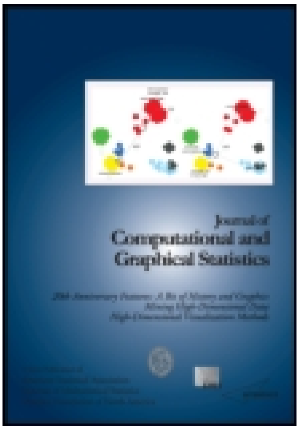

Journal of Computational and Graphical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

The Bayesian Change Point and Variable Selection Algorithm: Application to the $\delta^{18}O$ Proxy Record of the Plio-Pleistocene

Eric Ruggieri & Charles E. Lawrence

Accepted author version posted online: 05 Jul 2012.Published online: 12 Feb 2014.

To cite this article: Eric Ruggieri & Charles E. Lawrence (2014) The Bayesian Change Point and Variable Selection Algorithm: Application to the $\delta^{18}O$ Proxy Record of the Plio-Pleistocene, Journal of Computational and Graphical Statistics, 23:1, 87-110, DOI: 10.1080/10618600.2012.707852

To link to this article: http://dx.doi.org/10.1080/10618600.2012.707852

The Bayesian Change Point and Variable Selection Algorithm: Application to the δ^{18}O Proxy Record of the Plio-Pleistocene

Eric RUGGIERI and Charles E. LAWRENCE

In this article, we introduce the Bayesian change point and variable selection algorithm that uses dynamic programming recursions to draw direct samples from a very high-dimensional space in a computationally efficient manner, and apply this algorithm to a geoscience problem that concerns the Earth's history of glaciation. Strong evidence exists for at least two changes in the behavior of the Earth's glaciers over the last five million years. Around 2.7 Ma, the extent of glacial cover on the Earth increased, but the frequency of glacial melting events remained constant at 41 kyr. A more dramatic change occurred around 1 Ma. For over three decades, the "Mid-Pleistocene Transition" has been described in the geoscience literature not only by a further increase in the magnitude of glacial cover, but also as the dividing point between the 41 kyr and the 100 kyr glacial worlds. Given such striking changes in the glacial record, it is clear that a model whose parameters can change through time is essential for the analysis of these data. The Bayesian change point algorithm provides a probabilistic solution to a data segmentation problem, while the exact Bayesian inference in regression procedure performs variable selection within each regime delineated by the change points. Together, they can model a time series in which the predictor variables as well as the parameters of the model are allowed to change with time. Our algorithm allows one to simultaneously perform variable selection and change point analysis in a computationally efficient manner. Supplementary materials including MATLAB code for the Bayesian change point and variable selection algorithm and the datasets described in this article are available online or by contacting the first author.

Key Words: Direct posterior sampling; Dynamic programming recursions; Exact Bayesian inference in regression (EBIR); Mid-Pleistocene Transition; Glacial dynamics; Regression.

1. INTRODUCTION

The Earth's ice sheets have been melting and reforming for millions of years. Since the most abundant isotope of oxygen, ^{16}O, more readily evaporates from the oceans and falls

Eric Ruggieri is Assistant Professor, Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, PA 15282 (E-mail: *ruggierie@duq.edu*). Charles E. Lawrence is Professor of Applied Mathematics, Division of Applied Mathematics, Brown University, Providence, RI 02912 (E-mail: *Charles_Lawrence@brown.edu*).

Journal of Computational and Graphical Statistics, Volume 23, Number 1, Pages 87–110
DOI: 10.1080/10618600.2012.707852
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jcgs*.

as snow in the polar regions than a heavier isotope, ^{18}O, an increase in glacial ice will cause the concentration of ^{18}O in the oceans to increase relative to ^{16}O. Geoscientists capitalize on the resulting changes in the isotopic ratios of oxygen in the ocean to create a δ^{18}O ice volume proxy record from ocean sediment cores, which quantifies the amount of ice on the Earth at a specific time in the past (Hays, Imbrie, and Shackleton 1976; Imbrie and Imbrie 1980; Lisiecki and Raymo 2005; among others). Thus, the δ^{18}O record provides a way to study the patterns of glacial growth and destruction through time.

The Earth has undergone a gradual, but substantial cooling over the last five million years. This cooling engendered the formation of more permanent ice sheets over the Northern Hemisphere around 2.7 million years ago (Ma), reflected by an obvious increase in the amplitude of the δ^{18}O proxy values from ocean sediment cores (Lisiecki and Raymo 2005). Further cooling of the Earth likely contributed to the Mid-Pleistocene Transition (MPT; Raymo 1997; Tziperman and Gildor 2003; Raymo, Lisiecki, and Nisancioglu 2006) around 1 Ma. During the MPT, not only there was a further increase in the amplitude of the δ^{18}O proxy values, but also the periodicity of the glacial cycles apparently changed from 41 to 100 thousand years (kyr), sparking much debate in the geoscience literature and spawning several alternative models to explain this event. Imbrie and Imbrie (1980) noted this phenomenon, concluding that “to understand these long climatic records, it may be necessary to use models whose parameters vary with time”—indicating the existence of change points in the system. While the intensification of Northern Hemisphere glaciations and the MPT are two striking examples of distinct model changes, there may well be other unreported changes that are also important to the Earth’s climate history. Thus, when modeling how ice sheets respond to changes in solar insolation, their perceived forcing function (Milankovitch 1941/1969), the inference challenges are to infer the number and timing of change points, the appropriate model in each regime, and the parameters of these models.

This problem lives on a discrete high-dimensional space. There are $>1.2 \times 10^{17}$ different ways to place just six change points in the 2115 data points of the δ^{18}O proxy record (Lisiecki and Raymo 2005) before any considerations for variable selection are made. In general, given N data points, there are approximately $\binom{N}{k}$ ways to delineate $k+1$ regimes through the placement of k change points. In addition, given m variables, there are 2^m combinations of variables to select from within each regime. Therefore, brute-force attempts to study the δ^{18}O proxy record are futile. To address these challenges in the study of the Earth’s dynamic ice sheets, we introduce the Bayesian change point and variable selection algorithm, an algorithm that employs dynamic programming-like recursions to draw samples directly from this high-dimensional joint posterior space. Change point analysis seeks an answer to the question of when the Earth’s glacial systems have undergone regime changes, while variable selection seeks to infer which of several proposed predictor variables or combinations of them apply in each regime. Each climate regime detected by the algorithm is independently fit by a regression model using the predictors identified through the variable selection procedure.

Frequentist solutions to the change point problem in regression are often obtained by minimizing squared error or through likelihood ratio tests. For example, Tibshirani and Wang (2008) described the fused Lasso, a procedure designed to find the least-squares change point solution subject to an L1 penalty function. Olshen et al. (2004) introduced circular binary segmentation, which is a modification of the traditional binary segmentation algorithm that estimates the location of change points via a likelihood ratio statistic. The algorithm is efficient, O(Nlog$N$), because it iteratively divides the dataset according to the change points, but since the algorithm is greedy, there is no guarantee of finding the optimal solution. On the other hand, Muggeo and Adelfino (2011) searched for change points in comparative genomic hybridization (CGH) data using a piecewise constant model. They transformed the data into a cumulative sum, which is piecewise continuous, and then used an iterative model (Muggeo 2008) to identify the locations of the change points, subject to refinement. The R packages "strucchange" (Zeileis et al. 2002) and "changepoint" (Killick and Eckley 2011) implement several of the mainstream frequentist change point methods.

Additionally, dynamic programming has previously been employed in the frequentist setting to find guaranteed optimal solutions to high-dimensional change point problems as it reduces the O(N^k) calculation on the location of change points to a quadratic calculation O(N^2) (Hawkins 1976, 2001; Auger and Lawrence 1989; Bai and Perron 2003; Ruggieri et al. 2009). The downside to these algorithms is that they are unable to quantify the uncertainty associated with their solutions, both in the number and in the locations of the change points. Our Bayesian approach to the change point problem is designed to remedy these limitations.

As for Bayesian algorithms, Barry and Hartigan (1993) introduced the product partition model and developed an exact, recursive solution that is O(N^3). Parameter values are estimated using the probabilities that a "block" of the data is included in the partition. Barry and Hartigan (1993) also introduced a more efficient [O(N)] Markov chain Monte Carlo (MCMC) approach to the change point problem, but their work was limited to detecting changes in the mean.

To date, Gibbs sampling (e.g., Carlin, Gelfand, and Smith 1992; Stephens 1994; Western and Kleykamp 2004) and MCMC (e.g., Barry and Hartigan 1993; Green 1995; Lavielle and Lebarbier 2001; Erdman and Emerson 2008) approximations have dominated probabilistic solutions to the change point problem. Carlin, Gelfand, and Smith (1992) developed a method for detecting a single change point, while both Stephens (1994) and Western and Kleykamp (2004) could detect multiple change points for a general regression model. Green (1995) introduced the reversible jump MCMC that can jump between parameter spaces of differing dimensionality, while Lavielle and Lebarbier (2001) focused on finding changes in the mean. Chopin (2007) noted that "MCMC samples for change point models typically have a O(N^2) computational cost, while their convergence properties tend to deteriorate for larger values of N." More recently, Erdman and Emerson (2008) developed an O(N) (per iteration) MCMC procedure based on the product partition model of Barry and Hartigan (1993). An MCMC algorithm that is O(N) per iteration becomes of O($I \times N$), where I is the number of iterations. However, "bcp" is limited to detecting changes in the mean. Although there have been many heuristics developed to identify the number of iterations required to achieve convergence, in all but a few special cases the convergence of MCMC algorithms, including those developed for the change point problem, cannot be assured.

