Stoch Environ Res Risk Assess (2010) 24:917–923
DOI 10.1007/s00477-010-0386-z

ORIGINAL PAPER

Normalized Mahalanobis distance for comparing process-based stochastic models

C. L. Winter

Published online: 26 March 2010

Abstract We investigate a method based on normalized Mahalanobis distance, *D*, for comparing the performance of alternate stochastic models of a given environmental system. The approach is appropriate in cases where data are too limited to calculate either likelihood ratios or Bayes factors. Computational experiments based on simulated data are used to evaluate *D*'s ability to identify a "true" model and to single out good models. Data are simulated for two populations with different "signal–noise ratios" (*S*/*N*) The expected value of *D* is decomposed to evaluate the effects of normalization, model bias, and model correlation structure on *D*'s discriminatory power. Normalization compensates for the advantage one model may have over another due to technical features of its hypothesized correlation structure. The relative effects of bias and correlation structure vary with *S*/*N*, model bias being most important when *S*/*N* is relatively high and correlation structure increasing in importance as *S*/*N* decreases.

Keywords Model comparison · Mahalanobis distance · Model uncertainty

1 Introduction

Policy-makers, natural resource managers, environmental scientists, economists and stakeholders increasingly rely on computational models to analyze and manage environmental systems. Models are commonly used to understand environmental systems, to allocate local and regional resources, to trade off competing goods like biodiversity versus economic returns, and to assess the risks and impacts of global effects like climate change. Since models of an environmental system only approximate reality, they can, and do, take on multiple forms, which naturally raises the question of how to compare them. Although an extensive statistical literature on model comparison methods exists (for instance, Akaike 1973; Schwarz 1978; Spiegelhalter et al. 2002; Neuman 2003), those methods generally require strong assumptions about the forms of probability distributions for observed data and system parameters that can be difficult to justify on the basis of the limited information usually available (Christakos 2003). This is especially true when models are based on detailed descriptions of environmental systems consisting of interacting physical, chemical, anthropogenic, and biological processes whose basic forms are given by coupled partial differential equations and related mathematical structures.

In this paper, *D*, a simple alternative metric based on Mahalanobis' distance criterion, is proposed for testing complicated models in data-poor settings. It is appropriate to settings where there is insufficient data to assume specific forms for Bayes factors (likelihoods) and statistics depending on them. Groundwater modeling, climate predictions, and environmental risk assessments are examples of settings that are frequently data-poor, where system variables are too sparsely sampled to specify probability distributions per se. In some ways *D* is the mirror image of Bayesian model comparisons. For instance, Akaike (1973) defines a test based on maximum likelihood, a strong modeling assumption, but applies it to a weak class of models: autoregressive-moving average models. On the

C. L. Winter (✉)
Department of Hydrology and Water Resources,
University of Arizona, Tucson, AZ, USA
e-mail: winter@email.arizona.edu

C. L. Winter
National Center for Atmospheric Research, Boulder, CO, USA

other hand, we investigate a weak test that can be applied to strong (i.e., complicated) models.

1.1 Example

The challenge of analyzing and predicting groundwater flow and contaminant transport provides a good example of the kinds of models we have in mind. In most cases, the states and parameters of a groundwater system are observed at only a few locations, and their complexity is usually summarized by representing system parameters and states as random fields with complex correlation structures. By now, multiple methods for characterizing and quantifying parametric uncertainty exist, while stochastic models of groundwater dynamics range from continuum approaches based on Darcy's Law and Fickian diffusion (Winter et al. 1984; Zhang and Neuman 1995; Guadagnini et al. 2003) to continuous time random walks (Berkowitz et al. 2006; Dentz et al. 2008), to fractional dynamics (Benson et al. 2004).

The complexity of a groundwater model is not exhausted by specifying its form. For instance, Tartakovsky and Neuman (1998) give conditional moment equations and recursive approximations for groundwater flow based on Darcy's Law. In one of the simpler versions of their model, the expected state of pressure head, $\psi(x, t)$, is approximated by a first-order perturbation expansion, $\psi(x, t) \approx \psi^{(0)}(x, t; K_G) + \psi^{(1)}(x, t; K_G, \rho_K)$, that depends on the geometric mean of hydraulic conductivity, $K_G(x)$, and its spatial correlation, $\rho_K(x_1, x_2)$. Here the model consists of the Darcian assumption, the perturbation approximation, the specific forms of $\psi^{(0)}(x, t; K_G)$ and $\psi^{(1)}(x, t; K_G, \rho_K)$, and the forms specified for K_G and ρ_K. The model can be further complicated by including groundwater storativity, forcing functions and initial and boundary conditions as additional stochastic parameters.

