Ensemble cross-validation

→ How it works
→ Why it is needed
→ How it can fail

Thomas Navarro
Joel Bedard, Mark Buehner, Arlan Dirkson, Daisuke Hotta

2026 EnKF workshop

Double EnKF

Houtekamer & Mitchell 1997

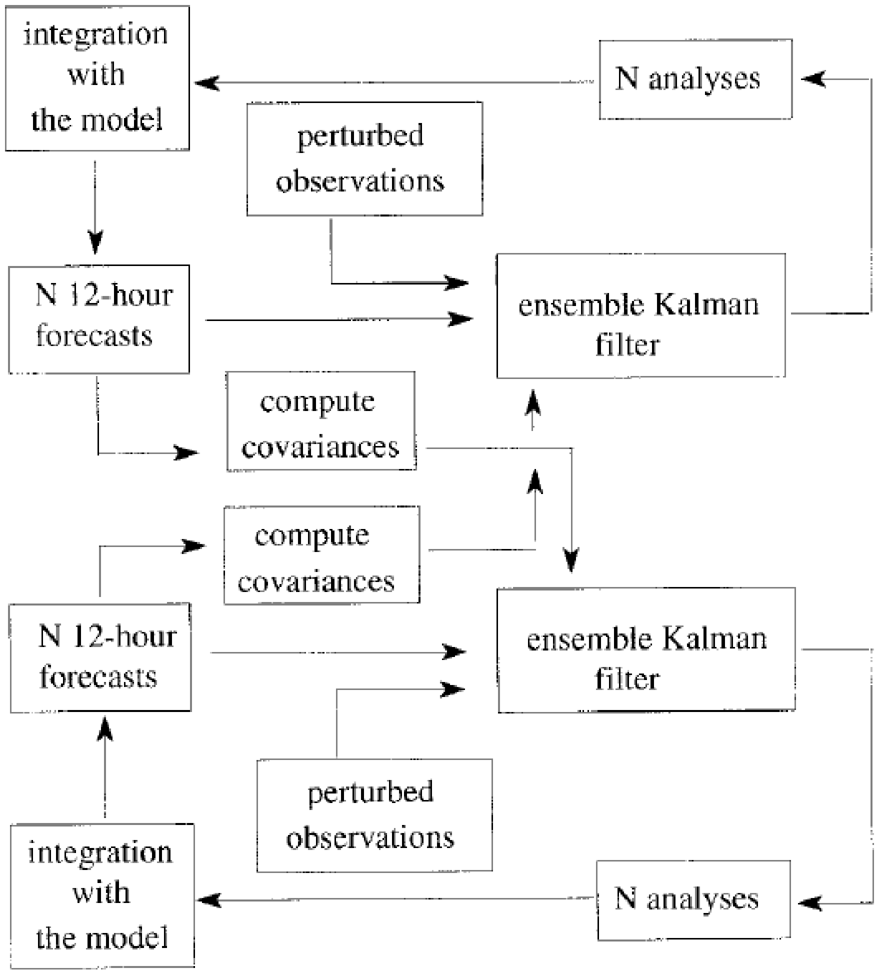
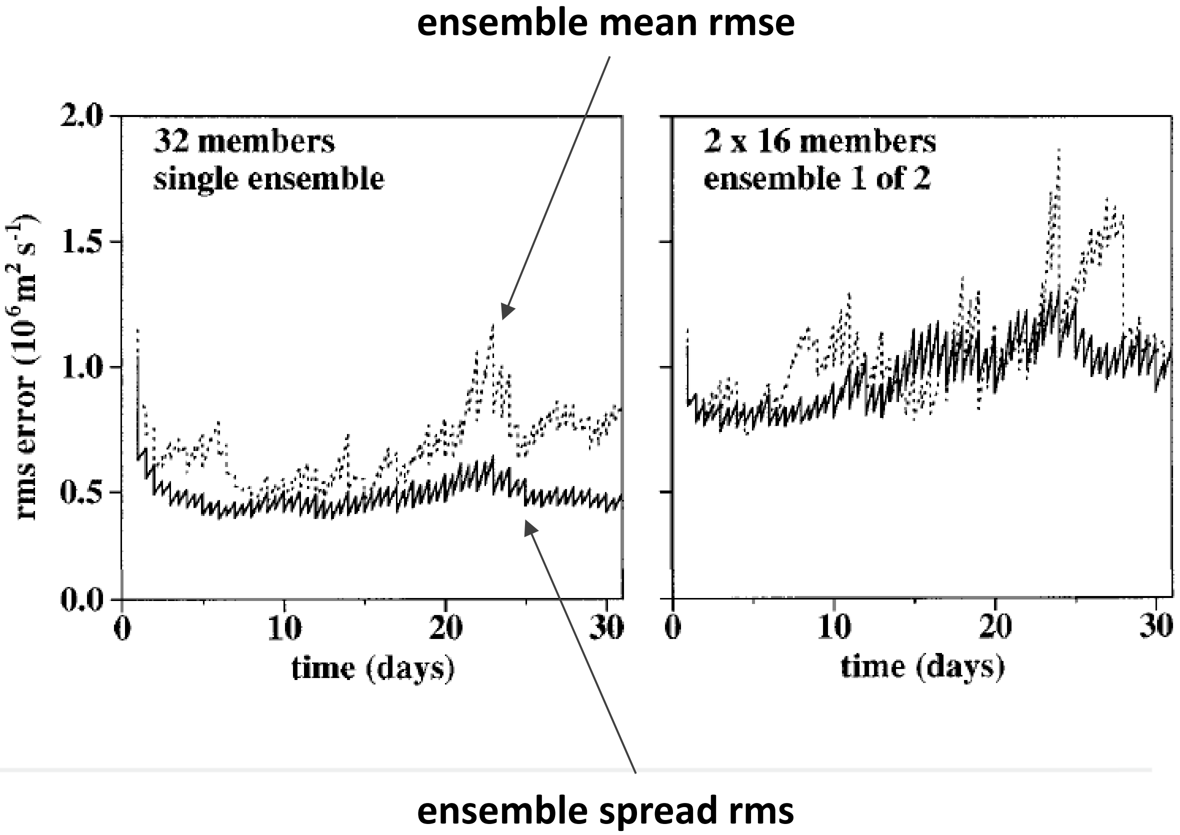


FIG. 2. A setup with a pair of ensemble Kalman filters. Data are assimilated into each ensemble using the covariances computed from the other ensemble.

__Inbreeding__: the analysis is calculated with the same ensemble perturbation used to estimate the Kalman gain.



Double EnKF

- Double EnKF improvements are entirely explained by the finite size of the ensemble. Inbreeding is just one of many terms *(van Leeuwen 1999)*

$$\mathbf{P}_1^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_1$$

$$\mathbf{P}_2^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_2$$

$$\begin{aligned}\mathbf{P}_{estimated}^{\mathrm{a}} &= \mathbf{P}_{true}^{\mathrm{a}} \\ &+ (\mathbf{I} - \mathbf{K}_2\mathbf{H})\boldsymbol{\epsilon}_2\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_2\mathbf{H}^{\mathrm{T}} + \mathbf{W}\right)^{-1}\mathbf{H}\boldsymbol{\epsilon}_1\left(\mathbf{I} - \mathbf{H}^{\mathrm{T}}\mathbf{K}_2^{\mathrm{T}}\right) \\ &+ \cdots + O(\|\boldsymbol{\epsilon}_{\mathrm{i}}\|^3)\end{aligned}$$

Double EnKF

- Double EnKF improvements are entirely explained by the finite size of the ensemble. Inbreeding is just one of many terms *(van Leeuwen 1999)*

$$\mathbf{P}_1^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_1$$

$$\mathbf{P}_2^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_2$$

$$\begin{aligned}\mathbf{P}^{\mathrm{a}}_{estimated} &= \mathbf{P}^{\mathrm{a}}_{true} \\ &+ (\mathbf{I} - \mathbf{K}_2\mathbf{H})\boldsymbol{\epsilon}_2\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_2\mathbf{H}^{\mathrm{T}} + \mathbf{W}\right)^{-1}\mathbf{H}\boldsymbol{\epsilon}_1\left(\mathbf{I} - \mathbf{H}^{\mathrm{T}}\mathbf{K}_2^{\mathrm{T}}\right) \\ &+ \cdots + O(\|\boldsymbol{\epsilon}_{\mathrm{i}}\|^3)\end{aligned}$$

- **CROSS-VALIDATION:** Generalization of the double EnKF to multiple sub-ensembles *(Houtekamer et al. 2009, Sacher & Bartello 2008)*

Double EnKF

- Double EnKF improvements are entirely explained by the finite size of the ensemble. Inbreeding is just one of many terms *(van Leeuwen 1999)*

$$\mathbf{P}_1^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_1$$

$$\mathbf{P}_2^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_2$$

$$\begin{aligned}\mathbf{P}_{estimated}^{\mathrm{a}} &= \mathbf{P}_{true}^{\mathrm{a}} \\ &+ (\mathbf{I} - \mathbf{K}_2\mathbf{H})\boldsymbol{\epsilon}_2\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_2\mathbf{H}^{\mathrm{T}} + \mathbf{W}\right)^{-1}\mathbf{H}\boldsymbol{\epsilon}_1\left(\mathbf{I} - \mathbf{H}^{\mathrm{T}}\mathbf{K}_2^{\mathrm{T}}\right) \\ &+ \cdots + O(\|\boldsymbol{\epsilon}_{\mathrm{i}}\|^3)\end{aligned}$$

- **CROSS-VALIDATION:** Generalization of the double EnKF to multiple sub-ensembles *(Houtekamer et al. 2009, Sacher & Bartello 2008)*

- Cross-validation may prevent analysis covariance error underestimation, and alleviate the need for inflation *(Houtekamer & Mitchell 1997, Houtekamer et al. 2009, Buehner 2020)*