Recursive dynamic programming-like algorithms that employ recursions to complete sums required to marginalize over high-dimensional discrete variables provide a means to guarantee convergence. The most famous of these, the hidden Markov model (HMM), has also been used to solve the change point problem in a probabilistic setting (Chib 1998;

Pesaran, Pettenuzzo, and Timmermann 2006). The transition from one state to another in an HMM is similar to the placement of a change point, while the emissions of an HMM are representative of the predictions of a regression model. However, HMMs are predicated on the recurrence of each emission model an unknown number of times. As a result, its emission models are not local to any given data substring or regime. Moreover, in the oft-cited algorithm of Chib (1998), the number of regimes must be prespecified. Given k change points, Chib reparameterized the change point model in terms of a unidirectional HMM that is required to begin in state 1 and terminate in state $k+1$. A prior that imposes a specified number of change points can lead to potentially undesirable behavior at the end of the sample (Koop and Potter 2009). Koop and Potter (2007, 2009) developed extensions of Chib (1998) that allow for an unknown number of breaks, most recently by developing a more noninformative uniform prior. Additionally, the parameters of an HMM cannot be estimated or inferred within the recursion itself and are typically estimated via the iterative Baum–Welch algorithm. Because these models may have multiple local modes, their convergence cannot be assured. In contrast, since each regime in a change point problem is described by its own set of parameters (thus they are all local), dynamic programming recursions are available to assure convergence to global optima in frequentist settings, or to permit direct inferences in Bayesian settings, in a single pass through the data. HMMs can be extended to avoid point estimates of parameters and permit full Bayesian inferences by replacing the expectation-maximization (EM) algorithm with a corresponding Gibbs sampling algorithm, but such an approach would of course incur the convergence limitations of MCMC algorithms.

Our approach to the change point problem is a product partition model that is fundamentally different from both MCMC and HMM approaches to solving the change point problem. It extends the work of Liu and Lawrence (1999) and Fearnhead (2006) by allowing the data segments between change points to be fit by regression models. Moreover, it is one of the few that uses dynamic programming-like recursions to complete the sums and integrals over all the unknown parameters required in finding the normalizing constants for the overall problem and all of its subproblems. The local nature of these calculations permits direct Bayesian inference on the parameters of every segment of the dataset while maintaining a manageable time and space complexity, $O(N^2)$. Thus, one important contribution of this article is its use of a dynamic programming-like algorithm to solve the Bayesian change point regression problem. In addition to being able to draw samples directly and independently from the ensemble of all change point solutions, the algorithm also permits direct Bayesian inference on the set of predictor variables in every data segment using the exact Bayesian inference in regression (EBIR) algorithm (Ruggieri and Lawrence 2012). The EBIR algorithm also employs a recursion to address the computational challenges associated with the variable selection component of this problem. Together, the Bayesian change point and variable selection algorithm capitalizes on the conditional independence feature of the change point problem to draw inferences directly and independently from the joint posterior space of all the unknowns of this problem. To the best of our knowledge, no other algorithm simultaneously performs variable selection with change point analysis in a computationally efficient manner.

The rest of this article is organized as follows. To facilitate understanding, we begin in Section 2 by describing the Bayesian change point algorithm for a fixed set of predictor variables. Section 3 describes the EBIR procedure and its integration into the Bayesian change point algorithm, extending the algorithm to the case where each regime may be modeled by a different subset of the predictor variables. In Section 4, we employ a simulated dataset that acts as proof of principle that the computer code works and returns the expected results. In Section 5, we apply the Bayesian change point and variable selection algorithm to a δ^{18}O proxy record of the Plio-Pleistocene. Section 6 provides discussion and conclusions.

2. METHODS: THE BAYESIAN CHANGE POINT ALGORITHM

Given the dependent variable, Y, and m known predictor variables $X_1, \ldots, X_m$, linear regression methods are based upon the statistical model

$$Y = \sum_{l=1}^{m} \beta_l X_l + \varepsilon, \tag{1}$$

where β_l is the lth regression coefficient and ε is a random error term. The goal here is to build a piecewise regression model whose regime boundaries are the change points. In the context of the δ^{18}O proxy record, the X_l's are periodic (sinusoidal) functions whose coefficients, β_l, may or may not change from one regime to the next, as each regime will be independently fit by a regression model.

Let N be the total number of data points, $X = [X_1, \ldots, X_m]$ be the matrix of predictor variables, and $Y = [Y_1, \ldots, Y_N]$ be the vector of response variables. Define $Y_{i:j} = \{Y_i, Y_{i+1}, \ldots, Y_{j-1}, Y_j\}$, $1 \leq i < j \leq N$, to be a substring (i.e., subset or regime) of the response variables in the dataset; $X_{i:j}$ is defined in a similar manner. Let σ^2 be the residual variance and $\{C\}$ be the set of change points whose locations are $c_0 = 0,\ c_1, \ldots, c_k,\ c_{k+1} = N$. The Bayesian change point algorithm is concerned with making inferences from the joint distribution

$$\begin{aligned} f(Y, \beta, \sigma^2, \{C\}\,|X) = & \left[\prod_{i=0}^{k} f\left(Y_{(c_i+1):c_{i+1}} \middle| \beta_i, \sigma_i^2, c_i, c_{i+1}, X_{(c_i+1):c_{i+1}}\right) \right. \\ & \left. \times\ f\left(\beta_i \middle| \sigma_i^2, c_i, c_{i+1}, X_{(c_i+1):c_{i+1}}\right) f\left(\sigma_i^2 \middle| c_i, c_{i+1}, X_{(c_i+1):c_{i+1}}\right) \right] \\ & \times\ P(\{C\} = c_0, c_1, \ldots, c_k, c_{k+1}), \end{aligned} \tag{2}$$

where $\beta_i = \{\beta_1, \beta_2, \ldots, \beta_m\}$ (regime index "i" omitted) and σ_i^2 are the regression parameters for the ith substring. This represents a varying coefficients model where the vector β_i and σ_i^2 are constant within each regime. The key to avoiding the combinatorial growth in the number of change point solutions is to break the problem into a series of progressively smaller subproblems, the smallest of which, the inference of a single change point, can easily be solved. The full solution can then be found by efficiently piecing together these solutions.

There are three steps to the algorithm:

1. *Calculating the probability density of the data* $f(Y_{i,j}|X_{i:j})$: To place a single change point, we must first calculate $f(Y_{i:j}) = f(Y_{i:j}|X_{i:j})$ for each and every possible

substring of the data, $Y_{i:j}$ (details below). Let $X = [X_1, X_2, \ldots, X_m]$ be the (sub)set of regressors included in the regression model. To facilitate the explanation, assume that X is fixed (this assumption is relaxed via variable selection in Section 3). Our regression model assumes that the error terms, ε, are independent, mean zero, and normally distributed random variables. Therefore, the likelihood function for a substring of the data, $Y_{i:j}$, is $f(Y_{i:j}|\beta, \sigma^2, X_{i:j}) \sim N(X_{i:j}, \beta, \sigma^2 I)$, where I is the identity matrix. Conjugate priors are chosen for the prior distributions on the vector of amplitudes, $\beta = \{\beta_1, \beta_2, \ldots, \beta_m\}$, and the error variance, σ^2. Specifically, β is multivariate normal [$\beta \sim N(0, \sigma^2/k_0)$], and $\sigma^2 \sim \text{Scaled} - \text{Inverse}\chi^2(v_0, \sigma_0^2)$, where k_0 is a scale parameter relating the variance of the regression coefficients to the residual variance, while v_0 and σ_0^2 act as pseudo data points—v_0 pseudo data points of variance σ_0^2 (essentially, unspecified training data gleaned from prior knowledge of the problem). Putting these together, the marginal probability for a substring of the data, $Y_{i:j}$, is

$$f(Y_{i:j}|X_{i:j}) = \iint f(Y_{i:j}|\beta, \sigma^2, X_{i:j}) f(\beta|\sigma^2, X_{i:j}) f(\sigma^2|X_{i:j}) d\beta\, d\sigma^2.$$

Let n be the number of data points in a substring, $v_n = v_0 + n$, $\beta^* = (X_{i:j}^T X_{i:j} + k_0 I)^{-1} X_{i:j}^T Y_{i:j}$, and $s_n = (Y_{i:j} - X_{i:j}\beta^*)^T (Y_{i:j} - X_{i:j}\beta^*) + k_0 \beta^{*T}\beta^* + v_0\sigma_0^2)$. Integration yields

$$f(Y_{i:j}) = f(Y_{i:j}|X_{i:j}) = \frac{\left(v_0\sigma_0^2/2\right)^{v_0/2}\Gamma(v_n/2)(k_0)^{m/2}}{\Gamma(v_0/2)(s_n/2)^{v_n/2}(2\pi)^{n/2}\left|X_{i:j}^T X_{i:j} + k_0 I\right|^{1/2}}. \quad (3)$$

This quantity is calculated and then stored in memory for all possible substrings of the data, $Y_{i:j}$, with $1 \le i < j \le N$. The dependence on X is hereafter suppressed.

2. *Forward recursion (dynamic programming)*: Starting from one end of the time series, we can find the probability of any prefix of the data, $Y_{1:j}$, containing one change point by multiplying together the probabilities of two nonoverlapping substrings [calculated in Equation (3)] and summing over all possible placements of the change point. Let $P_k(Y_{1:j}) = P_k(Y_{1:j}|X_{1:j})$ be the probability density of the first j observations of the data containing k change points, given the regression model. When $k = 1$, this gives $P_1(Y_{1:j}) = \sum_{v=1}^{j-1} f(Y_{1:v}) \times f(Y_{v+1:j})$. The position of this first change point has now been marginalized out; no further information about its location is needed to solve the full problem.

