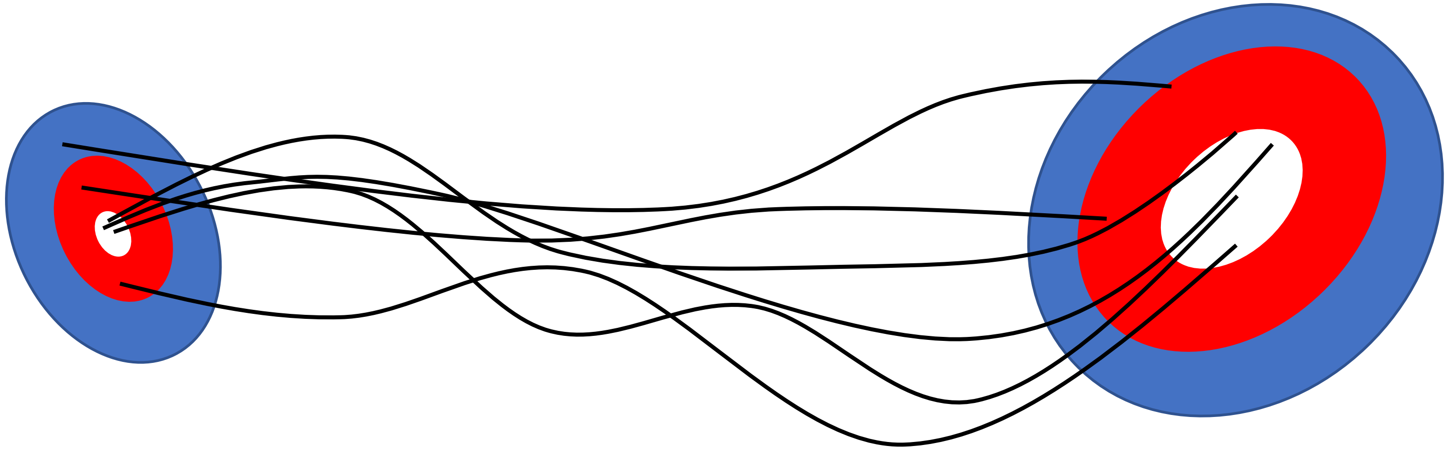
The Ensemble Kalman filter



Part II: Practicalities

Yumeng Chen (Based on notes by Ross Bannister, Alison Fowler, and Ivo Pasmans)

Data-assimilation training course. 9th-14th June, University of Reading

Quick recap

- Recall that the ensemble covariance matrix at time step k is given by with N ensemble members.

$$\mathbf{P}_k^{\mathrm{f}} = \frac{1}{N-1}\mathbf{X}_k'^{\mathrm{f}}\left(\mathbf{X}_k'^{\mathrm{f}}\right)^{\mathrm{T}} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_k^{(i),\mathrm{f}} - \bar{\mathbf{x}}_k^{\mathrm{f}}\right)\left(\mathbf{x}_k^{(i),\mathrm{f}} - \bar{\mathbf{x}}_k^{\mathrm{f}}\right)^{\mathrm{T}}$$

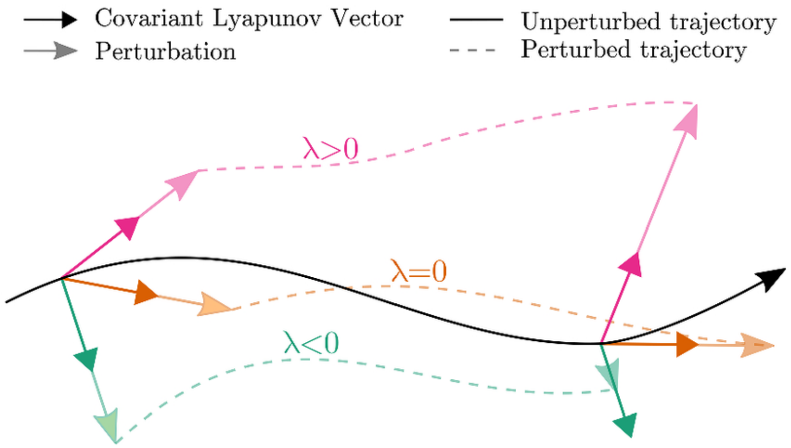
- We can write a similar expression for the analysis error covariance matrix in terms of the analysis perturbations

$$\mathbf{P}_k^{\mathrm{a}} = \frac{1}{N-1}\mathbf{X}_k'^{\mathrm{a}}(\mathbf{X}_k'^{\mathrm{a}})^{\mathrm{T}} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_k^{(i),\mathrm{a}} - \bar{\mathbf{x}}_k^{\mathrm{a}}\right)\left(\mathbf{x}_k^{(i),\mathrm{a}} - \bar{\mathbf{x}}_k^{\mathrm{a}}\right)^{\mathrm{T}}$$

where $\mathbf{X}' \ \in \mathbb{R}^{n\times N}$ is the **perturbation matrix**.

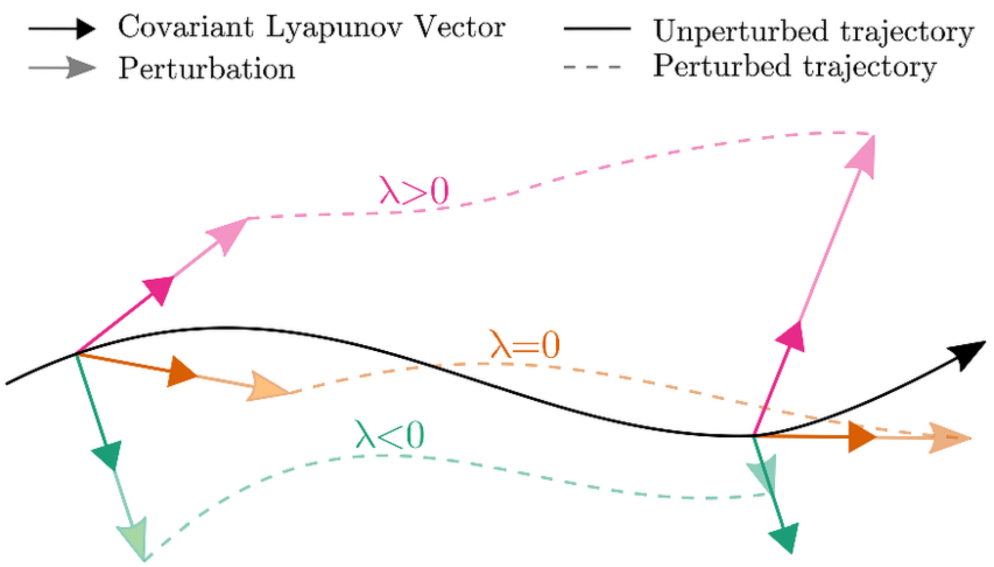
How much ensemble is sufficient?

