USER GUIDE FOR RUNNING BASE v2.3

User guide correspondence or for troubleshooting, contact Darren Giling (darren.giling@idiv.de)

Supplementary Material for:

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.

http://onlinelibrary.wiley.com/doi/10.1002/lom3.10011/abstract

1. Introduction

This is a guide to estimate single-station whole-stream metabolic rates from diel dissolved oxygen (DO) curves using the BASE program described by Grace *et al.* (2015). Please refer to and cite Grace *et al.* (2015) for the code, and note the updates detailed in Section 2.

   Grace *et al.* (2015) 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 parameter estimation is performed with the JAGS software (Plummer 2003). The model does not require experience with JAGS, but to gain a better understanding of the methods and outputs (e.g. for checking model convergence), we recommend consulting introductory texts on Bayesian methods (e.g. McCarthy 2007; Kéry 2010).

2. Update history

- *BASE v2.3 (January 2018)*: Multiple changes (see full instructions for details):
  - *This will be the last update for this repo – we are shifting to a package that will be available from github via dev.tools() – stay tuned.*

- o Simplified file handling. Only one input csv file required; users are no longer required to separate into files for individual days. However, separated files will still work (but are not recommended if using new smoothing options. The previous "separate files to days" code is now obsolete.

- o Added an option to smooth dissolved oxygen time series with a fast Fourier transform, and/or PAR time series with moving average.

- o Placed a restriction on the prior distribution of K. It is now truncated at a maximum of 40 day^{-1} to avoid chains getting caught in bad parameter spaces.

- o Placed a reasonable initial value for K (2 day^{-1}) to aid convergence.

- o Implemented a procedure to automatically update unconverged chains.

- *BASE v2.2 (March 2017):*

  - o Fix to the prior distribution for tau, which is now: `tau ~ dgamma(1,0.001)`. It was found that the tau prior must be more uninformative under some circumstances. The previous, more constrained prior resulted in low variation in the sampled values that are used to calculate the fit metric PPP in some datasets. Consequently, PPP was overly sensitive and indicated that well-fitting models should be rejected. In some datasets, this fix may reduce the precision of parameter estimates and will affect PPP values.

- *BASE v2.1 (October 2016)*

  - o Added an output of instantaneous rates

- *BASE v2.0 (July 2016)*: Major update including:

  - o Changes to the diel DO model structure following Song *et al.* (2016). Results published by Song *et al.* (2016) showed that BASE v1 underestimated metabolic rates in some cases due to two differences in the model formulation compared to other aquatic metabolic models (e.g. Hall and Tank 2005; Van de Bogert et al. 2007; Hanson et al. 2008; Holtgrieve et al. 2010):

1. BASE v1 used a 'stepwise' approach to model changes in DO concentration (Δ[DO]) between successive measurements rather than DO concentration ([DO]) directly.

2. BASE v1 used the measured DO concentration ([DO]$_{measured}$) to estimate oxygen deficiency for reaeration rates instead of the modelled DO concentration ([DO]$_{measured}$).

Given the findings of Song *et al.* (2016), these inconsistencies in the BASE model were rectified:

BASE v1:

$$\Delta[DO]_t/\Delta t = AI_t^p - R\left(\theta^{(T_t-\bar{T})}\right) + K_{DO}.\left(1.0241^{(T_t-\bar{T})}\right).([DO]_{sat,t} - [DO]_{meas,t})$$

BASE v2:

$$[DO]_{t+1} = [DO]_t + AI_t^p - R\left(\theta^{(T_t-\bar{T})}\right) + K_{DO}.\left(1.0241^{(T_t-\bar{T})}\right).([DO]_{sat,t} - [DO]_{modelled,t})$$

Here, *t* indicates the timestep, *A* is a constant, *p* is an exponent describing incident light use, Θ describes temperature dependence of respiration, *T* is water temperature and *sat*, *meas* and *mod* indicate [DO] at saturation, observed concentration and modelled concentration. Refer to Song *et al.* (2016) for a description of the assumptions underlying the alternative models. The updated model of BASE v2 improved fit (visual and R^2) of modelled DO to observed values and reduced uncertainty in parameter estimates of the validation dataset of Grace *et al.* (2015). There was greater agreement between the estimates from BASE v2 and BaMM (Holtgrieve et al. 2010), the 'accurate' method of Song *et al.* (2016), for a wide range of stream characteristics and range of metabolic rates (Figs. 1 & 2). Some remaining differences may be due to differences in the structure of the photosynthesis-irradiance (PI) curve used in each model.