Double EnKF

- Double EnKF improvements are entirely explained by the finite size of the ensemble. Inbreeding is just one of many terms *(van Leeuwen 1999)*

$$\mathbf{P}_1^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_1$$

$$\mathbf{P}_2^b = \mathbf{P}_{true}^b + \boldsymbol{\epsilon}_2$$

$$\begin{aligned}\mathbf{P}_{estimated}^{\mathrm{a}} &= \mathbf{P}_{true}^{\mathrm{a}} \\ &+ (\mathbf{I} - \mathbf{K}_2\mathbf{H})\boldsymbol{\epsilon}_2\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_2\mathbf{H}^{\mathrm{T}} + \mathbf{W}\right)^{-1}\mathbf{H}\boldsymbol{\epsilon}_1\left(\mathbf{I} - \mathbf{H}^{\mathrm{T}}\mathbf{K}_2^{\mathrm{T}}\right) \\ &+ \cdots + O(\|\boldsymbol{\epsilon}_{\mathrm{i}}\|^3)\end{aligned}$$

- **CROSS-VALIDATION:** Generalization of the double EnKF to multiple sub-ensembles *(Houtekamer et al. 2009, Sacher & Bartello 2008)*

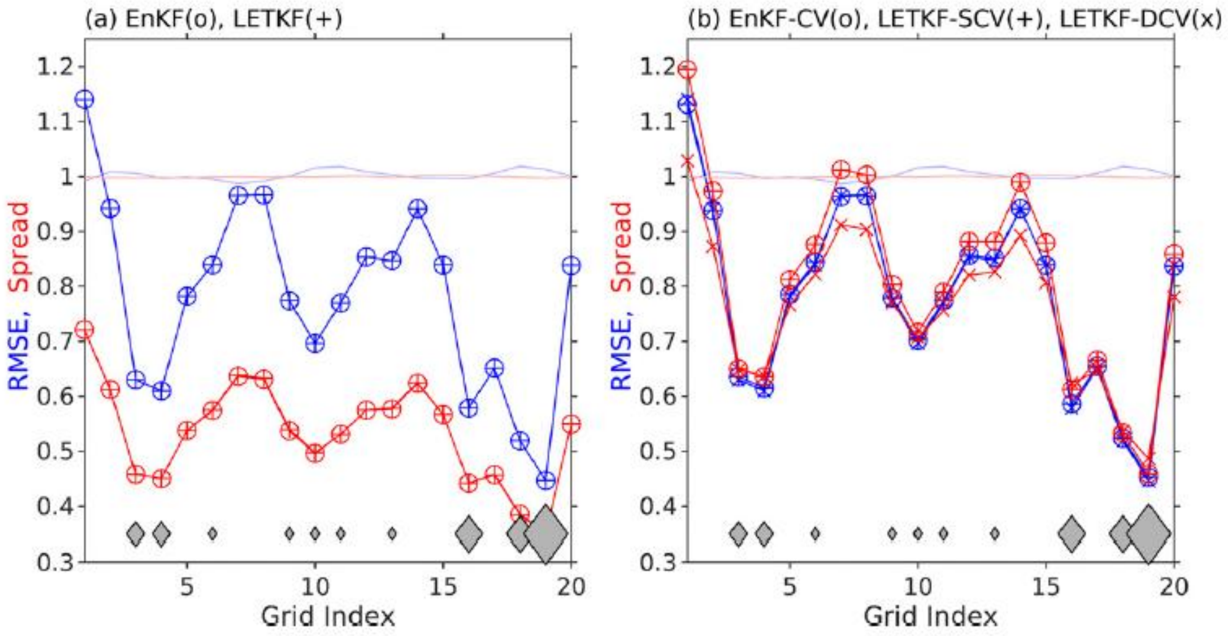
- Cross-validation may prevent analysis covariance error underestimation, and alleviate the need for inflation *(Houtekamer & Mitchell 1997, Houtekamer et al. 2009, Buehner 2020)*

- Cross-validation used continuously since 2005 for operational forecast at Environment Canada.

LETKF-DCV

Deterministic Local Ensemble Transform Kalman Filter with Cross-Validation

Formulation and test of a deterministic/stochastic LETKF with cross-validation *(Buehner et al. 2020)*, based on the Gain-form of the LETKF used for modulated ensembles *(Bishop et al. 2017, Lei et al. 2018)*



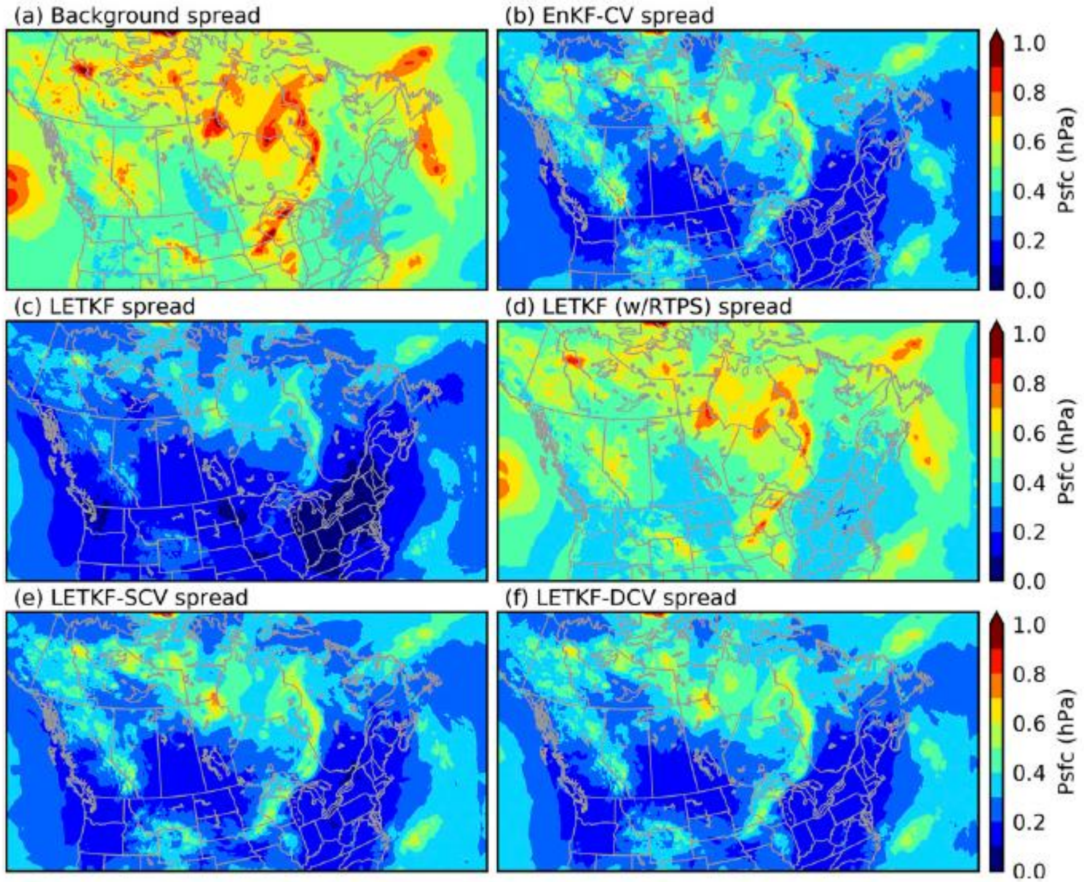


FIG. 5. Ensemble spread stddev from (a) the first background ensemble and also from the first analysis ensemble in the 10-day regional NWP ensemble DA experiments produced by (b) EnKF-CV, (c) LETKF, (d) LETKF with RTPS, (e) LETKF-SCV, and (f) LETKF-DCV for surface pressure (in units of hPa).

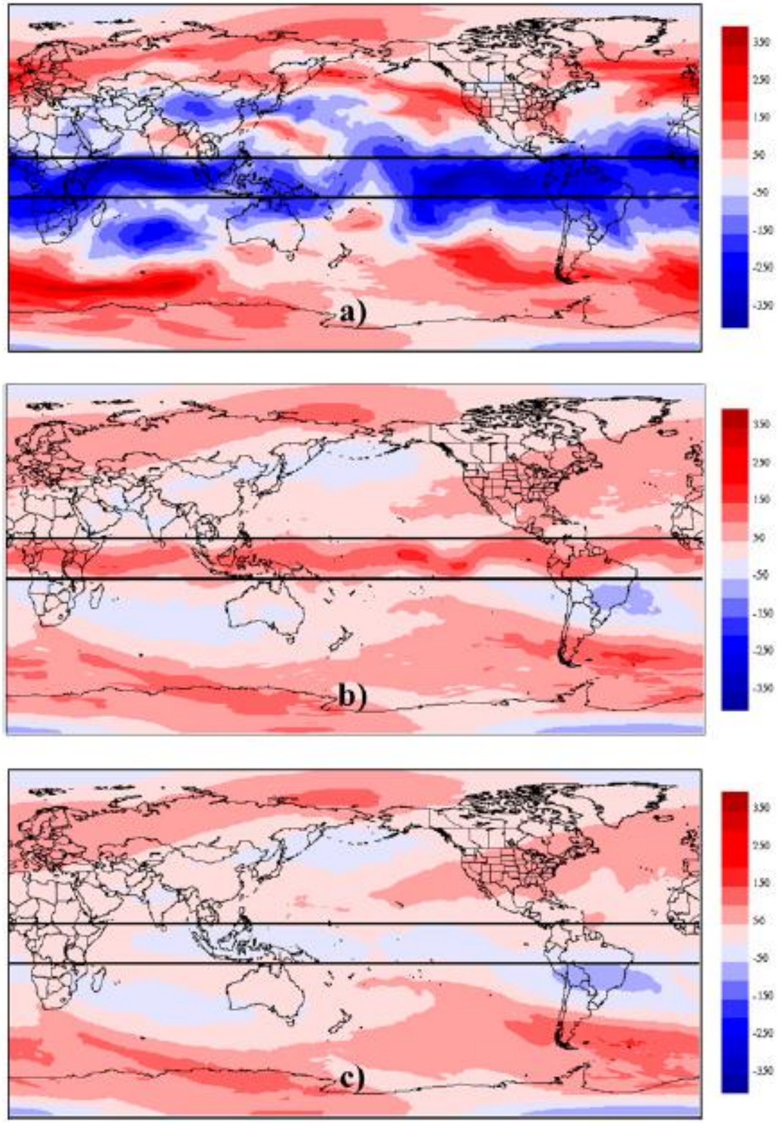


FIG. 1. Zonal wind (kt) at ~3 hPa valid at 1200 UTC 12 Oct 2021. (a) Background state from ensemble member 254 of a new version of the GEPS; (b) as in (a), but for ensemble member 234; and (c) the analysis from the operational GDPS. The black lines correspond to the parallels at 10°N and 10°S.

