User guide for package ‘BASEmetab’

Darren Giling, Ralph Mac Nally, Nick Bond & Michael Grace

2018-10-30

Maintainer: Darren Giling darren.giling@idiv.de

Version: 3.0

Description: Estimate Single station Whole-stream Metabolic Rates from Diel Dissolved Oxygen (DO) Curves. See Grace et al. (2015) for more details. NOTE: JAGS must be installed prior to use for the package to work.

Contents

BASEmetab (package)

Description

BASEmetab is a package for the batch analysis of single-station diel dissolved oxygen data to simultaneously estimate primary production, respiration and reaeration rates of streams and rivers. The analysis is run in a Bayesian framework, allowing for prior information to be considered during the estimation of model parameters. The package also contains other useful functions for preparing and processing time-series data from streams and rivers.

Details

The main function of the BASEmetab package is `bayesmetab`. This function implements the daytime regression model of Grace et al. (2015) for dissolved oxygen (DO) time series. See Grace et al. (2015) for a background to the subject and a full description of the model, but note the subsequent revisions to the model that are detailed below in 'Updates'. Grace and Imberger (2006) provide a guide to the practical aspects of measuring diel dissolved oxygen and PAR data in streams and rivers, including recommendations on probe drift that must be corrected prior to estimating metabolic rates.

The parameter estimation is performed with the JAGS software, which must be separately installed (Plummer 2003; http://mcmc-jags.sourceforge.net/). The model does not require experience with JAGS, but familiarity with metrics of MCMC chain convergence is recommended (see e.g. McCarthy 2007).

Installing BASEmetab

To install run the following code

```
# install devtools package
  install.packages("devtools")

# install BASEmetab package
  devtools::install_github("dgiling/BASEmetab")

  # Remove the package zip after installation
  unlink("BASEmetab.zip")

  #load library
  library(BASEmetab)

  #load R2jags to connect to JAGS
  library(R2jags)
```

Updates

The bayesmetab function was previously available (since 2015) as raw R and JAGS code referred to as "BASE" at https://github.com/dgiling/BASE. The version numbers of the now-obsolete raw BASE were continued for the BASEmetab package. Version 3.0 of the BASEmetab package is an implementation of the functionality in the raw code BASE v2.3.

The model has been modified since the publication of Grace et al. (2015) and the initial code upload. Importantly, the structure of the underlying dissolved oxygen model was updated following Song et al. (2016). Song et al. (2016) showed that the model implemented by Grace et al. (2015) underestimated metabolic rates in some cases. This was due to two differences in the 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). First, Grace et al. (2015) used a 'stepwise' approach to model the change in DO concentration between time t and t+1, rather than modelling concentration directly. Second, the original BASE model used the measured DO concentration ([DO]measured) to estimate oxygen deficiency for reaeration rates instead of the modelled DO concentration ([DO]measured). These inconsistencies were removed in a subsequent version of the raw code (BASE v2.0):

BASE v1:

$$\frac{\Delta[DO]}{\Delta t} = AI_t^p - R(\theta^{(T_t-\bar{T})}) + K_{DO}(1.0241^{(T_t-\bar{T})})([DO]_{sat,t} - [DO]_{meas,t}) \tag{1}$$

BASE v2:

$$[DO]_{t+1} = [DO]_t + AI_t^p - R(\theta^{(T_t-\bar{T})}) + K_{DO}(1.0241^{(T_t-\bar{T})})([DO]_{sat,t} - [DO]_{mod,t}) \tag{2}$$

Here, t indicates the timestep, A is a constant, p is an exponent describing incident light use, *theta* describes temperature dependence of respiration, T is water temperature and *sat*, *meas* and *mod* indicate [DO] at saturation, observed concentration and modelled concentration, respectively. Refer to Song et al. (2016) for a description of the assumptions underlying the alternative models.

These updates improved agreement of GPP and ER estimates (but not K) between the BASE code and BaMM (Holtgrieve et al. 2010) (Table 1), which was used as the 'accurate' method of Song et al. (2016). Some remaining differences may be due to differences in the structure of the photosynthesis-irradiance (PI) curve used in each model. Further, the updates to BASE greatly improved model fits and considerably reduced the coefficient of variation in BASE metabolic estimates (e.g. the CV of GPP estimates was 0.28 and 0.05 for BASE v1.0 and v2.0, respectively).