   To find the probability density of a prefix with two change points, $P_2(Y_{1:j})$, we multiply together the probability density of a prefix containing one change point, $P_1(Y_{1:v})$, and a nonoverlapping substring that fills out the rest of the prefix, $f(Y_{v+1:j})$ (both previously calculated), and then sum over all possible placements of the second change point: $P_2(Y_{1:j}) = \sum_{v=1}^{j-1} P_1(Y_{1:v}) \times f(Y_{v+1:j})$. Again, the location of the second change point has now been marginalized out as no further information about its position is needed. The process continues, $P_k(Y_{1:j}) = \sum_{v=1}^{j-1} P_{k-1}(Y_{1:v}) \times f(Y_{v+1:j})$, until the full problem is solved.

Therefore, $P_k(Y_{1:j})$ is calculated recursively as

$$P_1(Y_{1:j}) = \sum_{v<j} f(Y_{1:v})\, f(Y_{v+1:j}), \tag{4}$$

$$P_k(Y_{1:j}) = \sum_{v<j} P_{k-1}(Y_{1:v})\, f(Y_{v+1:j}), \tag{5}$$

for $j = 1, 2, \ldots, N$. Because we assume a uniform distribution on the locations of the change points (see below), we do not yet have to factor in the probability of a change point at position v. However, use of a nonuniform prior requires its incorporation into the forward recursion step. Inferences concerning the unknown parameters, including the number and locations of the change points, can be made by sampling directly from the posterior distribution on the quantities of interest using a stochastic backtrace algorithm.

3. *Stochastic backtrace*: Here, we use Bayes' rule to draw samples directly from the posterior distribution via a stochastic version of the dynamic programming-like recursions that takes advantage of marginalization for each subproblem created in the forward recursion step. Each of the posterior distributions described below has an exact representation that is straightforward to sample from.

   To have a completely defined partition function (or normalization constant),

$$f(Y_{1:N}) = \sum_{k=0}^{k_{\max}} \sum_{c_1,\ldots,c_k} f(Y_{1:N}|K=k, c_1,\ldots,c_k) \times P(c_1,\ldots,c_k|K=k) \times P(K=k), \tag{6}$$

two additional quantities need to be specified: (1) a prior distribution on the number of change points, $P(K = k)$; and (2) a prior distribution on the locations of the change points, $P(c_1, c_2, \ldots, c_k|K = k)$. For (1), a priori, we place half the probability mass on zero change points $[P(K = 0) = 0.5]$ and assume a uniform prior on a positive number of change points $[P(K = k) = 0.5/k_{\max}]$, where $k_{\max}$ is the maximal number of allowed change points. For (2), we employ a noninformative uniform prior on $P(c_1, \ldots, c_k|K = k)$, that is, all change point solutions with exactly k change points are equally likely. Let N_k be the number of change point solutions with exactly k change points, then $P(c_1, \ldots, c_k|K = k) = 1/N_k$. When there are no restrictions on the distance between adjacent change points, N_k is approximately equal to $\binom{N}{k}$. This combinatorial prior directly accounts for the greater number of solutions as the number of change points increases. Taken together, $P(K = 0) = 0.5$ and for $k > 0$, $P(K = k, c_1, \ldots, c_k) = 0.5/(k_{\max})(N_k)$. With the normalization constant specified, we can now draw samples of the parameters of interest:

   (a) Sample a number of change points: The forward recursion calculates the density of the entire dataset, $Y_{1:N}$, given k change points, $P_k(Y_{1:N}) = f(Y_{1:N}|K = k)$. Using Bayes' rule, the posterior distribution on the number of change points, given the data, is

$$f\,(K=k|Y_{1:N})=\frac{P_k\,(Y_{1:N})\,P(c_1,\ldots,c_k|K=k)P(K=k)}{f(Y_{1:N})},\tag{7}$$

with $f(Y_{1:N})$ defined in Equation (6).

(b) Sample the locations of the change points: Additionally, Bayes' rule can be used to assess the uncertainty related to the exact timing of a change. Let $c_{K+1}=N$, the last data point. Then, for $k=K,\ K-1,\ldots,1$, iteratively draw samples according to

$$f(c_k=v|c_{k+1})=\frac{P_{k-1}\,(Y_{1:v})\,f(Y_{v+1:c_{k+1}})}{\sum_{v\in[k-1,c_{k+1})}P_{k-1}\,(Y_{1:v})\,f(Y_{v+1:c_{k+1}})}.\tag{8}$$

(c) Sample the regression parameters for the interval between adjacent change points, c_k and c_{k+1}: Let $n=c_{k+1}-c_k$, the number of data points in a subinterval and let $\beta^*=(X^T_{(c_k+1):c_{k+1}}X_{(c_k+1):c_{k+1}}+k_0I)^{-1}X^T_{(c_k+1):c_{k+1}}Y_{(c_k+1):c_{k+1}}$, $s_n=(Y_{(c_k+1):c_{k+1}}-X_{(c_k+1):c_{k+1}}\beta^*)^T(Y_{(c_k+1):c_{k+1}}-X_{(c_k+1):c_{k+1}}\beta^*)+k_0\beta^{*T}\beta^*+v_0\sigma_0^2)$, and $v_n=v_0+n$. Using Bayes' rule one final time, we obtain

$$f(\beta|\sigma^2)\sim N\left(\beta^*,\left(X^T_{(c_k+1):c_{k+1}}X_{(c_k+1):c_{k+1}}+k_0I\right)^{-1}\sigma^2\right),\tag{9}$$

$$f(\sigma^2)\sim \text{Scaled}-\text{Inverse}\,\chi^2(v_n{}^{s_n}/_{v_n}).\tag{10}$$

Step 1 (calculating the probability density of the data) is O(N^2), the forward recursion step is O(kN^2), and the stochastic backtrace is O(kN). Therefore, the algorithm has a total time complexity of O(kN^2). In practice, the most time-consuming step is calculating the probability of the data, which depends in part on the complexity of the regression model.

3. METHODS: VARIABLE SELECTION VIA EBIR

Bayesian approaches to variable selection focus on finding the posterior distribution across the ensemble of candidate submodels. Approximations of the posterior space by stepwise regressions (see Miller 2002 and references therein) can leave open the question of local versus global optima and often do not address the uncertainty of including specific variables. The "spike and slab" model is first introduced by Mitchell and Beauchamp (1988). Here, the maximum a posteriori (MAP) estimator can be viewed as the "best" submodel, but the entire, exponentially growing space needs to be searched to find this estimator. Stochastic methods such as MCMC (Raftery, Madigan, and Hoeting 1997; Fernandez, Ley, and Steel 2001) and Gibbs sampling (George and McCulloch 1993) attempt to address this computational challenge. By approximating the posterior space, these approaches can address the issue of uncertainty of including specific regressors, but leave open the question of the length of the chain needed for convergence.

An exact representation of the posterior space for variable selection is always exponential, but an efficient calculation for each of the possible submodels reduces the time complexity to be similar to stochastic approximation methods (Ruggieri and Lawrence 2012). In this case, we address uncertainty without resorting to approximation techniques.

Step 1 of the Bayesian change point algorithm requires the calculation of the probability density of the data, $f(Y_{i:j}|X_{i:j})$, for a fixed subset of the regressors, X, and for each possible subinterval of the time series, $Y_{i:j}$, $1 \le i < j \le N$. The goal is to make inferences from the joint distribution, $f(Y, \beta, \sigma^2, \{C\}|X)$ [Equation (2)]. To relax the assumption of a fixed set of regressors, we carry out a variable selection procedure within each possible subinterval of the data, $Y_{i:j}$. Define A_m to be a vector of indicator variables for the inclusion or exclusion of each of the predictors from the set of predictor variables being considered. Given m predictor variables, there are 2^m possible subsets of variables to consider for each subinterval of the data, $Y_{i:j}$. Inferences on subsets of regressors for each subinterval can now be made from a more general joint distribution:

$$f(Y, \beta, \sigma^2, A_m, \{C\}|X) = f\left(Y, \beta, \sigma^2, \{C\}\,|A_m, X\right) \times f(A_m|X). \tag{11}$$

Once the probability density of the data has been calculated for each of the 2^m possible submodels, A_m, the choice of a submodel can be marginalized out in each substring of the data, $Y_{i:j}$, via model averaging:

$$f(Y_{i:j}) = \sum_{\text{all } A_m} f(Y_{i:j}|A_m, X_{i:j}) \times P(A_m|X_{i:j}), \tag{12}$$

where $P(A_m|X)$ is a product Bernoulli on the number of predictor variables included in the submodel, as defined below. The large number of subintervals that exist for any time series [O(N^2)] requires an efficient model selection procedure in order for the algorithm to be practical.

