What determines the success of ensemble data assimilation methods?

Femke C. Vossepoel

Max Frei and Nikolaj Mücke

TU Delft

Code available from `https://github.com/nmucke/non_gaussian_data_assim.git`

Outline

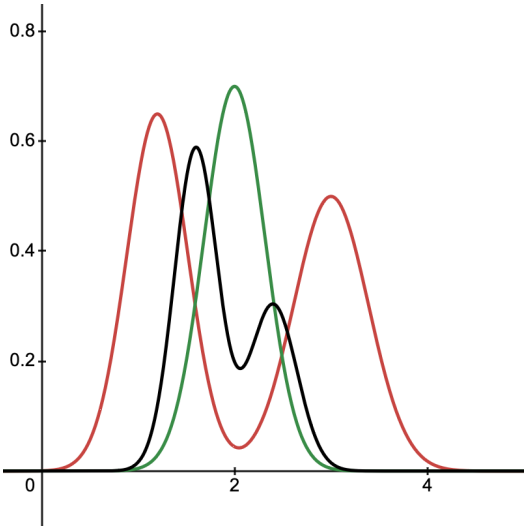
- What is for you the purpose of data assimilation?
- Reconstruction of posterior pdf: Kullback-Leibler divergence and other metrics
- Test case: Kuramoto Sivashinsky
- Information gain
- Implications
- Summary

Outline

- What is for you the purpose of data assimilation?
- Reconstruction of posterior pdf: Kullback-Leibler divergence and other metrics
- Test case: Kuramoto Sivashinsky
- Information gain
- Implications
- Summary

What is for you the purpose of data assimilation?

Reconstructing the posterior following Bayes' theorem



$$f(\mathbf{z}|\mathbf{d}) \propto f(\mathbf{d}|\mathbf{z}) f(\mathbf{z})$$

How we evaluate data assimilation today – and why it is not quite right

- In practice we judge a method by its fit to observations (RMSE), its forecast skill, and sometimes the ensemble spread.
- These mostly test the *point estimate* – the mean – not the full posterior distribution that ensemble / Bayesian DA is meant to deliver.
- A method can therefore look good (small RMSE) yet carry the wrong uncertainty: over-confident, mis-shaped, or with the wrong tails.

We want to evaluate the whole estimated posterior, not just its mean.

The dream*: estimate a posterior ≈ true posterior

The right yardstick compares the estimated posterior q with the true posterior p through the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(p \,\|\, q) = \int p(\mathbf{z}) \, \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \, d\mathbf{z} \;\geq\; 0, \qquad = 0 \iff q = p.$$

It vanishes only for the true posterior and penalises *every* error – location, spread, shape and tails.

But we can never compute it. We do not have the true posterior: only a *single* realisation of the truth and *noisy observations* at a few sensor locations. A KL estimate would need many samples of the (unknown) true posterior.

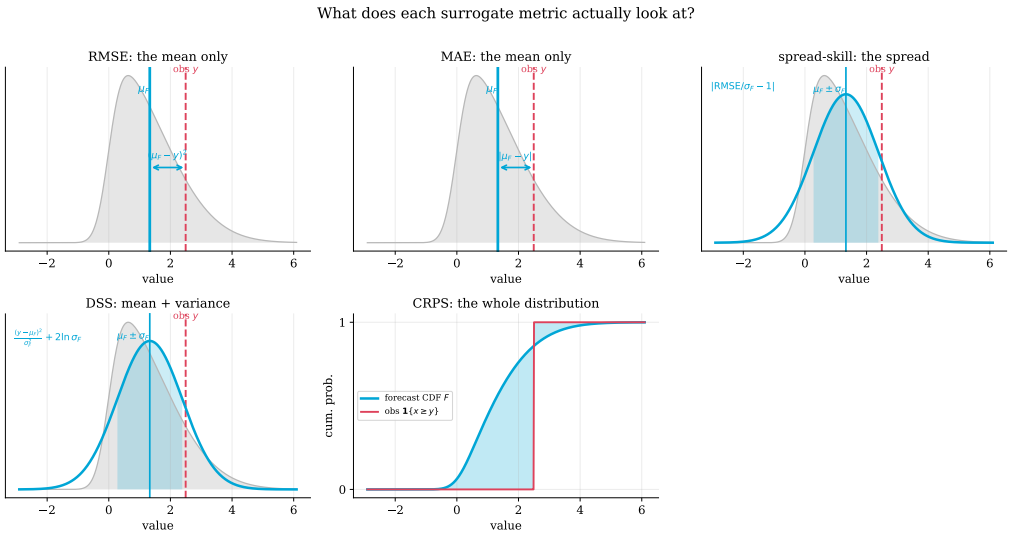
What we can do: hold-out observations and evaluate *surrogate* metrics

