Appendix

Vectors are denoted by bold lowercase letters (e.g., $\mathbf{x}$) and matrices by bold uppercase letters (e.g., $\mathbf{X}$). T denotes the transpose. A boldface $\mathbf{f}$ in $\mathbf{f}(\mathbf{x})$ indicates that $\mathbf{f}$ returns a vector for a given value of the vector $\mathbf{x}$. Individual vector or matrix elements are denoted by subscript indices (e.g., X_{ij}).

The Bayesian Solution to Inverse Problems.—The central part of Bayesian inference relies on the product rule of probability theory (Jaynes 2003), given by

$$p(A, B \mid C) = p(A \mid C)\, p(B \mid A, C), \tag{A.1}$$

where $p(A, B \mid C)$ is the joint conditional probability that propositions A and B are true given that proposition C is true. Because $p(B, A \mid C) = p(A, B \mid C)$, we obtain Bayes's Theorem:

$$p(A \mid B, C) = \frac{p(A \mid C)\, p(B \mid A, C)}{p(B \mid C)}. \tag{A.2}$$

Equation A.2 defines the posterior probability density function (PPD) on A, which allows inferences to be made on A taking into account B and C. If we substitute proposition A with a given model parameterization $\mathbf{m}$, proposition B with the observed data $\mathbf{d}$, and C with prior information I, we have essentially the PPD solution to the inverse problem:

$$p(\mathbf{m} \mid \mathbf{d}, I) = \frac{p(\mathbf{m} \mid I)\, p(\mathbf{d} \mid \mathbf{m}, I)}{p(\mathbf{d} \mid I)}. \tag{A.3}$$

In equation A.3, $p(\mathbf{m} \mid I)$ is the prior distribution, which represents prior knowledge of the model parameters, the geological setting, the ecology of the organism, and other factors. $p(\mathbf{d} \mid \mathbf{m}, I)$ quantifies how probable the observed data $\mathbf{d}$ are for a choice $\mathbf{m}$ of model parameters. It is often called the likelihood function when expressed as a function of $\mathbf{m}$ given $\mathbf{d}$, and it incorporates the data-model misfit function (see below). The denominator $p(\mathbf{d} \mid I)$, sometimes called the ''evidence,'' is the integral of the product of the prior and the likelihood over all possible values of $\mathbf{m}$ (i.e., the prior expectation of the likelihood), and represents a normalizing factor that makes the total probability equal one. $p(\mathbf{d} \mid I)$ is not a function of the model parameters in $\mathbf{m}$ (i.e., they are ''integrated out''), and for the purpose of parameter estimation it is often denoted by a constant k^{-1}. To simplify notation, the conditional term $\mid I$ is dropped from here on, the prior is denoted $\rho(\mathbf{m})$, and $p(\mathbf{d} \mid \mathbf{m})$ is expressed as the likelihood $L(\mathbf{m} \mid \mathbf{d})$. Thus, the PPD can be written as

$$p(\mathbf{m} \mid \mathbf{d}) = k\, \rho(\mathbf{m})\, L(\mathbf{m} \mid \mathbf{d}). \tag{A.4}$$

This quantity is taken to represent the information available on the model, and its calculation depends on the data, any prior information, and the error statistics, all of which are discussed in the paper.

Shape Quantification.—The Zahn and Roskies (1972) normalized shape function is given by

$$\phi^*(l) = \phi(l) + \frac{2\pi l}{L}, \tag{A.5}$$

where l is the cumulative arc length along the outline, L is total arc length, and $\phi(l)$ is the cumulative angular deviation, such that $\phi(0) = 0$ and $\phi(L) = -2\pi$ for any closed curve. The normalized variant $\phi^*(l)$ quantifies shape in terms of its deviation from a circle. Given a matrix $\mathbf{Z}$ of standardized (zero mean, unit variance) shape functions (columns) for a sample of specimens (rows), an estimate of the mean shape $\bar{\mathbf{z}}$ can be obtained by singular value decomposition of $\mathbf{Z}$,

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{A.6}$$

where the columns of $\mathbf{U}$ represent empirical shape functions (i.e., eigenshapes; Lohmann 1983), $\mathbf{D}$ is the diagonal matrix of singular values, and the columns of $\mathbf{VD}$ are projections of specimens onto the eigenshapes. The first eigenshape is taken as an estimate of the mean shape.