- A dynamical system with n number of states can have up to n different modes. Typically, only $N < n$ modes can lead to increased error from small errors in the initial condition.
  - In the weather systems, unstable convective storms and stable high pressure systems (imagining “anticyclonic gloom”) can occur simultaneously.
  - Mathematically, these modes are quantified by Lyapunov exponents.



How much ensemble is sufficient?

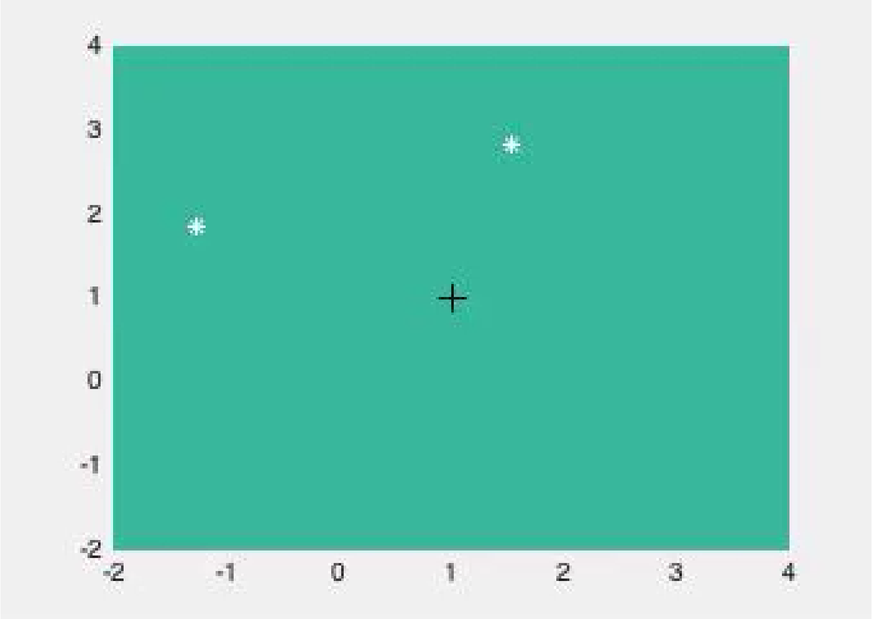
- We just need more than N number of ensemble members if
  - we know the number of positive Lyapunov exponents
  - the system is well-observed
  - the model is linear
  - the model has no random errors
  - a deterministic Kalman filter (e.g., ETKF) is used
- Practically, we'd like to have as many ensemble members as possible



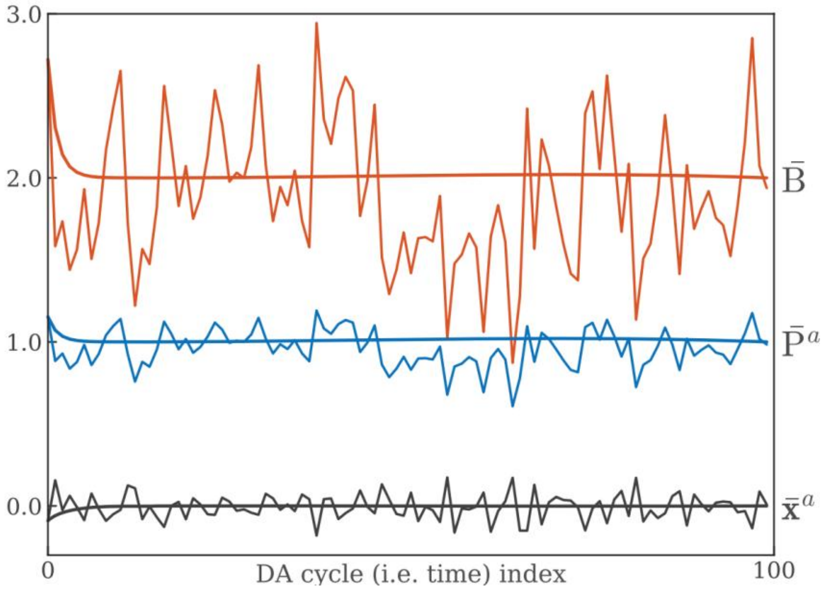
Bocquet, M., Gurumoorthy, K. S., Apte, A., Carrassi, A., Grudzien, C., and Jones, C. K. R. T.: Degenerate Kalman Filter Error Covariances and Their Convergence onto the Unstable Subspace, SIAM/ASA Journal on Uncertainty Quantification, 5, 304–333, https://doi.org/10.1137/16M1068712, 2017. a, b, c

Sampling error

- The ensemble Kalman Filter theory assumes that the ensemble is large enough to give an accurate estimate of the sample mean and covariance, $\bar{\mathbf{x}}$ and $\mathbf{P}$.
- Sampling errors can be caused by:
    - Limited ensemble members due to constraints from computational resources
    - Model nonlinearity



Even for a two variable model a large sample size is needed to accurately estimate the mean and covariance: $\boldsymbol{x} \sim \mathcal{N}([1,1], \boldsymbol{I}_2)$



Sampling errors caused by non-linear model even if Gaussianity is preserved (jittery lines) compared to the linear model (smooth lines)

