Forecasting Skill of Model Averages

C. L. Winter[1,2] and Doug Nychka[2]

September 17, 2008

Abstract. Given a collection of science-based computational models that all estimate states of the same environmental system, we compare the forecast skill of the average of the collection to the skills of the individual members. We illustrate our results through an analysis of regional climate model data and give general criteria for the average to perform more or less skillfully than the most skillful individual model, the "best" model. The average will only be more skillful than the best model if the individual models in the collection produce very different forecasts; if the individual forecasts generally agree, the average will not be as skillful as the best model.

.

Section 1. Introduction

Scientific models of environmental systems are based on accepted physical, chemical and biological principles of energy and mass transfer. The goal of a science-based environmental system model is to approximate selected states of the system, and science-based computational models are often used to forecast system states. Weather prediction, climate studies, and analyses of groundwater flow and transport provide examples too numerous to list. The accuracy of model forecasts must be assessed when management and policy decisions are based on them. In general, forecasts are assessed by comparing predicted states to observations taken over given periods, locations, or both. A wide range of evaluation measures has been used for assessment, including correlations between observed states and forecasts (*Epstein and Murphy*, 1989; *Murphy*, 1988), anomaly correlation coefficients (*Wilks*, 2005), ranked probability score (*Epstein*, 1969; *Murphy*, 1971), receiver operating characteristic under the curve (*Swets*, 1973), Peirce skill score (*Pierce*, 1884), potential predictability (*Boer*, 2004), various information criteria (*Neuman*, 2003; Ye *et al.*, 2008), odds ratio skill score (*Thornes and Stephenson*, 2001), and mean-square error (*Wilks*, 2005), which is the basis of square-error skill scores used in weather forecasting (*Murphy*, 1988 and 1996) and is also the basis for the analysis in this paper.

Assessing model forecasts is complicated by the existence of alternative models that produce different estimates of system states. The question of how to accommodate differences among models naturally arises, and one response has been to average model forecasts, the idea being that an average can achieve a consensus among individuals that

1 – Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721

2 – National Center for Atmospheric Research, Boulder, CO, 80305

emphasizes their points of agreement. In both climate (*Lambert and Boer*, 2001; *Palmer et al*., 2004; *Latif et al*., 2006) and groundwater applications (*Ye et al.*, 2004, 2005, 2008; *Bevin*, 2006; *Poeter and Anderson*, 2007), the (weighted) average of a collection of models seems to produce better forecasts than any individual model when evaluated by standard measures of skill. For example, the average of models is in general the "… 'best' model in its ability to simulate current climate, at least in terms of typical second order measures such as mean square differences, spatial correlation, and the ratio of variances" (*Boer*, 2004). Nonetheless, the reasons for that are not completely understood (*Latif et al.*, 2006).

We use observed mean-square error,

$$MSE(X,Y) = \frac{1}{T}\sum_{t=1}^{T}(X_t - Y_t)^2 . \qquad (1)$$

to compare the relative skill of individual models and their averages because it is a standard measure of forecast skill (e.g., *Murphy*, 1988, *Epstein and Murphy*, 1989; *Murphy*, 1996; *Lambert and Boer*, 2001; *Boer*, 2004; *Palmer et al.*, 2004; *Wilks*, 2005; *Poeter and Anderson*, 2005; *Ye et al.*, 2008). Observed *MSE* is a strong measure of skill in the sense that it directly compares model forecasts, $X_t \in X = (X_1,\ldots,X_T)$ to observed state variables, $Y_t \in Y = (Y_1,\ldots,Y_T)$ obtained over an interval of length T. The observed variables may be direct observations of the system state or a related variable, for instance, some combination of the principal components of the observed state and/or model. From here on "skill" observed means square error skill.

