Reviewer comments in black, our reponses in red.

Review by A. J. Geer

This is a very interesting paper that investigates the theoretical limits to wind extraction from tracers in the stratosphere. The "perfect model / perfect observation" approach is informative in this context. With the perfect approach it is possible to almost completely reconstruct the wind fields. After adding observation error, the results are not quite so good but they are still good enough to encourage further work on tracer assimilation. However, I think there are fundamental problems in the way background error has been treated in this study. The experiments should be re-run to correct this. It is unlikely the conclusions will be fundamentally different, but it is necessary to check, and for the credibility of the results, the data assimilation setup needs to be as correct as possible.

Major issues

1) The background error has been set up as follows (section 2.2): "Various values of the horizontal correlation lengths and background error standard deviations were tested to maximize tracer-wind extraction". This is the crux of the whole paper and the authors need to provide more detail. I imagine the authors would have been trying to maximise the wind extraction potential (WEP) in the first cycle of experiment 1. In this approach, the background errors would have been appropriate for exactly the combination of initial conditions, observation errors, and observations found in the first cycle of experiment 1. However, they are unlikely to have been appropriate for the system as it evolved during cycling, or for the other experiments where observation errors were larger. After 10 days of cycling, the errors in the forecast fields in experiments 1 and 2 are radically different: Figs. 7 and 9 show wind errors of around 2m and 20m respectively. The background errors used in the data assimilation should reflect this. If not, the data assimilation system is suboptimal and the WEPs calculated after 10 days are incorrect. Alternatively, if the background error tuning has been done to maximise the 10-day WEPs, the single-cycle WEPs are incorrect.

The best way to address this issue is to tune the background errors for each experiment, and to tune them differently for the single cycle results and for the cycled results.

The reviewer is correct that we did "tune" the results for a single cycle and for a single experiment (Experiment 1). Higher WEP values can therefore be obtained by tuning separately for each experiment and separately for one cycle versus 10 days of cycling. Our original approach was to use fixed values for the background wind error (4 m/s) and background tracer error (0.1 for ozone, 0.01 for nitrous oxide, and 0.3 for water vapor) for all experiments and for all cycles over the 10-day period. The fixed value did not respond to the decreasing background wind and tracer errors as the analysis improved. We decided in this revision, rather than tuning to fixed values, to allow the background error standard deviations to evolve over the course of 10 days, basing them on the global standard deviation of the differences between the analysis and the truth in the previous cycle. This provides a "self-tuning" approach that can be applied consistently for all experiments. We applied this method to the zonal and meridional wind and to the tracer field. The geopotential height background error variances adjust automatically by the dynamical coupling in P_0^b. We kept the horizontal correlation length scales fixed for all experiments as specified in the paper (see Figure 1). Sensitivity tests showed little variation of 10-day WEP values for a reasonable range of length scales. The main impact on WEP was due to the background error variances. We also examined the latitude dependence of the background error to see if a latitude-varying background error would improve the analysis. Varying the background error with latitude did not significantly change the results.

2) It is asserted that (e.g. the abstract) "assimilation of very noisy observations may worsen the wind fields". In an optimal data assimilation system, this should not be possible. In other words, if the basic assumptions of 4D-Var are valid, particularly that observation and background error are correctly modelled, this cannot happen. Hence, it hints at a sub-optimal data assimilation system. The text seems to imply a misconception in the results of experiments 3 and 4 are presented. Here, large observation errors are applied and there is (end of section 4) "a significant worsening from the initial conditions when assimilating noisy data". The observations are not actually worsening the analysis.

The two endpoints on the scale of data assimilation quality are (A) when we have sufficient observations to analyse the real world exactly and (B) when we have no observations, and no matter how good the initial conditions, the model will drift away from reality and lose all skill after 10+ days. As you slowly increase the availability and quality of observations, there will come a point somewhere in the middle where the observations are just good enough to stop the model drifting away from reality completely. Even experiment 2 (realistic obs errors) appears to be below this point, since RMS height errors steadily increase through the 10 day period (Fig. 8) - indeed this is acknowledged in the last sentence of the conclusion. Hence, the observations in experiments 3 and 4 (very high obs errors) are totally insufficient to stop the model drifting away from reality.

