Prequential posteriors

Shreya Sinha-Roy
Department of Statistics, University of Warwick
shreya.sinha-roy@warwick.ac.uk

Joint work with *Dr. Ritabrata Dutta*, *Dr. Richard Everitt and Prof Christian Robert*

One World Approximate Bayesian Inference (OWABI) seminar, Feb 26, 2026

Data assimilation (DA)

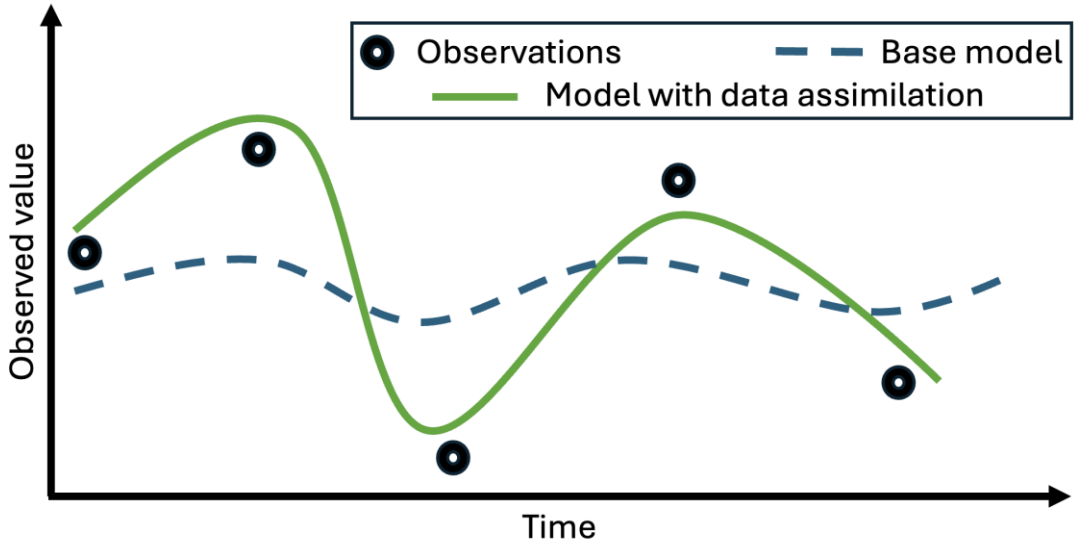


Fig 1: Data assimilation models (green) are helped by observations to produce more realistic forecasts.

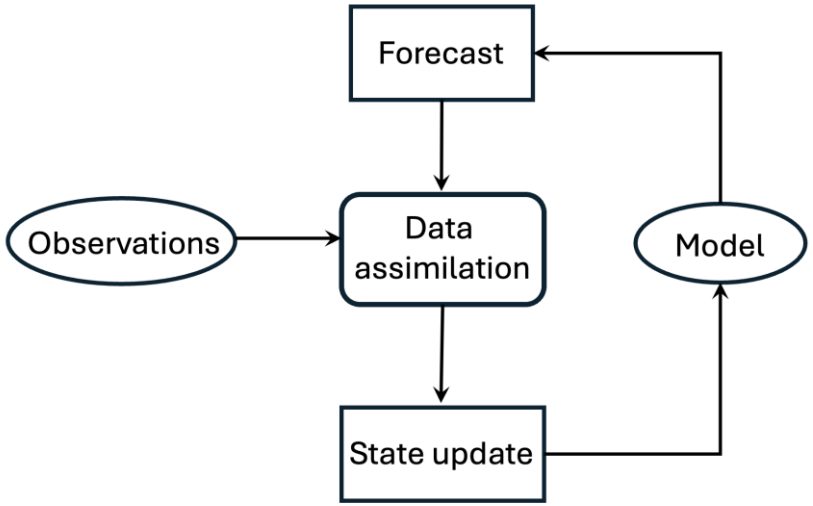


Fig 2: A flowchart explaining model update in data assimilation

Data assimilation (DA)

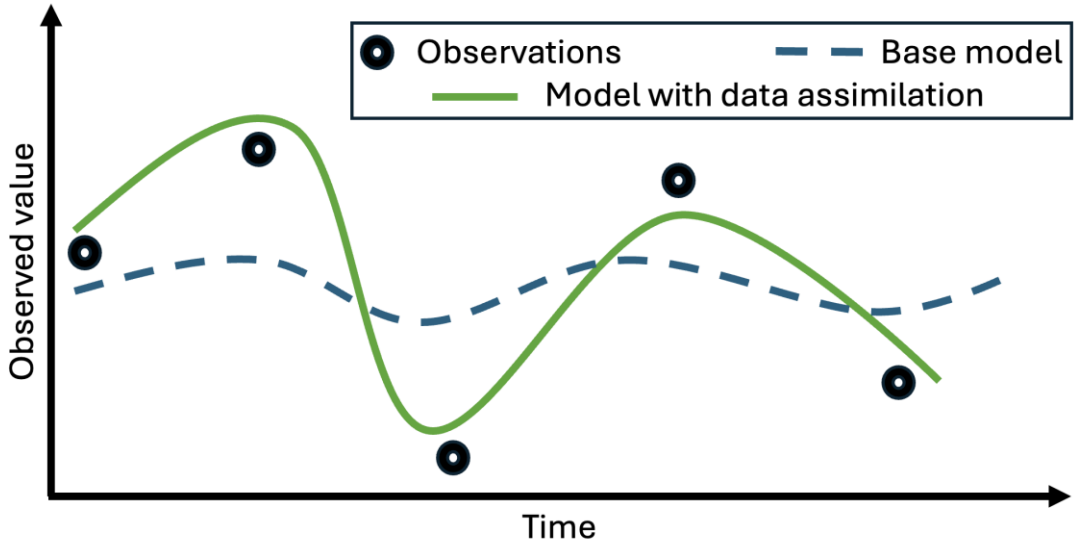


Fig 1: Data assimilation models (green) are helped by observations to produce more realistic forecasts.

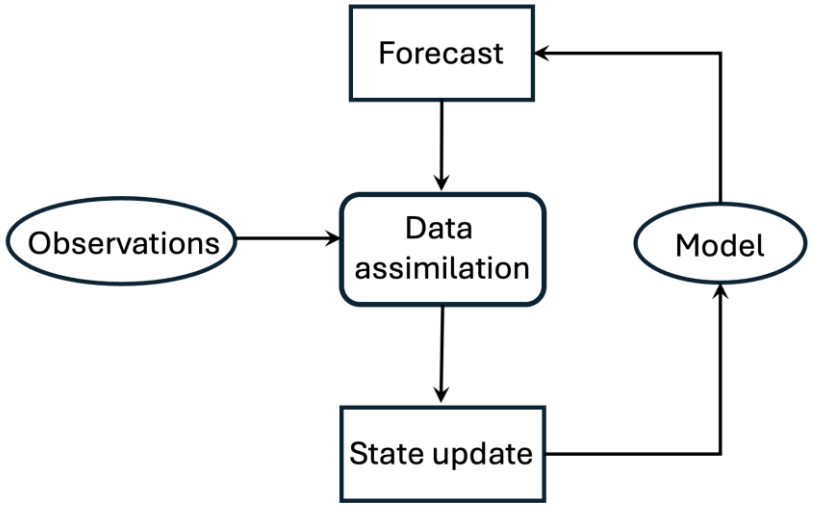


Fig 2: A flowchart explaining model update in data assimilation

- Traditional approach: Kalman filters (eg, KF, EnKF, EKF, Gen Bayes KF etc) assuming a state-space model (SSM).
- Recently deep generative forecasting models (DGFM) like GenCast (2024) etc. have gained popularity.

Data assimilation (DA)

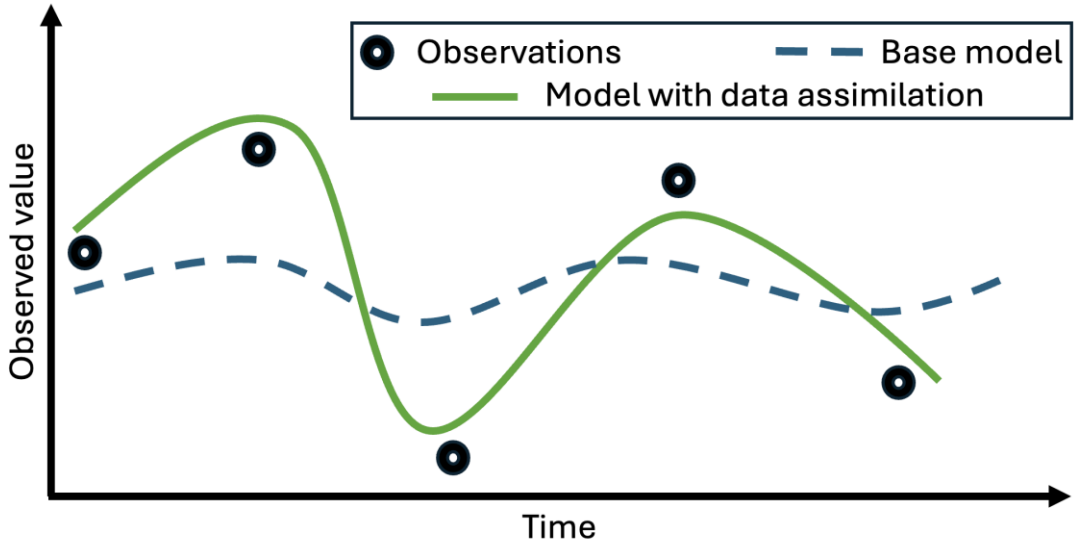


Fig 1: Data assimilation models (green) are helped by observations to produce more realistic forecasts.