1.2 Second-moment models for stochastic environmental processes

Abstracting from the groundwater flow example, a stochastic environmental process, $Y(t) \sim (P_Y, \Pi_Y)$, consists of a family of unknown finite-dimensional distributions, $P_Y = \text{Prob}[Y(t_1) < y_1, \ldots, Y(t_N) < y_N]$, parameterized by a set of parameters, Π_Y, whose values are uncertain. Among the uncertain parameters of $Y(t)$ are its expected value, $\psi_Y(t)$, and covariance, $\rho_Y(t_1, t_2)$. Often all that is known about the system are (1) some combination of prior data, theory, and expert opinions that allow a set, M, of alternative models to be specified and (2) an independent set of observations, $Y^{Obs} = \{Y(t_1), \ldots, Y(t_n)\}$, available for validating model performance. Data may be too limited to estimate P_Y, or probability distributions for Π_Y, in which case strong assumptions must be made to compute likelihood ratios or Bayes factors. A weaker alternative is to specify a model $m \in M$ by means of its estimates of the system state, $\psi_m(t)$, and the state's covariance, $\rho_m(t_1, t_2)$, and then quantify uncertainty in terms of intervals or distances based on the hypothesized moments. That is the approach examined here. The model hypothesis, H_m, reduces to equality of the first two moments,

$$H_m : \psi_Y(t) = \psi_m(t) \text{and } \rho_Y(t_1, t_2) = \rho_m(t_1, t_2). \tag{1}$$

In this paper we do not propose a statistical test of H_m since that would require essentially the same data needed to compute likelihoods and Bayes factors. Instead, we define a metric in terms of ψ_m and ρ_m that can be used to test the distance between a model and observations.

1.3 Normed Mahalanobis distance

We examine a distance measure based on Mahalanobis' notion that the distance between Y^{Obs} and a corresponding vector of state estimates, should reflect ρ_m as well as the difference between Y^{Obs} and ψ_m^{Obs},

$$D\left(Y^{Obs}, \psi_m^{Obs}\right) = \sqrt{\left(\psi_m^{Obs} - Y^{Obs}\right)^T \underset{n\times n}{\rho_m^{-1}} \left(\psi_m^{Obs} - Y^{Obs}\right) + \log \underset{n\times n}{|\rho_m|}} \tag{2}$$

Here $\underset{n\times n}{\rho_m} = (\rho_m(t_1, t_2))$ is the symmetric $n \times n$ matrix of covariances corresponding to the observation points and $\underset{n\times n}{|\rho_m|}$ is its determinant. By including ρ_m in the calculation of the distance, we can distinguish between models that propose the same mean performance but have different correlation structures.

The first term under the square root is $MD_m^2 = \left(\psi_m^{Obs} - Y^{Obs}\right)^T \rho_m^{-1} \left(\psi_m^{Obs} - Y^{Obs}\right)$, the square of Mahalanobis' distance (Mahalanobis 1936). We drop the subordinate $n \times n$ from now on except where it is needed for clarity. The second term regularizes D so as not to penalize models that hypothesize covariances whose inverses are relatively well-aligned with $\psi_m^{Obs} - Y^{Obs}$.

1.4 Outline

The method of analysis in this paper is computational and analytical, consisting in the first place of evaluating hypothesis (1) for each of $m = 1, \ldots, 6$ models with respect to simulated data sets, $Y(t)$, in a one-dimensional space. The dimensionality does not significantly limit conclusions. The moments, $\psi_Y(t)$ and $\rho_Y(t_1, t_2)$, of $Y(t)$ are known, and one model is "true" in the sense that $\psi_m(t) = \psi_Y(t)$ and $\rho_m(t_1, t_2) = \rho_Y(t_1, t_2)$. This is, of course, impossible to arrange when dealing with real data. The data samples are drawn from one of two populations, each with a different "signal–noise ratio" (S/N): $S/N_1 = 4$, $S/N_2 = 1$. Here $S/N = \frac{\max_t |\psi_Y - \psi_Y(t)|}{\sigma_Y^2}$ is the ratio of the maximum departure

