Supplementary material to "Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering"

Marcin Jurek* Matthias Katzfuss*†

S1 Neighborhood averaging

In the main paper, we assume from Section 3 onwards that observations are conditionally independent given the latent process, i.e. that

$$y_i \perp y_j \mid \mathbf{x} \qquad i, j \in \mathcal{I}. \tag{S1}$$

While the results in Proposition 2 do not hold under an arbitrary dependence structure, condition (S1) can be somewhat relaxed. In particular, recall from (3) that the density of the latent process under the HV approximation can be written as

$$\hat{p}(\mathbf{x}) = \prod_{m=0}^{M} \prod_{j_1,\ldots,j_m} p(\mathcal{X}_{j_1,\ldots,j_m} | \mathcal{A}_{j_1,\ldots,j_m}).$$

Define $\mathcal{I}_{j_1,\ldots,j_m} = \{i : x_i \in \mathcal{X}_{j_1,\ldots,j_m}\}$ to be the indices of the grid points which correspond to variables in $\mathcal{X}$. Then let $\mathcal{Y}_{j_1,\ldots,j_m} = \{y_i : i \in \mathcal{I}_{j_1,\ldots,j_m}\}$.

Now, if we assume that

$$\mathcal{Y}_{j_1,\ldots,j_\ell} \perp \mathcal{Y}_{i_1,\ldots,i_m} \mid \mathbf{x} \tag{S2}$$

instead of (S1), we can carry out the proof of Proposition 2 with the only change being that we consider sets of variables $\mathcal{X}_{j_1,\ldots,j_\ell}, \mathcal{X}_{i_1,\ldots,i_m}$ and $\mathcal{Y}_{j_1,\ldots,j_\ell}, \mathcal{Y}_{i_1,\ldots,i_m}$ instead of x_j, x_i and y_j, y_i.

This allows for observations corresponding to average values of the process over a certain neighborhood. For example, if only averages over domain regions $\mathcal{D}_{j_1,\ldots,j_m}$ are available, we could assume that $\mathcal{Y}_{j_1,\ldots,j_m}$ has only one element equal to

$$y_{j_1,\ldots,j_m} = \left(\frac{1}{\#\mathcal{X}_{j_1,\ldots,j_m}} \sum_{x \in \mathcal{X}_{j_1,\ldots,j_m}} x \right) + v_{j_1,\ldots,j_m},$$

with $v_{j_1,\ldots,j_m} \sim \mathcal{N}(0, \tau^2_{j_1,\ldots,j_m})$.

*Department of Statistics and Data Science, University of Texas at Austin
†Corresponding author: katzfuss@gmail.com, Department of Statistics, Texas A&M University

S2 A particle filter for unknown parameters

The distributions and matrices in model (12)–(13) may depend on parameters $\boldsymbol{\theta}_t$ at each time t, which we have implicitly assumed to be known thus far. We now discuss the case of a (small) number of unknown parameters $\boldsymbol{\theta}_t$. Specifically, $\boldsymbol{\mu}_{0|0}$ and $\boldsymbol{\Sigma}_{0|0}$ may depend on $\boldsymbol{\theta}_0$, and the quantities $\{g_{t,i}\}$, $\mathcal{E}_t$, and $\mathbf{Q}_t$ at each time t may depend on $\boldsymbol{\theta}_t$. There are two main approaches to simultaneous filtering for the state $\mathbf{x}_t$ and the parameters $\boldsymbol{\theta}_t$: state augmentation and Rao-Blackwellized filters (Doucet and Johansen, 2011). The main idea behind the former is to include $\boldsymbol{\theta}_t$ in the state vector $\mathbf{x}_t$ and to modify the evolution and the model error matrices accordingly, but this approach is known to work poorly in certain cases (e.g., DelSole and Yang, 2010; Katzfuss et al., 2020). Thus, following Jurek and Katzfuss (2021), we now present a Rao-Blackwellized filter in which integration over $\mathbf{x}_t$ is performed based on our HVL approximation.

Writing $\boldsymbol{\theta}_{0:t} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_t)$, the integrated likelihood at time t is given by

$$p(\mathbf{y}_{1:t}|\boldsymbol{\theta}_{0:t}) = p(\mathbf{y}_1|\boldsymbol{\theta}_{0:1}) \textstyle\prod_{k=2}^{t} p(\mathbf{y}_k|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}_{0:k}). \tag{S3}$$

It is well known that

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}) = \frac{p(\mathbf{y}_t, \mathbf{x}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t})}{p(\mathbf{x}_t|\mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t})} = \frac{p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t})}{p(\mathbf{x}_t|\mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t})}, \tag{S4}$$

where $p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t)$ is available in closed form from (12), and the forecast and filtering distributions can be approximated using the EKVL, to obtain

$$\mathcal{L}_t(\boldsymbol{\theta}_{0:t}) := \hat{p}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t)\mathcal{N}(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})}{\mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})}. \tag{S5}$$

The normal densities can be quickly evaluated for given parameter values $\boldsymbol{\theta}_{0:t}$, because Algorithm 5 calculates sparse Cholesky factors of their precision matrices. For $t = 1$, the term $\mathcal{L}_1(\boldsymbol{\theta}_{0:1}) := \hat{p}(\mathbf{y}_1|\boldsymbol{\theta}_{0:1})$ can be approximated in a similar way using $\boldsymbol{\mu}_{0|0}$ and $\boldsymbol{\Sigma}_{0|0}$. The particle-EKVL filter is given by Algorithm S1, assuming that the parameter priors are given by $f_0(\boldsymbol{\theta}_0)$ and then recursively by $f_t(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$.