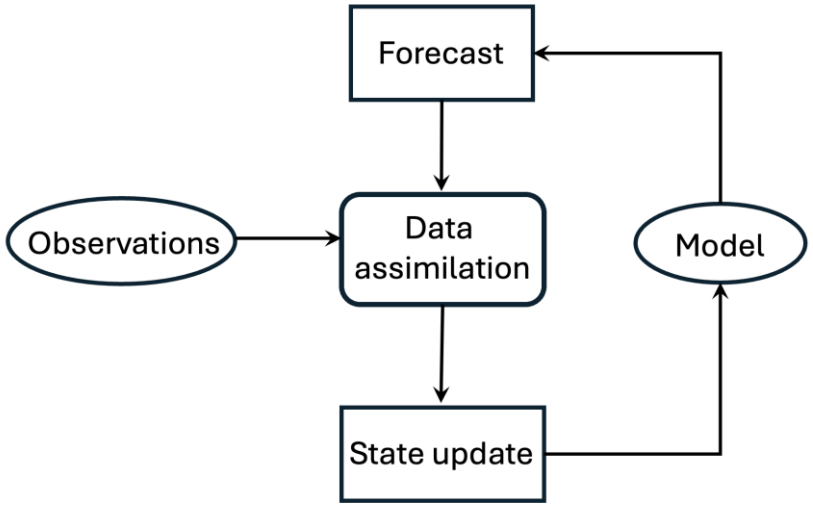


Fig 2: A flowchart explaining model update in data assimilation

- Traditional approach: Kalman filters (eg, KF, EnKF, EKF, Gen Bayes KF etc) assuming a state-space model (SSM).
- Recently deep generative forecasting models (DGFM) like GenCast (2024) etc. have gained popularity.

Model	DA
SSM	✓
DGFM	✗

What is a DGFM ?

Deep generative forecasting models (DGFMs)

- We observe a temporal process up to time $t-1 \rightarrow Y^{t-1} = (Y_1, Y_2, \ldots Y_{t-1}) \sim \mathbb{P}$.
- We want a probability distribution for $Y_t \rightarrow$ *'probabilistic forecasting'*

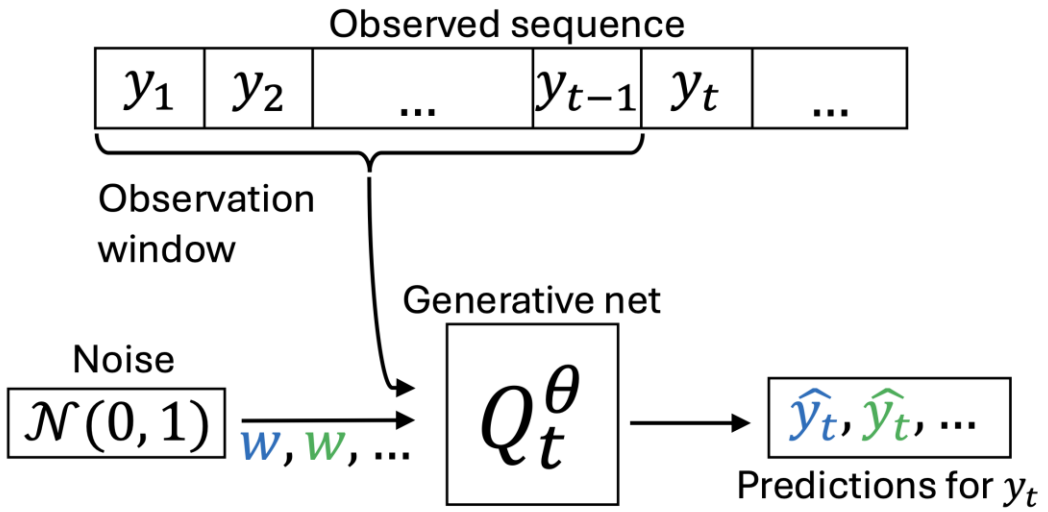


Fig 3: Q_t^θ is a conditional generative model which predicts y_t given the past observations $y_1, y_2, \cdots y_{t-1}$.

Deep generative forecasting models (DGFMs)

- We observe a temporal process up to time $t-1 \rightarrow Y^{t-1} = (Y_1, Y_2, \dots Y_{t-1}) \sim \mathbb{P}$.
- We want a probability distribution for $Y_t \rightarrow$ *'probabilistic forecasting'*

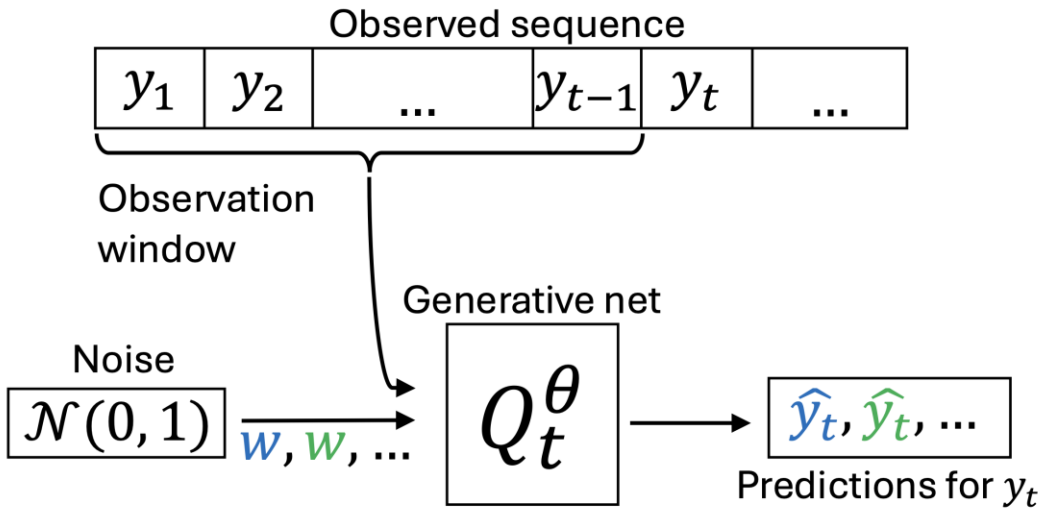


Fig 3: Q_t^θ is a conditional generative model which predicts y_t given the past observations $y_1, y_2, \cdots y_{t-1}$.

- Our simulator model $\mathbb{Q}^\theta$ is defined by a class of conditional probability distributions which produces the forecast $\hat{Y}_t$ conditioned on Y^{t-1} for $t = 1, 2, \dots$

$$\hat{Y}_t \sim Q_t^\theta, \qquad Q_t^\theta = \mathbb{Q}^\theta(\cdot | Y^{t-1})$$

How to train a DGFM ?

- Given the observations $Y^T = (Y_1, Y_2, \dots Y_T)$, we define the (predictive sequential) prequential loss,

$$L_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta)$$

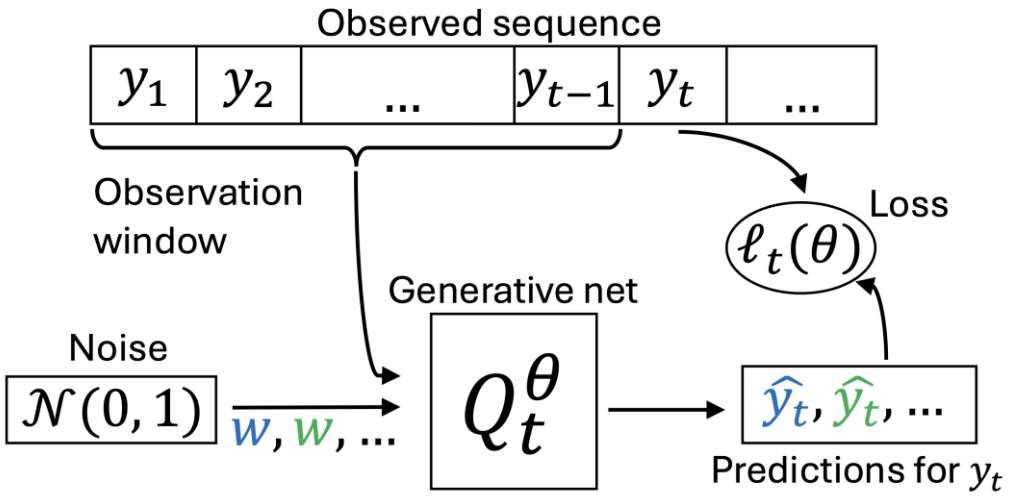


Fig 4: The loss ℓ_t^θ is calculated between the predictions from the model Q_t^θ and the observed y_t.