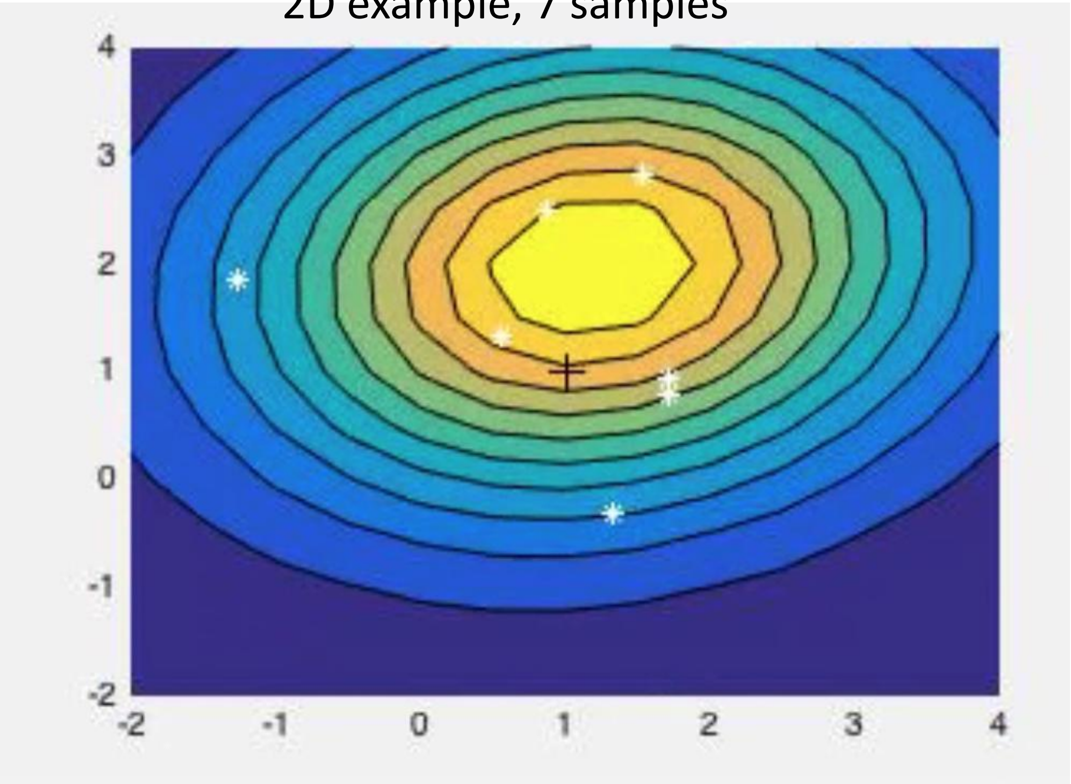
Raanes PN, Bocquet M, Carrassi A. Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Q J R Meteorol Soc*. 2019; 145: 53–75. **https://doi.org/10.1002/qj.3386**

Consequences of under sampling in EnKF

There are numerous consequences to under sampling in the EnKF

1. There may be a **bias** in the **ensemble mean**.

2D example, 7 samples



Consequences of under sampling in EnKF

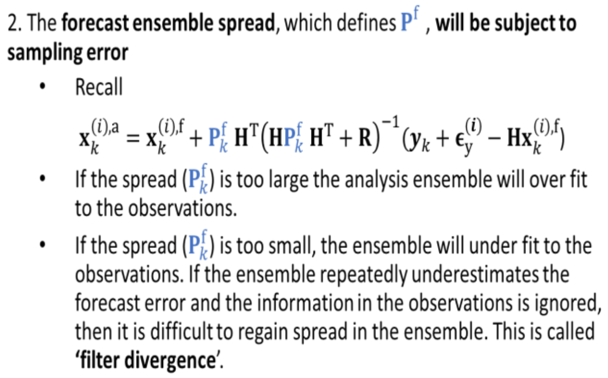
2. The **forecast ensemble spread**, which defines $\mathbf{P}^{\mathrm{f}}$, **will be subject to sampling error**

- Recall

$$\mathbf{x}_k^{(i),\mathrm{a}} = \mathbf{x}_k^{(i),\mathrm{f}} + \mathbf{P}_k^{\mathrm{f}}\,\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_k^{\mathrm{f}}\,\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\left(\boldsymbol{y}_k + \boldsymbol{\epsilon}_{\mathrm{y}}^{(\boldsymbol{i})} - \mathbf{H}\mathbf{x}_k^{(i),\mathrm{f}}\right)$$

- If the spread ($\mathbf{P}_k^{\mathrm{f}}$) is too large the analysis ensemble will over fit to the observations.
- If the spread ($\mathbf{P}_k^{\mathrm{f}}$) is too small, the ensemble will under fit to the observations. If the ensemble repeatedly underestimates the forecast error and the information in the observations is ignored, then it is difficult to regain spread in the ensemble. This is called **'filter divergence'**.

2D example, 7 samples



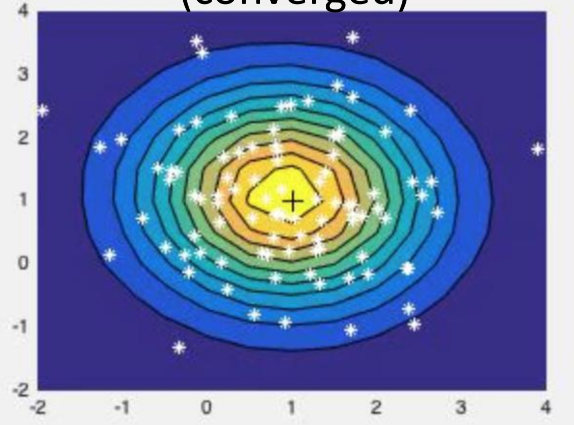
2. The **forecast ensemble spread**, which defines $\mathbf{P}^{\mathrm{f}}$, **will be subject to sampling error**

