BIGMACS v.4.7.1 Manual

Taehee Lee[1], Devin Rand[2], Lorraine Lisiecki[2], Geoffrey Gebbie[3], Charles Lawerence[4]

[1]Statistics Department, Harvard University, Cambridge, USA
[2]Department of Earth Science, University of California Santa Barbara, Santa Barbara 93106, USA
[3]Physical Oceanography Department, Woods Hole Oceanographic Institute, Falmouth, 02543, USA
[4]Applied Mathematics, Brown University, Providence, 02906, USA

Table of Contents

1. Introduction

BIGMACS can construct radiocarbon age models, benthic δ^{18}O-aligned age models, multiproxy age models (age models which leverage age information from both dating techniques), and benthic δ^{18}O stacks. Modeled sedimentation rates are realistically constrained using an empirically derived prior distribution based on the observed sedimentation rates of 37 radiocarbon dated cores (Lin et al., 2014). The stacking capability allows users to construct improved benthic d18O alignment targets that share similar water mass histories with their input cores.

Here we provide instructions to download BIGMACS, prepare the required input files, construct age models or a stack, and analyze the results. While we give a brief explanation of each parameter, most settings should not be changed from the default values. It is our hope that BIGMACS provides its users with a less subjective and more realistic method to construct ocean sediment core age models and benthic δ^{18}O stacks.

Happy chronology building!

2. Installation

INSTRUCTIONS TO DOWNLOAD THE SOFTWARE PACKAGE, VERSION OF MATLAB THAT IS COMPATIBLE, REQUIRED TOOL BOXES.

3 Preparing input files

To construct age models or a stack, create a new folder in BIGMACS/Inputs with the desired run name. BIGMACS will query all the necessary settings and proxy data from text files saved within this folder and the default folder (‘BIGMACS\Default’). All proxy data text files should be tab delimited and include column headers. See BIGMACS\Inputs\DNEA for example file structure and formatting. Settings that remain unchanged from their default values do not need to be specified in the run folder. See BIGMACS\Defaults for the default settings text files.

3.1 Proxy Data

Create a new folder titled ‘records’ in the run folder. Then create a folder for each input core within the records folder (e.g., BIGMACS/Inputs/[*run name*]/records/[*core name*]. These folders will hold the proxy data for each core and any core specific settings specified in ‘setting_core.txt’. Each core folder can store the following text files depending on the available proxy data and desired settings.

1. **C14_data.txt**: stores the radiocarbon and calibration data for each core and has the following six columns: (1) depth m, (2) ^{14}C age kyr, (3) ^{14}C standard deviation kyr, (4) reservoir age kyr, (5) reservoir age standard deviation kyr, and (6) the desired calibration curve (‘1’ for Intcal20, ‘2’ for Marine20, ‘3’ for SHCal20, and ‘4’ for a custom curve).

2. **d18O_data.txt**: stores the δ^{18}O data for each core and has two columns: (1) depth m, and (2) δ^{18}O ‰.

3. **additional_ages.txt**: This text file stores any additional age information to be included in the age model construction process (e.g., tephra layers, tie points, magnetic reversals, etc.). The user has the option to model each additional age as a Gaussian distribution or a uniform distribution. This text file has 4 columns: (1) depth m, (2) age kyr, (3) error kyr, and (4) mode. To model a given age as a Gaussian distribution, enter a ‘1’ in column four. A ‘0’ will indicate a uniform distribution. The error specified in column 3 will be the standard deviation for the Gaussian distribution and the half-width for the uniform distribution, such that the uniform distribution is centered on the value in column 2 and has the range column2 +/- column4.

4. **additional_depths.txt**:

5. **setting_core.txt**: This file does not need to be included if all settings are the same as the default settings. For a description of each setting, see the appendix below. For an example of this text file with the default settings BIGMACS\Defaults\setting_core.txt.

If the user is constructing a stack, all cores must have benthic d18O data. However, during age model construction, input cores with different combinations of proxy data can be included in the same run. If a core does not have a certain proxy type, that text file does not need to be included in the core folder.

3.2 Run Settings

The following are the text files that set the run-specific settings. For an example of these text files with their default values, see BIGMACS\Defaults. If you are not constructing a stack, you do not need to include setting_stacking.txt. If you are not changing the settings of a text file from the default values, you do not need to include that text file in the run folder.