❑ Pacchiardi, L., Adewoyin, R. A., Dueben, P., & Dutta, R. (2024). Probabilistic forecasting with generative networks via scoring rule minimization. Journal of Machine Learning Research, 25(45), 1-64

How to train a DGFM ?

- Given the observations $Y^T = (Y_1, Y_2, \dots Y_T)$, we define the (predictive sequential) prequential loss,

$$L_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta)$$

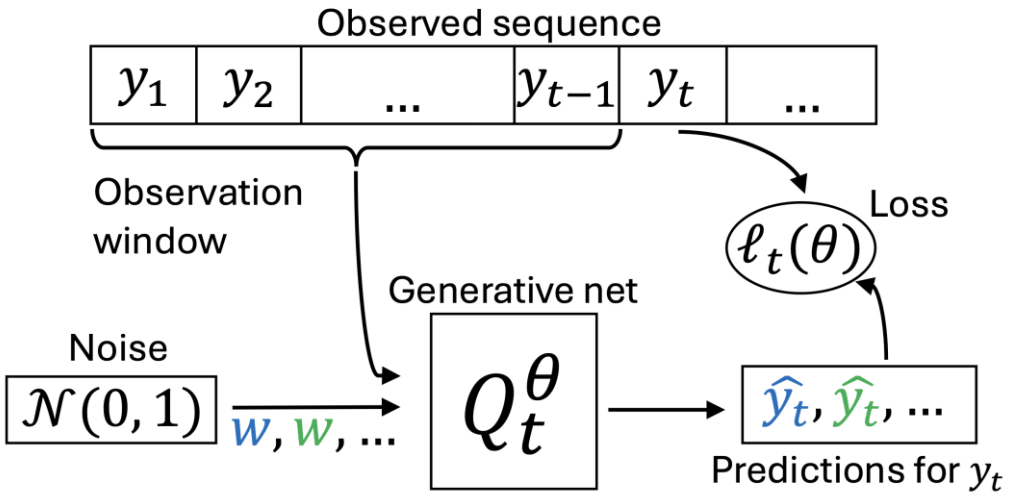


Fig 4: The loss ℓ_t^θ is calculated between the predictions from the model Q_t^θ and the observed y_t.

Prequential loss minimizer: $\hat{\theta}_T = arg\,min_{\theta \in \Theta}\ L_T(\theta)$

✓ Outlier-robust adversarial-free training

Can we prove consistency property for $\hat{\theta}_T$?

❑ Pacchiardi, L., Adewoyin, R. A., Dueben, P., & Dutta, R. (2024). Probabilistic forecasting with generative networks via scoring rule minimization. Journal of Machine Learning Research, 25(45), 1-64

Marginal versus conditional expected loss

Traditional approach: Marginal expected loss

$$\tilde{L}_T(\theta) = \sum_{t=1}^{T} \mathbb{E}\ell_t(\theta) = \sum_{t=1}^{T} \int \ell_t(\theta)\, p(y)\, dy$$

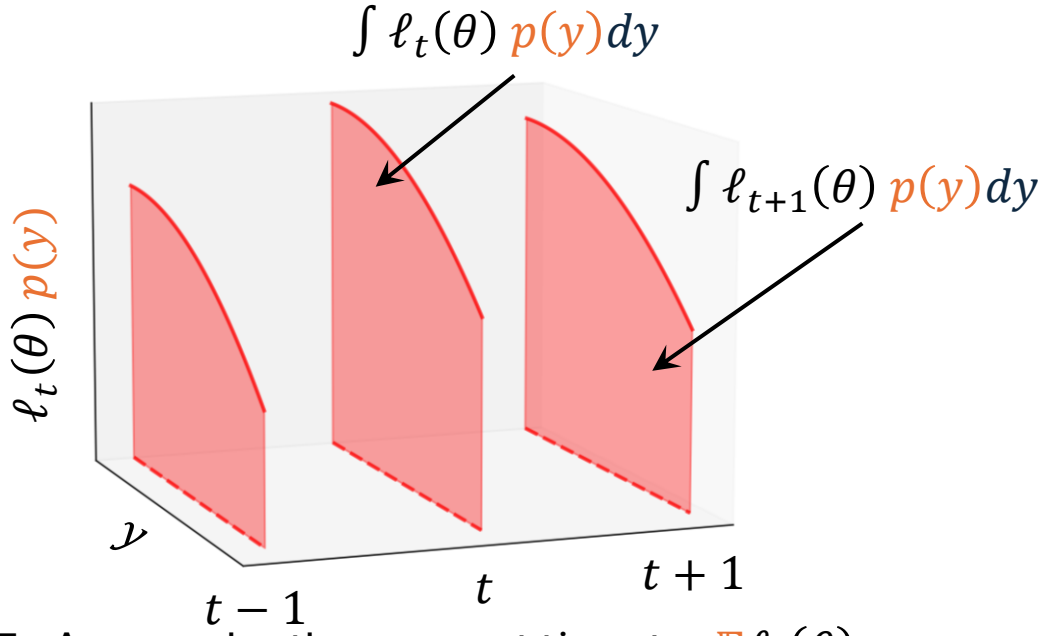


Fig 5 : Area under the curve at time t = $\mathbb{E}\ell_t(\theta)$.

Marginal versus conditional expected loss

Traditional approach: Marginal expected loss

$$\tilde{L}_T(\theta) = \sum_{t=1}^{T} \mathbb{E}\ell_t(\theta) = \sum_{t=1}^{T} \int \ell_t(\theta)\, p(y)\, dy$$

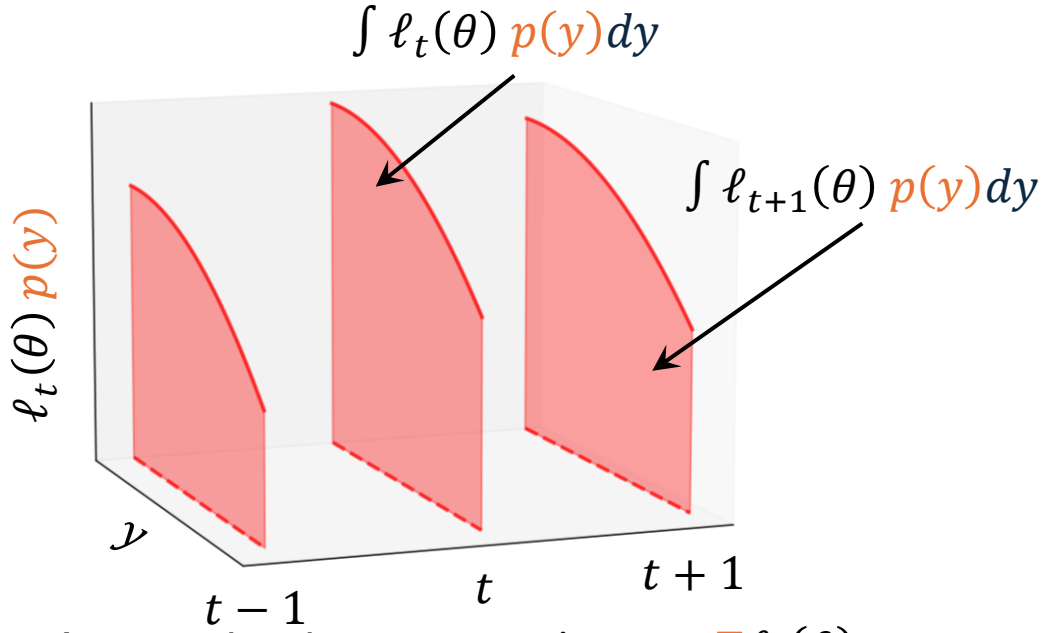


Fig 5 : Area under the curve at time t = $\mathbb{E}\ell_t(\theta)$.

- ✓ Consistency through ULLN.

