REVIEWER # 2

We are grateful to the reviewer for the thoughtful comments and suggestions to our manuscript. We have compiled a revised version and in the following provide a point-by-point reply to all issues raised.

COMMENT # 2.1

> *My background is mostly in data assimilation but I am little familiar with surface energy fluxes.*
>
> *The manuscript by Pirk and co-authors introduces existing data assimilation methodology (plus a new assimilation method hybridising two previous methods) into a new application area of surface energy fluxes observations by new autonomous technology (drones). The topic is interesting and has practical outcomes for the best use of drones and the further exploitation of a promising technology.*
>
> *The manuscript is very nicely written and is at a very mature stage already, making an enjoyable read. The methods are well presented, evaluated rigorously and the results make a convincing case to take the methodology forward. I only have a few minor questions.*

Reply:

Thanks for the nice comments!

COMMENT # 2.2

> *On the data assimilation side I appreciate the introduction of the PIES scheme, which is original to my knowledge. The PIES scheme does not seem to bring much improvement and the authors are open on the shortcomings of the method. What I am missing is a sentence explaining the reasoning behind the PIES scheme: why replace the penultimate iteration of the ES-MDA method and not other ones? Otherwise the comparison of the assimilation methods is done in a correct way. An indication of their respective computational costs would be useful as a perspective.*

Reply:

We followed the reviewer's suggestion, which is related to Comment #1.5 by Reviewer 1, and expanded on the motivation and reasoning behind the PIES scheme as well as the computational costs of the respective DA schemes used in our study. To clarify the notation, given that we now use the symbols N (normal distribution) N_a

(number of assimilation cycles) and N_r (number of LES runs in an experiment), we have changed the symbol for number of ensemble members from N to N_e throughout the manuscript.

Changes:

2.1.4 Data assimilation schemes

...

~~In practice, it may~~ Importance sampling is more effective the closer the proposal is to the target posterior distribution (8). So in theory it would be better to use the posterior estimate from the final (rather than penultimate) iteration of the ES-MDA for the proposal in PIES, but this would come at a high computational cost of requiring an additional round of runs of the LES ensemble. The motivation for pursuing the PIES scheme is that the ES-MDA produces a biased approximation of the posterior for non-linear forward models (16). Although this bias is typically less severe than that of non-iterative ensemble Kalman methods (2), it would nonetheless be advantageous to find efficient methods to reduce it. PIES is a straightforward translation of the scheme of (20) to iterative ensemble smoothers such as the ES-MDA. As such, PIES can be viewed as a simple extension of the ES-MDA that does not necessarily impose any noticeable computational burden and might improve performance. As with all particle methods, the effective sample size can be used to diagnose degeneracy in the ensemble of particles (21). A low ($\ll N_e$) effective sample size indicates degeneracy due to the fact that the proposal is too far from the target posterior. ...

4.3 Data assimilation schemes for turbulent transport

...

The majority of this computational burden stems not primarily from the update steps themselves, but rather from the need to iteratively run an ensemble of LES. The cost of running a single LES with PALM given our experimental setup is on average in the order of 50 CPU hours. The cost of running PALM with a particular parameter combination varies considerably given the adaptive timestep in PALM, but this average cost gives an indication of the considerable computational effort involved. As such, the computational cost of the DA schemes can be measured directly in terms of the number of runs of LES (N_r) required to infer the posterior flux estimates. Herein, these fluxes are parameters rather than states, so we do not strictly need to run posterior predictions, thus lowering the computational costs. Still, the ES-MDA with $N_a = 2$ iterations and $N_e = 100$ ensemble members requires $N_r = N_a \times N_e = 200$ LES. The PIES scheme requires exactly the same number of LES as the ES-MDA. The non-iterative ES and PBS schemes, on the other hand, have a lower cost of $N_r = N_e = 100$ LES. Performing these DA schemes together in the same experiment, i.e. with the same prior ensemble, has a lower cost than running them separately. In particular, while running the ES-MDA all the other schemes can effectively be run for free as benchmarks without the need for any additional LES. The total number of LES undertaken in this study was nonetheless considerable given that we performed 16 synthetic experiments and 18 real experiments, each with $N_r = 200$, amounting to a total of around 6800 LES. It is worth noting that this is still considerably less than the cost of a single Markov Chain Monte Carlo experiment, which typically requires in the order of 10^5 model evaluations. Nonetheless, the cost of these simulations placed a considerable constraint on the number of experiments we could perform to explore an otherwise vast space of design choices that should be investigated in future studies.