Let m_{inc} be the number of variables included in submodel A_m and let m_{exc} be the number of excluded variables ($m_{\text{inc}} + m_{\text{exc}} = m$, the total number of predictor variables). Associated with included variables is a "wide" prior variance parameter, k_{inc}, and associated with the excluded variables is the narrow prior variance parameter, k_{exc}. Let I_{A_m} be a diagonal matrix with either k_{inc} or k_{exc} on the diagonal, corresponding to whether or not a specific variable is included in the submodel being considered. Furthermore, define $v_N = v_0 + N$, $\beta^* = (X_{i:j}^T X_{i:j} + I_{A_m})^{-1} X_{i:j}^T Y_{i:j}$, and $s_N = (Y_{i:j} - X_{i:j}\beta^*)^T (Y_{i:j} - X_{i:j}\beta^*) + \beta^{*T} I_{A_m} \beta^* + v_0\sigma_0^2$. Finally, let p_{inc} be the probability of including a variable and let p_{exc} be the probability of excluding a variable ($p_{\text{inc}} + p_{\text{exc}} = 1$). Thus, the prior distribution on a submodel, A_m, is defined as $P(A_m) = p_{\text{inc}}^{m_{\text{inc}}} p_{\text{exc}}^{m_{\text{exc}}}$. The probability density of our data now takes the following form:

$$f(Y_{i:j}) = f(Y_{i:j}|X_{i:j}) = \sum_{\text{all } A_m} \iint f(Y|\beta, \sigma^2, A_m, X_{i:j}) f(\beta|\sigma^2, A_m, X_{i:j})$$
$$\times f(\sigma^2|A_m, X_{i:j})\, d\beta d\sigma^2 P(A_m|X_{i:j}),$$

or

$$f(Y_{i:j}) = \frac{(v_0\sigma_0^2/2)^{v_0/2}\Gamma(v_n/2)}{\Gamma(v_0/2)(2\pi)^{n/2}} \sum_{\text{All } A_m} \frac{\left(p_{\text{inc}}^2 k_{\text{inc}}\right)^{m_{\text{inc}}/2} \left(p_{\text{exc}}^2 k_{\text{exc}}\right)^{m_{\text{exc}}/2}}{(s_n/2)^{v_n/2} \left|X_{i:j}^T X_{i:j} + I_{A_m}\right|^{1/2}}. \tag{13}$$

The key to the EBIR algorithm (Ruggieri and Lawrence 2012) is to quickly derive $f(Y|\text{submodel } b)$ from $f(Y|\text{submodel } a)$. Note how only a determinant, matrix inversion (involved in the calculation of s_n), and sum on the number of included variables, needs to be completed for each submodel A_m, as the rest of the terms in the density function are independent of A_m. Therefore, the largest computational burden involved in the calculation of $f(Y_{i:j})$ is the evaluation of a matrix inverse and a matrix determinant. Consider a full binary tree whose depth is the number of possible variables where one branch adds/deletes a variable, while the other branch makes no changes to the current model. The EBIR algorithm works through the tree, performing a calculation only when the add/delete branch is traversed, and reduces matrix inversion from $O(m^{2.376})$ by the Coppersmith and Winograd (1990) algorithm to $O(m^2)$ and matrix determinant calculations from $O(m^{2.376})$ to O(1). Figure 1 provides visual aid with three possible variables. Each leaf on the tree holds the probability of one submodel: "0" represents an "excluded" variable while "1" represents an "included" variable. See Ruggieri and Lawrence (2012) for the full implementation details of EBIR. The EBIR calculation [Equation (13)] replaces Step 1 (calculating the probability density of the data), given above, carried out for each possible substring of the data. The updated algorithm is described in Appendix A (online supplementary materials).

4. PROOF OF CONCEPT: A SIMULATED DATASET

A simulated dataset was generated as proof of principle that the computer code works and numerical issues addressed so that the algorithm returns the expected results. Assume that we have 10 possible predictors: $X_1(t) = \sin(2\pi t/20)$, $X_2(t) = \sin(2\pi t/30)$, $X_3(t) = \sin(2\pi t/40)$, $X_4(t) = \sin(2\pi t/50)$, $X_5(t) = \sin(2\pi t/60)$, $X_6(t) = \sin(2\pi t/70)$,

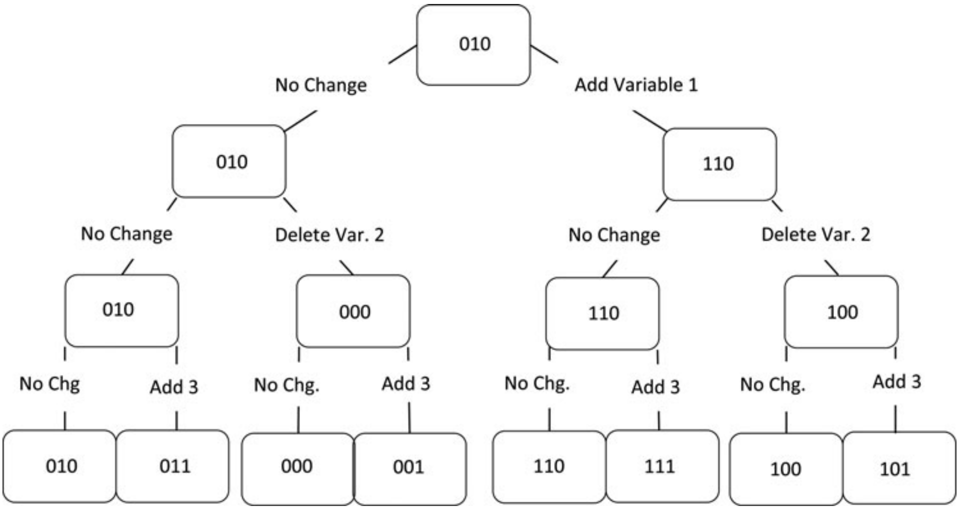

Figure 1. The recursive structure of the EBIR algorithm. Starting with any of the eight possible submodels from a set of three variables (in this case, 010), a move "left" keeps the current submodel, while a move "right" either adds or deletes the indicated variable, as appropriate. Only movements to the "right" in the tree require a calculation to be made. The recursive structure of this tree allows for a reduction in computational complexity by efficiently generating a child from its parent node.

$X_7(t) = \sin(2\pi t/80), X_8(t) = \sin(2\pi t/100), X_9(t) = \sin(2\pi t/150)$, and $X_{10}(t) = \sin(2\pi t/200)$. The dependent variable, Y, consists of 1000 randomly generated observations with four uniformly distributed change points. Regression coefficients are obtained from a mixture of normal distributions [0.5 × N(1,2) + 0.5 × N(−1,2)] and then Gaussian white noise of various levels is added [∼N(0,1), N(0, 1.5), and N(0,2)]. Using MATLAB R2011a's built-in random number generator, the following dataset was built:

$$Y(t) = \begin{cases} 3.3115X_2(t) - 0.2222X_7(t) + 0.7684X_8(t), & 1 \le t < 309, \\ -0.1844X_2(t) - 0.6663X_3(t) + 0.555X_4(t), & 309 \le t < 505, \\ -0.1844X_2(t) - 0.6663X_3(t) + 0.555X_4(t), & 505 \le t < 648, \\ 0.0468X_1(t) + 2.2412X_5(t) - 1.3785X_7(t) + 1.5937X_{10}(t), & 648 \le t < 751, \\ 2.3593X_3(t) - 1.1507X_5(t) - 1.3162X_7(t) - 0.3304X_9(t), & 751 \le t \le 1000. \end{cases}$$

The Bayesian change point and variable selection algorithm was run with $k_{\max} = 10$ and the parameter settings described in Appendix B (online supplementary materials).

Five hundred samples were drawn directly from the posterior distribution according to Step 3 (stochastic backtrace). The results for a noise level of 1.0 are shown in Figures 2 and 3. The posterior probability of selecting the proper number of change points, 4, is >0.99 for a noise level of 1.0 and >0.90 for a noise level of 1.5; this final change point is found less often when the noise is increased to 2.0. Figure 2 shows the recreated model in conjunction with the actual data and added noise. The locations of the sampled change points are indicated at the bottom of the figure. The height of the "spikes" is indicative of the number of times a change point is selected at that exact location. Tall "spikes" indicate relative certainty in the timing of a change point while wider "spikes" indicate a corresponding amount of uncertainty in the timing of the change point. As the level of noise increases, the change point locations remain accurate, but their distribution becomes

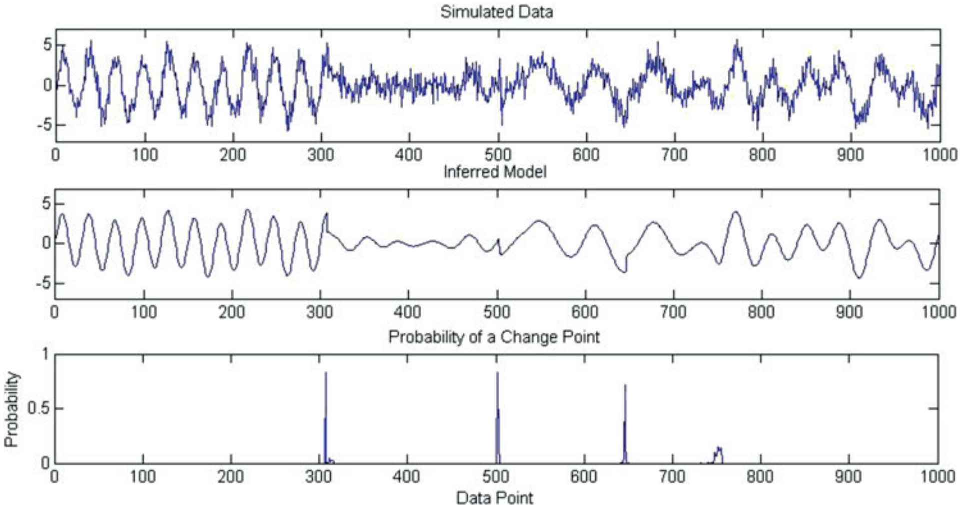

Figure 2. Simulated data and inferred model with change point locations indicated. The overall R^2 for a white-noise level of 1.0 is 0.790.