o Change in software for MCMC algorithm from OpenBUGS (Lunn et al. 2000) to JAGS (Plummer 2003) for platform-independence and to allow for parallel core processing of the three chains, greatly increasing computational speed. Fastest results will be achieved on a processor with at least 3 cores.
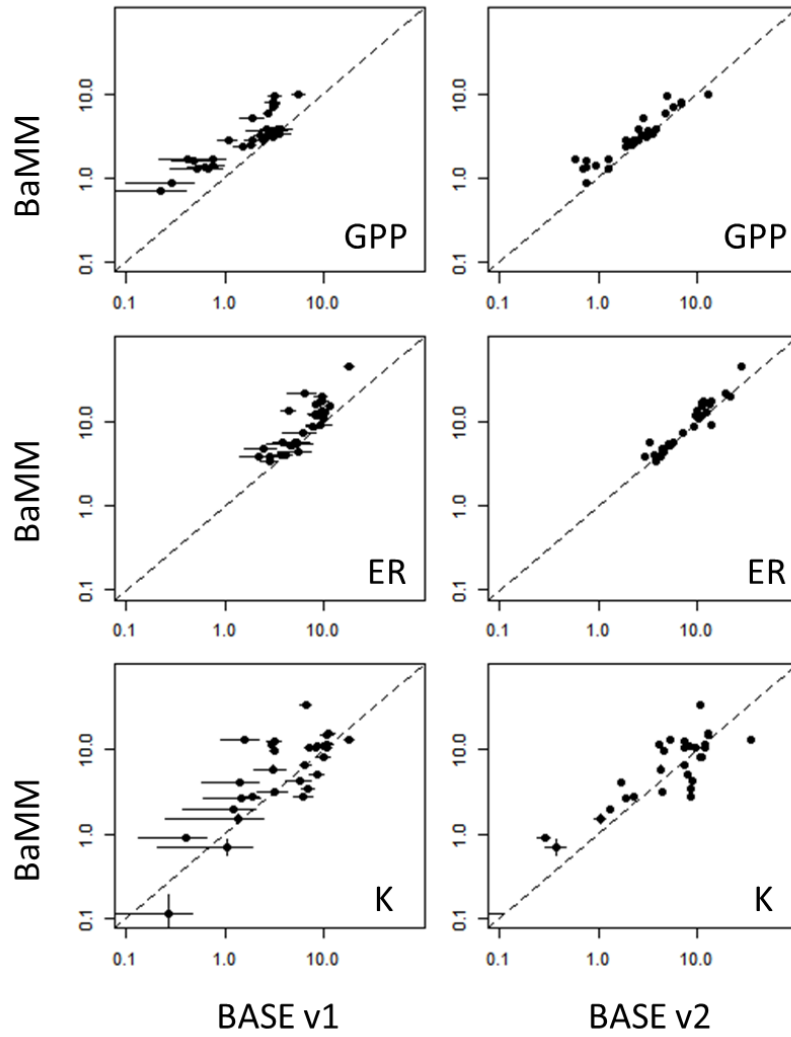
Figure 1. Comparisons between BASE (3-parameter model) and BaMM (Holtgrieve et al. 2010) for converged estimates of log-transformed GPP (mg O_2 L^{-1} day^{-1}), ER (mg O_2 L^{-1} day^{-1}) and K (day^{-1}). Dashed line indicates the 1:1 relationship. Data are those validation streams from Grace *et al.* (2016) with a measurement interval of 5 or 10 minutes. Note reduced uncertainty in BASE v2 estimates compared to BASE v1.