LETKF-DCV: Operational application

(Caron et al. 2022)

- 256 members split in 8 sub-ensembles
- One sub-ensemble catastrophically diverged with possibles causes:
  - Deterministic algorithm
  - Low amount of observations
  - Model sensitivity

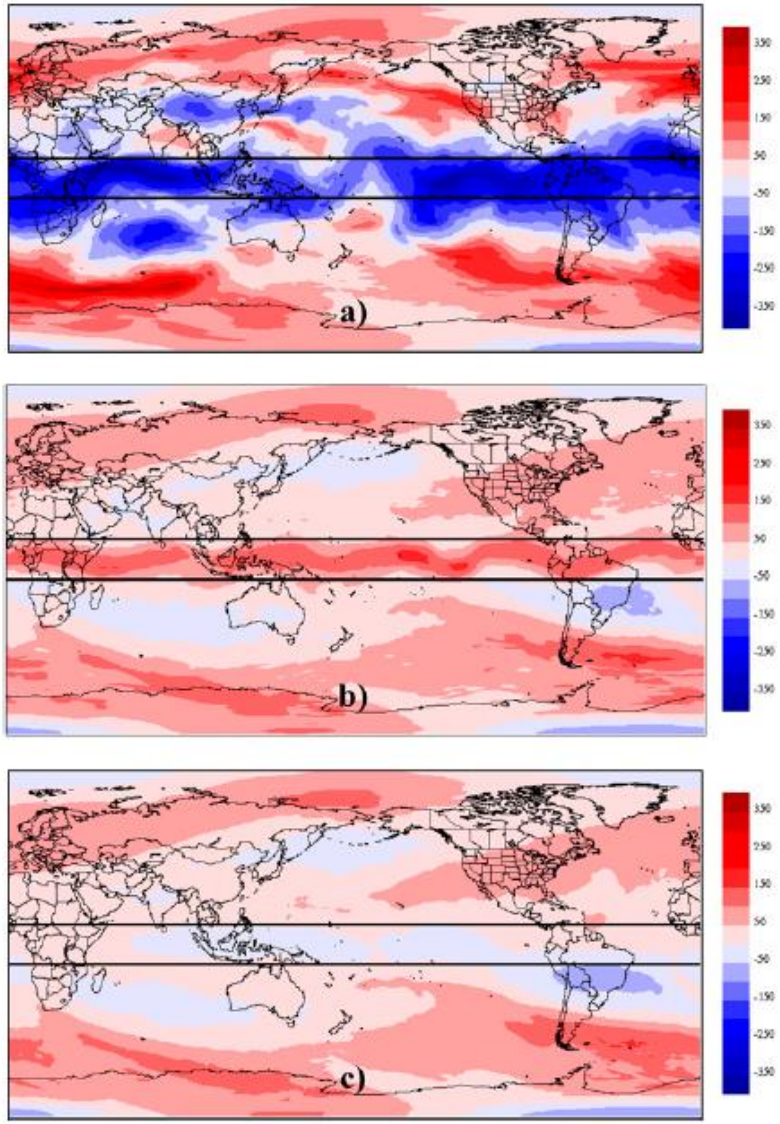


FIG. 1. Zonal wind (kt) at ~3 hPa valid at 1200 UTC 12 Oct 2021. (a) Background state from ensemble member 254 of a new version of the GEPS; (b) as in (a), but for ensemble member 234; and (c) the analysis from the operational GDPS. The black lines correspond to the parallels at 10°N and 10°S.

LETKF-DCV: Operational application

(Caron et al. 2022)

- 256 members split in 8 sub-ensembles
- One sub-ensemble catastrophically diverged with possibles causes:
  - Deterministic algorithm
  - Low amount of observations
  - Model sensitivity
- Solution: Randomize the selection of sub-ensembles at each analysis step
- Both the LETKF-DCV and stochastic EnKF-CV exhibited several divergence events that self-resolved.

Main questions

1. What does cross-validation do?

2. How does it address inbreeding?

3. Why can it fail?

Stochastic EnKF - Notation

Model size	M
Ensemble size	N
Number of observations	P
Observation error	**R**
Background perturbation in model space	$\delta\mathbf{x}_\mathrm{i}^\mathrm{b} = \mathbf{x}_\mathrm{i}^\mathrm{b} - \overline{\mathbf{x}^\mathrm{b}}$
Background perturbation in observation space	$\delta\mathbf{y}_\mathrm{i}^\mathrm{b} = \mathrm{H}(\mathbf{x}_\mathrm{i}^\mathrm{b}) - \overline{\mathrm{H}(\mathbf{x}^\mathrm{b})}$
Matrix of background perturbations	$\mathbf{X}^\mathrm{b} = (N-1)^{-1/2}[\ldots\ \delta\mathbf{x}_\mathrm{i}^\mathrm{b}\ \ldots]$
Matrix of background perturbations in observation space	$\mathbf{Y}^\mathrm{b} = (N-1)^{-1/2}[\ldots\ \delta\mathbf{y}_\mathrm{i}^\mathrm{b}\ \ldots]$

Stochastic EnKF equations

(Bishop et al. 2017, Hotta & Ota 2021, Snyder & Hakim 2022, etc …)

Singular value decomposition

$$\mathbf{R}^{-1/2}\mathbf{Y}^{\mathrm{b}} = \mathbf{U}\mathbf{\Sigma}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}$$

$$\mathbf{\Sigma}^{\mathrm{b}} = \mathrm{diag}(\sigma_{\mathrm{i}}^{\mathrm{b}})$$

Stochastic EnKF equations

(Bishop et al. 2017, Hotta & Ota 2021, Snyder & Hakim 2022, etc …)

Singular value decomposition

$$\mathbf{R}^{-1/2}\mathbf{Y}^{\mathrm{b}} = \mathbf{U}\boldsymbol{\Sigma}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}$$

$$\boldsymbol{\Sigma}^{\mathrm{b}} = \mathrm{diag}(\sigma_{\mathrm{i}}^{\mathrm{b}})$$

Kalman gain

$$\mathbf{K} = \mathbf{X}^{\mathrm{b}}\left(\mathbf{Y}^{\mathrm{b}}\right)^{\mathrm{T}}\left(\mathbf{Y}^{\mathrm{b}}\left(\mathbf{Y}^{\mathrm{b}}\right)^{\mathrm{T}} + \mathbf{R}\right)^{-1}$$

$$\mathbf{K} = \mathbf{X}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{b}}\left(1 + \left(\boldsymbol{\Sigma}^{\mathrm{b}}\right)^{2}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}$$

Stochastic EnKF equations

(Bishop et al. 2017, Hotta & Ota 2021, Snyder & Hakim 2022, etc …)

Singular value decomposition

$$\mathbf{R}^{-1/2}\mathbf{Y}^{\mathrm{b}} = \mathbf{U}\boldsymbol{\Sigma}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}$$

$$\boldsymbol{\Sigma}^{\mathrm{b}} = \mathrm{diag}(\sigma_{\mathrm{i}}^{\mathrm{b}})$$

Kalman gain

$$\mathbf{K} = \mathbf{X}^{\mathrm{b}}\left(\mathbf{Y}^{\mathrm{b}}\right)^{\mathrm{T}}\left(\mathbf{Y}^{\mathrm{b}}\left(\mathbf{Y}^{\mathrm{b}}\right)^{\mathrm{T}} + \mathbf{R}\right)^{-1}$$

$$\mathbf{K} = \mathbf{X}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{b}}\left(1 + \left(\boldsymbol{\Sigma}^{\mathrm{b}}\right)^{2}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}$$

Innovation for i^{th} member

$$\mathbf{d}_{\mathrm{i}} = \mathbf{y}^{\mathrm{o}} + \boldsymbol{\epsilon}_{\mathrm{i}}^{\mathrm{o}} - \mathrm{H}\left(\mathbf{x}_{\mathrm{i}}^{\mathrm{b}}\right)$$

Projection in ensemble space

$$\widetilde{\mathbf{d}}_{\mathrm{i}} = \mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{d}_{\mathrm{i}}$$

Stochastic EnKF equations

(Bishop et al. 2017, Hotta & Ota 2021, Snyder & Hakim 2022, etc …)

Singular value decomposition

$$\mathbf{R}^{-1/2}\mathbf{Y}^{\mathrm{b}} = \mathbf{U}\mathbf{\Sigma}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}$$

$$\mathbf{\Sigma}^{\mathrm{b}} = \mathrm{diag}(\sigma_{\mathrm{i}}^{\mathrm{b}})$$

Kalman gain

$$\mathbf{K} = \mathbf{X}^{\mathrm{b}}\left(\mathbf{Y}^{\mathrm{b}}\right)^{\mathrm{T}}\left(\mathbf{Y}^{\mathrm{b}}\left(\mathbf{Y}^{\mathrm{b}}\right)^{\mathrm{T}} + \mathbf{R}\right)^{-1}$$

$$\mathbf{K} = \mathbf{X}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{\mathrm{b}}\left(1 + \left(\mathbf{\Sigma}^{\mathrm{b}}\right)^{2}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}$$

Innovation for i^{th} member

$$\mathbf{d}_{\mathrm{i}} = \mathbf{y}^{\mathrm{o}} + \boldsymbol{\epsilon}_{\mathrm{i}}^{\mathrm{o}} - \mathrm{H}\left(\mathbf{x}_{\mathrm{i}}^{\mathrm{b}}\right)$$