- Recall

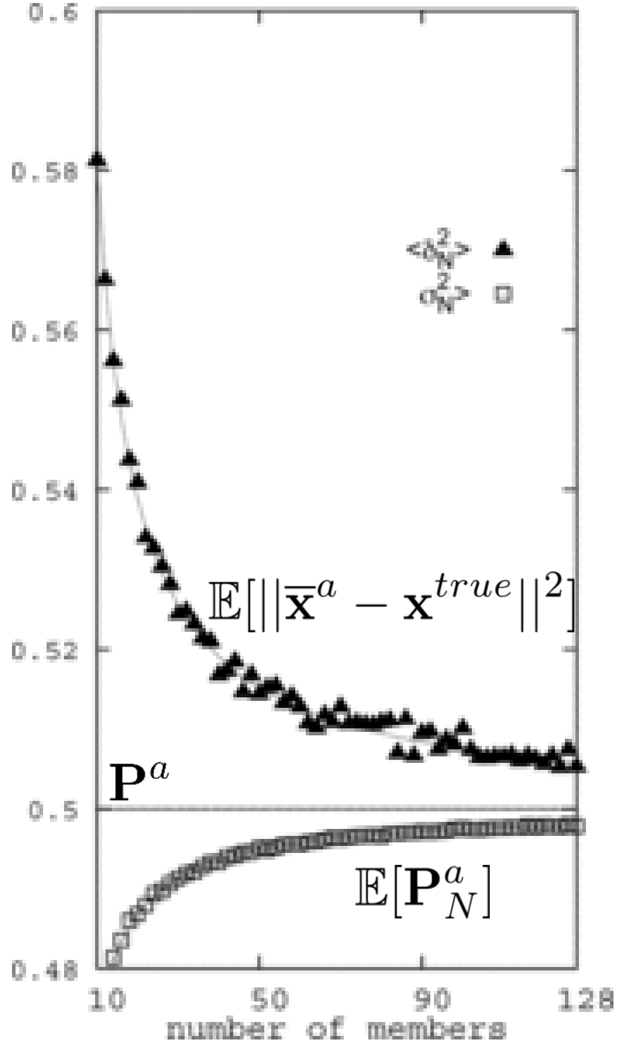
$$\mathbf{x}_k^{(i),\mathrm{a}} = \mathbf{x}_k^{(i),\mathrm{f}} + \mathbf{P}_k^{\mathrm{f}}\,\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_k^{\mathrm{f}}\,\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\left(\boldsymbol{y}_k + \boldsymbol{\epsilon}_{\mathrm{y}}^{(\boldsymbol{i})} - \mathbf{H}\mathbf{x}_k^{(i),\mathrm{f}}\right)$$

- If the spread ($\mathbf{P}_k^{\mathrm{f}}$) is too large the analysis ensemble will over fit to the observations.
- If the spread ($\mathbf{P}_k^{\mathrm{f}}$) is too small, the ensemble will under fit to the observations. If the ensemble repeatedly underestimates the forecast error and the information in the observations is ignored, then it is difficult to regain spread in the ensemble. This is called **'filter divergence'**.

2D example, 100 samples (converged)

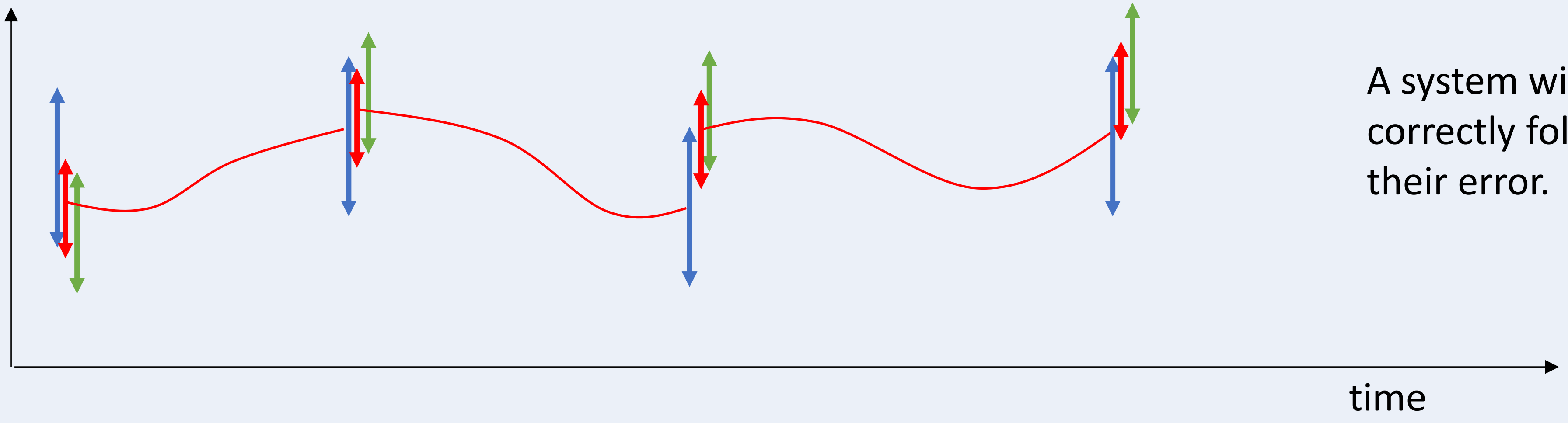


Variance bias is function of ensemble size



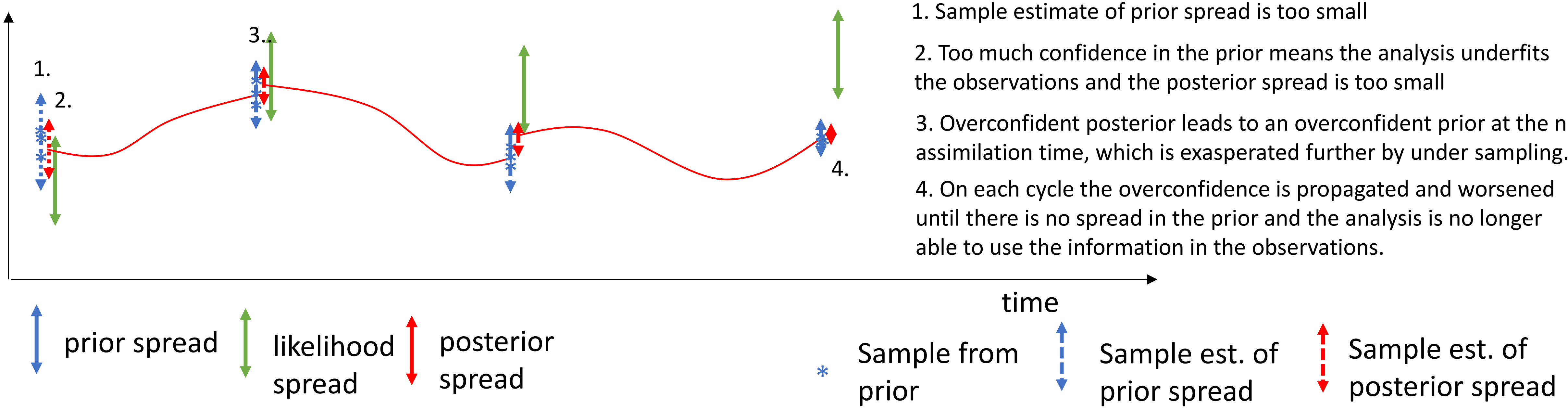
- In limit $N \to \infty$ the mean squared error and ensemble variance match on average.
- For small ensembles, the ensemble spread underestimates the analysis uncertainty.
- Smaller ensembles result in less error removal.

