Maximum likelihood Bayesian averaging of air flow models in unsaturated fractured tuff

Eric Morales-Casique[1], Shlomo P. Neuman[2], Velimir V. Vesselinov[3]

[1] *Department of Earth Sciences, Faculty of Geosciences, University of Utrecht P.O. Box 80021 3508 TA Utrecht, The Netherlands*

[2] *Department of Hydrology and Water Resources, The University of Arizona, 1133 E James E. Rogers Way Tucson, AZ 85721, USA*

[3] *Los Alamos National Laboratory, EES-6, MS T003, Los Alamos, NM 87545, USA*

Abstract We use log permeabilities and porosities obtained from single-hole pneumatic packer tests in six boreholes drilled into unsaturated fractured tuff near Superior, Arizona, to postulate, calibrate and compare five alternative variogram models (exponential, exponential with linear drift, power, truncated power based on exponential modes, and truncated power based on Gaussian models) of these parameters based on four model selection criteria (*AIC*, *AICc*, *BIC* and *KIC*). As all four criteria favour the first three of these variogram models, we adopt the three favoured models to parameterize log air permeability and porosity across the site via kriging in terms of their values at selected pilot points and at some single-hole measurement locations. For each of the three variogram models we estimate log air permeabilities and porosities at the pilot points by calibrating a finite volume pressure simulator against two cross-hole pressure data sets from sixteen boreholes at the site. The traditional Occam's window approach in conjunction with *AIC*, *AICc*, *BIC* and *KIC* assigns a posterior probability of nearly 1 to the power model. A recently proposed variance window approach does the same when applied in conjunction with *AIC*, *AICc*, *BIC* but spreads the posterior probability more evenly among the three models when

used in conjunction with *KIC*. We compare the abilities of individual models and MLBMA, based on both the Occam's window and variance window approaches, to predict space-time pressure variations observed during two cross-hole tests other than those employed for calibration. Individual models with the largest posterior probabilities turned out to be the worst or second worst predictors of pressure in both validation cases. Correspondingly, some individual models predicted pressures more accurately than did MLBMA. MLBMA was far superior to any of the individual models in one validation test and second to last in the other validation test in terms of predictive coverage and log scores.

Keywords Bayesian model averaging; air flow; inverse modelling; maximum likelihood

Correspondence author:

Eric Morales-Casique; morales@geo.uu.nl; phone (+31) 30 2535139; fax: (+31) 30 2535030

INTRODUCTION

Hydrologic analyses typically rely on a single conceptual-mathematical model. Yet hydrologic environments are open and complex, rendering them prone to multiple interpretations and mathematical descriptions. Adopting only one of these may lead to statistical bias and underestimation of uncertainty. Thus, hydrologists have developed several approaches to weigh and average predictions generated by alternative models (Neuman 2003; Neuman and Wierenga 2003; Ye et al. 2004; Poeter and Anderson 2005; Beven 2006; Refsgaard et al. 2006).

Bayesian model averaging (BMA) (e.g., Hoeting et al. 1999) provides an optimal way to combine the predictions of several competing models and to assess their joint predictive uncertainty. Neuman (2003) proposed a maximum likelihood (ML) version of BMA (MLBMA) that renders it compatible with ML methods of model calibration (Carrera and Neuman 1986; Hernandez et al. 2006) even in cases where prior information about the parameters is not available (such information being a prerequisite for the use of BMA). In the framework of MLBMA, if Δ is a quantity one wants to predict given a discrete set of data $\mathbf{D}$, then its posterior (conditional) mean and variance are (Neuman 2003)

$$E\left[\Delta \mid \mathbf{D}\right] = \sum_{i=1}^{K} E\left[\Delta \mid M_i, \hat{\mathbf{b}}_i, \mathbf{D}\right] p(M_i \mid \mathbf{D}) \tag{1}$$

$$Var\left[\Delta \mid \mathbf{D}\right] = \sum_{i=1}^{K} Var\left[\Delta \mid \mathbf{D}, M_i, \hat{\mathbf{b}}_i\right] p(M_i \mid \mathbf{D}) + \sum_{i=1}^{K} \left(E\left[\Delta \mid \mathbf{D}, M_i, \hat{\mathbf{b}}_i\right] - E\left[\Delta \mid \mathbf{D}\right]\right)^2 p(M_i \mid \mathbf{D}) \tag{2}$$

where model M_i has parameters $\mathbf{b}_i$, $E\left[\Delta \mid M_i, \hat{\mathbf{b}}_i, \mathbf{D}\right]$ and $Var\left[\Delta \mid \mathbf{D}, M_i, \hat{\mathbf{b}}_i\right]$ are posterior mean and variance of Δ under the *i*-th alternative model, and $\hat{\mathbf{b}}_i$ is a maximum likelihood estimate of

$\mathbf{b}_i$ based on the likelihood $p\left(\mathbf{D}\,|\,\hat{\mathbf{b}}_i, M_i\right)$. The posterior probability of the *i*-th alternative model, $p(M_i\,|\,\mathbf{D})$ is approximated on the basis of Occam's window by (Ye et al. 2004)

$$p(M_i\,|\,\mathbf{D}) = \left[\exp\left(-\Delta IC_i\right) p(M_i)\right] \Big/ \left[\sum_{j=1}^{M} \exp\left(-\Delta IC_j\right) p(M_j)\right] \tag{3}$$

where $\Delta IC_i = IC_i - IC_{\min}$, $IC_i = KIC_i$ being the Kashyap (1982) information criterion for the *i*-th model and $IC_{\min}$ is the minimum value among the models. Alternatively, posterior model weights are sometimes assigned by setting IC_i equal to information theoretic criteria (Poeter and Anderson 2005; Ye et al. 2008) such as *AIC* (Akaike 1974), *AICc* (Hurvich and Tsai 1989) or the Bayesian criterion *BIC* (Schwarz 1978). Ye et al. (2008) explain that *KIC* is the only one among these criteria which validly discriminates between models based not only on the quality of model fit to observed data and the number of model parameters but also on how close are the posterior parameter estimates to their prior values and the information contained in the observations.

