LaVEnDAR response to reviewers

We thank the reviewers for their careful attention to detail on this manuscript. Their comments have undoubtedly helped strengthen the paper.

RC2:

The reviewer commented that it was not clear if this was the first application of LaVEnDAR or not and queried what would be needed to apply this technique to other problems.

This is indeed the first application of LaVEnDAR. We have added text to the abstract and introduction to make this clear page 1 line 4

“In this paper we present the first application of LaVEnDAR, implementing the framework with the JULES land surface model.”

and page 2 line 33.

“In this paper we present the first application of the Land Variational Ensemble Data Assimilation fRamework (LaVEnDAR) for implementing the hybrid technique of Four-Dimensional Ensemble Variational Data Assimilation (4DEnVar) with land surface models.”

We have also added a new section on the implementation of LaVEnDAR including which modules would need to be changed for application to another problem on page 8 and line 5.

“In order to implement 4DEnVar we construct an ensemble of parameter vectors and then run the process model for each unique parameter vector over some predetermined time window. We then extract the ensemble of model-predicted observations from the ensemble of model runs and compare these with the observations to be assimilated over the given time window. In our code (Pinnington, 2019) we implement the method of 4DEnVar with JULES using a set of Python modules. The data assimilation routines and minimization are included in fourdenvar.py. This part of the code does not need to be modified to be used with a new model. Model specific routines for running JULES are found in jules.py and run_jules.py. JULES is written in FORTRAN with its parameters being set by FORTRAN namelist (NML) files; jules.py and run_jules.py operate on these NML files updating the parameters chosen for optimisation. The data assimilation experiment is setup in experiment_setup.py with variables set for output directories, model parameters, ensemble size and functions to extract observations for assimilation. The module run_experiment.py runs the ensemble of models and executes the experiment as defined by experiment_setup.py. Some experiment specific plotting routines are also included in plot.py. More information and a tutorial can be found at https://github.com/pyearthsci/lavendar.

To use another model in this framework new wrappers would have to be written to mimic the functionality of jules.py and run_jules.py and allow for multiple model runs to be conducted while varying parameters. The module run_experiment.py would need to be updated to account for these new wrappers and functions to extract the observations for assimilation included in experiment_setup.py. Although we have used Python here to implement a stand-alone setup of LaVEnDAR we envisage that the technique could be added to existing workflow systems such as Cylc (Oliver et al., 2019) or the Predictive Ecosystem Analyzer (PEcAn) (LeBauer et al., 2013)."

P1L15-17: The reviewer pointed out that both land surface and atmospheric models are deterministic.
We agree that our description of land surface and atmospheric models here is incorrect and have updated the text accordingly at page 1 and line 15.

"Most land surface models will converge to a steady state; their state vector tends toward an equilibrium defined by forcing variables (i.e. the meteorology experienced by the model) and the model parameters. This is quite unlike fluid dynamics models used for the atmosphere and oceans, which exhibit chaotic behaviour; a small change in their initial state can lead to large deviations in the state vector evolution with time."

P1L19: The reviewer pointed out a typo and thought we were overstating the problem of parameter estimation.
We have corrected the typo and moderated our statement of the problem on page 1 and line 18.
"Consequently, for some land surface applications parameter estimation can have greater utility than state estimation. This manuscript deals primarily with the problem of parameter estimation in land surface models, although the technique we introduce could easily be used to for state estimation problems too."

P2L9: The reviewer suggested that allowing parameters to change in time was a way of accounting for model structural inadequacies.
We have modified the text to reflect this at page 2 and line 8.

"However, this is not true for land surface models where parameters are much less well understood. Indeed these parameters can be allowed to change over time within a developing ecosystem or when an ecosystem is subject to a disturbance event to account for model structural inadequacies."

P2L12: The reviewer thought it was worth mentioning emulator methods here also.
We have added comment on these methods as requested at page 2 line 22

"There is also a growing interest in model emulation, (Gómez-Dans et al., 2016; Fer et al., 2018), these techniques are extremely efficient but require some initial construction of the emulator."

P2L14: non-Gaussianity not a word maybe "non-Gaussian error" instead?
Updated.

P2L21: I'm surprised the paper is adopting the position that parameters should be static in time after arguing just 12 lines ago that parameters change over time.

Our intention was to argue *against* time varying parameters and we obviously did not succeed in that very clearly as Reviewer 1 also commented on this. As described in our response to R#1 comment 1 we have deleted the text around this as it caused confusion and, ultimately, did not motivate the development of the DA tool we have presented.