Our goal is to compare the skill of forecasts of a system state at t made by $m = 1, \ldots, M$ individual models, $X_t^{(m)}$, to the skill of their average,

$$\overline{X}_t = \sum_{m=1}^{M} w_m X_t^{(m)}, \text{ where } \sum_m w_m = 1 \text{ and } w_m \geq 0 \text{ for all } m. \qquad (2)$$

In many climate applications, e.g., *Meehl et al.* (2007), model weights are uniform, $w_m = 1/M$, and we use uniform weights in our climate example (Section 3). However, that is not necessary. In geohydrology, e.g., *Neuman* (2003), *Ye et al*. (2004, 2005, and 2008), weights are often chosen by Bayesian methods (*Hoeting et al*., 1999). We make no assumptions about weights, beyond those just stated in (2), to derive our general results (Section 4), which are therefore independent of the method used to choose weights. In some approaches, for instance *Palmer et al*. (2004), the values $X_t^{(m)}$ are themselves the result of stochastic averaging, but that also does not affect our analysis. Krishnamurti and his colleagues have considered weighted ensembles where some of the weights may be negative (e.g., *Krishnamurti et al*., 2009), but that approach raises several questions that go beyond the scope of this paper and so we only consider nonnegative weightings.

We show by example (Section 3) and analysis (Section 4) that a forecast produced by averaging outputs from a collection of *M* models will be more skillful than the forecast of any individual model, $X^{(m)}$, only if the models in the collection do not correspond too closely. In other words, for the average to be more skillful than any $X^{(m)}$, it is necessary that the collection of models include a diverse set of distinctly different forecasts. Second, if the forecasts of individual models are too similar, the average will produce worse forecasts than the most skillful individual model. Intuitively, averaging in this case dilutes the best forecast with other, similar forecasts that are not as good.

Section 2. Components of *MSE*($\overline{X}$, *Y*)

Using $w_m \geq 0$ for all *m*, the *MSE* of the average of a collection of models,

$$MSE(\overline{X},Y) = \frac{1}{T}\sum_{t=1}^{T}(\overline{X}_t - Y_t)^2 = \sum_m w_m^2 MSE(X^{(m)},Y) + \sum\sum_{m\neq m'} w_m w_{m'} R_{m,m'}, \qquad (3)$$

depends on the *MSE*s of the individual models, as well as the *correspondences* between models,

$$R_{m,m'} = \frac{1}{T}\sum_t (X_t^{(m)} - Y_i)(X_t^{(m')} - Y_t). \qquad (4)$$

The correspondence has an obvious geometrical interpretation,

$$R_{m,m'} = \sqrt{MSE(X^{(m)},Y)MSE(X^{(m')},Y)}\cos\theta_{m,m'} \qquad (5)$$

due to its dependence on the angle, $\theta_{m,m'}$ between the vectors

$$\vec{Z}^{(m)} = (X_1^{(m)} - Y_1, \ldots, X_T^{(m)} - Y_T) \qquad (6)$$

and $\vec{Z}^{(m')}$ in $\Re^T$. We use this to motivate our general results in Section 4. If the models are unbiased, *i.e.,* $E[X_t^{(m)}] = Y_t$, the correspondence is proportional to the correlation between models and has similar mathematical properties. But it should be emphasized that we are focusing on the model results as a fixed set of outcomes and are not assuming additional probability structure in this problem. From now on we write $MSE(\vec{Z}^{(m)}) = MSE(X^{(m)},Y)$ when it is convenient.

Section 3. North American Climate Example

To illustrate these relationships, we consider the departures,

$$\vec{Z}^{(m)} = (Z_{WNA}^{(m)}, Z_{CNA}^{(m)}, Z_{ENA}^{(m)}) \tag{7}$$

produced by 19 global climate models ($m = 1,\ldots,19$) when estimating normal winter temperature for Western North America (*WNA*), Central North America (*CNA*) and Eastern North America (*ENA*). The normal is defined for the period 1970-1999, and winter consists of December, January and February. The data are normalized by summing over months and through the normal period 1970-1999,

$$Z_{i,j}^{(m)} = \frac{1}{3}(Z_{i,j,Dec}^{(m)} + Z_{i,j,Jan}^{(m)} + Z_{i,j,Feb}^{(m)}) \text{ and } Z_i^{(m)} = \frac{1}{30}\sum_{j=1970}^{1999} Z_{i,j}^{(m)}. \tag{8}$$

Here we use *i* = *WNA*, *CNA*, or *ENA* instead of *t* to emphasize that the data are normalized and range over regions.