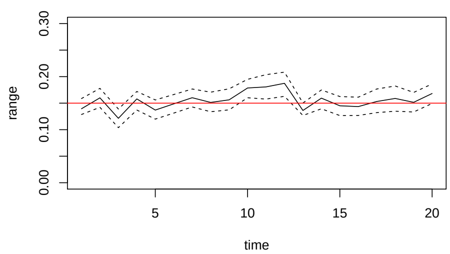
We demonstrate the accuracy of the particle filter for inferring the range parameter λ of the exponential covariance function within the framework described in Section 6.3, except that we assumed that all grid cells have observations associated with them (i.e., $n_t = n$). We consider the parameter $\theta_t = \log \lambda_t$, assuming independent and identical prior distributions $f_t(\theta_t) = \mathcal{N}(\log(0.15), 0.5^2)$ at all time points. We use proposal distributions of the form $q_t(\theta_t|\theta_{t-1}) = \mathcal{N}(\theta_{t-1}, 0.1)$. Using $N_p = 1{,}000$ particles, we obtained the results summarized in Figure S1. We conclude that Algorithm S1 can be successfully used for parameter estimation.

Algorithm S1: Particle-EKVL filter

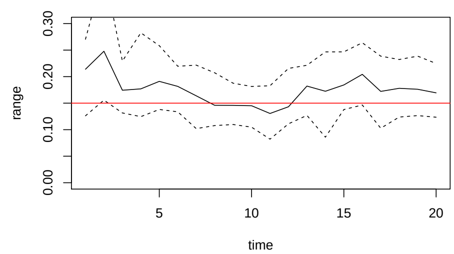
Input: $\mathbf{S}$, $\boldsymbol{\mu}_{0|0}$, $\boldsymbol{\Sigma}_{0|0}$, $\{(\mathbf{y}_t, \mathcal{E}_t, \mathbf{Q}_t, \{g_{t,i}\}) : t = 1, 2, \ldots\}$, priors $\{f_t\}$, proposal distributions $\{q_t\}$, desired number of particles N_p

Result: $\{(\boldsymbol{\theta}_t^{(l)}, w_t^{(l)}, \boldsymbol{\mu}_{t|t}^{(l)}, \mathbf{L}_{t|t}^{(l)}) : l = 1, \ldots, N_p\}$, such that $\hat{p}(\boldsymbol{\theta}_t, \mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{l=1}^{N_p} w_t^{(l)} \delta_{\boldsymbol{\theta}_t^{(l)}}(\boldsymbol{\theta}_t)\mathcal{N}_n(\mathbf{x}_t|\boldsymbol{\mu}_{t|t}^{(l)}, \mathbf{L}_{t|t}^{(l)}\mathbf{L}_{t|t}^{(l)\top})$

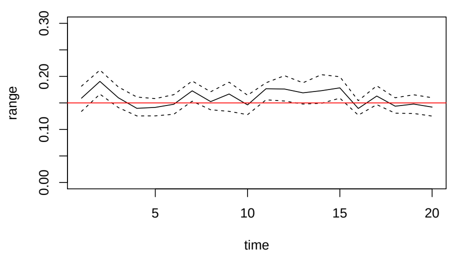
for $l = 1, 2, \ldots, N_p$ **do**
  Draw $\boldsymbol{\theta}_0^{(l)} \sim f_0(\boldsymbol{\theta}_0)$ and set weight $w_0^{(l)} = 1/N_p$
  Compute $\mathbf{L}_{0|0}(\boldsymbol{\theta}_0^{(l)}) = \text{ichol}(\boldsymbol{\Sigma}_{0|0}(\boldsymbol{\theta}_0^{(l)}), \mathbf{S})$
  Compute $\boldsymbol{\mu}_{0|0}^{(l)}(\boldsymbol{\theta}_0^{(l)})$ and $\mathbf{U}_{0|0}(\boldsymbol{\theta}_0^{(l)}) = \mathbf{L}_{0|0}^{-\top}(\boldsymbol{\theta}_0^{(l)})$
end for
for $t = 1, 2, \ldots$ **do**
  for $l = 1, 2, \ldots, N_p$ **do**
    Draw $\boldsymbol{\theta}_t^{(l)} \sim q_t(\boldsymbol{\theta}_t^{(l)}|\boldsymbol{\theta}_{t-1}^{(l)})$
    Calculate $\mathbf{A}_t^{(l)} = \frac{\partial \mathcal{E}_t(\mathbf{y}_{t-1}, \boldsymbol{\theta}_t^{(l)})}{\partial \mathbf{y}_{t-1}}\Big|_{\mathbf{y}_{t-1} = \boldsymbol{\mu}_{t-1|t-1}(\boldsymbol{\theta}_{t-1}^{(l)})}$
    Forecast: $\boldsymbol{\mu}_{t|t-1}^{(l)} = \mathcal{E}_t(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\theta}_{t-1}^{(l)})$ and $\mathbf{L}_{t|t-1}^{(l)} = \mathbf{A}_t^{(l)}\mathbf{L}_{t-1|t-1}^{(l)}$
    For (i, j) s.t. $\mathbf{S}_{i,j} = 1$: $\boldsymbol{\Sigma}_{t|t-1;i,j}^{(l)} = \mathbf{L}_{t|t-1;i,:}^{(l)}(\mathbf{L}_{t|t-1;j,:}^{(l)})^\top + \mathbf{Q}_{t;i,j}(\boldsymbol{\theta}_t^{(l)})$
    Update: $[\boldsymbol{\mu}_t^{(l)}, \mathbf{L}_{t|t}^{(l)}] = \texttt{HVL}(\mathbf{y}_t, \mathbf{S}, \boldsymbol{\mu}_{t|t-1}^{(l)}, \boldsymbol{\Sigma}_{t|t-1}^{(l)}, \{g_{t,i}(\boldsymbol{\theta}_t^{(l)})\})$
    Calculate $\mathcal{L}_t(\boldsymbol{\theta}_{0:t}^{(l)})$ as in (S5)
    Update particle weight $w_t^{(l)} \propto w_{t-1}^{(l)}\mathcal{L}_t(\boldsymbol{\theta}_{0:t}^{(l)}) f_t(\boldsymbol{\theta}_t^{(l)}|\boldsymbol{\theta}_{t-1}^{(l)})/q_t(\boldsymbol{\theta}_t^{(l)}|\boldsymbol{\theta}_{t-1}^{(l)})$
    return $\boldsymbol{\mu}_{t|t}^{(l)}, \mathbf{L}_{t|t}^{(l)}, \boldsymbol{\theta}_t^{(l)}, w_t^{(l)}$
  end for
  Resample $\{(\boldsymbol{\theta}_t^{(l)}, \boldsymbol{\mu}_{t|t}^{(l)}, \mathbf{L}_{t|t}^{(l)})\}_{l=1}^{N_p}$ with weights $\{w_t^{(l)}\}_{l=1}^{N_p}$ to obtain equally weighted particles (e.g., Douc et al., 2005)
end for