Experience indicates (and our results below confirm) that Eq. (3) tends to assign posterior probabilities or model weights of nearly 1 to one (the best) model and nearly zero to all other models. Tsai and Li (2008) suggest that this is because Occam's window is often too narrow to accommodate models that are not the best but still potentially acceptable. As a remedy, they propose to rely on a broader variance window obtained upon scaling ΔIC_i in Eq. (3) by a factor α selected subjectively by the analyst based on a desired level of significance, which determines the size of the variance window: $\alpha = c/\sqrt{n}$, where *n* is the number of observation data and *c* is a coefficient which depends on the window size and desired significance level.

We use log permeabilities and porosities obtained from single-hole pneumatic packer tests in six boreholes drilled into unsaturated fractured tuff near Superior, Arizona, to postulate,

calibrate and compare five alternative variogram models of these parameters based on *AIC*, *AICc*, *BIC* and *KIC*. As all four criteria favour the first three of these variogram models, we adopt the three favoured models to parameterize log air permeability and porosity across the site via kriging in terms of their values at selected pilot points and, optionally, at some single-hole measurement locations. For each of the three variogram models we estimate log air permeabilities and porosities at the pilot points by calibrating a finite volume pressure simulator against two cross-hole pressure data sets from sixteen boreholes at the site. Finally, we compare the abilities of individual models and MLBMA, based on both the Occam's window and variance window approaches, to predict space-time pressure variations observed during two cross-hole tests other than those employed for calibration.

BACKGROUND ON THE APACHE LEAP RESEARCH SITE

The previous University of Arizona Apache Leap Research Site (ALRS) near Superior, Arizona is a block of unsaturated fractured tuff measuring $64 \times 55 \times 46$ m (Fig. 1). The test site includes sixteen boreholes, three vertical (V1, V2, V3) and thirteen inclined at 45° (X1, X2, X3, Y1, Y2, Y3, Z1, Z2, Z3, W1, W2, W2A, W3). Several pneumatic cross-hole tests were conducted at the ALRS (Illman et al. 1998; Illman and Neuman 2001); a summary of the conditions for each test is presented in Table 1. For inverse calibration we selected the cross-hole tests labelled PP4 and PP5; we validated the calibrated models by predicting pressure variations during cross-hole tests, PP6 and PP7. During each test air was injected into a given interval and responses were monitored in 13 relatively short intervals (0.5–2 m) and 24 relatively long intervals (4–42.6 m)

shown in Fig. 1. The hydrologic parameters controlling air-flow are air permeability k and air-filled porosity ϕ, both attributed largely to air-filled fractures transecting water-saturated porous tuff.

Table 1 Cross-hole tests conditions at ALRS (Illman et al. 1998)

Test	Flow regime	Injection Interval		Injection Rate (kg/s)
		Location	Length (m)	
PP4	Const. Rate	Y2-2	2	10^{-3}
PP5	Step*	X2-2	2.2	10^{-4}
PP6	Step*	Z3-2	2	10^{-4}
PP7	Step*	W3-2	1.2	10^{-4}

* Only data from the first stage was included

ALTERNATIVE GEOSTATISTICAL MODELS OF AIR PERMEABILITY AND AIR-FILLED POROSITY

$\mathbf{Log_{10}}$ *k*. Ye et al. (2004) used MLBMA to investigate the geostatistical properties of log air permeability k (m^2) at ALRS by postulating several alternative variogram models based on 184 data of $\log_{10}k$ obtained via steady-state interpretation of stable pressure data from pneumatic injection tests in 1-m long intervals along six boreholes, V2, W2A, X2, Y2, Y3 and Z2 in Fig. 1 (Guzman et al. 1996). Ye et al. (2004) fitted seven variogram models (power *P*, exponential *E*, exponential with first order drift *E1*, exponential with second order drift *E2*, spherical *S*, spherical with first order drift *S1*, and spherical with second order drift *S2*) to this data set using

the adjoint state maximum likelihood cross validation (ASMLCV) method of Samper and Neuman (1989) in conjunction with universal kriging and generalized least squares methods. They found that the first three models (*P*, *E* and *E1*) consistently dominated in terms of their posterior model probability. We expanded their list of models to include truncated power models based on Gaussian (*Tpg*) and exponential (*Tpe*) modes (Di Federico and Neuman 1997), fitted the variogram models using the same data set and the same procedure, computed the values of four model selection criteria (*AIC*, *AICc*, *BIC* and *KIC*) and computed the corresponding posterior model probability. Table 2 lists the results of this analysis, where posterior probabilities or (in the case of *AIC* and *AICc*) model weights are based on equal prior probabilities $p(M_k)$ (the neutral choice) for all models. Model *E1* is associated with the smallest negative log-likelihood value *NLL* (e.g. Carrera and Neuman 1986) and thus provides the best fit to the data. When using Occam's window, model ranking varies depending on the information criterion. Whereas *AIC* and *AICc* strongly prefer *E1* and *P* in this order over all other models, *BIC* strongly prefers *P*. On the other hand, *KIC* shows a slight preference for *E1* over *P* while considering *E* to be a not much less promising option. Whereas in terms of *NLL* the truncated power models, *Tpg* and *Tpe*, fit the sample variogram as well as does *P* (Fig. 2), they are ranked lower by all four model selection criteria due to their larger number of parameters. *KIC* is the only such criterion showing a clear preference for *Tpg* over *Tpe*. Alternatively, a variance window of size $4\sigma_D$ and a significance level of 5%, leads to $\alpha = 0.078$ and posterior probabilities that are distributed more evenly among all models, and the difference in magnitude between probabilities based on different information criteria is reduced.

