Supplementary Material

Grace, M.R., Giling, D.P., Hladyz, S., Caron, V., Thompson, R.M. and Mac Nally, R. (2015). Fast processing of diel oxygen curves: estimating stream metabolism with BASE (BAyesian Single-station Estimation). Limnology & Oceanography: Methods, 13, 103–114.

USER GUIDE FOR RUNNING BASE

User guide correspondence: Darren Giling, giling@igb-berlin.de

Overview

1. *Introduction*

This is a user guide to implementing the BASE program described by Grace *et al.* (2014) to estimate single-station whole-stream metabolic rates from diel dissolved oxygen (DO) curves. For details, please read and cite Grace *et al.* (2014).

The manuscript provides a brief overview of whole-stream metabolism methods and a description of the Bayesian estimation model. Here, we describe how to structure the input data and to run the model to calculate metabolic rates from diel DO curves; processing can be done in batch mode. Running the model is straightforward, but requires familiarity with R (R Development Core Team 2011). The fitting is performed with the OpenBUGS software (Lunn et al. 2009). The model does not require experience with OpenBUGS, but to gain a better understanding of the methods and outputs (e.g. for checking model convergence), we recommend consulting the introductory texts *Bayesian Methods for Ecology* (McCarthy 2007) and *Introduction to WinBUGS for Ecologists* (Kéry 2010).

2. *Required software and model files*

Download and install R (http://www.r-project.org/) and OpenBUGS (http://www.openbugs.net/w/FrontPage). Ensure you have at least R version 2.15 and OpenBUGS version 3.2 or newer because the model is NOT compatible with older versions of R and/or OpenBUGS.

Step 1

Extract the entire zipped 'BASE' folder to a location on your hard drive. This folder contains this user guide, the R and BUGS scripts, and the subfolders 'input' (data) and 'output' (results).

Do not alter the folder structure inside the 'BASE' folder.

3. *Code description*

There are two files (located in the 'BASE' folder) required for the model:

Script 1: *Script-1_run-OpenBUGS.R*

Script 2: *Script-2_metab-model.txt*

Script 1 is an R script used to define the diel data vectors and OpenBUGS model parameters within the R environment. The Bayesian model is called to run in OpenBUGS, looping iteratively through each data file (diel time-series) defined in the 'input' folder (see Section 4). Results are written to file after each data-file iteration in the 'output' folder (see Section 6).

Script 2 (the model file) is the OpenBUGS model description that takes the data and parameters packaged by Script 1 to run the Bayesian model. Temperature and salinity corrections are made, and then the daytime regression model is fitted to measured DO data; key outputs are estimates of A (constant used to calculate GPP), R (instantaneous respiration rates) and K (the reaeration coefficient). See Grace *et al.* (2014) for a full description of the daytime regression model.

4. *Input file location and format*

An example data set of diel curves is provided with the download in the 'input' folder. We recommend familiarizing yourself with the model by first using this data series, and subsequently using one of the comma-separated values (.csv) files as a template to input your own data. This will ensure correct header formatting.

Rates can be estimated for multiple diel time-series in one model run. Each DO diel time-series is provided in a separate csv file within the 'input' folder. Each data file must be a 24-hour time series of DO measurements (5 or 10 minute data intervals are commonly used). *Dissolved oxygen concentration should be corrected for probe drift prior to running the metabolic model (Grace and Imberger 2006).* The first data point should be midnight (24h00) of the day of interest and the last data point should be midnight of the following day. See Section 7 for description of code to split long time series into separate files.

Example input:

Date	Time	I	tempC	DO.meas	atmo.pressure	salinity
2012-04-05	0:00:00	0	16.32	8.11	0.976886	0.071737
2012-04-05	0:05:00	0	16.29	8.112	0.976886	0.071737
2012-04-05	0:10:00	0	16.27	8.107	0.976886	0.071737
...	...	...	...	...	...	...
2012-04-05	23:55:00	0	16.68	8.077	0.977577	0.071737
2012-04-06	0:00:00	0	16.67	8.064	0.977281	0.071737

Where:

I Photosynthetic active radiation (PAR; in μ mol m^{-2} s^{-1}).

tempC Stream water temperature (in degrees Celsius).