Projection in ensemble space

$$\widetilde{\mathbf{d}}_{\mathrm{i}} = \mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{d}_{\mathrm{i}}$$

Analysis increment for i^{th} member

$$\mathbf{x}_{\mathrm{i}}^{\mathrm{a}} - \mathbf{x}_{\mathrm{i}}^{\mathrm{b}} = \mathbf{K}\,\mathbf{d}_{\mathrm{i}}$$

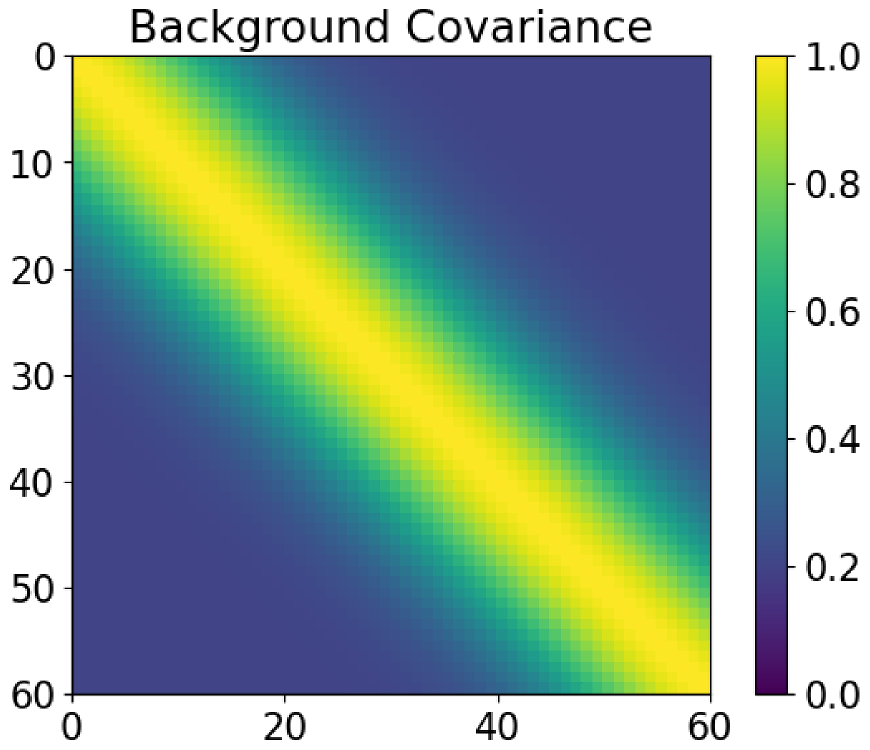
$$\mathbf{x}_{\mathrm{i}}^{\mathrm{a}} - \mathbf{x}_{\mathrm{i}}^{\mathrm{b}} = \mathbf{X}^{\mathrm{b}}\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}^{\mathrm{b}}\left(1 + \mathbf{\Sigma}^{\mathrm{b}}\left(\mathbf{\Sigma}^{\mathrm{b}}\right)^{\mathrm{T}}\right)^{-1}\widetilde{\mathbf{d}}_{\mathrm{i}}$$

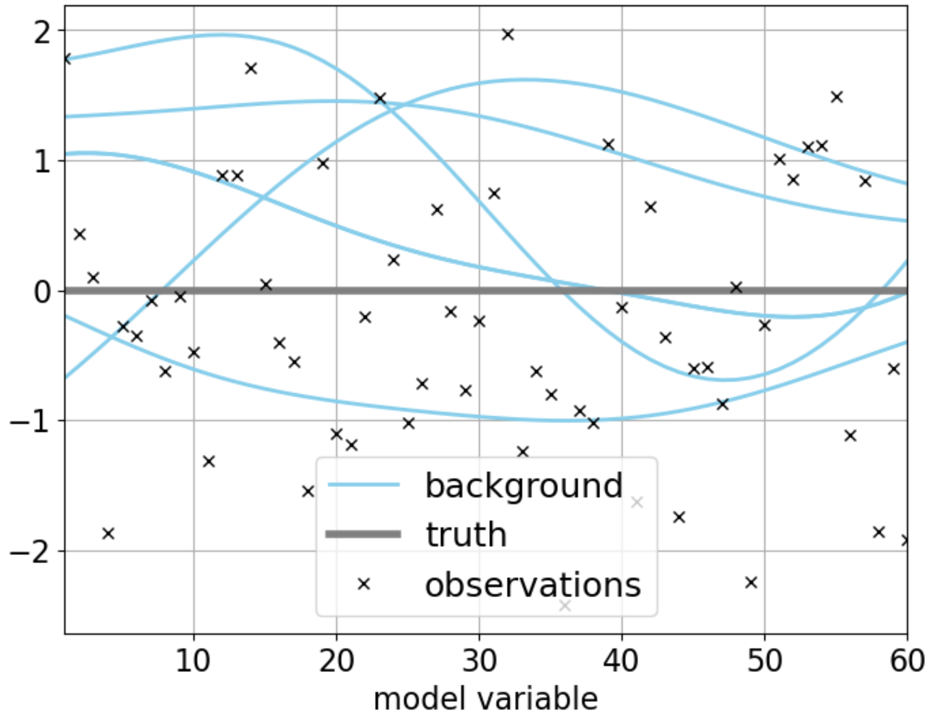
Idealized experiment

- $\mathbf{R} = \mathbf{I}$
- Prescribed background covariance matrix
- Observation operator is the identity; one observation per grid point
- <u>Goal:</u> Assess ONE analysis step with and without cross-validation

Idealized experiment

- $\mathbf{R} = \ \mathbf{I}$
- Prescribed background covariance matrix
- Observation operator is the identity; one observation per grid point
- <u>Goal:</u> Assess ONE analysis step with and without cross-validation





Innovation in model space

M = 60 ; 2000 iterations

Innovation for i^{th} member

$$\mathbf{d}_{\mathrm{i}} = \mathbf{y}^{\mathrm{o}} + \boldsymbol{\epsilon}_{\mathrm{i}}^{\mathrm{o}} - \mathrm{H}\left(\mathbf{x}_{\mathrm{i}}^{\mathrm{b}}\right)$$

$$\widetilde{\mathbf{d}}_{\mathrm{i}} = \mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{d}_{\mathrm{i}}$$

$$\mathbf{d}_{i\parallel} = \mathbf{R}^{1/2}\mathbf{U}\widetilde{\mathbf{d}}_{\mathrm{i}}$$

$$\mathbf{d}_{\mathrm{i}} = \mathbf{d}_{i\parallel} + \mathbf{d}_{i\perp}$$

Innovation in model space

M = 60 ; 2000 iterations

Innovation for ith member

$$\mathbf{d}_{\mathrm{i}} = \mathbf{y}^{\mathrm{o}} + \boldsymbol{\epsilon}_{\mathrm{i}}^{\mathrm{o}} - \mathrm{H}\left(\mathbf{x}_{\mathrm{i}}^{\mathrm{b}}\right)$$

$$\widetilde{\mathbf{d}}_{\mathrm{i}} = \mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{d}_{\mathrm{i}}$$

$$\mathbf{d}_{i\parallel} = \mathbf{R}^{1/2}\mathbf{U}\widetilde{\mathbf{d}}_{\mathrm{i}}$$

$$\mathbf{d}_{\mathrm{i}} = \mathbf{d}_{i\parallel} + \mathbf{d}_{i\perp}$$

Background perturbation for ith member

$$\mathbf{d}_{\mathrm{i}} = \mathbf{y}^{\mathrm{o}} + \boldsymbol{\epsilon}_{\mathrm{i}}^{\mathrm{o}} - \overline{\mathbf{y}^{\mathrm{b}}} + \delta\mathbf{y}_{\mathrm{i}}^{\mathrm{b}}$$

$$\widetilde{\mathbf{d}}_{\mathrm{i}} = \mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\left(\mathbf{y}^{\mathrm{o}} + \boldsymbol{\epsilon}_{\mathrm{i}}^{\mathrm{o}} - \overline{\mathbf{y}^{\mathrm{b}}}\right) + \widetilde{\delta\mathbf{y}_{\mathrm{i}}^{\mathrm{b}}}$$

$$\widetilde{\delta\mathbf{y}_{\mathrm{i}}^{\mathrm{b}}} = \mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\delta\mathbf{y}_{\mathrm{i}}^{\mathrm{b}}$$

$$\delta\mathbf{y}_{\mathrm{i}\parallel}^{\mathrm{b}} = \mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_{\mathrm{i}}^{\mathrm{b}}}$$

Innovation in model space

M = 60 ; 2000 iterations

Innovation for ith member

$$\mathbf{d}_\mathrm{i} = \mathbf{y}^\mathrm{o} + \boldsymbol{\epsilon}_\mathrm{i}^\mathrm{o} - \mathrm{H}\left(\mathbf{x}_\mathrm{i}^\mathrm{b}\right)$$

$$\tilde{\mathbf{d}}_\mathrm{i} = \mathbf{U}^\mathrm{T}\mathbf{R}^{-1/2}\mathbf{d}_\mathrm{i}$$

$$\mathbf{d}_{i\parallel} = \mathbf{R}^{1/2}\mathbf{U}\tilde{\mathbf{d}}_\mathrm{i}$$

$$\mathbf{d}_\mathrm{i} = \mathbf{d}_{i\parallel} + \mathbf{d}_{i\perp}$$

Background perturbation for ith member

$$\mathbf{d}_\mathrm{i} = \mathbf{y}^\mathrm{o} + \boldsymbol{\epsilon}_\mathrm{i}^\mathrm{o} - \overline{\mathbf{y}^\mathrm{b}} + \delta\mathbf{y}_\mathrm{i}^\mathrm{b}$$