Table 2 ASMLCV results for $\log_{10} k$

	Power *P*	Exponential *E*	Exponential, 1st order drift *E1*	*Tpe*	*Tpg*
Numb. Parameters	2	2	6	3	3
Numb. Observations	184	184	184	184	184
Sill/Coefficient	0.29	0.72	0.51	0.08	0.12
Integral scale/Exponent	0.46	1.84	1.24	0.23	0.23
Lower cutoff				1.14×10^{-5}	1.56×10^{-4}
NLL	352.19	361.01	341.57	352.19	352.19
AIC	356.19	365.01	353.57	358.19	358.19
Ranking	2	5	1	4	3
p_{AIC}, %, $\alpha=1$	18.33	0.22	67.97	6.73	6.74
p_{AIC}, %, $\alpha=0.078$	21.79	13.93	24.90	19.69	19.69
AICc	356.25	365.07	354.04	358.32	358.32
Ranking	2	5	1	4	3
p_{AICc}, %, $\alpha=1$	21.07	0.26	63.70	7.48	7.50
p_{AICc}, %, $\alpha=0.078$	21.54	15.26	23.48	19.86	19.86
BIC	362.62	371.44	372.86	367.83	367.83
Ranking	1	4	5	2	3
p_{BIC}, %, $\alpha=1$	85.80	1.04	0.51	6.33	6.32
p_{BIC}, %, $\alpha=0.078$	24.94	17.67	16.71	20.34	20.34
KIC	369.58	370.15	369.45	385.77	371.10
Ranking	2	3	1	5	4
p_{KIC}, %, $\alpha=1$	30.43	22.90	32.42	0.01	14.24
p_{KIC}, %, $\alpha=0.078$	22.44	21.94	22.55	11.92	21.14
Average log scores	49.6	47.8	48.3	69.6	53.6

Notes: All $p(M_i|\mathbf{D})$ were computed assuming $p(M_i) = 1/5$.
The first order drift is given by $f(\mathbf{x}) = a_0 + a_1x + a_2y + a_3z$, with coefficients determined in the manner of (Ye et al 2004) are $a_0 = -15.1805$, $a_1 = 0.03717$, $a_2 = 0.01061$ and $a_3 = 0.04633$.

$\mathbf{Log_{10}\ \phi}$. We conducted a similar geostatistical analysis of 109 log air-filled porosity ($\log_{10}\phi$) data obtained by type-curve interpretation of the recovery phase of single-hole tests conducted on a nominal scale of 1 m (Illman 2005). As there appears to be no discernible cross-correlation between the $\log_{10}\phi$ and $\log_{10}k$ data we analyzed each set separately. Four alternative variogram models were postulated for $\log_{10}\phi$: exponential *E*, spherical *S*, truncated power based on Gaussian *Tpg* and exponential *Tpe* modes. Fig. 3 depicts the models fitted to the sample

variogram and Table 3 lists the corresponding statistics. In terms of *NLL* the truncated power models *Tpe* and *Tpg* fit the data almost equally well and somewhat more closely than do *E* and *S*. Posterior probabilities based on Occam's window and *AIC*, *AICc* and *BIC* rank the two truncated power models as best. However, *KIC* ranks *E* much higher than all other models. By using a variance window of size $4\sigma_D$ at a significance level of 5% ($\alpha = 0.1$), posterior probabilities are distributed more evenly among the models but the ranking is not changed.

Predictive capability of variogram models. We evaluate the predictive capability of variogram models for $\log_{10} k$ and $\log_{10} \phi$ by computing the log scores of the cross-validation errors in the manner of Ye et al. (2004). The data set was split into two parts, eliminating the data corresponding to one borehole at a time, obtaining ML estimates of the parameters and using these to predict the eliminated data. The quality of the predictions was evaluated by the log scores. We repeated the procedure for each data set for $\log_{10} k$ and $\log_{10} \phi$. The log score $-\ln p(\mathbf{D}^v \mid M_k, \mathbf{D}^c)$ (Volinsky et al. 1997), approximated by $-\ln p(\mathbf{D}^v \mid M_k, \hat{\mathbf{b}}_k, \mathbf{D}^c)$ (Ye et al. 2008), is a measure of the predictive capability of a model. The lower the predictive log score of model M_k based on data $\mathbf{D}^c$ (the calibration data set), the smaller the amount of information in $\mathbf{D}^v$ (the validation data set) not covered by model M_k based on $\mathbf{D}^c$. The log score of a model is given by

$$-\ln p(\mathbf{D}^v \mid M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}^c) = \frac{Nv}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{Nv}\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{Nv}\frac{\left(\hat{D}_i^v - D_i^v\right)^2}{\sigma_i^2} \qquad (4)$$

where Nv is the number of data points in $\mathbf{D}^v$, $\hat{D}_i^v$ and σ_i^2 are the i-th kriged variable and the kriging variance, respectively, based on the parameter estimates $\hat{\mathbf{b}}_k$ for model M_k. The results

for average predictive log scores are listed in the last row of Table 2 for $\log_{10} k$ and Table 3 for $\log_{10} \phi$. For $\log_{10} k$, models *E, E1* and *P* have log scores ranging from 47.8 to 49.6, while the log scores of *Tpg* and *Tpe* are considerable larger, 53 and almost 70, respectively. For $\log_{10} \phi$, model *E* has the lowest log scores with 36.2, models *S* and *Tpg* have log scores of about 40 and *Tpe* has largest log socres. Based on these results we retain only models *E*, *E1* and *P* to parameterize $\log_{10} k$ while we retain only model *E* to parameterize $\log_{10} \phi$.

Table 3. ASMLCV results for $\log_{10} \phi$

	Exponential *E*	Spherical *S*	*Tpe*	*Tpg*
Numb. Parameters	2	2	3	3
Numb. Observations	109	109	109	109
Sill/Coefficient	0.25	0.27	0.08	0.08
Integral scale/Exponent	1.03	0.46	0.37	0.35
Lower cutoff			0.29	0.25
NLL	181.1	189.9	174.7	175.3
AIC	185.1	193.9	180.7	181.3
Ranking	3	4	1	2
$p_{AIC},\ \%,\ \alpha = 1$	5.89	0.07	53.17	40.86
$p_{AIC},\ \%,\ \alpha = 0.1$	24.37	15.59	30.47	29.56
AICc	185.2	194.0	181.6	181.5
Ranking	3	4	2	1
$p_{AICc},,\ \%,\ \alpha = 1$	7.21	0.09	45.57	47.14
$p_{AICc},,\ \%,\ \alpha = 0.1$	24.71	15.81	29.67	29.82
BIC	191.0	199.8	189.6	190.1
Ranking	3	4	1	2
$p_{BIC},,\ \%,\ \alpha = 1$	21.51	0.27	44.25	33.98
$p_{BIC},,\ \%,\ \alpha = 0.1$	26.59	17.01	28.55	27.84
KIC	190.9	202.2	199.0	222.7
Ranking	1	3	2	4
$p_{KIC},,\ \%,\ \alpha = 1$	98.00	0.33	1.67	~0
$p_{KIC},,\ \%,\ \alpha = 0.1$	41.23	23.23	27.33	8.21
Average log scores	36.2	40.1	54.8	39.3

Posterior probabilities were computed assuming $p(M_k) = 1/4$.