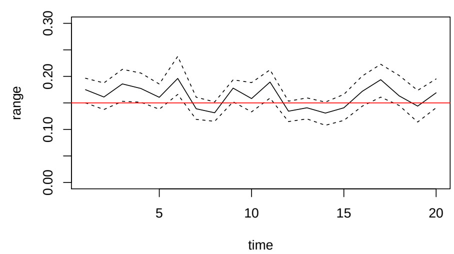
(a) Gaussian data



(b) binary data



(c) Poisson data



(d) gamma data

Figure S1: Distribution of particles corresponding to the range parameter λ. The solid black line represents the mean and the dashed lines stand for the 10-th and 90-th percentile. The red horizontal line is the true value.

S3 Nonlinear evolution with non-Gaussian data

In the setting of the most general model (12)-(13) with nonlinear evolution $\mathcal{E}_t$, we considered a complicated model (Lorenz, 2005, Sect. 3) that realistically replicates many features of atmospheric variables along a latitudinal band, and whose different variants are commonly used as a test-bed for other data assimilation techniques (e.g., Ott et al., 2004). The model dynamics are described by

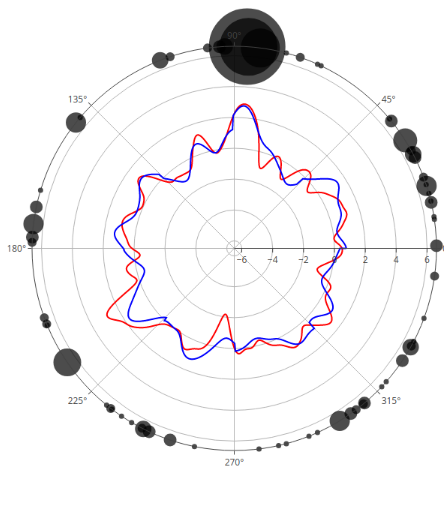
$$\tfrac{\partial}{\partial t}\tilde{x}_i = \tfrac{1}{K^2}\textstyle\sum_{l=-K/2}^{K/2}\sum_{j=-K/2}^{K/2} -\tilde{x}_{i-2K-l}\tilde{x}_{i-K-j} + \tilde{x}_{i-K+j-l}\tilde{x}_{i+K+j} - \tilde{x}_i + F, \tag{S6}$$

where $\tilde{x}_{-i} = \tilde{x}_{n-i}$, and we used $K = 32$ and $F = 10$. By solving (S6) on a regular grid of size $n = 960$ on a circle with unit circumference using a 4-th order Runge-Kutta scheme, with five internal steps of size $dt = 0.005$, and setting $x_i = b\tilde{x}_i$ with $b = 0.2$, we obtained the evolution operator $\mathcal{E}_t$. We also calculated an analytic expression for its derivative $\nabla\mathcal{E}_t$, which is necessary for Algorithm 5.

