #2 for detailed comments and suggestions for improving the original manuscript. We answer the main comments in the following sections (minor and technical comments are included in the revised manuscript):

...The weakness that I see in the technique used to evaluate the models and their relative posterior model probability is the fact that there is not penalty factor accounting for adding too many parameters. So, the models with the highest number of parameters, as expected, have the highest posterior model probability. I will make also the point in the specific comments, but I would strongly suggest to make a simple and fast test of the performances of the different models using a statistic which account for the number of parameters added (the simplest I have in mind is AIC or AICc).

We partially agree with this comment. As mentioned by the reviewer there is no (explicit) penalty term included in the likelihood function when the number of parameters increases. Since the likelihood is linked to model performance, models with more parameters usually show the highest model likelihood. A penalizing term, however, can be included in the framework of the GLUE-BMA method through the prior model probabilities. If the analyst's prior knowledge about the plausibility of the alternative conceptualizations is sound enough to define non-uniform prior model weights, the latter could be used to penalize those models including a high number of parameters. The inclusion of this prior knowledge rests on the Bayesian paradigm and it is seen as an advantage in the context of the used methodology. This knowledge should reflect the analyst's prior perception about how likely the alternative conceptualizations are. Related to this, in a recent article (Rojas et al., 2009c) we have shown that including proper prior knowledge about the alternative conceptual models will reduce the predictive variance and outperform the case when selecting uniform prior model probabilities. The question here, however, translates into how to efficiently define these prior model probabilities to reflect this penalizing factor. Answering this question, however, is beyond the scope of this article. Some guidelines can be found in Ye et al. (2008b).

Reviewer #2 suggests doing simple tests using model selection criteria (e.g. AIC or AICc) to assess model performances accounting for the number of parameters added. We must emphasize two points for answering this comment. First, it is not the aim of this article to select "the best model" out of an ensemble of model candidates, i.e. solve a model selection problem. Rather the objective is to assess the uncertainty arising from the definition of an ensemble of plausible conceptual models for the PTA, i.e. solve a predictive multimodel problem. Second, we have shown in a recent work (Rojas et al. 2009a) that using different model selection criteria to approach the posterior model weights used for multimodel aggregation will likely produce misleading and conflicting results. This is mainly due to the differences in how alternative model selection criteria (e.g. AIC, AICc, BIC, KIC) penalize model complexity, value prior information on parameter estimates, or interpret the quality of the available dataset D. Although Ye et al. (2008a) have shown clear advantages of KIC compared to the other model selection criteria, the controversy about using different selection criteria in the framework of multimodel methodologies is not settled yet.

Working with model selection criteria, to comply with the principle of parsimony, alternative conceptualizations will be ranked differently, and consequently, they will be given different posterior model probabilities. Different posterior model probabilities will lead to different estimations of conceptual model uncertainty and predictive uncertainty. In the framework of multimodel methodologies (and applications) this is critical and can not be neglected.

As an example, Table R.1 shows the results obtained in a multimodel application for the aquifer underlying the Walenbos Nature Reserve in Belgium (Rojas et al. 2009a). In this application we used 3 conceptual models (M1, M2 and M3) and 4 model selection criteria to estimate posterior model weights based on Ye et al (2008a). Conceptualizations considered an increasing number of parameters and a fixed number of observations.

Table R.1: Summary of posterior model probabilities using alternative model selection for models M1, M2, and M3

	Conceptual models		
	M1	M2	M3
Nr observations	51	51	51
Nr parameters	4	6	8
SWSR*	180.95	182.18	158.18
MLOFO**	64.95	64.93	57.73
LN |F|***	-122.75	-117.88	-102.18
p(Mk)	1/3	1/3	1/3
AIC	74.59	78.93	75.73
Rank	1	3	2
P(Mk|D)	**0.596**	**0.068**	**0.337**
AICc	75.92	81.54	80.12
Rank	1	3	2
P(Mk|D)	**0.845**	**0.051**	**0.104**
BIC	84.25	92.46	93.11
Rank	1	2	3
P(Mk|D)	**0.972**	**0.016**	**0.012**
KIC	-5.99	-6.68	-10.48
Rank	3	2	1
P(Mk|D)	**0.085**	**0.119**	**0.796**

* SWSR: Sum of weighted squared residuals.

