We would like to thank both reviewers for their careful reading and thoughtful comments. Our responses are given inline as follows.

The authors have greatly improved the clarity, conciseness, and focus of the manuscript. It reads well and should be considered for publication with minor revisions at detailed below.

As a general technical comment, try to avoid 2-sentence paragraphs.

I think it would be a useful addition to the presentation to have the first figure of the manuscript as an example representation of the model domain, similar to Figure 10, but just showing a snapshot of an example flow and surface height field that is representative of the model's behavior under the given surface forcing.
We have added three representative plots: one at the initial time, one after 30m, and one after 30h. (Fig.1)

p.2, line 3:
I suggest to change the citation format from " (Abarbanel (2013))." to "(Abarbanel 2013)." for this and all following similarly-formatted citations for conciseness.
Thank you. We were not aware this was an option.

p.2, line 8:
This is a 2-sentence paragraph, the first of which is a run-on sentence. I suggest breaking the first sentence up to be more concise.
p.2, line 10:
Remove "In other words," since what follows is adding information rather than rephrasing the previous statement.
Paragraph now reads:
To clarify these ideas we refer to the observability study given by~\cite{mwr13}, which evaluated the performance of familiar nudging methods on chaotic, shallow water flow. The flow was simulated on a β-plane defined by a square grid with uniform spacing N_Δ, periodic boundary conditions, and driven by Ekman pumping. Poor predictions were obtained unless the height variable h and at least one of the two velocity variables u, v at each of the $N_\Delta \times N_\Delta$ grid points were measured. In other words, accurate forecasts required direct observation of roughly 70% of the $3N_\Delta^2$ dynamical variables.

p.2, line 23:

"These outcomes suggest that time delays may be useful for reducing the number of required observations to meet the practical constraints of operational NWP"

I suppose there is some degree of interpretation here as to what qualifies as a 'countable' observation. I would call a single instrument making 10 consecutive measurements at the same point 10 observations.

In that sense, the use of time-delay methods could make better use of a fixed number of observations that might currently be ignored.

What needs clarification is whether L refers to the number of distinct observation locations, or the total number of distinct measurements in a time window? Please make sure this is clear.
The former is correct, except that it is the number of observation locations times the number of observations made at that location, since for instance h,u,v are all measured at the same gridpoint. We have clarified this in the second paragraph of p1:
Here we consider an idealized situation where a perfect dynamical model describes the deterministic time evolution of a set of D state variables. We assume that L measurements are made at uniform time intervals Δt within an observation window of length T, so the total number of distinct measurements is $L\times(T/\Delta t + 1)$.
Also, to the second paragraph of section 2.
The total number of measurements in the observation window is $L\times(N+1)$.

p. 2, line 27:
Connect line 27 into the paragraph on the next line.
Done.

p.3, line 25:
Is capital Phi the state transition matrix in this context? Could you state its role in this case?
Yes, it is the linearized state transition matrix. We have modified the sentence to read:
where $\vec{\Phi}(t,t') = \partial\vec{x}(t)/\partial\vec{x}(t')$ is the linearized state transition matrix. Its time evolution is described by the variational equation

p. 4, line 29:
eq 4: Do you intend for tau to be negative so that adding increasing multiples of tau goes backwards in time? It may be more clear to assume tau is positive and subtract multiples of tau to illustrate the delay aspect of the approach. If not, you may want to make it clear immediately that your time 'delays' are actually forward in time given your indexing framework.
Yes, the 'delays' here are forward in time. We have modified this paragraph as follows:
...and τ is the delay, which here is assumed to be a positive integer multiple of Δt. Also, note that the term `delay' here is not used in its usual sense. Rather, this method involves an advanced formulation, which for positive τ incorporates observations at later times. Both formulations are acceptable however.

We also modified the discussion of the connections with 4DVar on p5-6.

p.4, eqns 4, 5:
Thanks, this is much clearer than the previous presentation.

p. 5, line 20:

It may be helpful to clarify the definition of 'delay' when the delay is first mentioned.
Agreed, see response above. P.4 line 29

p. 6, line 2:
Change "nudging [using] truncated" to "nudging [uses] truncated"
Fixed.

p.6, line 14:
This isn't a complete sentence: "Namely, the shallow water equations."
Fixed.
Instead, we focus its application to a core geophysical model: the shallow water equations.

p.6, line 18:
"data is taken" to "data are taken"
Fixed.

p. 7, line 24:
"by measuring the average growth rate of random perturbations."

I think a little more detail is needed here. Are you using a 'bred vector' type method? If so, the scales captured are determined by the magnitude of the random perturbations, the rescaling interval, and the norm used. These parameters impact the scales of the instabilities that are amplified and identified by the algorithm. Please provide more detail, as it is difficult to interpret this description.

For bred vectors, e.g., see Toth and Kalnay (1997)
http://journals.ametsoc.org/doi/full/10.1175/1520-0493(1997)125%3C3297%3AEFANAT%3E2.0.CO%3B2

The details of this calculation are given in our Whartenby 2013 paper.