Using the new “self-tuning” approach described above, we performed experiments for a wide range of imposed observation errors in order to try to address the “two endpoints” that you mention. We decided to remove the “perfect” observation (no imposed error) experiments from the paper, which were unrealistic, since we had to artificially inflate the background and observation errors to get convergence. Instead of having a “perfect” experiment, we reduced the observation errors as small as possible, until the minimization did not converge to a solution (within the limit of 100 iterations). This “BEST” solution represents endpoint (A). The resulting best WEP obtained for each tracer is 90%. That this number is not 100% likely reflects some non-optimality of the DAS and limitations in the numerics of the solver, rather than limitations in the tracers themselves. There are several possibilities that could limit the system including the assumption of globally constant background error variances, limitations due to use of TLM, and use of approximate dynamical balances in the wind-geopotential cross-covariances.

Endpoint (B) is examined by applying increased observation errors, up to 100% of the global mean value for each tracer. We show in a new figure (Figure 10) that even when applying large observation errors (up to 100% of the global mean) the impact on the winds is positive. We note that in the original submission, we thought that Experiments 3 and 4 (large imposed observation errors) worsened the winds, since the WEP was negative. However, when we calculated WEP for no data assimilation, we found that there actually was improvement due to tracer assimilation with large errors (i.e., the WEP was less negative than in the case of no data). We discuss the impact relative to the case of no observations in the revision.

The crucial point comes back to the background errors: this is a suboptimal system and there might be much greater wind extraction in the high-obs-error case if the background error variances were much larger. These kind of issues have been coming up in the preparations for early-20c reanlyses (e.g. Whitaker, 2009) where it has been shown that 4D-Var can reconstruct the global weather from very sparse observations from the early 1900s, but only if the background errors are sufficiently large. Trying to use background errors appropriate to the present day observing system would extract nearly no information from the sparse observations.

As discussed above, we examined this problem in much more detail and attempted to optimize the system for both low- and high-ob error cases using a consistent self-tuning approach. The MLS results are then presented in the context of the whole range of possible errors, which is very helpful.

To sum up major issues 1 and 2, the authors need to really carefully think about the way the background errors are constructed in their experiment. Should they be constructed so as to be appropriate to a current NWP system with full observations? Probably not, given this is a "perfect" setup and the paper attempts to see if tracer observations alone can constrain the wind fields. But certainly they need to be separately optimised to maximise wind extraction in all the different cases (e.g. single-cycle, 10-day cycling, and the varying range of observation errors in experiments 1-4).

As discussed above, we attempted to optimize wind extraction by specifying the background wind error for each experiment using the differences between the previous analysis and the nature run (i.e., the truth). This approach provides a self-tuning that allows the background errors to decrease as the analysis improves over the 10-day cycle.

1) The paper is focused on stratospheric wind extraction but it needs to be put in context of tropospheric water vapour assimilation, which has been improving wind fields through the tracer-effect in operational systems for at least the last 20 years (e.g. Andersson et al., 1994 , Peubey and McNally, 2009). Perhaps even the title of the paper could be more precisely targeted as: "Wind extraction potential from the assimilation of stratospheric 03 ..."

This is a good point. We changed the title of the paper accordingly and emphasized in several places that we are focusing on stratospheric wind extraction. We also include references to the two papers mentioned to put our paper into context of other work.

2) Section 2.2 on the data assimilation system is lacking a few details. See the major points particularly, but also:

a) It would be great to have a few sentences on the "accelerated representer approach" without the reader having to consult the references - i.e., how does it vary from other 4D-Var algorithms?

We added some more information on the accelerated representer (AR) approach. It minimizes the cost function in observation space using a solver and post-multiplier approach. Although we assume a perfect model, that is not a requirement for AR.

b) Please discuss the approach of constructing a perturbation model for the TL yet using a line-by-line approach for the adjoint. Why not construct both the TL and the AD with the line-by-line approach? At least then the code can be checked using a standard adjoint test.

The TLM was constructed by perturbing the fields in the original nonlinear SWM, which was originally constructed without the 4D-Var problem in mind. Given the spectral form of the SWM, the TLM construction required only minor modifications to the full nonlinear model. The adjoint, on the other hand, was constructed line-by-line from the TLM and checked at each point with standard adjoint test as you suggest. We clarified this in the revision.

c) The BECM (page 25298) could be explained in a little more detail, addressing key questions such as "is it represented explicitly (e.g. full matrix form)?" It would be really great to see the matrix expanded in the text to show the sub-blocks (e.g. wind-wind, wind-height). That would be really helpful in summarising which correlations have been modelled and which have been ignored.