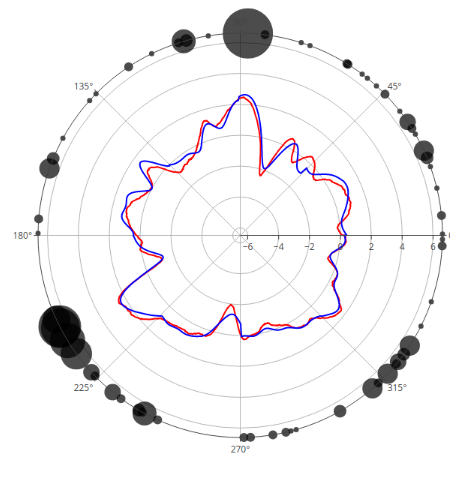
To complete our state-space model (12)-(13), we assumed $\mathbf{Q}_{t;i,j} = 0.2\exp(-\|\mathbf{s}_i - \mathbf{s}_j\|/0.15)$, we randomly selected $n_t = n/10 = 96$ observation indices at each t, and we took the initial moments $\boldsymbol{\mu}_{0|0}$ and $\boldsymbol{\Sigma}_{0|0}$ to be the corresponding sample moments from a long simulation from (S6). We simulated 40 datasets from the state-space model, each at $T = 20$ time steps, and for each of the four exponential-family likelihoods, using $\tau^2 = 0.2$ in the Gaussian case.

Figure S2 shows sample trajectories from this Lorenz model and shows that the trajectories can be successfully recovered using the EKVL filter (Algorithm 5), even based on non-Gaussian data.

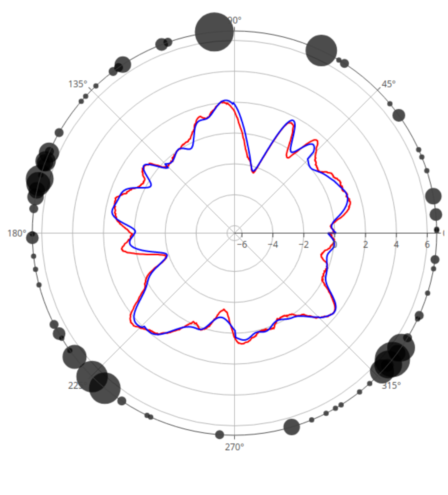
For each simulated data set, we applied the three filtering methods described in Section 6.1. We used $N = 39$, and for the EKVL filter we set $J = 2$, $M = 7$, and $|\mathcal{X}_{j_1,\ldots,j_m}|$ equal to 5, 5, 5, 5, 6, 6, 6, respectively, for $m = 0, 1, \ldots, M-1$, and $|\mathcal{X}_{j_1,\ldots,j_M}| \leq 1$. The average scores over the 40 simulations are shown in Figure S3. Our method (HV) compared favorably to the low-rank filter and provided excellent approximation accuracy as evidenced by very low RRMSPE and dLS scores.



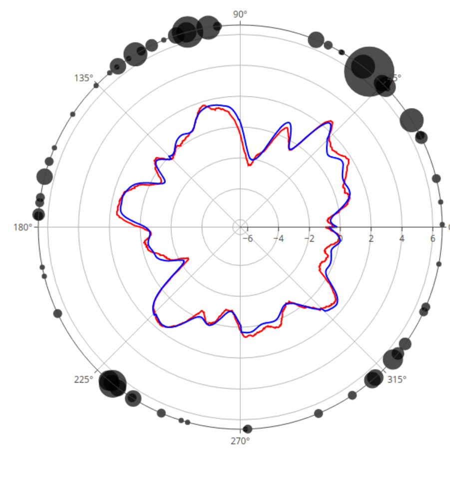
(a) $t = 1$



(b) $t = 5$



(c) $t = 10$



(d) $t = 15$

Figure S2: Simulation from the Lorenz model described in Section S3 (blue line) with Poisson observations (black circles). Size of the circle denotes the value of the observation. Red line is the filtering mean obtained using the HV filter. The figure shows that even using (non-Gaussian) Poisson observations, over time the filter is able to successfully recover the latent process.

