Review 2

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 number s 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. "

Response:

Thanks a lot for the very valuable comments. Our answers are briefly listed below:

This auto-calibration software can be used for all the water quality variables included in the CAEDYM model (e.g. TP, TN, NO3, NH4, PO4, CHLA). The only thing needs to do for that is set up all the variable information in the configure file, for example, the number of variables, variable names, the observation length of variables, and the weighting factor for each variable. The statistical assessment parameter (e.g. RMSE) will be the sum of RMSE multiplied by the weighting factor for each variable. In this paper, we just presented a DO case study to show the procedure for how the software works.

There is a switch (1 or 0) for calibrating the physical (par), chemical (chm), biological (bio) and sediment (sed) process or not. The model user can choose which file parameters need to be calibrated or not or he/she can choose all the files for calibration with switch 1 for each file.

It is true that the success of this software'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. The parameter value ranges can be mostly found from related literatures. The more iterations, the more possibility to find the "optimized" parameter set. The same with the accuracy of observations.

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?

Response:

Actually all the parameters can be potentially calibrated but some parameters are not necessary. For example, the stoichiometric parameters, the density, size and critical shear stress of different types of sediment, some non-sensitive parameters. So if the mode user doesn't know which parameter is sensitive or not, I would recommend this parameter be included in calibration. If a parameter is not included in the calibration, the minimum and maximum values should be the same value and this value can be obtained from relevant literatures or observations or experiments.

All the parameters calibrated in this case study were listed in the attached table required by the reviewer 1.

· 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.)

Response:

Thanks for the comments. Answers below.

Was each parameter chosen independently of the others?

Yes, a random number for each parameter is independent.

Were any particular probability distributions assumed when generating random values?

No. Just random number.

· **Line 25: Would it help to refer to Table 1 here, to identify the parameters and values used?**

Response (now 5:32):

Thanks for the comments. Tab. 1 is refered at 5:34

· **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. "**

Response (now 5:40-6:2):

Thanks for comments. We have made explained the "combined RMSE" before this sentence (The combined RMSE is the sum of RMSEs for chlorophyll a and DO with each RMSE multiplied by an arbitrarily chosen weighing factor varying from 0 to 1. see 6:2). "A single parameter file was chose…" has been changed to "The four DYRESM-CAEDYM parameter files (par, bio, chm and sed) were chosen which minimised the combined RMSE of these variables with corresponding weighing factors between the model simulations and measured values." (6:2 – 6:4).

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) .**

Response:

Thanks for comments. We did not check the model performance for temperature simulation for the validation runs but we did that for the calibration. The DYRESM model has been previously successfully applied to water temperature simulations. In this paper, its performance in the calibration process was also good and moreover we focused on the software development and application with a case study of hypoxic event. That's way we didn't check its performance for temperature simulation in the validation.

Discussion section:

· **Consider including some further discussion of how the timing of hypoxic events related to strength of stratification.**

Response:

Thanks a lot for the nice comments. It looks very scientifically interesting to provide substantial information about how the timing of hypoxic events related to strength of

stratification. There will be a lot of work to do, for example, how to define the hypoxic events, how to decide the timing of hypoxic events, and how to quantitatively calculated the strength of stratification (e.g. lake number). These parameters are strongly dependent on hydrodynamics and meteorological conditions. We wouldn't include this work in this paper since we are just addressing the principles of the auto-calibration for DYRESM-CAEDYM with a case study of hypoxic events in this paper, but would be very happy to work on another paper about that.

· **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} ?**

Response:

I presume that the problem might be from the model itself. I talked to Professor David Hamilton about this problem before and he had no idea either. However, we could fix the problem by setting the "Photo-respiration phytoplankton DO loss" as negative value in the "chm" file although this is not the reality.

I have applied DYRESM-CAEDYM to a Chinese reservoir (Lake Qiandao) and got the same problem with under prediction of DO concentration. So I am still thinking about where the problem is possibly from.

Technical corrections

Page 2, line s 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 .

Response:

Thanks for the nice comments.

We have added "a" to the first reference paper and "b" the second one. We have also correspondingly changed the citation at 2:14 and 2:28.

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.

Response:

"This" has been changed to "The" at 2:28

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

Response:

Done at 3:12. Thanks for the comments.

