Environment and Climate Change Canada | Environnement et Changement climatique Canada

# Potential and limitations of assimilating multiple collocated datasets with “optimal” error estimates

Statistical view


Annika Vogel[1,2] and Richard Ménard[1]

[1] Air Quality Research Division, Environment and Climate Change Canada

[2] Rhenish Institute for Environmental Research (RIU) at the University of Cologne, Germany

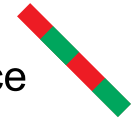

Canada

BIRS workshop, 2023-03-23

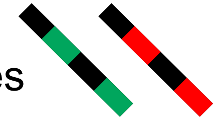
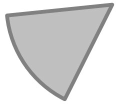
# THE ERROR ESTIMATION PROBLEM IN DATA ASSIMILATION

- <u>2 datasets</u>

+1

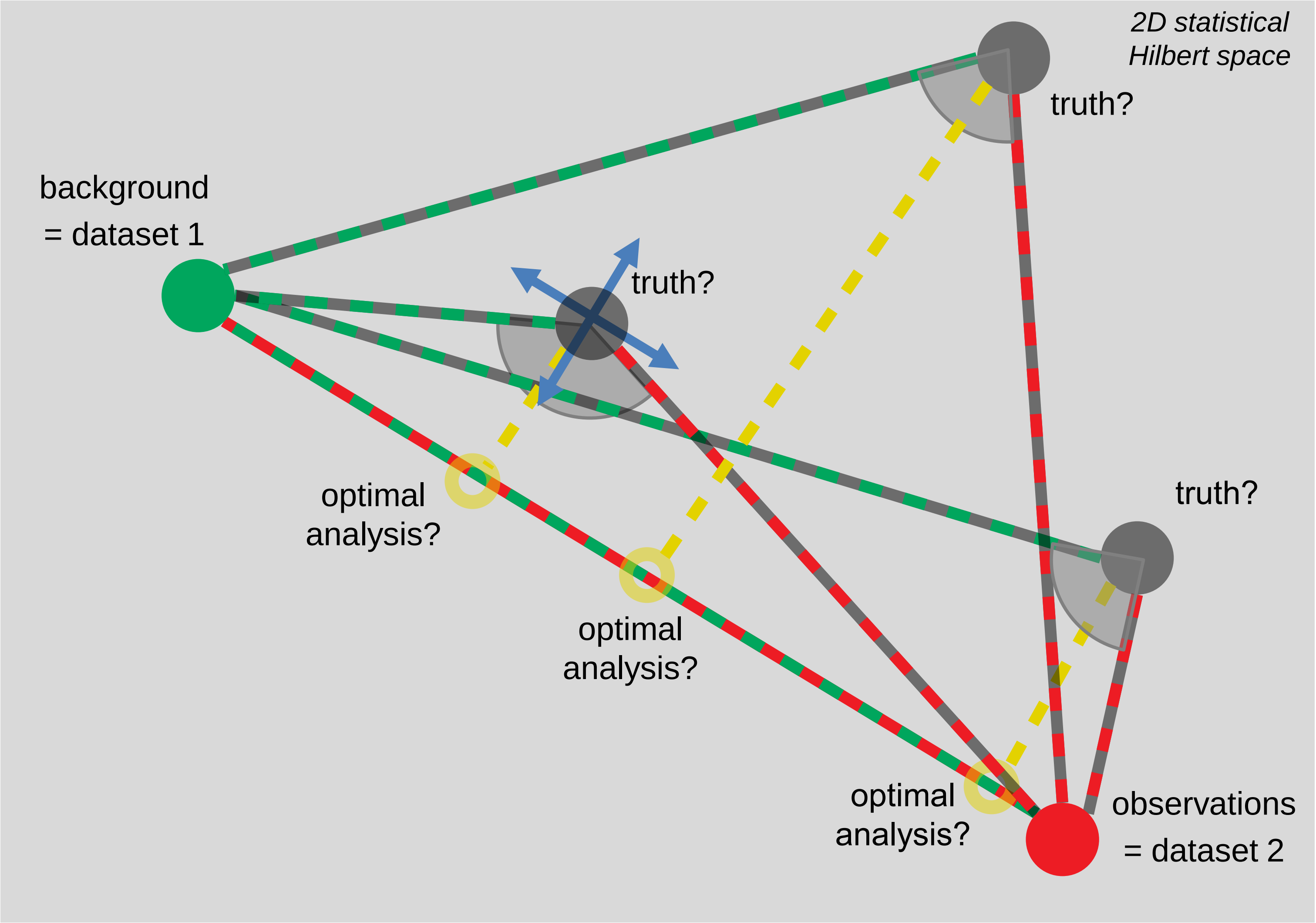
- Given:
  - 1 innovation covariance 

-3

- Unknown:
  - 2 error covariances 
  - 1 error cross-correlations 

-2

Underdetermined!




