We thank reviewer #3 for comments and suggestions on our manuscript entitled “Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal – North Chile”. In the following sections we answer main comments raised by the reviewer (minor and technical comments are included in the revised version of the manuscript).

While field applications of newly developed methods are always interesting and important to further evaluate and develop the method, the field application presented in this study does not appear to provide much insight for evaluation and/or development of the GLUE-BMA-based model. In other words, this manuscript appears to be more like an application rather than for theoretical advancement.

We agree with this comment. In three previous articles (already published and under review) we have further developed theoretical aspects of the GLUE-BMA method. The relevance of prior information about the conceptualizations is discussed in Rojas et al. (2009c), the value of different conditioning data to constrain conceptual model uncertainty is presented in Rojas et al. (2009b), and a real-world application using GLUE-BMA and alternative model selection criteria to estimate posterior model probabilities (model weights) is discussed in Rojas et al. (2009a). As last step, and as pointed out by the reviewer, this work reports the application of the method to a real-world groundwater modelling exercise. Thus, no further theoretical advancement was developed in this article.

1. Objective of this study

It is unclear about the author’s purposes of conducting this real-world modelling. While the authors explained significance of conducting groundwater flow modelling for the PTA, it would be useful to explain what particular problems were tackled for the real-world application. For example, for the variables simulated by the flow models, BMA performance is different and model uncertainty behaves also in different ways. If the authors can focus on the most important variable for the PTA models and have more in-depth discussion, this manuscript would be more interesting. For example, if the outflow (Figure 8e) is the most important variables for this modelling, there may be no need to consider the alternative models.

Corrected. See revised manuscript. The main purpose of this article is to illustrate the effects of neglecting conceptual model uncertainty in a real-world modelling exercise. We demonstrated that by working with a suite of reasonable (and previously developed) conceptualizations to model the regional aquifer of the Pampa del Tamarugal, more robust uncertainty estimations can be obtained. The ultimate goal was to quantify the uncertainty arising from the definition of these alternative conceptualizations, i.e. between-model variances.

2. Effect of parameter distribution

The ranges of parameters listed in Table 3 are very large, varying in several orders of magnitude. Given that the uniform parameters distributions are assumed in this study, the deviation, between the results of this study and those of previous studies (e.g. those shown in Figures 6 and 8) might be related to the parameter distribution. On the other hand, it appears that parameter correlations are not considered in the sampling. This *may also affect the modelling results. For example, in the context of model calibration, the estimates of hydraulic conductivity and recharge are correlated. Ignoring this correlation may yield biased results.*

Parameter ranges defined in Table 3 are realistic values obtained from previous studies (DGA-UChile, 1988; JICA-DGA-PCI, 1995; DSM, 2002, Rojas and Dassargues, 2007; DICTUC, 2008). These limits are defined to guarantee a physically meaningful sampling of the parameter space. For example, parameters TRANSP, EVAP, EXTD, CH_N, CH_S and K are fully based on preliminary estimations or calibration of previous models. For parameters RECH and RECH_BAS the ranges are defined to ensure a realistic sample of potential recharge values according to the hydrologic regime of the study area

Despite some parameter ranges shown in Table 3 span several orders of magnitude, samples were efficiently selected by the Metropolis-Hastings algorithm and clear zones of attraction for them were defined. Figure 6 shows the zones of attraction for a series of model results. We doubt that the deviations between results of this study and those of previous studies (Figure 6 and 8) are related to the uniform parameter distribution used for sampling. Although we can not provide any theoretical proof, we believe that working with non-uniform priors might have a more severe impact on the deviations of previous studies and the results from the GLUE-BMA methodology. We selected uniform distributions to keep the analysis at a neutral level, i.e. combining flat priors with the information conveyed by the data through the model likelihoods.

As pointed out by the reviewer, we did not consider parameter correlations in the sampling. We acknowledge that parameter correlations may yield biased results (see revised manuscript), however, an analysis of the effects of these correlations on the GLUE-BMA results is beyond the scope of this article.