1. **setting_alignment.txt**
    - data_type (‘both’, ‘d18O’, ‘C14’): Indicates the proxy data that should be used in the run. The additional age data will be used independent of the data type selected. If ‘both’ is selected, all available proxy data saved in each core folder will be used. If constructing a stack, the user must select either ‘both’ or ‘d18O’.
    - Islearn_transition (‘global’ or ‘no’): Indicates whether the state change probabilities saved in ‘transition_parameter.txt’ should remain fixed or optimized throughout the run. Options are ‘global’ which indicates optimization and ‘no’ which will preserve the state change probabilities.
    - nSamples_learning (number): Indicates the number of sample age models that are used to learn parameters. Default value is 100.
    - nSamples_drawing (number): Defines the number of sampled age paths used to draw the final age model. The default is 1000.
    - Stack_min (number): defines the minimum allowed age for the sampled age paths. Default is –inf.
    - Stack_max (number): defines the maximum allowed age for the sampled age paths. Default is inf.

2. **setting_stacking.txt**
    - variance (‘homoscedastic’ or ‘heteroscedastic’): Defines whether the variance of the constructed stack is constant (‘homoscedastic’) or time-dependent (‘heteroscedastic’). The default is ‘heteroscedastic’.
    - kernel_function (‘OU’, ‘SE’, ‘M25’, ‘M15’): Specifies the kernel covariance function that will be used during stack construction. The settings correspond to Ornstein-Uhlenbeck, square-exponential, Matérn kernel with the order 5/2, and Matérn kernel with the order 3/2.
    - start_age (number): Defines the start age of the new constructed stack. Default is –inf, meaning the start age of the new stack will be identical to the start age of the initial stack.
    - end_age (number): Defines the end age of the new constructed stack. Default is –inf, meaning the start age of the new stack will be identical to the start age of the initial stack.

3. **stack.txt**: The initial stack for stack construction or the alignment target for age model construction. Users can enter their own stack using a text file with three columns: (1) age kyr, (2) δ^{18}O ‰, (3) standard deviation ‰. The user can also use one of the default stacks provided, which are the regional 150 kyr stacks from Lisiecki & Stern (2016) and the LR04 stack. To use one of the default stacks, write the abbreviated name of the stack in ‘stack.txt’. For example, to use the deep North Atlantic stack, just write DNA in the stack.txt input file. The abbreviated names for each stack are the names of the text files saved in ‘BIGMACS\Defaults\Regional_Stacks

4. **list_of_figures.txt**: A list of figures for BIGMACS to save in the output folder. ‘age_vs_d18O’ plots the shifted and scaled δ^{18}O aligned to the target. ‘age_vs_depth’ plots the age models with radiocarbon data, additional age data, and indicates the depths of the δ^{18}O data. ‘age_vs_sedrate’ plots the sedimentation rate, the time-dependent average sedimentation rate, and the log of the normalized sedimentation rate. ‘stack_summary’ plots the final stack with the shifted and scaled δ^{18}O data from each core, the standardized δ^{18}O residuals, and the new stack vs. the initial stack.

5. **hyperparameters.txt**: We discourage users from changing the settings saved in this text file. For a description of each setting, see the appendix.

6. **Transition_parameter.txt**: We discourage users from changing the settings saved in this text file. For a description of each setting, see the appendix.

4 Running BIGMACS

BIGMACS is run from the script BIGMACS/main.m. Set the variable 'inputMode' equal to 'stacking' if you want to construct a stack or 'alignment' if you want to construct age models. The variable 'inputFile' indicates the specific run and should be set equal to the same name as the run folder. Once ready, execute the script. For a single core, constructing an age model should take only a few minutes. Constructing a stack from a handful of cores can take hours. The exact run time depends on the number and length of cores, the resolution of the proxy data, and the length of the stack.

5 Output Files

All output files will be saved in a folder titled 'BIGMACS/Outputs/[*run name_proxy type*]. The output folder will include the following text files.

1. Age models for each core will be saved in the folder BIGMACS/Outputs/[*run name_proxy type*]/ages. Each text file will contain depth, 95% and 68.2% confidence bands, the mean and the median.
2. If radiocarbon ages were used, the calibrated radiocarbon ages will be saved in BIGMACS/Outputs/[*run name_proxy type*]/C14_ages.
3. All figures will be saved in BIGMAS/Outputs/[*run name_proxy type*]/figures.
4. If a stack was constructed, the final stack will be saved in a text file titled 'stack.txt'. In addition, two types of stack samples will be saved in the text files stack_samples_mean.txt and stack_samples_noisy.txt The file stack_samples_mean.txt includes averages of the Gaussian process regressions while stack_samples_noisy.txt include sample stack paths that include the appropriate variance.
5. All input settings, proxy data, and finalized parameters are also saved in 'results.mat'. This file includes five structures.
    a. CI_C14: contains the calibrated radiocarbon ages, including the calendar age samples for each age.
    b. Samples: includes the final MCMC age model samples and an outlier flag for each data point on each age model sample (with a '1' indicating an outlier)
    c. Data: contains all proxy data for each core, the final δ^{18}O shift and scale parameters, the learned average sedimentation rate 'R', final state change probabilities (phi_I', 'phi_C', and 'phi_M'), and all input settings from 'setting_core.txt'
    d. Setting: the input settings from 'setting_alignment.txt' and 'setting_stacking.txt'
    e. 'param' contains all hyperparameter settings and learned hyperparameter values during the run time.
    f. 'target' contains the final stack ('stack'), the calibration curve ('cal_curve'), the initial target stack ('init_stack'), and the stack samples ('stack_sample').