CALIBRATION OF AIRFLOW MODELS

Following Vesselinov et al. (2001a; 2001b) we calibrate a finite volume pressure simulator (FEHM; Zyvoloski et al. 1999) against cross-hole pressure data using a parameter estimation code (MPEST; V. Vesselinov, personal communication; a parallelized version of PEST, Doherty 1994). Additional elements of the calibration process include geostatistical interpolation of $\log_{10}k$ and $\log_{10}\phi$ via kriging (GSTAT; Pebesma and Wesseling 1998) and *a posteriori* averaging of pressure at grid nodes along packed-off pressure monitoring intervals. Details of the simulation grid, the air-flow equation and its solution can be found in Vesselinov et al. (2001a); here we merely mention that the upper boundary condition was set to constant barometric pressure; monitoring intervals in which observed pressure showed a clear influence of atmospheric pressure fluctuations are not considered in the analysis.

We parameterize $\log_{10} k$ and $\log_{10} \phi$ geostatistically and estimate their values by inverse calibration at selected pilot points (de Marsily et al. 1984). We then project these estimates (with or without the available 184 1-meter scale $\log_{10} k$ measurements) by kriging onto a grid. In the case of $y = \log_{10} k$ the projection is done through $y^* = \sum_{i=1}^{Npp} \lambda_i y_i + \sum_{j=1}^{Na} \lambda_j y_j$ where y^* is the value at any point within the simulated block, y_i are unknown values (parameters) at N_{pp} pilot points, y_j are known values at N_a measurement points, and λ_i and λ_j are kriging weights. Following Vesselinov et al. (2001a; 2001b) we set $N_{pp} = 32$; 29 pilot points are placed at the centers of pressure monitoring intervals (small dots in Fig. 1) and 3 are offset from the center of the injection interval to better represent air-flow. Of the 184 1-m $\log_{10} k$ data 18 correspond to locations at pilot points and are included as priors in the manner discussed below, thus $N_a = 166$.

Inversion entails minimizing the negative log-likelihood criterion (Carrera and Neuman 1986)

$$NLL(\mathbf{b}) = \frac{\Phi_s}{\sigma_s^2} + \frac{\Phi_p}{\sigma_p^2} + \left(N_s + N_p\right)\ln(2\pi) + N_s \ln\sigma_s^2 + \ln\left|\mathbf{Q}_s^{-1}\right| + N_p \ln\sigma_p^2 + \ln\left|\mathbf{Q}_p^{-1}\right| \qquad (5)$$

where **b** is a vector of M parameters to be estimated, N_s is the number of observed state variables, N_p is the number of prior parameter values, $\Phi_s = \mathbf{r}_s^T \mathbf{Q}_s \mathbf{r}_s$ is a generalized sum of square residuals of the state variable, $\Phi_p = \mathbf{r}_p^T \mathbf{Q}_p \mathbf{r}_p$ is a generalized sum of square residuals of the parameters, $\mathbf{Q}_s$ and $\mathbf{Q}_p$ are corresponding weight matrices (considered known), and σ_s^2 and σ_p^2 are scalar multipliers (nominal variances, considered unknown) of the covariance matrices $\mathbf{C}_s = \sigma_s^2 \mathbf{Q}_s^{-1}$ and $\mathbf{C}_p = \sigma_p^2 \mathbf{Q}_p^{-1}$ of measurements errors associated with state variables and prior parameter values, respectively. Whereas it is possible to consider temporal correlations between pressure measurements in each monitoring interval, we presently treat them as being uncorrelated with zero mean and a uniform variance. We adopt a similar assumption with regard to log permeability measurements, disregarding spatial or cross-correlations between any of the data, thereby rendering $\mathbf{Q}_s$ and $\mathbf{Q}_p$ diagonal.

Since σ_s^2 and σ_p^2 are independent of $\log_{10}k$ and $\log_{10}\phi$ values (parameters) at the pilot points, minimizing (4) with respect to these latter parameters is equivalent to minimizing $\Phi = \Phi_s + \mu\Phi_p$ while treating $\mu = \sigma_s^2 / \sigma_p^2$ as an unknown. We perform this minimization using the regularization capability of PEST. In regularisation mode (Doherty 1994) PEST minimizes $\Phi_p^{\mu} = \mu\Phi_p$ subject to $\Phi_s \le \Phi_s^l$ (in practice $\Phi_s = \Phi_s^l$) where Φ_s^l is typically set by the user to a

value slightly higher than the minimum value of Φ_s obtained without regularization (i.e., upon setting $\mu = 0$). During each optimization step the program computes iteratively a value of μ (treating it as a reciprocal Lagrange multiplier) which insures that $\Phi_s = \Phi_s^l$ and then minimizes Φ_p^μ. We repeat the process for various Φ_s^l till *NLL* attains its minimum, yielding ML estimates of μ and the pilot point values.

A first-order approximation of the covariance $\boldsymbol{\Sigma}$ of parameter estimates $\hat{\mathbf{b}}$ is given by (Carrera and Neuman 1986)

$$\boldsymbol{\Sigma}(\hat{\mathbf{b}}) = \left[\frac{1}{\sigma_s^2} \mathbf{J}^T \mathbf{Q}_s \mathbf{J} + \frac{\mathbf{Q}_p}{\sigma_p^2} \right]^{-1}_{\mathbf{b}=\hat{\mathbf{b}}} \tag{6}$$

where **J** is a Jacobian matrix. If the estimate $\hat{\mu}$ of μ is optimal (as we take it to be) then ML estimates of the nominal variances are given by $\hat{\sigma}_s^2 = \Phi_s\left(\hat{\mathbf{b}}\right)/\left(N_s + N_p\right)$ and $\hat{\sigma}_p^2 = \hat{\sigma}_s^2/\hat{\mu}$. An alternative (not employed here) would be to specify $\hat{\mu}$, compute $\hat{\mathbf{b}}$ by minimizing $\Phi = \Phi_s + \hat{\mu}\Phi_p$, obtain ML estimates of the nominal variances according to $\hat{\sigma}_s^2 = \Phi_s\left(\hat{\mathbf{b}}\right)/N_s$ and $\hat{\sigma}_p^2 = \Phi_p\left(\hat{\mathbf{b}}\right)/N_p$, recompute $\hat{\mu} = \hat{\sigma}_s^2/\hat{\sigma}_p^2$ and repeat the process till *NLL* attains its minimum (Carrera and Neuman 1986).