> *... we selected two sets of very close initial conditions and integrated the two solutions of the equations forward in blocks of 4000 time steps of 0.01 h each. During each block of 40 h, we tracked the RMS distance between the two slightly perturbed solutions to the shallow-water equations and evaluated the largest Lyapunov exponents (LE) by approximating the growth of this error as e LE3t . Following the standard procedure in studies of nonlinear dynamics, when this error grew large, we rescaled the values of the two orbits to have a very small value of their distance and then tracked growth of the distance between the same two orbits for another block of 40 h in time. We did this for 80 blocks of time.*

A reference has been added in the text.
With these fixed parameters the shallow water flow is chaotic, and the largest Lyapunov exponent for this flow is estimated to be $\lambda_{max} = 0.0325/h \approx 1/31h$ by measuring the average growth rate of random perturbations. The details of this calculation are given by \citet{mwr}.

p. 8, table 1:
Could you clarify - is the spacing delta X and delta Y consistent across all resolution experiments {16,32,64}, or is the total domain size constant?
Total domain size is constant. Added clarification in p4 of section 4.
The total domain size is constant $800 \times 800~\text{km}$ and periodic boundary conditions are enforced.

p. 8, table 1:
The authors list the Coriolis parameter f0 and the Rossby parameter (meridional derivative). Please also list the corresponding latitude for this f value (e.g. at the center of the model domain).
Central latitude of the beta plane ~3.6 deg has been added to the table

p.9, lines 18-24:
Please mention that the results with D_M=7 reached high accuracy in some cases, dependent on initial conditions. These 'boundary' cases are interesting to identify for future study.
Added the following:
In addition, with the choice $D_M=7$ some initial conditions synchronized while others did not. Further analysis of this `boundary' case is an interesting area for future study.

p. 10, line 30:
Change "The same [striking] improvement" to "The same improvement"'
The use of extreme adjectives of this type does not enhance the presentation. The results speak for themselves and the readers can determine whether or not they find them 'striking'.
We agree, the adjective has been removed.

p. 10, line 25:
I'd like to see one case additional where the noise is large enough to break the synchronization.
Added the following:
Furthermore, results fail to synchronize when the magnitude of the noise gets large enough. This transition occurs at roughly $\sigma = \{1.3,2.0\}$ for $D_M = \{8,10\}$ respectively.

p. 11, line 10:
"The dynamics of drifters are described as two-dimensional fluid parcel motion on the surface of the water layer"
Technically, wouldn't the drifters modeled in these equations be representative of the vertically integrated layer velocity? This may be a better interpretation anyway, since in reality drifters are usually representative of a given depth. For example, the GDP drifters are drogued at 15m. Surface floating drifters may be adversely impacted by winds and thus not accurately represent the near surface currents.
We have changed our statement to read

...as two-dimensional fluid parcel motion near the surface of the water layer

p. 11, line 14:
“Hybrid measurements are incorporated into the time delay nudging method by combining the grid variables and the collective drifter positions”
Is it the drifter position you are appending, or is it really the drifter id and dimension label as a type of coordinate space for the drifter data? The position data should be analogous to the velocity and height measurements, while drifter id’s are analogous to the model grid points.
The 2D drifter positions are appended to the state be vector. They provide analogous velocity and height information in between gridpoints. But that information is not used directly, say by approximating the drifter velocity and interpolating it to the nearest gridpoint. Rather the drifter information is transferred to the estimate through its dynamical model dr/dt = {u(t),v(t)}. For the simulations, the u,v fields had to be interpolated from the true solution at the drifter positions. Whereas in reality, this data will just be provided.

Does this answer your question?

p. 11, line 16:
As a warning, the boldface capital R is typically used in DA to represent the observation error covariance matrix, and so the choice to use it in this context may cause unnecessary confusion.
Thanks for pointing this out. It has been changed to \xi.

p. 12, line 4-9:
I’d like to see one more case where the approach to forming the initial conditions matching the previous cases is used, and sufficient drifters are added to achieve synchronization. This would make it easier to compare the findings in this section with results from earlier sections.
We ran these cases, but the results are inconsistent and depend on where the drifters are initialized. As such we have decided to leave these results out of the paper. Obviously there is more work to be done here. However, we have modified two paragraphs in this section to make this more clear.
We consider cases with $N_D = 0$ (no drifters), as well as $N_D = 20$ and $N_D = 64$, in addition to $L = 208$ and $L = 128$ grid observations. Example plots showing the initial locations for $N_D = 20$ and $N_D=64$ are given in Figure 8. While many runs were successful using the same setup described above, the results were somewhat dependent on where the drifters were initialized. These results are not shown.
The consistency of the results improved by choosing the initial estimate to only in magnitude from the true solution, rather than in both both phase and frequency as was done above. Specifically, for the results reported below the initial conditions of the data ...