Image: Sacher, W., & Bartello, P. (2008). Sampling errors in ensemble Kalman filtering. Part I: Theory. *Monthly Weather Review*, *136*(8), 3035-3049.



A system without sampling error can correctly follow the observations within their error.

Illustration of sampling error leading to filter divergence



1. Sample estimate of prior spread is too small

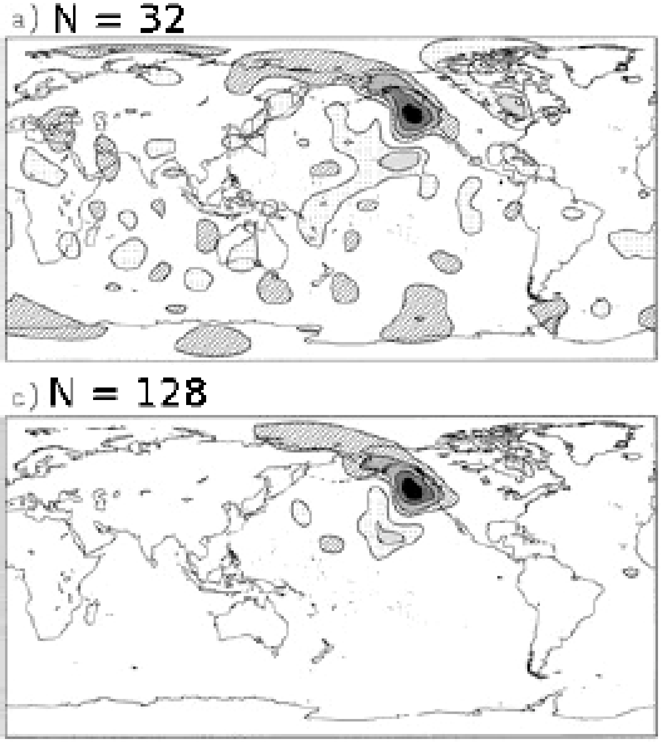
2. Too much confidence in the prior means the analysis underfits the observations and the posterior spread is too small

3. Overconfident posterior leads to an overconfident prior at the next assimilation time, which is exasperated further by under sampling.

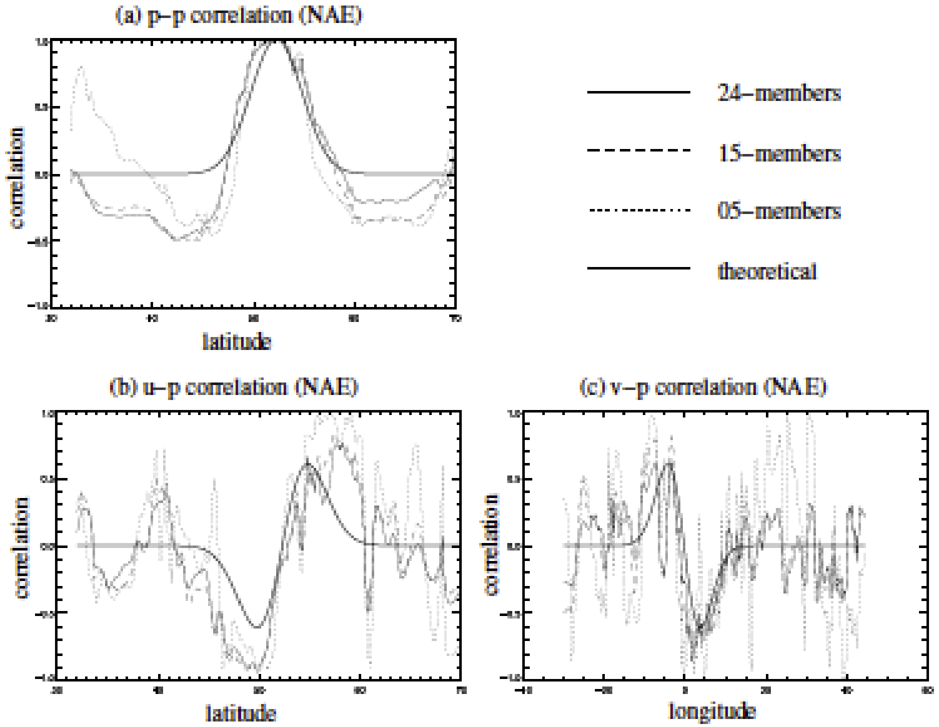
4. On each cycle the overconfidence is propagated and worsened until there is no spread in the prior and the analysis is no longer able to use the information in the observations.

Consequences of under sampling in EnKF

3. The **correlation will be subject to sampling error**. Implying that observations can influence regions and variables that they shouldn't.



From Houtekamer & Mitchell (1998)



From Bannister, Migliorini & Dixon (2011)

Consequences of under sampling in EnKF

4. $\mathbf{P}_k^{\mathrm{f}}$ **is rank deficient**

Remember that $\mathbf{P}^{\mathrm{f}} \in \mathbb{R}^{n\times n}$ and $\mathbf{X'}^{\mathrm{f}} \in \mathbb{R}^{n\times N}$, so rank($\mathbf{P}_k^{\mathrm{f}}$) = rank($\mathbf{X'}_k^{\mathrm{f}}$) $\leq N-1 \ll n$.