CALIBRATION OF AIR FLOW MODELS

Elsewhere we have tested three approaches to the calibration of air-flow models with and without prior information (Morales-Casique et al. 2008). Here we focus on the use of prior information during the calibration process. We calibrate $\log_{10}k$ and $\log_{10}\phi$ at 32 pilot points against observed pressures, fixing variogram parameters from Tables 2 and 3, including 18 measurements of $\log_{10}k$ at pilot points as priors in Φ_p and incorporating the remaining 166 of $\log_{10}k$ values in the kriging process. The kriged $\log_{10}k$ field is based on three alternative variogram models *E1*, *E* and *P*, while the kriging of $\log_{10}\phi$ is based only on *E*. We calibrate the model jointly against pressure data from cross-hole tests PP4 and PP5. As noted earlier, we set $\mathbf{Q}_s = \mathbf{I}$ and $\mathbf{Q}_p = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Computed pressures are compared with measured values during each test in Figures 4 and 5. Overall, the calibrated models fit the observed data reasonably well in most intervals.

Table 4 shows the results of calibrating our models jointly against pressure data from cross-hole tests PP4 and PP6. In terms of *NLL* the best fit was obtained with log *k* variogram model *P* and the worst with model *E1*. Whereas *AIC*, *AICc* and *BIC* rank the models in this same order, *KIC* ranks *E1* higher than *E*. Posterior probabilities based on *AIC*, *AICc* and *BIC* are similar and so we list only those corresponding to *BIC* and *KIC*. Using Ockam's window leads to a preference for *P* at the virtual exclusion of the remaining two models regardless of which criterion is used. Using a variance window ($\alpha = 0.049$ which corresponds to a variance window of size $4\sigma_D$ and a significance level of 5%) also leads to a similar preference for *P* by *BIC* but a less pronounced preference for this model by *KIC*. Below we use both sets of posterior probabilities obtained with *KIC* to test the abilities of individual models, and MLBMA, to predict pressures observed during cross-hole tests PP6 and PP7.

Table 4 Results of joint calibration of cross-hole tests PP4 and PP5.

Model	*E1*	*E*	*P*
Parameters at pilot points	64	64	64
Variogram parameters for $\log_{10}k$ and $\log_{10}\phi$ (fixed)	8	4	4
Number of pressure data	462	462	462
NLL	2451	2305	2176
AIC	2579	2433	2304
Rank *AIC*	3	2	1
AICc	2599	2454	2324
Rank *AICc*	3	2	1
BIC	2846	2701	2571
Rank *BIC*	3	2	1
KIC	2702	2725	2673
Rank *KIC*	2	3	1
P_{BIC} %, $\alpha = 1$	2E-58	6E-27	99.99
P_{KIC} %, $\alpha = 1$	6E-05	5E-10	99.99
P_{BIC} %, $\alpha = 0.049$	0.11	3.91	95.99
P_{KIC} %, $\alpha = 0.049$	27.81	15.70	56.50

Model selection criteria: AIC = Akaike; AICc = Modified Akaike; BIC = Bayesian; KIC = Kashyap.
P_{IC} = posterior probability based on model information criteria *IC* for a given variance window (α = 1 corresponds to Occam's window).

PREDICTION OF PRESSURES DURING CROSS-HOLE TESTS PP6 AND PP7

Air injection during cross-hole tests PP6 and PP7 (Illman et al. 1998) took place into different intervals, and at different rates, than those in tests PP4 and PP5 (Table 1). Inverse calibration against pressure data from the latter two tests yielded ML estimates $\hat{\mathbf{b}}$ of the parameters and a covariance matrix of the corresponding estimation errors (6). To obtain corresponding statistics of the state variable, in this case air-pressure, one must either linearize the flow equation or solve it for numerous random realizations of the parameter vector $\mathbf{b}$ about

its ML estimate $\hat{\mathbf{b}}$. We have chosen the second option and conducted Monte Carlo simulations assuming the estimation error $\left(\hat{\mathbf{b}}-\mathbf{b}\right)$ to be multivariate Gaussian with zero mean and covariance $\boldsymbol{\Sigma}(\hat{\mathbf{b}})$ in the vicinity of $\hat{\mathbf{b}}$. This allowed us to generate random realizations of $\mathbf{b}$ using standard methods such as Cholesky factorization of $\boldsymbol{\Sigma}(\hat{\mathbf{b}})=\mathbf{U}^T\mathbf{U}$ followed by random draws of $\mathbf{b}=\mathbf{U}\boldsymbol{\zeta}$ where $\boldsymbol{\zeta}$ is a vector of standard uncorrelated normal variables (Clifton and Neuman 1982). Following this procedure we have generated 150 realizations of the parameter vector and solved the forward problem for each of them. In some cases the nonlinear solver failed to converge; the corresponding partial results were discarded. Our results are thus based on 119, 67 and 97 MC runs with models *E1*, *E* and *P*, respectively for test PP6 is based on and 104, 62 and 92 runs for test PP7. In addition to predicting pressure with individual models, we also generated MLBMA predictions by (1) and (2) based on posterior model probabilities in Table 5 obtained with a variance window.