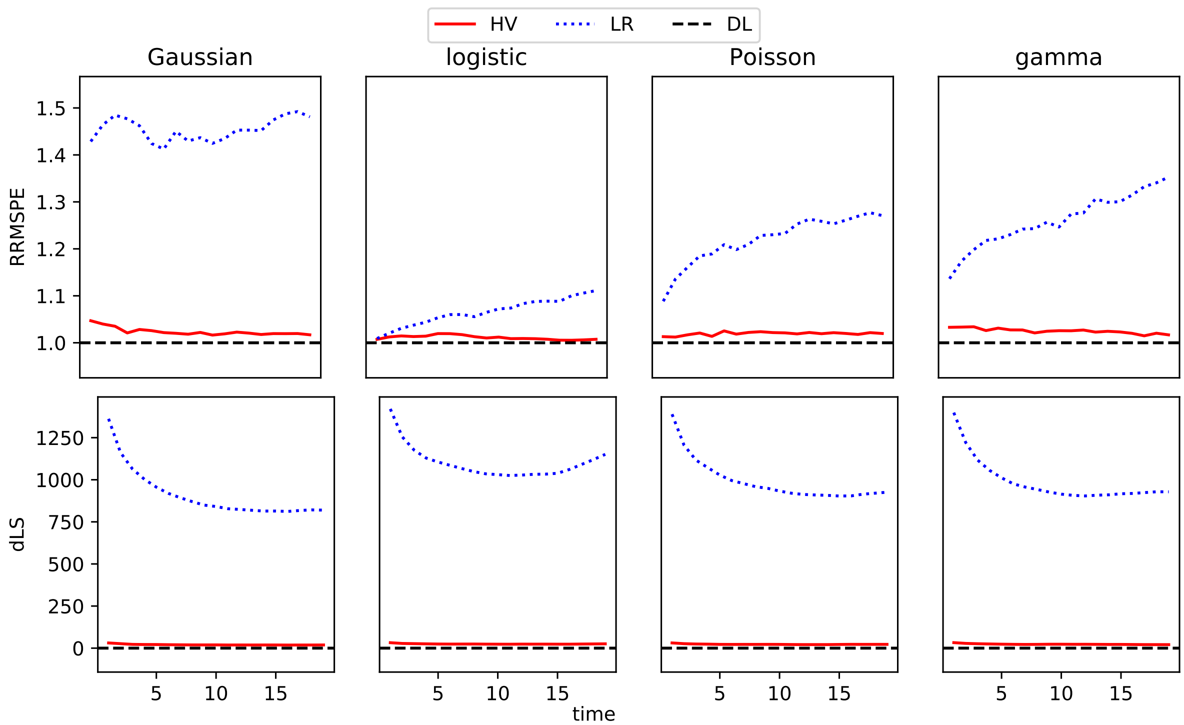
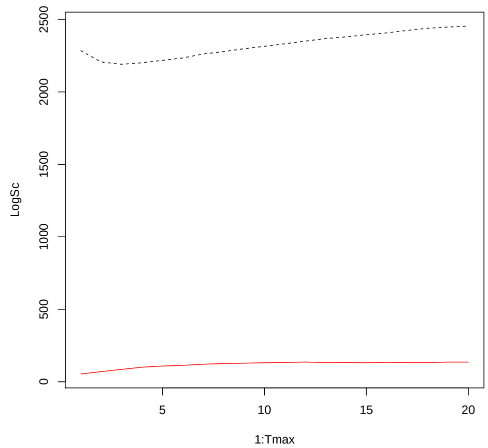


Figure S3: Accuracy of filtering distribution $\mathbf{x}_t|\mathbf{y}_{1:t}$ for the Lorenz model in Section S3

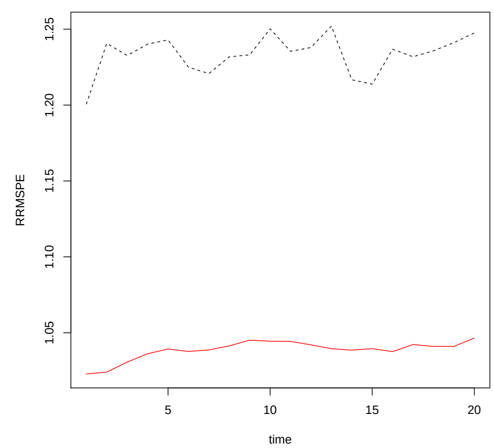
S4 Comparison with the ensemble Kalman filter

To compare the HV filter to the ensemble Kalman filter (EnKF; e.g., Katzfuss et al., 2016), we performed filtering inference using both filters to the linear advection-diffusion model and the Lorenz model, described in Sections 6.3 and S3, respectively. We restricted ourselves to Gaussian data only, because there does not exist a canonical extension of the EnKF to non-Gaussian data. In order to ensure a fair comparison, we used 50 ensemble members in EnKF (compared to $N \leq 41$ for HV) and we adjusted the tapering radius so that the covariance matrix has at most 50 nonzero elements in a row.

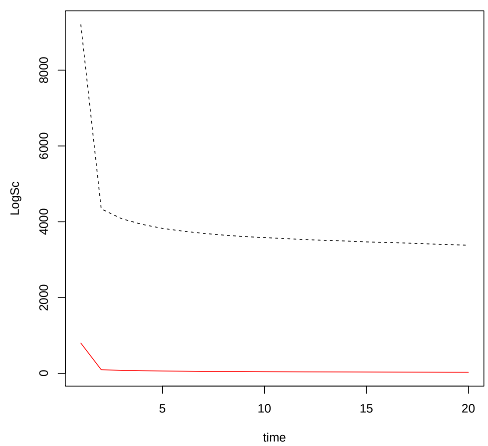
We used the same scoring methods as described in Section 6.1. Looking at the results presented in Figure S4, we conclude that the HV filter can substantially outperform the EnKF. While EnKF methods can be very effective, even in the presence of nonlinear evolution, they often struggle with long-range correlations in the forecast covariance, because EnKF methods typically require localization (e.g., tapering the empirical covariance matrix). Our study illustrates this problem.



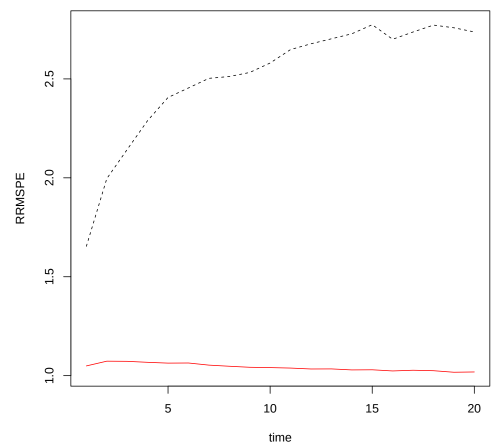
(a) Log-scores for the advection-diffusion model



(b) RRMSPE for the advection-diffusion model



(c) Log-scores for the Lorenz model



(d) RRMSPE for the Lorenz model

Figure S4: Comparison of filtering results using EnKF and the HV filter. The dashed black lines represent the EnKF scores and the solid red lines represent the HV filter scores.