Marginal versus conditional expected loss

Traditional approach: Marginal expected loss

$$\tilde{L}_T(\theta) = \sum_{t=1}^{T} \mathbb{E}\ell_t(\theta) = \sum_{t=1}^{T} \int \ell_t(\theta)\, p(y)\, dy$$

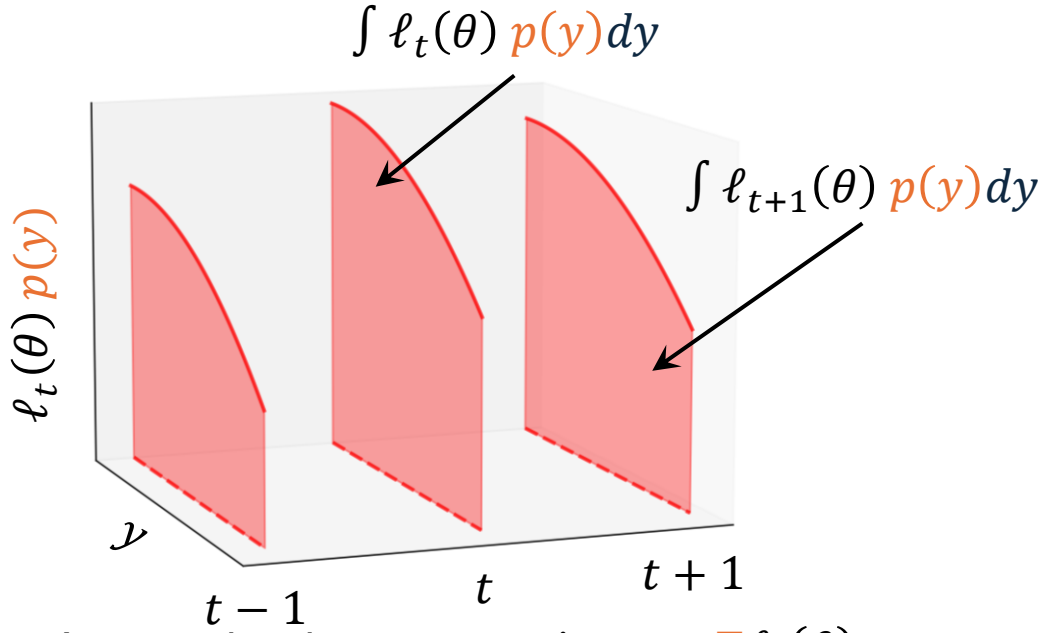


Fig 5 : Area under the curve at time t = $\mathbb{E}\ell_t(\theta)$.

- ✓ Consistency through ULLN.

Predictive ~~Traditional~~ approach: ~~Marginal~~ expected loss

Conditional

$$L_t^*(\theta) = \sum_{t=1}^{T} \mathbb{E}_{t-1}\ell_t(\theta) = \sum_{t=1}^{T} \int \ell_t(\theta)\, p_t(y|\, y_{1:t-1})\, dy$$

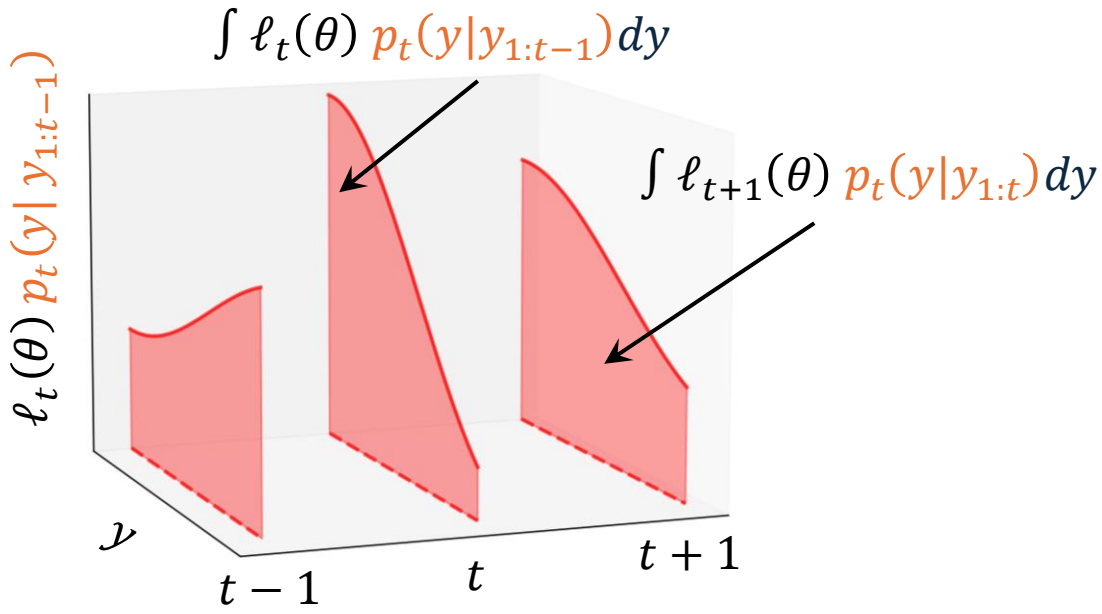


Fig 6 : Area under the curve at time t = $\mathbb{E}_{t-1}\ell_t(\theta)$.

Marginal versus conditional expected loss

Traditional approach: Marginal expected loss

$$\tilde{L}_T(\theta) = \sum_{t=1}^{T} \mathbb{E}\ell_t(\theta) = \sum_{t=1}^{T} \int \ell_t(\theta)\, p(y)\, dy$$

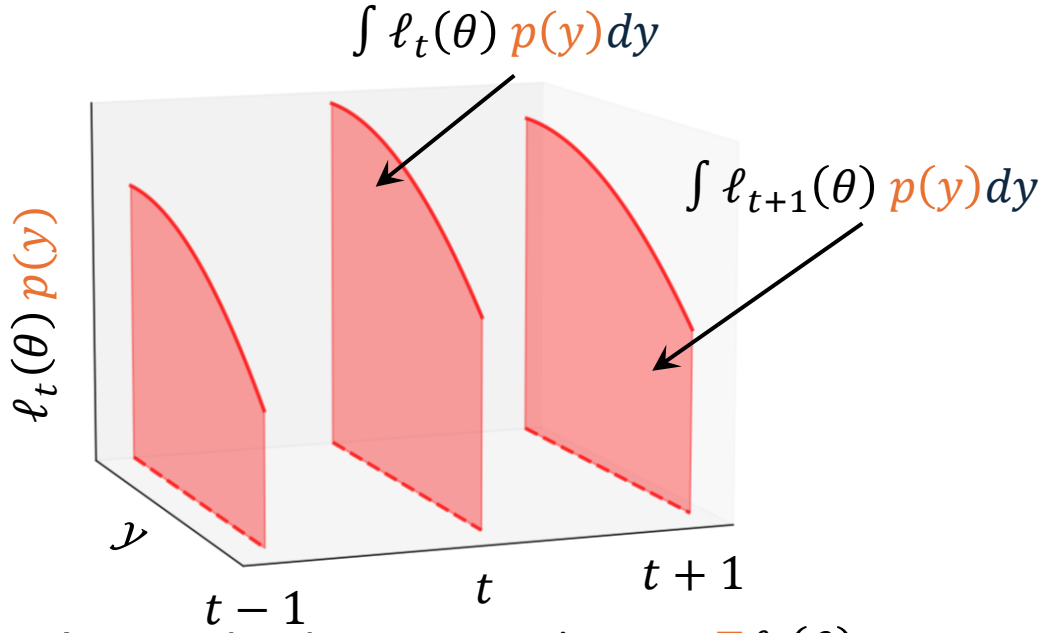


Fig 5 : Area under the curve at time t = $\mathbb{E}\ell_t(\theta)$.

✓ Consistency through ULLN.

Predictive ~~Traditional~~ approach: ~~Marginal~~ expected loss

Conditional

$$L_t^*(\theta) = \sum_{t=1}^{T} \mathbb{E}_{t-1}\ell_t(\theta) = \sum_{t=1}^{T} \int \ell_t(\theta)\, p_t(y|\, y_{1:t-1})\, dy$$

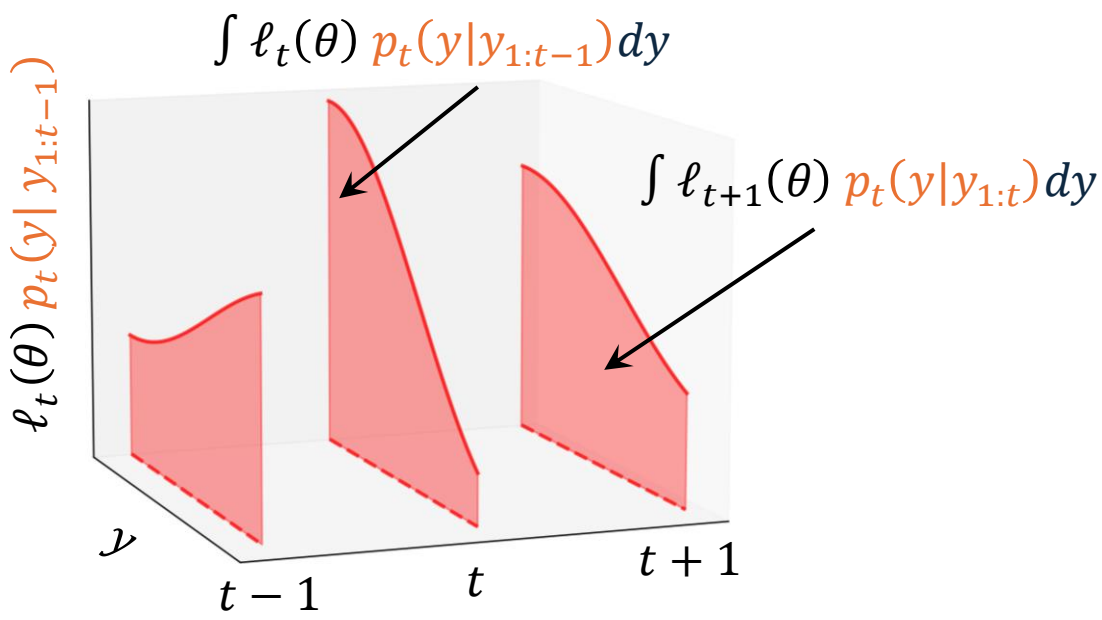


Fig 6 : Area under the curve at time t = $\mathbb{E}_{t-1}\ell_t(\theta)$.