DO.meas Measured dissolved oxygen concentration (in mg L^{-1}).

atmo.pressure Measured atmospheric pressure in atmospheres. Can be constant (i.e. fill every time interval with same value) and inferred from stream altitude if barometric data is unavailable. A default of 1 can be used if pressure and altitude are unknown.

salinity Water salinity (in ppt). Can be constant (i.e. fill every time interval with same value) or a time-series. A default of 0 can be used in salinity is low and unknown.

IMPORTANT:

- Correct column headings are vital (they are case-sensitive).
- All columns must contain data for each time interval.

  Ensure correct date and time formatting. Date must be separated by dashes (-), and not by slashes.
- Time must be formatted to hh:mm:ss within input csv files.

5. *Running the model*

5.1 Default model

Step 2

Open R

Open the R script *Script-1_run-OpenBUGS.R* by selecting 'File', then 'Open script...'

The first time you run the model you will need to install several R packages – these are shown at the beginning of Script 1. Select 'Packages' menu, then 'Install package(s)'. Chose a local mirror site for the download and select the 'coda' and 'R2OpenBUGS' packages.

Sections of the code in one or both scripts must be updated to adjust the model for your system and data. This section (5.1) describes the minimum lines of code that must be specified to run the default (3-parameter) model (these lines are all within Script 1). The model can be customized by including prior information for K or changing how some model parameters are estimated, if desired (described in Section 5.2).

The capital letters below (A, B and C) are paired with corresponding letters in the code of Script 1 where the code is to be updated.

Step 3: Amend code at lines A, B and C in Script 1

(A) Define the location of 'BASE' folder

Tell R where to find the unzipped folder here. Replace "[your directory]" in the code at (A) with the location of the 'BASE' folder on your disc. Use forward slashes to indicate folder levels. Windows explorer uses back slashes, so you will have to change these. For example:

```
folder.location <- "C:/Desktop/Analysis"
```

(B) Define the measurement interval

Define the measurement interval of your DO time-series (in seconds).

(C) Define the number of model iterations

Define the total number of Bayesian model iterations and number of burn-in ('settling') iterations. By default this is set to 2000 iterations with 1000 burn-in, which should be sufficient in most cases. This can be reduced (e.g. to 200/100) to quickly test if the model is functioning properly before proceeding with the full analysis. The number of required iterations can be assessed by inspecting the convergence statistics in OpenBUGS (see section 6.2).

Step 4: Run the model

Once the code at A, B and C in Script 1 has been amended to your requirements, start the model by running the entire Script 1 within R. Select the 'Edit' menu, then 'Run all'.

This will call the OpenBUGS program, which will operate in the background, and will take at least several minutes for each diel cycle (i.e. file in the input folder). The script will loop through each file in the input folder without further user input. You do not need to run Script 2 manually. Pressing ESC in R will stop the model from looping to the next file. Results are saved after each file is completed. The model outputs are described in Section 6. The user can check the progress of the model by seeing how many fitting plots have been written (see description Section 6). Do not open the results csv file while the model is running.

5.2 Optional model customization

By default, BASE is set to estimate GPP, ER and K simultaneously (i.e. a 3-parameter model) and the parameters theta and p have fixed values.

There is an option to change these defaults:

- Priors for K: estimated or measured K

  K can be estimated from the model and data with uninformative priors ($K \geq 0$), or you can inform the priors with mean and uncertainty of a measured K (e.g. if measured using SF_6 injections).

- Fixed or estimated theta or p

  The constants for temperature dependence and light saturation (theta and p) can be estimated from the model and data (within narrow, realistic bounds), which may enhance model fit. Theta or p are estimated along with GPP, ER and K, making a 4-parameter (theta or p is estimated) model or 5-parameter (theta and p are estimated) model. Selecting the most appropriate model is described in section 6.3 – Model selection.