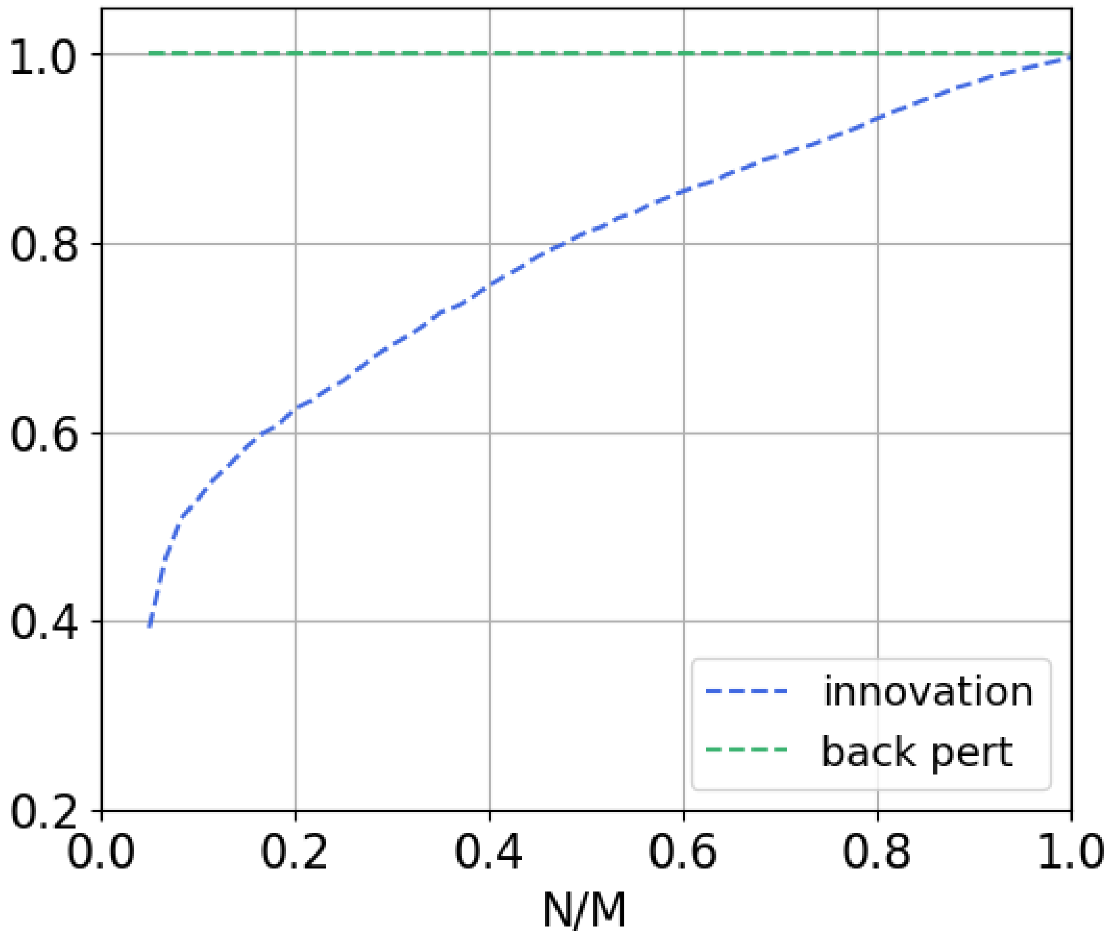
$$\tilde{\mathbf{d}}_\mathrm{i} = \mathbf{U}^\mathrm{T}\mathbf{R}^{-1/2}\left(\mathbf{y}^\mathrm{o} + \boldsymbol{\epsilon}_\mathrm{i}^\mathrm{o} - \overline{\mathbf{y}^\mathrm{b}}\right) + \widetilde{\delta\mathbf{y}_\mathrm{i}^\mathrm{b}}$$

$$\widetilde{\delta\mathbf{y}_\mathrm{i}^\mathrm{b}} = \mathbf{U}^\mathrm{T}\mathbf{R}^{-1/2}\delta\mathbf{y}_\mathrm{i}^\mathrm{b}$$

$$\delta\mathbf{y}_{\mathrm{i}\parallel}^\mathrm{b} = \mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_\mathrm{i}^\mathrm{b}}$$

$$\frac{\left\|\mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_\mathrm{i}^\mathrm{b}}\right\|}{\left\|\delta\mathbf{y}_\mathrm{i}^\mathrm{b}\right\|}$$

$$\frac{\left\|\mathbf{R}^{1/2}\mathbf{U}\tilde{\mathbf{d}}_\mathrm{i}\right\|}{\left\|\mathbf{d}_\mathrm{i}\right\|}$$



Innovation in model space

M = 60 ; 2000 iterations

Innovation for i^{th} member

$$\mathbf{d}_i = \mathbf{y}^o + \epsilon_i^o - H(\mathbf{x}_i^b)$$

$$\tilde{\mathbf{d}}_i = \mathbf{U}^T\mathbf{R}^{-1/2}\mathbf{d}_i$$

$$\mathbf{d}_{i\parallel} = \mathbf{R}^{1/2}\mathbf{U}\tilde{\mathbf{d}}_i$$

$$\mathbf{d}_i = \mathbf{d}_{i\parallel} + \mathbf{d}_{i\perp}$$

Background perturbation for i^{th} member

$$\mathbf{d}_i = \mathbf{y}^o + \epsilon_i^o - \overline{\mathbf{y}^b} + \delta\mathbf{y}_i^b$$

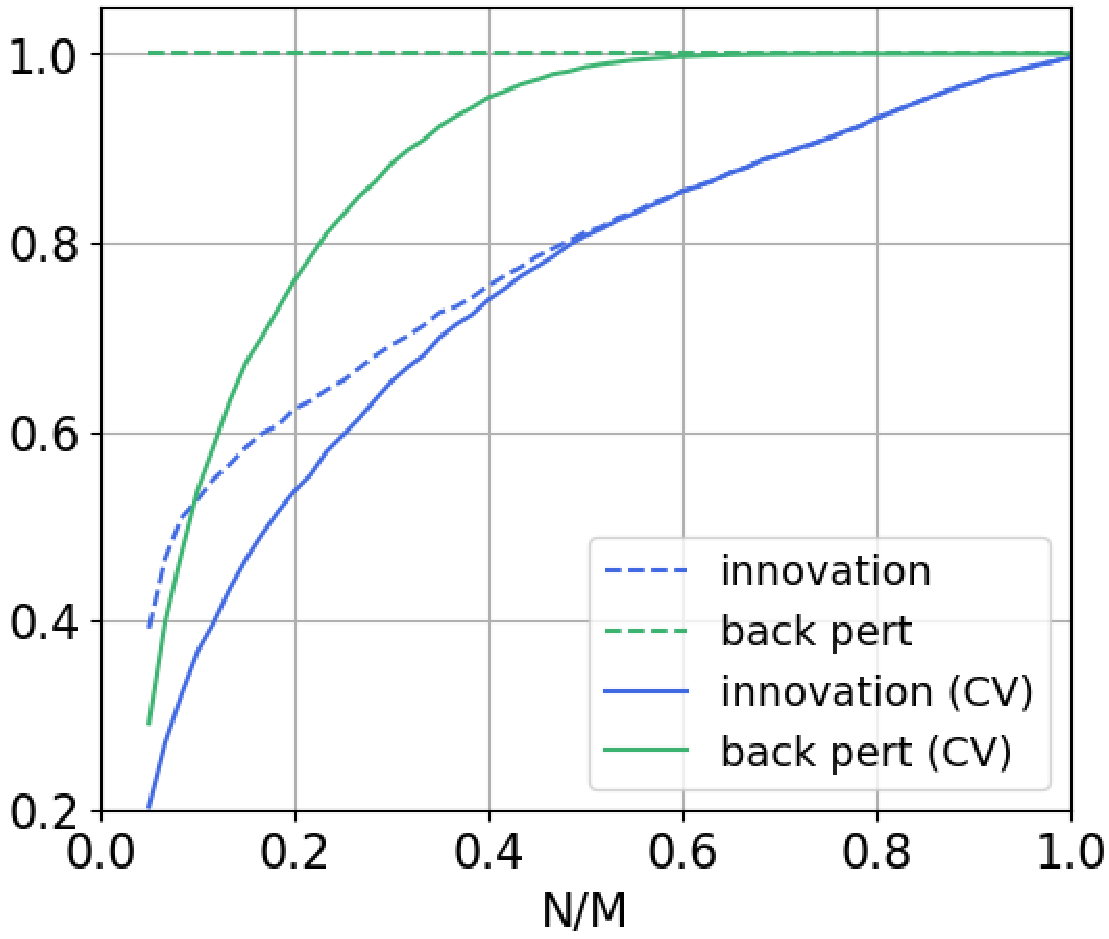
$$\tilde{\mathbf{d}}_i = \mathbf{U}^T\mathbf{R}^{-1/2}\left(\mathbf{y}^o + \epsilon_i^o - \overline{\mathbf{y}^b}\right) + \widetilde{\delta\mathbf{y}_i^b}$$

$$\widetilde{\delta\mathbf{y}_i^b} = \mathbf{U}^T\mathbf{R}^{-1/2}\delta\mathbf{y}_i^b$$

$$\delta\mathbf{y}_{i\parallel}^b = \mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_i^b}$$

$$\frac{\left\|\mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_i^b}\right\|}{\left\|\delta\mathbf{y}_i^b\right\|}$$

$$\frac{\left\|\mathbf{R}^{1/2}\mathbf{U}\tilde{\mathbf{d}}_i\right\|}{\left\|\mathbf{d}_i\right\|}$$



Innovation in ensemble space

Analysis increment for i^{th} member

$$\mathbf{x}_\mathrm{i}^\mathrm{a} - \mathbf{x}_\mathrm{i}^\mathrm{b} = \mathbf{K}\,\mathbf{d}_\mathrm{i}$$

$$\mathbf{x}_\mathrm{i}^\mathrm{a} - \mathbf{x}_\mathrm{i}^\mathrm{b} = \mathbf{R}^{1/2}\mathbf{U}\boldsymbol{\Lambda}^\mathrm{a}\widetilde{\mathbf{d}}_\mathrm{i}$$