We hold out a few validation observations and score the estimated posterior against the observed value y (the only ground truth we have):

Root Mean Square Error RMSE	$\sqrt{\mathbb{E}[(y-\mu_F)^2]}$
Mean Absolute Error MAE	$\mathbb{E}[\,\lvert y-\mu_F\rvert\,]$
Spread–skill	$\left\lvert \mathrm{RMSE}/\sigma_F - 1 \right\rvert$
Dawid–Sebastiani Score, DSS	$\mathbb{E}[(y-\mu_F)^2]/\sigma_F^2 + 2\ln\sigma_F$
Continuous Ranked Probability Score	$\mathbb{E}\lvert X-y\rvert - \frac{1}{2}\mathbb{E}\lvert X-X'\rvert = \int \big(F(x) - \mathbf{1}\{x \geq y\}\big)^2 dx$
CRPS, Gneiting and Raftery (2007)	

None of these is the KL we want – so we must ask: do they behave like KL?

What does each surrogate metric actually measure?

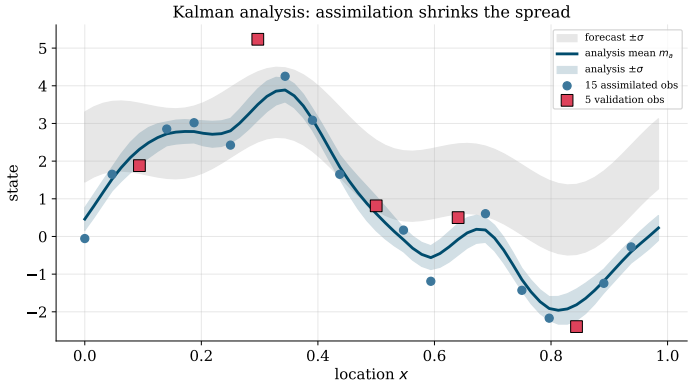


RMSE and MAE use only the mean; spread–skill, DSS adds the variance; only CRPS uses the whole distribution. The faint grey shape is what the moment-based metrics ignore.

A linear problem where the true posterior is known

To test the surrogates against the *real* KL we need a case where the true posterior is available.

- With a linear model and observation operator and Gaussian noise, one assimilation step gives the exact posterior (a Kalman update).
- 20 observations: 15 assimilated, 5 held out; the posterior is defined even at the held-out observations.

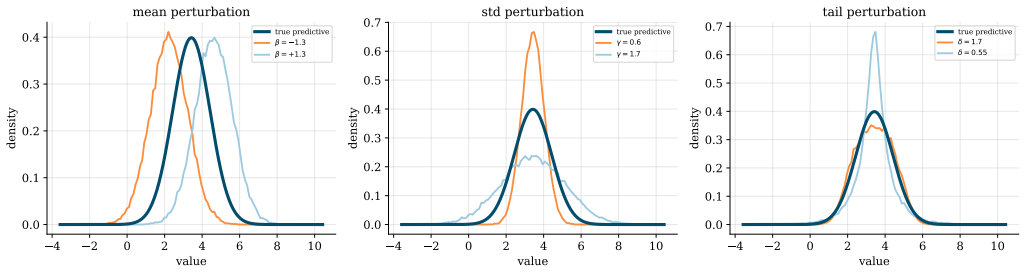


Perturbing the true posterior into “methods”

