Biogeosciences Discuss., 9, C4129–C4132, 2012
www.biogeosciences-discuss.net/9/C4129/2012/

Biogeosciences Discussions

Interactive comment on "Reducing the model-data misfit in a marine ecosystem model using periodic parameters and Linear Quadratic Optimal Control" *by* M. El Jarbi et al.

M. Dowd (Referee)

mdowd@mathstat.dal.ca

Received and published: 21 September 2012

This primary contribution of this paper is the incorporation of seasonally varying parameter values for a model of plankton ecology/biogeochemistry. Moreover, this is done over multiple years allowing identification of an annual cycle for parameters. A depth-resolved PNZD model is used along with the BATS observations. By allowing for temporal changes about baseline parameter values, the model is able to fit the data extremely well. The unstated caveat here is that there is a great deal of prior information that is used in terms of a baseline parameter set which provides a reference about which the annual cycle is fit. This is along with the fact that varying a large

numbers of parameters simultaneously allows the flexibility for the model to match the data (which could be interpreted as over-fitting). That said, the results – in the form of annual cycles of parameters – are encouraging and look plausible. Annual cycles for parameters are an important means to make these models more realistic, and to learn about the underlying unresolved processes. The optimal control used here relies on linearization. It would not be my first choice as a data assimilation approach – it is specialized and a bit awkward to implement, and restrictive in its assumptions. Again, that said, it does the job and provides for a proof of concept for the central notion of time-varying ecological parameters. I think the same approach could be done with a more modern variational or ensemble approach and this is the direction that I would encourage future work to take. The paper makes an important contribution to the field. It is well-written and the estimation method is solid. It needs a number of issues (see below) discussed. However, I feel it is well worthy of publication.

I offer more detailed comments below, which could be incorporated into a revised version.

1. There is a strong prior knowledge imposed on this problem, which plays a large part in explaining the success of the approach. The cost function, eqn (12), measures the deviations of the state and parameters from a reference values, and minimizes those by adjusting parameters in time. The reference used for the state are the observations, and so this part measures weighted squared discrepancy between observations and model predictions. The parameter part measures the weighted squared discrepancy between the parameters and the baseline values in Table 1. Parameters are estimated to minimize these (subject to satisfying the linearized model and satisfying the periodicity "constraint" on parameters). This setup is a powerful mechanism to preventing parameter drift since you effectively impose the annual mean value for each parameter, and estimate only seasonal anomalies. This also gets around the well-known parameter identifiability problem for ecological models by fixing the parameters in relatively narrow ranges, so paramters don't start trading off against one another to improve fit.

The paper should be very clear and up front about this aspect – it is what makes the approach work and why the results look so good. It seems like, in practice, the procedure requires good static parameters to start, and then it can get deviations (seasonal cycles) that better fit the data. These things should be made clear.

2. The linear optimal control method is not a state-of-the-art methodology. Such approaches were experimented with quite extensively in the early days of data assimilation. For example, the extended Kalman filter relies on such an assumption. Linearization approaches have been superseded by variational and ensemble methods that incorporate fully the nonlinearity in the forward model (the main use of linearization now – or the so-called tangent linear model – is to produce a gradient to aid in minimizing the cost function). The linear optimal control approach does works, but it is awkward to implement and I doubt many data assimilation practitioners would use it. There is nothing wrong with it, however, and I think it is suitable for illustrating the value of time dependent parameters, and flexible seasonal cycles for. I can see how the essence of the approach could be implemented using more modern data assimilation methods, and this is what I find interesting. Maybe some discussion on this is warranted.

3. The sensitivity of the results to changes in weighting "R" is interesting. This measures the extent to which parameters are allowed to vary, traded off against the fit of the model to the data (since Q is being fixed). This is, of course, is yet another parameter that can be tuned to help better fit the data. In Figure 5, it is shown that it does not make much difference in the state estimates. But it yields quite different values for the parameter evolution in subsequent Figures– their cycles are often similar in shape but their magnitudes and offsets are different. I think, ultimately, you want a more quantitative criteria for estimating R. There are two ways I have approached this in similar problems: through cross-validation, and by estimating the ratio of R to Q explicitly. Maybe mention?

4. Is there an issue of over-fitting here? When I see model predictions (for a model

with a lot of parameters) fitting data exactly (as in Figure 2), over-fitting is the first thing that would come to mind for any Statistician. You are estimating parameter anomalies at every time step. The matrix Ricatti equations can be cast as a sequence of linear regressions, where here the number of parameters likely exceeds number of observations. But at the same time things are tied together by the dependence structure in time so reducing the effective degrees of freedom. There is a lot of flexibility in the time varying parameters, and in general you are trading temporal smoothness off against fit via R. This is more of a discussion point, as an obvious issue when things (state estimates) look too good to be true.

5. Section 4.1 A better metric to assess the effectiveness of the seasonally varying parameters might be predictive skill, rather than just fit of the model state to the observations. The flexibility in the parameter values guarantees a good fit, but a more true test is: can you predict future states, or even retrospectively fill in withheld observations (i.e. cross-validation).

6. Section 2.1. I am a little unclear on the model – it states that is uses output from a basin scale model of the North Atlantic but then states it simulates the water column at one place (the BATS site, I presume). I think it is just a 1D location specific model. This should be clarified.

7. Theorem 1. Note sure if you should state this as a theorem here. It is a well-established result and should likely be presented as an clear step by step algorithm, to aid in implementation by interested readers.

8. Figure 3: How did the optimized static parameters compare with the Table 1 values? I presume the fit is better, since you likely used Table 1 parameters as starting values for the optimizer.

9. Figure 4: The model-data misfit looks strange and blocky. Why is this?

Interactive comment on Biogeosciences Discuss., 9, 10207, 2012.