Table 1. Correlations between metabolic estimates made with BaMM (Holtgrieve et al. 2010) and the BASE raw code v1.0 and v2.0. Data are estimates from the validation dataset of Grace et al. 2015 with measurement intervals of 5 and 10 minutes, which cover a wide range of stream characteristics. The model is the default three-parameter estimation (p and θ are fixed).

Model parameterization	Parameter	Coefficient (mean ± SD)	R^2	n
Grace et al. 2015 (raw code v1.0)	GPP	0.76 ± 0.06	0.85	27
	ER	1.03 ± 0.09	0.84	27
	K	0.88 ± 0.10	0.77	27
Updated model (raw code v2.0, package)	GPP	0.82 ± 0.05	0.89	31
	ER	1.08 ± 0.07	0.90	31
	K	0.77 ± 0.10	0.64	31

In addition to this major update, minor changes have been made to the initial values and prior distributions of some of the parameters in the model to aid convergence in a wide range of stream and river systems. These include:

- Prior for K is truncated at 40 day^{-1}
- Prior for tau was changed to be more uninformative. Previously this prior was too constrained in some circumstances. One implication of this was that the PPP values were occasionally overly sensitive and indicated well-fitting models should be rejected.

Finally, JAGS is now used for the estimation instead of OpenBUGS (Lunn et al 2000), to improve analysis speed by computing the three MCMC chains in parallel with R2jags::jags.parallel.

References

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.

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.

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.

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).

Song, C., W. K. Dodds, M. T. Trentman, J. Ruegg, 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.

bayesmetab (function)

Description

bayesmetab is the main function of the BASEmetab package. This function calls JAGS to perform the single-station stream metabolism analysis.

Useage

```
bayesmetab(data.dir , results.dir, interval, n.iter=20000, n.burnin=n.iter*0.5,  update.chains
= TRUE, extra.iter=1, smooth.DO=0, smooth.PAR=FALSE, K.init = 2, K.est = TRUE, K.meas.mean
= 0, K.meas.sd = 4,  p.est=FALSE, theta.est=FALSE, instant=TRUE)
```

Arguments

data.dir	relative or absolute path to the folder containing csv input data files to be read.
results.dir	relative or absolute path to the output folder where results (plots and tables) will be written.
interval	Integer. The time interval in seconds (e.g. 600 seconds)
n.iter	Integer. Number of MCMC iterations (default = 20000)
n.burnin	Integer. Number of iterations of MCMC chains to delete
update.chains	Logical. Should the chains automatically update once if not converged? (default = TRUE)
extra.iter	Numeric. Number of extra iterations to run if chains are not converged, as multiple of n.iter (default = 1 times)
smooth.DO	Numeric. Proportion of high-frequency fluctuations to filter with fast Fourier transform (default = 0)
smooth.PAR	Logical. Should PAR be smoothed with a moving average? (default = FALSE)
K.init	Numeric. Initial value of chains for K (day^{-1}). Reasonable estimate aids convergence (default value = 2)
K.est	Logical. Should K be estimated with uninformative priors? (default = TRUE)
K.meas.mean	Numeric. Mean for informed normal prior distribution when `K.est` = FALSE
K.meas.sd	Numeric. Standard deviation for informed normal prior distribution when `K.est` = FALSE
p.est	Logical. Should p be estimated? (default = FALSE)
theta.est	Logical. Should theta be estimated? (default = FALSE)
instant	Logical. Should a table of instantaneous rates be written? (default = FALSE)

Details

Required data

Rates can be estimated for multiple diel time-series in one execution of the function. 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). Missing timesteps are not allowed; incomplete days will be excluded from analyses (the number of rows must equal number of periods in a day of the specified interval; e.g. 144 rows for 600-second intervals). The csv files must include the column names Date, Time, I, tempC, DO.meas, atmo.pressure and salinity. These are case sensitive. See data 'Yallakool' for example.

The measurement interval must be the same across all days and input files run in any single analysis. Data should be arranged so that the first point of each day is midnight.

Model iterations

The number of MCMC iterations is set to 20000 iterations with 10000 burn-in, which should be sufficient in most cases. The number of required iterations can be assessed visually and by inspecting the convergence statistics (see Return). If update.chains=TRUE, the chains will be automatically updated once if convergence criteria are not met (no extra burn-in). The number of additional iterations is the number of iterations (`n.iter`) multiplied by extra.iter (default = 1; the same number of iterations again). A larger number of iterations or longer burn-in may aid model convergence in some cases.