To alter the defaults, adjust the lines of code (described below) in Script 2 (Y and Z). This is performed by commenting (i.e. adding "#") or un-commenting (removing "#") the appropriate lines of code for the model you are running. Script 2 should be opened in Notepad (or similar), ensuring that you save any changes before running the model. Note that the alternative lines must have one commented and one uncommented line; the code will fail if both are commented or both are uncommented. Start the model by running all of Script 1 in R (Script 2 is not run manually).

(Y) Priors for K

Inform OpenBUGS if you are using informative or uninformative priors here.

(Y1) Uninformative priors

(Y2) Informative priors

(Z) Treatment of theta and p

Inform OpenBUGS if theta and p should be treated as fixed or estimated.

(Z1) theta and p fixed

(Z2) theta and p estimated

NOTE: You can choose to treat either p or theta as estimated and the other as fixed by selecting the appropriate combination of code from lines Z1 and Z2.

6. *Model outputs*

6.1 Results table (BASE_results.csv)

The results table ('BASE_results.csv', located in the 'output' folder) provides the means and standard deviations for the metabolic rates and other parameters estimated by the model. Each row of the csv file is the result for one input file. Rates of 'GPP' (daily gross primary production) and 'ER' (daily ecosystem respiration; calculated from the instantaneous rates R) are expressed in mg O_2 L^{-1} day^{-1}). The reaeration coefficient ('K') is expressed in day^{-1}.

IMPORTANT: Rename or move the results file and plots if you wish to keep the results because they will be overwritten the next time you run the model.

6.2 Assessing model convergence and fit (quick guide pg 10)

Assessing model convergence

It is vital to ensure that the MCMC chains of model parameters have adequately converged to a stationary distribution or the model may be inappropriate. The 'R-hat' statistic gives an indication of convergence. Values close to 1 indicate good convergence, while values >1.1 indicate poor mixing. R-hat for model parameters (A, R, K, theta, p and GPP) are included in the results table. Poor mixing can (but not always) be improved by increasing the number of iterations. The output

csv contains a column titled ‘convergence.check’, which tests whether all the R-hat values are < 1.1 and can be used to quickly assess convergence (returns ‘fine’ when all R-hats < 1.1).

The results table returns the effective number of parameters (‘pD’). This value normally should be positive. Negative pD may indicate the posterior mean is not a good measure of the posterior distribution, and there is likely an issue with the model.

Convergence also can be visually assessed by examining that the distributions are stationary and chains are well mixed on the MCMC trace plots in OpenBUGS (see McCarthy 2007 or Kéry 2010 for examples). To inspect the MCMC chains you must inform the model to not close OpenBUGS after finishing the iterations on each file. To do this, locate the section of code in the lower half of Script 1with the comment:

```
# Set debug = T below
```

and replace the debug argument F (for false) with T (for true) in the brackets below.
Note that to continue to the next csv file it is necessary to manually close OpenBUGS after inspecting plots, so that this option is best used to double-check convergence on a subset of files before running all files with `debug = F`.

Assessing model fit

There are three measures of model fit included in the results table: (1) the posterior predictive p-value (PPP), the (2) R^2 value, and (3) the residual mean square error (rmse). The PPP compares lack of fit of the model to the actual data against lack to fit to a distribution of possible model discrepancies by using data simulated from the parameterized model (Gelman et al. 1996). A PPP value (the PPfit.mean column in the results table) of close to 0.5 indicates a very plausible model, while values <0.1 or >0.9 indicate that the parameterized model is not a plausible explanation of the observed data. The correlation (R^2) between the observed and modelled DO data is reported in the results table. Due to the temporal and correlated nature of the DO time-series, the R^2 may be high when model estimates are consistently above or below measured values. In these cases, the

poor fit may be indicated by the residual mean square error (rmse) and maximum run length fraction (mrl.fraction). The rmse is specific to the magnitude of the dataset and should be assessed against models from days at the same site. The rmse is expressed relative to the point-to-point variation in the dataset by the output column 'rmse.relative'. The maximum run length fraction (mrl.fraction) is the proportion of time occupied by the longest run of values for which the estimated DO is below or above the measured DO. A high maximum run length proportion may indicate consistent over- or under-estimation of DO and plots should be inspected.