Figures 6 and 7 compare predicted pressures averaged over all MC simulations against observed pressure for cross-hole tests PP6 and PP7. Each plot includes average predicted pressure from models *E1*, *E* and *P* plus the MLBMA estimate. For some data records average predicted pressure is close to the observed data; in other cases the prediction is poor, particularly at the injection interval (Z32 for PP6 and W32 for PP7) where models *E* and *P* over-predict pressure by orders magnitude while model *E1* under-predicts it. In addition, prediction is poor for all models at interval X1 in test PP7 (Figure 7), were observed pressure shows a large pressure response to injection in interval W32; evidence of this connectivity was absent in the calibration tests PP4 and PP5, and thus was not captured in the estimated parameters. We attribute this poor prediction in part to the extreme heterogeneity of the fractured tuff at the site and our disregard of barometric pressure fluctuations during the tests. We also predicted pressure

for both tests, PP6 and PP7 based on a single model run with the best parameter estimates $\hat{\mathbf{b}}$. Predicted pressures from a single run constitute a biased estimate of the ensemble mean pressure and do not provide information about the variance of the estimate. The results are shown in Figures 8 and 9 for tests PP6 and PP7, respectively. As before the prediction is poor at the injection intervals Z32 for PP6 and W32 for PP7, and at X1 for PP7, but now all models consistently under predict pressure at those intervals. Table 5 compares both estimates of pressure based on the sum of the squared errors SSE. Average predicted pressure based on MC simulations leads to one model clearly outperforming the other two by orders of magnitude. Results from a single run on the other hand show SSEs of the same order of magnitude. Excluding intervals with poor predictions (marked as B in Table 5) leads to model *P* being the most accurate in test PP6 and model E1 in test PP7. MLBMA is second in test PP6 while is third (MC simulations) and first (single run) in test PP7.

Table 5 Sum of squared errors (SSE)

Test	Prediction Method	*Option*	*E1*	*E*	*P*	*MLBMA*
PP6	MC simulations	A	5.86.E+03	4.24.E+06	4.07.E+05	4.48.E+05
		B	14.29	6.28	3.96	5.78
	Single run with $\hat{\mathbf{b}}$	A	1.25.E+04	1.26.E+04	1.22.E+04	1.23.E+04
		B	24.44	23.96	15.45	18.23
PP7	MC simulations	A	3.81.E+03	2.71.E+05	5.85.E+10	1.87.E+10
		B	43.43	1332.55	68.64	69.50
	Single run with $\hat{\mathbf{b}}$	A	4.37.E+03	2.74.E+03	4.45.E+03	4.04.E+03
		B	47.04	48.34	49.45	37.74

A – Includes all data records; B – Excludes records from Z32 in PP6 and W32 and X1 in PP7

We evaluate the predictive capabilities of each model and of MLBMA by computing their log scores and predictive coverage. The log score is computed by (4) with $\hat{D}_i^v$ and σ_i^2 are

the *i*-th sample mean and variance of predicted pressure based on MC realizations of the parameter estimates $\hat{\mathbf{b}}_k$ for model M_k. The predictive log score of MLBMA is (Ye et al. 2008)

$$-\ln p(\mathbf{D}^v \mid \mathbf{D}^c) = -\ln \sum_{k=1}^{K} p(\mathbf{D}^v \mid M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}^c)\, p(M_k \mid \mathbf{D}^c) \tag{7}$$

Table 6 lists the predictive log score of each model and MLBMA based on the variance window approach for both validation tests PP6 and PP7. Overall model *E1* has the lowest log score of the models and MLBMA for both validation tests, despite being ranked second by *KIC* and third, with 0.1% posterior probability, by *BIC* (Table 4). The main source of predictive error for model *E1* is the injection interval (Z32) in test PP6, while for test PP7 the main source are intervals X1 (large predictive errors) and Z1 (very small variance, $\sigma^2_{Z1} \sim 10^{-9}$ and significant predictive error, thus the log score penalizes it). For the remaining models and MLBMA the ranking changes for each validation test; MLBMA ranks second and third in test PP6 and PP7, respectively. The largest log score for MLBMA and models *P* and *E* comes from the injection intervals (Z32 in test PP6 and W32 in test PP7) and X1 for test PP7 where these models and MLBMA have large prediction errors (Figures 6 and 7). Excluding low prediction intervals (Z32 in PP6 and W32, Z1 and X1 in PP7, results denoted by Total B in Table 6) model *E1* ranks first in test PP6 and last in PP7; in turn, MLBMA ranks second and first in PP6 and PP7, respectively.

Table 6 Predictive log scores for validation tests PP6 and PP7

	PP6				PP7			
	E1	E	P	MLBMA	E1	E	P	MLBMA
Total	3.86E+03	1.33E+05	7.31E+04	7.12E+04	5.69E+05	1.17E+06	1.85E+10	1.76E+10
Rank	1	4	3	2	1	2	4	3
Total B	302	7877	2612	613	17466	9627	6563	3828
Rank B	1	4	3	2	4	3	2	1
X1	12.10	228.61	29.92	21.87	262902	160764	860129	230284
X21	3.15	12.37	2.92	2.96	12.11	9.98	7.76	7.46
X22	8.61	72.14	30.98	17.72	6.54	8.93	23.78	16.61
X23	3.87	4.35	4.34	4.11	7.75	51.31	7.88	15.70
X3	0	0	0	0	4.94	13.05	5.70	6.06

Y12	0	0	0	0	0	0	0	0
Y21	5.80	7.08	5.88	5.78	138.52	3.17	102.60	11.07
Y22	17.21	1051.60	8.92	9.37	312.11	1584.71	421.18	339.41
Y23	9.69	199.73	37.57	20.52	13.14	3.16	41.77	5.62
Y31	2.00	1.88	1.87	1.88	30.90	2.87	14.81	3.88
Y32	0	0	0	0	3.61	3.29	5.44	2.86
Y33	4.41	4.14	4.44	4.36	5.89	18.46	10.87	7.93
Z1	0	0	0	0	288598	3.88	11.07	4.52
Z21	5.06	4.18	4.28	4.45	74.98	4.27	10.25	7.47
Z22	0	0	0	0	12.96	4.17	4.35	3.78
Z23	0	0	0	0	9.44	4.43	4.05	3.80
Z24	0	0	0	0	8.48	6.31	3.95	4.18
Z31	25.72	44.68	116.32	79.80	40.56	7.72	13.08	13.02
Z32	3560	124722	70499	70538	48.65	108.15	8.63	11.44
Z33	0	0	0	0	16158.55	12.16	5.73	5.94
V1	11.03	41.15	11.22	11.61	7.86	7.30	16.54	8.69
V22	23.53	1299.36	310.17	65.82	89.17	973.72	772.89	199.12
V31	33.60	404.82	134.49	74.37	82.83	59.35	48.71	56.15
V32	57.17	3910.46	1638.93	180.89	125.58	2345.18	1420.19	307.88
V33	5.28	18.04	16.15	9.00	10.30	16.12	98.75	6.67
W1	4.82	339.52	64.06	9.22	211.10	632.94	221.89	233.96
W2A1	5.15	38.06	23.25	11.43	0	0	0	0
W2A2	9.58	145.30	67.56	21.93	7.36	6.24	13.55	7.95
W2A3	5.33	12.84	10.44	6.97	3.85	4.94	11.81	3.92
W2A4	3.12	4.69	3.14	3.12	0	0	0	0
W31	46.00	31.74	85.12	45.39	39.21	3734.80	3267.21	2537.42
W32	0	0	0	0	158.01	997467	1.847E+10	1.762E+10

Total B and Rank B correspond to predictive log scores excluding low prediction intervals (Z32 in PP6 and W32, Z1 and X1 in PP7).