P3L31: The reviewer commented that GPP is not an observation and using this data in the assimilation should be treated with extreme caution.
We agree with this and have added text at page 4 line 5 to add caution.

"It is important to note that GPP is not an observation *per se* and is derived by partitioning the net carbon flux using a model which is likely to be inconsistent with the process model we are assimilating the data into."

P4L14-15: The reviewer thought our notation was confusing here and suggest we change the i subscript to a t. They also asked if we need a subscript on the model, *f,* and if this represented the model changing in time.
We agree that just using a t subscript may make things clearer for the reader, we have made this change throughout the manuscript. The subscript on f is not representing the model changing in time but repeated applications of the model to update the state to the desired time step. This then forms the basis for the matrix notation in equation (13).

P4L25: The reviewer asked if this structure would change the time invariance on p when accounting for process error.
It is possible to set up the assimilation system to include process error, but it has not been done in this case. Equation 5 deals only with the formation of the augmented state-vector. The reviewer is correct however that this point in the system is where we would add the process error if required. This would result in variation in p with time but it would be possible to prescribe a small variance in the process error to keep the change in p minimal, if this was the desired behaviour. We have added text to the paper discussing this at page 5 line 3.

"Process error could be included in equation (5) by specifying an additional term, but in this application is neglected."

P6 L6: The reviewer commented that more detail would be beneficial here.
We agree extra description is helpful here and have included this at page 6 line 11.

"For certain applications the prior error covariance matrix **B** can become large, ill-conditioned and difficult to invert. As a result minimising the cost function in equation (11) and finding the optimised model state/parameters can be slow. To ensure the 4DVar cost function converges as efficiently as possible and to avoid the explicit computation of the matrix **B** the problem is often preconditioned using a control variable transform (Bannister, 2016). We define the preconditioning matrix **U** by,"

P7L17: Here you say the adjoint is still present, but this is the first mention of an adjoint in the Methods. Needs further explanation.
We have added description of the adjoint earlier in the methods section page 5 line 24.

“$\mathbf{M}^T_{t,0}$ is the model adjoint propagating the state backward in time (this is required for efficient minimisation of the cost function using gradient descent techniques).”

Figure 1: The reviewer thought this figure was not helpful.
We have removed this figure.

P9L6: The reviewer wanted to know why we picked the parameters we did for the experiments and if there was uncertainty analysis conducted that attributed model uncertainty to these specific parameters.
We have included a sentence on this in the text page 9 line 13. See also response to comment P9L11/P10/L2-4.

“These seven parameters have an effect on the crop's seasonal growth cycle and its photosynthetic response to meteorological forcing data. The choice of parameters was motivated by the analysis of Williams et al. (2017) who found that they were least able to constrain these parameters with the available data”
Williams, K., Gornall, J., Harper, A., Wiltshire, A., Hemming, D., Quaife, T., Arkebauer, T., and Scoby, D.: Evaluation of JULES-crop performance against site observations of irrigated maize from Mead, Nebraska, Geosci. Model Dev., 10, 1291-1320, https://doi.org/10.5194/gmd-10-1291-2017, 2017.

P9L11/P10L2-4: The reviewer asked what the reasoning was behind the choice of parameter variance values for both twin and Mead experiments and that the priors and observation errors be given some justification.
We agree that a more rigorous approach could be taken to assigning the parameter uncertainties. As the analysis of Williams et al. (2017) showed all parameters to be poorly constrained with available data in a more traditional model calibration study we applied a blanket variance to all parameters. Reviewer 1 also asked for justification of observation errors. We have included extra text on this at page 10 line 13.

“We apply the same variance to all parameters here as the analysis of Williams et al. (2017) showed these parameters to all be poorly constrained with the available data in a more traditional model calibration study. In reality it is unlikely that all parameters will have the same variance but in the absence of additional information and for the purposes of this demonstration we used $(0.25\times x_b)^2$ […] We prescribe a 5% standard deviation for canopy height and leaf area index errors and a 10% standard deviation for errors in GPP. These uncertainties are rough estimates that we considered adequate for demonstrating our system, but for any specific application the errors estimates should be determined more carefully. However, our uncertainties are consistent with Schaefer et al. (2012) who found an uncertainty of 1.04 g C m-2 day-1 to 4.15 g C m-2 day-1 (scaling with flux magnitude) for estimates of GPP, Raj et al. (2016) who found an uncertainty in the order of 10% for daily estimates of GPP and Guindin-Garcia et al. (2012) who found a standard error of 0.15 m2 m-2 for destructively sampled green LAI at the Mead flux site.”