When bayesmetab is run, jags.parallel calls JAGS, which runs in the background (R will pause for some minutes). This may take several minutes per diel cycle depending on the number of iterations. A progress bar is only shown if the chains are updated when unconverged (i.e. when `update.chains=TRUE`), as these additional iterations do not occur in parallel (and so are slower).

Optional data smoothing

A fast Fourier transform can be used to filter out high-frequency fluctuations in DO (Gallegos et al. 1977), potentially enhancing the diel signal of interest. 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 with `smooth.PAR = TRUE`. Smoothing is applied to each csv file separately, so it is recommended to only use if your data are continuous periods longer than one day. Smoothing should not be used if the dates within each csv file are not contiguous, as smoothing will then be applied inappropriately across jumps in time.

Optional customisation of priors and estimated parameters

`K.est` defaults to TRUE, meaning that K is estimated from the model and data with uninformative priors (but realistically bound; $0 \leq \mathrm{K} \leq 40$ day^{-1}). Alternatively, K can be informed with a measured or inferred mean and uncertainty (e.g. based on SF_6 injections or stream morphology). In this case, `K.est` should be set to FALSE, and the mean and standard deviation provided to `K.meas.mean` and `K.meas.sd`, respectively. This mean and standard deviation will be used in the normally-distributed prior for *K*.

By default, the function estimates A, R and K (i.e. a 3-parameter model), while the parameters θ and p have fixed values. Alternatively, θ and/or p may also be estimated within narrow, physically realistic bounds (i.e. 4- or 5-parameter model), which may enhance model fit. A 4- or 5- parameter model can be run by switching `p.est` and/or `theta.est` to TRUE.

Return

Dataframe of metabolic results and metrics for model convergence and fit. This dataframe, a table of instantaneous rates and plots for validating model fit and smoothing are also written to the pathway specified by results.dir.

Description of columns in results dataframe:

Column	Description
File	Name of the csv file
Date	Date of the diel cycle
GPP.mean	Estimated mean rate of gross primary production (mg O_2 L^{-1} day^{-1})
GPP.sd	Standard deviation of estimated GPP (mg O_2 L^{-1} day^{-1})
GPP.median	Median of GPP posterior distribution (mg O_2 L^{-1} day^{-1})
ER.mean	Estimated mean rate of ecosystem respiration (mg O_2 L^{-1} day^{-1})
ER.sd	Standard deviation of estimated ER (mg O_2 L^{-1} day^{-1})
ER.median	Median of ER posterior distribution (mg O_2 L^{-1} day^{-1})

Column	Description
NEP.mean	Estimated mean rate of net ecosystem production (mg O_2 L^{-1} day^{-1})
NEP.sd	Standard deviation of estimated NEP (mg O_2 L^{-1} day^{-1})
NEP.median	Median of NEP posterior distribution (mg O_2 L^{-1} day^{-1})
PR.mean	Estimated mean P:R ratio (unitless)
PR.sd	Standard deviation of estimated P:R
PR.median	Median of P:R posterior distribution
K.mean	Estimated mean reaeration (K) rate (day^{-1})
K.sd	Standard deviation of estimated reaeration (K) rate (day^{-1})
K.median	Median of reaeration (K) rate posterior distribution(day^{-1})
theta.mean	Mean value of θ (equals 1.07177 if `theta.est=FASLE`)
theta.sd	SD of θ (equals 0 if `theta.est=FALSE`)
theta.median	Median of θ posterior distribution
A.mean	Mean estimated A
A.sd	SD of estimated A
A.median	Median of estimated A posterior distribution
p.mean	Mean value of exponent p (equals 1 if `p.est=FALSE`)
p.sd	SD of exponent p (equals 0 if `p.est=FALSE`)
p.median	Median of p posterior distribution
R2	Coefficient of determination between observed and modelled DO. Can be inflated due to the temporally autocorrelated nature of the timeseries.
PPP	Posterior predictive p-value. 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 value close to 0.5 indicates a very plausible model, while values <0.1 or >0.9 indicate the model is an implausible explanation of the observed data.
rmse	Residual mean square error. The rmse is specific to the magnitude of the dataset and should be assessed against models from days at the same site.
rmse.relative	rmse expressed relative to the point-to-point variation in the dataset
mrl.fraction	Maximum run length fraction. The proportion of timesteps 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, which may nonetheless produce a high R2.
ER.K.cor	Correlation between sampled values of ER and K.
convergence.check	Logical check for whether all $\hat{R}$ values are < 1.1 (see 'model validation' below)
A.Rhat	Gelman-Rubin statistic ($\hat{R}$) for A. Assesses convergence of chains. Values close to 1 indicate good convergence, while values >1.1 indicate poor mixing of the chains.
K.Rhat	Gelman-Rubin statistic ($\hat{R}$) for K
theta.Rhat	Gelman-Rubin statistic ($\hat{R}$) for θ
p.Rhat	Gelman-Rubin statistic ($\hat{R}$) for p
R.Rhat	Gelman-Rubin statistic ($\hat{R}$) for R
GPP.Rhat	Gelman-Rubin statistic ($\hat{R}$) for GPP
DIC	Deviance Information Criterion. Assessment of how well the model will predict a replicate dataset taking into account model complexity. Can be used to select whether a 3- 4- or 5-parameter model is the most parsimonious. Lower DIC is desirable, with a difference of ≥ 5 between models generally indicating that the model with lower DIC best predicts the data. May be negative.
pD	Effective number of parameters. 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.
totDailyLight	Sum of PAR