Comment # 2.3

> *The synthetic experiments results seem to argue against the random exploration flight strategy, although for a reason related to the data assimilation technique (their effective observation errors are smaller). The authors should insist that their experiments do not disqualify the random flight strategy but may want to devise their observation representation errors more carefully.*

Reply:

We completely agree with this point, which is related to comment #1.4 by Reviewer 1, and clarified that our random exploration strategy used no temporal averaging in all relevant sentences of the manuscript.

Changes:

Abstract

. . .

Sampling strategies prioritizing space-time exploration ~~instead of temporal averaging~~without temporal averaging, instead of hovering at fixed locations while averaging, enhance the non-linearities in the forward model and can lead to biased flux results with ensemble-based assimilation schemes.

2.1.3 Drone measurements, observations and errors (in the added paragraph about the error covariance matrix in relation to comment #1.4 by Reviewer 1)

. . .

The second type involves random exploration where no averaging is performed such that $S = 1$.

4.2 Possible improvements

...

The results indicate that both methods can constrain the surface fluxes, but random exploration without averaging multiple measurements for an observation can give biased flux results. These biases are likely due to shortcomings of the assimilation schemes used when dealing with strongly non-linear forward models rather than the sampling strategy itself, and so could be alleviated by improving the assimilation algorithms.

COMMENT # 2.4

> *There is only one difference between the synthetic case and the real observations case and that is the independent versus correlated H and LE parameters. The authors do not come back to this difference in the discussions: does the correlation of parameters work well or should it be done differently?*

Reply:

We agree that the effect of prior parameter correlations should be brought up again in the discussion. We propose to add the following sentences to Section 4.2 (Possible improvements).

Changes:

4.2 Possible improvements

...

Our framework allows to add further information to the priors through correlations between individual parameters, which we only used for H and LE in our field experiments. The effect of these prior parameter correlations was mostly a slightly more effective exploration of the parameter space, but future studies could investigate how this feature can be used to reduce the computational costs with expensive models like LES.

COMMENT # 2.5

> *The authors also use several statistical metrics to evaluate the methods from the classical RMSE and bias to the CRPS and KLD. It would be interesting to have the authors recommendation on how useful or redundant these metrics are in practice.*

Reply:

In our view, the presented metrics are all useful, as they quantify different aspects of the parameter distributions. RMSE and bias are standard metrics, typically used to compare the fit of point estimates. We also use the less known CRPS to compare the fit of the entire ensemble to the known true values. In that sense, CRPS captures a similar property as RMSE and bias, but we still think it's valuable to report all of them. Lastly, KLD compares the posterior and prior, and can be thought of as a measure of information gain, which is different from the fit of the data. In sum, we would recommend to report all four metrics in studies like ours.

Changes:

2.2 Synthetic experiments

. . .

These four metrics quantify different aspects of the fit and information gain of parameter distributions and can hence give a more holistic evaluation of a synthetic experiment.

COMMENT # 2.6

> ***Detailed comments and typos:***
>
> *-l.15: "variance" is missing "minimum variance".*

Reply:

Technically, the degenerate posteriors of the PBS and PIES schemes have the minimum variance, but this is not desirable in this case. So we propose to clarify this sentence using "well-calibrated posterior uncertainty" instead of "low bias and variance".

Changes:

Abstract

. . .

It is shown that an iterative ensemble smoother outperforms both the non-iterative ensemble smoother and the particle batch smoother in the given problem, yielding ~~low bias and variance posterior distributions~~ well-calibrated posterior uncertainty with continuous ranked probability scores of 12 W m^{-2} for both H and LE with standard deviations of 37 W m^{-2} (H) and 46 W m^{-2} (LE) for a 12 min vertical step profile by a single drone.