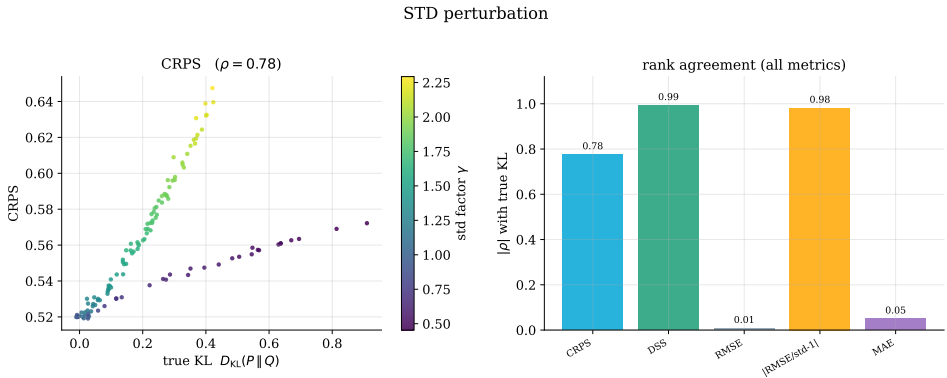
We turn the exact posterior into 200 imperfect “methods” by perturbing it in four ways – alone and in every combination:

- mean (bias), std (spread), tail (heaviness), skew (asymmetry). The tail and skew knobs keep the *mean and variance fixed*.

Three ways we perturb the true posterior

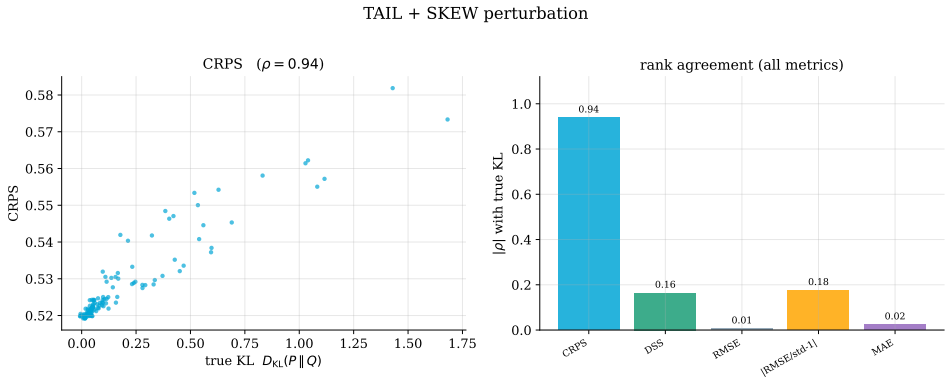


Std perturbation: does CRPS track the true KL?



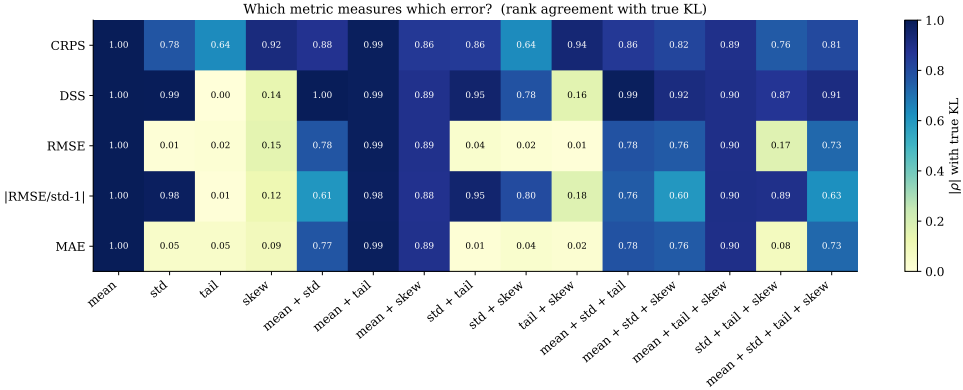
Std perturbation only. CRPS rises with the true KL (ρ = 0.78). The bar shows all metrics: CRPS, DSS and spread–skill catch a spread error, while RMSE and MAE are blind – they use only the mean.

Tail + skewness: does CRPS track the true KL?



Tail + skew (both leave the mean and variance unchanged). Only CRPS tracks the true KL ($\rho = 0.94$); the moment-based metrics – DSS (0.16), spread–skill, RMSE / MAE (≈ 0) – are blind to pure shape errors.