Fossil Abundance and Preservation.—Let d denote water depth, and g denote grain size. Let μ_d be the species' preferred depth, σ_d the depth tolerance, μ_g the preferred grain size, and σ_g the grain size tolerance. If we assume that the species' response to each environmental variable is uncorrelated, the depth and grain size parameters describe a bivariate Gaussian density function representing the probability of occurrence of an organism given the d and g values in a sample:

$$f(d, g) = \frac{1}{2\pi\sigma_d\sigma_g}\exp\left(-\frac{1}{2}\left[\frac{(d - \mu_d)^2}{\sigma_d^2} + \frac{(g - \mu_g)^2}{\sigma_g^2}\right]\right). \tag{A.7}$$

Local abundance n is found by scaling the peak of $f(d, g)$ to the peak abundance $n_{\max}$. The number of individuals preserved as fossils in a sample, K, is determined by (a) local abundance n, where each individual is considered a binomial (success-failure) trial, and (b) per capita preservation probability q, representing the intrinsic fossilization potential of the organism as well as the probability of collection. Because n is generally very large relative to q, we can model this as a Poisson process with density function

$$f(K, nq) = \frac{e^{-nq}(nq)^K}{K!}. \tag{A.8}$$

Multivariate Phenotypic Evolution.—If the phenotypic distribution of traits is multivariate Gaussian, then the evolution of the multivariate mean phenotype can be expressed as

$$\Delta\bar{\mathbf{z}} = \boldsymbol{\beta}\mathbf{G}, \tag{A.9}$$

where $\Delta\bar{\mathbf{z}}$ is the change in population mean phenotype in one generation, $\mathbf{G}$ is the additive genetic covariance matrix, and $\boldsymbol{\beta}$ is the vector of selection coeffcients acting on each of the phenotypic traits (Lande 1976, 1979; Arnold et al. 2001). Equation (A.9) can be rewritten as

$$\Delta\bar{\mathbf{z}} = \boldsymbol{\beta}\mathbf{H}\mathbf{P}, \tag{A.10}$$

where $\mathbf{H}$ is a matrix of trait heritabilities that transforms the phenotypic covariance matrix $\mathbf{P}$ into $\mathbf{G}$. Because we can estimate $\mathbf{P}$ from fossil samples, multivariate evolution can be modeled by using the product $\boldsymbol{\beta}\mathbf{H}$ as a parameter vector. This is the approach taken by Polly (2004), who also described how a more effcient, reduced-dimension simulation can be performed by projecting equation A.10 into principal component shape space. The $\mathbf{P}$-matrix can be expressed in terms of its principal components (eigenvectors) and eigenvalues,

$$\mathbf{P} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T, \tag{A.11}$$

where $\mathbf{U}$ represents the eigenvectors of $\mathbf{P}$ (the loadings of the original variables on the principal component axes), and $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{P}$. The matrix $\mathbf{U}$ can be used as a rotation matrix to project the shape onto the principal component axes, and the evolution of the projected mean shape (Polly 2004) is given by

$$\begin{aligned}\Delta\bar{\mathbf{z}}^* = \Delta\bar{\mathbf{z}}\mathbf{U} = \boldsymbol{\beta}\mathbf{H}\mathbf{P}\mathbf{U} = \boldsymbol{\beta}\mathbf{H}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{U}\\ = \boldsymbol{\beta}\mathbf{H}\mathbf{U}\boldsymbol{\Lambda} = (\boldsymbol{\beta}\mathbf{H})^*\boldsymbol{\Lambda}.\end{aligned} \tag{A.12}$$

The result can be rotated back into the original coordinate space by post-multiplying with $\mathbf{U}^T$. Using $\mathbf{S}^* = (\boldsymbol{\beta}\mathbf{H})^*$ as a vector of projected selection differentials, different modes of evolution can be simulated by drawing the elements of $\mathbf{S}^*$ from different underlying distributions. Furthermore, it is possible to limit the number of elements of $\mathbf{S}^*$ needed based on the eigenvalues associated with each principal component vector.

Defining Data-Model Misfit.—Given a set of model parameters $\mathbf{m}$, the forward model generates a predicted stratophenetic series $\mathbf{g}(\mathbf{m})$, which is compared to the observed stratophenetic series $\mathbf{d}$. This data vector includes the number of fossils per sam-

ple (most samples are barren) and the mean sample shape function for each sample. Data-model misfit involves the sum of separate misfit measures for each of these two data types. In the observed data vector $\mathbf{d}$, the mean shape functions of samples without fossils are not defined and are omitted from the misfit calculation. In the predicted data vector $\mathbf{g}(\mathbf{m})$, the barren sample mean shape functions are defined as zero vectors, to allow misfit calculation when the observed shape is defined. Under the Gaussian error statistics assumption, a large number of pure ''noise'' stratophenetic series can be generated and their covariance matrix calculated as described by Gouveia and Scales (1998). Assuming that the temporal correlation in the data uncertainty is negligible, only the main diagonal of the covariance matrix (noise variance) is needed. This noise vector $\boldsymbol{\epsilon}^2$ thus contains both sample size error and shape error.

Let N_m be the number of model parameters, N_o the number of samples, N_q the number of preserved samples in the observed data (stratophenetic series length), and N_ϕ the number of points along the shape outline (phenotypic variables). A data-model misfit measure can then be constructed as

$$\chi^2(\mathbf{m}) = \frac{1}{\nu_q}\sum_{i=1}^{N_o}\frac{(d_i - g(\mathbf{m})_i)^2}{\epsilon_i} + \frac{1}{\nu_\phi}\sum_{j=N_o+1}^{N_qN_\phi}\frac{(d_j - g(\mathbf{m})_j)^2}{\epsilon_j}, \quad \text{(A.13)}$$

where $\nu_q = N_o - N_m$ and $\nu_\phi = N_qN_\phi - N_m$ are the (approximate) degrees of freedom. The likelihood function can be written as

$$L(\mathbf{m}\,|\,\mathbf{d}) = k\exp\left(-\frac{1}{2}\chi^2(\mathbf{m})\right), \quad \text{(A.14)}$$

where k is a normalizing constant (see eqs. A.3 and A.4).