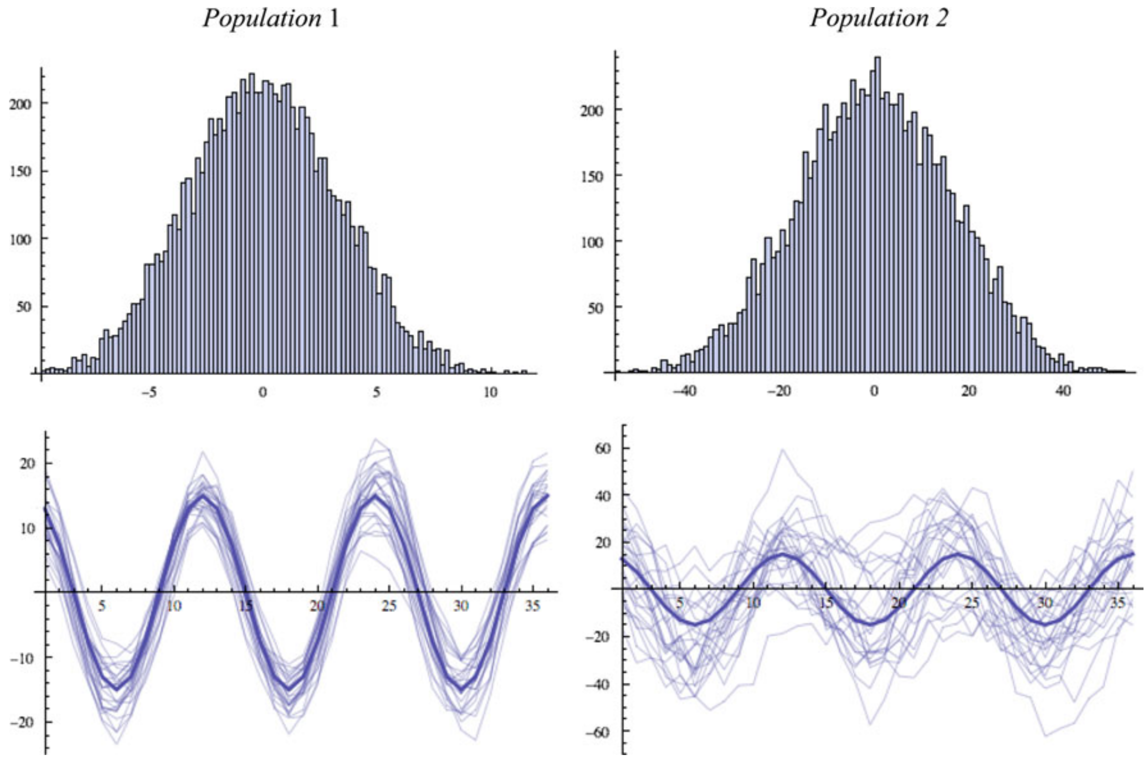

Fig. 1 Properties of the simulated samples. *Above* Histograms of the 9000 (= 25 × 360) sample values of Y'(t) from each population. Sample means and standard deviations are given in Table 1. *Below* The first 36 values for each member of the 25 simulations of $Y(t)$ by population plotted with the mean, $\psi_Y(t)$, which is in *bold*

of $\psi(t)$ from the grand mean over time, $\psi_Y = \frac{1}{T}\int_0^T \psi_Y(t)\mathrm{d}t$, to the variance of individual departures, $\sigma_Y^2 = E[(Y(t) - \psi_Y(t))^2]$. The individual departure variance is constant, $\sigma_Y^2(t) = \sigma_Y^2 = \text{const}$, due to the way we construct our examples. There are 25 samples of 360 observations for each population. In Sect. 2, we describe the method used to generate the sample data, and we define the different models in terms of their moments. Models are compared in Sect. 3 on the basis of the simulated data and also by analytically decomposing $E[D_m^2]$, the expected squared distance, into terms that depend on the normalization $(\log|\rho_m|)$, model bias $(\Psi_m(t) - Y(t))$, and correlation structure (ρ_m). Performance is related to S/N in Sect. 3 and also in Sect. 4, which summarizes the paper.

2 Simulated data and models

Computational examples are based on samples of 25 series from each of two different populations, Populations 1 and 2, that differ on their standard deviations (variances) and hence their signal–noise ratios. Each population is simulated by $Y(t) = \psi_Y(t) + Y'(t)$ where

$$E[Y(t)] = \psi_Y(t) = 15\cos[2\pi t/12] \tag{3}$$

is periodic with period 12 and amplitude 15. Such a function might represent the mean of monthly temperature in a region with an average annual range of 30°. Plots of histograms of $Y'(t)$ and variations about the mean $\psi_Y(t)$ are shown in Fig. 1.

