We thank reviewer #1 for comments and suggestions to improve our paper entitled: "Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal – North Chile" by R. Rojas et al.

We answer the main comments in the following sections:

Pag 5891 line 9-15. It appears that your posterior model weights do not account for the number of parameters of the model. If the integrated likelihood and the prior probability of two models are equal your method assigns the same posterior model weight. I believe that in this case the model with less number of parameter should be preferred in according with the principle of parsimony. Is this correct? The authors should comment on this.

We partially agree with this comment. Posterior model weights implicitly account for the number of parameters through the likelihood function. It is likely that a model with more parameters will have a higher likelihood value, as obtained from model performance. If both prior and integrated model likelihoods are equal for alternative conceptualizations, these conceptualizations will receive the same posterior model weight. The latter means that the evidence provided by the data did not support any single model compared to the rest, independently of the number of parameters for each model. We acknowledge, however, that the GLUE-BMA method does not penalize more complex models (i.e. models with a higher number of parameters) through the model likelihood. However, this can be done through the definition of non-uniform prior model probabilities. These non-uniform prior model probabilities could penalize alternative conceptual models on the basis of the number of parameters, complexity, plausibility or any other criteria followed by the analyst. An analysis of this type, however, is beyond the scope of this article.

The principle of parsimony is the general guideline for penalizing more complex models and for selecting one model over the others. Different model selection criteria using different penalizing terms could be used to select a model as the number of parameters increases. We must emphasize, however, that this work is not a model-selection exercise, rather the aim is to assess the uncertainty arising from the definition of an ensemble of plausible conceptual models for the PTA, i.e. a multimodel exercise.

In the case of uncertainty analysis for multi-modelling applications the use of model selection criteria (to comply with the principle of parsimony as suggested by the reviewer) can produce misleading and conflicting results. In a recent work, we investigated the use of different model selection criteria, namely, AIC, AICc, BIC, and KIC, in a multimodel application for the aquifer underlying the Walenbos Nature Reserve in Belgium (Rojas et al. 2009). Conceptualizations used in that work considered an increasing number of parameters. Results of that work are summarized in Table R.1.

Results from Table R.1 show that different model selection criteria assign (significantly) different posterior model probabilities. This had a drastic impact on the estimation of the predictive variance (uncertainty) (as estimated from equation (4) of the manuscript) in function of alternative model selection criteria. In the context of a multimodel analysis this is crucial and can not be neglected. On the basis of these results we argued that seems more reasonable to work with posterior model probabilities (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 posterior model weights approximated from model selection criteria using penalizing terms for model complexity. Although Ye et al., (2008) have presented an enlightening discussion about the merits and drawbacks of alternative model selection criteria in multimodel applications, the discussion about using an alternative criterion over the others is far from being settled yet. Considering this, we believe that estimations of posterior model weights based on an extensive exploration of the likelihood surface are more robust than local approximations based on model selection criteria.

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.

Conceptual model uncertainty is often neglected in groundwater modelling by working with a single conceptualization solely, which is equivalent to assign all posterior model weight to a particular model. This is usually done by using model selection criteria in the case of multiple conceptualizations, supported by a guiding principle as parsimony. In recent works we have shown, however, that neglecting conceptual model uncertainty will produce biased and under-dispersive uncertainty estimations, and that using model selection criteria to approach posterior model weights might lead to conflicting results in multi-modelling applications.

P5899 line 12. The authors describe the cell size of the models. I suggest to write explicitly if the models are three-dimensional and report the thickness of the cell in 1-layer and 2-layers models. The authors should state if they performed any sensitivity analysis to the cell size and if not show that their cell sizes are adequate for the problem at hand.

Model M1 (a and b) is three-dimensional with varying cell thicknesses defined by stratigraphic units Q3 and Q4 (see Rojas and Dassargues, 2007) whereas models M2, M3 and M4 are two-dimensional. We did not perform a sensitivity analysis on the cell size. We defined the cell size on the basis of the previous models developed for the PTA and based on pragmatic reasons to make the problem computationally tractable. For the latter, we performed preliminary runs to estimate the computation time for individual runs.

P5899 line 24. To improve the clarity of the paper I suggest to add more information about the 42 observed heads. Do the hydraulic head measurements belong to the aquifer Q3 or Q4? What are the minimum and the maximum value that you observed? On which basis do you assume that the standard deviation of observed heads is 10 m?

Head measurements are obtained for both aquifers as observation wells are screened at multiple sections. The range of head measurement values is between 915 masl and 1033 masl.

We performed a series of preliminary runs to test the implementation of the M-H algorithm. The standard deviation defined for the heads is the basis for the rejection criterion implemented in the GLUE-BMA method and, therefore, has a significant impact on the procedure to explore the sampling space using the M-H algorithm. Small standard deviation values made the algorithm excessively slow by defining a too stringent rejection criterion. We sequentially increased the value of the standard deviation from 2.5m (value obtained from Rojas et al., 2008) until we reached a trade-off between computational time and number of "behavioural" simulators in the subset A_k. This value was 10m which allowed defining the rejection threshold as 30m. Considering the dimensions of the modelled domain, and the range of observed heads (915-1033 masl), we considered this value as acceptable for the problem at hand.