[Vogel & Ménard, egusphere-2022-996]

# THE ERROR ESTIMATION PROBLEM IN DATA ASSIMILATION

- <u>2 datasets</u>

- Given:
  +1 – 1 innovation covariance

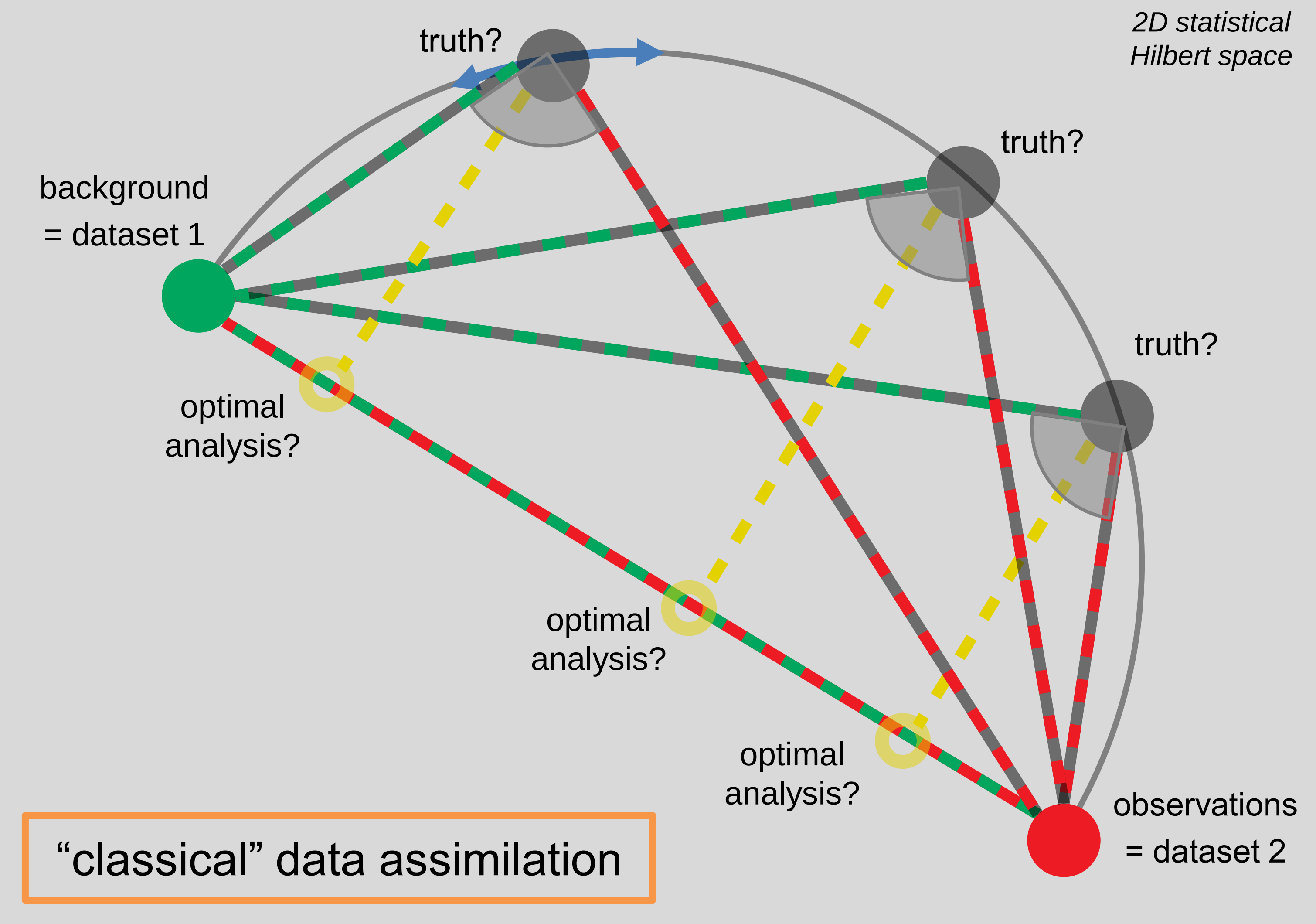
- Unknown:
  -3 – 2 error covariances
  – 1 error cross-correlations

- Assumptions:
  +1 – uncorrelated errors

-1

Still underdetermined!

– but analysis error only function of analysis state




[Vogel & Ménard, egusphere-2022-996]

# ANALYSIS WITH CORRELATED ERRORS – KF EQUATIONS

- Standard KF analysis equations (for 2 datasets) assume uncorrelated errors between BG and OBS:

$$x_a = (\mathbb{1} - \mathbf{K}_1\mathbf{H}_1)\, x_b + \mathbf{K}_1 y_1$$

$$\mathbf{K}_1 = \mathbf{B}\mathbf{H}_1^T \left(\mathbf{H}_1\mathbf{B}\mathbf{H}_1^T + \mathbf{R}_1\right)^{-1}$$

$$\mathbf{A} = (\mathbb{1} - \mathbf{K}_1\mathbf{H}_1)\,\mathbf{B}\left(\mathbb{1} - \mathbf{H}_1^T\mathbf{K}_1^T\right) + \mathbf{K}_1\mathbf{R}_1\mathbf{K}_1^T$$

$x_b$ background, $y_1$ observation, $x_a$ analysis state
**B** background, $\mathbf{R}_1$ observation, **A** analysis error covariance
$\mathbf{H}_1$ observation operator, $\mathbf{K}_1$ Kalman Gain

What if the errors are not fully uncorrelated?