Column	Description
aveDailyTemp	Mean water temperature
interval	Analysis metadata. Specified measurement interval (seconds)
smooth.DO	Analysis metadata. Specified degree of smoothing for DO (numeric)
smooth.PAR	Analysis metadata. Specified smoothing of PAR (logical)
n.iter	Analysis metadata. Specified number of iterations.
n.burnin	Analysis metadata. Specified burnin of burnin iterations.

Model validation

Metabolic estimates from poor-fitting or unconverged models are unreliable. As a general rule, a validated model meets the requirements that all parameters have converged (all $\hat{R} < 1.1$) and PPP is between 0.1 and 0.9 (closer to 0.5 is better). Checking the mrl.fraction or visually confirming for lack of consistent over- or underestimation is also strongly recommended (see below).

In addition to the results dataframe, the function may optionally return a separate table of metabolic rates for each timestep (instantaneous_rates.csv), written to the results.dir. 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 calculated according to the instantaneous PAR and the daily mean estimate of A and p. The units are per timestep, e.g. mg O_2/L/600seconds for the example dataset.

As additional validation tools, two multi-panel jpeg files are printed in the results.dir for each diel period (with the system time of the analysis appended on the file name).

The first figure shows MCMC traceplots and scatterplots of the data and model fit (Figure 1). The first five panels of this figure are trace plots for the samples of parameters A, p, R, K and θ. Convergence is indicated by the three chains (indicated by red, green and blue) being well mixed (i.e. overlapping) and stationary. When set as fixed values the trace plots for p and θ show a horizontal lines (i.e. no variation in the samples). The other plots are scatter plots of DO (showing measured, smoothed if applicable and the model fit), PAR (measured and smoothed if applicable) and temperature. 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 assumptions of the model; the results are only as a good as the underlying data collection. For example, a sharp increase or decrease in temperature or DO may indicate another source of water entered the system, or a lack of diel signal may indicate low biological activity compared to reaeration.

The second plot shows density plots of the posterior distributions of GPP, ER, A, p, K and theta (Figure 2). These plots can be used as an additional check of convergence of the three chains. For each rate or parameter, the three distributions (one for each chain) should be similar. Further, the density plots can be used to assess whether the posterior distributions are normally distributed. If they are not normally distributed, the median may be a better representation of the parameter estimate if further inference is to be made (e.g. to compare reaeration or PI curve parameters among sites).

Examples

```
##View example data set
#set path to example data.
data.dir <- system.file("extdata", package = "BASEmetab")
ex.data <- read.csv(file.path(data.dir, "Yallakool_example.csv"))
head(ex.data)
tail(ex.data)

##Run Example
```