Confusion matrix: which metric measures which error



Rank agreement $|\rho|$ with the true KL over all 15 perturbation combinations. Mean: every metric. Spread: CRPS, DSS, spread–skill. Tail & skew: only CRPS. RMSE / MAE see location alone – CRPS is the one surrogate that reflects the whole posterior.

Outline

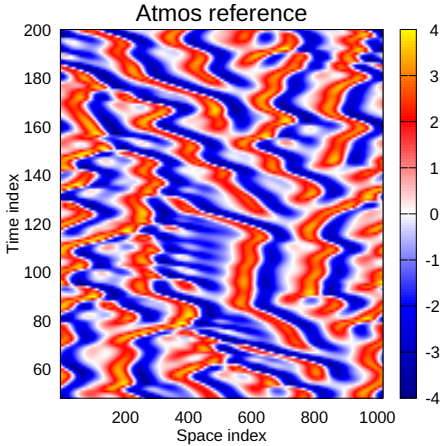
- What is for you the purpose of data assimilation?
- Reconstruction of posterior pdf: Kullback-Leibler divergence and other metrics
- Test case: Kuramoto Sivashinsky
- Information gain
- Implications
- Summary

Example with (uncoupled) multiscale Kuramoto-Sivashinsky model

$$\tau_a \frac{\partial A}{\partial t} = -\frac{\lambda_a}{2}\frac{\partial A^2}{\partial x} - \lambda_a^2 \frac{\partial^2 A}{\partial x^2} - \frac{\lambda_a^4}{2}\frac{\partial^4 A}{\partial x^4}$$

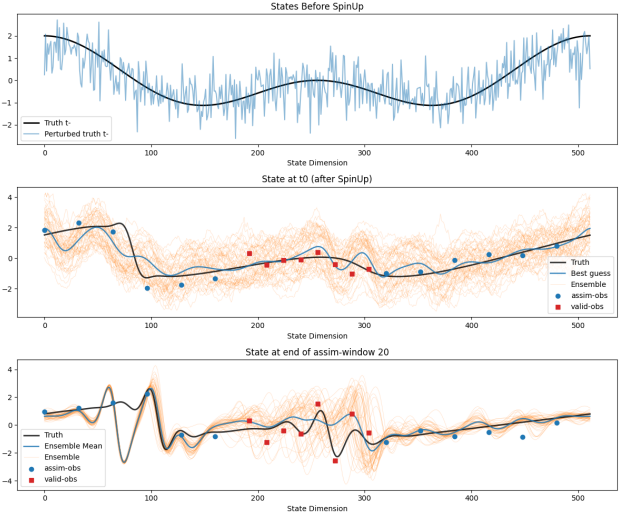
We apply EnKF for synthetic data-assimilation experiments.

Example with (uncoupled) multiscale Kuramoto-Sivashinsky model

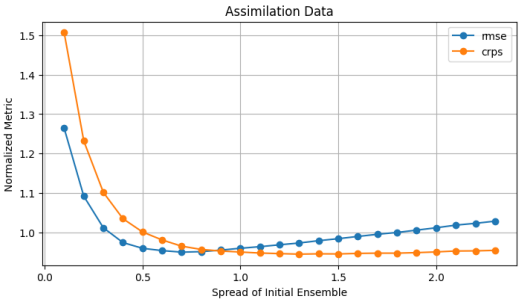


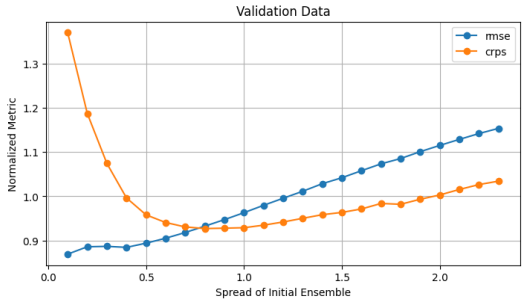
- choose an initial value for for an atmospheric state (based on some cosine functions)
- define a synthetic *truth*
- sample observations
- simulate atmospheric variability with the KS model
- assimilate some of the observations, validate with others