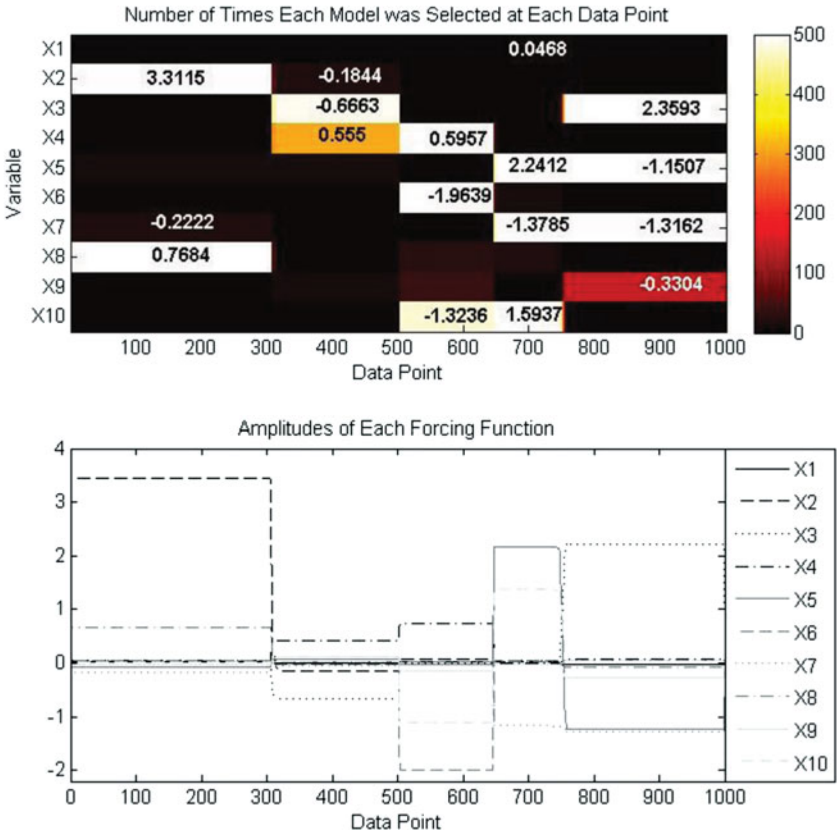

Figure 3. Variables selected for inclusion in the simulated dataset. (a) Of the 500 sampled solutions, the heat map displays the number of times that each variable is selected at each data point along with that variable's true regression coefficient; (b) the average inferred coefficient of each of the 10 possible regressors at each data point.

more spread around the true timing of the change. The overall average R^2 for a Gaussian white-noise level of 1.0 is 0.790, but drops to 0.716 for a white-noise level of 1.5, and 0.641 for a white-noise level of 2.0. A perfect recreation of the input model would yield R^2 values of 0.7811, 0.7041, and 0.6409, respectively, implying that the Bayesian change point and variable selection algorithm is able to faithfully reproduce the input, while fitting only a tiny amount of the added noise.

Figure 3 illustrates the results of variable selection. Figure 3(a) is a heat map displaying the number of times that each variable was selected at each data point along with the true regression coefficient. Small amplitude inputs (i.e., X_1 in $Y_{648:750}$) become overwhelmed by the noise in the system and so they are not selected by the algorithm. As discussed by George and McCulloch (1993), changing the parameters k_{inc} and k_{exc} can alter how conservative the algorithm is in its selection. Corresponding heat maps for larger values of white noise are structurally similar, although increases in the amount of added noise cause fewer of the variables to be selected in each subinterval. Figure 3(b) shows the average regression coefficient for each variable selected at each data point. These sampled coefficients match well with their true values for the variables that are often selected [Figure 3(a)]. However, the coefficient for a variable that is not often selected will be underestimated in proportion to the number of times it was sampled when averaged across all 500 sampled solutions.

5. APPLICATION: δ^{18}O PROXY RECORD OF THE PLIO-PLEISTOCENE

Today, most theories of ice sheet dynamics are a variation of Milankovitch theory (1941), which loosely states that ice sheets respond linearly to the amount of solar insolation (i.e., energy) received at the top of the Earth's atmosphere at 65°N latitude during the summer. Thus, we can use regression to model the glacial dynamics represented in the δ^{18}O proxy record as a function of solar insolation.

Changes in the amount of solar insolation received by the Earth are caused by variations in the Earth's orbit around the Sun. This motion can be described by three parameters: obliquity (tilt), precession (wobble), and eccentricity (ellipticity), each of which is nearly periodic. Whereas variations in precession ($\sim$23 kyr) and obliquity ($\sim$41 kyr) can alter the amount of solar insolation received during the summer at 65°N latitude by 30 W/m^2 and 15 W/m^2, respectively, variations in eccentricity ($\sim$100 kyr) alter the solar insolation budget by less than 1 W/m^2 (out of a total $\sim$500 W/m^2). Because the eccentricity signal in the solar insolation record is not sufficient to force the $\sim$100 kyr glacial cycles observed in the δ^{18}O proxy record after the MPT (Hays, Imbrie, and Shackleton 1976), several alternative hypotheses have been developed to explain the emergence of 100 kyr glaciation at this time. Here, we use the Bayesian change point and variable selection algorithm to compare the following four theories of ice sheet formation and destruction using sinusoidal approximations to each of the orbital components:

1. *A linear response to Milankovitch forcing*: Milankovitch forcing is composed of precession, obliquity, and eccentricity. Precession is represented by a 23 kyr sinusoid, obliquity by sinusoids at 41 and 53 kyr, and eccentricity by a triplet of sinusoids at 95, 124, and 404 kyr. Each of these frequencies represents the strongest spectral components for each of the orbital forcing functions (Berger and Loutre 1991; Imbrie et al. 1992).

2. *Harmonics of obliquity*: The harmonics (or bundles) of obliquity hypothesis (Huybers and Wunsch 2005; Liu, Cleaveland, and Herbert 2008) claims that ice sheets terminate with every second or third obliquity cycle after the MPT, whereas ice sheets terminated with every obliquity cycle prior to the MPT. Therefore, the second and third obliquity cycles are represented by 82 and 123 kyr sinusoids, respectively.

3. *Harmonics of precession*: The harmonics (or bundles) of precession hypothesis (Raymo 1997; Ridgwell, Watson, and Raymo 1999) claims that ice sheets grow when summer insolation is unusually low for a full precession cycle and once established, do not last beyond the next increase in summer insolation. Thus, ice sheets will terminate at the fourth or fifth precession cycle. The harmonics of precession are represented by sinusoids at 92, and 115 kyr, corresponding to the fourth and fifth precession cycles, respectively.

4. *Orbital inclination*: Muller and MacDonald (1995, 1997) claimed that the narrow spectral peak of orbital inclination at 100 kyr is a good match to the narrow spectral peak at 100 kyr in the δ^{18}O data. Therefore, we represent the orbital inclination hypothesis as a single 100 kyr sinusoid.

The question now becomes whether or not the full models, including the regressors that represent precession and obliquity, are necessary to model the proxy record. To reflect the uncertainty in these theories about the inclusion of all terms in a given interval, we allow our EBIR procedure to exclude regressors from a model to obtain a more realistic fit. Given these four models, the sinusoids used as input to the algorithm are 23, 41, 53, 82, 92, 95, 100, 115, 123, 124, and 404 kyr. The 123 and 124 kyr sinusoids have frequencies equivalent to three decimal places. As a result, they will be combined into a single sinusoid for this analysis. Because the linear response of ice sheets to precession (23 kyr) and obliquity (41 and 53 kyr) has not been questioned, these sinusoids are permitted to enter in each of the four models described above without restriction. However, because each of these hypotheses is exclusive of the others, any submodel containing sinusoids from two exclusive hypotheses is eliminated from consideration. For example, a permitted submodel cannot contain both the second harmonic of obliquity and the fourth harmonic of precession or the 100 kyr sinusoid for orbital inclination. Consequently, our 2^{11} possible submodels are reduced to 112 for each possible subinterval. A priori, we assume that all submodels are equally likely, that is, $p_{\text{inc}} = p_{\text{exc}} = 0.5$, and we set $v_0 = 10$ and $\sigma_0^2 =$ variance of the δ^{18}O proxy record (see Appendix B for a full description). The final parameter to specify is $k_{\max}$. While geoscientists expect at least two change points in δ^{18}O record, there is a serious doubt that there can be more than six change points. So, we set $k_{\max}$ equal to six. Additionally, because geoscientists are primarily interested in changes in the forcing functions of the climate system rather than the long-term cooling trend that exists in the δ^{18}O proxy record, we removed this long-term cooling trend from the dataset using an exponential function before analyzing the time series [Figure 4(a)]. If one were also interested in studying changes in trend, then additional predictor variables can be added to the regression model.

