Supplement to "mechanistic site-based emulation of a global ocean biogeochemical model for parametric analysis and calibration" by J. C. P. Hemmings, P. G. Challenor & A. Yool

7th June 2014

This supplement provides reference data and MarMOT input data that can be used to reproduce the 1-D simulator experiments described in the main article. Instructions are given below for reproducing the evaluation experiments for the mechanistic site-based emulator with direct and indirect uncertainty quantification. All files provided in the directory `data` are ASCII text files.

S.1 MarMOT software availability

MarMOT 1.1 was released on 21st November 2013 under the CeCILL Free Software License Agreement. The software is designed for use on UNIX-based systems, including LINUX and MacOSX. A tar archive containing the MarMOT 1.1 open source distribution can be downloaded from the National Oceanography Centre's web site at http://noc.ac.uk/project/marmot or supplied by the corresponding author on request (e-mail: jhemmings@wessexenv.mail1.co.uk). The software release includes a set of command line tools for handling MarMOT-compatible data tables. Full documentation and test data are included with the distribution.

S.2 Reference data

The file `siteinfo` gives a table with one record for each site. Sites are identified by labels of the form **`site`** = `jAAAiBBB`, where `AAA` is a 3 digit number giving the *j* co-ordinate of the site on the NEMO grid and `BBB` is a 3 digit number giving the *i* co-ordinate. The NEMO grid has a nominal horizontal resolution of 1°. The variables **`nav_lat`** and **`nav_lon`** give the geographic location corresponding to the NEMO grid point. The variables **`nom_lat`** and **`nom_lon`** give a nominal geographic location for the site. Nominal locations are points on a 1° grid. Sites were selected by choosing the nearest NEMO grid point to the nominal site locations.

The file `ref_nemo10m` contains the surface chlorophyll output data from the 3-D target model reference ensemble (units are mg m^{-3}). These are the NEMO-MEDUSA 'truth' data for evaluating the performance of the 1-D simulator array and quantifying uncertainty in the simulator output. The variable **`member`** is used to indicate the parameter vector number. Variable **`t`** gives the time in days since the start of 1997, **`year`** gives the calendar year and **`tofyr`** is the time in days since the start of the current year.

S.3 MarMOT input data

A MarMOT experiment is configured using an input file containing an *experiment control table*. The control table gives the names of the other input files required and the output files that are to be produced. (An additional file `gfan.log` is automatically written in the current directory.) Each input file referenced in a control table contains one of the following tables.

- *parameter set item table*
- *gridded domain item specification table*
- *case table*
- *output variable selection table*

Parameter set item tables are used to set scalar values that are constant for a simulation. Gridded domain items are used to set up the vertical grid and the physical and biogeochemical environments. A case table is used to indicate which simulations are to be run, in terms of site and ensemble member identifiers. Output variable selection tables are used to specify variables to be included in the output tables.

Each gridded domain item specification table has an associated *item data table* in a separate file.

Further details, including table formats and recognized variables, can be found in the MarMOT user guide. Templates for the input data files used in this study are available in the directory `templates`. (A full set of templates can be automatically generated using MarMOT, if required.) The template file for each type of input table includes all variables that are recognized in that type of table. It also gives descriptions and units for each variable.

S.3.1 Informed simulator array

The following files are used to set up the 1-D experimental framework for the informed simulator array. The simulator array is run for each trial parameter vector, identified in the relevant tables by the variable **`member`** = `1…10`.

The file `ex_10m_isim` contains the control table, which references the following files from the `in` and `vars` directories.

Parameter set item tables:

- **`model3f`** = `in/pset_medusa_nemo10` contains the model parameter table (gives the representative sample of vectors from the MEDUSA parameter space; see `gfan.template.model3f` for descriptions and units).
- **`optionf`** = `in/option_medusa_pertrt` contains the option parameter table (gives simulation options).
- **`taxisf`** = `in/taxis_base1997` contains the time axis parameter table (defines time axis with origin at start of 1997).
- **`timeperiodf`** = `in/tperiod_97-98` contains the time period parameter table (defines start and finish times).
- **`environf`** = `in/environ_upper` contains the environment parameter table (gives maximum water column depth and latitude).