The analysis increments are in the sub-space spanned by the forecast ensemble

- The analysis increments are given by

$$\mathbf{x}_k^{(i),\mathrm{a}} - \mathbf{x}_k^{(i),\mathrm{f}} = \mathbf{P}_k^{\mathrm{f}}\,\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_k^{\mathrm{f}}\,\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}(\boldsymbol{y}_k + \boldsymbol{\epsilon}_{\mathrm{y}}^{(\boldsymbol{i})} - \mathbf{H}\mathbf{x}_k^{(i),\mathrm{f}})$$

- The analysis increments are therefore a linear combination of the forecast error perturbations.
- Therefore, even if the observations indicate otherwise, the analysis is restricted to space spanned by the ensemble which has at the most a dimension of *N-1*.
- The error covariance matrix cannot represent all modes of forecast/analysis error.

Possible solutions

1. Use more ensemble members (see Miyoshi et al. 2014)
2. Re-centre the ensemble around a deterministic analysis e.g., from 4DVar.
    - Addresses problem of bias in the ensemble mean
3. Ensemble inflation
    - Addresses problem of filter divergence
4. Localization
    - Addresses problem of spurious correlations
    - Splits problem into quasi-independent problems, increasing the rank of forecast perturbation matrix.
5. Combine ensemble with variational approaches (see afternoon's lectures)
    - These are known as hybrid methods

Focus of this lecture

Ensemble inflation

Ways to inflate

- Additive inflation (Mitchell and Houtekamer, 2000; Corazza *et al*., 2003)
    - At each model time step add a random perturbation using similar ideas to representing model error given in the last lecture

$$\mathbf{x}_k^{(i)} = M_{t_{k-1}\to t_k}(\mathbf{x}_{k-1}^{(i)}) + \boldsymbol{\eta}_k^{(i)}, \text{ where } \boldsymbol{\eta} \sim N(\mathbf{0},\mathbf{Q})$$

- Multiplicative inflation (Anderson and Anderson, 1999)

$$\mathbf{P}^{\mathrm{f}}_{\text{inflated}} = (1+\rho)^2\mathbf{P}^{\mathrm{f}}, \qquad \rho > 0$$

- Relaxation to prior ensemble (Zhang *et al*., 2004; Whitaker and Hamill, 2012)
    - Only accept part of the spread reduction proposed by

$$\mathbf{P}_k^{\mathrm{a}} = (\mathbf{I}-\mathbf{K}_k\mathbf{H})\,\mathbf{P}_k^{\mathrm{f}}$$

-
    - Two methods: relaxation to prior perturbation and relaxation to prior spread

Relaxation

- Relaxation to prior perturbation (Zhang & Sun, 2004):

$$\mathbf{x}'^{a,(n)} \leftarrow (1-\alpha)\mathbf{x}'^{a,(n)} + \alpha\mathbf{x}'^{f,(n)} \text{ with } \mathbf{x}'^{(n)} = \mathbf{x}^{(n)} - \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}^{(m)}$$

  - Does not work with rotation in ETKF as there is not a one-to-one relation between forecast and analysis members.

- Relaxation to prior spread (Whitaker & Hamill, 2012):

$$\mathbf{x}_i'^{a,(n)} \leftarrow \mathbf{x}_i'^{a,(n)}\left(\alpha\frac{\sigma_i^b - \sigma_i^a}{\sigma_i^a} + 1\right)$$

if $\alpha = 0, x_i'^{a,(n)} = x_i'^{a,(n)}$; if $\alpha = 1, x_i'^{a,(n)} = x_i'^{b,(n)}$

Zhang F, Snyder C, Sun J. 2004. Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Weather Rev.* 132: 1238–1253.
Sakov, P., & Oke, P. R. (2008). A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus A: Dynamic Meteorology and Oceanography*, *60*(2), 361-371.
Whitaker, J. S., & Hamill, T. M. (2012). Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, *140*(9), 3078-3089.

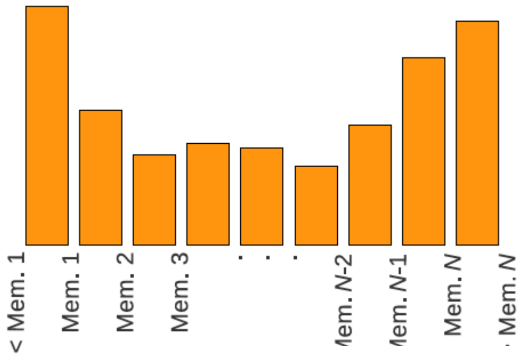
Tuning the inflation factor – validation of ensemble spread

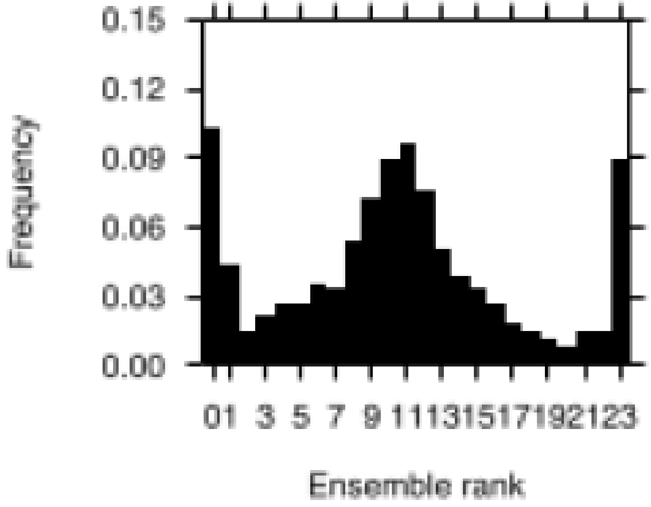
Method 1: rank histograms (Hamill, T., 2001):

- For the ensemble to be reliable it is assumed that it is sampling the same distribution as the truth.
  - The ensemble and true distribution lead to the same number of samples for each interval of the PDF
- A rank histogram is constructed by considering a point in space that is well observed.
  - The values of the ensemble members at that point are ranked from highest to lowest creating N-1 bins.
  - Then each observation is binned to give a frequency diagram.