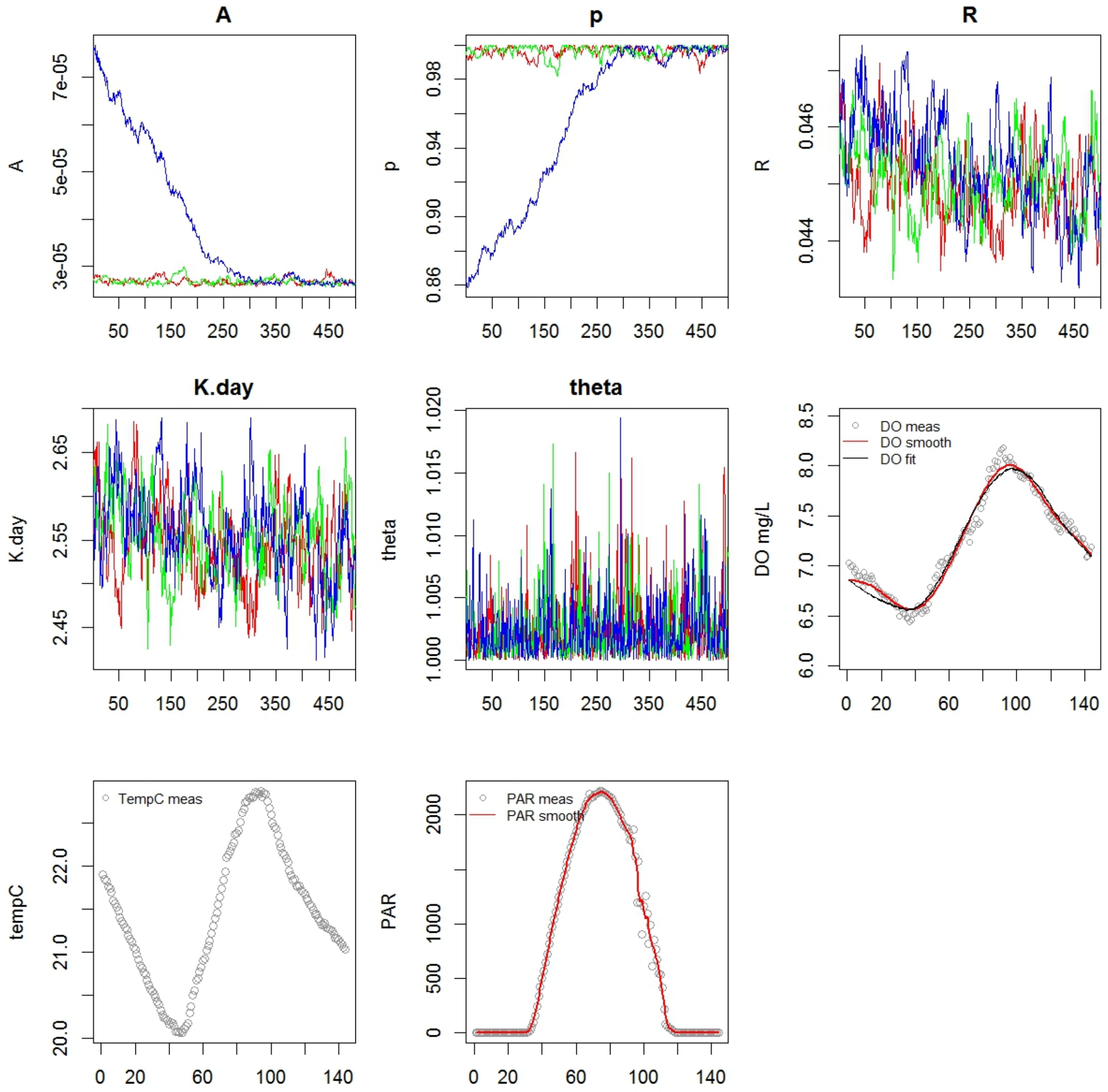

Figure 1: Example of the trace and fit plots that are printed to the specified output directory for model validation. MCMC traceplots show the three chains in red, green and blue. In this example, the chains for A and p are not converged, indicating the model requires more iterations with a longer burn-in. Parameter θ is better mixed and more stationary than R or $K.day$. The scatterplots show DO concentration (measured, smoothed and fit), water temperature and PAR (measured and smoothed) over the diel period.