The analysis of the δ^{18}O proxy record (Lisiecki and Raymo 2005) by the Bayesian change point and variable selection algorithm is shown in Figures 4 and 5. Five hundred samples are drawn directly from the posterior distribution according to Step 3 (i.e., stochastic backtrace), to characterize the shape of the posterior space. Figure 4(b) shows the fit of our model (averaged over the 500 samples) to the δ^{18}O data [Figure 4(a)] with the marginal probability of a change point indicated as "spikes" in Figure 4(c). As before, the height of these spikes indicates the number of times that a given data point was selected as a change point, while the width of these spikes indicates the uncertainty in the timing of a change point. For some of the change points, the timing is relatively certain. For example, the change point at 338 thousand years ago (ka) is well defined with 0.610 of the posterior mass at exactly 338 ka [95% credibility limit (the Bayesian analog to a confidence interval) = 335–339 ka]. On the other hand, one of the change points identified with the MPT is less precisely located with 39.6% of the posterior mass at exactly 788 ka (95% credibility limit = 788–814 ka). In this analysis, the MPT is identified as a pair of change points, with the second being bimodal and centered at 1.22 Ma. Additionally, the intensification

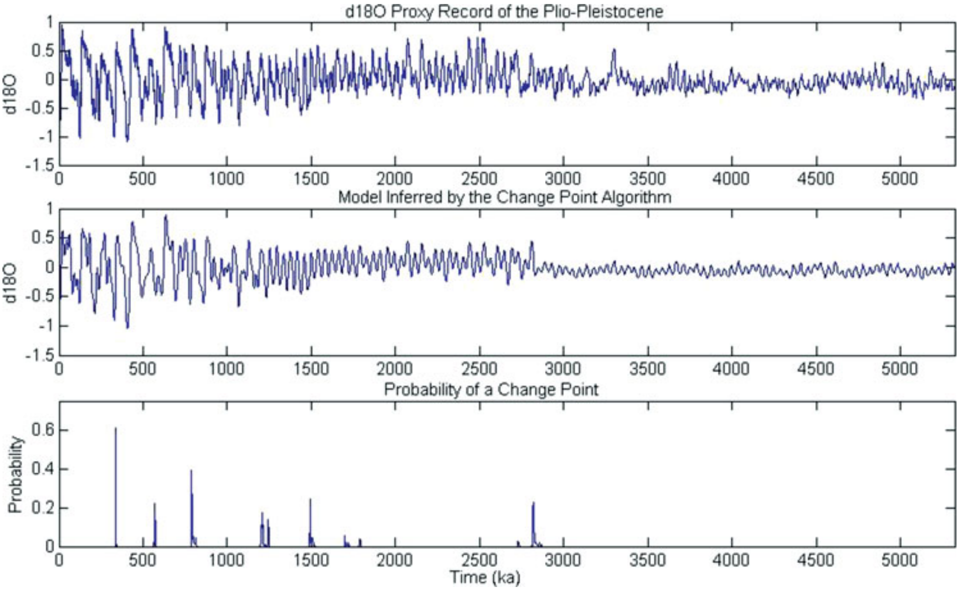

Figure 4. The δ^{18}O proxy record of the Plio-Pleistocene with its inferred model and the posterior probability of a change point. (a) A 5.3 million year global record of ice volume on the Earth compiled by Lisiecki and Raymo (2005) after removing the long-term cooling trend via an exponential function. (b) The model inferred by the Bayesian change point and variable selection algorithm. (c) The probability of a change point at a specific point in time and the uncertainty in its location can be determined by the height and width of the "spikes." The overall R^2 is 0.788.

of Northern Hemisphere glaciations is also bimodal, with 5.6% of its probability mass centered at 2.73 Ma and the remaining 94.4% centered at ~2.82 Ma.

Of course, since the algorithm draws change points jointly from their posterior distribution, this analysis is not restricted to an examination of marginal distributions. This

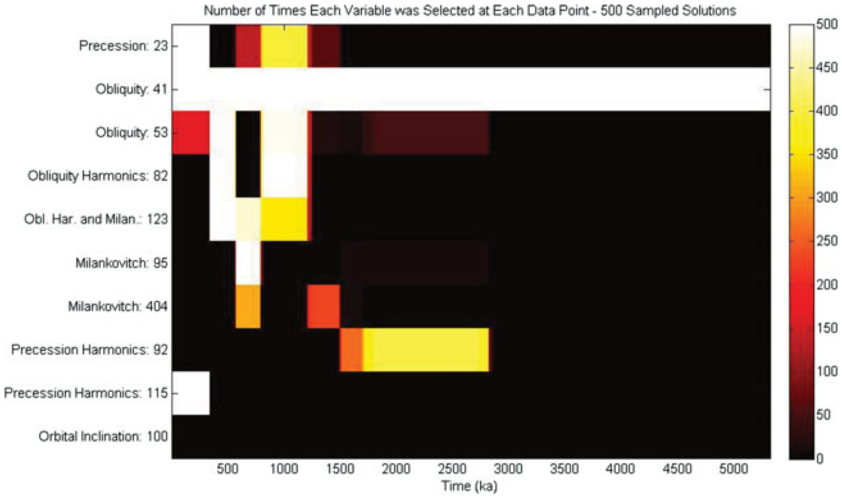

Figure 5. Heat map of the variables selected to fit the δ^{18}O proxy record.

is especially important if the change point locations are not independent of one another, which can call into question the convergence of an MCMC algorithm or the "optimality" of a greedy solution. In general, any algorithm that adds, subtracts, or alters one change point at a time may fail to find all sets of plausible solutions.

The double peak around 1.22 Ma suggests that the posterior distribution of change points may contain more than one class of favored solutions. Recall that in a Bayesian setting, inferences are based on the posterior probability distribution of the unknown parameters. Accordingly, there may be more than one location in the posterior space that has substantial posterior mass, as has been previously discovered through clustering in another discrete high-dimensional setting, the prediction of RNA secondary structures (Ding, Chan, and Lawrence 2005). Furthermore, since the most probable point estimate (i.e., the MAP) is optimal under a zero-one loss function, it is likely to have a low probability in high-dimensional spaces and may be located far from all regions that have high probability mass in the posterior space (Carvalho and Lawrence 2008). In such cases, alternative estimators, such as centroid estimators, are likely to return more representative estimates (Carvalho and Lawrence 2008). Following the lead of Ding, Chan, and Lawrence (2005), we investigate the potential of a multimodal posterior space by clustering our 500 joint samples of change point solutions using a k-means clustering algorithm (Seber 1984) with five clusters to obtain a better understanding of the shape of the posterior space. Each cluster is characterized by its centroid estimator (Carvalho and Lawrence 2008) and the credibility limits around these estimators, which represent the smallest hypersphere around the centroid that contains 95% of the posterior space (Webb-Robertson, McCue, and Lawrence 2008; Newberg and Lawrence 2009) (Table 1). Choosing fewer clusters acts to combine clusters 1–4, while choosing more clusters acts to split the existing clusters into nearly identical groups. As shown in Table 1, there are three interesting patterns to note:

1. The first three change point locations in all five clusters are essentially the same.
2. Cluster 5 has change points at 1.24 Ma and 1.73 Ma, while the other clusters lack these change points. In fact, for each sampled solution where there is a change point at 1.24 Ma, there is also a change point at 1.73 Ma, and very rarely is the 1.73 Ma change point included in any other sample. Marginally, these two points are a part of bimodal pairs.
3. The change points that identify the intensification of Northern Hemisphere glaciations at 2.73 Ma are almost always paired with change points at 1.21 and 1.49 or 1.50 Ma, which are the bimodal counterparts to the change points at 1.24 and 1.73 Ma described above.

Thus, the clustering of samples from the joint posterior distribution clearly illustrates the multimodality of the posterior space. Clustering algorithms such as k-means may find local rather than globally optimal solutions. However, multiple runs were conducted with little difference in the results.

Table 1. Clustering the change point solutions. The 500 change point solutions sampled from the posterior distribution according to Step 3 (stochastic backtrace) were distributed into five clusters using a *k*-means clustering algorithm. Listed above are the centroid and credibility limits (in parentheses) for each of the five clusters, rounded to the nearest ka

Cluster no.	Size of cluster	Change point 1	Change point 2	Change point 3	Change point 4	Change point 5	Change point 6
1	217	338 (335–340)	569 (561–573)	792 (788–812)	1208 (1202–1220)	1492 (1490–1510)	2820 (2730–2852)
2	80	338 (335–339)	568 (560–574)	794 (788–821)	1214 (1202–1244)	1507 (1496–1515)	2820 (2812–2830)
3	26	338 (336–339)	569 (560–573)	794 (788–808)	1210 (1200–1244)	1494 (1488–1502)	2730 (2725–2740)
4	33	338 (335–340)	569 (560–573)	798 (788–816)	1208 (1202–1222)	1496 (1490–1515)	2855 (2840–2870)
5	144	338 (335–339)	569 (562–573)	794 (788–812)	1242 (1204–1246)	1732 (1697–1792)	2820 (2730–2860)

Of the 500 sampled solutions, Figure 5 indicates the number of times that each variable was selected for inclusion at each data point. From this figure, several conclusions can be drawn:

1. The 41 kyr sinusoid representing obliquity is selected for every data point in at least 499 of the 500 sampled solutions.
2. While precession (23 kyr) may contribute more to solar insolation than any other term, it shows up strongly in the posterior only in the most recent subinterval (0–338 ka), with a posterior probability near 1, and in the MPT (790 ka to 1.22 Ma), with a posterior probability of roughly 0.8.
3. Terms exclusively from the Milankovitch model (95 and 404 kyr) are absent in the posterior except for the short interval from about 575 to 790 ka.
4. Before the onset of permanent glaciers in the Northern Hemisphere (∼2.8 Ma), only obliquity (41 kyr) is inferred by any of the 500 sampled solutions. Thus, none of the four models are appropriate for fitting this "41 kyr world" prior to 2.8 Ma. Additionally, with the exception of the 92 kyr sinusoid during 1.5–2.8 Ma, the 41 kyr sinusoid is the only frequency regularly chosen for the entire period prior to the MPT. Therefore, the MPT can be viewed as a period of transition from a "41 kyr world" to a more complex subharmonic model involving longer wavelengths, previously called the "100 kyr world."
5. The orbital inclination model (100 kyr) was not chosen in more than 1 of 500 sampled solutions at any point in time. Thus, the representation of the late Pleistocene glacial era as a "100 kyr world" is not supported in our findings.

