We want to thank all the reviewers for providing helpful questions and comments. Here we comment on and clarify some of your questions/concerns, but all comments will of course be addressed when we revise the paper.

Reviewer #1

The connection with PCG samplers and the formal justification of the blocking strategy: It is possible to motivate the marginalized versions of PG/PGAS using the PCG framework of [2], but in the basic setting when θ is fully marginalized, this is not necessary. Simply using cSMC on the marginalized target for the state trajectory will yield samples from the correct stationary distribution. We do, however, use the PCG framework for motivating the blocking strategy. To see that this is correct, set $X_1 = x_{0:B}$, $X_2 = x_{B+1:B+L}$, $X_3 = x_{B+L+1:T}$, $Y = y_{1:T}$. The block sampler essentially implements the following Gibbs sweep to sample from $p(X_1, X_2, X_3, \theta|Y)$

$$X_1, X_2^* \sim p(X_1, X_2|X_3, Y, \theta), \quad X_2, X_3 \sim p(X_2, X_3|X_1, Y), \quad \theta \sim p(\theta|X_1, X_2, X_3, Y).$$

Step 1 and 3 are standard Gibbs steps. In step 2 , θ is collapsed, which can be thought of as adding a draw of θ from its full conditional to yield a draw from $p(X_2, X_3, \theta|X_1, Y)$. Since θ is not conditioned on in step 3, removing θ from step 2 is valid and the scheme is correct. Similar arguments can be applied for the actual sampler, which makes use of cSMC kernels in step 1 and 2. We agree that the validity of the block sampler was not very clearly shown in the paper, and we have therefore added a detailed proof in the supplement.

Comparison PG/mPG/PGAS/mPGAS: PGAS/mPGAS are now implemented in Birch. We have chosen to focus on the performance improvement offered by marginalization. On the toy model we observe a clear improvement from marginalizing both for PG and PGAS, but also from using PGAS/mPGAS compared to using PG/mPG. However, for the VBD model we observe a clear improvement from marginalizing, but using PGAS/mPGAS instead of PG/mPG gives no clear improvement. Results supporting these claims have been added to the supplement.

Computational cost: There are indeed some extra computations for the marginalized methods, but the overhead is quite small. For the toy model, with N=500, using the tic-toc timer in MATLAB we get: PG 1231.5 s mPG 1430.7 s PGAS 1260.7 s mPGAS 1566.1 s. Note that the code has not been optimized.

Line 45-56, misleading discussion: This discussion is mainly intended as a pedagogical motivation for why marginalization is useful. PG is the standard approach in many cases, but is limited by the "ideal" Gibbs it approximates. What we propose to do instead is, like you point out, to approximate the "ideal" collapsed Gibbs sampler using marginalized versions of PG/PGAS. Note that "ideal" here does not mean optimal, but refers to the hypothetical non-particle version. We have clarified this in the paper, and to avoid confusion we have changed the word "ideal" to "hypothetical".

Reviewer #2

Using backward sampling? Yes that is possible, however, [1] argue that ancestor sampling is more suitable for non-Markovian models and we have therefore chosen to focus on PGAS.

Path degeneracy: Indeed, ancestor sampling helps to reduce the effect of path degeneracy. Furthermore, the MCMC nature of mPG/mPGAS means that we can "revisit" and update states at early time steps, which is not possible in purely "online" methods, which also mitigates the effect of path degeneracy.

Reviewer #3

Extend the evaluation with more large models and on more datasets: The VBD implementation has been extended to daily data (instead of weekly) and we have updated relevant figures in the paper with these. All conclusions are the same as before. Regarding running times, see the reply to Reviewer #1.

Clearer definitions: (1) Definitions of h,s, and A are the same as for the exponential family (defined in the supplement). For clairity we have added definitions of all variables in the main text and a more detailed explanation of the restricted exponential family is now in the supplement. (2) $\tilde{w}^i$ is the weight of each possible ancestor trajectory, and is used in the resampling step to assign a new ancestor path to the reference trajectory. A detailed description and motivation can be found in [1]. We have clarified this in the paper.

Figure 1: Ideally we would like iid samples from the posterior distribution, in terms of the ACF of the samples it should be zero everywhere except for lag 0. Figure 1 illustrate that mPGAS yields samples with a lower autocorrelation (closer to iid) than what is attainable with standard PGAS, even for a low number of particles ($N = 50$).

Path degeneracy of PGAS (line 81): We viewed path degeneracy as the issue that the mixing rate goes to zero as T becomes large (for fixed N), which is typically not the case for PGAS. However, the mixing of PGAS is indeed affected by the mixing properties of the model. Typically, it decreases to a *non-zero constant* as T becomes large. To avoid confusion we have changed the wording from "unaffected" to "more robust".

References

[1] F. Lindsten, M. I. Jordan, and T. B. Schön. Particle Gibbs with ancestor sampling. *JMLR*, 15:2145–2184, 2014

[2] D. A. van Dyk and T. Park. Partially collapsed gibbs samplers. *JASA*, 103(482):790–796, 2008