We thank the reviewers for their invested time and valuable suggestions, which we will gladly incorporate to improve the paper. We first respond to two comments made by all reviewers and then address additional comments separately.

Novelty: As pointed out by the reviewers, the proposed model and algorithm have several components that are similar to previously published work. While we did attempt to highlight the novel contributions and the overlap with prior work in the related work and methods sections, all reviewers raised questions around this point, so we provide further clarification about these similar components below and we will update the relevant sections to point out the novelty.
i) **Non-linear parametrisation of a GLS**: The KVAE and DeepState parametrise the GLS through a *deterministic* RNN and use inputs that have only partial information to predict the SSM parameters. That is, DeepState uses only controls $\mathbf{u}_t$ (open loop) and KVAE uses *samples* of the *pseudo-observations* $\mathbf{z}_t$. In contrast, our *probabilistic* switches transition is conditioned on the controls $\mathbf{u}_t$ and the state variable $\mathbf{x}_t$ for which uncertainty is preserved (no sampling) and which includes dynamics information (in contrast to $\mathbf{z}_t$). Furthermore, we perform *probabilistic inference* via (Rao-Blackwellised) SMC rather than assuming that the SSM parameters can be predicted with absolute certainty.
ii) **State-to-switch recurrence**: While previous work (Barber 2006, Linderman 2017, Becker-Ehmck 2019) has also used a form of recurrence, the details differ substantially. The approach of Linderman 2017 is unfortunately not suitable for optimization with SGD since the Polya-Gamma distribution can not be reparametrised. Becker-Ehmck 2019 samples the state variable $\mathbf{x}_t$ and Barber 2006 makes a sample-based or deterministic approximation. In contrast, our conditionally linear Gaussian state-to-switch recurrence allows us to maintain the full distribution.
iii) **Auxiliary variable with a decoder-type neural network**: Our approach differs from the same component in the KVAE in the inference algorithm and choice of proposal distribution / variational approximation. Whereas the KVAE uses an encoder-based variational approximation, we apply SMC and use a similar encoder only as part of the proposal distribution, i.e. by taking the product with the conditional distribution $p(\mathbf{z}_t, \mathbf{s}_t|\mathbf{z}_{1:t-1}, \mathbf{s}_{1:t-1})$ (where the states $\mathbf{x}_{t-1}$ are marginalised out). This is well motivated by the structure of the optimal proposal distribution (see comments below) and similar arguments could have been made for the variational approximation in the KVAE.

Algorithm box and visualisation: These are good suggestions, we will use the extra page to add an algorithm box with references to a visualisation of the graphical model and the computational structure of the inference procedure.

Reviewer 1:
i) We agree that evaluating the **impact of the Rao-Blackwellisation** is interesting and will provide additional results.
ii) Regarding the potential **drawback of the linear latent transitions**, note that *conditionally* linear systems can approximate non-linear systems through linearisation around the current state. We will expand on this in the paper.
iii) Wrt. the **optimal proposal distribution**, cf. Sec. 3.3.3 and the numerator of Eq. (11). The latter two conditionals are known; we therefore choose the same factorisation and only replace the likelihood term by a Gaussian approximation (through an encoder). It is optimal if the likelihood is indeed Gaussian. We will discuss this in more detail.
iv) We were not aware of Poyiadjis et al. (2011). This gradient estimate seems related to black-box variational inference in the sense that both are based on a log-derivative trick. It potentially yields higher-variance estimates, so there might be a variance-bias trade-off. This is an interesting research question that we try to answer for the camera-ready version.
v) Thank you for carefully spotting undefined quantities and unclear statements, we will clarify these in the paper. All of the crossed (diagonal line) terms cancel due to independence, no additional assumptions are made.

Reviewer 2:
i) It is indeed possible to extend the approach to **hierarchies of state variables**, e.g. by interleaving non-linear, sampled switches with linear states. The details are not as straightforward, but we plan to develop this extension in future work.
ii) **Scaling the approach to larger models** should be feasible. The decoder-type likelihood in Sec. 5.3 already uses a CNN and the non-linear switch transitions could also make use of more recent architectures that use gating or attention, although the latter requires to alleviate the Markovian assumption.
iii) Regarding the **strong baseline DeepAR**, we will add a short motivation/discussion about other practical benefits of SSMs compared to AR models, e.g. in applications such as anomaly detection and when handling missing data. AR models need to impute missing or anomalous data and thus accumulate errors. In contrast, SSMs maintain uncertainty information and provide a principled way to ignore missing or anomalous data, i.e. by omitting the Bayes update (Kalman update step). Note also that DeepAR as originally described can handle only univariate time-series and can thus not be directly applied to the scenario from Sec. 5.3.
iv) We will discuss the related work you pointed out which is indeed very relevant.

Reviewer 3:
i) Please see above for detailed comments on the **novelty**.
ii) The purpose of the **pendulum toy experiment** in Sec. 5.1 is to qualitatively demonstrate the necessity of the state-to-switch recurrence. For a quantitative evaluation and comparison with related methods please refer to Secs. 5.2 and 5.3. We agree that a comparison with DVBF, DKN and others would have been interesting, but we decided to focus on the most closely related methods, i.e. KVAE and DeepState.