Consistency ??

Prediction focused view of consistency via conditional expectation

Lemma 1 : [Martingale uniform law of large numbers]
Under some regularity conditions, a uniform law of large numbers (ULLN) for martingales holds which implies with probability one under $\mathbb{P}$,

$$sup_{\theta \in \Theta} \frac{1}{A_T} |L_T(\theta) - L_T^*(\theta)| \to 0 \text{ as } T \to \infty.$$

where A_T is a normalizing constant.

- Skouras, K., Dawid, A. (2000), Consistency in misspecified models. Research Report No.218, Department of Statistical Science, University College London

Prediction focused view of consistency via conditional expectation

Lemma 1 : [Martingale uniform law of large numbers]
Under some regularity conditions, a uniform law of large numbers (ULLN) for martingales holds which implies with probability one under $\mathbb{P}$,

$$sup_{\theta \in \Theta} \frac{1}{A_T} |L_T(\theta) - L_T^*(\theta)| \to 0 \text{ as } T \to \infty.$$

where A_T is a normalizing constant.

$$d\left(\hat{\theta}_T, \theta_T^*\right) \to 0, \text{with probability one under } \mathbb{P}, \text{ where } \theta_T^* = arg\ min_{\theta \in \Theta} L_T^*(\theta)$$

$\hat{\theta}_T$ converges to a data dependent limit which minimizes the conditional predictive risk.

- Skouras, K., Dawid, A. (2000), Consistency in misspecified models. Research Report No.218, Department of Statistical Science, University College London

Data assimilation for DGFM?

Data assimilation for DGFMs?

- How to define a posterior distribution on the parameter space when the likelihood function for the simulator model is not tractable?

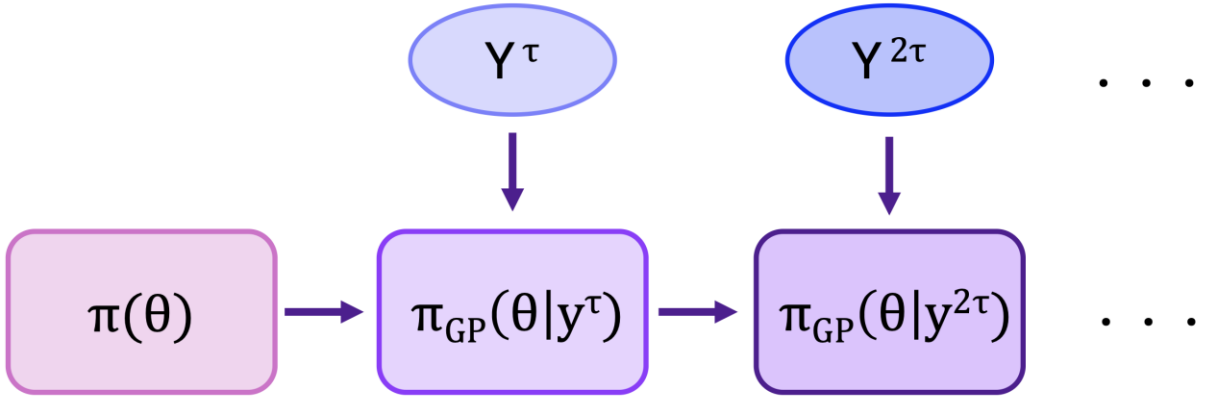


Fig 7: Data assimilation through posterior update.

??

posteriors for DGFMs

Generalized Bayesian posterior

$$\pi_G(\theta|y) \propto \pi(\theta)e^{-\gamma \cdot \ell(\theta,y)}$$

$\ell(\theta, y) \rightarrow$ Loss
$\gamma > 0 \rightarrow$ Scale of ℓ wrt the prior

Prequential loss for DGFM

Can simulate data: $x \sim \mathbb{Q}^{\theta}$

But cannot evaluate $p(\, y \mid \theta\,) \rightarrow$ **standard posterior not available**

Prequential posteriors for DGFMs

Generalized Bayesian posterior

$$\pi_G(\theta|y) \propto \pi(\theta)e^{-\gamma \cdot \ell(\theta,y)}$$

$\ell(\theta, y) \rightarrow$ Loss
$\gamma > 0 \rightarrow$ Scale of ℓ wrt the prior

Prequential loss for DGFM

Can simulate data: $x \sim \mathbb{Q}^{\theta}$

But cannot evaluate $p(\, y \mid \theta\,) \rightarrow$ **standard posterior not available**

Prequential posterior : $\pi_P(\theta|y^T) \propto \pi(\theta)e^{-\gamma L_T(\theta)}$

where $\pi(\theta)$ is a prior distribution on the model parameters and γ is a free parameter.

❑ Sinha-Roy, S., Everitt, R. G., Robert, C. P., Dutta, R. (2025). Prequential posteriors. Preprint arXiv:2511.17721.

Bernstein-von Mises (BvM) theorem for the prequential posterior

- Suppose, there exists a function $L(\theta)$ such that as $T \to \infty$ we have, $\frac{1}{A_T} L_T^*(\theta) \to L(\theta)$ uniformly with probability one under $\mathbb{P}$.

$$\theta^* = \ arg\, min_{\theta \in \Theta}\ L(\theta)$$

Bernstein-von Mises (BvM) theorem for the prequential posterior

- Suppose, there exists a function $L(\theta)$ such that as $T \to \infty$ we have, $\frac{1}{A_T} L_T^*(\theta) \to L(\theta)$ uniformly with probability one under $\mathbb{P}$.

$$\theta^* = \ arg\, min_{\theta \in \Theta}\ L(\theta); \qquad d\left(\hat{\theta}_T, \theta^*\right) \to 0\ as\ T \to \infty$$

Theorem 1: [BvM theorem for prequential posterior]

Under some regularity conditions, letting q_T be the density of $\sqrt{A_T}\ (\theta\ - \hat{\theta}_T)$ when $\theta \sim \pi_{GP}(\theta\ |y^T)$, we have with probability one under $\mathbb{P}$,

$$\int_\Theta |q_T(\theta) - \ \mathcal{N}(\theta|\ 0, H^{-1})| d\theta\ \to 0 \text{ as } T \to \infty.$$

where $\frac{1}{A_T} L_T''\ (\theta^*) \to H$, as $T \to \infty$.

Example loss function: Proper scoring rules

- Examples of some loss functions: Proper scoring rules like log score, energy score, kernel score etc.
- Energy score:

$$S(\mathbb{Q}^\theta, y) = 2 \cdot E_{\mathbb{Q}^\theta} \left|\left|X - y\right|\right|_2^\beta - E_{\mathbb{Q}^\theta} \left|\left|X - X'\right|\right|_2^\beta; \ \mathrm{X}, \mathrm{X}' \sim \mathbb{Q}^\theta, \beta \in (0, 2).$$

- Calculating the expectation is difficult but we can simulate from the model $\mathbb{Q}^\theta$: $x \sim \mathbb{Q}^\theta$.
- We use these simulations to unbiasedly estimate the score as,

$$\hat{S}(\mathbb{Q}^\theta, y) = \frac{2}{m} \sum_{i=1}^{m} \left|\left|x_i - y\right|\right|_2^\beta - \frac{1}{m(m-1)} \sum_{i,j=1}^{m} \left|\left|x_i - x_j\right|\right|_2^\beta$$

where, $x_1, x_2, \ldots x_m \sim \mathbb{Q}^\theta$

Sequential Inference ?

