Review of: "Auto-calibration of a one-dimensional hydrodynamic-ecological model (DYRESM 4.0-CAEDYM 3.1) using a Monte Carlo approach: simulations of hypoxic events in a polymictic lake", by Liancong Luo, David Hamilton, Jia Lan, Chris McBride and Dennis Trolle

18 November 2017

General comments

As someone who has used coupled hydrodynamic-ecosystem water quality models (including previous versions of DYRESM and CAEDYM) to provide science information for assisting lake managers in making decisions, I welcome this contribution by Luo et al. Aside from potentially making calibration of complex water quality models more efficient, it makes calibration less subjective, which ultimately adds to the credibility of model results. This paper describes a tool that is shown to be efficient and effective, producing convincing results in a well-documented case study. I found the paper to be clearly written, concise and well-focused, with good descriptions of technical aspects the optimization technique and associated statistics, and of the lake processes associated with the case study.

I found the literature review in the Introduction to be comprehensive and helpful, in its coverage of both water quality models and optimization methods used for model calibration, and in its description of the strengths and weaknesses of the various optimization methods. I thought the authors satisfactorily explained the reasons for their choice of MCS as an appropriate method for their DYRESM-CAEDYM applications. I agree with the authors' observation that, compared with rainfall-runoff models, there are very few examples in the literature that describe optimization methods for calibration of complex water models that contain large numbers of parameters and state variables for empirical algorithms based on biogeochemical rates – hence the need for studies like the one presented here.

I felt that the restriction of the optimization application to simulation of dissolved oxygen and prediction of hypoxic events was sensible in this prototype stage of development for an auto-calibration tool. In practice, however, phytoplankton productivity and biomass, chlorophyll-a concentrations, and nitrogen and phosphorus dynamics are often the ultimate focus of interest in water quality modelling. These latter processes and variables are accounted for in the present work in terms of the parameters listed in Table 3, but only as they affect oxygen concentrations, not measures of chlorophyll, nitrogen or phosphorus.

I wonder if the authors would care to comment (or speculate) briefly on the possible future development and applications of their auto-calibration approach to cases where it is important not only to model oxygen dynamics reliably, but also to predict one or more other measures of chlorophyll, nitrogen and/or phosphorus concentration. For example, how many other parameters would this bring into the auto-calibration process? Is there a number beyond which the procedure would not perform satisfactorily? How could the approach be extended if there were more than one variable (oxygen or temperature in the present work) of major concern in a given optimization run? Would the user have to be content with choosing the single most important variable to optimize on? Would a stepwise approach, such as described here for first optimizing temperature prediction, then oxygen dynamics, be appropriate? I realize that this is complicated by the fact that many of the same parameters that affect oxygen prediction also affect chlorophyll, nitrogen and phosphorus

concentrations. This is in contrast to the application presented in the paper where the two sets of parameters (one set for temperature, another for oxygen) are independent of each other. Perhaps an iterative strategy could be used? Or does that defeat the purpose of auto-calibration in terms of reducing time-consuming iterative procedures?

I felt that the authors' comment at the bottom of page 10 regarding the value of their approach, as qualified by: the need for experience, knowledge and expertise with the water quality model; the lake processes concerned; and the accuracy of the field data available for calibration – provided a fair assessment of both the contribution and limitation of their work: "The success of its [the auto-calibration's] application is strongly dependent on prior knowledge about parameter value ranges, the number of iterations performed which is closely related to the computer's performance capability and the accuracy of observations, but it has great potential to reduce the repetitive model iterations that are required using traditional trial-and-error calibration."

Specific comments (mostly relating to questions of clarification)

Page 5, Auto-calibration procedure for DYRESM-CAEDYM: I think the authors might consider adding some further explanation to this section to clarify some of the details of their auto-calibration procedure. Questions that arose in my mind as I read this section included:

- CAEDYM parameters not included in the optimization – How many other parameters were there? How were their values chosen? Were any changes made to these values as the calibration proceeded?
- Random search module – Could the authors supply a little more detail about how the random search model worked? For example, was each parameter chosen independently of the others? Were any particular probability distributions assumed when generating random values? (It seems that this question is answered later on page 10 [see comment below for page 5, lines 23-25] but it might be helpful to include this here.)
- Line 25: Would it help to refer to Table 1 here, to identify the parameters and values used?
- Lines 34-36: Could the authors consider supplying a little more explanation or detail about how "A single parameter file was chosen which minimised the combined *RMSE* of these variables with different weighing factors between the model simulations and measured values."