- KF analysis equations with correlations (for 2 datasets)*:

error cross-covariance: $\boxed{\mathbf{X}_{b;1} := \overline{\epsilon_B \cdot \epsilon_1^T}}$

$$x_a = (\mathbb{1} - \mathbf{K}_1\mathbf{H}_1)\, x_b + \mathbf{K}_1 y_1$$

$$\mathbf{K}_1 = \left(\mathbf{B}\mathbf{H}_1^T \underline{- \mathbf{X}_{b;1}}\right)\left(\mathbf{H}_1\mathbf{B}\mathbf{H}_1^T \underline{- \mathbf{H}_1\mathbf{X}_{b;1} - \mathbf{X}_{b;1}^T\mathbf{H}_1^T} + \mathbf{R}_1\right)^{-1}$$

$$\mathbf{A} = (\mathbb{1} - \mathbf{K}_1\mathbf{H}_1)\,\mathbf{B}\left(\mathbb{1} - \mathbf{H}_1^T\mathbf{K}_1^T\right) \underline{+ (\mathbb{1} - \mathbf{K}_1\mathbf{H}_1)\,\mathbf{X}_{b;1}\mathbf{K}_1^T + \mathbf{K}_1\mathbf{X}_{b;1}^T\left(\mathbb{1} - \mathbf{H}_1^T\mathbf{K}_1^T\right)} + \mathbf{K}_1\mathbf{R}_1\mathbf{K}_1^T$$

– cross-covariance terms in Kalman gain and analysis error covariance matrix

* partly equivalent to serially correlated observation errors [Daley 1992, MWR,v120,pp.164]

# ANALYSIS WITH CORRELATED ERRORS – KF EQUATIONS

- Standard KF analysis equations (for 2 datasets) assume uncorrelated errors between BG and OBS:

$x_b$ background, $y_1$ observation, $x_a$ analysis state
**B** background, $\mathbf{R}_1$ observation, **A** analysis error covariance
$\mathbf{H}_1$ observation operator, $\mathbf{K}_1$ Kalman Gain

innovation covariance: $\boxed{\mathbf{\Gamma}_1 := \mathbf{H}_1\mathbf{B}\mathbf{H}_1^T + \mathbf{R}_1}$

$$\mathbf{K}_1 = \mathbf{B}\mathbf{H}_1^T\mathbf{\Gamma}_1^{-1} \qquad \mathbf{A}^{-1} = \mathbf{B}^{-1} + \mathbf{H}_1^T\mathbf{R}_1^{-1}\mathbf{H}_1$$

What if the errors are not fully uncorrelated?

- KF analysis equations with correlations (for 2 datasets)*:

substituted error covariances: $\boxed{\tilde{\mathbf{B}} := \mathbf{B} - \mathbf{X}_{b;1}\mathbf{H}_1^{+T} \quad \tilde{\mathbf{R}}_1 := \mathbf{R}_1 - \mathbf{X}_{b;1}^T\mathbf{H}_1^T \quad \tilde{\mathbf{A}} := \mathbf{A} - \mathbf{X}_{b;1}\mathbf{H}_1^{+T}}$

\+ pseudo-inverse

generalized innovation covariance: $\boxed{\begin{aligned}\tilde{\mathbf{\Gamma}}_1 &:= \mathbf{H}_1\mathbf{B}\mathbf{H}_1^T \underline{- \mathbf{H}_1\mathbf{X}_{b;1} - \mathbf{X}_{b;1}^T\mathbf{H}_1^T} + \mathbf{R}_1 \\ &= \mathbf{H}_1\tilde{\mathbf{B}}\mathbf{H}_1^T + \tilde{\mathbf{R}}_1\end{aligned}}$

asymmetry (of error cross-covariance): $\boxed{\mathbf{Y}_{b;1} := \mathbf{X}_{b;1} - \mathbf{H}_1^+\mathbf{X}_{b;1}^T\mathbf{H}_1^T}$

$$\mathbf{K}_1 = \tilde{\mathbf{B}}\mathbf{H}_1^T\tilde{\mathbf{\Gamma}}_1^{-1} \qquad \tilde{\mathbf{A}}^{-1} = \tilde{\mathbf{B}}^{-1} + \mathbf{H}_1^T\tilde{\mathbf{R}}_1^{-T}\mathbf{H}_1 \underline{- \mathbf{H}_1^T\tilde{\mathbf{R}}_1^{-T}\mathbf{Y}_{b;1}^T\tilde{\mathbf{B}}^{-1}}$$

– equivalent form of substituted Kalman gain

– additional "asymmetry-term" in substituted analysis error covariance

* partly equivalent to serially correlated observation errors [Daley 1992, MWR,v120,pp.164]

# ANALYSIS WITH CORRELATED ERRORS - SENSITIVITIES

- Sensitivity of KF analysis variance to error correlation:
  - scalar (**H** → 1) $x_{b;1} = \rho\, \sigma_b\, \sigma_1$