6. Time Complexity

All age models and stacks presented in Lee & Rand et al., () were constructed on a standard desktop machine. However, longer stacks constructed from a large number of high resolution cores may have run times that require a computing cluster. Here we provide time complexity equations to give users a run-time estimate.

Age models are constructed in parallel and the time complexity depends on the number of input cores (L) and the number of available CPU processors (C). During age model construction, parameter values are estimated first and then ages are sampled. Parameter estimation requires the particle smoothing algorithm, the Metropolis Hastings algorithm, and the Baum-Welch Expectation Maximization algorithm. Once parameters are estimated, ages are sampled with the particle smoothing algorithm and Metropolis-Hastings algorithms. Particle smoothing requires two steps: a forward step and a backward step. During the forward step the time complexity is quadratic to the number of particles (P, default is 100) and linear to the number of proxy observations (N). The backward step is linear to the number of particles, proxy observations, and age model samples (M_0, default is 100). The Metropolis Hastings algorithm has a time complexity linear to the number of proxy observations, steps until the burn-in phase (B, default is 500) and age model samples (M_0). The total time complexity for a single iteration to learn parameter values is equal to $\mathcal{O}\left(\frac{\mathrm{L}}{\mathrm{C}}(\mathrm{P}^2\mathrm{N} + \mathrm{PNM}_0 + \mathrm{BNM}_0)\right)$. Once the parameters are estimated, ages are sampled. If the number of age model samples is set to M (the default is 1000) and the maximum number of iterations in parameter estimation is equal to R (default is 10), the total time complexity for age model construction is equal to $\mathcal{O}\left(\frac{\mathrm{L}}{\mathrm{C}}R(\mathrm{P}^2\mathrm{N} + \mathrm{PNM}_0 + \mathrm{BNM}_0) + \frac{\mathrm{L}}{\mathrm{C}}(\mathrm{P}^2\mathrm{N} + \mathrm{PNM} + \mathrm{BNM})\right)$. The multiproxy age model for GIK13289-2 (which has 30 δ^{18}O data points and 12 radiocarbon ages) took approximately 86 seconds to run on a standard desktop machine.

Stack construction iterates between an age model construction step and a stack updating step. The latter consists of kernel parameter estimation, δ^{18}O outlier classification, heteroscedastic variance estimation and the Gaussian process regression. Kernel parameter estimation requires a fixed number of iterations (S, default is 3000), with each iteration having a time complexity quadratic to the number of induced pseudo-inputs (fixed to N_0, sampled every 0.5 kyr, see S5 for details) and linear to the number of proxy observations. Time complexity for outlier classification is linear to the number of age model samples and proxy observations. Heteroscedastic variance estimation requires computations proportional to the number of age model samples and quadratic to the number of total proxy observations. Finally, the Gaussian process regression has a time complexity linear to the number of sampled age paths, total proxy observations and the length of the stack (K), and quadratic to the number of induced pseudo-inputs. Therefore, the total time complexity for one stack updating step is $\mathcal{O}\left(\mathrm{T}(\mathrm{SN}_0^2\mathrm{LN} + \mathrm{LNM}_0 + \mathrm{L}^2\mathrm{N}^2\mathrm{M}_0 + \mathrm{N}_0^2\mathrm{LNM}_0\mathrm{K})\right)$.

The stack construction algorithm includes A (default is 5) stack updating steps, and each update includes a new set of age models (i.e., an age model construction step). Thus the total time complexity to construct a stack is equal to $\mathcal{O}\left(\mathrm{A}\left(\frac{\mathrm{L}}{\mathrm{C}}R(\mathrm{P}^2\mathrm{N} + \mathrm{PNM}_0 + \mathrm{BNM}_0) + \mathrm{T}(\mathrm{SN}_0^2\mathrm{LN} + \mathrm{LNM}_0 + \mathrm{L}^2\mathrm{N}^2\mathrm{M}_0 + \mathrm{N}_0^2\mathrm{LNM}_0\mathrm{K})\right)\right)$. The DNEA stack, (which contains 6 cores, 2,112 δ^{18}O data points, 150 radiocarbon ages, and extends to 150 kyr) has a total run time of 1.8 hours.