Model validation: Was any check made on model performance for temperature simulation for the validation runs (as it apparently was in Burger et al. 2008)? Or was this done only for DO? It might be good to clarify this at some stage, perhaps in section 3.2.2, or perhaps earlier, e.g. in section 3.1 (Physical parameter selection set).

Discussion section:

- Consider including some further discussion of how the timing of hypoxic events related to strength of stratification.
- Do the authors have any explanation for the under-prediction of DO concentration by the model in the validation run when measured DO concentrations were above 8 mg L^{-1}?

Technical corrections

Page 2, lines 11-12 and line 25: There are two Li et al. 2013 publications listed in the References; perhaps the authors can assign designations for 2013a and 2013b.

Page 2, line 25: Consider replacing "This traditional calibration procedure …", with "The traditional calibration procedure …"; in lines 21-24 the authors list a number of procedures, and it is not clear which one "This" refers to.

Page 3, line 8: "evaluate" – should be "evaluates".

Page 5, line14: Can the authors specify the values of the two depths at which samples were collected?

Page 5, lines 23-25: Consider specifying here the type of probability distribution used for the random search module; text on page 10,lines 17-21, indicates the this was a uniform distribution, and no further distributions were used, but it would be helpful to also provide this information earlier on when the auto-calibration procedure is being described.

Page 6, Equations 1 and 2: Should Q_i be O_i?

Page 7, Table 2: Table 2 is to be inserted after line 5, but Table 2 does not appear to be referred to anywhere in the text. Possibly it could be referred to at the end of the first sentence in line 4, in which the variation in model performance with depth is discussed.

Page 7, line 26, value for simulation minimum DO concentration: The text specifies that the minimum DO concentration from the simulation was 2.46 mg L^{-1}, but in Fig. 3 the minimum appears to be less than this, around 2.0 mg L^{-1} for the event being described. However, the time scale of Figures 3 and 4 is not so easy to follow – see comment below under "Figures 3 and 4".

Page 7, line 30: Value for minimum measured DO concentration – it appears from Fig. 3 that minimum measured DO concentration for the fourth and fifth hypoxic events was less than the value 0.72 mg L^{-1} specified in the text.

Page 8, lines 20-21: Consider specifying the months that "spring" refers to. Also, the water level decrease mentioned in the text does not seem to be shown in the bottom panel of Fig. 2, where the top of the figure seems to have been cut off – the top is straight and horizontal, and in the scale for elevation the tick for 18m has been labelled as 20m.

Page 9, lines 9-10: Are the five DO depletion events that are referred to for the calibration? Or for the validation? Or both?

Page 9, line15: Consider replacing "in the upper ranges of values…" with "greater than values…".

Page 9, line 25: Should "Burger et al. 2007" be "Burger et al. 2008"?

Figure 2: Time scale labels are missing. See also comment above (page 8, lines 20-21) – the top of the bottom panel showing simulation results seems to have been cut off (the top is straight and horizontal), and in the scale for elevation it appears that the tick for 18m has been labelled as 20m.

Figures 3 and 4, Time scale: I recommend that time scales explicitly showing the date in a day, month and year format (e.g. 13 Jul 07 for 2007194) be provided in addition, or instead of, the YYYYDDD format (presumably the format used by DYRESM) shown. This would make it easier for the reader to relate the description in the text, which refers mostly to conventional dates and only occasionally to day numbers. It also makes it easier for the reader to recognize seasonal influences. I also suggest that the authors consider adding a legend to the plots showing the difference between lines used for simulated and measured DO concentrations; this difference is included in the figure caption, but a legend would help make the plot more self-contained.

Reference section: Check alphabetical order for citations in the References (e.g., Alarcon precedes Antenucci; Copetti follows Chung;Cui follows Cox).

Page 12, line18, upper case B for Bombardelli

Page 14, lines 30-34: should there be 2013a and 2013b for the two Li et al. references? (Also noted above in comment for page 2, lines 11-12, 25.)