Gridded domain items:

- **zlevelf** = in/zlevel contains the vertical grid specification table. zlevel.dat contains the vertical grid data table (see gfan.template.zleveldataf).

- **initf** = in/init_10m contains the initial profile specification table. init_10m.dat contains the initial profile data table (gives initial state in the form of vertical profiles for tracers and composition ratios; see gfan.template.initdataf).

- **ftf** = in/ft_std contains the scalar forcing specification table. ft_std.dat contains the scalar forcing data table (gives time series of surface irradiance and dust deposition; see gfan.template.ftdataf).

- **fktf** = in/fkt_std contains the physical profile forcing specification table. fkt_std.dat contains the physical profile forcing data table (gives time series of vertical profiles of vertical diffusion, vertical velocity and temperature; see gfan.template.fktdataf).

- **fkt2f** = in/fktpert_sqrt_10m contains the biogeochemical profile forcing specification table. fktpert_sqrt_10m.dat contains the biogeochemical profile forcing data table (gives time series of vertical profiles of applied perturbations for each tracer in square root units; see gfan.template.fktdataf; for untransformed tracer units see gfan.template.initdataf).

Case table:

- **casef** = in/case_10m contains the case table for running the simulator at all 12 sites for all 10 parameter vectors.

Output variable selection tables:

- **outtdayvarf** = vars/outvar_scal contains the variable list for the scalar output table.
- **outktdayvarf** = vars/outvar_prof contains the variable list for the profile output table.

See gfan.template.outtdayvarf and gfan.template.outktdayvarf for complete lists of available variables. Output tables are written to the directory outisim as indicated in the control table. This directory is expected to exist prior to running the experiment.

Experiments can be run without the effects of lateral advection by removing or commenting out the **fktpertf** variable in the control table in ex_10m_isim.

S.3.2 Uninformed simulator array

For an uninformed simulator experiment, the simulator array is run for each trial parameter vector with its biogeochemical environment equal to the mean environment based on the other 9 parameter vectors.

The file `ex_10m_usim` contains the control table. Output tables are written to an existing directory `outusim`.

The control table references two modified input items:

- **`initf`** = `in/init_10m_9mav`
- **`fkt2f`** = `in/fktpert_sqrt_10m_9mav`

The input files for these items are not supplied but can be derived as follows. The new initial profile item can be derived from the input item used for the informed simulator array (Section 3.1 above):

- The item specification table (file `init_10m_9mav`) should be an identical copy of the original (file `init_10m`). Number of records = 120 (12 sites × 10 parameter vectors).

- The item data table (file `init_10m_9mav.dat`) should have the same variables and number of records as the original (file `init_10m.dat`) but the initial state values should be replaced by the appropriate mean values based on 9 parameter vectors. Number of records = 4440 (120 instances × 37 depth levels).

The new biogeochemical profile forcing item is derived in the same way:

- The item specification table (file `fktpert_sqrt_10m_9mav`) should be an identical copy of the original (file `fktpert_sqrt_10m`). Number of records = 120 (12 sites × 10 parameter vectors).

- The item data table (file `fktpert_sqrt_10m_9mav.dat`) should have the same variables and number of records as the original (file `fktpert_sqrt_10m.dat`) but the initial state values should be replaced by the appropriate mean values based on 9 parameter vectors. Number of records = 648240 (120 instances × 37 depth levels × 146 time points).

S.3.3 Direct uncertainty quantification for the uninformed simulator array

For each of the 10 trial parameter vectors, we need a separate uncertainty quantification based on uninformed simulator performance for the other 9 parameter vectors, so we run a 90-member ensemble. For each member, the uninformed simulator array requires a mean environment determined from the target model output for the remaining 8 parameters.

The 90 ensemble members are identified by labels of the form **`member`** = `expXXsYY` where `XX` is a 2 digit experiment number, equal to the number of the trial parameter vector, and `YY` is a 2 digit simulation number, equal to the number of the parameter vector used in the simulation.

To avoid the need to handle very large files, some of the tables involved are split into 10 separate files, one for each experiment.

The file `ex_90m_uqd.XX` contains the control table for experiment `XX`. The file `case_90m.XX` contains the corresponding case table. Output tables are written to an existing directory `outuqd.XX`. The variable **`model3key`** in the case file indicates which parameter vector is to be used.