The model results are taken from the coordinated modeling effort supporting the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (*Meehl et al.*, 2007). The model output fields were re-gridded to a common 5 degree grid and compared to the observational data set from the Climate Research Unit (CRU), East Anglia, and the Hadley Centre, UK MetOffice (*Jones et al*., 1999). More details of these data and a global analysis of this multi-model sample can be found in (*Tebaldi et al.*, 2005; *Tebaldi and Knutti*, 2007; *Jun et al.*, 2008; *Knutti and Nychka,* 2008). The regions *WNA*, *CNA*, and *ENA* are a subset of the Giorgi and Mearns divisions (*Giorgi and Mearns*, 2002).

The model departures from regional normals and *MSEs* are shown in *Table* 1 where models correspond to rows and regions to columns. The last two rows are average departures over 1) the full suite of models, $\overline{Z}_{19}$, including the most skillful, and 2) the suite of models with the most skillful left out, $\overline{Z}_{18}$. As noted, models are uniformly weighted in this section, $w_m = 1/M$ for all *m*, where *M* = 18 or 19 depending on which average is being considered.

Looking at the suite of all 19 models first, the average, $\overline{X}_{19}$, is not as skillful as the best model, *m* = 16, for these data: $MSE(X^{(min_{19})}, Y) = MSE(X^{(16)}, Y) = 0.08$ while $MSE(\overline{X}_{19}, Y) = 0.39$. When model 16 is removed from the set of models, $MSE(X^{(min_{18})}, Y) = MSE(X^{(2)}, Y) = 0.47$ while the average of the reduced set, $\overline{X}_{18}$, yields $MSE(\overline{X}_{18}, Y) = 0.45$. These two cases illustrate the sensitivity of model evaluations to the set of models. There are perhaps other factors such as the specific observation interval and choice of regions, however, our analysis does not address these points.

Table 1 -- Departures (℃) and MSE for North America.

Model ID	WNA ($Z_{m,1}$)	CNA ($Z_{m,2}$)	ENA ($Z_{m,3}$)	MSE	Minimum
1	-2.02	-0.44	1.07	1.80	
2	0.37	-1.11	-0.17	0.47	18 Models
3	0.48	-1.35	-0.22	0.70	
4	-3.85	-3.15	-2.29	10.00	
5	-3.03	-2.25	-0.81	4.97	
6	0.66	1.53	2.10	2.39	
7	-2.39	0.54	0.51	2.09	
8	-1.66	-1.01	-0.69	1.41	
9	0.69	1.89	4.28	7.44	
10	-0.37	0.61	1.10	0.57	
11	1.27	-1.70	-0.37	1.54	
12	-0.26	1.60	2.01	2.22	
13	0.73	1.23	-0.53	0.78	
14	-0.50	2.34	1.48	2.65	
15	-0.91	-0.54	-1.49	1.11	
16	0.31	0.27	0.27	0.08	19 Models
17	-2.16	-1.96	-2.77	5.39	
18	-1.77	-2.93	0.08	3.90	
19	-3.57	-2.75	0.12	6.76	
$\bar{Z}_{19}$	-0.95	-0.48	0.19	0.39	0.08
$\bar{Z}_{18}$	-1.02	-0.52	0.19	0.45	0.47

Additional insight can be gained by decomposing

$$MSE(\bar{Z}) - MSE(\vec{Z}^{(min)}) = \sum_m w_m^2 (MSE(\vec{Z}^{(m)}) - MSE(\vec{Z}^{(min)}))$$

$$+ \sum \sum_{m \neq m'} w_m w_{m'} (R_{m,m'} - MSE(\vec{Z}^{(\min)}))$$

$$= m^2 + r \,, \qquad (9)$$

which makes clear that the model average is more skillful than the best model when $r \leq -m^2$, i.e., when the models do not correspond too much. Values of r and m corresponding to the full suite of nineteen models with model 16 included are $r = 0.16$ and $-m^2 = -0.15$ [*Table* 2]. This is a case where the best model is so much more skillful than all others that averaging only dilutes its individual skill. When model 16 is removed, there is then enough disagreement among the remaining models for the average to perform better than any of the rest: $r = -0.17$ and $-m^2 = -0.15$ [*Table* 2].