S5 Visual artifacts in the HV approximation

The hierarchical conditional-independence assumptions in the HV approximation mean that certain dependence relations are not captured accurately. For example, in the toy example presented in Figure 1a, the correlation of points that are close to the boundary between $\mathcal{D}_1$ and $\mathcal{D}_2$ is only captured through black point on the leftmost plot (see caption under Figure 1a). We show a particularly bad example of these artifacts in Figure S5, where a horizontal line corresponding to the first subregion split is clearly visible.

While HV is certainly not a perfect spatial approximation method, the “blockiness” of predictions is greatly alleviated when filtering over time. Figure S6 shows how the visual artifacts quickly disappear in the case of an advection-diffusion model and Section S3 illustrates the same for the Lorenz model. See Section 3.4 for more details.

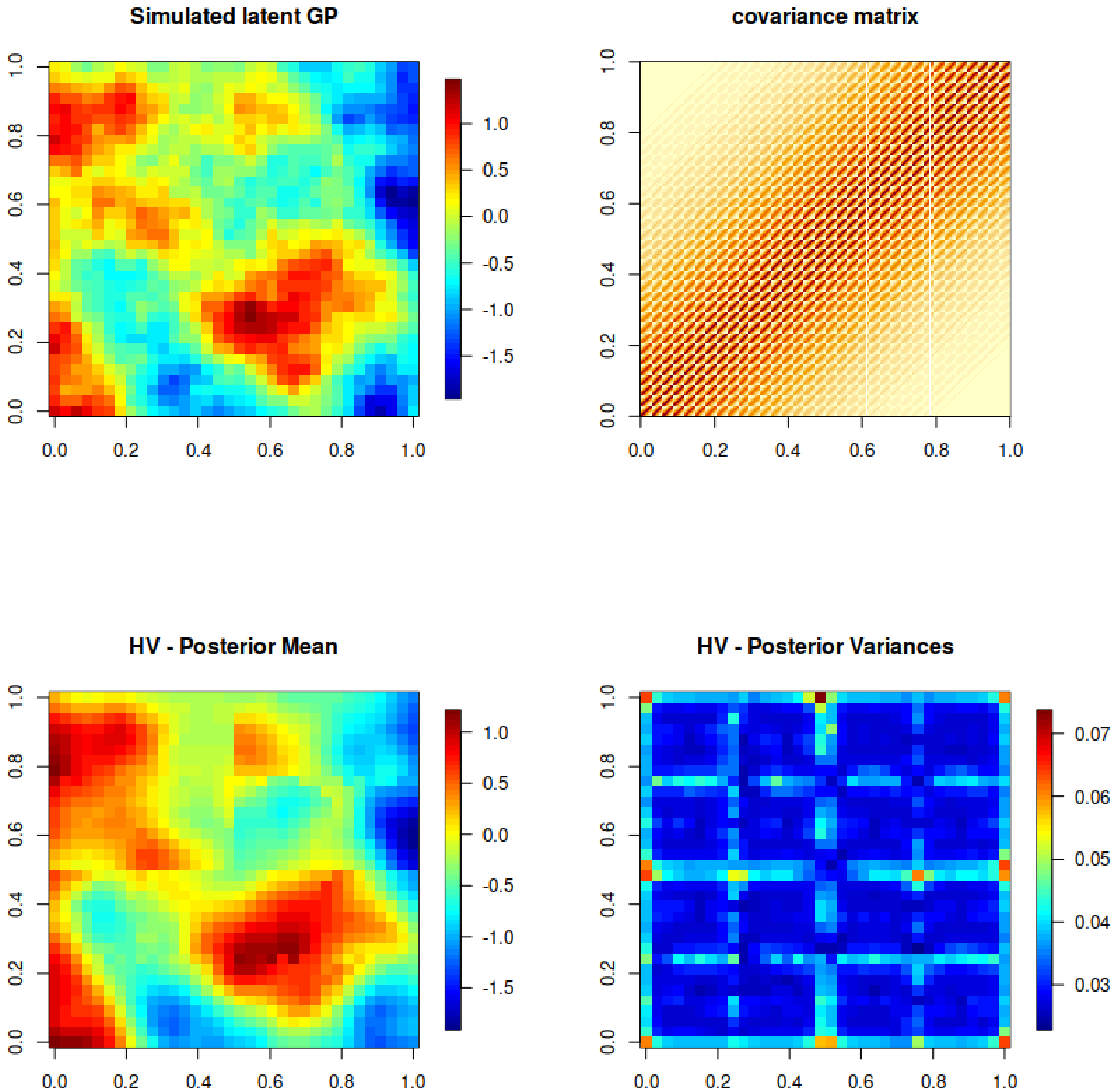
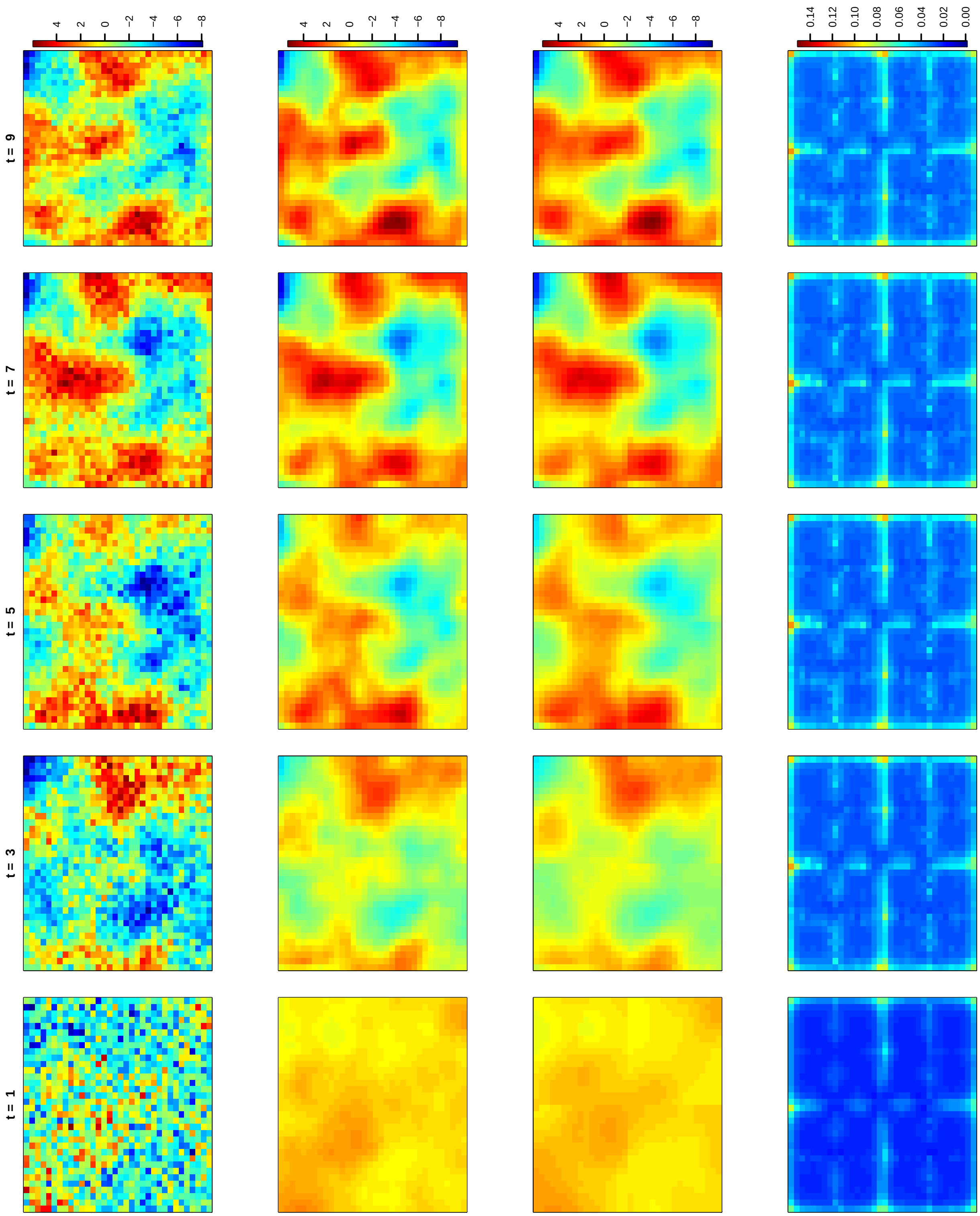