The control table for experiment `XX` references 3 modified input items:

- **`model3f`** = `in/pset_medusa_nemo10a`
- **`initf`** = `in/init_90m_8mav`
- **`fkt2f`** = `in/fktpert_sqrt_90m_8mav.XX`

The file `pset_medusa_nemo10a` is identical to the original parameter set item file `pset_medusa_nemo10` except that the variable name **`member`** is changed to **`model3key.`** This is because explicit referencing is required, in place of contextual referencing, since the member label does not now match the parameter set number.

The new initial profile item files and the files for the 10 biogeochemical profile forcing items are not supplied. They can be derived as follows.

The new initial profile item can be derived from the input item used for the informed simulator array (Section S.3.1):

- The item specification table (file `init_90m_8mav`) should have the same variables as the original (file `init_10m`) but should contain 90 members for each site instead of 10. The variable **`member`** takes the values indicated above. Number of records = 1080 (12 sites × 90 members).

- The item data table (file `init_90m_8mav.dat`) should have the same variables as the original (file `init_10m.dat`) but should contain 90 members per site, matching those in the specification table. The initial state values should be replaced by the appropriate mean values based on 8 parameter vectors. Number of records = 39960 (1080 instances × 37 depth levels).

The new biogeochemical profile forcing items are likewise derived from the input item used for the informed simulator array. For experiment `XX`:

- The item specification table (file `fktpert_sqrt_90m_8mav.XX`) should have the same variables as the original (file `fktpert_sqrt_10m`) but should contain the 9 members for each site for experiment number `XX`. Once again, the variable **`member`** takes the values indicated above. Number of records = 108 (12 sites × 9 members).

- The item data table (file `fktpert_sqrt_90m_8mav.XX.dat`) should have the same variables as the original (file `fktpert_sqrt_10m.dat`) but should contain 9 members per site, matching those in the specification table. The applied perturbation values should be replaced by the appropriate mean values based on 8 parameter vectors. Number of records = 583416 (108 instances × 37 depth levels × 146 time points).

S.3.4 Indirect uncertainty quantification for the uninformed simulator array

Indirect uncertainty quantification for the uninformed simulator requires performance statistics for the informed simulator, together with statistics for the parametric environment residual derived from an uncertainty analysis with 100 different environment realizations. The informed simulator statistics are derived by comparison of 9-member subsets of informed simulator output (see Section S.3.1) with the target model truth (see Section S.2). The configuration for the environmental uncertainty analysis is described below.

For each of the 10 trial parameter vectors, we need statistics based on a 100-member simulator array ensemble with that trial parameter. 1000 runs of the simulator array are therefore required. The environment realizations are generated using a statistical environment model based on the other 9 parameter vectors.

A recommended configuration of the analysis uses 10 control files, 1 for each trial parameter vector experiment. Separate case tables are used for each site to produce a total of 120 output data sets, each containing output for 100 members. The ensemble members for each experiment are identified by **`member`** `= 1…100` and the variable **`model3key`** in the case tables indicates which parameter vector to use. Member numbers are re-used between experiments.

For an experiment `XX`, the control table is set up in a file `ex_ens_uqi.XX`. The corresponding case tables are contained in 12 files with names of the form `case_ens.jAAAiBBB.XX`. Output tables are written to 120 existing directories with names of the form `outuqi.jAAAiBBB.XX`. Files `ex_ens_uqi.01` and `case_ens.j162i268.01` are provided as examples.

New initial profile item tables and biogeochemical profile forcing item tables are required containing environment data generated from the result of an EOF analysis. The input data for the EOF analysis can be derived from the original item data files `init_10m.dat` and `fktpert_sqrt_10m.dat`.

The required files are:

- 10 files `init_ens.XX` with number of records = 1200 (12 sites × 100 members)
- 10 files `init_ens.XX.dat` with number of records = 1200 (1200 instances × 37 depth levels)
- 120 files `fktpert_sqrt_ens.jAAAiBBB.XX` with number of records = 100 (100 members)
- 120 files `fktpert_sqrt_ens.jAAAiBBB.XX.dat` with number of records = 540200 (100 instances × 37 depth levels × 146 time points)

As before, the tables in these files should have the same variables as the originals.