The model fit also can be confirmed visually using the fitting plots. The model prints a jpg plot of the measured (empty circles) and predicted (black line) DO curve for each diel period and saves them in the 'fitting plots' folder of the 'output' folder. These plots can be used to visually confirm curve fits and quickly identify any discrepancies in the data or model. The file name of each plot is the same as for the input file for those data.

Quick guide for assessing model convergence and fit

To check the model is reliable:

- Ensure all parameters have converged (R-hats < 1.1; convergence.check = 'fine')
- Check PPfit is between 0.1 and 0.9 (closer to 0.5 is best)
- Check pD is positive
- Check maximum run length is not large and visually confirm the model is appropriate

6.3 Model selection

The results table returns the Deviance Information Criterion (DIC), an assessment of how well the model will predict a replicate dataset. DIC takes into account the complexity of

the model and can be used for model selection (for example from the 3- or 5-parameter model). DIC may be negative, and lower DIC is desirable.

There is no hard rule, but we recommend a difference in DIC of ≥ 5 between models is evidence that the model with lower DIC best predicts the data because values exceeding 5 correspond to 10-fold or greater support for the model with the lower DIC.

For example:

3-parameter (default) model: DIC = -1797 *alternative, simpler model*

5-parameter (customized) model: DIC -1807 *'best', more complex model*

In this case, the difference in DIC between the alternative and 'best' model is 10 (-1797 − -1807). We would rule the 5-parameter model (with lower DIC) provides a substantially better prediction of the data than the 3-parameter model, despite its additional complexity (two additional parameters estimated).

More information of DIC and pD in BUGS can be found here:

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml

7. *Additional code*

7.1 'Separate files to days' code

The code and folders inside the 'Separate files to days' folder can be used to quickly convert a single csv file with long-term sonde data (i.e. many days or months) into separate csv files for each date, which are required by the metabolism model. The input structure for the 'Separate files to days' code is the same as described in Section 4 and an example file is included. The code will split the data according to the 'Date' column and include one extra row from the next date, with the assumption this is the midnight data point. The user needs only to open the script in R, update the input and output directory lines of code, and then run the entire script.

7.2 'Plots to troubleshoot poor fits' code

This short code will plot measured DO (mg L^{-1} and % saturated), modelled DO (mg L^{-1}, with uncertainties), temperature (°C) and PAR (I; μ mol m^{-2} s^{-1}) together in one multi-panel figure. This code can be run after completing a model fit (it uses the 'metab' results object) to view the results from the last file run by the model. This can be a useful troubleshooting exercise if the results indicate a poor model fit. The user can easily assess if there are any inconsistencies in the data that may indicate a violation in the assumption that reaeration, GPP and ER are the only processes contributing to change in DO. For example, a sharp increase or decrease in temperature or DO may indicate another source of water started to enter the system, or a lack of increase in O_2 percent saturation during daylight hours may indicate low biological activity compared to reaeration.

References

Gelman, A., X.-L. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica **6:** 733-807.

Grace, M. R., D. P. Giling, S. Hladyz, V. Caron, R. M. Thompson, and R. Mac Nally. 2014. Fast processing of diel oxygen curves: estimating stream metabolism with BASE (BAyesian Single-station Estimation). Limnology & Oceanography: Methods.

Grace, M. R., and S. J. Imberger. 2006. Stream Metabolism: Performing & Interpreting Measurements, p. 204. Water Studies Centre Monash University, Murray Darling Basin Commission and New South Wales Department of Environment and Climate Change. Accessed at http://www.sci.monash.edu.au/wsc/docs/tech-manual-v3.pdf.

Kéry, M. 2010. Introduction to WinBUGS for Ecologists: A Bayesian Approach to Regression, ANOVA and Related Analyses. Academic Press.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best. 2009. The BUGS project: Evolution, critique, and future directions. Stat. Med. **28:** 3049-3067.

McCarthy, M. A. 2007. Bayesian Methods for Ecology. Cambridge University Press.

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.