Reviewer #2

- *Main points: The perfect model and observations may not be sufficient in supporting the conclusions reached in this manuscript. Even though the reviewer agrees with the authors 4DVAR has its potential in oceanic state estimation, their case is simply too perfect for convincing the readers, particularly those from the non-variational analysis community, like EnKF or 4DEnVar. The reviewer suggests more realistic experiments and recommend this manuscript for major revision.*

Thanks to the reviewer for the interesting and helpful perspective. The experimental setup may not be as idealized as originally thought. For example, the synthetic observations generated in this study are not perfect, and we detail their generation in the response to the next point below.

Regarding the perfect model assumption, our equations have been formulated in analogy to the ECCO (Estimating the Climate and Circulation of the Ocean) state estimation equations (*Stammer et al.*, 2002), as our goal is to develop diagnostics regarding when and why the Lagrange multiplier method succeeds or fails. In the ECCO formulation, the ocean model equations are typically treated as perfect in the ocean interior, and errors are permitted in the surface gridcells with air-sea forcing. Here, we permit these errors and attribute them to errors in the external forcing.

One aspect of the analysis that we have improved is the first-guess of the forcing field. In a case study where the first-guess of the forcing is zero, the results are similar to the original case. This is reported in a new Section 3.5 and a new Figure 8. The new section follows.

> The previous examples in Section 3 proceed with prior information that the forcing is periodic with an accurate magnitude and phase. A good analogy is the regular forcing of solar insolation on the ocean surface. Here, we test the performance of the Lagrange multiplier method with inaccurate prior information about the forcing, as is a more realistic analogy to the uncertainty of air-sea fluxes. In particular, our first guess of the forcing, $f_0(t)$, is systematically biased by decreasing b from 1.5 to 0.75 rad s^{-2}. The trajectory driven by inaccurate forcing is no worse than the previous cases with accurate forcing due to the dominance of the chaotic dynamics of system (Figure 8). Using the same observations as shown in Figure 3, we find that the chaotic pendulum trajectory is tracked over multiple nonlinear timescales despite this more stringent test. In this case, however, the forcing estimate still contains errors relative to the true forcing calculated with $b = 1.5$ rad s^2, and some high-frequency structures remain in $f(t)$ (see "improved first guess" in bottom panel, Figure 8). If instead the Lagrange multiplier method is started from the standard first guess, a smoother and more accurate estimate of the forcing is obtained at the expense of not fitting the data as well (see "final estimate" in bottom panel). Any remaining irregular structures can be handled by imposing temporal correlations as was done in Section 3.4. If such measures are not taken, the investigator must take care to decide what elements of the forcing represent true variability and which are compensating for model error. In our simple system of equations, model errors and forcing errors are mathematically equivalent. In state estimates with eddy-resolving GCMs, however, smallscale forcing variability is found near oceanic fronts and the investigator must determine on a case-by-case basis to what extent it reflects real variability.

Convincing the non-variational analysis community to adopt the Lagrange multiplier method would be a challenging task, but outside the goals of our work. It is clear that the motivation and goals of the manuscript need to be made more explicit. The Lagrange multiplier method is popular in oceanography due to automatic adjoint model compilers and strategies to reduce computer memory consumption. Much time and effort has been spent to develop this technique in real-world scenarios, yet it is unclear whether this method should be applied to eddy-resolving models and how long the time window should be. For the Lagrange multiplier method to be successful in state-of-the-art ocean models, two major issues need to be addressed: (1) the high dimensionality of the forward model and estimation problem, and (2) the nonlinearity of ocean models at increasingly fine resolution. Issue

(1) has been overcome by groups such as the ECCO Consortium. Here we focus on (2). It is true that issues may arise by the combined effect of (1) and (2), but first we attempt to isolate the effect of nonlinearity.

With this problem in mind, it is logical to find a numerical model that can be thoroughly understood and one that is highly nonlinear. It is not the goal of this manuscript to use a state-of-the-art numerical model. We believe that these expectations should be set at the outset, so we include the following in the Introduction.

> Because the effect of nonlinearity is seen as the major roadblock for application of the Lagrange multipler method, we isolate this effect by choosing a model that is highly nonlinear but low-dimensional: the forced, chaotic pendulum (Section 2). Toy models are worth revisiting because the dynamics are comparatively simple to understand, and they have strongly influenced when the Lagrange multiplier method has been deployed to realistic ocean problems. We will show that previous toy models have sometimes been misinterpreted.