** MLOFO: Maximum likelihood objective function observations.

*** Ln|F|: Natural log of the determinant of the Fisher matrix.

As suggested by reviewer #2 we could consider AIC and AICc. Although the ranking of models is the same, the posterior model probabilities (p(Mk|D)) are rather different for models M1 (59.6% and 84.5% for AIC and AICc, respectively). This difference resulted in predictive variances differing in one order of magnitude for key groundwater flow components (e.g. inflows to the Nature Reserve) when working with AIC- or AICc-based posterior model weights. Additionally, when using AICc-based model weights conceptual model uncertainty accounted for 16% of the total uncertainty, whereas using AIC-based model weights this contribution was 36%. If we consider BIC or KIC (compared to AIC and AICc), the ranking of models and, even more important, the posterior model weights are significantly different.

These results show that, even for the case of the simplest statistics accounting for the number of parameters added (AIC or AICc), using alternative model selection criteria to estimate posterior model probabilities might result in misleading and conflicting results in multi-modelling applications. We argued that seems more reasonable to work with model weights (used for multimodel aggregation) obtained from the sampling of the full hyperspace dimensioned by conceptual models, parameters and forcing data vectors than working with model weights obtained from model selection criteria using penalizing terms for model complexity. In the event that a penalizing term must be considered, this can be done through the definition of non-uniform prior model weights. As mentioned, however, an analysis of this kind is beyond the scope of this work.

I suggest to rework on section 5 and make the explanations of the different results easier to understand by mean of bullet list, or analysis by area, or any other way that authors feel appropriate.

Corrected. See revised manuscript.

P5885, lines 1-5. The recharge seems to be the key issue for the model representations. It will be explained later, but it would be useful to have at this point the different recharges.

Corrected. See revised manuscript.

P5885, line 7. What is the implication of neglecting that recharge? It will be presented later but it is of interest here.

Corrected. See revised manuscript.

P5886, lines 15-20. Which are exactly the observations used?

Observations used varied among the alternative models. In general, previous studies used a common set of head measurements (40-55), estimated discharges for the transpiration from the forested areas (years 1960, 1988 and 1993), and estimated discharges for the evaporation from salares (salt pans) present in the study area for years 1960, 1988, 1983.

P5887, lines 10-15. No flow observations in the calibration set? How this affects the calibration results in terms of nonuniqueness?

Since the study area is located in the Atacama's Desert, there are no flow observations. Values for some flow components (e.g. transpiration and evaporation) were obtained as models' result from previous studies. These estimations, however, are not included in this application for conditioning simulations as they are considered relatively uncertain. We agree with the reviewer that adding flow observations will reduce the problem of non-uniqueness and probably will constrain the likelihood response surfaces as demonstrated by Rojas et al. (2009b), thus, improving the performance of the GLUE-BMA approach. However, for the present application only a suite of head measurements, which may often be the only information available to perform multimodel applications, was available to assess the model performance.

P5888, lines 15-20. No penalty for too many parameters with respect to the number of observations?

See answer comment #1. As explained, a penalty term could be included in the form of non-uniform prior model probability distributions. The definition of such priors, however, is beyond the scope of this article.

P5892, lines 10-13. Explain.

Corrected. See revised manuscript.

P5898, lines 23-28. Did you run any sensitivity analysis on the selected parameters? Did you see parameter correlations problems? Also models M2 have a much bigger number of parameters and I am not sure about having a successful calibration of those models having only heads observations.

For this work we did not perform a sensitivity analysis on the selected parameters. In a previous work, Rojas and Dassargues (2007) did an extensive sensitivity analysis for a groundwater flow model equivalent to model M2 used in this application. The sensitivity analysis in that work was performed for the 22 hydraulic conductivity zones, elevation of constant head at the south boundary, evaporation rate, extinction depth for evaporation process and recharge flow rates. From that analysis, the most sensitive parameters were the recharge rates and the hydraulic conductivities, which showed some degree of correlation given the steady-state nature of the model developed. The elevation of the southern boundary condition (CH_S) showed moderate sensitivity whereas parameters related to the evaporation process were relatively insensitive.