The P_0^b is not represented in full matrix form, since it is too large to be stored in memory, but is rather re-calculated for each iteration. Several terms in the P_0^b are pre-computed to make the calculation efficient. The P_0^b includes all nine cross-correlations of the three dynamical state variables (u, v, and height). There is no cross-correlation between tracer and the dynamical variables in the BECM. We have included additional text on how these were computed along with a schematic of P_0^b to indicate the correlations that are modeled (see Table 1).

d) The paper needs to provide some basic diagnostics to help the reader see if the assimilation system is optimal or not, for example Chi squared or Jo/n

We calculated chi-squared using the formula in Menard et al. (2000), for which the statistical expectation value should be the number of observations. These two numbers agreed to within

1.5% (averaged over the 10-day cycle) for all of our experiments. We indicate this in the revision.

3) In the discussion (Sec. 5) reference is made to the fact that the ozone-wind adjoint has been cut at ECMWF. It could be mentioned that this was because of biases between the modelled and observed ozone fields in the stratosphere. The only way the model could adjust to the observed ozone fields was to make major, erroneous changes to winds and temperatures in the upper stratosphere. Hence, biases are one of the major practical obstacles to overcome before we can extract winds from tracers in the stratosphere.

This is a good point. We will comment on this in the revision.

Andersson, E., Pailleux, J., Thépaut, J.-N., Eyre, J. R., McNally, A. P., Kelly, G. A. and Courtier, P. (1994), Use of cloud-cleared radiances in three/four-dimensional variational data assimilation. Q.J.R. Meteorol. Soc., 120: 627–653. doi: 10.1002/qj.49712051707

Peubey, C. and McNally, A.P. (2009), Characterization of the impact of geostationary clear-sky radiances on wind analyses in a 4D-Var context. Q.J.R. Meteorol. Soc., 135: 1863–1876. doi: 10.1002/qj.500

Whitaker, Jeffrey S., Gilbert P. Compo, Jean-Noël Thépaut, 2009: A Comparison of Variational and Ensemble-Based Data Assimilation Systems for Reanalysis of Sparse Observations. Mon. Wea. Rev., 137, 1991–1999. doi: http://dx.doi.org/10.1175/2008MWR2781.1

Review by T. Milewski

This article addresses the potential of analyzing the stratospheric wind flow by assimilating different trace gases in a simplified and idealized context. It builds upon previous research exploring the idea of wind extraction using chemical data assimilation, and provides new innovative material by showing the impact of assimilating different (passive) tracers in a more theoretical context than recent studies. It also offers an estimate of the maximum wind extraction potential (WEP, a new diagnostic, to the best of the reviewer's knowledge) that can be expected given certain set of parameters, which is interesting information for developping future data assimilation systems. The authors also provide informative comments on the limits and caveats of the idealized experiment settings, as well as giving some results sensitivity on the particular tracer/dynamical situation.

The reviewer only has minor comments: 1) p25296: "In the DAS algorithm, the horizontal flow is converted from the forecast model variables (vorticity and divergence) to zonal and meridional wind (u and v)". Can you provide some reasoning for this switch ? Have you tested the assimilation with vorticity and divergence ? The question comes up considering that you state in the Discussion section that correlations between tracers and potential vorticity may prove useful.

Vorticity and divergence provide convenient coordinates for the global spectral shallow water model, but the modeling of background errors is more straight-forward using the measurable variables u and v (you can't measure vorticity and divergence directly). For example, we can assume globally constant background error variances for u and v with some degree of confidence. It is more difficult to know how to specify the errors for vorticity and divergence. We have tested the assimilation with vorticity and divergence using very simple error specification (e.g., globally constant), and have not found improvement relative to u and v. We also have considered trying streamfunction and velocity potential. With 4-D Var any of these combinations can be used. Further work in this area would be very interesting, including cross-correlations of tracer with potential vorticity.

2) bottom of p25301: change to "where no random error is added to the observations"

We have replaced the "perfect" observation experiment to the "BEST" case with the lowest possible errors, so this phrase has been removed.