6. DISCUSSION AND CONCLUSIONS

The Bayesian change point and variable selection algorithm makes two important contributions to the study of change point problems: (1) the algorithm generalizes the dynamic programming Bayesian change point algorithms of Liu and Lawrence (1999) and Fearnhead (2006) to include regression analysis; and (2) the algorithm combines variable selection with a change point technique to provide increased flexibility in statistical modeling. While other Bayesian algorithms may be able to perform change point analysis in a general regression setting, we know of no other algorithm that simultaneously performs variable selection. The EBIR algorithm facilitates allowing variables to be added or deleted as the data suggest and thus yields a more realistic model of the underlying phenomenon, especially at the boundaries of two regimes. The algorithm is therefore well suited to study datasets where the locations of the change points, the parameters of the model, and the included predictor variables are suspected to change through time, such as the δ^{18}O proxy record of the Plio-Pleistocene. Two major changes to this record have been thoroughly discussed in the literature: the intensification of glaciations in the Northern Hemisphere around 2.7 Ma and the MPT around 1 Ma, where not only did the glaciations increase in magnitude, but they changed in frequency as well. Thus, the former was a change in the parameter values, while the latter was a change in the included predictor variables.

The Bayesian change point algorithm is built upon a product partition model introduced by Barry and Hartigan (1993) using the same recursive procedure as its least-square counterpart (Auger and Lawrence 1989; Ruggieri et al. 2009), but remedies its shortcomings related to uncertainty estimates. The recursions that Barry and Hartigan (1993) used to obtain an exact solution to the change point problem are similar to those proposed here, except that their recursions stem from both ends of the dataset, rather than just one end. Their exact algorithm is $O(N^3)$ as opposed to our $O(N^2)$ exact algorithm because of differences in how the parameter values are calculated. First, instead of sampling, Barry and Hartigan (1993) calculated an expected value at each data point based on the probability that a given substring is included in the partition. Since there are $O(N^2)$ possible substrings, the parameter estimates are $O(N \times N^2) = O(N^3)$ instead of $O(kN)$, as in our approach. The R package "bcp" (Erdman and Emerson 2008) uses an MCMC approximation to Barry and Hartigan (1993) that is $O(N)$ per iteration. A second difference between the algorithms is that Barry and Hartigan (1993) only detected a change in the mean, whereas our Bayesian regression model characterizes the full posterior solution in all its potential complexity, including multiple high posterior regions. "bcp" shares this limitation.

Given the probabilistic framework of the Bayesian change point algorithm, inferences can now be made about the number of change points, their locations, and the parameters of the regression model. Moreover, we can also relax the assumption of a fixed set of predictors. These inferences replace the need for an arbitrary choice of the number of change points and the use of maximum likelihood estimates for the parameters of the regression model observed in the least-square setting. Four of the change points identified in this analysis are similar to those found by Ruggieri et al. (2009), specifically, the change points centered at ∼338, 790, 1208, and 2820 ka. The middle two change points correspond to a pair of changes that surround the MPT, while the final change point is consistent with the onset of Northern Hemisphere glaciation. Also, by considering the ensemble of sampled solutions, we now have an indication about the abruptness of change. Uncertainty in the placement of a change point indicates the presence of a gradual change.

The Bayesian change point algorithm in its current form is an example of interrupted regression (Marsh and Cormier 2001) and thus does not have a continuity constraint. Therefore, the algorithm allows for both gradual and abrupt changes to take place. Given the debate in the geosciences community concerning abrupt versus gradual changes, a model of this form is well suited to the analysis of glacial cycles. If continuity is desired, spline regression techniques can instead be used, but the problem then becomes nonlinear in its parameters and approximation techniques must instead be used (Marsh and Cormier 2001). The model selection procedure outlined here, and more generally the change point algorithm itself, is not restricted to sinusoidal functions. The algorithm can be adapted to fit simpler (i.e., the mean) or more complex functions so long as the density function for the residual error can be calculated in any subinterval for all allowable submodels.

Simulation results (Figures 2 and 3) show highly accurate placement of change points, but a conservative variable selection algorithm. Of the variables that are chosen, their inferred amplitudes are quite similar to their true values. On the other hand, the contributions of some variables are overwhelmed by the added noise and therefore are not selected. Longer intervals between change points will, in general, lead to more accurate inferences.

The results of the Bayesian change point and variable selection algorithm on the δ^{18}O proxy record of the Plio-Pleistocene can be sensitive to the values chosen for the prior parameters υ_0 and σ_0^2, an effect not observed in the simulated data. Specifically, the larger the product of these two parameters, the fewer is the number of change points chosen by the algorithm. The prior distribution on the error variance is analogous to adding pseudo data points of a given residual variance and helps to bound the likelihood function. In the maximum likelihood setting, this effect is similar to penalized likelihood techniques (Ciuperca, Ridolfi, and Idier 2003). As noted by Fearnhead (2006), the choice of prior distribution can affect the number, but not the distribution of the change point positions. See Appendix B for a full description of these parameters.

Overall, six change points fit 78.8% of the variance in the δ^{18}O proxy record of Lisiecki and Raymo (2005). By clustering the 500 sampled solutions, we obtain a clearer understanding of the shape of the posterior space. In particular, we find that our bimodal change point locations are not independent. While the centroids of the clusters for the three change points since the MPT are nearly identical, it appears that our change points prior to the MPT tend to occur in pairs. For example, a change point at 1.24 Ma is paired with a change point at 1.73 Ma, while a change point at 1.21 Ma is paired with a change point at ~1.50 Ma.

The number of samples can be kept small (relative to an MCMC approach) because these samples are drawn directly and independently from the posterior distribution. With MCMC algorithms, including those designed to address the change point algorithm, convergence is required to obtain samples from the target distribution, but rigorous procedures to assure such convergence are unavailable. This stands in stark contrast with the algorithm presented in this article. Since all required sums and integrals can be simultaneously completed, all samples of all unknowns from this algorithm are drawn directly and independently from the target distribution. Samples drawn directly from their posterior distributions avoid the convergence issues associated with MCMC algorithms and permit direct assessment of uncertainly in the models, parameter values, and hidden variables.

The models being used to study the δ^{18}O proxy record are obviously a simplification of the real-life phenomena associated with the Earth's glacial system. However, capturing the dominant periodicities in the proxy record and determining when they change is of importance to geoscientists as they often describe ice volume in terms of glacial cycles (see, e.g., Raymo 1997; Ridgwell, Watson, and Raymo 1999; Huybers and Wunsch 2005). A more complete model could take into account local autocorrelations in the data, time-varying amplitudes and phases, and uncertainties in the dating of the proxy record, which could account for some of the nonstationarities in the data.

In the post-MPT era, the strength of the contributions of each term is not specified by the authors of the four theories being compared. Because a regression model is used for this analysis, each of the models attempts to describe the full pattern of variation in the δ^{18}O record, rather than solely describing the timing or frequency of deglaciations as in their original formulations. Thus, these models may overspecify the intent of their authors. Accordingly, attention is focused on the inclusion/exclusion of model components.

Given the omnipresence of an obliquity term in the δ^{18}O record, it is perhaps better to think of the interval before the onset of continent-sized glaciers in the Northern Hemisphere around 2.7 Ma as an interval that is exclusively described by obliquity rather than a distinct "41 kyr world." The interval between 2.7 Ma and the MPT appears to be a transitional period in which the subharmonics of precession terms begin to play an important role. After the MPT, models that include multiples of the 41 kyr obliquity signal and the 23 kyr precession signal dominate, running counter to the concept of a "100 kyr world" described as a single 100 kyr glacial cycle. Together, these findings support the notion that the MPT represents a transition from a glacial regime that responded linearly to the obliquity component of solar insolation to a subharmonic glacial regime that mechanistically has been described by the nonlinear phase-locking models proposed by Tziperman et al. (2006). The nonlinear phase-locking models allow the frequency of individual glacial cycles to change through time due to variations in insolation forcing and can account for the presence of both obliquity and precession subharmonics found in the proxy record. Given that none of the existing theories by themselves are sufficient to describe the ice volume proxy data after the MPT, further modeling, perhaps including proxies for other components of the system such as carbon dioxide and methane, will be required to understand the evolution in the Earth's glacial system over the last five million years.

SUPPLEMENTARY MATERIALS

1. **MATLAB code for the Bayesian change point and variable selection algorithm** (Bayes_Chgpt_VS.zip). This file contains the MATLAB code (main script and supporting functions) needed to run the algorithm described in this article. Also included are the two datasets (simulation and δ^{18}O proxy record) described in this article and an additional small simulation suitable for a quick review. A detailed readme file (README.txt) describes each of the functions and datasets included in the .zip file as well as run times for each of the datasets.
2. **Appendix A:** The Bayesian change point and variable selection algorithm— implementation details of the combined Bayesian change point and variable selection algorithm.
3. **Appendix B:** Parameter settings for the Bayesian change point and variable selection algorithm—a detailed description of each of the parameters, their default values, and discussions on how changes in these values may affect results.
4. **Appendix C:** Glossary of terms—a brief description of all variables used in the article.

ACKNOWLEDGMENTS

The authors thank the associate editor and the two anonymous reviewers whose comments helped to greatly improve this article. Our work was supported in part by a grant from the National Science Foundation, NSF/CMG 1025438 (C.E. Lawrence, PI).

[Received October 2011. Revised June 2012.]

