Geosci. Model Dev. Discuss., 8, C2533–C2538, 2015
www.geosci-model-dev-discuss.net/8/C2533/2015/

Geoscientific Model Development Discussions
Open Access

Interactive comment on "DasPy 1.0 – the Open Source Multivariate Land Data Assimilation Framework in combination with the Community Land Model 4.5" *by* X. Han et al.

Anonymous Referee #2

Received and published: 19 October 2015

DasPy 1.0 – the Open Source Multivariate Land Data Assimilation Framework in combination with the Community Land Model 4.5

Geoscientific Model Development Discussions, Han et al., 2015

The paper presents a publicly available source code package to assimilate observations into the CLM for state and/or parameter updating. (The source code was effectively available at the time of reviewing this paper, which was greatly appreciated.) While the platform is an applaudable development for this research group, it is oversold in this paper and it not anywhere as well tested and as general as other existing land data assimilation systems. Whereas the authors suggest that their framework would fill

a gap in existing land data assimilation systems, I see nothing that would outperform the existing systems or be more advanced than what exists already. Instead, there are a lot of hypothetical statements about what may be possible with the system without any proof or quantitative results. The data assimilation example in this paper is chosen rather poorly: it is fairly simple and illustrates that the assimilation "does nothing".

My suggestion would be that the authors rethink the paper, and

- give a more objective perspective on the current state-of-the-art of land data assimilation systems and the position of DasPy

- present either (A) a complex multi-source and multi-variate data assimilation example with a thorough step-by-step validation of the (1) ensemble errors in time and space, (2) data assimilation diagnostics (increments, innovations), and (3) independent validation with in situ observations, or (B) review and show all the tested possibilities of DasPy with quantitative results and highlight how they add value over other existing systems.

Detailed comments:

p.7397

L.5: observation data from multiple state variables → observation data related to multiple state variables

L.8: most reasonably working land data assimilation systems have a multi-variate updating (e.g. a simple case would be multiple layers of soil moisture or temperature) scheme. A multi-variate updating scheme is no novelty, but a standard approach.

Instead, a multi-source, multi-scale data assimilation framework (i.e. simultaneous assimilation of various observation types) would be a step ahead. Is DasPy able to deal with multi-scale data assimilation in its current version, i.e. assimilate observations at several resolutions simultaneously, downscale observations, etc.? Please add to the text.

p.7340

This is a mix of land data assimilation systems, atmospheric assimilation systems, large operational and small research systems. Where do the authors fit in here? It makes no sense to compare DasPy with something like the operational atmospheric system, nor with a synthetic research system like ATLAS. At the other hand, why are the relevant global land data assimilation systems from ECMWF and NASA missing?

My guess is that DasPy can be compared to e.g. DART or LIS, which are both freely available data assimilation systems and are used extensively for multiple * land * data assimilation purposes, often at limited domains (like DasPy). The question now is what is it that DasPy offers and DART or LIS does not have? In general, don't these systems have much more to offer than DasPy and should that not be rightly recognized in a publication?

Many of the systems listed between L.3-L.13 are highly effective large scale, multi-scale and multi-source, and multi-variate. I can only see the implementation of parameter updating as something that is tested in DasPy and maybe not fully tested in the existing systems (although, several have tested it, e.g. LIS and many research data assimilation systems).

p.7402

L2-L9: a lot is repeated in the previous page L1-L5, please consolidate

L26: "data assimilation is handled separately for each model grid cell".

- How does DasPy deal with overlapping areas of influence when spatial extrapolation is performed? Are domain halos implemented? (I only briefly screened the source code and did not immediately find anything that would account for overlapping update areas)

- What is the maximum total simulation domain area tested with DasPy?

- Can DasPy run on various grid layouts (e.g. cube sphere, EASE grid,...)? It looks like it is hardwired to run on a regular latitude-longitude grid only. Please mention that in the paper.

p.7403

There is a difference between correlation localization and simulating the spatial error correlations. The text on this page mixes both concepts.

L. 6: "spatial correlation characteristics of model state variables and observation data" → should be * error * correlations. This has nothing to do with error localization, this refers to limiting the spatial error correlations in the background error covariance matrix.

L. 15: The Gaspari-Cohn method is not used for spatial error characterization, but instead to localize the spatial error correlations.

L. 25: please explain all variables; I could not understand this. Eq.(1) has a perfectly identical value at the right hand side for each ensemble member.

p.7404

Is CLM effectively restarted with a new file after every update? What is the restart frequency? Is the restart frequency flexible for each type of observations?

What is meant by "The main disadvantage is the loss of computational efficiency, however, the binary NetCDF file format used by CLM could compensate for loss of computational performance."? How does the compensation happen?

p.7405

Are there any spatial or temporal error correlations in the current DasPy version, and if so, where exactly? L. 18 suggest that there are no spatio-temporal correlations, but p. 7403 is all about spatial error correlations and the localization thereof. p. 7406 hints to other data assimilation systems, not something that is implemented in DasPy. If there are no spatial correlations, then I do not understand what p. 7403 wants to convey.

p. 7406

CMEM = Community Microwave Emission Modelling * Platform *

p.7407

- L.24: replace “direct measurements of soil moisture” with “soil moisture retrievals”

- How does DasPy deal with biases in each of the various observation types?

- Can DasPy deal with spatially variable observation errors? Spatially correlated observation errors? Error correlations between various assimilated observation types? Please comment on observation errors.

p.7408

L. 5: please add references to support that DasPy can be “easily extended for. . ..” (i.e. quantitatively tested cases).

p.7410

- what is the difference between f(h) and h, or better, what is f(.)?

- what is the difference between lambda and lambda_vap. Please explain all variables.

p.7411

- Please illustrate how is the information in MODIS LST partitioned between Vcmax, Fdrai, Qdrai and all state variables? E.g. what are the update statistics (e.g. standard deviation in increments) to all individual components?

- Is MODIS LAI used as input to CLM, and also updated? How does this work? Does DasPy overwrite the MODIS LAI input with new/updated files as CLM is restarted? Is LAI perturbed as described in section 2.4?

- If MODIS LAI underestimates the true LAI, was any bias estimation turned on in DasPy?

Section 3.4

- Please some measure of statistical significance to the results?

- Fig. 7-11: the connector lines have no physical meaning. Perhaps switch the x-axis to show the various validation sites and show the various experiments as different bars or symbols for each experiment?

Interactive comment on Geosci. Model Dev. Discuss., 8, 7395, 2015.