The Prior Model Covariance Matrix.—From the definition of the variance of a uniform distribution (Wackerly et al. 2002), we can express the prior model covariance matrix as

$$C_{i,j}^{prior} = \begin{cases}\frac{1}{12}\Delta m_i{}^2 & \text{if } i = j\\ 0 & \text{otherwise,}\end{cases} \quad \text{(A.15)}$$

where Δm_i is the range of the ith parameter.

Bayesian Integrals.—The 1-D marginal posterior distribution of parameter m_i is found by integrating the PPD over the remaining dimensions N_m of the parameter space:

$$p(m_i\,|\,\mathbf{d}) = \int\cdots\int p(\mathbf{m}\,|\,\mathbf{d})\prod_{\substack{k=1\\k\neq i}}^{N_m} dm_k. \quad \text{(A.16)}$$

The posterior model covariance matrix is given by

$$C_{i,j}{}^M = \int m_i m_j p(\mathbf{m}\,|\,\mathbf{d})\,d\mathbf{m} - \langle m_i\rangle\langle m_j\rangle, \quad \text{(A.17)}$$

where $\langle m_i\rangle$ is the posterior mean model for the ith parameter. If the PPD is Gaussian then the mean model is located at the peak of the PPD, but for nonlinear problems the shape of the PPD can be complex, and $\mathbf{C}^M$ is less easily interpreted (Tarantola 1987). A model correlation matrix is calculated by dividing each element $C_{i,j}{}^M$ by the product of the posterior standard errors of parameters m_i and m_j. Furthermore, C^M can be used to obtain a model resolution matrix:

$$\mathbf{R} = \mathbf{I} - \mathbf{C}_{prior}^{-1}\mathbf{C}^M, \quad \text{(A.18)}$$

where $\mathbf{C}_{prior}^{-1}$ is the inverse prior covariance matrix (eq. A.15), and $\mathbf{I}$ is the identity matrix. A dimensionless resolution matrix is obtained by multiplying each element of the resolution matrix by the ratio of the prior standard errors of the parameters (Sambridge 1999b).

The Bayesian integrals are calculated using Monte Carlo numerical integration techniques implemented in the Neighbourhood Algorithm, and readers are referred to Sambridge (1999a,b) for details.

Convergence of the Gibbs Sampler.—All the Bayesian quantities reported in the paper rely on the convergence of the Gibbs sampler used to resample the ensemble and generate a distribution that follows the approximate PPD. The convergence of the Gibbs sampler can be evaluated by calculating a potential scale reduction (PSR) factor, which is deemed acceptable if the PSR values for all variables are less than 1.2 (Gelman et al. 1995). For the numerical integrations in this paper, the Gibbs sampler showed good convergence, with PSR values for all the parameters between 1.01 and 1.12, with a median of 1.04.

Literature Cited

Arnold, S. J., M. E. Pfrender, and A. G. Jones. 2001. The adaptive landscape as a conceptual bridge between micro- and macroevolution. Genetica 112–113:9–32.

Gelman, A. B., J. S. Carlin, H. S. Stern, and D. B. Rubin. 1995. Bayesian data analysis. Chapman & Hall/CRC Press, Boca Raton, Fla.

Gouveia, W. P. J., and J. A. Scales. 1998. Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis. Journal of Geophysical Research 103-B2:2759–2779.

Jaynes, E. T. 2003. Probability theory—the logic of science. Cambridge University Press, Cambridge.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

———. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33: 402–416.

Lohmann, G. P. 1983. Eigenshape analysis of microfossils: a general morphometric procedure for describing changes in shape. Mathematical Geology 15:659–672.

Polly, P. D. 2004. On the simulation of the evolution of morphological shape: multivariate shape under selection and drift. Palaeontologia Electronica 7(2)7A:1–28.

Sambridge, M. 1999a. Geophysical inversion with a neighbourhood algorithm. I. Searching a parameter space. Geophysical Journal International 138:479–494.

———. 1999b. Geophysical inversion with a neighbourhood algorithm. II. Appraising the ensemble. Geophysical Journal International 138:727–746.

Tarantola, A. 1987. Inverse problem theory: methods for data fitting and parameter estimation. Elsevier, Amsterdam.

Wackerly, D. D., W. Mendenhall III, and R. L. Scheaffer. 2002. Mathematical statistics with applications, 6th ed. Duxbury, Pacific Grove, Calif.

Zahn, C. T., and R. Z. Roskies. 1972. Fourier descriptors for plane closed curves. IEEE Transactions on Computers C-21: 269–281.