Comment on gmd-2021-33
Anonymous Referee #2

Referee comment on "A Model-Independent Data Assimilation (MIDA) module and its applications in ecology" by Xin Huang et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-33-RC2, 2021

Summary

Huang et al created a model-independent data assimilation module and applied their module to different types of applications in ecology. This article is timely as data assimilation has become an urgent need but remains difficult for ecologists to implement. The details of implementation are extensive, but the benefits of the approach are somewhat repetitive throughout the manuscript. This type of tool is of great use to many types of modelers in ecology who don’t want to invest the time to add their model to a system. The details and video are very helpful for non-specialists in model calibration. I think there should be more papers like this so there can be more research using data assimilation techniques. However, I do have some concerns about MIDA which are described below. I think this paper should be published with revisions addressing Section 2.4.

General Comments

I have made several specific comments that are listed below. My three major comments are 1) that the statistical model used for model calibration in Section 2.4 "Step: Execution of data assimilation" is not well defined. It’s quite unclear to me what the authors are using for model calibration therefore it is difficult to evaluate the calibration exercises themselves. I’m not convinced that you are really showing posterior distributions in the figures because of the description of the algorithm in the methods. What is the actual statistical model used for model calibration? More details on my questions about this section are in my specific comments section 2) My second major comment is that Fer 2018 and PEcAn were entirely left out of this manuscript but should certainly be included in several places.

Specific Comments

Line 26: what kinds of states?

Line 38: I think there’s a word missing “model ... the land component”

Line 42: Doesn’t the easy implementation potentially make this more 'black box'?
Line 58: Citation for invasive coding?

Line 75: Missing link sentence between “, 2009). ... DA was”

Line 78: Not sure about “data-worth”

Line 100: Include Predictive Ecosystem Analyzer (PEcAn)

Line 120-121: Not sure what is meant by this sentence

Line 131: Nice!

Line 176: Hinders?

Line 178: How does MIDA know how to write out model specific configurations?

Line 183 transfers -> transfer

Line 184: organize -> organizes

Line 184: So the “different files” are like lookup tables?

Line 194: Would be cool to have an illustration of dynamic initialization *

Line 209: In think you mean “inference”

Line 210: Before talking about the sampling algorithm, it would be good for the reader to know about the model formulation like the likelihood, prior, etc.

Line 245: What kind of MLE? What’s the model formulation? What are you maximizing over here? Why do you get maximum likelihoods and posteriors?

Line 250: At this point, I’m still confused how you define your drivers and model settings in general? For a particular model you usually have a parameter file that gets read by the executable. So are you rewriting those parameter files in step 2?

Line 263: Make sure to cite all software packages.

Line 314: Helpful examples. Really curious where 302 years of leaf area data come from!

Line 343: “complex reasons” is somewhat vague

Line 499: Citation?

Line 504: Not sure about “first” unless you make your definition of model agnostic more specific

Figure 1: Very nice!

Figure 6: the Cis appear homogenous. Can MIRA deal with heteroskedasticity?

Figure 7: the colors in the legend have an extra character that should be removed

Figure 4, 5, and 8: these posterior distributions do not look converged to me. Could you also include the mcmc chains as well here or in the supplements?