The random departures have mean zero $E[Y'(t)] = 0$, and are stationary, $\rho_Y(t_1, t_2) = E[Y'(t_1)\ Y'(t_2)] = \rho_Y(|t_1 - t_2|)$. Their variances are, respectively, $\sigma_{Y_1} = 3.38$ $(\rho_{Y_1}(0) = 11.43)$ and $\sigma_{Y_2} = 16.48$ $(\rho_{Y_2}(0) = 271.43)$. Hence, Population 2 has a standard deviation that is about the same as the amplitude of the mean $(S/N_2 \approx 1)$, while the standard deviation of Population 1 is only about one quarter the mean amplitude $(S/N_1 \approx 4)$. The correlation structure of the two populations is otherwise the same (Fig. 2).

2.1 Simulated data

Realizations of $Y'(t)$ are generated by smoothing a sequence of independent, identically distributed uniformly random variables, U_t, defined on the intervals $-15 \leq U_{t_1} \leq 15$ and $-75 \leq U_{t_2} \leq 75$, respectively, for Populations 1 and 2. Hence, $E[U_i] = 0$ for $i = 1, 2$ and $\sigma_{U_1} = 8.94$, $\sigma_{U_2} = 43.59$. For each population, smoothing is accomplished by applying a uniform kernel, $k(t) = 1$

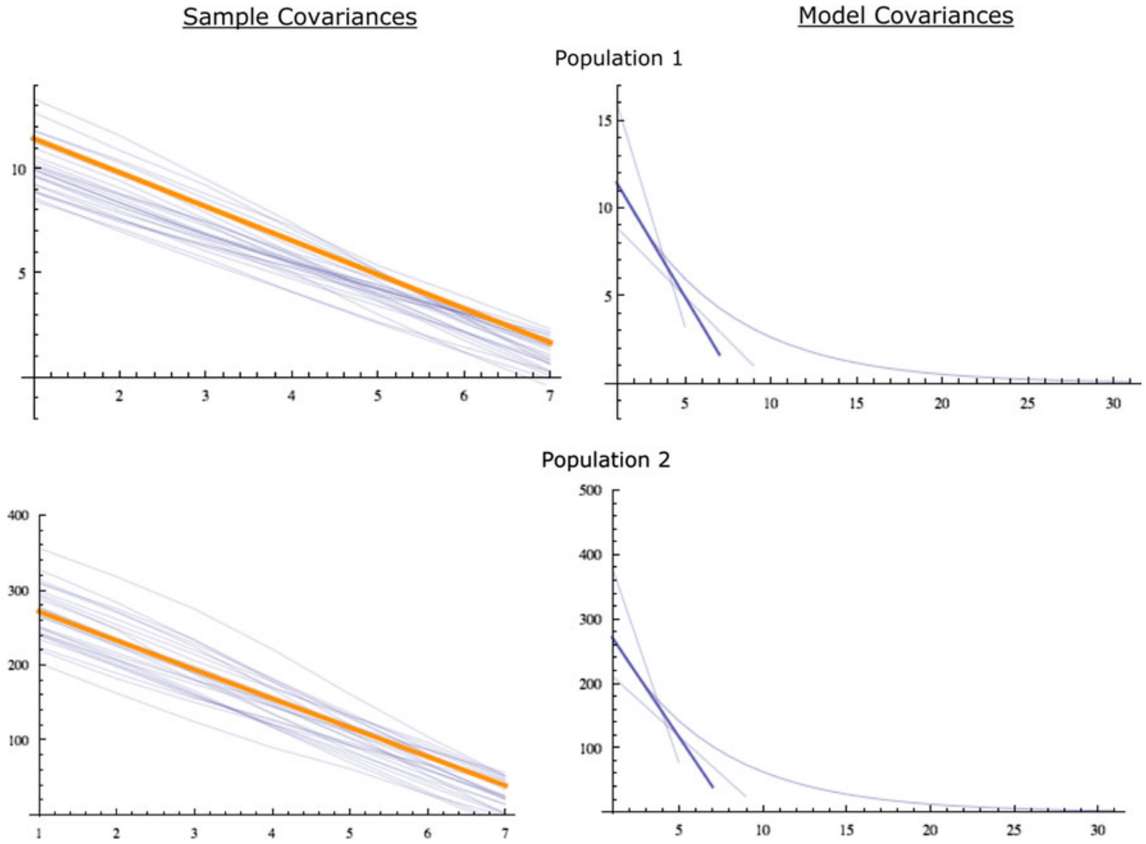

Fig. 2 Covariances. *Left* Sample covariances plotted against the population covariance (in *bold*). ρ_{Y_1} above and ρ_{Y_2} below. *Right* Model covariances against the population covariance, again in *bold*. Note the overlap in the initial part of the range between ρ_Y and the exponential covariance, ρ_4