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

Response (5:19):

One is 1 meter below surface and the other one is 19 meters below surface. We have added "(1 meter and 19 meters below surface)" after "two depths" in the text.

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.

Response:

Thanks for the comments. 5:23-5:25 is now probably 5:29-5:31 referring to "A random search module was then run for all remaining parameters to produce files with combinations of parameters which could then be used to generate independent runs of the DYRESM hydrodynamic module." Which has been changed according to previous comments. The random module produces random number without any consideration on probability distribution.

*Text 10:17-10:21 (now 10:25-10:29) is "Random Monte Carlo simulation, as adopted in our study, has the advantage of being easily incorporated into model code and programming, and can also include adequate consideration of "equifinality" of water quality models with large sets of parameters, without the need for the user to make assumptions regarding parameter distributions (as a simple uniform parameter distribution within the defined range is used)." **Here we just say the Random Monte Carlo simulation has the advantage of being easily incorporated into model code and programming and DOES NOT NEED TO CONSIDER PARAMETER DISTRIBUTIONS.***

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

Response:

Changes have been made in the text.

Page 7, Table 2: Table 2 is to be inserted after line 5, but Table2 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.

Response:

Thanks a lot for letting me know this mistake. Tab. 2 has been inserted to the end of 7:7-7:8 (see below).

The Pearson correlation coefficient (r) between model output and measured temperature over all depths exceeded 0.98 with a RMSE of < 0.71 °C (Tab. 2).

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 ".

Response (now 7:34 – 7:35):

In the text, it is "The first occurred during 7 Sep–1 Oct 2007 with minimum DO of 3.95 mg L^{-1} (simulation 2.46 mg L^{-1}) on 24 Sep (Julian day 267)". That means the minimum observed DO in this hypoxic event was 3.95 mg L^{-1} and the corresponding modeled DO was 2.46 mg L^{-1} on 24 Sep. It was not talking about the modeled minimum DO.

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.

Response (now 7:39):

Thanks for the valuable comments. The bottom DO started to decrease from 3 Dec. 2008 and the lowest DO concentration (0.72 mg L^{-1}) was found on 10 Dec 2008. On the next day, the bottom DO increased until 13 Dec. 2008 and began to decrease on 14 Dec. 2008. The minimum value (0.32 mg L^{-1}) was found 16 Dec. 2008 and then the observed DO increased dramatically from 0.32 mg L^{-1} to 7.5 mg L^{-1} in two days. However, the simulations followed the similar pattern but were much higher than the observations. The simulated DO was 7.7 mg L^{-1} on 16 Dec. 2008 while the observed value was 0.32 mg L^{-1}. So we thought the problem was probably from the meteorological conditions driving DYRESM-CAEDYM during 13 – 16 Dec 2008 but there is no any other available meteorological data. So we excluded this period from the fourth hypoxic event time because we really didn't know why there was huge difference between the observations and simulations at that time. So the fourth hypoxic event was from 3-10 Dec. 2008. In the fifth event, the lowest DO was 0.5 mg L^{-1} on 11 Jan 2009 after my double check.

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.

Response (now 8:30):

Thanks for the comments. "Spring" has been changed to "January and February". We have replaced the figure with the originally produced figures, which looks clearer . The figure caption has been changed to "Comparison of observed (top) and simulated (bottom) water column temperature based on daily data from 13 Jul. 2007–13 Jan. 2009."

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

Response (now 9:16):

Thanks for the comments and sorry for the confusion. In the text, "in the calibration process" has been added to the end of "Simulated bottom DO represented observations well" at 9:16.

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

Response (now 9:22):

Done at 9:22.

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

Response (now 9:32):

Yes it is "2008". Done at 9:32

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 h as been labelled as 20m.

Response:

I replaced the figure by the original picture produced by Modeller 2.0 which looks clearer. I would be very happy if this clearer figure doesn't meet the journal standard.

Figures 3 and 4, Time scale: I recommend that time scales explicitly showing the date in a day , month an d 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.

Response:

The format has been changed according to the comments. Now it is at "mm/dd/yy" format which will be easier to know the season information. Legend has been added to the plottings.

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

Response:

The reference order has been adjusted.

Page 12, line18, upper case B for Bombardelli

Response (12:22):

Done at Line 12:22.

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 .)

Response:

"a" and "b" have added to the two references and the citation has also been changed in the text.