P9 L12: The reviewer stated that the observational noise was much too low in the twin experiments and that it would be beneficial to repeat the twin experiment with large uncertainties.
Reviewer 1 had a similar comment that we should repeat the twin experiments using the same error statistics as in the real-world Mead experiment, for which we replied:
The purpose of the twin-experiments is to demonstrate that we can retrieve correct parameters when we have high confidence on the observations and priors. When observations and priors are less certain (as is the case in the real-world experiment) retrieving the "true" parameter values is not guaranteed and hence is a less clear test of the data assimilation system. However, we agree that using the same uncertainties can also be informative and have now included the suggested experiment in supplementary material and added reference in the main text at page 10 line 7.

"We also include a twin experiment using the same error statistics as those used for the real data experiments at the Mead site (outlined in section 2.4.2) in supplementary material section S1.1."

P19L8: The reviewer again queried the choice of parameters asking why fd was selected if model outputs were not sensitive to it.
See comment P9L6. Although we have not been able to recover this parameter we believe this is a good example of an instance where unobservable parameter/state members become a problem.

P20 L4-14: The reviewer asked us to discuss the other sources of uncertainty not included in this study and how these might be included in the 4DEnVar framework.
We have included a discussion of the other sources of error noted by the reviewer and how these might be included in the DA system at page 20 line 16.

"In this study we have only considered the uncertainty in the parameters and initial conditions and not the uncertainty in forcing data, random effects (parameter variability) or uncertainty in the process model (Dietze, 2017). The inclusion of these additional sources of error would avoid the ensemble converging too tightly around any given value. In order to include uncertainty in the forcing data it would be necessary to run each ensemble member with a different realisation of the driving meteorology. Process error could be included in equation (5) resulting in a new term in the 4DEnVar cost function in equation (24) containing a model error covariance matrix, it has also been shown that these different types of uncertainty could be built into the observation error covariance matrix R (Howes et al., 2017). If estimates to these sources of error are not available the use of methods such as ensemble inflation (Anderson and Anderson, 1999), a set of techniques where the ensemble spread is artificially inflated, will help alleviate problems of ensemble convergence."

P20L8: The reviewer asked us to include posterior covariances (or correlations) as a supplement to this discussion.
We have included the posterior correlation matrix in supplementary material and referenced this at page 20 line 11.

"From table 3 we can see this issue for the two parameters controlling photosynthetic response with the posterior slightly over-predicting α and under-predicting *neff*, as different combinations of these parameters can produce the same trajectory for the observed target variables. The effect of equifinality can be seen more clearly for the posterior ensemble correlation matrix included in Figure S7 of the supplementary material."

P20L13: The reviewer suggested we had not explained ensemble inflation adequately.
We have expanded on this in the text, see comment P20L4-14.

P21L3: The reviewer asked us to include a description of localization or drop it.
We have included a brief description of localization at page 21 line 16

"Methods of ensemble localisation (Hamill et al., 2001), where distant correlations or ensemble members are down-weighted or removed, could be used to improve prior estimates."

P21L17: "[...] parameter estimation as it is often more important [...]". The reviewer commented that the matter is not settled.
We have removed this sentence.

P21L22: (i) The reviewer noted that we should not equate process error with stochastic noise or inflation (ii) The reviewer asked for more discussion of how iteration would work in LaVEnDAR and that we consider adding an additional year for validation.
(i) We have removed discussion of stochastic noise and inflation here. (ii) We have added more discussion on the iteration at page 22 line 6

"This would require additional modules to be written within LaVEnDAR which would handle the starting and stopping of the process model. It would also require that the implemented model was able to dump the full existing model state and then be restarted with an updated version of this state (as is possible in JULES). In this iterative framework accounting for model error would also become more important."

and also included an additional year of validation (i.e. a hindcast) in supplementary material, where we have run our posterior ensemble for 2008 across 2009. We have added a reference to this at page 20 and line 34

"By conducting a hindcast for 2009 (shown in supplementary material Figure S6 and table S2) we also find the retrieved posterior ensemble improves the fit to the unassimilated observations in the subsequent year, with an average reduction in RMSE of 54% when compared with the prior estimate."