Inference via sequential Monte Carlo (SMC)

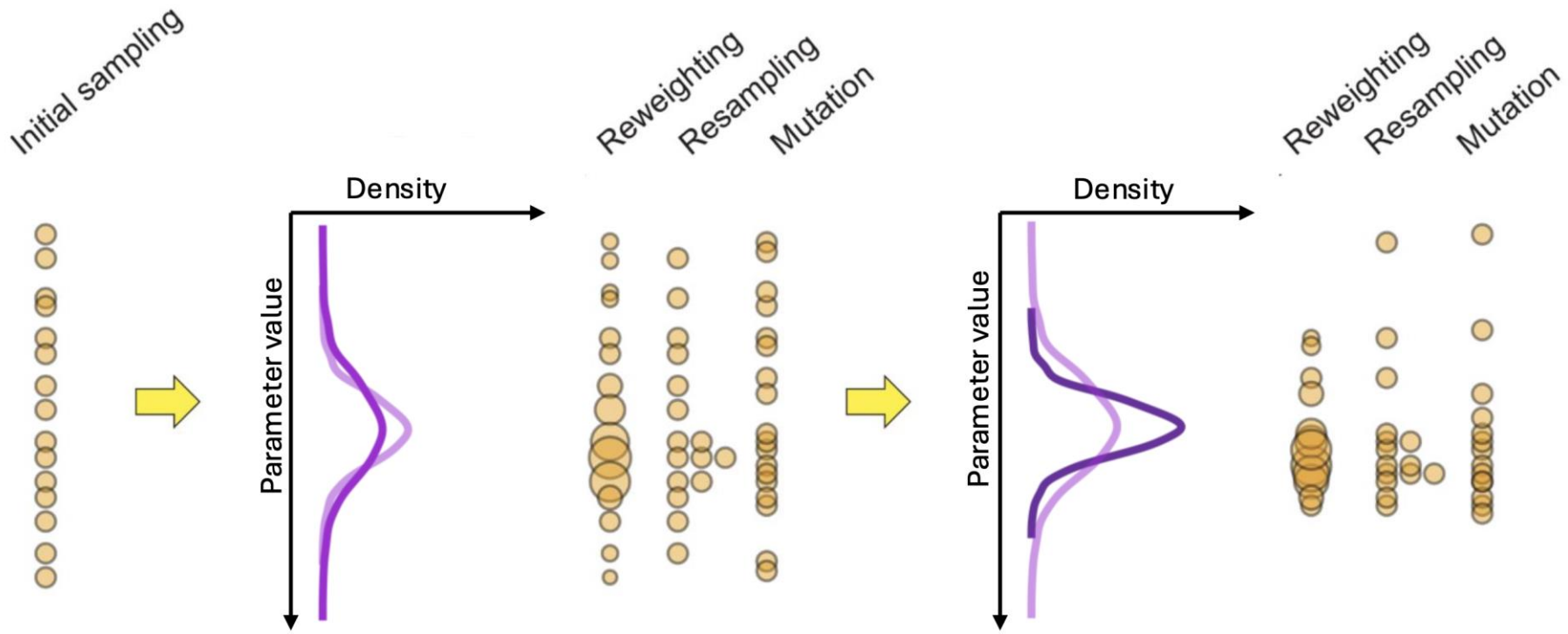


Fig 8: We use SMC to propagate the particles from proposal distribution(pink) to the target posterior (purple). We define intermediate target distributions (curve in lighter shade) by tempering the target.

Inference via sequential Monte Carlo (SMC)

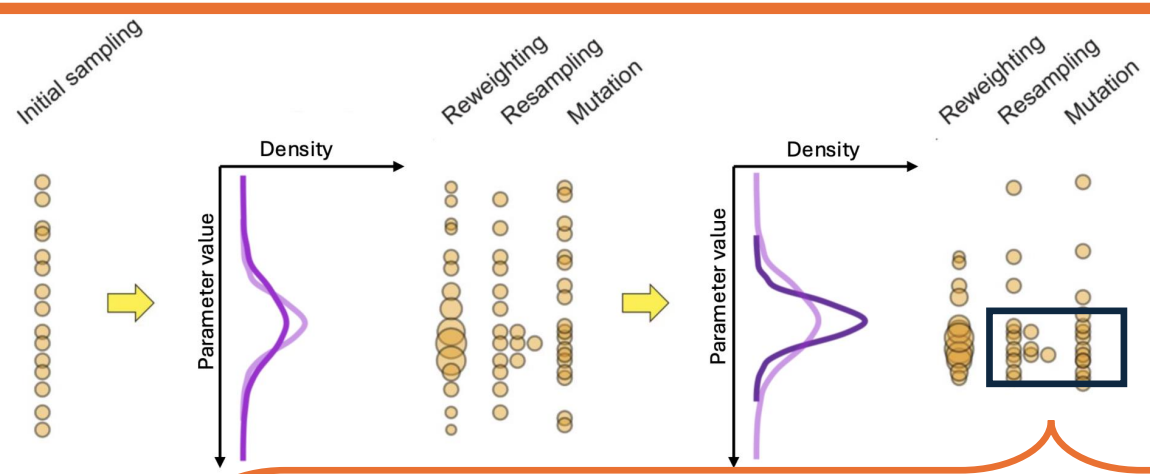


Fig 8: We use SMC to propagate the particles from proposal distribution(pink) to the target posterior (purple). We define intermediate target distributions (curve in lighter shade) by tempering the target.

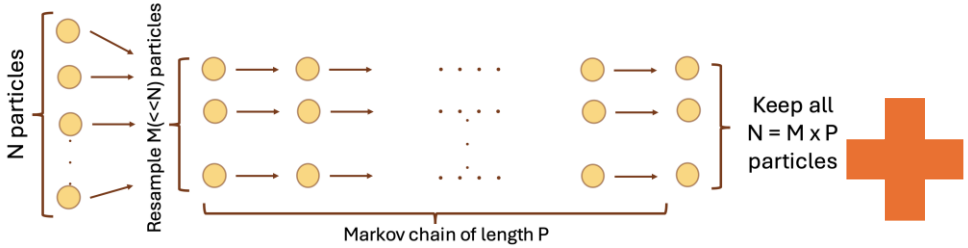


Fig 9: Mutation of the resampled particles in a wastefree style using a preconditioned forward kernel

Algorithm 1: Preconditioned forward kernel

Input: Learning rate η, parameters σ, λ, noise $\{\zeta_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)\}$

Initialize: $\boldsymbol{\theta}_0$, $\mathbf{u}_0 = \sqrt{\eta} \cdot \mathcal{N}(0, \mathbf{I}_p)$, $\boldsymbol{\alpha}_0 = \sqrt{\eta} \cdot C$, $\mathbf{v}_0 = \mathbf{0}$

for $n = 1, 2, \ldots P$ **do**

Compute stochastic gradient: $\tilde{\mathbf{f}}_n \leftarrow \nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}_{n-1})$;

$\mathbf{v}_n \leftarrow \sigma \mathbf{v}_{n-1} + \frac{1-\sigma}{T^2} \tilde{\mathbf{f}}_n \odot \tilde{\mathbf{f}}_n$;

$\mathbf{g}_n \leftarrow 1 \oslash \sqrt{\lambda + \sqrt{\mathbf{v}_n}}$;

$\boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_{n-1} + \frac{1}{2} \mathbf{g}_n \odot \mathbf{u}_{n-1}$;

$\boldsymbol{\alpha}_n \leftarrow \boldsymbol{\alpha}_{n-1} + \frac{1}{2} (\mathbf{u}_{n-1} \odot \mathbf{u}_{n-1} - \eta)$;

$\mathbf{u}_n \leftarrow \exp(-\boldsymbol{\alpha}_n / 2) \odot \mathbf{u}_{n-1}$;

$\mathbf{u}_n \leftarrow \mathbf{u}_n - \eta \cdot \mathbf{g}_n \odot \tilde{\mathbf{f}}_n + \sqrt{2 \cdot \mathbf{g}_{n-1} \eta^{3/2}} \odot \zeta_n$;

$\mathbf{u}_n \leftarrow \exp(-\boldsymbol{\alpha}_n / 2) \odot \mathbf{u}_n$;

$\boldsymbol{\alpha}_n \leftarrow \boldsymbol{\alpha}_n + \frac{1}{2} (\mathbf{u}_n \odot \mathbf{u}_n - \eta)$;

$\boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_n + \frac{1}{2} \mathbf{g}_n \odot \mathbf{u}_n$;

end

- Dau, H.D., Chopin N. Waste-free sequential Monte Carlo. JRSS B: Statistical Methodology, 84(1):114–148, 2022
- Chen C. et al. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In Artificial Intelligence and Statistics,1051–1060. PMLR, 2016

Prequential posteriors in practice

Time series data

Lorenz 96

- Toy representation of atmospheric behaviour with fast (x) and slow (y) evolving variables.

The original process

$$\dot{y}(k) = -y(k-1)\big(y(k-2) - y(k+1)\big) - y(k) + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} x(j),$$

$$\dot{x}(j) = -cbx(j+1)\big(x(j+2) - x(j-1)\big) - cx(j) + \frac{hc}{b}\, y\left(\frac{j-1}{J} + 1\right);$$

here $y(k)$ and $x(j)$ denote the k th and the j th component of the corresponding variables for $k = 1, 2, \ldots K$; $j = 1, 2, \ldots JK$; with $K = 8, J = 32, h = 1, b = 10, c = 10, F = 20$.

We integrate the system using fourth order Runge-Kutta (RK4) method and **keep only the y values as observations.**

Time series data

Lorenz 96

- Toy representation of atmospheric behaviour with slow (y) and fast evolving variables.
- **DGFM**: A single-layer generative gated recurrent unit (GRU) with last 10 values as covariate.
- Energy score (proper scoring rule) to define the prequential Posterior.
- We define an episode of length 100 observations and update the posterior at every episode.