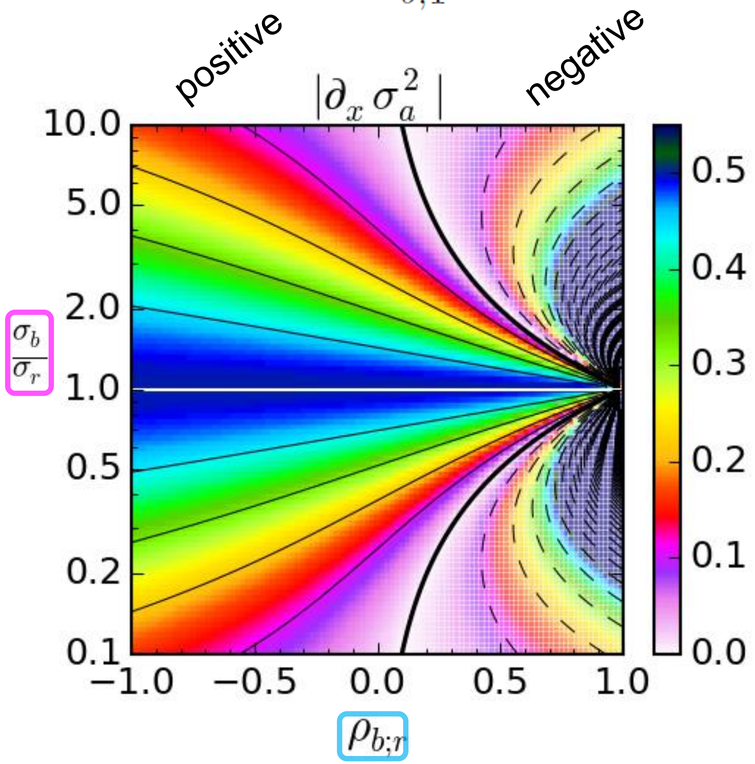
background variance · observation variance · cross-variance

$$\frac{\partial \sigma_a^2}{\partial \rho} = 2\, \sigma_b \sigma_1 \left(\sigma_b^2 - x_{b;1}\right)\left(\sigma_1^2 - x_{b;1}\right)\left(\sigma_b^2 - 2\, x_{b;1} + \sigma_1^2\right)^{-2}$$

- Sensitivity of KF analysis variance to error cross-variance:
  - define ratio of standard-deviations $\alpha := \frac{\sigma_b}{\sigma_1}$

$$\frac{\partial \sigma_a^2}{\partial x_{b;1}} = 2\,(\alpha - \rho)\left(\alpha^{-1} - \rho\right)\left(\alpha - 2\,\rho + \alpha^{-1}\right)^{-2}$$

- Only a function of error correlation and ratio
- Symmetric around max. at $\alpha = 1$
- Decreasing with differences between BG and OBS
- Almost linear for small differences ($\alpha \in \left[\frac{1}{3}; 3\right]$)
- Negative sensitivities for high correlations
  = when cross-variance > smaller variance
- Rapidly decreasing negative sensitivity
  for high correlation and moderate $\alpha \in \left[\frac{1}{6}; 6\right]$



# ASSIMILATION FROM THREE DATASETS

- 3 datasets

- Given:
  - 3 innovation covariances

+3

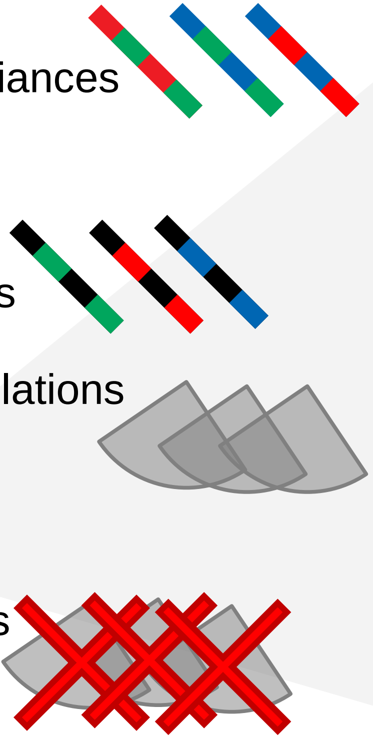
- Unknown:
  - 3 error covariances
  - 3 error cross-correlations

