Review of “Optimizing models of the North Atlantic spring bloom using physical, chemical and bio-optical observations from a Lagrangian float” by Bagniewski et al.

In this manuscript the authors assimilate Lagrangian float data into three models of biological carbon export, with the aim of finding “the optimal biological model variant for simulating the bloom and quantifying the associated carbon export, and to investigate the importance of including diatom aggregation triggered by low silicate concentrations”. Their findings include (1) None of the models can be formally rejected based on their misfit with the available observations, but the model with diatom aggregation has a slightly better fit, and (2) the models give similar export at 100m, but different rates of export at 600m.

This is an interesting topic and the authors certainly have an excellent data set to use in their modeling work. Unfortunately, however, I am not convinced that the authors have successfully demonstrated that the model with diatom aggregation has a better fit and is thus the ‘optimal’ model. There is also text in the Discussion that seems to indicate misunderstandings in regards to past work on this topic of using data assimilation to select an ‘optimal’ ecosystem model.

General comments:

A major conclusion of this study is that the most complex model, i.e. the one with the greatest number of optimized parameters, fit the available data the best. This is based on the fact that the cost functions for the three models (in order of increasing complexity) are 5.6, 5.0 and 4.6. However, no uncertainties are provided for these cost functions. The observations used in the cost are associated with a certain degree of uncertainty. These should be able to be used to define at least a minimum uncertainty in the cost function. How do we know that 5.6 is a significantly greater cost than 4.6? Since this is a major conclusion of the paper, uncertainty estimates are critical. Additionally, it would be interesting to know more about how large these costs were prior to assimilation. Did the costs of the three models start out equally large?

Another relevant issue here, and one that has been demonstrated and described at length in a number of recent papers but is not discussed here, is the fact that as the number of optimized parameters grows, models tend to be less able to reproduce unassimilated data, i.e. data collected from different times or locations or from different instruments. This is a key point that is currently missing from this analysis. Here the most complex model is deemed the ‘best’ because of the fact that that it produces the lowest cost function; however this analysis would be more convincing if instead of simply assessing the skill of the models based on the model-data misfit of the assimilated data, the authors withheld some of the data from the assimilation, and assessed the skill of the model based on these unassimilated data (Gregg et al., 2009). If the authors defined skill in this way, they would be very likely to see that the skill of their models would decrease as more and more parameters are optimized. Admittedly some unassimilated *qualitative* data are used to judge the models (p. 8500), but these comparisons aren't provided to the reader in the form of a table or figure; instead the reader simply must take the authors' word that the 2p2s model fits these qualitative data the best. In order to be convinced that the 2p2s model is indeed the 'optimal' model, I would have liked to have seen a quantitative comparison to unassimilated data.

The fact that the model with the greatest number of free parameters fit the model the best is not a surprising conclusion. When fitting a model to data, the model with the greatest number of tunable parameters will nearly always fit the data the best. This point has also been made in many recent papers, including in the Friedrichs et al. (2006) and (2007) papers that are cited here. In fact I find it disappointing to see that the authors cite Friedrichs et al. (2006) as an example of a study that found that the more complex models performed worse than the simpler models. In fact, when all parameters were optimized, the simplest and most complex models in this study did not perform significantly differently (except when using the F2 forcing which was shown to have problems). The fact that the most complex model did not perform better than the simple model was merely an indication that the cost function was stuck in a relative minimum as a result of attempting to optimize too many parameters.

On p. 8497, Friedrichs et al. (2007) is cited as an example of a study where "the model does not adequately represent the system that is studied" because only 2-4 parameters were optimized. There seems to be some misunderstanding here. The issue in the Friedrichs et al. (2007) study was that the models had the highest skill (when skill is judged by comparing model results to unassimilated data) when fewer parameters were optimized. Friedrichs et al. (2006) demonstrated that the cost function is indeed much lower when 10-20 parameters are optimized (Experiment 1, Table 1) than when 2-6 parameters are optimized (Experiment 2, Table 1). However, if we look at the model-data misfit for unassimilated data (J_P), we see this misfit is lower when fewer parameters are optimized. The choice of assimilating only 2-4 parameters for each of the twelve models in no way indicated that none of the models adequately represented the system.

There are a number of other papers that are relevant here. Including a discussion of these would have strengthened this manuscript:

Anderson, T. R., 2005. Plankton functional type modeling: running before we can walk? J. Plankton Res., 27, 1073– 1081.

Gregg, W.W. et al., 2009. Skill assessment in ocean biological data assimilation. Journal of Marine Systems 76, 16–33.

Pahlow, M. et al., 2008. Adaptive model of plankton dynamics for the North Atlantic. Progress in Oceanography, 76, 151-191.

Ward, B.A. et al., 2010. Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. Journal of Marine Systems 81, 34-43.

A few specific comments:

Eqn. 19: It would be helpful to define the delta function here.

Eqn. 24: How do the units work here? Please provide values and units of the sigmas used here, as well as the number of observations of each type of variable. The authors state that each variable contributes equally to the cost, but they should point out that this is only true before assimilation. Also, how could this be true for all the models? Since this equation is defined to be the equation of the cost function F_R should be included here.

p. 8490, line 2: Why are cysts not included in PON?

p. 8490, line 11: Again, F_R has the same weight before assimilation, but presumably not after assimilation? It would be helpful to see the individual components of the cost function after assimilation.

Table 3: Caption should be changed to make it clear that non-optimized parameters are listed here too. Also, it is important for the reader to be able to see the a priori values as well as the a posteriori values for each of these parameters.

Figures 4-7: These figures would be clearer if symbols (rather than a dashed line) were used for the observations. Then the reader would clearly be able to see the temporal resolution of the data. (Daily below MLD and every minute or two within the MLD?)

p. 8495, line 16: I would appreciate having more information about how the initial conditions were set for the different models. Why is the initial condition of zooplankton different for the three models?

Table 4: For comparisons with other export calculations, it would be helpful to have export reported as the average molC/m2 per day, rather than the total for 34 days.

p. 8496: Given that the major conclusions of the manuscript focus on export rates, a figure showing a time series of export (perhaps divided into its components) for each model would be very appropriate and helpful.