Interpretation:

- Concave shape- the ensemble is under spread
- Convex shaped- the ensemble is overspread
- Flat- the ensemble is correctly spread
- Asymmetric- the ensemble is biased





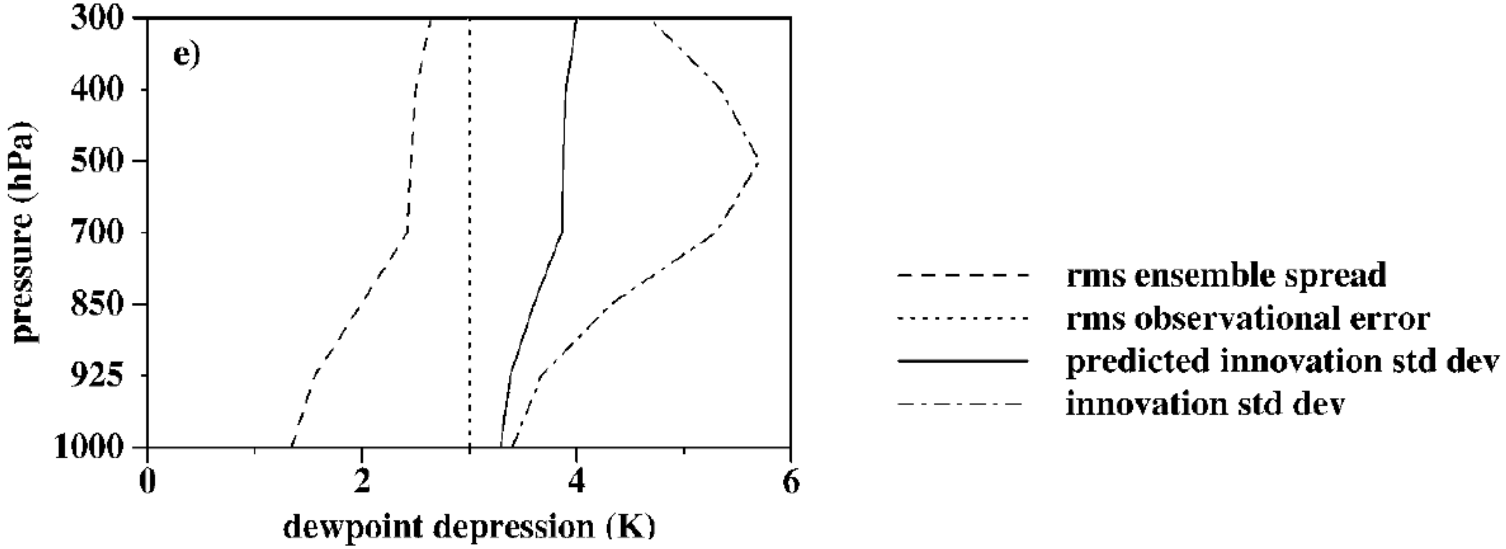
Rank histograms for surface precipitation rate rate. From Migliorini et al. (2011).

Tuning the inflation factor – validation of ensemble spread

Method 2: Covariance matching

- This checks that the sum of the spread in the background ensemble and observation error variance match with the variance of the innovations (e.g. Houtekamer et al. 2005).

$$E\left[\left(\boldsymbol{y}-\mathbf{Hx}^{\mathrm{f}}\right)\left(\boldsymbol{y}-\mathbf{Hx}^{\mathrm{f}}\right)^{\mathbf{T}}\right]\approx \mathbf{R}+\mathbf{HP}^{\mathrm{f}}\ \mathbf{H}^{\mathrm{T}}$$



The idea of covariance matching has led to various schemes for adaptive covariance inflation e.g. Kotsuki et al. 2017

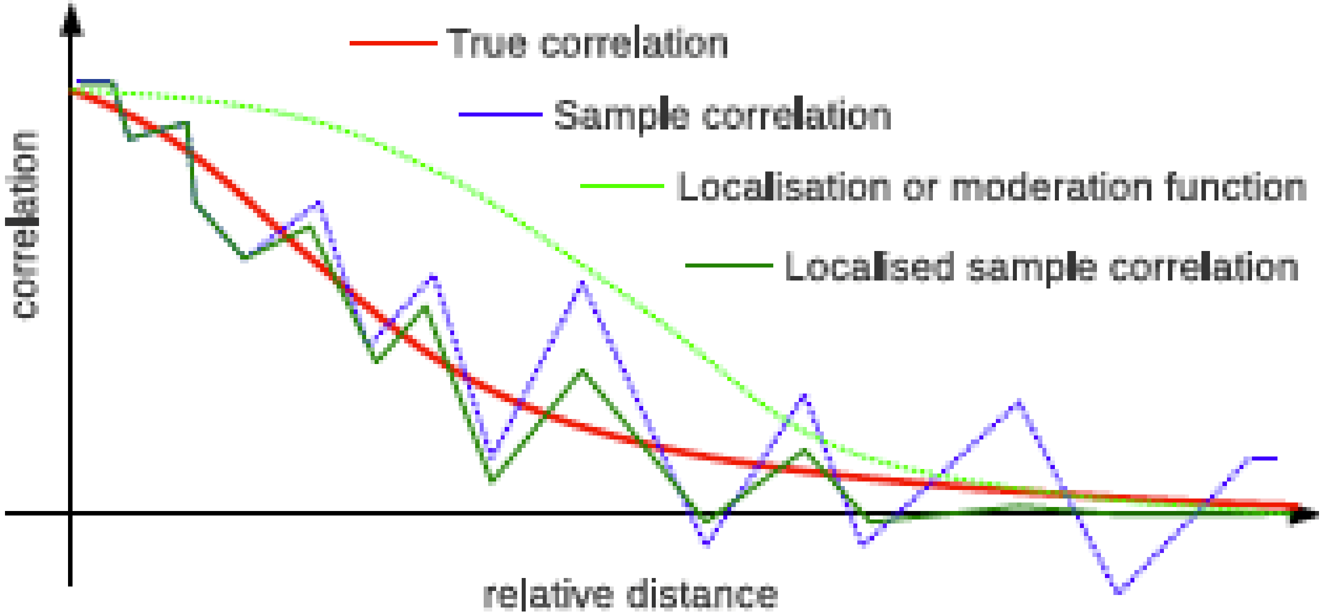
FIG. 5. Comparison of error amplitudes that have been averaged over a 10-day experimental period. Shown are the predicted innovation std dev (solid) that should match the observed innovation std dev (dashed–dotted). The predicted std dev is computed from the rms observational error (dotted) and the rms ensemble spread (dashed).