Figure 2. Comparisons between BASE (5-parameter model) and BaMM (Holtgrieve et al. 2010) for converged estimates of log-transformed GPP (mg O_2 L^{-1} day^{-1}), ER (mg O_2 L^{-1} day^{-1}) and K (day^{-1}). Dashed line indicates the 1:1 relationship. Data are those validation streams from Grace *et al.* (2016) with a measurement interval of 5 or 10 minutes. Note reduced uncertainty in BASE v2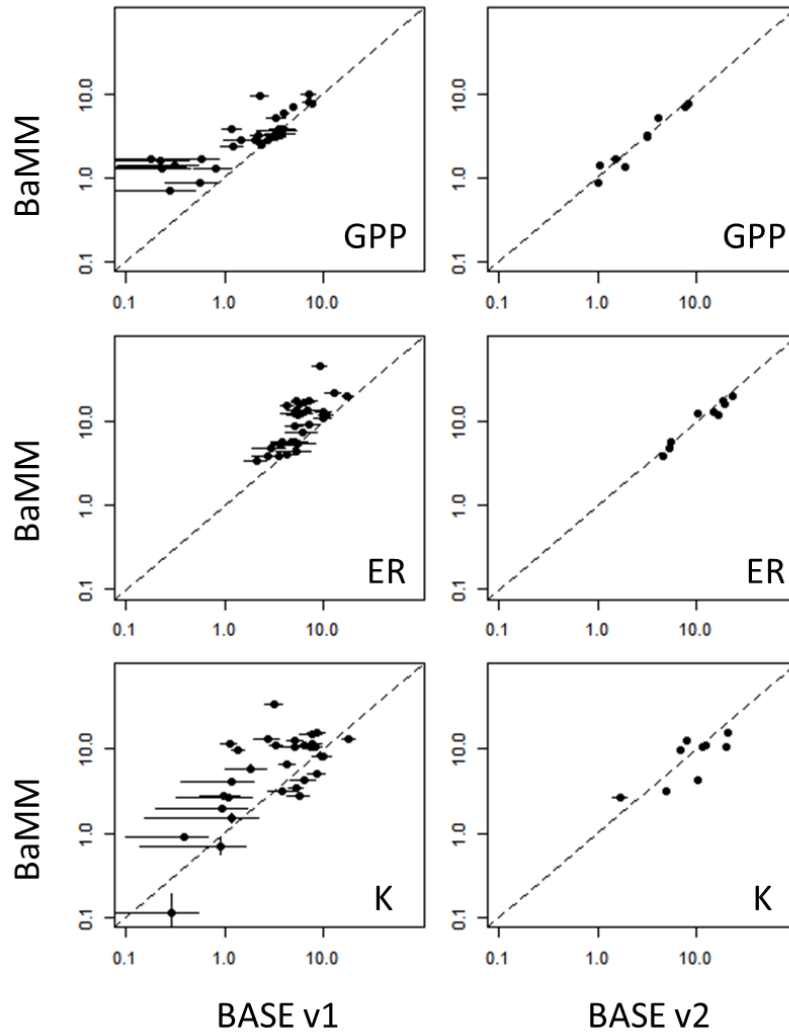 estimates compared to BASE v1. Models converged less frequently in the 5-parameter BASE v2 model than in the BASE v1 model.

- *BASE v1.0 (June 2014):* Code uploaded to github with publication of Grace *et al.* (2015)

3. Required software

Download and install R (http://www.r-project.org/) and JAGS (http://mcmc-jags.sourceforge.net/).

Update software if necessary.

Step 1

Extract the zipped 'BASE_v2.3' folder to a location on your hard drive. This folder contains the R

and JAGS code, and the subfolders 'input' (data) and 'output' (results). You may create new

folders, but do not alter the folder structure inside the 'BASE' folder.

---

4. Code description

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

      <u>Script 1:</u> *Call_BASE_v2.3.R*

      <u>Script 2:</u> *BASE_metab_model_v2.3.txt*

<u>Script 1</u> is an R script used to define the diel data vectors and Bayesian model parameters within the R

environment. The Bayesian model is then called from R to run in JAGS (occurs in the background),

looping iteratively through each data file (e.g. different sites) in the 'input' folder (see Section 5).

Results are written to file after each model fit in the 'output' folder (see Section 7).

<u>Script 2</u> is the JAGS code specifying the model. The code that takes the data and parameters packaged

by <u>Script 1</u> to run the Bayesian model. Temperature and salinity corrections are made, and 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. (2015) for a full description of the daytime regression model.

5. *Input file location and format*

An example set of diel curves (six days) 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 the comma-separated values (.csv) files as a template to input your own data. This will ensure correct formatting.

Rates can be estimated for multiple diel time-series (i.e.) in one model run. Each csv file in the 'input' folder may contain one or more full 24-hour time series of DO measurements (5 or 10 minute data intervals are commonly used). By default, incomplete days will be excluded from analyses. For example, each day must contain 144 data rows if the measurement interval is 600 seconds. The time interval of measurements must be identical across all days and input files run in any single analysis. Note that the code is currently set up to estimate daily rates (but could be modified to model multiple of days of data at a time). There is no necessity to estimate rates over a defined 24-hour window, but the example data uses midnight to midnight by convention. In this case, the first data point should be midnight (00:00) of the day of interest.