P5905 line 23. Could you explain better what you mean by 'synthetic piezometers'? To improve the method reliability I suggest to leave out of the calibration process some point of the dataset D. In this way you could do a real validation instead of a pseudo-validation that you did. It would be more interesting to show how the mean of the full BMA prediction, conditioned on available information, reproduced measured values at a set of validation points, the measurements laying or not within the corresponding envelops of uncertainties.

Synthetic piezometers are defined as points where no conditioning data either head or hydraulic conductivity measurements are included in the dataset (**D**) used for conditioning the multimodel simulations. The objective of including a suite of synthetic piezometers was to assess the relevance of uncertainty arising from the alternative conceptualizations at points not included in **D** (heads + conductivity measurements). We are aware of the cross-validation approach suggested by the reviewer. However, in our manuscript we are not working with an exact conditioning technique, i.e. observed heads are not exactly reproduced since the GLUE-BMA method allows a departure based on

the rejection criterion, thus, exploring the global likelihood response surface cut-off at those thresholds. Thus, strictly speaking, the validation technique proposed by the reviewer can also be thought of as a pseudo-validation method in the context of the GLUE-BMA method. If the conditioning of heads would have been exact (i.e. if the variances at the observed points would have been zero in Table 5 of the manuscript), the cross-validation method proposed by the reviewer could be applied, however, this is not the case.

P5906 line 2. In Table 5 the authors report only the head variances at observation wells. I suggest the authors to report also the observed heads and the full BMA prediction with the corresponding envelops of uncertainties at these points. Of course it is possible to see the observed heads in Figure 7, but it is difficult to see if the observed head lies within the envelops of uncertainties of the estimated one. In Table 6 the authors consider points, that they called 'synthetic piezometers', the variance could be very small or very large but they could not know how the estimated head is close to the observed one. I suggest the authors to give a physical explanation of the results that they show here. Moreover, how the information about the variance can be used from an applicative point of view?

Corrected see revised manuscript.

The summary for the observed heads and synthetic piezometers is shown in Tables R.2 and R.3.

Table R.2: Summary of observed heads, BMA prediction and predictive variances for groundwater heads at observation wells.

	Observation well												
	162	237	263	276	281	286	290	294	C-30	315	D-60	A-13	133
Observed [m]	980.4	967.8	971.2	957.5	954.9	951.3	951.9	924.5	1131.4	993.5	997.1	1111.2	972.0
BMA prediction ± std. dev. [m]	968.1 ± 6.5	966.4 ± 4.9	964.7 ± 5.5	950.6 ± 5.6	945.7 ± 6.0	942.1 ± 6.0	937.5 ± 5.1	932.8 ± 7.9	1130.1 ± 11.6	975.8 ± 7.8	987.6 ± 9.3	1125.3 ± 14.5	976.5 ± 6.4
Total variance [m²]	41.8	23.6	30.0	31.9	36.5	36.2	26.5	62.8	134.8	60.3	89.6	211.7	41.2

Table R.2: Summary of BMA prediction and predictive variances at synthetic piezometers.

	Synthetic piezometers										
	P1	P2	P3	P4	P7	P10	P11	P12	P13	P14	P15
BMA prediction ± std. dev. [m]	1131.0 ± 13.8	1101.2 ± 18.6	1043.3 ± 17.3	1009.8 ± 14.5	981.1 ± 7.5	972.3 ± 5.5	973.2 ± 7.8	965.8 ± 8.7	943.3 ± 9.4	942.3 ± 10.1	929.7 ± 12.1
Total variance [m²]	191.0	345.9	299.7	209.7	54.6	30.7	60.9	76.5	88.6	102	146.9

The reviewer is right in pointing out that values of the variance might be very small or very large at the synthetic piezometers. As discussed above, however, the objective of defining the synthetic piezometers was to assess the contribution of conceptual model uncertainty at points not contained in the dataset **D**. Therefore, independently of the variance value, we were interested in the ratio between-model to total variance at these points.

P5912 line 3. The point number 4 is related to the point number 1. I suggest the authors to merge the two conclusions or to write them in two points which come one right after the other.

Corrected. See revised manuscript.

P5912 line 18. The authors conclude that the relevance of conceptual model uncertainty is more significant for spatial data not included as conditioning points. I believe that this statement is not a major conclusion and I would suggest not to highlight it. I think that this conclusion can not be deduced on the basis of the example shown (e.g. if you choose other 'pseudo-validation' points closest to the conditioning points it could happen that you do not appraise any difference between different conceptual models, is this correct?

We agree with this comment. The relevance of conceptual model uncertainty, however, as seen from the results is more important for variables not included as calibration targets. In general, for heads the uncertainty is relatively low compared to predictions of groundwater flows or other variables not used for calibration. The latter is in full agreement with the results of Harrar et al. (2003) and Troldborg et al. (2008).

References

Harrar, W., Sonnenberg, T. and H., Henriksen (2003) Capture zone, travel time, and solute transport predictions using inverse modelling and different geological models, Hydrogeology Journal, 11:536-548, doi:10.1007/s10040-003-0276-2.

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., Feyen, L. and A. Dassargues, (2008) Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, Water Resources Research, 44, W12418, doi:10.1029/2008WR006908.

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

Troldborg, L., Refsgaard, J., Jensen, K. and P., Engesgaard (2007) The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system, Hydrogeology Journal, 15:843-860, doi:10.1007/s10040-007-0192-y.

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