Time series data

Lorenz 96

- Toy representation of atmospheric behaviour with slow (y) and fast evolving variables.
- **DGFM**: A single-layer generative gated recurrent unit (GRU) with last 10 values as covariate.
- Energy score (proper scoring rule) to define the prequential Posterior.
- We define an episode of length 100 observations and update the posterior at every episode.

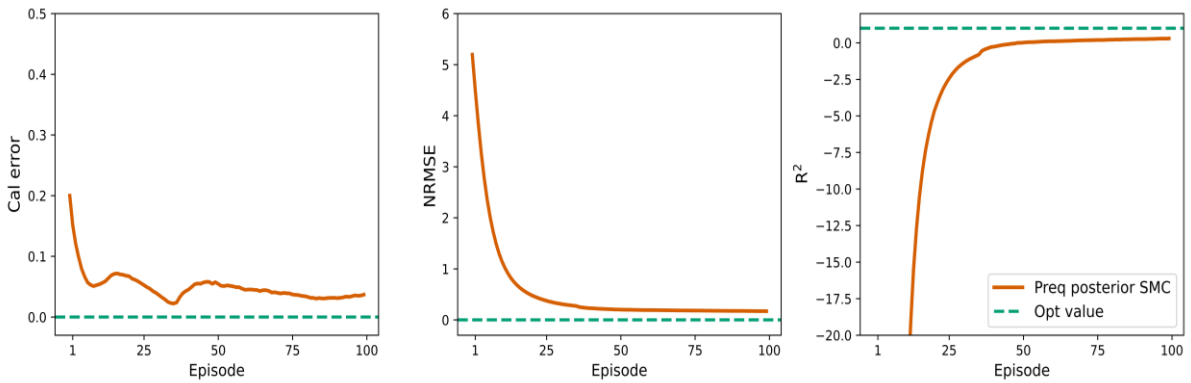


Fig 10: We measure the predictive accuracy of the posterior predictive (orange line) after every episodic update in terms of calibration error, normalised RMSE (NRMSE), and R^2 calculated on a test dataset of length 2000. For reference, the maximum achievable values of the metrics are shown in green dashed line.

- **Successful data assimilation** : We can see improved predictive performance of the posteriors over training episodes.

Lorenz 96:
Preq posterior + SMC vs Kalman filter

The original process

$$\dot{y}(k) = -y(k-1)\big(y(k-2) - y(k+1)\big) - y(k) + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} x(j)$$

$$\dot{x}(j) = -cbx(j+1)\big(x(j+2) - x(j-1)\big) - cx(j) + \frac{hc}{b}\, y\left(\frac{j-1}{J} + 1\right)$$

with $K = 8,\ J = 32,\ h = 1,\ b = 10,\ c = 10,\ F = 20.$

State space model (SSM) representation

$$\dot{v}(k) = -v(k-1)\big(v(k-2) - v(k+1)\big) - v(k) + F_k \quad \text{[latent]}$$

$$y(k) = v(k) + \psi_k \quad \text{[noisy obs]}$$

with $F_k \sim \mathcal{N}(20, 1)$, $\psi_k \sim \mathcal{N}(0, 1)$.

Lorenz 96:
Preq posterior + SMC vs Kalman filter

The original process

$$\dot{y}(k) = -y(k-1)\big(y(k-2) - y(k+1)\big) - y(k) + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} x(j)$$

$$\dot{x}(j) = -cbx(j+1)\big(x(j+2) - x(j-1)\big) - cx(j) + \frac{hc}{b}\, y\left(\frac{j-1}{J} + 1\right)$$

with $K = 8,\ J = 32,\ h = 1,\ b = 10,\ c = 10,\ F = 20.$

State space model (SSM) representation

$$\dot{v}(k) = -v(k-1)\big(v(k-2) - v(k+1)\big) - v(k) + F_k \quad \text{[latent]}$$

$$y(k) = v(k) + \psi_k \quad \text{[noisy obs]}$$

with $F_k \sim \mathcal{N}(20, 1)$, $\psi_k \sim \mathcal{N}(0, 1)$.

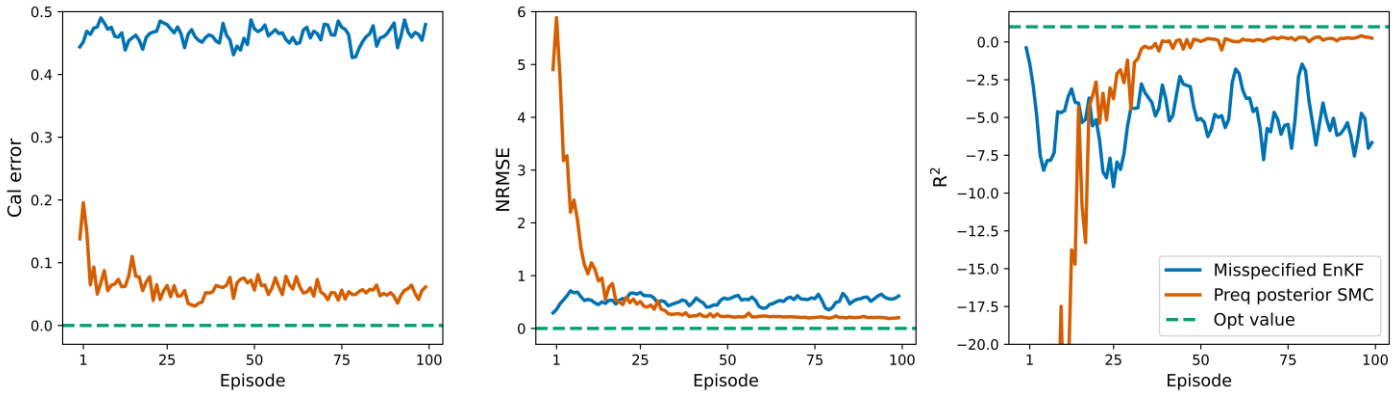


Fig 11: Posterior predictive performance across episodes for the Lorenz 96 model. **The prequential posterior (orange curve) improves with additional data, while the misspecified EnKF (blue curve) shows little change**, as measured by calibration error, normalised RMSE, and coefficient of determination R^2. The diagnostics are computed on the next episode (of length $\tau = 100$) of the same dataset.

Real life weather data WeatherBench (European region)

- Data : 500 hPa geopotential height
- **DGFM**: generative U-net, conditioned on last 3 values.
- Episode: 362 datapoints (~1 year)

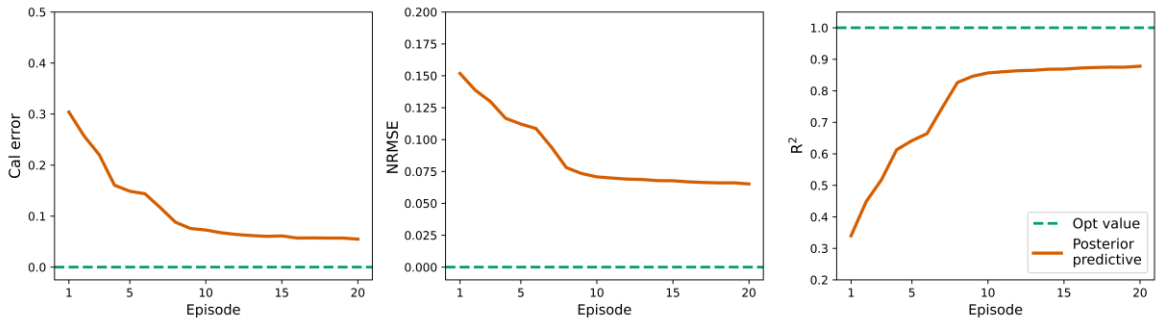


Fig 12: We measure the predictive accuracy of the posterior predictive (orange line) after every episodic update in terms of calibration error, normalised RMSE (NRMSE), and R^2. For reference, the maximum achievable values of the metrics are shown in green dashed line.

Real life weather data WeatherBench (European region)

- Data : 500 hPa geopotential height
- **DGFM**: generative U-net, conditioned on last 3 values.
- Episode: 362 datapoints (~1 year)

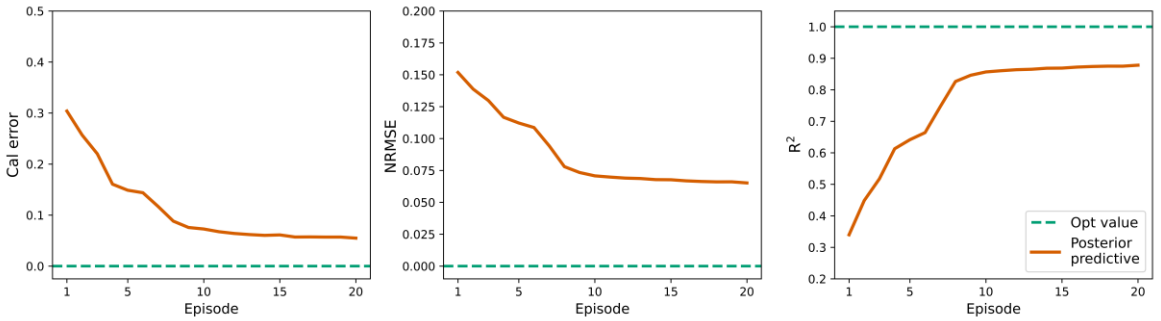


Fig 12: We measure the predictive accuracy of the posterior predictive (orange line) after every episodic update in terms of calibration error, normalised RMSE (NRMSE), and R^2. For reference, the maximum achievable values of the metrics are shown in green dashed line.