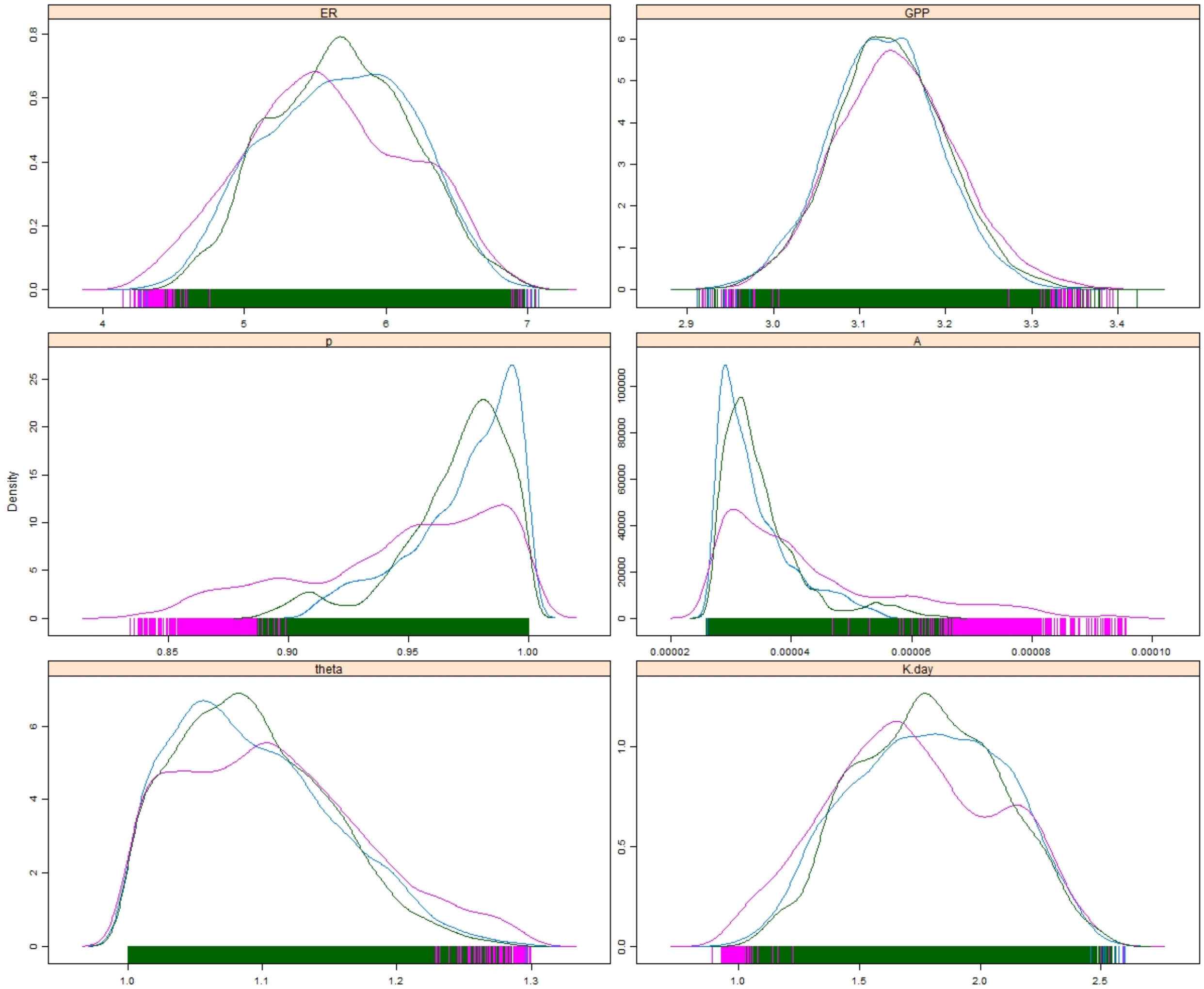

Figure 2: Example of the density plots that are printed to the specified output directory for model validation. In this example, the parameters p and A are skewed, so that the reported mean will not be an accurate representation of the distribution for any subsequent calculations. Convergence could also be better in this example.

```
#set output directory to Output folder in current working directory.
results.dir <- file.path(getwd(), "Output")
if (dir.exists(results.dir)){} else {
dir.create(results.dir)}

#run model
results <- bayesmetab(data.dir, results.dir, interval=600)
```

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.

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.

Yallakool_example (dataset)

Description

An example set of diel curves from Yallakool Creek, New South Wales, Australia. These data can be used to familiarize yourself with the model and as a template for data input.

Useage

```
data.dir <- system.file("extdata", package = "BASEmetab")
ex.data <- read.csv(file.path(data.dir, "Yallakool_example.csv"))
head(ex.data)
tail(ex.data)
```

Details

The format for csv/dataframes must contain the case sensitive column names Date, Time, I, tempC, DO.meas, atmo.pressure and salinity, as shown below:

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:

Parameter	Description
Date	Ideally in format yyyy-mm-dd. Cannot contain slashes.
Time	Ideally hh:mm:ss. Number of time periods must equal a complete 24 hour period for each day of analysis. Additional leading and trailing part-days may be included for smoothing purposes but will be ignored by function bayesmetab. There can be no missing measurements, but interpolating short periods has been successfully applied (Obrador et al. 2014; Giling et al. 2017).
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 if salinity is low and unknown.

References

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

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.