Figure S5: Spatial field reconstructed using the HV approximation as in Algorithm 2. The true field was simulated using a Matérn covariance function with smoothness 1.5, range 0.2 and marginal variance 1.0. Observations are as in (1) with $\tau_i^2 = 0.5$. For the HV approximation, we used $N = 79$, with the other parameters being equal to the default values selected by the `GPvecchia` package. Using the notation from Section 6, these were $M = 6$, $J = 2$, and $|\mathcal{X}_{j_1,\dots,j_m}| = 12, 12, 12, 12, 13, 13$, respectively, for $m = 0, 1, \dots, M-1$, and $|\mathcal{X}_{j_1,\dots,j_M}| \leq 5$.



(a) data (b) latent (c) filtering mean (d) std. deviation

Figure S6: Filtering for the advection-diffusion model from Section 6.3. The edge effects in the filtering mean, visible at $t = 1$, disappear at further timepoints. The artifacts in uncertainty quantification persist, but become slightly less pronounced over time.

S6 Further details of the data application

For the satellite-data analysis in Section 7, in order to find appropriate parameter values, we first conducted an exploratory study considering each day as an independent, purely spatial problem. We assume model (1) - (2) with covariance matrix $\mathbf{\Sigma}$ derived from a Matérn covariance function with smoothness 1.5, range λ_t and marginal variance σ_t^2. Using the HV approximation with conditioning sets of size $N = 80$, we then estimated the parameters separately for each day by maximizing the approximate likelihood. The estimation results are shown in Table S1.

t	1	2	3	4	5	6	7	8	9	10	11
σ^2	18.7	46.46	45	45.9	21.1	43.9	31.1	46.9	51.7	0.75	19.7
λ	2.017	2.324	2.727	2.191	1.866	2.325	1.923	2.659	2.319	0.716	1.932

Table S1: Spatial parameter estimates for the satellite data at 11 time points

We calculated the average marginal variance $\bar{\sigma}^2$ and range $\bar{\lambda}$, and used them as corresponding parameters in the covariance function determining $\mathbf{\Sigma}_{0|0}$. We used the same function to generate the model-error covariance $\mathbf{Q}_t$ matrix, except that we take the marginal variance to be $\sigma^2_{\mathbf{Q}} = \frac{1}{1-\alpha^2}\bar{\sigma}^2$.