COMMENT # 2.7

- l.239 "in a cyclic manner": do you mean the model domain is cyclic?

Reply:

No, in this case we mean that the local mean gradients, i.e. the difference between the mean values at two (vertical) measurement levels, are also calculated for the first measurement location, but using the difference to the last measurement level. We propose to clarify this with the following change in the sentence.

Changes:

2.1.3 Drone measurements, observations and errors

...

This is done in a cyclic manner through the measurement locations, so that the local gradient at the first position is calculated as the difference to the last location.

COMMENT # 2.8

- l.245: I miss an argument that the temporal representativity is somehow related to the spatial representativity of the observations, the discrepancy between the size of the instrument on the drone and the LES cell dimension.

Reply:

We agree that the spatio-temporal aspect of representativeness errors between observations and model should be clarified. Key to this issue is the rotor wash from the drone that mixes the air around the drone, making its measurements less localized – and more representative for spatial scales similar to our LES grid spacing. We propose to add the following sentence to the paragraph in Section 2.1.3.

Changes:

2.1.3 Drone measurements, observations and errors

...

The related spatio-temporal representativeness errors are affected by the rotor wash from the drone that mixes the air around the drone and makes its measurements more representative for spatial scales similar to the LES grid spacing.

COMMENT # 2.9

> *- l.255: Can you explain why 2 m/s errors on wind speed is "conservative"?*

Reply:

The studies we refer to for estimation of horizontal wind speeds from Inertial Measurement Unit data of multi-copter drones (e.g., Palomaki et al., 2017) report measurement uncertainties of less than 0.5 m/s. Since we did not evaluate this uncertainty for our drones, we decided to use a larger value of 2.0 m/s, to avoid underestimating this uncertainty.

Changes:

2.1.3 Drone measurements, observations and errors

...

For the horizontal wind speed U, the standard deviation for the measurement error is conservatively estimated to be 2.0 m s^{-1}. Other studies using Inertial Measurement Unit data of multi-copter drones for wind estimation report measurement uncertainties of less than 0.5 m s^{-1} (27), but since we did not evaluate this uncertainty for our drones, we decided to use a somewhat larger value to avoid underestimating this uncertainty.

COMMENT # 2.10

> *- l.283: "the" is missing before EnKF.*

Reply:

Fixed, thanks.

COMMENT # 2.11

> *- l.420: "mean local differences": this notion is maybe familiar to the surface flux community but I would needed a little definition (horizontal gradient? Positive in which direction?)*

Reply:

We apologize for this confusion, which is caused by inconsistent semantics in this case. We meant to refer to the "local mean gradients" as defined in the manuscript, but wrote "mean local differences".

Changes:

3.2 Field experiments

. . .

The measured mean values and ~~mean local differences~~ local mean gradients are generally well reproduced by the posterior LES ensemble.

COMMENT # 2.12

- l.616 and 617: "an $\mathrm{d} \times \mathrm{N}$ *matrix" should sound better as "a* $\mathrm{d} \times \mathrm{N}$ *matrix".*

Reply:

Since "m" in "matrix" is a consonant, the use of "a" over "an" should be correct.

References

[1] G. Evensen, F. C. Vossepoel, and P. J. van Leeuwen, *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*. Springer Textbooks in Earth Sciences, Geography and Environment, Cham: Springer International Publishing, 2022.

[2] A. A. Emerick and A. C. Reynolds, "Ensemble smoother with multiple data assimilation," *Computers & Geosciences*, vol. 55, pp. 3–15, June 2013.

[3] K. Aalstad, S. Westermann, T. V. Schuler, J. Boike, and L. Bertino, "Ensemble-based assimilation of fractional snow-covered area satellite retrievals to estimate the snow distribution at Arctic sites," *The Cryosphere*, vol. 12, p. 247–270, 2018.

[4] A. M. Stuart, "Inverse problems: A Bayesian perspective," *Acta Numerica*, vol. 19, pp. 451–559, May 2010.