Analysis error covariance eigenvalues

$$\boldsymbol{\Lambda}^\mathrm{a} = \left(\boldsymbol{\Sigma}^\mathrm{b}\right)^2\left(1+\left(\boldsymbol{\Sigma}^\mathrm{b}\right)^2\right)^{-1} = \mathrm{diag}\left(\frac{{\sigma_\mathrm{i}^\mathrm{b}}^2}{1+{\sigma_\mathrm{i}^\mathrm{b}}^2}\right)$$

$$\mathrm{DFS} = \mathrm{trace}(\boldsymbol{\Lambda}^\mathrm{a})$$

Innovation in ensemble space

Analysis increment for i^th member

$$\mathbf{x}_i^a - \mathbf{x}_i^b = \mathbf{K}\,\mathbf{d}_i$$

$$\mathbf{x}_i^a - \mathbf{x}_i^b = \mathbf{R}^{1/2}\mathbf{U}\boldsymbol{\Lambda}^a\widetilde{\mathbf{d}}_i$$

Analysis perturbation for i^th member

$$\delta\mathbf{x}_i^a = \mathbf{x}_i^a - \overline{\mathbf{x}^a}$$

$$\delta\mathbf{x}_i^a = (\mathbf{I} - \mathbf{K})\delta\mathbf{y}_i^b + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\delta\mathbf{x}_i^a = (\mathbf{I} - \mathbf{K})\mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\delta\mathbf{x}_i^a = \mathbf{R}^{1/2}\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda}^a)\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$$

Analysis error covariance eigenvalues

$$\boldsymbol{\Lambda}^a = \left(\boldsymbol{\Sigma}^b\right)^2\left(1 + \left(\boldsymbol{\Sigma}^b\right)^2\right)^{-1} = \operatorname{diag}\left(\frac{{\sigma_i^b}^2}{1 + {\sigma_i^b}^2}\right)$$

$$\mathrm{DFS} = \operatorname{trace}(\boldsymbol{\Lambda}^a)$$

Innovation in ensemble space

M = 60 ; 2000 iterations

Analysis increment for ith member

$\mathbf{x}_i^a - \mathbf{x}_i^b = \mathbf{K}\,\mathbf{d}_i$

$\mathbf{x}_i^a - \mathbf{x}_i^b = \mathbf{R}^{1/2}\mathbf{U}\boldsymbol{\Lambda}^a\widetilde{\mathbf{d}}_i$

Analysis perturbation for ith member

$\delta\mathbf{x}_i^a = \mathbf{x}_i^a - \overline{\mathbf{x}^a}$

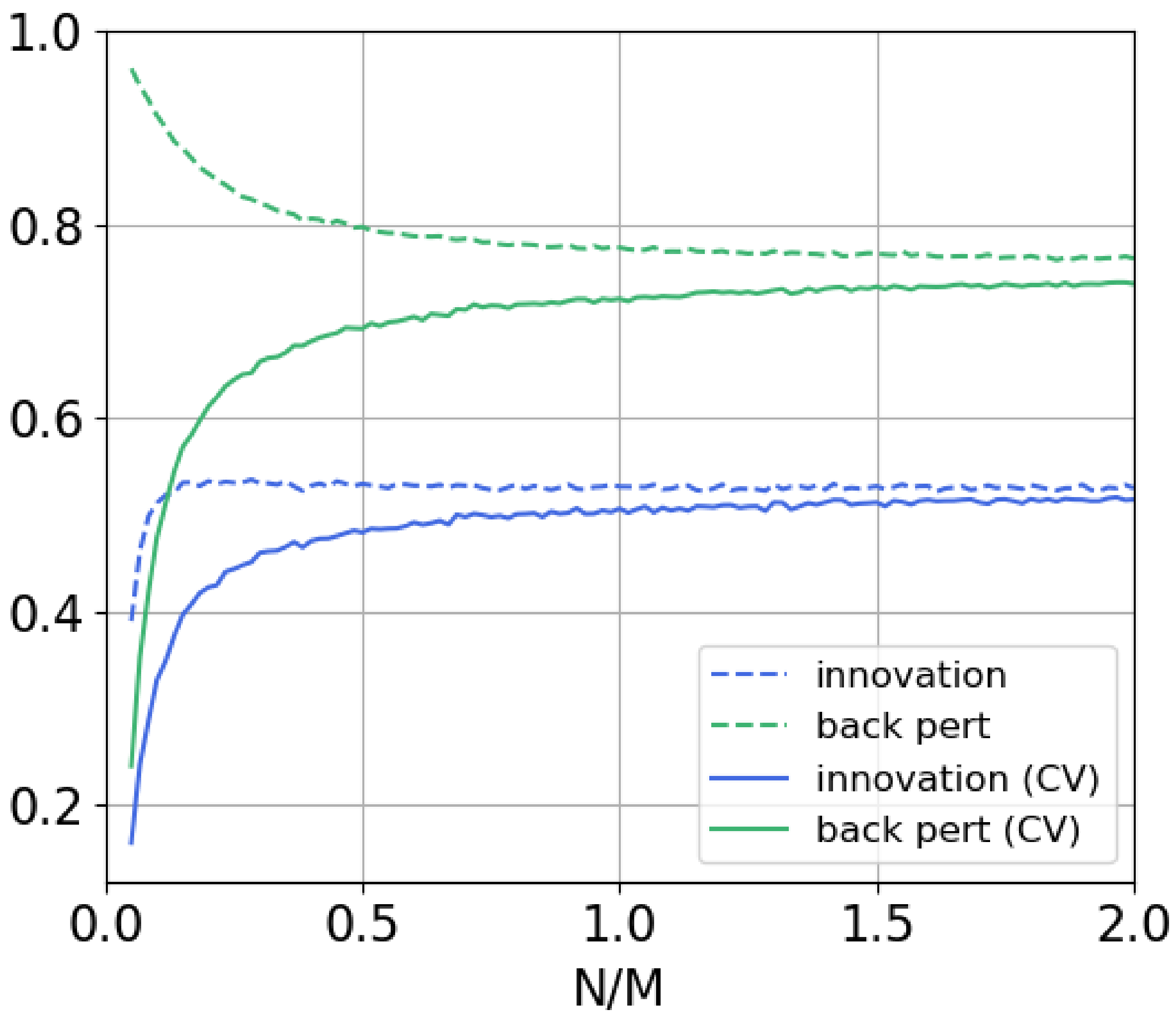
$\delta\mathbf{x}_i^a = (\mathbf{I}-\mathbf{K})\delta\mathbf{y}_i^b + \mathbf{K}\,\boldsymbol{\epsilon}$

$\delta\mathbf{x}_i^a = (\mathbf{I}-\mathbf{K})\mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$

$\delta\mathbf{x}_i^a = \mathbf{R}^{1/2}\mathbf{U}(\mathbf{I}-\boldsymbol{\Lambda}^a)\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$

$$\frac{\left\|\boldsymbol{\Lambda}^a\widetilde{\delta\mathbf{y}_i^b}\right\|}{\left\|\widetilde{\delta\mathbf{y}_i^b}\right\|}$$

$$\frac{\left\|\boldsymbol{\Lambda}^a\widetilde{\mathbf{d}}_i\right\|}{\left\|\widetilde{\mathbf{d}}_i\right\|}$$



Innovation in ensemble space

M = 60 ; 2000 iterations

Analysis increment for ith member

$$\mathbf{x}_i^a - \mathbf{x}_i^b = \mathbf{K}\,\mathbf{d}_i$$

$$\mathbf{x}_i^a - \mathbf{x}_i^b = \mathbf{R}^{1/2}\mathbf{U}\boldsymbol{\Lambda}^a\widetilde{\mathbf{d}}_i$$