REFERENCES

Auger, I. E., and Lawrence, C. E. (1989), "Algorithms for the Optimal Identification of Segment Neighborhoods," *Bulletin of Mathematical Biology*, 51, 39–54. [89,105]

Bai, J., and Perron, P. (2003), “Computation and Analysis of Multiple Structural Change Models,” *Journal of Applied Economics*, 18, 1–22. [89]

Barry, D., and Hartigan, J. A. (1993), “A Bayesian Analysis for Change Point Problems,” *Journal of the American Statistical Association*, 88, 309–319. [89,105]

Berger, A. L., and Loutre, M. F. (1991), “Insolation Values for the Climate of the Last 10 Million Years,” *Quaternary Science Reviews*, 10, 297. [99]

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992), “Hierarchical Bayesian Analysis of Changepoint Problems,” *Applied Statistics*, 41, 389–405. [89]

Carvalho, L. E., and Lawrence, C. E. (2008), “Centroid Estimators for Inference in High-Dimensional Discrete Spaces,” *Proceedings of the National Academy of Sciences of the United States of America*, 105, 3209–3214. [102]

Chib, S. (1998), “Estimation and Comparison of Multiple Change-Point Models,” *Journal of Econometrics*, 86, 221–241. [89]

Chopin, N. (2007), “Dynamic Detection of Change Points in Line Time Series,” *Annals of the Institute of Statistical Mathematics*, 59, 349–366. [89]

Ciuperca, G., Ridolfi, A., and Idier, J. (2003), “Penalized Maximum Likelihood Estimator for Normal Mixtures,” *Scandinavian Journal of Statistics*, 30, 45–59. [106]

Coppersmith, D., and Winograd, S. (1990), “Matrix Multiplication via Arithmetic Progressions,” *Journal of Symbolic Computation*, 9, 251–280. [96]

Ding, Y. E., Chan, C. Y., and Lawrence, C. E. (2005), “RNA Secondary Structure Prediction by Centroids in a Boltzman Weighted Ensemble,” *RNA*, 11, 1157–1166. [102]

Erdman, C., and Emerson, J. (2008), “A Fast Bayesian Change Point Analysis for the Segmentation of Microarray Data,” *Bioinformatics*, 24, 2143–2148. [89,105]

Fearnhead, P. (2006), “Exact and Efficient Bayesian Inference for Multiple Changepoint Problems,” *Statistics and Computing*, 16, 203–213. [90,104,106]

Fernandez, C., Ley, E., and Steel, M. F. J. (2001), “Benchmark Priors for Bayesian Model Averaging,” *Journal of Econometrics*, 100, 381–427. [94]

George, E. I., and McCulloch, R. E. (1993), “Variable Selection via Gibbs Sampling” *Journal of the American Statistical Association*, 88, 881–890. [94,98]

Green, P. J. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” *Biometrika*, 82, 711–732. [89]

Hawkins, D. M. (1976), “Point Estimation of the Parameters of Piecewise Regression Models,” *Journal of the Royal Statistical Society,* Series C, 25, 51–57. [89]

——— (2001), “Fitting Multiple Change-Point Models to Data,” *Computational Statistics and Data Analysis*, 37, 323–341. [89]

Hays, J. D., Imbrie, J., and Shackleton, N. J. (1976), “Variations in the Earth’s Orbit: Pacemakers of the Ice Ages,” *Science*, 194, 1121–1132. [88,99]

Huybers, P., and Wunsch, C. (2005), “Obliquity Pacing of the Late Pleistocene Glacial Terminations,” *Nature*, 434, 491–494. [99,106]

Imbrie, J., Boyle, E. A., Clemens, S. C., Duffy, A., Howard, W. R., Kukla, G., Kutzbach, J., Martinson, D. G., McIntyre, A., Mix, A. C., Molfino, B., Morley, J. J., Peterson, L. C., Pisias, N. G., Prell, W. L., Raymo, M. E., Shackleton, N. J., and Toggweiler, J. R.. (1992), “On the Structure of Major Glaciation Cycles: 1. Linear Responses to Milankovitch Forcing,” *Paleoceanography*, 7, 701–738. [99]

Imbrie, J., and Imbrie, J. Z. (1980), “Modeling the Climatic Response to Orbital Variations,” *Science*, 207, 943–953. [88,88]

Killick, R., and Eckley, I. A. (2011), *“Changepoint: An R Package for Changepoint Analysis,”* R Package version 0.6 [online]. Available at *http://cran.r-project.org/web/packages/changepoint/index.html*. [89]

Koop, G., and Potter, S. M. (2007), “Estimation and Forecasting in Models With Multiple Breaks,” *Review of Economic Studies*, 74, 763–789. [90]

——— (2009), “Prior Elicitation in Multiple Change-Point Models,” *International Economic Review*, 50, 751–772. [90]

Lavielle, M., and Lebarbier, E. (2001), “An Application of MCMC Methods for the Multiple Change-Points Problem,” *Signal Processing*, 81, 39–53. [89]

Lisiecki, L. E., and Raymo, M. E. (2005), “A Pliocene-Pleistocene Stack of 57 Globally Distributed Benthic d18O Records,” *Paleoceanography*, 20, PA1003. [88,88,106]

Liu, J. S., and Lawrence, C. E. (1999), “Bayesian Inference on Biopolymer Models,” *Bioinformatics*, 15, 38–52. [90,104]

Liu, Z., Cleaveland, L. C., and Herbert, T. D. (2008), “Early Onset and Origin of 100-kyr Cycles in Pleistocene Tropical SST Records,” *Earth and Planetary Science Letters*, 265, 703–715. [99]

Marsh, L. C., and Cormier, D. R. (2001), *Spline Regression Models (*Sage University Papers Series on Quantitative Applications in the Social Sciences*, 07-137)*, Thousand Oaks, CA: Sage Publications. [105]

Milankovitch, M. (1969), *Canon of Insolation and the Ice-Age Problem (*Israel *Program for Scientific Translations, Translator)*, Washington, DC: U.S. Department of Commerce and National Science Foundation. (Original work published 1941). [88]

Miller, A. J. (2002), *Subset Selection in Regression* (2nd ed.), New York: Chapman & Hall. [94]

Mitchell, T. J., and Beauchamp, J. J. (1988), “Bayesian Variable Selection in Linear Regression” (with discussion), *Journal of the American Statistical Association*, 83, 1023–1036. [94]

Muggeo, V. M. R. (2008), “Segmented: An R Package to Fit Regression Models With Broken-Line Relationships,” *R News*, 8, 20–25. [89]

Muggeo, V. M. R., and Adelfino, G. (2011), “Efficient Change Point Detection for Genomic Sequences of Continuous Measurements,” *Bioinformatics*, 27, 161–166. [89]

Muller, R., and MacDonald, G. (1995), “Glacial Cycles and Orbital Inclination,” *Nature*, 377, 107–108. [100]

——— (1997), “Glacial Cycles and Astronomical Forcing,” *Science*, 277, 215. [100]

Newberg, L. A., and Lawrence, C. E. (2009), “Exact Calculation of Distributions on Integers, With Application to Sequence Alignment,” *Journal of Computational Biology*, 16, 1–18. [102]

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004), “Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data,” *Biostatistics*, 5, 557–572. [88]

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006), “Forecasting Time Series Subject to Multiple Structural Breaks,” *Review of Economic Studies*, 73, 1057–1084. [90]

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), “Bayesian Model Averaging for Linear Regression Models,” *Journal of the American Statistical Association*, 92, 179–191. [94]

Raymo, M. E. (1997), “The Timing of Major Terminations,” *Paleoceanography*, 12, 577–585. [88,99,106]

Raymo, M. E., Lisiecki, L. E., and Nisancioglu, K. H. (2006), “Plio-Pleistocene Ice Volume, Antarctic Climate, and the Global d18O Record,” *Science*, 313, 492–495. [88]

Ridgwell, A. J., Watson, A. J., and Raymo, M. E. (1999), “Is the Spectral Signature of the 100 kyr Glacial Cycle Consistent With a Milankovitch Origin?,” *Paleoceanography*, 14, 437–440. [99,106]

Ruggieri, E., Herbert, T., Lawrence, K. T., and Lawrence, C. E. (2009), “Change Point Method for Detecting Regime Shifts in Paleoclimatic Time Series: Application to d18O Time Series of the Plio-Pleistocene,” *Paleoceanography*, 24, PA1204. [89,105,105]

Ruggieri, E., and Lawrence, C. E. (2012), “On Efficient Calculations for Bayesian Variable Selection,” *Computational Statistics and Data Analysis*, 56, 1319–1332. [90,95,96]

Seber, G. A. F. (1984), *Multivariate Observations*, Hoboken, NJ: Wiley. [102]

Stephens, D. A. (1994), “Bayesian Retrospective Multiple-Changepoint Identification,” *Applied Statistics*, 43, 159–178. [89]

Tibshirani, R., and Wang, P. (2008), “Spatial Smoothing and Hot Spot Detection for CGH Data Using the Fused Lasso,” *Biostatistics*, 9, 18–29. [88]

Tziperman, E., and Gildor, H. (2003), “On the Mid-Pleistocene Transition to 100-kyr Glacial Cycles and the Asymmetry Between Glaciation and Deglaciation Times,” *Paleoceanography*, 18, 1001. [88]

Tziperman, E., Raymo, M. E., Huybers, P., and Wunsch, C. (2006), “Consequences of Pacing the Pleistocene 100 kyr Ice Ages by Nonlinear Phase Locking to Milankovitch Forcing,” *Paleoceanography*, 21, PA4206. [107]

Webb-Robertson, B. J. M., McCue, L. A., and Lawrence, C. E. (2008), “Measuring Global Credibility With Application to Local Sequence Alignment,” *PLoS Computational Biology*, 4, e1000077. [102]

Western, B., and Kleykamp, M. (2004), “A Bayesian Change Point Model for Historical Time Series Analysis,” *Political Analysis*, 12, 354–374. [89]

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002), “Strucchange: An R Package for Testing for Structural Change in Linear Regression Models,” *Journal of Statistical Software*, 7, 1–38. [89]