Example input (6 full days of data at 10-minute intervals):

Date	Time	I	tempC	DO.meas	atmo.pressure	salinity
2011-12-01	0:00:00	0	21.91	7.034	0.985816	0.2
2011-12-01	0:10:00	0	21.85	6.989	0.985816	0.2
2011-12-01	0:20:00	0	21.83	6.998	0.985816	0.2
...	...	...	...	...	...	...
2011-12-06	23:50:00	0	22.07	7.169	0.985816	0.2

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. Salinity does not play a large role in determining DO saturation in freshwaters; a default of 0 can be used in salinity is low and unknown.

IMPORTANT:

- Check column names match (they are case-sensitive).

- All columns must contain data for each time interval, i.e. *no missing data*. It may be possible to interpolate short periods of missing data, such as that resulting from logger maintenance (Obrador et al. 2014; Giling et al. 2017).

- Ensure dates are not separated with slashes ("/"). A good convention is yyyy-mm-dd. Time should be formatted to hh:mm:ss.

- *Dissolved oxygen concentration should be corrected for probe drift prior to running the metabolic model (Grace and Imberger 2006).*

6. *Running the model*

6.1 *Default model*

Step 2

Open the R script *Call_BASE_v2.R* by selecting 'File', then 'Open script...'

The first time you run the model you will need to install several R packages. They can be installed using the following code:

```
install.packages("coda")

install.packages("R2jags")

install.packages("zoo")
```

Sections of the code in one or both scripts must be updated to adjust the model for your system and data. This section 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 6.2).

The capital letters listed below (A-D) 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. Replace "[your directory]" in the code at (A) with the location of the 'BASE_v2' 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 20000 iterations with 10000 burn-in, which should be sufficient in most cases for the 3-parameter model. This can be reduced (e.g. to 200) to quickly test if the model is functioning properly before proceeding with the full analysis. The number of required iterations can be assessed visually and by inspecting the convergence statistics (see section 7.2).

(D) Optional - set the smoothing behaviour

Dissolved oxygen and PAR data are often noisy due to physical factors such as light intensity being patchy in space and time due to clouds and canopy cover. A fast Fourier transform may be used to filter out high-frequency fluctuations in DO (Gallegos et al. 1977), potentially enhancing the diel signal of interest. The value `hf.cutoff` is used to set the proportion of low pass filtering (0-1), which is 0 by default (no smoothing). A value of 0.88 was used by Oliver and Merrick (2006). Additionally, smoothing of PAR with a moving average across five time periods can be added by changing the logical object smooth.PAR to TRUE. Plots showing an example of the smoothing will be saved in the 'validation plots' folder (e.g. Fig. 3). *Smoothing should not be used if the dates within each csv file are not contiguous, as smoothing will then be applied across jumps in time.*
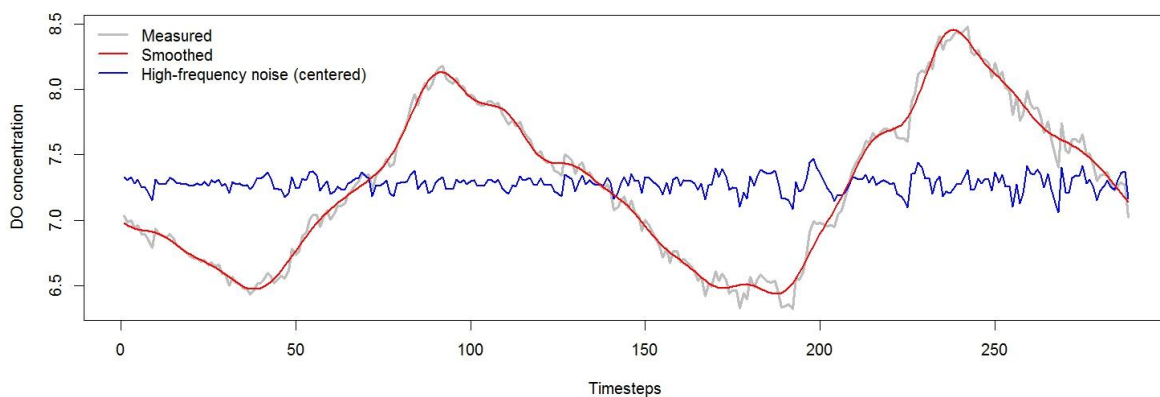

Figure 3. Example of smoothing over 48 hours of dissolved oxygen data with a low pass frequency of 88% (`hf.cutoff <- 0.88`), following Oliver and Merrick (2006).

