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.

RC1:

"Their assertion that an additional benefit of the approach is that it identifies "true" parameters [...]". The reviewer suggested that our assertion in the introduction that an additional benefit of the technique is that it identifies "true" parameters which are static in time was tenuous and superfluous

Our brief comment on time varying parameters was perhaps not well thought through - and this was picked up by both reviewers. Our intention was to distinguish between the result we would have obtained using a filter rather than a variational technique. However, we agree with R#1 that this was superfluous. To improve the clarity of the manuscript we have removed these sentences.

"[...] it would be very good to have a small amount of addition information as to how the system is actually implemented". The reviewers main request was for the addition of information on the implementation of the technique

We agree with the reviewer that this was an obvious omission in the manuscript and have added a new section describing the implementation in JULES and what would be required for a more general implementation in other models on page 8 at 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)."

P9 L12: What does 2% Gaussian noise mean?
We have reworded this at page 10 line 4.

"Perturbed using Gaussian noise with a standard deviation of 2% of the synthetic truth value."

P10L2 The reviewer asked why we chose to have lower uncertainties in the twin vs. the real experiment and thought it might be more informative to use the same uncertainties for the twin as in the real experiment.
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."

P10L3 The reviewer commented that our justification of observational errors, in particular GPP, was insufficient
We agree that our justification of errors in GPP data is simplistic, and for the purpose of making robust scientific inference a greater level of effort would be required to quantify the uncertainty in the flux data. We have added references to the text demonstrating that a 10% error on GPP could be a reasonable choice on page 10 line 17.

"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."

P10L17/P10L20/P20L28 The reviewer asked for the inclusion of a description of how model harvest date and harvestable material are calculated
We have now included a brief description of how harvest date and harvestable material is calculated in the model and cited a paper that provides details of the complete algorithm at page 3 line 25.

"Crop development is governed by a crop development index which increases as a function of crop-specific thermal time parameters with the crop being harvested when the development index crosses certain thresholds. The crop grows by accumulating daily NPP and partitioning this between a set of carbon pools (havestable material, leaf, root, stem and reserve), equations for JULES-crop can be found in Williams et al. (2017) appendix A1."

P11L3 The reviewer commented that some of the parameter priors (particularly mu) did not seem to be very normally distributed.
Here we are only sampling 50 ensemble members for the prior and then fitting a curve to the parameter histogram. We are also sampling a bounded distribution to ensure none of the sampled prior parameters are below zero (except for delta which is negative). Due to the small number of samples and the boundedness the prior distributions sometimes don't appear very Gaussian.

P15L7 Reviewer asked what was meant by "only capturing 5 of the 11 observations"
Here we were referring to how many observations the ensemble mean passed through, rather than how many observations were captured by the ensemble spread. We agree this was not at all clear and have updated the wording accordingly on page 15 line 6.

"From Figure 7 we can see that the prior mean underestimates LAI, reaching a much lower peak than observations, despite this the technique finds a posterior mean estimate that agrees well with all but 2 LAI observations (in September and October)."

P15L15 If LAI agrees with observations, but leaf C does not, this implies SLA is incorrect, but this is one of the parameters being optimized, or at least a coefficient controlling it? What is the suggestion of this for the model?
It is likely that the optimized parameters controlling SLA are compensating for error in the parameters controlling the partitioning of NPP into the leaf carbon pool. This allows us to achieve the correct leaf area with the incorrect leaf carbon content. We have added discussion on this at page 15 line 16.

"The fact that we can find good agreement for LAI with a poorer fit to leaf carbon content is likely due to the optimised parameters controlling specific leaf area compensating for errors in model parameters controlling the partitioning of net primary productivity into the leaf carbon pool. This allows us to achieve the correct leaf area but with the incorrect leaf carbon content."

P20L2 How will the correlations in the prior error covariance matrix be determined/ estimated?
For the purposes of testing the 4DEnVar technique we did not consider error covariances in our prior information. We have published on this topic previously and include a brief discussion of this at page 20 line 3 in the current manuscript.

"Alternatively including correlations in the prior error covariance matrix would provide information to update *fd* even when the assimilated observations are not impacted by changes in this parameter. It has been shown that suitable correlations can be diagnosed by

sampling from a set of predetermined ecological dynamical constraints and taking the covariance of an ensemble run forward over a set time window (Pinnington et al. 2016).”

P20L4 To what extent is this ensemble collapse a function of (over optimistic) observation error?
Ensemble collapse here was a poor choice of phrasing - we did not mean it in the traditional sense used in data assimilation. We meant only that the ensemble had converged around a particular value rather than “collapsed” which is normally used to indicate all ensembles occupying the same space. We have modified the text at page 20 line 6.

P21L4 A brief discussion of the steps required to extend this framework to models running on spatial grid regionally/globally in addition to a need for localization would be very beneficial, including any potential limitations.
We have added this discussion on page 21 line 18

“In order to extend this framework to model runs over a spatial grid we will need a method to sample prior parameter distributions regionally or globally, it would then be possible to conduct parameter estimation experiments over a region, either on a point by point basis or for the whole area at once. Considering a large area would increase the parameter space and require more ensemble members. Localisation in space could help to reduce the number of ensemble members required.”

P21L9 The reviewer asked how this framework would be used in a cycling system
The use of a cycling system is probably more appropriate to state assimilation rather parameter estimation. Once parameter estimation had been conducted the framework could be set up for state estimation and then cycled on a timescale suitable for the desired target variable and data availability. We have elaborated on this in the text page 21 line 27.

“While posterior parameter estimates could be used in future studies with their associated uncertainties we envisage that cycling of the assimilation system will be more appropriate for state estimation (after initial parameter estimation) where the system could be cycled on a timescale suitable for the required state variable and data availability.”

P21L20
Typo sentence modified.