Analysis perturbation for ith member

$$\delta\mathbf{x}_i^a = \mathbf{x}_i^a - \overline{\mathbf{x}^a}$$

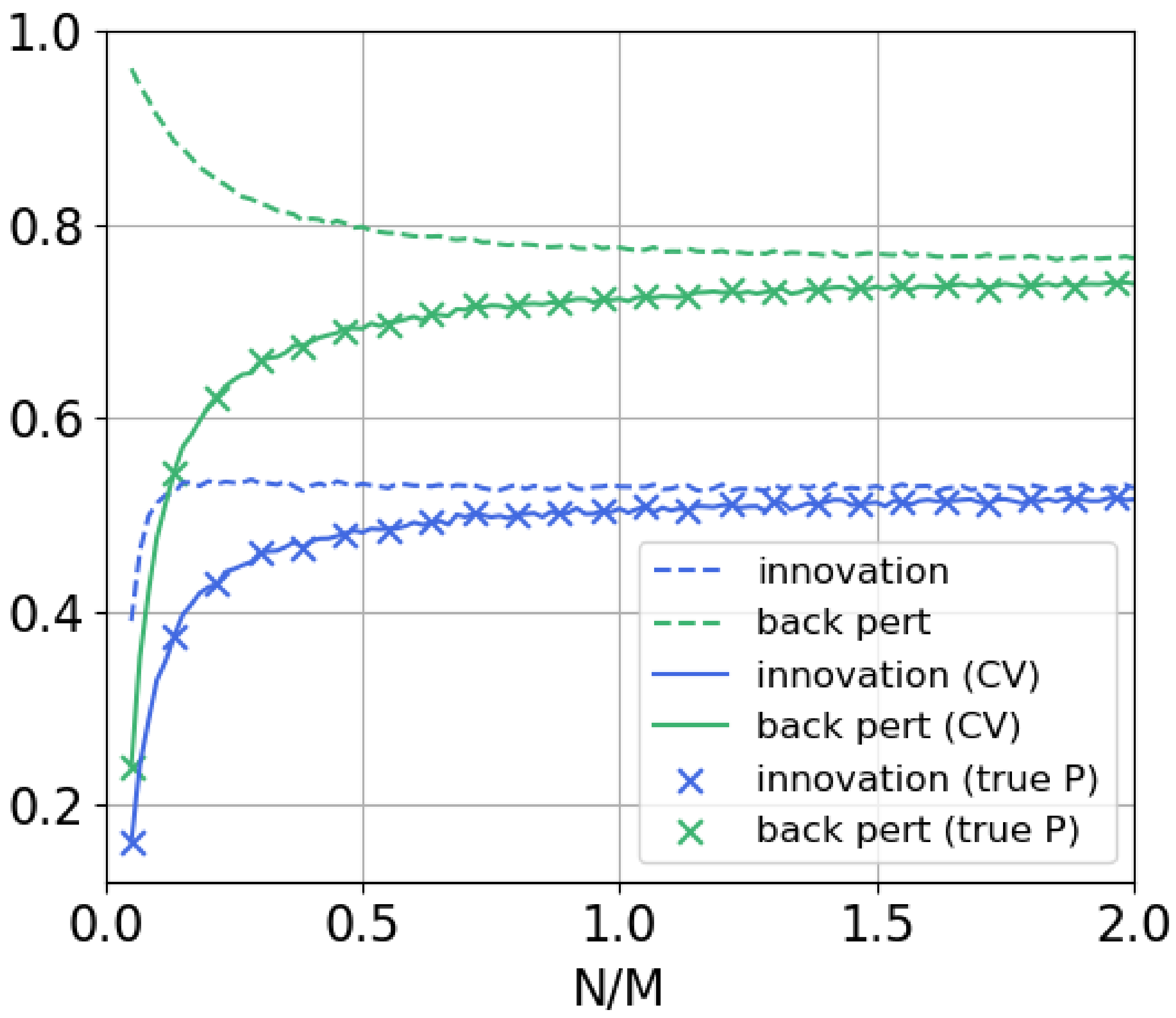
$$\delta\mathbf{x}_i^a = (\mathbf{I}-\mathbf{K})\delta\mathbf{y}_i^b + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\delta\mathbf{x}_i^a = (\mathbf{I}-\mathbf{K})\mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\delta\mathbf{x}_i^a = \mathbf{R}^{1/2}\mathbf{U}(\mathbf{I}-\boldsymbol{\Lambda}^a)\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\frac{\left\|\boldsymbol{\Lambda}^a\widetilde{\delta\mathbf{y}_i^b}\right\|}{\left\|\widetilde{\delta\mathbf{y}_i^b}\right\|}$$

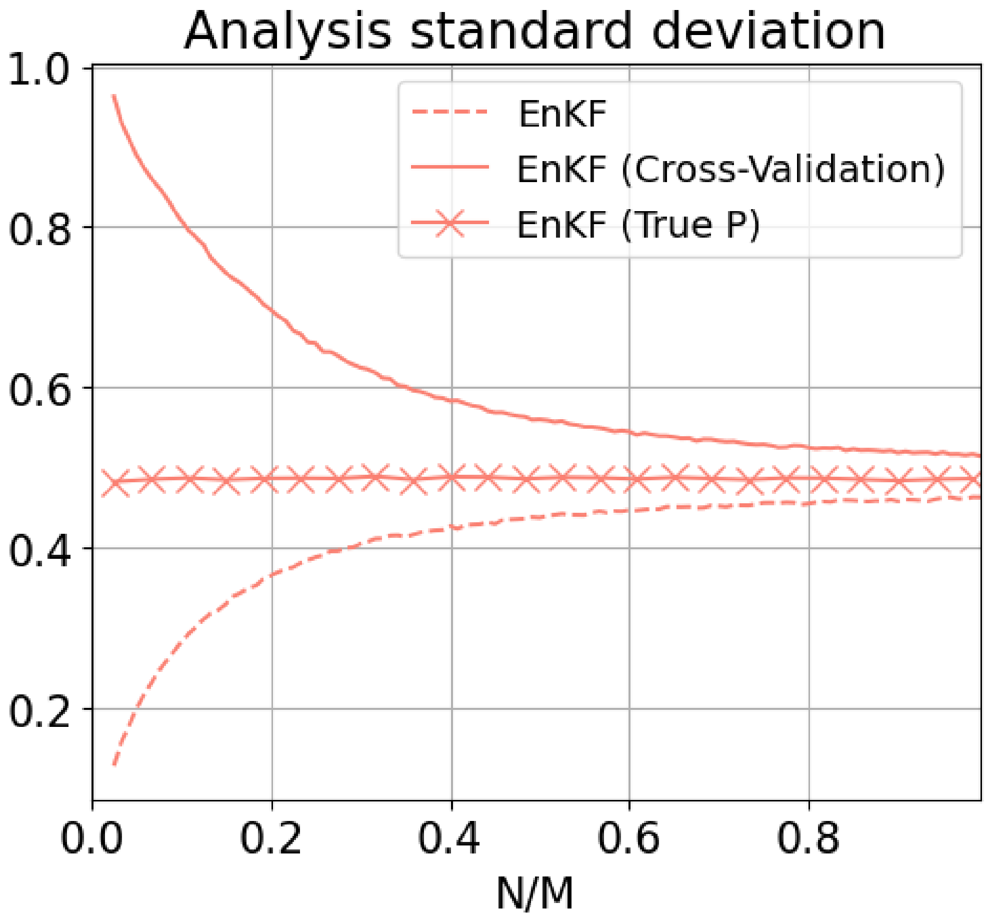
$$\frac{\left\|\boldsymbol{\Lambda}^a\widetilde{\mathbf{d}}_i\right\|}{\left\|\widetilde{\mathbf{d}}_i\right\|}$$



Analysis ensemble

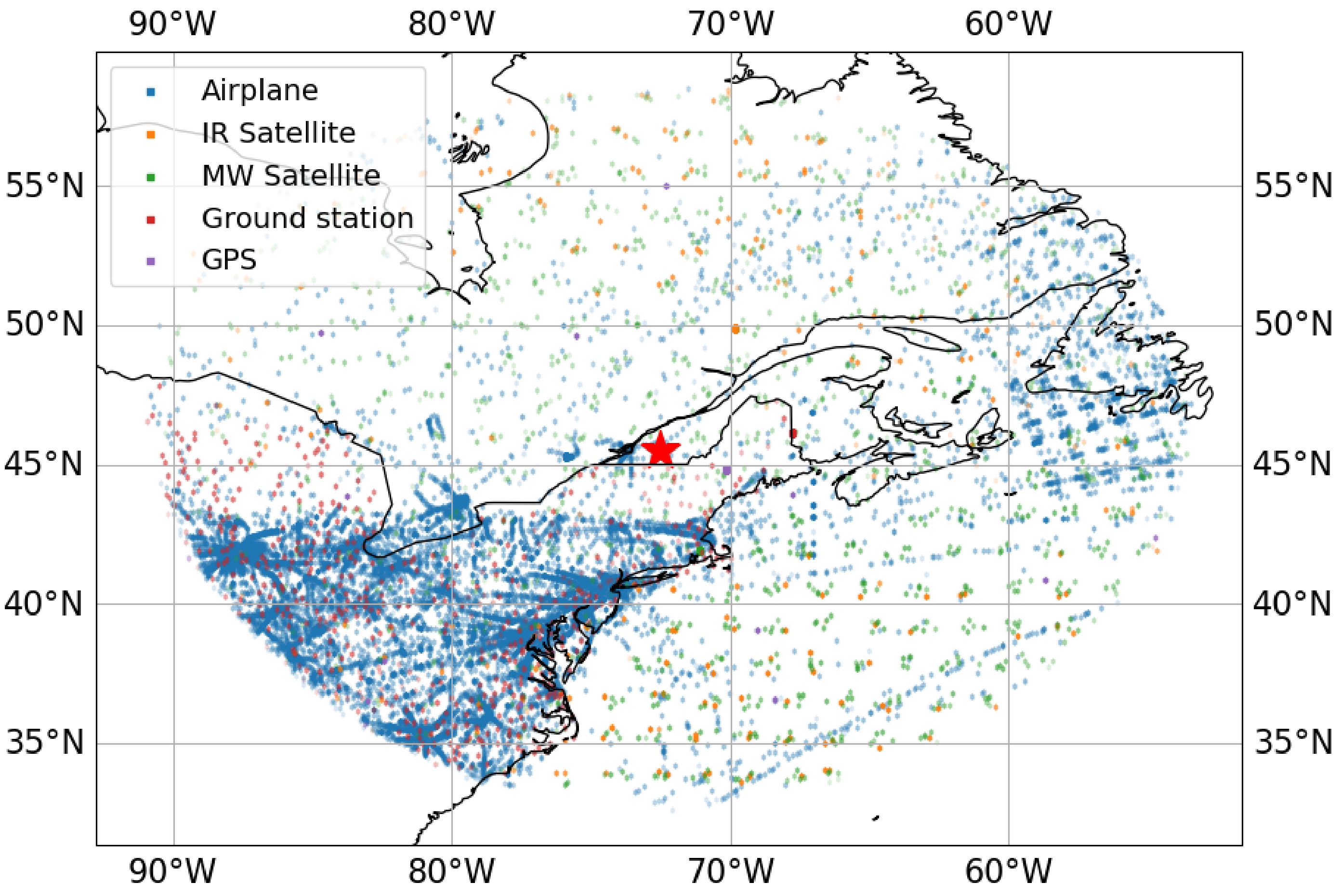
M = 60 ; 1000 iterations

Analysis perturbation for i$^{\text{th}}$ member	$\delta \mathbf{x}_i^a$
Analysis mean for i$^{\text{th}}$ member	$\overline{\overline{\mathbf{x}_i^a}} = \mathbf{K}_{j\neq i}\,\overline{\mathbf{d}_{j\neq i}}$
Analysis mean	$\overline{\mathbf{x}_a} = \frac{1}{N}\sum_i \overline{\overline{\mathbf{x}_i^a}}$
$(N-1)\sigma_{a,CV}^2 = \sum_i (\delta \mathbf{x}_i^a)^2 + \sum_i \left(\overline{\overline{\mathbf{x}_i^a}} - \overline{\mathbf{x}_a}\right)^2 + 2\sum_i \left(\overline{\overline{\mathbf{x}_i^a}} - \overline{\mathbf{x}_a}\right)\delta \mathbf{x}_i^a$	
For any distribution, $\mathrm{Var}\left(\overline{\overline{\mathbf{x}_i^a}} - \overline{\mathbf{x}_a}\right) = \frac{N-1}{N}\sigma_{a,true}^2$	
$\mathbb{E}[\sigma_{a,CV}^2] = \left(1 + \frac{1}{N}\right)\mathbb{E}[\sigma_{a,true}^2]$	