Table 2 –Effect of Model Collection on Performance

Number of Models	$-m^2$	r	MSE_{Ave}	MSE_{min}
19	-0.15	0.16	0.39	0.08
18	-0.15	-0.17	0.45	0.47

Section 4. General Results

Our general results consist of a sufficient condition for the model average to perform less skillfully than the best model, and a necessary condition for the average to perform better. Each result only requires that the weights be non-negative and sum to 1 (Eqn 2).

Result 1. *If the models correspond too closely, the average is less skillful than the best model*. Intuitively, this is because the other models dilute the performance of the best model in this case. Referring to Eqn 9, it is clear that $MSE(\overline{Z}) > MSE(Z^{(min)})$ if $R_{m,m'} > MSE(Z^{(min)})$, for all m and m'.

Another way for the best model to be more skillful than the average is for the best to be much more skillful than all the other models, i.e., $MSE(\vec{Z}^{(min)}) << MSE(\vec{Z}^{(m)})$ for all $m \neq min$, as was the case in the climate example. In that case, $MSE(\overline{Z}_{n+1}) - MSE(Z^{(min)}) \cong MSE(\overline{Z}_n) \geq 0$. However, Result 1 shows it is not necessary for one model to be much more skillful than the rest for the average to be less skillful; it is enough for all the models to correspond at a level that only depends on $MSE(Z^{(min)})$.

Result 2. *The average is more skillful than the best model only if some individual models do not correspond too much.* This is derived from Result 1 as a proof by contradiction (*modus tollens*).

To emphasize the role of geometry, we illustrate the results in $\Re^T$, the T-dimensional vector space of model forecasts [*Figures* 1 and 2]. Letting $\overline{Z} = (\overline{Z}_1,\ldots,\overline{Z}_T)$ and taking $\vec{Z}^{(m)}$ from Eqn 6, the MSE are equivalent to squared lengths in $\Re^T$, $MSE(\overline{Z}) = \frac{1}{T}\left\|\overline{Z}\right\|^2$ and $MSE(Z^{(m)}) = \frac{1}{T}\left\|\vec{Z}^{(m)}\right\|^2$, while $R_{m,m'} = \cos\theta_{m,m'}$ is the vector product $\vec{Z}^{(m)} \bullet \vec{Z}^{(m')}$. In the Figures, $T = 3$ for convenience of illustration, so each $\vec{Z}^{(m)} = (Z_1^{(m)},Z_2^{(m)},Z_3^{(m)})$. The thin vectors represent the performances of different models, while the thick vector is the average, $\overline{Z}$, their lengths corresponding to $MSE(\overline{Z})$ and $MSE(Z^{(m)},Y)$ respectively.

Figure 1 shows a typical case of Result 1 leading to $MSE(\overline{Z}) > MSE(Z^{(min)})$. The models all vary similarly about $\vec{Y}$, as evidenced by their approximate colinearity, but $\vec{Z}_{\min}$

performs much better than the others. The result of averaging is to "stretch" $\bar{Z}$ away from $\vec{Z}_{\min}$ in the general direction of the other models. Figure 2 illustrates Result 2, a case when $MSE(\bar{Z}) \leq MSE(Z^{(min)})$. The role of the anti-correspondence requirement,

$$\cos\theta_{m,m'} < \frac{MSE(Z^{(min)})}{\sqrt{MSE(Z^{(m)})MSE(Z^{(m')})}} \text{ for some } m,\ m' ,$$

is clear: $MSE(Z^{(min)}) > MSE(\bar{Z})$ because the two sets of vectors "pull" against each other to produce a reduced $\bar{Z}$.

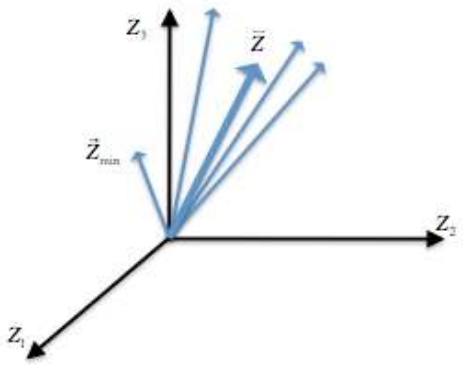

Figure 1 – Example of Result 1. Colinearity places $\bar{Z}$ beyond $\vec{Z}^{(min)}$.