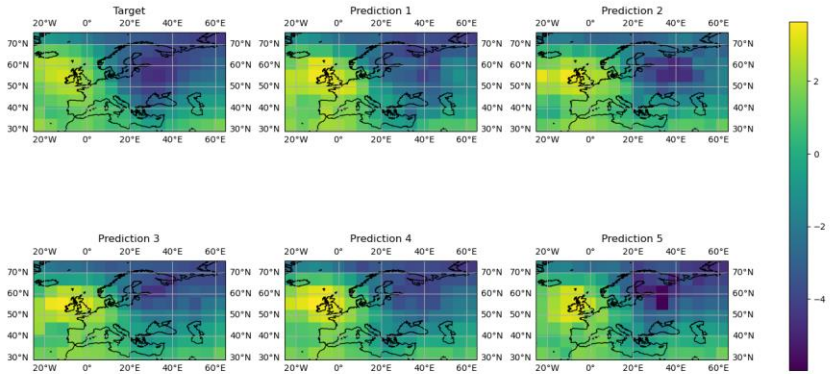


Fig 13:Top left: target. Rest are predictions from the final posterior samples.

- **Successful data assimilation** : We can see improved predictive performance of the posteriors over training episodes.

Application in model-based reinforcement learning (RL)

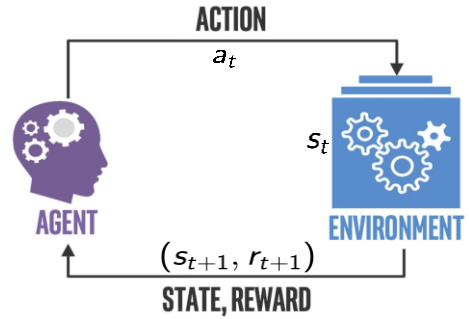


Fig 13: RL as a Markov decision process.

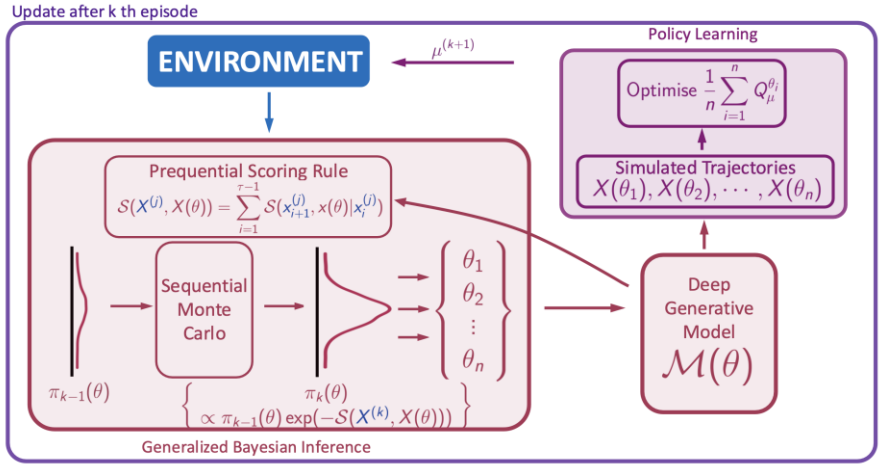


Fig 14: Diagram explaining Generalized Bayesian deep RL using Expected Thompson Sampling.

- Roy, S. S., Everitt, R. G., Robert, C. P., Dutta, R. (2024). Generalized Bayesian deep reinforcement learning. arXiv preprint arXiv:2412.11743.

Application in model-based reinforcement learning (RL)

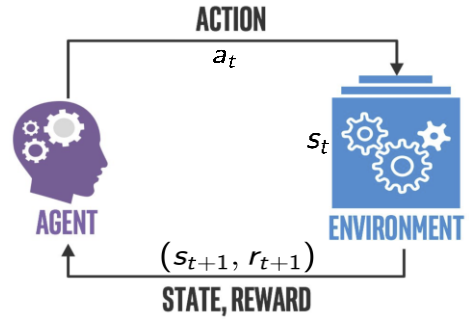
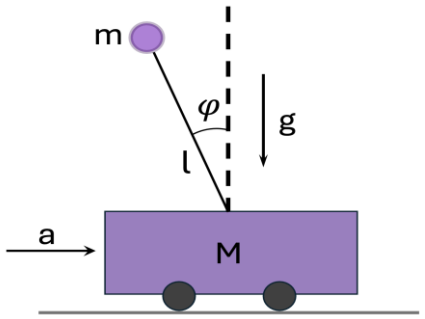
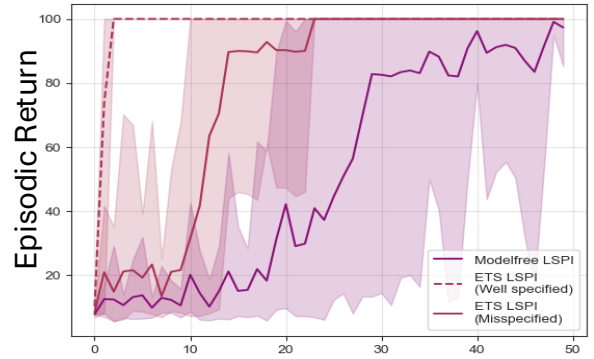


Fig 13: RL as a Markov decision process.



The pendulum balancing task



Number of training episodes

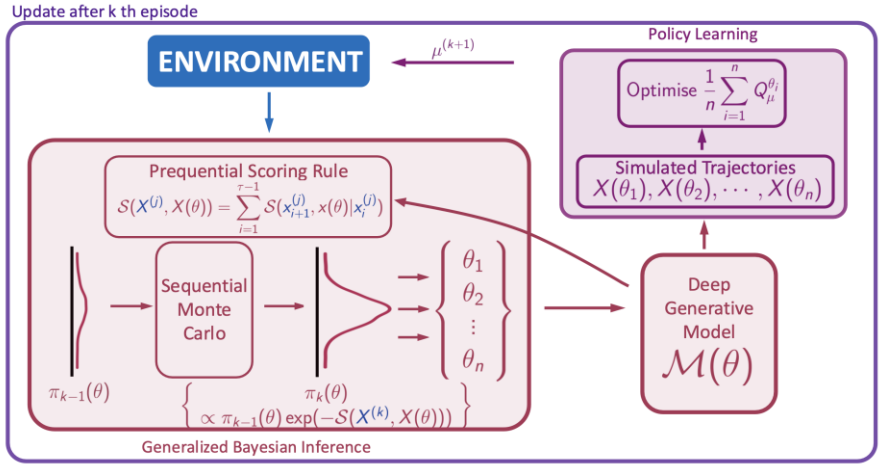
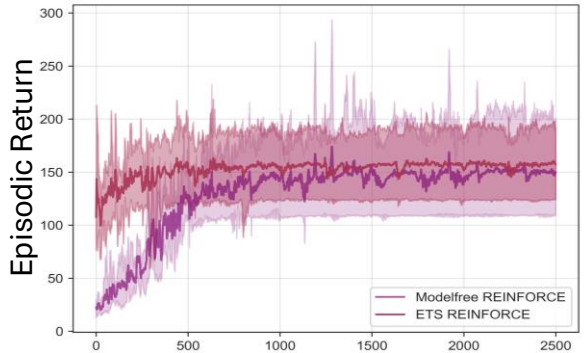
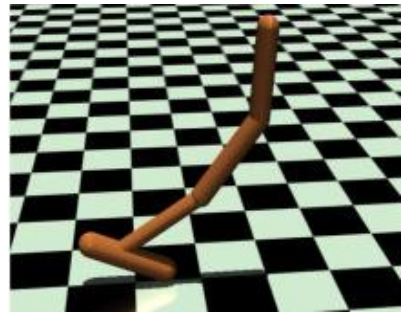


Fig 14: Diagram explaining Generalized Bayesian deep RL using Expected Thompson Sampling.



Number of training episodes



OpenAI Gymnasium Hopper expt

Fig 15: Examples where model based gen Bayes RL algorithm (red) learns an optimal policy faster than the modelfree(pink) approach.

- Roy, S. S., Everitt, R. G., Robert, C. P., Dutta, R. (2024). Generalized Bayesian deep reinforcement learning. arXiv preprint arXiv:2412.11743.

Conclusion

- A likelihood-free data assimilation framework for deep generative forecasting models (DGFMs) via prequential posteriors.
- Theoretical analysis of the prequential posterior, establishing predictive consistency under model misspecification.
- A scalable and computationally efficient SMC algorithm (Wastefree SMC + preconditioned kernel + symmetric splitting scheme) for training high-dimensional neural network parameters.

Thank you!