We also investigated the quality of the approximation of the uncertainty measured by pointwise standard deviations of the filtering distribution. The heatmaps showing these results are presented in Figure S7 for selected time points. Both methods under consideration (i.e., the LR filter and the HV filter) exhibited undesirable artifacts in the spatial pattern of the uncertainty; for HV, these were visible partition boundaries, whereas LR resulted in visible knots with circular pattern around them. The size of this distortion is difficult to evaluate precisely, as the true values are unknown. However, it seems that the magnitude of the underestimation of uncertainty in the case of the LR filter might be greater than its overestimation in the case of the HV.

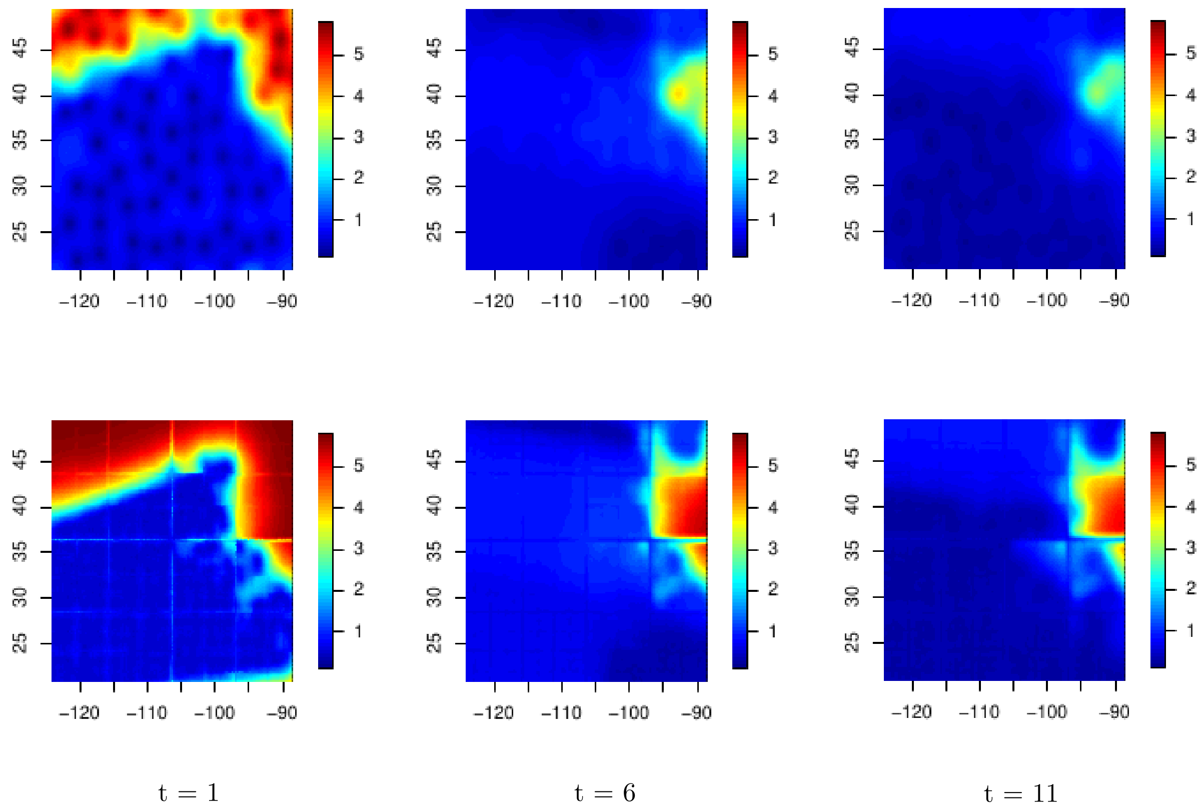


Figure S7: Standard errors of TPW filtering results at selected time points. The top row corresponds to the low-rank filter and the bottom row to the HV filter.

References

DelSole, T. and Yang, X. (2010). State and parameter estimation in stochastic dynamical models. *Physica D: Nonlinear Phenomena*, 239(18):1781–1788.

Douc, R., Cappé, O., and Moulines, E. (2005). Comparison of resampling schemes for particle filtering. *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69.

Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In Crisan, D. and Rozovskii, B., editors, *The Oxford Handbook of Nonlinear Filtering*, number December, pages 656–704. Oxford University Press.

Jurek, M. and Katzfuss, M. (2021). Multi-resolution filters for massive spatio-temporal data. *Journal of Computational and Graphical Statistics*.

Katzfuss, M., Stroud, J. R., and Wikle, C. K. (2016). Understanding the ensemble Kalman filter. *The American Statistician*, 70(4):350–357.

Katzfuss, M., Stroud, J. R., and Wikle, C. K. (2020). Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. *Journal of the American Statistical Association*, 115(530):866–885.

Lorenz, E. N. (2005). Designing chaotic models. *Journal of the Atmospheric Sciences*, 62(5):1574–1587.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., Kalnay, E., Patil, D. J., and Yorke, J. A. (2004). A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, 56:415–428.