GEPS: Global Ensemble Prediction System

Hybrid EnVar-LETKF



LETKF-DCV

$N = 256$

8 sub-ensembles

$M \sim P \sim 10^4$

$N/M \sim 2x10^{-2}$

Stochastic EnKF with cross-validation

Analysis perturbation for i$^{\mathrm{th}}$ member

$$\delta\mathbf{x}_{\mathrm{i}}^{\mathrm{a}} = \mathbf{x}_{\mathrm{i}}^{\mathrm{a}} - \overline{\mathbf{x}^{\mathrm{a}}}$$

$$\delta\mathbf{x}_{\mathrm{i}}^{\mathrm{a}} = (1 - \mathbf{K})\delta\mathbf{y}_{\mathrm{i}}^{\mathrm{b}} + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\delta\mathbf{x}_{\mathrm{i}}^{\mathrm{a}} = \delta\mathbf{x}_{\mathrm{i}}^{\mathrm{b}} - \mathbf{R}^{1/2}\mathbf{U}\boldsymbol{\Sigma}_{\mathrm{c}}^{\mathrm{b}}\left(\mathbf{I} + \left(\boldsymbol{\Sigma}_{\mathrm{c}}^{\mathrm{b}}\right)^{2}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\delta\mathbf{y}_{\mathrm{i}}^{\mathrm{b}} + \mathbf{K}\,\boldsymbol{\epsilon}$$

Deterministic LETKF with cross-validation

Analysis perturbation for sub-ensemble $\mathbf{Y}_{\mathrm{s}}^{\mathrm{b}}$ from its complement $\mathbf{Y}_{\mathrm{c}}^{\mathrm{b}}$

$$\mathbf{X}_{\mathrm{s}}^{\mathrm{a}} = \mathbf{X}_{\mathrm{s}}^{\mathrm{b}} - \mathbf{X}_{\mathrm{c}}^{\mathrm{b}}\mathbf{V}_{\mathrm{c}}\mathbf{D}\left(\boldsymbol{\Sigma}_{\mathrm{c}}^{\mathrm{b}}\right)^{-1}\mathbf{U}_{\mathrm{c}}^{\mathrm{T}}\mathbf{U}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}^{\mathrm{b}}\mathbf{V}_{\mathrm{s}}^{\mathrm{T}}$$

$$\mathbf{D} = (N_c - 1)^{-1}\left(\mathbf{I} - \left(\mathbf{I} + \left(\boldsymbol{\Sigma}_{\mathrm{c}}^{\mathrm{b}}\right)^{2}\right)^{-1/2}\right)$$

$$0 \leqslant d_{ii} \leqslant (N_c - 1)^{-1}$$

Conclusions

Inbreeding happens for low N/M ratio
The analysis ensemble spread is **underestimated.**

Conclusions

Inbreeding happens for low N/M ratio
The analysis ensemble spread is **underestimated.**

Cross-validation solves inbreeding.
Cross-validation gives the **expected analysis perturbation.**
The expected analysis spread is **overestimated by a simple scalar term.**

Conclusions

Inbreeding happens for low N/M ratio
The analysis ensemble spread is **underestimated.**

Cross-validation solves inbreeding.
Cross-validation gives the **expected analysis perturbation.**
The expected analysis spread is **overestimated by a simple scalar term.**

LETKF with deterministic cross-validation is more complicated.
The analysis ensemble spread depends on the singular values of the sub-ensemble and its complement.

Conclusions

Inbreeding happens for low N/M ratio
The analysis ensemble spread is **underestimated.**

Cross-validation solves inbreeding.
Cross-validation gives the **expected analysis perturbation.**
The expected analysis spread is **overestimated by a simple scalar term.**

LETKF with deterministic cross-validation is more complicated.
The analysis ensemble spread depends on the singular values of the sub-ensemble and its complement.

Open questions
Assess cross-validation impact on analysis mean.
Test against real case in GEPS.
Exact conditions for cross-validation divergence (observation density, sensitivity, randomization of sub-ensembles).

Two job opportunities at Environment Canada

- One postdoctoral fellow
- One research scientist

- Expertise in one or more of the following areas:
  - Ensemble and/or variational data assimilation methodologies
  - Remote sensing, with emphasis on radiative transfer processes and algorithms
  - Machine learning and artificial intelligence applications in geosciences
  - Observation quality control and bias correction techniques

- Contact Dr. Joel Bedard (*joel.bedard@ec.gc.ca*)

Perturbation in ensemble space

M = 120 ; N = 30 ; 2000 iterations

With cross-validation, the background perturbation is spread across more coordinates in ensemble space.

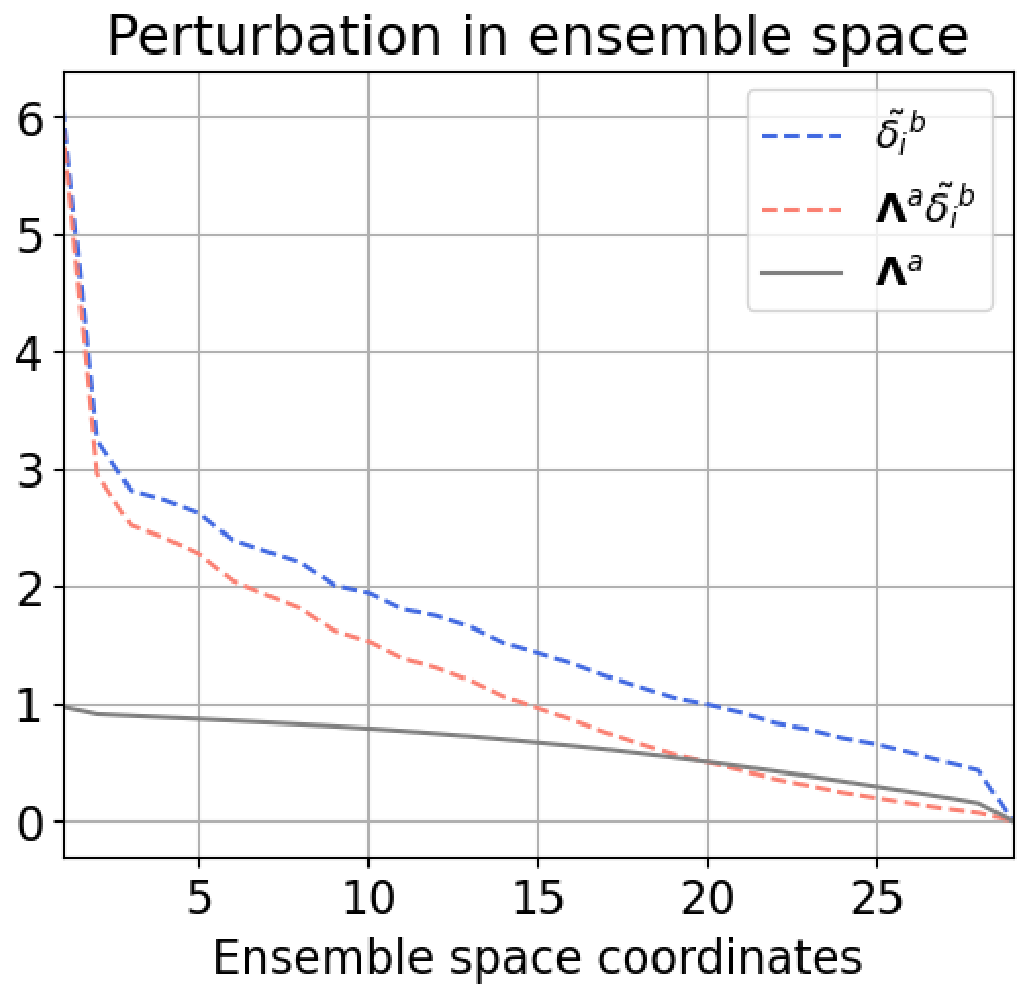
Analysis perturbation for i^{th} member

$$\delta\mathbf{x}_i^a = \mathbf{x}_i^a - \overline{\mathbf{x}^a}$$

$$\delta\mathbf{x}_i^a = (\mathbf{I} - \mathbf{K})\delta\mathbf{y}_i^b + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\delta\mathbf{x}_i^a = (\mathbf{I} - \mathbf{K})\mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$$

$$\delta\mathbf{x}_i^a = \mathbf{R}^{1/2}\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda}^a)\widetilde{\delta\mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$$



Perturbation in ensemble space

M = 120 ; N = 30 ; 2000 iterations

With cross-validation, the background perturbation is spread across more coordinates in ensemble space.

Analysis perturbation for i^th^ member

$\delta \mathbf{x}_i^a = \mathbf{x}_i^a - \overline{\mathbf{x}^a}$

$\delta \mathbf{x}_i^a = (\mathbf{I} - \mathbf{K})\delta \mathbf{y}_i^b + \mathbf{K}\,\boldsymbol{\epsilon}$

$\delta \mathbf{x}_i^a = (\mathbf{I} - \mathbf{K})\mathbf{R}^{1/2}\mathbf{U}\widetilde{\delta \mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$

$\delta \mathbf{x}_i^a = \mathbf{R}^{1/2}\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda}^a)\widetilde{\delta \mathbf{y}_i^b} + \mathbf{K}\,\boldsymbol{\epsilon}$