Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-33-RC1, 2021

Comment on gmd-2021-33

Anonymous Referee #1

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-RC1, 2021

Review of "A model independent data assimilation (MIDA) module and its applications in ecology"

Authors: Huang et al.,

Summary

The authors identify the growing role data assimilation is playing in reducing parametrisation errors / uncertainty thus supporting data and/or model structural selection for terrestrial ecosystem models to improve predictive skill. The authors also identify an important obstacle to wider uptake is data assimilation frameworks being built around the requirements of a specific model, i.e. they are not easily applicable to alternate models. In response the authors have created a data assimilation framework which aims to be model-independent reducing the coding barrier.

Overall I don't have any objections to this paper being published more or less as is. The Github repository appears well organised with code that is commented to help people learn the framework. I would like a little more detail of is contained within the "black box" placed in the SI to give those who are interested this information.

Introduction

The introduction make a clear description of the state of data assimilation applications in ecology and identifies a key obstacle (i.e. substantial programming skill / investment in

time) to wider engagement within the ecological community.

Methods

A clear description of the process MIDA is given. I wonder about the trade-offs between the usually very fast exchange of information achieved when writing an interface vs the more user friendly approach described here? Not an objection to your approach but genuinely curious.

What isn't quite so clear is the details of how MIDA knows which information in the existing model output files corresponds to its observations. For example, the namelist.txt must contain information on the variable names used to describe the observations and their corresponding output variable generated by the model? These must still vary depending on the model being used? A screen shot showing the interface which is populated with an example would make this really clear.

On a similar point. The models need to be able to read the parameters from a file. The MIDA framework must then be able to write out the proposed parameters in a unique format for each model, is that correct?

L322-330: Could you add a link to further details in SI for this section? The reason I ask is that Haario et al., (2001) steps based on the weighted (e.g. beta) combination of the multivariate Gaussian and a minimum step size scaled by a value drawn from a Gaussian distribution of mean = 0, sd = 1. The multivariate Gaussian being derived from the covariance matrix for the parameters adjusted by an optimal scaling parameter (e.g. 2.38 / npars^0.5). The weighting between the two steps (beta ~0.05) and the minimum step size. So which of these variables (or something else entirely) for example is you "jump scaling"?

I like the inclusion of a screenshot of the software but I think it would be useful to have an example which has been filled in to help guide the potential user. Alternatively showing an example of the namelist.txt might be informative.

Applications with MIDA

Case 1: This doesn't really impact the validity of the paper but just something I noticed and wanted to raise as it should really be clarified. The DALEC model is stated as having 5 C pools but also to having a Growing Degree Days phenology model. However, the Williams et al., (2005) model doesn't have phenology model (i.e. continuous allocation / evergreen). DALEC was split into deciduous and evergreen versions in Fox et al., (2009)

as part of the reflex project adding a 6th pool and the GDD model. The example DALEC code provided on the MIDA Github shows a alternate version of the model where leaf C is not dependent on GPP (and thus the system is not mass balanced). This is a distraction from the main point of demonstrating your DA system. Please make the origin of the code clear as it doesn't match that found in the citations given. For example, the paper cited with the 5 pools Ricciuto et al., (2011) describes a model called LoTEC. Is this a DALEC derivative?

Case 3: It is really nice to have the comparison between multiple hypothesised model structures tested here. I think this is a really nice demonstration of the sort of simple experiments that can be achieved but are really useful.

Discussions

No comment, really happy.

Technical comments

L140: "DA is a statistical *approach...*" – there are many different algorithms for DA whether for state update or parameter estimation (in this case). I think it would be clearer refer to it as an "*approach*". I can see that you are trying to talk about your specific approach so maybe "The DA approach embeded within MIDA..."

L176: "*hinders*" or "*hides*"?

L454: "This model *simulates...*"