Setup of Kuramoto Sivashinsky identical twin experiment



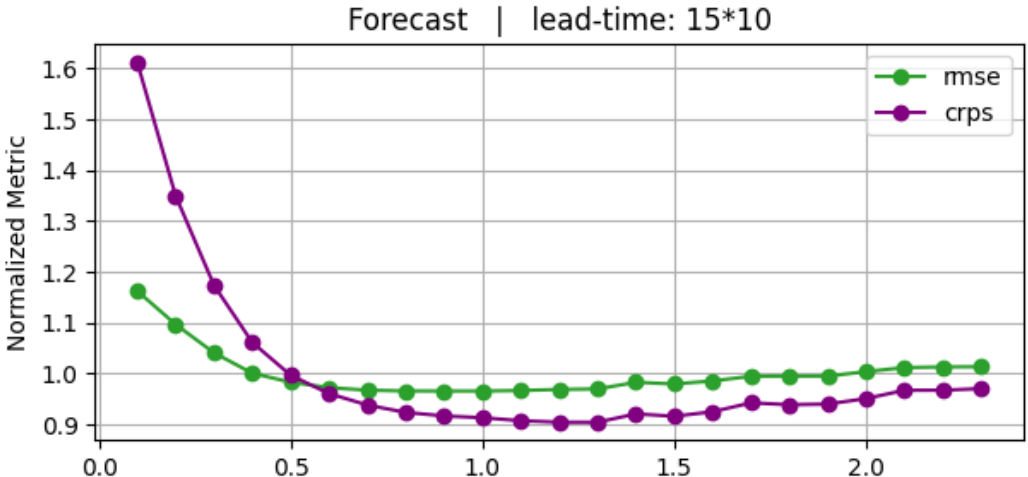
Hyperparameter tuning: CRPS versus RMSE





- Aim: find optimal spread

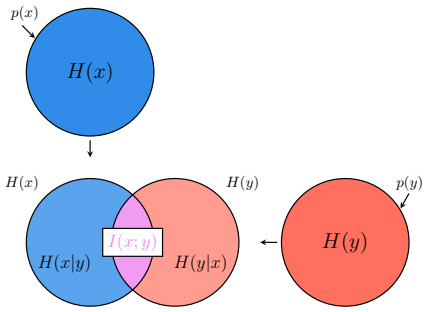
Hyperparameter tuning: CRPS versus RMSE

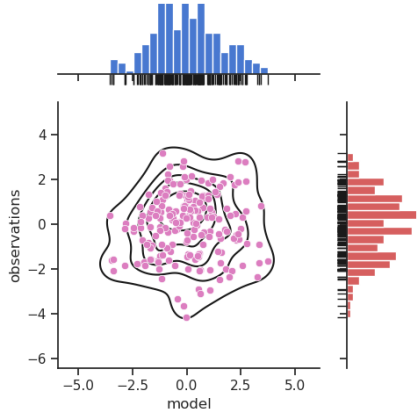


Outline

- What is for you the purpose of data assimilation?
- Reconstruction of posterior pdf: Kullback-Leibler divergence and other metrics
- Test case: Kuramoto Sivashinsky
- Information gain
- Implications
- Summary

Entropy, mutual information





From Kim and Vossepoel (2024), after Nearing et al. (2018)

Entropy, mutual information

Entropy:

$$H(x) = -\sum_i p(x_i) \log\left(p(x_i)\right) . \tag{1}$$

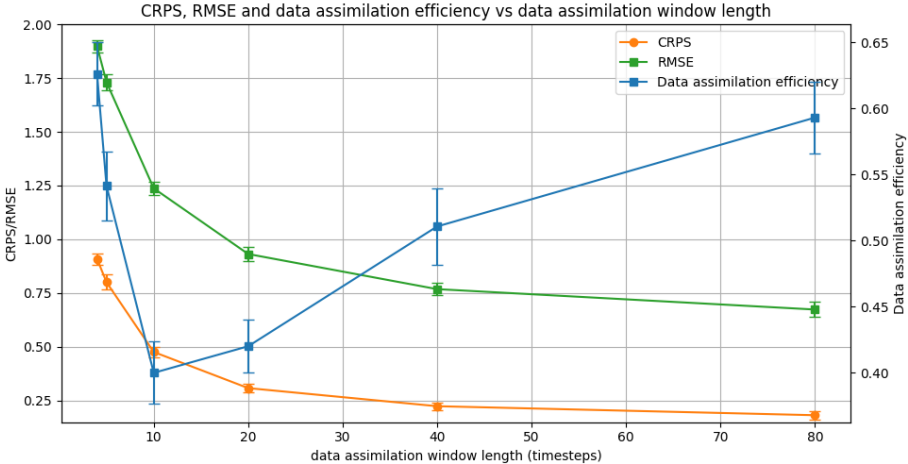
Mutual information written as a function of the entropy and the joint entropy $H(z, x)$:

$$I(z; x) = H(x) + H(z) - H(z, x). \tag{2}$$

Data assimilation efficiency $\mathcal{E}_{DA}$. $\mathcal{E}_{DA}$ is then defined as the ratio of the posterior mutual information $I(z; \hat{x})$ and the total information in model and observations $I(z; x, y)$,

$$\mathcal{E}_{DA} = \frac{I(z; \hat{x})}{I(z; x, y)} \tag{3}$$

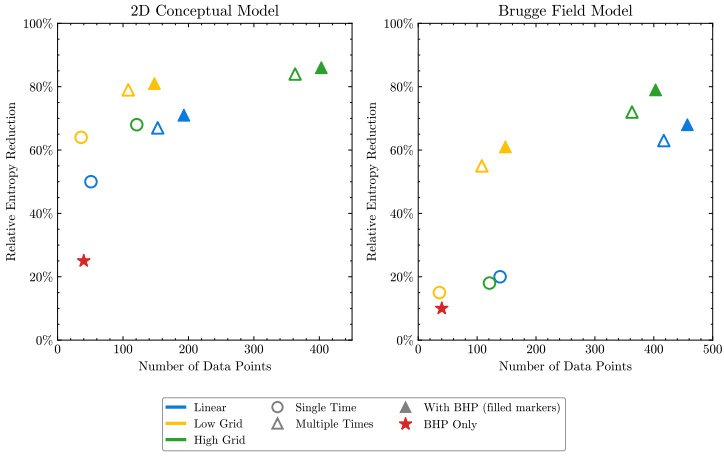
Resulting data-assimilation efficiency for KS test case (averaged over 100 experiments, 50 ensemble members each)



Implications

- Continuous Ranked Probability Score (CRPS) is a highly informative metric
- Evaluation of metrics from information theory provide another perspective
- All methods are equal, some are more equal than others
- Using multiple metrics to evaluate a method's performance helps to tune hyper-parameters

Example: Entropy reduction as a function of the number of observations in high-dimensional model



- using entropy reduction to compare performance of data assimilation in conceptual and realistic reservoir model
- testing different spatial densities of deformation observations for monitoring gas storage
- testing additional types of observations

From Seabra1 et al. (2026), to appear in *"Subsurface Data Assimilation" (ISBN 9780443415432)*, November 2026

Summary

- In Bayesian data assimilation, the goal is (or should be) to reconstruct the *true* posterior
- Since we don't know the true posterior, we cannot use the Kullback-Leibler divergence to assess the performance of the method
- Metrics like MAE and RMSE provide only limited information on the data-assimilation performance
- Continuous Ranked Probability Score (CRPS) is a highly informative metric
- Evaluation of metrics from information theory provide another perspective
- In the end, it is up to the data assimilation practitioner to evaluate

Bibliography

Gneiting, T. and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102 (477):359–378, 2007.

Kim, S. S. R. and F. C. Vossepoel. On spatially correlated observations in importance sampling methods for subsidence estimation. *Computational Geosciences*, 28(1):91–106, 2024. doi:10.1007/s10596-023-10264-9. URL https://doi.org/10.1007/s10596-023-10264-9.

Nearing, G., S. Yatheendradas, W. Crow, X. Zhan, J. Liu, and F. Chen. The efficiency of data assimilation. *Water resources research*, 54(9): 6374–6392, 2018. doi:10.1029/2017WR020991.

Seabra1, G. S., I. Saifullin1, D. Voskov, and F. C. Vossepoel. Geomechanical proxy models and data assimilation: Evaluating monitoring strategies in geological carbon sequestration. *to appear in "Subsurface Data Assimilation" (ISBN 9780443415432)*, 2026.