[5] M. A. Iglesias, J. H. Law, and A. M. Stuart, "Ensemble Kalman methods for inverse problems," *Inverse Problems*, vol. 29, no. 4, p. 045001, 2013.

[6] C. Schillings and A. M. Stuart, "Analysis of the Ensemble Kalman Filter for Inverse Problems," *SIAM Journal on Numerical Analysis*, vol. 55, no. 3, pp. 1264–1290, 2017.

[7] M. Iglesias and Y. Yang, "Adaptive regularisation for ensemble kalman inversion," *Inverse Problems*, vol. 37, no. 2, p. 025008, 2021.

[8] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[9] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[10] E. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. doi:10.1017/CBO9780511790423.

[11] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 ed., 2013.

[12] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, "Data assimilation in the geosciences: An overview of methods, issues, and perspectives," *WIREs Climate Change*, vol. 9, Sept. 2018.

[13] G. Evensen, F. C. Vossepoel, and P. J. van Leeuwen, *Data Assimilation Fundamentals*. Springer, 2022.

[14] M. Katzfuss, R. S. Stroud, and C. K. Wikle, "Ensemble Kalman Methods for High-Dimensional Hierarchical Dynamic Space-Time Models," *Journal of the American Statistical Association*, vol. 115, no. 530, pp. 866–885, 2020.

[15] R. M. Neal, "Sampling from multimodal distributions using tempered transitions," *Statistics and Computing*, vol. 6, pp. 353–366, Dec. 1996.

[16] A. S. Stordal and A. H. Elsheikh, "Iterative ensemble smoothers in the annealed importance sampling framework," *Advances in Water Resources*, vol. 86, pp. 231–239, Dec. 2015.

[17] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart, "Interacting langevin diffusions: Gradient structure and ensemble kalman sampler," *SIAM Journal on Applied Dynamical Systems*, vol. 19, pp. 412–441, Jan. 2020.

[18] E. Cleary, A. Garbuno-Inigo, S. Lan, T. Schneider, and A. M. Stuart, "Calibrate, emulate, sample," *Journal of Computational Physics*, vol. 424, p. 109716, Jan. 2021.

[19] O. Dunbar, A. Duncan, A. Stuart, and M.-T. Wolfram, "Ensemble Inference Methods for Models With Noisy and Expensive Likelihoods," *SIAM Journal on Applied Dynamical Systems*, vol. 21, no. 2, pp. 1539–1572, 2022.

[20] N. Papadakis, E. Mémin, A. Cuzol, and N. Gengembre, "Data assimilation with the weighted ensemble Kalman filter," *Tellus A: Dynamic Meteorology and Oceanography*, vol. 62, pp. 673–697, Jan. 2010.

[21] N. Chopin and O. Papaspiliopoulos, *An Introduction to Sequential Monte Carlo*. Springer, 2020.

[22] G. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*. Springer, 1988.

[23] K. Aalstad, S. Westermann, and L. Bertino, "Evaluating satellite retrieved fractional snow-covered area at a high-Arctic site using terrestrial photography," *Remote Sensing of Environment*, vol. 239, p. 111618, 2020.

[24] L. D. van der Valk, A. J. Teuling, L. Girod, N. Pirk, R. Stoffer, and C. C. van Heerwaarden, "Understanding wind-driven melt of patchy snow cover," *The Cryosphere*, 2022.

[25] P. L. Finkelstein and P. F. Sims, "Sampling error in eddy correlation flux measurements," *Journal of Geophysical Research: Atmospheres*, vol. 106, pp. 3503–3509, Feb. 2001.

[26] E. Bassi, “From Here to 2023: Civil Drones Operations and the Setting of New Legal Rules for the European Single Sky,” *Journal of Intelligent & Robotic Systems*, vol. 100, pp. 493–503, Nov. 2020.

[27] R. T. Palomaki, N. T. Rose, M. van den Bossche, T. J. Sherman, and S. F. J. De Wekker, “Wind Estimation in the Lower Atmosphere Using Multirotor Aircraft,” *Journal of Atmospheric and Oceanic Technology*, vol. 34, pp. 1183–1191, May 2017.