Review 2

This study explores the use of history matching and the not-ruled-out-yet (NROY) parameter space to calibrate land surface model parameters and quantify uncertainty on model outputs. The method is demonstrated in a twin experiment with the ORCHIDEE model (i.e. using model-generated "observations" with known parameters), and compared to parameter optimisation and uncertainty characterisation by a gradient-based method and a global-search method.

The paper is very clearly written, and provides a valuable demonstration of the history matching / NROY method. The comparison with the gradient-based and global search methods is an important part of the paper, as these methods have often been used to optimise parameters and characterise model uncertainty in land surface models. The comparison of the uncertainty range from the ensemble of 200 optimisations with Bpost and with the HM range is also good to see. I like the exploration of different metrics. The results are significant and I recommend publication of the paper with minor revision to address the comments below.

We would like to thank the reviewer for taking the time to read and comment on the manuscript. They make some interesting points, which we have incorporated into the manuscript as best we can.

Specific comments:

Line 11 - "the true parameters are contained in the posterior distribution" - is this guaranteed with history matching, or do you find that it occurs in this example and could there be cases where it doesn't?
The fact that it contains true parameters shows how well HM works. This sentence can fail to be true if the emulators are bad/biased leading one to rule out good parts of space because the emulators are not being a suitable (probabilistic) representation of the model. Furthermore, misspecification of errors can also lead to ruling out the whole space: if Var[f(x)] -> 0 and one had set e.g. Var[e] and Var[eta] too small. However, it is possible to revisit these tolerances to keep something, which should then retain the 'truth' - but only if the emulator is good enough.

Line 73 - VarDA (and the term variational) is used here to include both the gradient-based method and the genetic algorithm. I am used to using the term 'variational' to describe gradient-based methods, in contrast to terms like Monte Carlo, global search, or stochastic to describe a genetic algorithm. I wasn't able to find a definitive definition of variational, and it seems to be used differently by different authors (e.g. Santos et al. (2013, doi: 10.1590/S2179-84512013005000012) contrasting variational methods and genetic algorithms, and Schmehl et al (2011, doi: 10.1007/s00024-011-0385-0) describing a genetic algorithm variational approach). Nonetheless, I think it is worth considering whether a different term would be better to describe the two parameter optimisation methods (e.g. ParOpt for Parameter Optimisation, or ParEst or PE for Parameter Estimation) to avoid any possible confusion.

This is a tricky point since, as the reviewer points out, in the literature, variational can both be interpreted as “minimising the cost using gradient-based methods” and simply “minimising the cost function” (regardless of technique). In all cases, the main focus of the literature is usually on constructing the cost function, and therefore variational is commonly used to refer to the type of cost function used more than the method used in the minimisation. While reviewers suggestions for alternative terms have merit, we feel they are a) too vague, especially given the fact we want optimisation with uncertainties (as described by the 4D-Var equation) and b) one could argue that history matching could fit into the realm of of parameter estimation as the goal is still to find viable parameter sets by ruling out the unlikely ones. We have decided the keep the term “VarDA” but with more transparency on L128 (in accordance with RC1), which hopefully will ease confusion:

> **Note that here we use the term variational to describe the form of the cost function minimised. While the classical approach to minimise this function relies on gradient-based methods, in the absence of gradient information, other methods have increasingly been used to find the optimum. This has led perhaps to an abusive use of term “variational”, however, we feel here it helps to group, via a common cost function, the two minimisation approaches we wish to compare to the history matching approach.**

Line 156 - this equation assumes σi is the same error for all observations in each stream

That is correct, the statement has been added to the text.

Line 165 - define E

We have added the following to line 167:

> **where E is the expectation and Var is the variance.**

Figure 1 - I don't understand the text between the first pink shape and the first purple shape "For 1st wave χ=χNROY". I would understand it if it defined the first wave χNROY=χ.

Apologies for the confusion, the reviewer is correct the equation needs to be reversed. The figure has been amended as such.

Line 225 - write 10,000 rather than 1e4

Done

Line 313 - Q10 is the most constrained parameter *relative to the prior range*.

Added

Line 340 - could remind the reader here that 200 GA optimisations were used e.g. "Instead, multiple GA optimisations were preferable (we used 200), which is extremely costly."

Added

Figure 7 - Is Min/Max the quotient of the min and max of the data? Please define exactly what this metric is. What is the number beside the panel caption (i.e. 0.35 beside Min/Max, 0.06 beside Spring gradient etc)? Vertical gray lines could be added to the timeseries for b) at Feb and Apr and c) at Aug and Sep to point out the months used in the metrics. The constraint of initial carbon stocks, is that something that is often observed?

Line 383 - be consistent in using min/max or amplitude to describe that metric.

Amplitude on L383 changed to min/max

Line 383 - Do you need to weight the different metrics when they are combined in HM?

Not in the traditional sense. We have added the following to L385:

> **Note that these metrics are not weighted (in the traditional sense) when combined. Instead, the weighting occurs through the individual errors used to set up the experiments.**

Is there noise added to the observations used for the alternative metrics? In a twin experiment, without noise on the observations, it is easy to see how the parameters could be much better constrained than in a realistic case with actual observations.

Yes, the noise on the observation (described in L200) is applied throughout.

Line 459 - In the context of land surface models, a stepwise approach to separate calibration of the fast and slow processes (as described at line 85) would benefit from this feature of the HM.

Good point, we have added the following to the text:

> Furthermore, HM’s iterative nature means we can add different data when and as they become available. **It also lends itself well to a stepwise approach to calibration, allowing us to separately constrain, for example, the fast and slow processes of the model.**

In contrast with the RC1 comment, I do believe that the VarDA part of the paper is important, as it reflects the way parameters are often optimised and uncertainties quantified in land surface models. Personally, I like the style of discussing the meaning of some of the results given in the Results section, rather than leaving all of that discussion to the Discussion section, but I guess that is a matter of style.

Thank you for saying this, we agree that both approaches have their merits. Here we have decided to keep some of the discussion with the results, leaving space in the “Discussion” section to consider the pros and cons of both techniques more broadly.