We now also emphasize upfront that the development of a new state-of-the-art data assimilation technique is not the goal of this work, either. Instead, we wish to evaluate the current use of the Lagrange multiplier method. Now, the Introduction makes this explicit.

> Rather than developing a new state-of-the-art data assimilation technique, we proceed by taking the existing Lagrange multipler method and developing diagnostics regarding when and why it succeeds or fails, as evaluated by the ability to fit observations. Relative to the initialization problem, the prospects for a successful state estimate are shown to be improved in the boundary control problem, even if one uses a highly nonlinear model such as the forced, chaotic pendulum (Section 3).

Thus, a practical goal of this work is to convince those groups that already use the Lagrange multiplier method to reconsider the range of scenarios in which they apply the method.

- *1. What are the values of ? From the true solution? I am afraid if ideal observations are used, it does not imply the conclusions made for this ideal model to be useful.*

  Stochastic noise is used to generate synthetic data, mimicking the imperfection of ocean observations. This is a common approach that has appeared in ocean state estimation studies such as Tziperman et al. (1992), who used the same iterative adjoint method on a simplified ocean GCM (the momentum equations are balanced and the nonlinear advection is neglected). The manuscript states the following.

  > We consider an "identical twin" experiment where the true solution is known (solid line, Figure 1), and we observe the pendulum angle episodically through time with normally-distributed random errors of standard deviation, $\sigma_\theta = 0.5$ rad. In most oceanographically-relevant cases, observations have already been collected over some fixed time interval ($0 \leq t \leq T$). Here, observations, $y(t)$, are taken at a set of N_y evenly-spaced times with an time interval of $\Delta t_y = T/(N_y - 1)$.

- *2. Page 4, line 23, The statement of . . .the quantity inside curl brackets vanishes is not generally true. To do so, there needs an additional term, penalizing the constraint.*

  The terms in the curly bracket vanish by definition of our time-stepping model in equation (2). There was an error in defining the external forcing term which may have caused confusion. The equation is deterministic and is now stated explicitly in equation (1). The weight matrix S_f in equation (5) is selected in the state estimation process to limit the difference of the improved guess from the first guess, similar to the formulation in Bennett (2002). The revised text reads as follows.

  > The motion of the forced pendulum is described by the deterministic equation (*Baker and*

Gollub, 1990),

$$\frac{d^2\theta}{dt^2} + \frac{1}{q}\frac{d\theta}{dt} + \frac{g}{l}\sin\theta = f(t), \tag{3}$$

where θ is the displacement angle from vertical, q is a damping coefficient, g is gravitational acceleration, l is the pendulum length, and $f(t)$ is an external forcing term. In turn, the external forcing has a first guess and a perturbation, $f(t) = f_0 + \delta f(t)$, where the first-guess is set to periodic forcing, $f_0(t) = b\cos(\omega_d t)$.

- *3. Page 5, line 23. That is where the problem is that such as observation and prior knowledge and freely-running forward model are not enough.*

  We agree with the reviewer that the first-guess may not always be sufficient to track a chaotic system. For this reason, we implement a $\chi - 2$ test, detailed in the next point-by-point response, that diagnoses the likelihood of success or failure.

- *4. For solving a global minimization problem of (6), the first guess is crucial as the authors stated in page 4 line 25-26. However, the improved initial guess of their work presented in Section 2.4 cannot guarantee the initial guess is good enough for converging to the global minimizer. The authors should at least present convincing arguments of why they believe their improved initial guess could reach their goal. To the reviewer, the improved initial guess may fall into the same valley as the original initial guess.*

  The reviewer correctly states that there is no guarantee that a solution will be a global minimizer. We discuss the underdetermined nature of the problem in Sec. 3.3, where we acknowledge that the solution is not unique. Instead, we focus on finding any acceptable fit. We view the problem as having two clear steps. It is a first step to find any solution that acceptably fits the data. Only then can we proceed to investigate the uniqueness of the solution. In real-world situations, the first step may be the only one that can actually be evaluated.

  To consider whether a solution is an acceptable fit, we include Figure 6 which details the size of the cost function for various numbers of controls and observations. By implementing a χ^2 posterior statistical test, we determine the ratio of success to failure for various parameter ranges. After running many trials, we do not guarantee the results for any particular number of observations and controls, but a clear pattern emerges. We suggest that the pattern of Figure 6 is explained by the basic metrics of controllability and observability. rather than the stability of the system. This is one novel result we are reporting, and the previous works suggested by the reviewers have not already made this point, nor do they appear to contradict it.