As explained before, flow observations were not available for the calibration. There are estimations which are based on previous models' results but they are used as reference values solely. We fully agree with the reviewer that non-uniqueness of parameters and equifinality problems would be reduced with the inclusion of flow-related observations. In a recent work (Rojas et al., 2009b) we have illustrated the value of flow-related observations in the context of the GLUE-BMA methodology.

P5904, lines 19-23. This conclusion may be driven by my concern on the number of parameters. Also Table 4 shows that none of the models is really showing a significantly higher model posterior probability.

As explained, model likelihoods are linked to model performances. Therefore, better performing models will have higher integrated model likelihoods. Since we are keeping the analysis neutral by considering uniform prior model probabilities, integrated model likelihoods are proportional to the posterior model weights.

Conceptual models using RSF theory to describe the spatial distribution of the hydraulic conductivity show high integrated likelihoods. If we consider conditioned realizations of the hydraulic conductivity field, the conceptualization shows higher posterior model probabilities.

We agree with the reviewer that Table 4 does not show significant differences in posterior model probabilities. This is due to the fact that models' simulations are conditioned solely on heads and on a relatively low number of hydraulic conductivity measurements given the area of the modelled domain (62 k-measurements for an aquifer of more than 5000 km^2). In a recent work (Rojas et al., 2009b) we have demonstrated that head measurements convey low information content to improve the discrimination among alternative conceptualizations contained in an ensemble M. This discrimination is drastically improved, however, by the inclusion of flow-related measurements and a sufficiently dense measurement network of hydraulic conductivity values for conditioning the K-field. Unfortunately, for the application to the PTA information to improve the discrimination among alternative conceptualizations was not available.

P5911-5912. Most of the conclusion can be revisited after making a test with a measure which accounts for number of parameters. Overall, considering the results obtained, which model would you select for future simulation and prediction?

See answer comment #1.

We must emphasize that it is not the aim of this work to select "the best model" out of an ensemble of model candidates, i.e. solve a model selection problem. Rather the objective is to assess the uncertainty arising from the definition of an ensemble of plausible conceptual models for the PTA, i.e. solve a predictive multimodel problem. By mimicking actual conceptual models developed in previous studies we demonstrated that conceptual model uncertainty is a relevant source of uncertainty. This is critical for the PTA where human pressure for water resources is considerably high and uncertainty due to climatic conditions is relatively important.

Clearly, the most conditioned model (M4) showed slightly higher posterior model probabilities. However, neglecting the other conceptualizations will likely produce biased and under-dispersive uncertainty estimations. Working with the ensemble prediction, on the other hand, we deliberately work with a suite of potential candidates spanning plausible model realizations. Raftery and Zhang (2003) have shown that BMA outperforms predictions of any other single member of an ensemble of potential model candidates for statistical models. Similar results were obtained by Rojas et al. (2008) for a synthetic groundwater system.

References

Raftery, A. and Y. Zhang, (2003) Discussion: Performance of Bayesian model averaging, Journal of the American Statistical Association, 98(464): 931-938.

Rojas, R. and A. Dassargues, (2007) Groundwater flow modelling of the regional aquifer of the Pampa del Tamarugal, northern Chile, Hydrogeology Journal, 15(3): 537-551, doi:10.1007/s10040-006-008406.

Rojas, R., Kahunde, S., Peeters, L., Batelaan, O., Feyen, L. and A., Dassargues, (2009a) Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling, submitted to Journal of Hydrology, under review.

Rojas, R., Feyen, L., Batelaan, O. and A. Dassargues, (2009b) On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling, submitted to Water Resources Research, under review.

Rojas, R., Feyen, L. and A. Dassargues, (2009c) Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling, Hydrological Processes 23(8): 1131-1146, doi: 10.1002/hyp.7231.

Ye, M., Meyer, P. and S. Neuman, (2008a) On model selection criteria in multimodel analysis, Water Resources Research, 44, W034228, doi:10.1029/2008WR006803.

Ye, M., Pohlmann, K. and J. Chapman, (2008b) Expert elicitation of recharge model probabilities for the Death Valley regional flow system, Journal of Hydrology 354(1-4): 102-115, doi:10.1016/j.jhydrol.2008.03.001.