-6

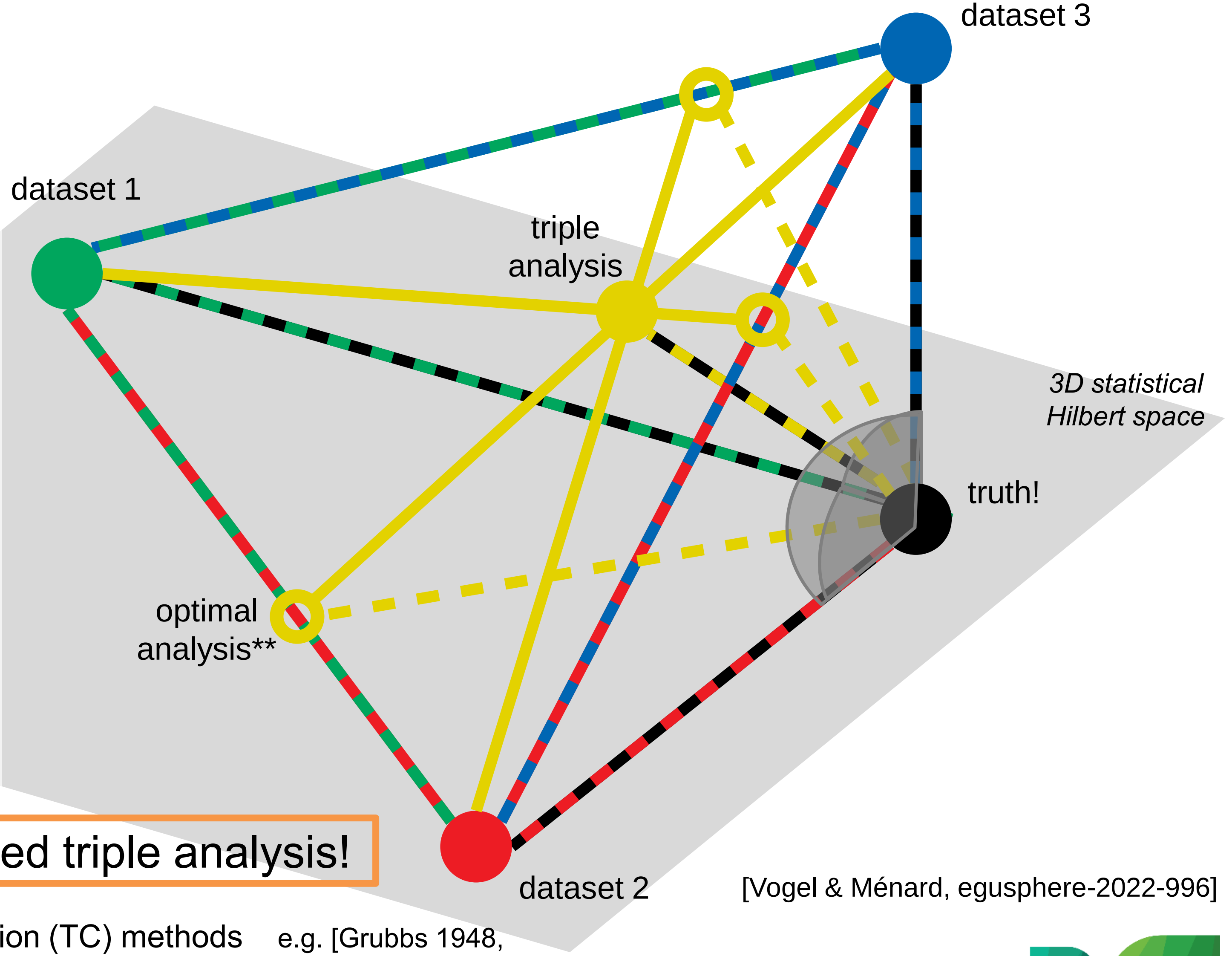
- Assumptions:
  - uncorrelated errors

+3

0



Fully determined!*




Improved triple analysis!

[Vogel & Ménard, egusphere-2022-996]

* equivalent to scalar 3 cornered hat (3CH) and triple collocation (TC) methods e.g. [Grubbs 1948, doi:10.1080/01621459.1948.10483261]; [Stoffelen 1998, doi:10.1029/97JC03180]; [Sjoberg et al. 2021, doi:10.1175/jtech-d-19-0217.1]

** equivalent to cross-validation approach [Ménard&Deshaies-Jacques 2018a&b, doi:10.3390/atmos9030086 & atmos9020070]

# GENERALIZED ANALYSIS FROM MULTIPLE DATASETS – KF EQUATIONS

What if we have more than two datasets?

- Direct KF analysis equations (for 3 datasets)*:

$$x_a = \underbrace{(\mathbb{1} - \mathbf{W}_1\mathbf{H}_1 - \mathbf{W}_2\mathbf{H}_2)}_{:=\mathbf{W}_b} x_b + \mathbf{W}_1 y_1 + \mathbf{W}_2 y_2$$

$$\mathbf{A} = \mathbf{W}_b\mathbf{B}\mathbf{W}_b^T + \mathbf{W}_1\mathbf{R}_1\mathbf{W}_1^T + \mathbf{W}_2\mathbf{R}_2\mathbf{W}_2^T$$

**W** weights

define Kalman gain-like matrices:

$$\boxed{\begin{aligned}\mathbf{K}_{d_1} &:= \mathbf{B}\mathbf{H}_1^T\left(\mathbf{H}_1\mathbf{B}\mathbf{H}_1^T + \mathbf{R}_1\right)^{-1} = \mathbf{B}\mathbf{H}_1^T\mathbf{\Gamma}_1^{-1}\\ \mathbf{K}_{d_2} &:= \mathbf{B}\mathbf{H}_2^T\left(\mathbf{H}_2\mathbf{B}\mathbf{H}_2^T + \mathbf{R}_2\right)^{-1} = \mathbf{B}\mathbf{H}_2^T\mathbf{\Gamma}_2^{-1}\end{aligned}}$$

e.g.: $\mathbf{W}_1 = (\mathbb{1} - \underline{\mathbf{W}_2\mathbf{H}_2})\,\mathbf{B}\mathbf{H}_1^T\left(\mathbf{H}_1\mathbf{B}\mathbf{H}_1^T + \mathbf{R}_1\right)^{-1}$

$$\mathbf{W}_1 = (\mathbb{1} - \mathbf{K}_{d_2}\mathbf{H}_2)\,\mathbf{K}_{d_1}\left(\mathbb{1} - \mathbf{H}_1\mathbf{K}_{d_2}\mathbf{H}_2\mathbf{K}_{d_1}\right)^{-1}$$