Table 1 Statistics of the simulated samples

	$\frac{1}{2}\sum_i Y_i^{Obs}$	σ_i^2	$\max_i\{\sigma_i^2\}$	$\min_i\{\sigma_i^2\}$	Mean Log Det
Sample set 1	0.01	10.22	13.34	8.44	8.37
Sample set 2	−0.05	268.93	355.861	202.12	31.1

if $-3 \le t \le 3$ and $k(t) = 0$ otherwise to the uniform variates so that

$$Y'(t) = \frac{1}{7}\sum_{\tau=t-3}^{t+3} U_\tau \tag{4}$$

The result is a sample drawn from a correlated random variable that is approximately normally distributed (Fig. 1) with covariance

$$\rho_Y(|\Delta t|) = \begin{cases} (7-|\Delta t|)\sigma_U^2/49 \\ 0 \quad \text{otherwise} \end{cases} \tag{5}$$

As noted, computational results are based on sets, $\{Y_i^{Obs}\}$, from Populations $i = 1$ or 2 of 25 simulated sample series, each series consisting of 360 elements, $Y_i^{Obs} = \{Y_{i,1}^{Obs}, \ldots, Y_{i,360}^{Obs}\}$. Table 1 gives basic sample statistics across all 25 series for each population. The last column is $\frac{1}{25}\sum_i \log|\rho_{Y_i}|$, the mean over the 25 samples Y_i^{Obs}. Sample covariances are plotted with $\rho_Y(|\Delta t|)$ from (Eq. 5) on the right side of Fig. 2.

2.2 Model suite

To investigate the properties of D computationally, we want a set of models that includes the "true" model and brackets it with other models illustrating the full range of behavior the terms in (2) can show. We also want the models to include forms for the covariance, like the exponential, that are in common use. Hence the model suite in this paper includes biased and unbiased models, a range of models with linear covariances like ρ_Y, and one with an exponential covariance (see Table 2). As noted, model forms are identical for Populations 1 and 2, except for their

Table 2 Sample statistics

	Model					
	1 (True)	2	3	4	5	6
Mean	Unbiased	Unbiased	Unbiased	Unbiased	$\bar{Y}(t)/2$	0.00
ρ_m	Length = 7	Length = 5	Length = 9	Exponential	Length = 7	Length = 7
Population 1						
Mean (SDM_m) Mahalanobis Distance	18.35 (0.35)	58.22 (12.03)	88.24 (14.41)	17.97 (0.63)	27.98 (0.84)	46.20 (0.96)
Mean D_m (SD$_m$)	23.22 (0.28)	61.98 (11.35)	88.40 (14.39)	27.23 (0.42)	31.29 (0.75)	48.34 (0.92)
Maximum D_m	23.76	85.16	127.29	28.51	32.95	50.37
Minimum D_m	22.55	46.26	67.42	26.36	30.15	46.78
Population 2						
Mean (SDM_m) Mahalanobis distance	18.76 (0.42)	66.92 (16.77)	88.70 (20.18)	18.37 (0.72)	19.22 (0.44)	20.61 (0.53)
Mean D_m (SD$_m$)	41.17 (0.19)	78.29 (14.30)	95.31 (18.90)	43.55 (0.30)	42.04 (0.26)	
Maximum D_m	41.54	104.32	153.03	44.29	41.78	42.55
Minimum D_m	40.86	56.96	66.09	43.04	41.03	41.52

standard deviations. Model 1 is the true model, $\psi_1 = \psi_{\mathrm{Y}}$ and $\rho_1 = \rho_Y$. Models $m = 2, 3, 4$ are unbiased, $\psi_{\mathrm{m}} = \psi_{\mathrm{Y}}$, while models $m = 5, 6$ have the same correlation structure as $Y(t)$, $\rho_{\mathrm{m}} = \rho_{\mathrm{m}}$, but are biased, $\psi_5(t) = \psi_Y(t)/2$ and $\psi_5(t) = 0$. For both populations, model 4 has an exponential covariance that was obtained by fitting $\rho_4(t) = \sigma_Y^2 \exp(-\alpha)$ to ρ_Y with $\alpha = 0.16$. Models 2 and 3 have linear covariances that are defined by kernels with lengths $l = 5$ or 9, respectively. Thus, their variances and covariances follow (6), except l and l^2 substitute for 7 and 49. These are the correct hypotheses for uniform smoothing kernels with these lengths. The results, incidentally, are unaffected if the covariances for these models are adjusted to variances of $\rho_{Y_1}(0) = 11.43$ and $\rho_{Y_1}(0) = 271.43$. The model covariances are shown on the right of Fig. 2.