Another measure of the predictive capabilities of a model is its predictive coverage, the percentage of observed data that fall within a given prediction interval around average predicted pressure. Prediction intervals were computed by confidence intervals assuming the errors are normally distributed and for a confidence level of 95%. Table 7 lists the results obtained based on 776 observed data for test PP6 and 829 for test PP7. Among individual models, model *P* has the best predictive coverage for test PP6, while for test PP7 it is second to model *E*. While MLBMA has a superior predictive coverage than any of the three individual models for test PP7, it is second to last for test PP6. Excluding as before low prediction intervals (Z32 in PP6 and

W32, Z1 and X1 in PP7) increases the predictive coverage of the models and MLBMA but does not change the rankings.

Table 7 Predictive coverage for validation tests PP6 and PP7

Interval	PP6				PP7			
	E1	E	P	MLBMA	E1	E	P	MLBMA
Total	5.54	9.54	10.16	7.63	6.13	18.04	12.96	31.48
Rank	4	2	1	3	4	2	3	1
X1	0.13	0	0.13	0	0	0	0	0
X21	16.62	0.13	46.39	20.49	1.81	93.24	0	8.08
X22	0.77	0	0.39	0.13	6.51	94.09	99.03	98.31
X23	14.30	14.82	17.40	12.76	0	0	0.36	93.61
X3	0	0	0.13	0	24.00	0.24	7.60	95.54
Y12	0	0	0	0	0.84	2.17	1.69	1.81
Y21	3.61	0	29.38	9.41	0	5.79	0	1.81
Y22	0.39	0	1.68	1.03	0	0.24	0.24	0.24
Y23	0.52	0.13	0.26	0.64	0.24	34.86	0.36	0.97
Y31	0.64	27.06	23.84	20.88	3.26	73.46	0.48	7.36
Y32	0	4.25	0.39	0.26	15.68	0.12	11.94	60.31
Y33	0	0	0	0	80.10	0.12	1.45	91.31
Z1	0	4.12	0	0	0	61.88	0.24	5.07
Z21	0	40.08	0	0	1.21	12.42	0.72	1.21
Z22	12.50	41.24	0	14.56	4.22	0	37.15	49.70
Z23	23.20	61.86	54.38	48.84	1.69	0	24.61	57.54
Z24	2.84	66.11	56.19	50.77	2.17	0	30.40	41.62
Z31	1.80	25.64	60.44	39.43	1.21	6.76	3.14	2.77
Z32	0	0	0.13	0.13	0.60	0.48	3.86	3.62
Z33	1.93	0	0	0	0	0	13.51	62.85
V1	1.29	1.29	1.16	2.06	2.90	6.15	0.84	3.14
V22	0	0	0	0	0.36	0.60	0.12	0.48
V31	0.13	0	0	0	0.36	0.36	0.60	0.36
V32	0	0	0	0	0	0.24	0.36	0.36
V33	0.39	0.39	0.13	0.52	0.72	79.01	0.36	21.23
W1	0.90	0.13	0	0.26	0.24	0.12	0.24	0.24
W2A1	33.25	15.85	14.82	1.16	11.70	61.76	97.47	94.21
W2A2	0	0	0	0	2.05	6.76	0.48	2.17
W2A3	2.71	0.26	0.13	0.26	22.32	2.65	3.26	63.57
W2A4	58.89	0.90	8.89	19.07	10.98	33.66	2.29	40.65
W31	0.26	0.52	0	0.26	0.97	0	72.01	97.23
W32	0.13	0.39	8.89	1.16	0	0	0	0

CONCLUSIONS

We have shown that it is possible to employ MLBMA in complex models, illustrated how to include prior information and applied the method to air flow models in unsaturated tuff. We calibrate $\log_{10}k$ and $\log_{10}\phi$ at selected pilot points against observed pressures in two pneumatic injection tests (PP4 and PP5) and including prior information about $\log_{10}k$. All of the calibrated models reproduce satisfactorily the observed data set. We computed model discrimination criteria and used them to compute posterior model probabilities based on Occam's window and a broader window variance. The first approach leads to selecting with probability of almost 1 the model with the lowest fitting error and neglecting the remaining models. When a variance window of $4\sigma_D$ is employed, this leads to more equilibrated posterior probabilities as long as these are computed using *KIC*. When posterior probabilities are computed using *AIC*, *AICc*, or *BIC* even the use of a variance window lead to the best model being assigned posterior probability of almost 1.

The results of the calibration were validated against an independent data set consisting of two cross-hole tests (PP6 and PP7) were injection took place in a different location than the data set used in the calibration. During this stage the model ranked second by *KIC*, and discarded by *AIC*, *AICc*, or *BIC*, yielded the most accurate results. Since only *KIC* based posterior probabilities recognized that no model was clearly dominant supports the asseveration by Ye et al. (2008) that *KIC* better accounts validly for the likelihood of prior parameter estimates and has the ability to discriminate between models based not only on their number of parameters and sample size but also on how close are the posterior parameter estimates to their prior values and how much expected information is contained, on average, in each observation.

We also evaluated the predictive capabilities of MLBMA based on tests PP6 and PP7. Predicted pressures using MLBMA were less accurate than some individual models, due to the fact that the individual model with the largest posterior probability was the worst or second worst predictor in both validation cases. In terms of predictive coverage, MLBMA was far superior to any of the individual models in one validation test and second to last in the other validation test.

We attribute the mixed results obtained to the fact that the medium is highly heterogeneous, hydrologic parameters depend not only on spatial location but also on the flow regime/pattern and that the model space used in our test was under-sampled, meaning that other plausible descriptions of spatial distribution of parameters were not included.