Page 12, line 21:
Please change:
“when the model is [wrong]” to “when the model is [imperfect]”
If the model is ‘wrong’ in practice then it cannot not be used for forecasting or data assimilation. It must

be reasonably accurate to produce any reliable forecast skill.
Fixed

page 12, line 22:
“this methodology provides some idea as to whether the model is at fault, or whether more observations are needed.”
I suppose I need more explanation for how the methodology provides this information. I understand that it indicates for a given model you can identify how many observations should be sufficient for synchronization. So are the authors extrapolating that then if the model does not synchronize when using real data then the model is to blame? There are subtleties in this process that make this statement seem like an oversimplification, particularly regarding the assimilation of satellite data.
The point we are trying to make is that, if synchronization occurs with simulated data but not with real data, the model may be the limiting factor. Whereas if it doesn't succeed with simulated data, given the constraints on observability, then the observations are likely the limiting factor. However our argument is muddled by the fact that (as you mention) the synchronization methods we've described here don't apply to situations where the measurement operator is nonlinear, such as for assimilation of satellite radiance data. We have changed this argument to read:
Since the successful analysis of simulated data a prerequisite for success with real data, this methodology provides a way to assess where one stands with respect to critical observability limits of the system at hand.

Figures 4,5,6:
It may just be my pdf rendering, but I see two plots on the top row but only one plot in the center of the bottom row. It appears that one is missing. Please double check to make sure there are reproduced as intended.
Yes, that was a `bug' in our latex code. Thank you for your thorough reading.

Figure 10:

It appears your domain is very close to the equator since there is no geostrophic-type flow due to the inclusion of the Coriolis term. Instead it appears you have a down-gradient flow, without the effects of rotation. In the future, I’d suggest focusing on a domain that is shifted more from the equator so there are significant contributions from both the geostrophic and ageostrophic components in your flow field.
Thanks for your feedback. We will take this into consideration.

Title: Estimating the State of a Geophysical System with Sparse Observations : Time Delay Methods to Achieve Accurate Initial States for Prediction

Recommendation: Minor Revision General Comment I think the authors have overall well addressed my original concerns and have indeed substantially modified and improved the manuscript also in the light of the other Reviewer' comments. I have only a few additional minor remarks for the Authors to address before acceptance. The comments that follow refer to the last Authors' reply and the numbering is made accordingly.

Specific Comment

Abstract. I understand that Ls is a lower boundary, so that synchronization is achieved when L ≥ Ls. If so please add this piece of information in the text since the name critical number does not fully explain that it is also a minimum number. Also, please explain what the authors mean by synchronization here; a word explanation will suffice.

Ls is no longer referenced in the abstract, but we have added the term `lower bound' to the fifth paragraph where it is introduced.

This lower bound was termed the critical number of measurements L_s required to synchronize the model with the data. Synchronization occurs when the error between the model state estimate and the data drops below a prescribed threshold, which is typically below the magnitude of the observation noise.

On point (4). The paper is still wrongly cited, as Pazo (2016) instead of Pazo et al. (2016), in line 7 page 6.

Sorry, this is fixed now.

On point (18). I could not find where these statements are given in the current version of the manuscript. Maybe they have been removed, which is fine with me. Please clarify this and point the exact portion of the text. If the authors refer to the sentences between lines 4-7 at page 9, those are fine with me but I do not straightforwardly see how they relate to the statements originally under discussion.

To the end of section 5.2 we have added

However, we have not developed a systematic way of choosing these values, and it is known from classical results on synchronization that the optimal choice depends on the number and distribution of observations.

We have removed the statement regarding the model resolution and the observation network.

On point (19). Here as well, I have problems to identify how and where in the text the authors have introduced modifications, if wished, to address my points. For instance, the current Section 5.1 does not seem to be related to the original Section 5.1, instead this seems to be the current Section 5.4. This makes me difficult to judge the changes.

Yes, the noise results were moved to their own section, which is now 5.4. The section was rewritten to make it more clear how noise is added to the data.

Regarding points (1) and (2), I could not find the equations in the current version of the manuscript. Are they still shown somewhere?

These equations were removed, they were confusing.

As for my point (3), I referred to an apparent inconsistency between the type of observations you used. From everywhere in the text is clear you are working with height measurements, but from the equation in the original Section 5.1, it seemed you also create synthetic observations of the velocity. This issue does not seem to be present anymore in the current version (Section 5.4).

Yes, noise is only added to the heights. There are no observations made on velocity. The misleading equations were removed.

Please indicate where in the text are placed the corrections mentioned in points (4) and (5).

References to C_height and C_data have been removed and the text in section 5.4 is now consistent with Fig 7.

On point (23). Please provide information on which actions, if any, you have taken to address this remark. By looking at the current version of the conclusion, it seems that all discussion on equivalence with the smoothers has been removed. Has it been placed elsewhere or modified somehow? Please clarify.

The discussion on the connection with 4DVar was revised considerably and moved to the end of Section 2. We plan to give a more detailed comparison between the two methods in a future paper.