3. In the titles of Figures 6 and 8, the authors referred to the results of "previous studies", but it is unclear which previous studies the authors are referred to.

Corrected. See revised manuscript.

4. Thinned samples

The thinning may affect the calculation of mean and variance, in particular when the thinning was conducted at regular frequency, irrelevant to the statistical distribution of the data. It may be worthwhile to investigate this effect.

We tried several thinning intervals (frequencies) before reaching a compromise between size of the chains and information content. For each of these trials we recomputed the first two moments to ensure consistency of the retained samples. We acknowledge the use of regular frequencies for the thinning of the samples, however, this was done on the basis of Gilks et al. (1995) and Sorensen and Gianola (2002).

5. Difference between the modelling results of BMA and single models

Although the authors presented modelling results in terms of the likelihood surfaces and the distribution functions of various variables at different locations. It would be useful to present the modelling results of the individual models (or of several most plausible models). This will give readers an overview of the difference of the model results in terms *of their spatial distributions. The spatial distribution would be complementary to the likelihood surfaces and the CDF's.*

Corrected. See revised manuscript.

6. Principle of parsimony

Since the posterior model probabilities calculated using the GLUE-BMA method are simply based on model-fit, the principle of parsimony is not considered. Would this affect the model averaging results?

As pointed out by the reviewer, posterior model weights are based on model-fit. The principle of parsimony (i.e. penalizing models based on complexity or the number of parameters), however, can be implemented by defining non-uniform prior model probabilities. These prior distributions should reflect the analyst's prior perception about the plausibility of the alternative conceptualizations and/or potentially a penalizing term due to model complexity. For this application, we kept the analysis at a neutral level by considering the information conveyed by the dataset D solely combined with uniform (flat) prior model probabilities. Definition and analysis of non-uniform priors to comply with the principle of parsimony is beyond the scope of this article.

As discussed in previous comments, using model selection criteria (e.g. AIC, AICc, BIC, KIC to comply with the principle of parsimony) to approximate posterior model probabilities used for multimodel aggregation may lead to conflicting results.

7. Standard deviation discussed in pages 5899 and 5900

It is understandable that the rule of three sigma is used, but it is unknown based on what the value of 10m is assumed. Is it reasonable to test validity of this assumption based on the field observations?

We defined the value of 10m based on pragmatic reasons where the successful implementation of the sampling algorithm was the main goal. 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. As a consequence, chains explored relatively small areas of the posterior space. 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. Keeping in mind the dimensions of the modelled domain and the range of variation for the observed heads (915-1033 masl) we considered this value as acceptable for the problem at hand.

8. Future research

Some discussion of the future research for this site may be interesting.

Corrected. See revised manuscript.

Minor comments

See revised manuscript.

References

DGA-UChile, (1988) Modelo de simulación hidrogeológico de la Pampa del Tamarugal. Centro de Información de Recursos Hídricos, Dirección General de Aguas, Santiago, Chile, 112 pp.

DICTUC, (2008) EIA proyecto Pampa Hermosa. Anexo VIII Hidrogeología. Accessed from https://www.e-seia.cl/expediente/expedientes.php?modo=ficha&id_expediente=3083858&idExpediente=3083858 on 04/02/2009.

DSM, (2002) Estación Cerro Gordo. Modelo de flujo numérico de aguas subterráneas. Santiago, Chile, 70 pp.

Gilks, W., Richardson, S. and D., Spiegelhalter, (1995) Markov Chain Monte Carlo in practice, Chapman & Hall/CRC, Boca Raton, Florida, 668 pp.

JICA-DGA-PCI, (1995) The study on the development of water resources in northern Chile. Supporting report B: Geology and groundwater, 216 pp, Japanese International Cooperation Agency, Dirección General de Aguas, and Pacific Consultants International, Santiago, Chile.

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.

Sorensen, D. and D. Gianola, D., (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics, First Ed., Springer-Verlag, New York, 740 pp.