- Sequential KF analysis equations (for 3 datasets)*:

$$x_{a_2} = \underbrace{(\mathbb{1} - \mathbf{K}_{a_2}\mathbf{H}_2)(\mathbb{1} - \mathbf{K}_{a_1}\mathbf{H}_1)}_{:=\mathbf{W}_b} x_b + \underbrace{(\mathbb{1} - \mathbf{K}_{a_2}\mathbf{H}_2)\,\mathbf{K}_{a_1}}_{:=\mathbf{W}_1} y_1 + \underbrace{\mathbf{K}_{a_2}}_{:=\mathbf{W}_2} y_2$$

$$\mathbf{W}_b = \mathbf{A}_2\mathbf{B}^{-1}$$

$$\mathbf{A}_2^{-1} = \mathbf{B}^{-1} + \mathbf{H}_1^T\mathbf{R}_1^{-1}\mathbf{H}_1 + \mathbf{H}_2^T\mathbf{R}_2^{-1}\mathbf{H}_2$$

$$\mathbf{W}_1 = \mathbf{A}_2\mathbf{H}_1^T\mathbf{R}_1^{-1} \qquad \mathbf{W}_2 = \mathbf{A}_2\mathbf{H}_2^T\mathbf{R}_2^{-1}$$

– summation-like generalization of analysis state and error covariance
– weight(/gain) reduced by weight of additional dataset, equivalent analysis-based form
– equivalence of direct and sequential form

* equivalent to multi-model KF e.g. [Logutov&Robinson 2005, doi:10.1256/qj.05.99]; [Narayan et al. 2012, doi:10.1016/j.jcp.2012.06.002]

# GENERALIZED ANALYSIS FROM MULTIPLE DATASETS – VAR EQUATIONS

What if we have more than two datasets?

- Direct 3D-var equations (for 3 datasets)*:

$$J_2(x) = \frac{1}{2}\left[\left(x - x_b\right)^T \mathbf{B}^{-1}\left(x - x_b\right) + \left(\mathbf{H}_1 x - y_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{H}_1 x - y_1\right) \underline{+ \left(\mathbf{H}_2 x - y_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{H}_2 x - y_2\right)}\right]$$

$$\nabla_x J_2(x) = \mathbf{B}^{-1}\left(x - x_b\right) + \mathbf{H}_1^T \mathbf{R}_1^{-1}\left(\mathbf{H}_1 x - y_1\right) \underline{+ \mathbf{H}_2^T \mathbf{R}_2^{-1}\left(\mathbf{H}_2 x - y_2\right)}$$

- Sequential 3D-var equations (for 3 datasets) assuming vanishing gradient in 1st assimilation:

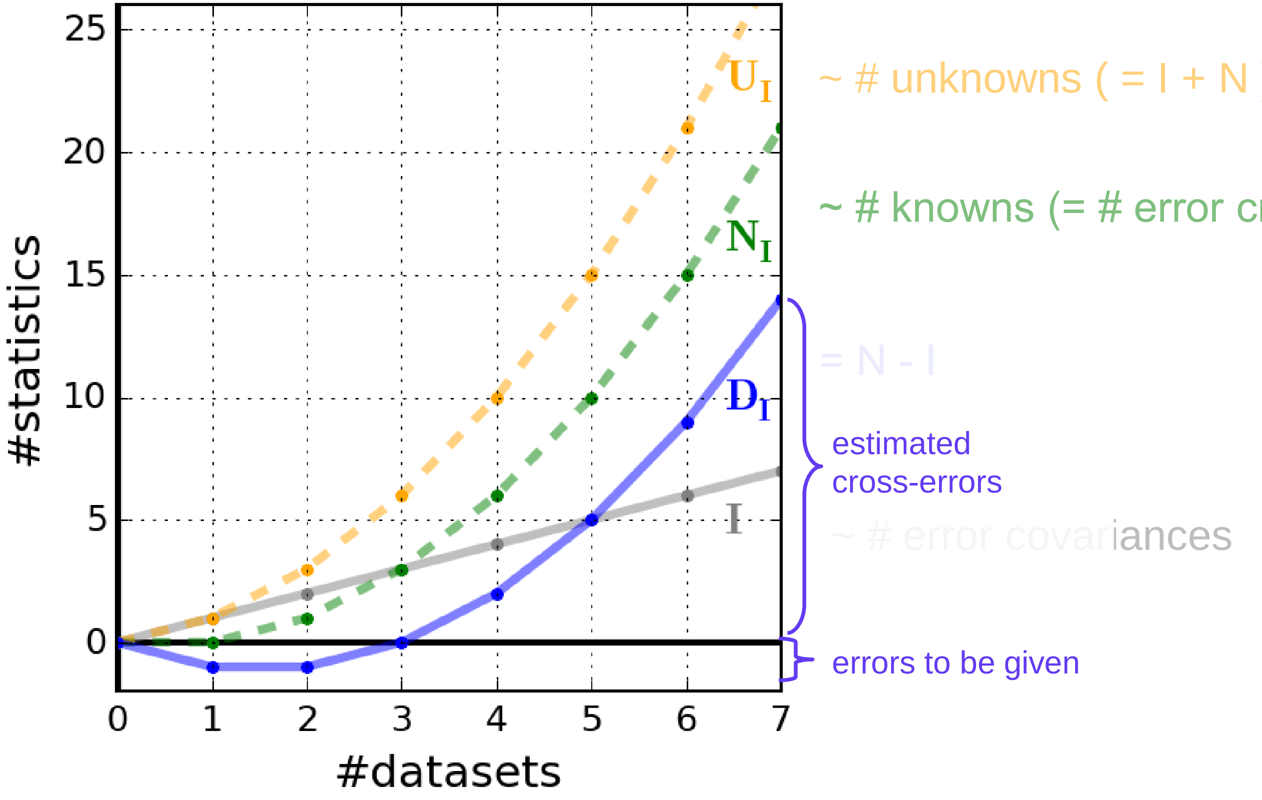
$$J_2(x) = \frac{1}{2}\Bigg[\underbrace{\left(x - x_b\right)^T \mathbf{B}^{-1}\left(x - x_b\right) + \left(\mathbf{H}_1 x - y_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{H}_1 x - y_1\right) \underline{+ \left(\mathbf{H}_2 x - y_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{H}_2 x - y_2\right)}}_{\rightarrow \text{dependent on optimization variable } x}$$
$$+ \underbrace{\boxed{\left(\mathbf{B}^{-1}x_b + \mathbf{H}_1^T \mathbf{R}_1^{-1} y_1\right)^T \left(\mathbf{B}^{-1} + \mathbf{H}_1^T \mathbf{R}_1^{-1}\mathbf{H}_1\right)^{-1}\left(\mathbf{B}^{-1}x_b + \mathbf{H}_1^T \mathbf{R}_1^{-1} y_1\right) - \left(x_b^T \mathbf{B}^{-1} x_b + y_1^T \mathbf{R}_1^{-1} y_1\right)}}_{\rightarrow \text{ independent of } x}\Bigg]$$