Step 4: Run the model

Once the code at A, B, C (and optionally D) in Script 1 has been amended to your requirements, start the model by running the entire Script 1 within R. This will call the JAGS program, which will operate in the background, and may take up to several minutes for each diel cycle. 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 fit. Results are saved after

each file is completed. The model outputs are described in Section 7. The user can check the progress of the model by seeing how many fitting plots have been written (see description Section 7). Do not open the results csv file while the model is running because further results cannot then be written.

6.2 Optional model customization

By default, BASE is set to estimate parameters *A*, *R* and *K* simultaneously (i.e. a 3-parameter model) and the parameters Ѳ (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. based on stream morphology or if measured using SF_6 injections).

- Fixed or estimated theta or p

  The constants for temperature dependence and light saturation (*Ѳ* and *p*) can be estimated from the model and data (within narrow, physically 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 7.3 – 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 model will produce an error 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 JAGS if you are using informative or uninformative priors here.

    (Y1) Uninformative priors

    (Y2) Informative priors

(Z) Treatment of theta and p

    Inform JAGS 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.

7. Model outputs

7.1 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 within a file is the result for one site and diel period. 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 in day^{-1}. The other results are described in section 8 below.

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.

7.2 Table of instantaneous rates

In a separate folder ('instantaneous rates'), another csv is written containing the rates of K, ER and GPP for each timestep on each day. This may be useful for users interested in metabolic variation on sub-daily timescales. The ER and K are temperature corrected based on the daily mean estimate of

theta, and GPP corrected according to instantaneous PAR and the daily estimate of A and p. The units are per timestep, e.g. mg O_2/L/600seconds for the example dataset.

7.3 Validation plots

Plots to assess model convergence and fit are printed in the 'validation plots' folder. See section 8 below.

8. Model validation and selection

8.1 Assessing convergence and fit

It is vital to ensure that the MCMC chains of model parameters have adequately converged to a stationary distribution; otherwise parameter estimates cannot be regarded as accurate. 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. 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). If chains for any one day do not converge, the program will update the chains once more by running another set of iterations. This additional set of iterations is not run in parallel like the first set, so will slow down the runtime of the analysis. The progress of these additional samples will be indicated with progress bars in the R console. *Poor mixing can (but not always) be improved by increasing the number of iterations. Additionally, if you are familiar with JAGS and R2jags, modifying prior distributions and/or initial values for parameters to suit your system may aid convergence.*

Convergence also can be visually assessed by examining that the distributions are stationary and chains are well mixed on the MCMC trace plots. The model prints a jpg multi-panel plot for each diel model in the 'validation plots' folder of the 'output' folder, including MCMC trace plots for *A, p, R, K* and *Θ*. The three chains should be converged (overlapping) and stationary (centred) (Fig. 4). When set as fixed values the trace plots for *p* and *Θ* show a horizontal lines with no variation.
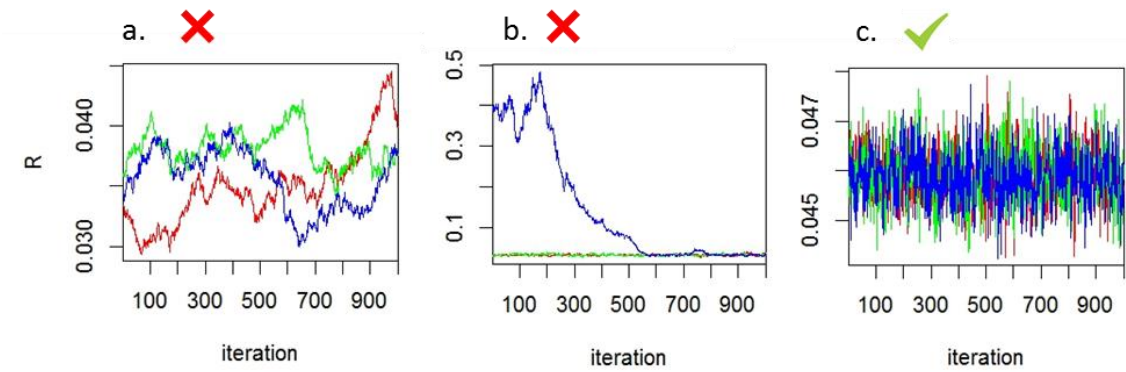
Figure 4. Example trace plots showing (a) three unconverged chains with poor mixing, (b) a longer burn-in period is required, and (c) converged and stationary chains.