Localisation

The aim of localisation is to restrict the influence of observations to just a physically realistic region.

Two ways of doing this are:

- Covariance localisation ($\mathbf{P}^{\mathrm{f}}$-localisation)
  - This modifies the forecast error covariance matrix to reduce long-range correlations.



- Domain localisation ($\mathbf{R}$-localisation)
  - This restricts observations which are allowed to influence each grid point.

Covariance localisation (Houtekamer and Mitchell, 2001)

- Instead of using directly, $\boldsymbol{\rho} \circ \mathbf{P}^{\mathrm{f}}$ is used.
- In practice cannot act on $\mathbf{P}^{\mathrm{f}}$ directly

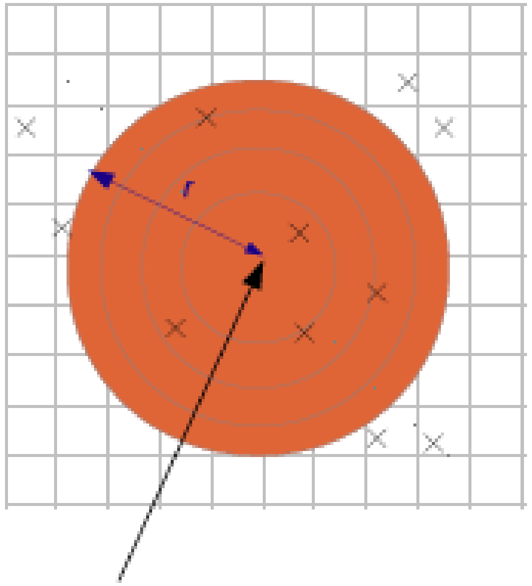
$$\mathbf{K} = \boldsymbol{\rho} \circ (\mathbf{P}^{\mathrm{f}} \ \mathbf{H}^{\mathrm{T}})\big(\boldsymbol{\rho} \circ (\mathbf{H}\mathbf{P}^{\mathrm{f}} \ \mathbf{H}^{\mathrm{T}}) + \mathbf{R}\big)^{-1}$$

  - Need to choose function $\boldsymbol{\rho}$ and length scales, this may be state-dependent
  - Not clear how to define distance between observations which have no clearly defined location in space, e.g. satellite observations
  - Not clear how to deal with multivariate covariances
  - Can affect the balance e.g. to conserve geostrophic balance length scale O(1000)km must be used in the horizontal

Domain localisation (Hunt et al. 2007)

- Each grid point performs an independent ETKF where only observations within a localisation radius are assimilated (LETKF)
- To avoid discontinuous DA corrections observational error variances are gradually inflated for observations far from selected grid point *i*, e.g.

$$f_{Rloc} = exp\left[\frac{d(i,j)^2}{2L^2}\right]$$



Example analysis point

where

- *d(i,j,)* is the distance between observation *i* and model grid point *j*.
- *L* is the localisation lengthscale.

Hybrid methods

Hybrid methods combine the best parts of the EnKF (flow-dependent P^f) with the best parts of variational methods (full rank B).

The earliest hybrid method was proposed by Hamil and Snyder (2000), in which the representation of the error covariance of the prior information is a weighted combination of the flow-dependent estimate from the EnKF, $\mathbf{P}^f$, and the full rank estimate used in variational methods $\mathbf{P}^s$

$$\mathbf{P}_{hybrid} = \beta \mathbf{P}^s + (1 - \beta)\mathbf{P}^f$$

where β is a tunable parameter.

Note localisation and inflation of the ensemble are still necessary.

Summary

- Ensemble data assimilation relies on a sample estimate of the mean and covariance of forecast distribution. This allows it to provide a flow-dependent estimate of the forecast uncertainty.
- If the ensemble size is much smaller than the size of the state then sampling error becomes an issue
    - Biases
    - Analysis increments lie in the subspace of the ensemble
    - Filter divergence
    - Spurious correlations
- To make ensemble DA practical need
    - Ensemble inflation
    - Localisation
    - …Hybrid methods

Further reading

- Anderson JL, Anderson SL. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Mon. Weather Rev. 127: 2741–2758.
- Bannister et al. 2011: Ensemble prediction for nowcasting with a convection-permitting model - II: forecast error statistics, Tellus 63A, 497-512 .
- Corazza et al.. 2003. Use of the breeding technique to estimate the structure of the analysis ‘errors of the day’. Nonlinear Processes Geophys. 10: 233–243.
- Greybush, 2011: Balance and ensemble Kalman filter localisation techniques. Mon. Wea. Rev., 139, 511-522.
- Hamill, 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Mon. Wea. Rev., 129,550-560.
- Hamill, 2006: Ensemble-based atmospheric data assimilation. In *Predictability of Weather and Climate*, Palmer T, Hagedorn R (eds). Cambridge University Press: Cambridge; 124– 156.
- Houtekamer and Zhang (2016) Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation . *Mon. Weather Rev*., **144**, 4489–4532
- Houtekamer and Mitchell 1998: Data assimilation using an ensemble Kalman Filter technique, Mon. Wea. Rev. 126, 796-811.
- Houtekamer and Mitchell 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. Mon. Wea. Rev., 129, 123–137
- Hunt et al. 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. Physica D, 230, 112–126,
- Kotsuki et al. 2017: Adaptive covariance relaxation methods for ensemble data assimilation: experiments in the real atmosphere. Q.J.R.Meteorol.Soc.143: 2001 – 2015.
- Migliorini et al. 2011: Ensemble prediction for nowcasting with a convection-permitting model - I: description of the system and the impact of radar-derived surface precipitation rates, Tellus 63A, 468-496.
- Mitchell and Houtekamer 2000. An adaptive ensemble Kalman filter. Mon. Weather Rev. 128: 416–433.
- Miyoshi et al. 2014: The 10,240-member ensemble Kalman filtering with an intermediate AGCM. Geophys. Res. Lett., 41, 5264-5271.
- Whitaker and Hamill 2012. Evaluating methods to account for system errors in ensemble data assimilation. Mon. Weather Rev. 140: 3078–3089.
- Zhang et al. 2004. Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. Mon. Weather Rev. 132: 1238–1253.