$$\nabla_x J_2(x) = \mathbf{B}^{-1}\left(x - x_b\right) + \mathbf{H}_1^T \mathbf{R}_1^{-1}\left(\mathbf{H}_1 x - y_1\right) \underline{+ \mathbf{H}_2^T \mathbf{R}_2^{-1}\left(\mathbf{H}_2 x - y_2\right)}$$

– summation-like generalization of cost function and gradient
– constant additional term in sequential cost function → equivalence

* equivalent to multi-model KF e.g. [Logutov&Robinson 2005, doi:10.1256/qj.05.99]; [Narayan et al. 2012, doi:10.1016/j.jcp.2012.06.002]

# GENERALIZED ERROR ESTIMATION PROBLEM




$I < 3$ underdetermined even under uncorrelated assumption

$I = 3$ fully determined under uncorrelated assumption

$I > 3$ overdetermined under uncorrelated assumption

$\rightarrow$ estimation of $\boxed{D_I}$ error correlations

<u>Consequences:</u>

- Relative number of assumed statistics reduces with increasing number of datasets.
- Absolute number of assumptions increases with number of datasets (system is never closed).

Estimation of all error covariances and <u>some</u> cross-correlations

[Vogel & Ménard, egusphere-2022-996]

# GENERALIZED ANALYSIS WITH CORRELATED ERRORS – KF EQUATIONS

What if we have more than two correlated datasets?

- Direct KF analysis equations with correlation (for **3** datasets):

substituted error covariances:

$$\tilde{\mathbf{B}}_1 := \mathbf{B} - \mathbf{X}_{b;1}\mathbf{H}_1^{+T} \qquad \tilde{\mathbf{B}}_2 := \mathbf{B} - \mathbf{X}_{b;2}\mathbf{H}_2^{+T} \qquad \tilde{\mathbf{\Gamma}}_{1;2} := \mathbf{H}_1\mathbf{B}\mathbf{H}_2^T - \mathbf{H}_1\mathbf{X}_{b;2} - \mathbf{X}_{b;1}^T\mathbf{H}_2^T + \mathbf{X}_{1;2}$$

$$\tilde{\mathbf{\Gamma}}_1 := \mathbf{H}_1\mathbf{B}\mathbf{H}_1^T - \mathbf{H}_1\mathbf{X}_{b;1} - \mathbf{X}_{b;1}^T\mathbf{H}_1^T + \mathbf{R}_1 \qquad \tilde{\mathbf{\Gamma}}_2 := \mathbf{H}_2\mathbf{B}\mathbf{H}_2^T - \mathbf{H}_2\mathbf{X}_{b;2} - \mathbf{X}_{b;2}^T\mathbf{H}_2^T + \mathbf{R}_2$$

$$\mathbf{W}_1 = \left(\tilde{\mathbf{B}}_1\mathbf{H}_1^T - \underline{\tilde{\mathbf{B}}_2\mathbf{H}_2^T\tilde{\mathbf{\Gamma}}_2^{-1}\tilde{\mathbf{\Gamma}}_{1;2}^T}\right)\left(\tilde{\mathbf{\Gamma}}_1 - \underline{\tilde{\mathbf{\Gamma}}_{1;2}\tilde{\mathbf{\Gamma}}_2^{-1}\tilde{\mathbf{\Gamma}}_{1;2}^T}\right)^{-1}$$

compare: 2 datasets $\mathbf{K}_1 = \tilde{\mathbf{B}}\mathbf{H}_1^T\tilde{\mathbf{\Gamma}}_1^{-1}$