Acknowledgements This work was supported jointly by U.S. National Science Foundation grant EAR-0407123 and the U.S. Department of Energy through a contract with Vanderbilt University under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III.

REFERENCES

Akaike H (1974) A new look at statistical model identification. *IEEE Trans. Autom. Control,* AC-19: 716– 722.

Beven K (2006) A manifesto for the equifinality thesis. *J. Hydrol.* 320:18– 36.

Carrera J, Neuman SP (1986) Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resour. Res.* 22:199–210.

Clifton P, Neuman SP (1982) Effects of kriging and inverse modeling on conditional simulation of the Avra valley aquifer in southern Arizona. *Water Resour. Res.* **18** (4):1215–1234.

de Marsily G, Lavedan C, Boucher M, Fasanino G (1984) Interpretation of interference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model. In: Verly G, David M, Journel AG, Marechal A (ed), *Geostatistics for Natural Resources Characterization*, NATO ASI Ser., Ser. C, 182, 831-849.

Di Federico V, Neuman SP (1997) Scaling of random fields by means of truncated power variograms and associated spectra. *Water Resour. Res*. **33**(5), 1075–1086.

Doherty J (1994) PEST Model-Independent Parameter Estimation, User Manual: 5th Edition. *Watermark Numerical Computing.*

Guzman AG, Geddis AM, Henrich MJ, Lohrstorfer CF, Neuman SP (1996) Summary of Air Permeability Data From Single-Hole Injection Tests in Unsaturated Fractured Tuffs at the Apache Leap Research Site: Results of Steady-State Test Interpretation, NUREG/CR-6360, *U.S. Nucl. Regul. Comm*., Washington, D. C.

Hernandez AF, Neuman SP, Guadagnini A, Carrera J (2006) Inverse stochastic moment analysis of steady state flow in randomly heterogeneous media. *Water Resour. Res*. **42**, W05425, doi:10.1029/2005WR004449.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. *Stat. Sci.* **14**(4):382–417.

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small sample. *Biometrika*, 76(2):99–104.

Illman WA (2005) Type curve analyses of pneumatic single-hole tests in unsaturated fractured tuff: Direct evidence for a porosity scale effect. *Water Resour. Res*. **41**, W04018, Doi:10.1029/2004WR003703.

Illman WA, Neuman SP (2001) Type curve interpretation of a cross-hole pneumatic injection test in unsaturated fractured tuff. *Water Resour. Res*. **37**(3):583–604.

Illman WA, Thompson DL, Vesselinov VV, Chen G, Neuman SP (1998) Single- and Cross-Hole Pneumatic Tests in Unsaturated Fractured Tuffs at the Apache Leap Research Site: Phenomenology, Spatial Variability, Connectivity and Scale. NUREG/CR-5559, *U.S. Nucl. Regul. Comm*. Washington, D. C.

Kashyap RL (1982) Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4(2):99-104.

Morales-Casique E, Neuman SP, Vesselinov VV (2008) Maximum likelihood Bayesian averaging of air flow models in unsaturated fractured tuff. In: Refsgaard et al. (ed) Calibration and Reliability in Groundwater Modelling: Credibility of Modelling, Proceedings of ModelCARE 2007 Conference held in Denmark, September 2007, *IAHS Publ*. 320, 70-75.

Neuman SP (2003) Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models. *Stochastic Environ. Res. Risk Assess*. **17**(5):291– 305, doi:10.1007/s00477-003-0151-7.

Neuman SP, Wierenga PJ (2003) A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites. NUREG/CR-6805, U.S. Nucl. Regul. Comm., Washington, D. C.

Pebesma EJ, Wesseling CG (1998) Gstat: a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences* **24**(1):17-31

Poeter EP, Anderson DA (2005) Multimodel ranking and inference in ground water modeling, *Ground Water* 43(4):597–605.

Refsgaard JC, van der Sluijs JP, Brown J, van der Keur P (2006) A framework for dealing with uncertainty due to model structure error. *Adv. Water. Resour.* 29:1586–1597.

Samper FJ, Neuman SP (1989) Estimation of spatial covariance structures by adjoint state maximum likelihood cross validation: 1, Theory. *Water Resour. Res.*, 25(3):351–362.

Schwarz G (1978) Estimating the dimension of a model. *Ann. Stat.* 6(2):461– 464.

Tsai FTC, Li X (2008) Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. *Water Resour. Res.*, 44, W09434, doi:10.1029/2007WR006576.

Vesselinov VV, Neuman SP, Illman WA (2001a) Three-dimensional numerical inversion of pneumatic cross-hole tests in unsaturated fractured tuff 1. Methodology and borehole effects. *Water Resour. Res*. **37**(12):3001–3018.

Vesselinov VV, Neuman SP, Illman WA (2001b) Three-dimensional numerical inversion of pneumatic cross-hole tests in unsaturated fractured tuff 2. Equivalent parameters, high-resolution stochastic imaging and scale effects. *Water Resour. Res*. **37**(12):3019–3042.

Volinsky CT, Madigan D, Raftery AE, Kronmal RA (1997) Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Applied Statistics* 46(4):433-448.

Ye M, Neuman SP, Meyer PD (2004) Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resour. Res*. **40**, W05113, doi:10.1029/2003WR002557.

Ye M, Meyer PD, Neuman SP (2008) On model selection criteria in multimodel analysis. Water Resour. Res. 44, W03428, doi:10.1029/2008WR006803.

Zyvoloski GA, Robinson BA, Dash ZV, Trease LL (1999) Models and methods summary for the FEHM application: A finite-element heat- and mass-transfer code. SC-194, *Los Alamos Natl. Lab*., Los Alamos, N. M.

List of Figures

Fig. 8 Pressure buildup (kPa) versus time (days) during cross-hole test PP6. Observed data = small circles. Predicted results from a single model run with the best parameters, for variogram models: *E1* = squares, *E* = triangles, *P* = big circles. Empty circles correspond to MLBMA prediction.

Fig. 9 Pressure buildup (kPa) versus time (days) during cross-hole test PP7. Observed data = small circles. Predicted results from a single model run with the best parameters, for variogram models: *E1* = squares, *E* = triangles, *P* = big circles. Empty circles correspond to MLBMA prediction.