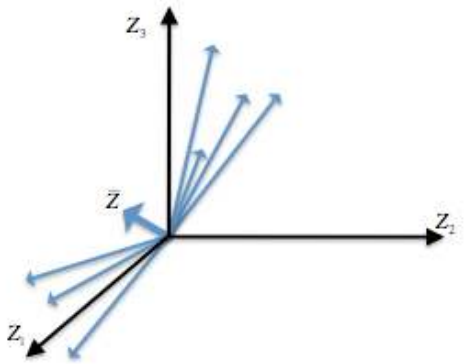

Figure 2 – Example of Result 2. Weak correspondence allows $MSE(\bar{Z},Y) < MSE(Z^{(min)},Y)$.

Section 5. Summary and Discussion

Alternate science-based computational models of a given environmental system always forecast system states that differ somewhat, and sometimes forecast states that differ considerably. Model averaging has been proposed as a means for dealing with differences among model forecasts, and it has been noted that in some cases the average of a collection of models produces "better" forecasts than any individual model in the collection (e.g., *Boer*, 2004). We have compared models and their averages on the basis of mean-square difference skill score, which is a standard assessment measure in weather forecasting, climate studies and groundwater hydrology. In addition to its ubiquity, no other measure of skill has the straightforward metric properties of mean square error, the normalized distance between observations and forecasts.

We investigated two climate examples to compare the sensitivity of the skill of an average of a collection of models to the most skillful model (the "best") in the collection.

The best model in the first example was so much more skillful than any other model that the average simply could not perform as well as the best. When the best model was removed from the first collection, the average performed better than any individual model in the reduced collection. In that case, none of the remaining climate models performed well enough to dominate the others. Furthermore, the models disagreed, thus giving an example of the general result (Result 2) that the average can be more skillful than the best model only if some models make markedly different forecasts. The example also illustrated the sensitivity of skill assessments to the collection of models, but we did not go more deeply into that point.

Our general results give a sufficient condition for the best model to be more skillful than the average (Result 1) and a necessary condition for the average to be more skillful than the best individual (Result 2). In general, 1) the average is less skillful than the best individual if the forecasts of individuals correspond closely to each other when compared to the skill of the best model, and 2) the average is more skillful than the best model only if the forecasts of some individuals do not correspond. These results depend just on simple geometric properties of the collection of models, and are independent of i) how the models make their forecasts and ii) the weighting scheme used to derive their average (except the weights must be positive and sum to 1).

These results, and the example, suggest a certain amount of caution should be applied when making strong claims that model averages are more skillful than individual models, at least if those claims are based on second-order performance measures like observed square-error skill. At the same time, the results should not be over-interpreted. In the first place, Result 2 does not indicate that collections of models should be assembled with the idea of maximizing the differences among individuals. The goal is to make skillful forecasts, not to merely have a collection of models whose average is more skillful on a set of observations than any individual model. At the same time, Result 1 does not imply that an average is never useful when models agree. A collection of good models (models based on reasonable physical assumptions and estimates of system parameters) can be expected to produce forecasts that correspond strongly with each other, but differ in their details. An average might be useful in that setting even if it is not as skillful as the best model on a set of observations.

The discussion of the relative skill of science-based environmental models and their averages has taken place so far in the absence of statistical tests. The complex probability distributions of environmental forecasts in realistic settings is one reason for that. What is needed is a statistical test (or tests) to evaluate hypotheses about the means of forecasts made by complex physics models whose uncertain physical are not Gaussian, or even necessarily unimodal (cf., *Rubin,* 1995; *Gomez-Hernandez and Wen*, 1998; *Winter and Tartakovsky*, 2000, 2002; *Christakos*, 2003; *Guadagnini et al.*, 2003; *Neuman and Wierenga*, 2003). In the absence of such tests, the results in this paper indicate the sources of apparent forecasting skills of model averages have a simple geometric explanation in some cases.

Section 6. References

Beven, K., "A manifesto for the equifinality thesis", *J. Hydrol.*, 320, 2006.

