We thank all the reviewers for their insightful and positive feedback. We are delighted they enjoyed reading the paper, and they found our approach to be interesting and useful. We are also very pleased the reviewers perceived the approach to have significant potential, beyond just performance gains.

Common comments: We agree that the clarity of qualitative experiments section could be improved by giving a better description of Figure 2 (R3) and adding missing axis labels (R5). The detailed description of these experiments had to be moved to the Appendix due to space limitations. We will make this section clearer in the final version.

We will now address the reviewers' comments individually.

Reviewer 3:
i) We will add discussion about the relation between our method and Kernel Kalman Filters, State Space Gaussian Process (GP) approach as well as GP latent variable model in the final version. We also note that GP-Copula [1] used in the evaluation is an example of non-Gaussian AR methods.
ii) *hyperparams ('observation' and 'state' noise variances) are sampled:* It is not the variances that are sampled but the random noise in order to sample from the forecast distribution. The observation and state variances are not hyperparameters but are actually predicted from the RNN, using the associated covariates. Here sampling is done because in applications it is preferable to represent the forecast distribution in the form of Monte Carlo samples [1, 8, 9].
iii) *The missing samples approach seems the same as that proposed for the KF by Shumway and Stoffer:* Thank you for the reference. We agree it is a standard procedure and will add a citation.

Reviewer 4:
i) *Placing the emission noise before the nonlinearity is a crucial move; omission of any discussion of the non-additive noise:* You are perfectly right that placing the noise before the non-linearity is crucial to obtain tractability for filtering. However, this does not imply that data generated from a process where additive noise is added after the non-linear function cannot be modelled: in fact we use this particular model (nonlinear transformation with additive non-Gaussian noise) to generate data and test our method in the qualitative experiments in Section 4.1 (refer to appendix C.1 for details). We will make this more explicit in the final version.
ii) *This technique can be applied as a drop-in replacement in many models:* Yes, we believe so.
iii) *Simple example:* There may be issues if the normalizing flow is highly skewed, however we have not encountered this issue in practice.
iv) *Qualitative experiments seemed particularly artificial:* While the main purpose of these experiments is a sanity check for the overall method, the artificial data generated still has all the intricacies of real world scenarios like non-Gaussian *time varying* observation noise coupled with a seasonal behaviour and non-linear dependencies between the time series. The complex nature of the time series observations are better highlighted in appendix Section C.1.2 when plotted against time. Also note that the generative model used to simulate artificial data is not exactly same as our model but instead uses non-Gaussian additive noise after nonlinear transformation of the state.
v) *Results of GP-Copula for some datasets differ between your paper and [1]:* The results in the paper of GP-Copula use a flawed variant of CRPS-Sum *without* normalizing the time-series beforehand as in the data, the time-series have drastically different scales, hence not properly evaluating the captured dependencies across time-series.
vi) Thanks for the detailed comments in the additional feedback section. We will incorporate these in the final version.

Reviewer 5:
i) *Partial observations:* The local-global instantiation presented in Section 3 could deal with partial observations, but would require marginalising over the missing dimensions resorting to a Monte Carlo approximation. Alternatively, if one is interested in dealing with partial observations, it is possible to consider an instantiation with a global LGM with non-diagonal covariance matrix modelling dependencies across time-series, and a normalising flow is applied locally to each time-series (as in our ablation study model f_t Local). In this case, the marginalisation of any missing set of time-series actually yields an analytic form for the filtering, smoothing and forecast distributions, and handling of partial observations can be efficiently dealt with. Note that many nonlinear methods do not achieve this. We will be happy to respond to this in the final version.
ii) *Representational capacity difference between having fully independent latent time series or not when normalizing flows are used to mix across latent time series?* Just as you have said, we believe that without loss of expressivity, the latent time-series can be considered independent, assuming that the normalizing flow itself is expressive enough. However it may be still be useful to consider the model in its general form for completeness, e.g. as one may have some prior knowledge in the form of the latent LGM that one may wish to encode explicitly.
iii) *Hyper-parameter optimization in the main paper:* We agree that although we mention that we use a validation set in the appendix Section C, this should also be contained in the main paper. This will be modified. In order to select the hyper-parameters, we use a validation set, which is derived from the training set: we cut the original training set in time into two parts, taking the most recent part for validation, so as to have a validation set of the same size as the test set.