Appendix

1. setting_core.txt

This text file stores core specific settings and should be saved in the core folders if default settings are not used.

- start_depth (number, m): Indicate the first depth to infer ages. The default is NaN.
- End_depth (number, m): Indicate the last depth to infer ages. The default is NaN.
- Initial_shift (number, ‰): Specify the initial shift for d18O alignment. The default is NaN.
- Initial_scale (number): Specify the initial scale for d18O alignment. The default is 1.
- Initial_average_sed_rate: Suggest an initial average sed rate.
- Islearn_shift ('yes' or 'no'): Specify whether the core specific shift and scale parameters are learned during alignment using the Baum-Welch Expectation Maximization algorithm.
- Islearn_scale ('yes' or 'no'): Specify whether the core specific shift and scale parameters are learned during alignment using the Baum-Welch Expectation Maximization algorithm.
- Islearn_average_sed_rate ('adaptive', no', 'constant'): Indicates how the average sed rate will be learned. If 'adaptive', BIGMACS will learn a smoothed depth-dependent average sed rate, 'constant' will cause the sedimentation rate to be held constant throughout the
- Lower_bound (number, kyr): The lower limit of age samples for this particular core. Default is NaN, implying the lower limit is equal to the first age of the target stack.
- Upper_boud (number, kyr): The upper limit of age samples for this particular core. Default is NaN, implying the upper limit is equal to the last age of the target stack.
- Min_resolution_mode ('off','absolute', 'relative'): The method by which BIGMACS defines the maximum length to a given sedimentation rate. Default is 'absolute', and thus the maximum length of a sedimentation rate is defined in 'min_resolution' and has units of meters. If set to 'relative', 'min_resolution' should be given as a percentage of the core length. The final option 'off' indicates there is no maximum length of a given sedimentation rate.
- Min_resolution (number): Sets the maximum length between depths for which a sedimentation rate is calculated. This setting creates additional depths in the core's age-depth model if the d18O and/or radiocarbon data have gaps larger than min_resolution. The unit of this number depends on the setting prescribed for 'min_resolution_mode'. The default is 0.145679 m, which was derived from the radiocarbon compilation from Lin et al., (2014).
- is_top_14C_inlier ('no' or 'yes'): this option specifies whether the shallowest ^{14}C information follows the Gaussian distribution ('yes') or Student's t-distribution ('no'). The default is 'yes'.
- lower_sed_rate (number) : The lowest allowable normalized sedimentation rate. Default is 1/8.
- Upper_sed_rate (number): The highest allowable normalized sedimentation rate. Default is 8.

2. hyperparameters.txt

We discourage users from changing these settings unless they have good reason to do so. These parameters involve setting the transition and emission models and training the hyperparameters.

- q: is the prior probability for being a δ^{18}O outlier of the benthic δ^{18}O record. The default value is 0.05, or 5%
- d: Number of standard deviations that δ^{18}O outliers are deviated from the alignment target. The default value is 3.
- nParticles: the number of particles in the initialization step of particle smoothing. Increasing this number might be useful for alignments or stacks longer than ~500 kyr but will increase runtime. Default is 500.
- v: The efficiency of the sampling in particle smoothing. Default value is …
- max_iters: the number of iterations in the Metropolis Hastings algorithm for sampling ages.
- h: the smoothness of the regression of the core's average sedimentation rate. The default value is 20 kyr. Lower values will cause less smoothing of the core's average sedimentation rate (i.e., allowing more absolute variability in sedimentation rate).

- a_d18O: Parameter that defines the generalized student’s t-distribution for the d18O emission model. The default value is 3.
- b_d18O: Parameter that defines the generalized student’s t-distribution for the d18O emission model. The default value is 4.
- a_C14: Parameter that defines the generalized student’s t-distribution for the C14 emission model. The default value is 10.
- b_C14: Parameter that defines the generalized student’s t-distribution for the C14 emission model. The default value is 11.

6. Defaults/transition_parameter.txt

- The default state change probabilities during age model construction. Rows and columns correspond to {‘Initial’,’Contraction’,’Steady’,’Expansion’} and {’Contraction’,’Steady’,’Expansion’}, respectively.