Boer, G.J., "Long time-scale potential predictability in an ensemble of coupled climate models," *Climate Dynamics,* 23, 2004.

Christakos, G., "Another look at the conceptual fundamentals of porous media upscaling," *Stochastic Environmental Research and Risk Assessment* 17 (5), 2003.

Epstein, E.S., "A scoring system for probability forecasts of ranked categories," *J. Appl. Meteorol*., 8, 1969.

Giorgi, F. and L. O. Mearns, "Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM cimulations via the "Reliability Ensemble Averaging" (REA) method," *J. Climate*, 15, 2002.

Gomez-Hernandez, J. J., and X.-H.Wen, "To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology," *Adv. Water Resour.*, 21(1), 1998.

Guadagnini, A. , L. Guadagnini, D. M. Tartakovsky and C. L. Winter, "Random domain decomposition for flow in heterogeneous stratified aquifer," *Stochastic Environmental Research and Risk Assessment* 17 (6), 2003

Jones, P.D., M. New, D.E. Parker, S. Martin and I.G. Rigor, "Surface air temperature and its variations over the last 150 years," *Reviews of Geophysics*, 37 (173),1999.

Jun, M., R. Knutti and D.W. Nychka, "Spatial analysis to quantify numerical model bias and dependence: How many climate models are there?" *J. Am. Stat. Assn.*, 103 (483), 2008.

Knutti, R., M. Jun, and D W. Nychka, "Local eigenvalue analysis of CMIP3 climate model errors," *Tellus A*, 60 (5), 2008.

Latif, M., M. Collins, H. Pohlmann and N. Keenlyside, "A review of predictability studies of Atlantic sector climate on decadal time scales," *J. Climate* 19, 2006.

Meehl, G.A., Stocker, T.F., Collins, W.D., Friedlingstein,P., Gaye, A.T., Gregory, J.M., Kitoh, A., Knutti, R., Murphy, J.M., Noda, A., Raper, S.C.B., Watterson, I.G., Weaver, A.J. and Zhao, Z. C., *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,* chapter *Global Climate Projections*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Murphy, A.H., "A note on the ranked probability score," *J. Appl.* Meteorol., 10, 1971.

Neuman, S.P., "Maximum likelihood Bayesian averaging of uncertain model predictions," *Stochastic Environmental Research and Risk Assessment* 17 (5), 2003.

Neuman, S.P. and P.J. Wierenga, "A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites (NUREG/CR-6805), Office of Nuclear Regulatory Research U. S. Nuclear Regulatory Commission Washington, DC, July 2003.

Peirce, C.S., "The numerical measure of the success of predictions. *Science,* 4**,** 1884.

Poeter, E. and D. Anderson, "Multimodel ranking and inference in ground water modeling", *Ground Water*, 43 (4), 2005.

Rubin, Y., "Flow and transport in bimodal heterogeneous formations," *Water Resources Research*, 31(10), 1995.

Swets J.A., "The relative operating characteristic in psychology," *Science* 182, 1973.

Tebaldi C., R. L. Smith, D. Nychka and L.O. Mearns, “Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles”, *J. Climate*, 18 (10), 2005.

Tebaldi C. and R. Knutti, “The use of the multi-model ensemble in probabilistic climate projections”, *Phil. Trans. Royal Soc. A*, 365 (1857), 2007.

Thornes J.E. and D.B. Stephenson, “How to judge the quality and value of weather forecast products,” *Meteorol. Appl.*, 8, 2001.

Wilks, D.S., *Statistical Methods in the Atmospheric Sciences*, 2nd Edition, Academic Press, 2005.

Winter, C.L. and D.M. Tartakovsky, "Mean flow in composite porous media", *Geophysical Research Letters*, 27 (12), 2000.

Winter, C.L. and D.M. Tartakovsky, "Groundwater flow in heterogeneous composite aquifers", *Water Resources Research*, 38 (8), 2002.

Ye, M., S. P. Neuman, and P. D. Meyer, “Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff”, *Water Resources Research*, 40 (5), 2004.

Ye, M., S. P. Neuman, P. D. Meyer, and K. F. Pohlmann, “Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff”, *Water Resources Research*, 41 (12), 2005.

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