3 Results and discussion

Computational results based on the 25 samples for each population are used to compare the ordinary Mahalanobis distance, MD_m, to the normalized distance, D_m (Table 2). We also decompose $E[D_m^2]$ into components that quantify effects of the correlation structure, ρ_m versus ρ_Y; the expected bias of ψ_m as weighted by ρ_m^{-1}; and the normalization effected by $\log|\rho_m|$ (Table 3).

3.1 Sample results

The columns of Table 2 correspond to models $m = 1, \ldots, 6$. For each population, the average Mahalanobis distance for model m,

Table 3 Decomposition of $E\left[D_m^2\right]$ along with $E\left[MD_m^2\right]$

	1	2	3	4	5	6
Population 1						
Trace	360.00	4507.04	8288.02	342.95	360.00	360.00
Bias	0.00	0.00	0.00	0.00	452.65	1810.60
Norm	202.16	437.49	27.42	418.46	202.16	202.16
$E\left[D_m^2\right]$	23.71	70.32	91.19	27.59	31.86	48.71
$E\left[MD_m^2\right]$	18.97	67.13	91.04	18.52	28.51	46.59
Population 2						
Trace	360.00	4507.04	8288.02	342.95	360.00	360.00
Bias	0.00	0.00	0.00	0.00	19.05	76.24
Norm	1342.50	1577.82	1167.74	1558.78	1342.50	1342.50
$E\left[D_m^2\right]$	41.26	78.00	97.24	43.61	41.49	42.17
$E\left[MD_m^2\right]$	18.97	67.13	91.04	18.52	19.47	20.89

$$\begin{aligned}\overline{MD}_m &= \frac{1}{25}\sum_{i=1}^{25} MD_{m,i} \\ &= \frac{1}{25}\sum_{i=1}^{25}\sqrt{(\psi_m^{Obs} - \psi_i^{Obs})^T \rho_m^{-1} (\psi_m^{Obs} - \psi_i^{Obs})} \qquad (6)\end{aligned}$$

is summed over Mahalanobis distances, $MD_{m,i}$, calculated for each member, Y_i^{Obs}, of the respective population samples, $I = 1, \ldots, 25$. The overbar indicates a sample average. The standard deviation of the averages,

$$\mathrm{SD}M_m = \sqrt{\frac{1}{24}\sum_i \left(MD_{m,i} - \overline{MD}_m\right)^2}, \qquad (7)$$

is given in parentheses. The mean normalized Mahalanobis distances, $\bar{D}_m$, their standard deviations, SD$_m$, maxima, and minima are calculated likewise.

The first point worth noting is that both D and MD are effective overall discriminators for each population. In both cases, they separate the models into one class that fits the data well, and another that does not: the good-fitting classes are Population 1, $m = 1, 4, 5$ and Population 2, $m = 1, 4, 5, 6$. The addition of $m = 6$ to the good class in the results for Population 2 is due to the much higher standard deviation σ_{y_2}, which reduces the effect of the mean signal. Also notable is that the biased models ($m = 5, 6$) are closer to the data on both MD and D than the unbiased models ($m = 2, 3$), and they have much smaller sample standard deviations with correspondingly tighter ranges.

Considering Population 1 first, models 1, 4 and 5 outperform the other three on both average Mahalanobis distance, $\overline{MD}_m$, and mean normalized distance, $\bar{D}_m$. Interestingly, model 4 is slightly closer on average to the data than the true model, model 1, when measured by ordinary MD. Indeed, the two models are easy to confuse since $\overline{MD}_1$ and $\overline{MD}_4$ are within a couple of sample standard deviations (0.35 for model 1, 0.63 for model 4) of each other. Model 4 also appears superior to model 1 when performance is gauged by comparing MD pairwise for individual members of the sample: a separate calculation shows $MD_4(Y_1^{Obs}) < MD_1(Y_1^{Obs})$ on 20 of the samples. These results indicate that MD is not as good a discriminator as D. Indeed, the statistics for D favor model 1, the true model, and are much less equivocal than those of MD. $\bar{D}_1$ is more than two sample standard deviations less than $\bar{D}_4$ and $D_4(Y_1^{Obs}) > D_1(Y_1^{Obs})$ on all 25 samples. Their ranges do not overlap either: $\max[D_1] = 23.76 < 26.36 = \min[D_4]$.