The model fit also can be confirmed visually using the validation plots, which show the measured (empty circles) and predicted (black line) DO curve for each diel period and the measured temperature and PAR data (Fig. 5). These plots can be used to visually confirm curve fits and quickly identify any discrepancies in the data or model. Inconsistencies in the data may indicate a violation in the assumption of the free-water DO method 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 DO percent saturation during daylight hours may indicate low biological activity compared to reaeration.
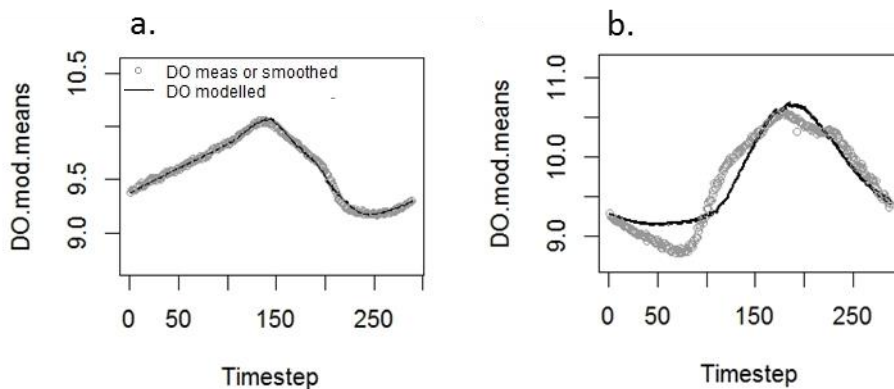

Figure 5. Examples of models with (a) good fit and (b) poor fit. *Note that even if model fit is visually adequate the parameter estimates will be unreliable if chains have not converged.*

The results table also 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.

Assessing model fit quantitatively

There are three quantitative measures of model fit included in the results table: (1) the posterior predictive p-value (PPP), (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.

Quick guide for model validation

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

- Visually confirm the model is appropriate

8.2 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). Lower DIC is desirable, and DIC may be negative.

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).

References

Gallegos, C. L., G. M. Ilornberger, and M. G. Kelly. 1977. A model of river benthic algal
   photosynthesis in response to rapid changes in light1. Limnol Oceanogr **22:** 226-233.

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

Giling, D. P. and others 2017. Delving deeper: Metabolic processes in the metalimnion of stratified
   lakes. Limnology & Oceanography **62:** 1288-1306.

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

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.

Hall, R. O., and J. L. Tank. 2005. Correcting whole-stream estimates of metabolism for groundwater
   input. Limnology and Oceanography-Methods **3:** 222-229.

Hanson, P. C., S. R. Carpenter, N. Kimura, C. Wu, S. P. Cornelius, and T. K. Kratz. 2008. Evaluation
   of metabolism models for free-water dissolved oxygen methods in lakes. Limnol. Oceanogr.
   Methods **6:** 454–465.

Holtgrieve, G. W., D. E. Schindler, T. A. Branch, and Z. T. A'mar. 2010. Simultaneous quantification
   of aquatic ecosystem metabolism and reaeration using a Bayesian statistical model of oxygen
   dynamics. Limnol Oceanogr **55:** 1047-1063.

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

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing **10:** 325-337.

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

Obrador, B., P. A. Staehr, and J. P. C. Christensen. 2014. Vertical patterns of metabolism in three contrasting stratified lakes. Limnol Oceanogr **59:** 1228-1240.

Oliver, R. L., and C. J. Merrick. 2006. Partitioning of river metabolism identifies phytoplankton as a major contributor in the regulated Murray River (Australia). Freshwater Biol **51:** 1131-1148.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003).

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.

Song, C., W. K. Dodds, M. T. Trentman, J. Rüegg, and F. Ballantyne. 2016. Methods of approximation influence aquatic ecosystem metabolism estimates. Limnol. Oceanogr. Methods **14:** 557–569.

Van De Bogert, M. C., S. R. Carpenter, J. J. Cole, and M. L. Pace. 2007. Assessing pelagic and benthic metabolism using free water measurements. Limnol. Oceanogr. Methods **5:** 145-155.