The results are similar for Population 2, except now $m = 5$ is the second-best model when measured by D, and $m = 6$ joins the class of good models. This reflects the relatively greater importance of model correlation structure compared to model bias as S/N decreases. Once again, normalization favors model 1, the true model, but now $\bar{D}_1$ and $\bar{D}_5$ are within a couple of sample standard deviations (0.19 for model 1, 0.20 for model 5) of each other and their ranges overlap. Models 4 and 6 are also close to $m = 1$, but are more than two sample deviations away from it. However, when D_1, D_4, D_5, and D_6 are compared pair-wise on individual samples D_1 is less than the others on all 25 samples. As S/N decreases, it will be harder to distinguish between models, which is to be expected in general.

$E[D^2(\psi_m^{Obs}, Y^{Obs})]$. Since MD is a quadratic form, we can quantify the relative effects on $E[D_m^2]$ of bias and correlation structure by breaking it into components (Lam 1973),

$$E\left[D_m^2\right] = Tr\left(\rho_m^{-1} \cdot \rho_Y\right) + \left(\psi_m^{Obs} - Y^{Obs}\right)^T \rho_m^{-1} \left(\psi_m^{Obs} - Y^{Obs}\right) + \log|\rho_m| \quad (8)$$

This separates $E[D_m^2]$ into relative contributions of terms depending on (i) the similarity between ρ_m and ρ_Y as evaluated through the trace of the matrix product, $Tr(\rho_m^{-1} \cdot \rho_Y)$, (ii) the expected bias as measured by the squared Mahalanobis distance between ψ_m and ψ_Y, $(\psi_m^{Obs} - Y^{Obs})^T \ \rho_m^{-1} \ (\psi_m^{Obs} - Y^{Obs})$, and (iii) the normalization effected by $\log|\rho_m|$. It is worth noting that $E[D_m^2]$ based on unbiased models only depends on factors related to ρ_m, $E[D_m^2] = Tr(\rho_m^{-1} \cdot \rho_Y) + \log|\rho_m|$, and the true model yields $E[D_1^2] = N + \log|\rho_Y|$. The sum of the first and second terms gives the expected Mahalanobis distance,

$$E[MD_m]^2 = Tr\left(\rho_m^{-1} \cdot \rho_Y\right) + (\psi_m - \psi_Y)^T \rho_m^{-1} (\psi_m - \psi_Y) \quad (9)$$

Models $m = 1, 5, 6$ have a trace of 360 (Table 3) because they have the same correlation structure as the true model, $\rho_m = \rho_Y$. Model 4 has a smaller trace (342.95) than the true model, clearly indicating the need for normalization. Models $m = 1, 2, 3, 4$ are unbiased, but the importance of bias decreases as S/N decreases: the bias on Population 2 goes down for models 5 and 6 because of the decreased S/N. Bias goes from accounting for 50% (80%) of $E[D_5^2]$ ($E[D_6^2]$) in Population 1 to 2% (7%) in Population 2. That effect is also evident in Fig. 1. This is the main reason for the improvement of $m = 5$ and 6 on $E[D_m^2]$ relative to $m = 1$ and 4 as we go from Population 1 to 2. $E[MD_4]$ is slightly less than $E[MD_4]$ for both populations, indicating the importance of normalization. Normalization yields $E[D_1^2] < E[D_4^2]$ for both populations. Additionally, the decreased importance of departures from the mean (through reduced S/N) gives $E[D_1^2] < E[D_5^2] < E[D_6^2] < E[D_4^2]$ for Population 2. At the same time the importance of the normalization increases with decreasing S/N since, e.g., $\log[\rho_5]/\log[\rho_4] = 0.48$ for Population 1 and $\log[\rho_5]/\log[\rho_4] = 0.86$ for Population 2, bringing D_4^2 closer to D_5^2.

4 Summary and conclusions

An important approach to modeling environmental systems is based on stochastic models of system dynamics drawn from basic physical, chemical and biological principles. This approach is common, for instance, in groundwater hydrology, weather prediction and climate modeling. Stochasticity is invoked in such models for at least two reasons. First, data for system parameters are usually sparse, while the parameters are almost always heterogeneous in space and/or time. Hence, there is a high degree of uncertainty about parameters, and therefore, about system states which are obtained by propagating parametric uncertainty through model dynamics. In many cases only enough data is available to estimate the first couple of moments of system states. Additionally, state estimates are often based on low-order perturbation approximations. Hence, even if a model were known to be exactly correct, sparse data combined with heterogeneity and low-order approximations would induce a high degree of uncertainty in state estimates.

Second, in most realistic settings, the form of a model is an approximation, and in most important settings, more than one model, m, is available. Hence, model form is itself an additional source of uncertainty. When enough data is available to estimate a probability distribution, $P_m[Y(t)]$, for the system states and also probability distributions, $P_m[\Pi_m]$, for model parameters, models can be compared by computing Bayes factors or simple likelihood ratios. This paper deals with the opposite case: settings in which data is so limited that only the first couple of moments of states, i.e., the expected state value ψ_m and the correlation structure ρ_m, can be estimated. It is still important to compare models in such settings, but we must be prepared to do so on a weaker basis.

Some kind of distance between observations of states, Y^{Obs}, and modeled state estimates, ψ_m^{Obs}, is an obvious candidate for a weak comparison criterion. It is desirable for the distance to take account of all the information about models available, that is, to account for both ψ_m and ρ_m. The most obvious candidate is Euclidean distance, $\|\psi_m^{Obs} - Y^{Obs}\|$ but it relies on just the first moment, ψ_m, so it cannot distinguish between models that have approximately the same bias but different correlation structure. The Mahalanobis distance, MD, on the other hand, includes ρ_m as well as ψ_m. In that sense, it uses all the information that is usually available when data is limited. However, the Mahalanobis distance as it is ordinarily defined is not exactly appropriate for model comparison, because differences between model correlation structures hypothesized by different models can be large. Hence, we have examined a variation, D, on the Mahalanobis distance that is normalized by $\log|\rho_m|$. D seems to compensate for the advantage one model may have over another (including over the "true" model) that arises due to "small" ρ_m.

We evaluated D's ability to quantify model performance in two ways. For one, we applied it to two populations of synthetic data with different signal–noise ratios (S/N) and compared it to MD. D consistently selected the true model out of the sample data, although its performance deteriorates as S/N decreases. This is to be expected from any distance measure since they all depend on S/N through the bias $\psi_m^{Obs} - Y^{Obs}$. Second, we decomposed $E[D^2]$ into components depending on model bias, correlation structure, and normalization and we compared it to $E[MD^2]$. From these results it is clear that normalization is effective precisely because it equalizes the effect of correlation structure. These results also quantify the comment just made about bias and S/N: as S/N decreases, so does the effect of bias.

Not too much should be expected of a relatively weak criterion like normalized Mahalanobis distance, D. Although it is true that D selects the true model in the computational experiments reported here, our most important finding is that D discriminates between "good" models and "weak" models reasonably well, and it does so for good reasons. The effect of bias is relatively greatest when it should be, i.e., when S/N is relatively large and the mean signal is clearest so that good estimates can be distinguished from poor ones on the basis of data. As S/N deteriorates, the role of bias decreases and D_m comes to depend more on ρ_m. Overall, the influence of both moments on the value of D_m is crucial, which is what we sought when looking for a measure that factors in both the means and correlation structures hypothesized by models.

Acknowledgement The author wishes to thank the reviewers and the editor for their many useful comments and corrections. He is especially grateful to Reviewer 2 and the editor for pointing out an important simplification to Eq. 2.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Cs′aki F (eds) Proceedings of the 2nd international symposium on information theory. Akad′emiai Kiad′ o, Budapest, pp 267–268

Benson D, Tadjeran C, Meerschaert M, Farnham I, Pohll (2004) Radial fractional-order dispersion through fractured rock. Water Resour Res 40:W12416 1–9

Berkowitz B, Cortis A, Dentz M (2006) Modeling non-Fickian transport in geological formations as a continuous time random walk. Rev Geophys 44: RG2003 1–49

Christakos G (2003) Another look at the conceptual fundamentals of porous media upscaling. Stoch Env Res Risk Assess 17:276–290

Dentz M, Scher H, Holder D (2008) Transport behavior of coupled continuous-time random walks. Phys Rev E 78:041110 1–9

Guadagnini A, Guadagnini L, Tartakovsky DM, Winter CL (2003) Random domain decomposition for flow in heterogeneous stratified aquifer. Stoch Env Res Risk Assess 17:394–407

Lam TY (1973) The algebraic theory of quadratic forms. W. A. Benjamin, Reading, MA

Mahalanobis P (1936) On the generalised distance in statistics. Proc Natl Inst Sci India 2:49–55

Neuman S (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch Env Res Risk Assess 17:291–305

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–466

Spiegelhalter D, Best N, Carlin B, van der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc B 64(Part 4):583–639

Tartakovsky D, Neuman S (1998) Transient flow in bounded randomly heterogeneous domains 1. Exact conditional moment equations and recursive approximations. Water Resour Res 34:1–12

Winter C, Neuman C, Neuman S (1984) A perturbation expansion for diffusion in a random velocity field. SIAM J Appl Math 44:411–424

Zhang D, Neuman S (1995) Eulerian-Lagrangian analysis of transport conditioned on hydraulic data. Water Resour Res 31:39–51