- Direct KF analysis equations with correlation (for **I** datasets):

$$\mathbf{W}_1 = \left[\mathbf{B}\mathbf{H}_1^T - \mathbf{X}_{b;1} - \sum_{j=2}^{I-1}\underline{\mathbf{W}_j}\left(\mathbf{H}_j\mathbf{B}\mathbf{H}_1^T - \mathbf{X}_{j;b}\mathbf{H}_1^T - \mathbf{H}_j\mathbf{X}_{b;1} + \mathbf{X}_{j;1}\right)\right]\left(\mathbf{H}_1\mathbf{B}\mathbf{H}_1^T - \mathbf{H}_1\mathbf{X}_{b;1} - \mathbf{X}_{1;b}\mathbf{H}_1^T + \mathbf{C}_1\right)^{-1}$$

...???

- solution consistent with uncorrelated form and correlated form for 2 datasets
- significant increase in complexity
- use sequential form for assimilating multiple correlated datasets

# GENERALIZED ANALYSIS WITH CORRELATED ERRORS - SENSITIVITIES

- Sensitivity of generalized analysis eq. :
  - scalar (**H** → 1)
  - background and N=I-1 observations with common variance
  - no cross-correlation among obs., common cross-correlation to background
  - define common ratio of standard-deviations

$$x_{b;r} = \rho\ \sigma_b\ \sigma_r \quad \forall\ r$$

$$\alpha := \frac{\sigma_b}{\sigma_r}$$

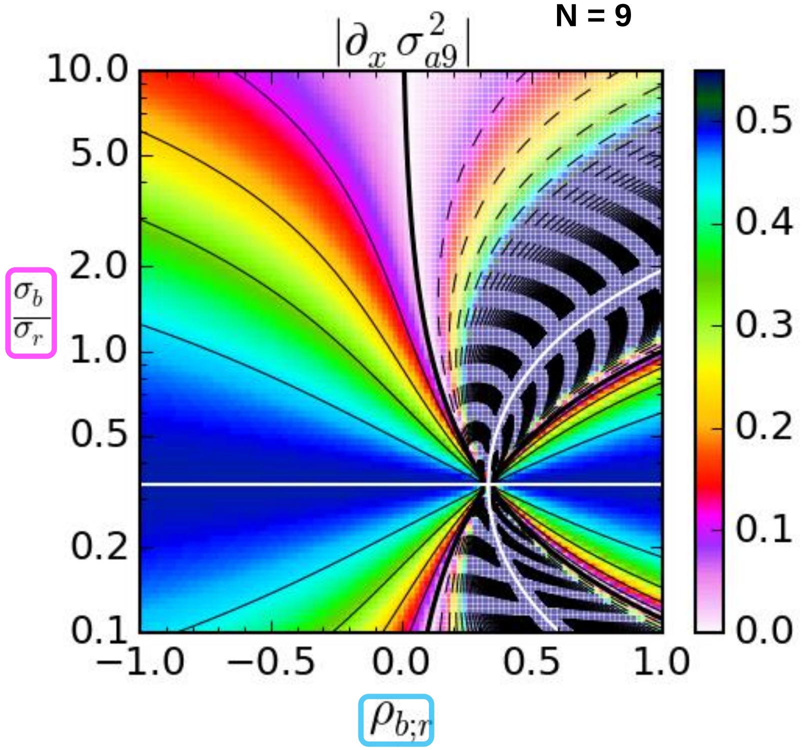
  - analysis error variance and cross-variance from obs.-only

$$\sigma_{ar}^2 = \left[\sum_{n=1}^{N}\left(\sigma_n^2\right)^{-1}\right]^{-1} = \frac{1}{N}\sigma_r^2 \quad , \quad x_{b;ar} = x_{b;1} + \sum_{n=2}^{N} K_n \underbrace{\cancel{\left(x_{b;n} - x_{b;1}\right)}}_{0} = x_{b;r}$$

scalar cross-variance

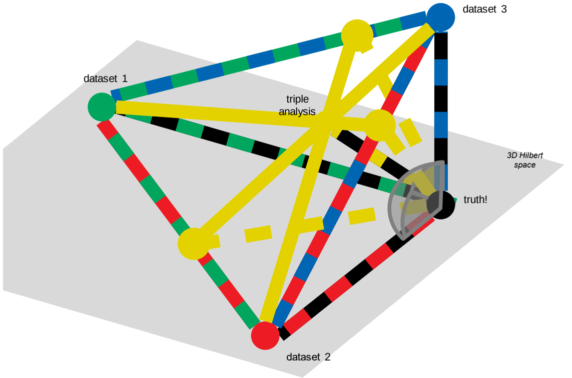
  - generalized sensitivity of final analysis error variance

$$\frac{\partial \sigma_a^2}{\partial x_{b;r}} = 2\left(\boxed{\alpha} - \boxed{\rho}\right)\left(\frac{1}{N\boxed{\alpha}} - \boxed{\rho}\right)\left(\boxed{\alpha} - 2\boxed{\rho} + \frac{1}{N\boxed{\alpha}}\right)^{-2}$$




- symmetry line decreases (were $\sigma_{ar}^2 = \sigma_b^2$)
- common line of vanishing sensitivity (were $\alpha = \rho$)
- significant negative sensitivities already for decreasing correlations

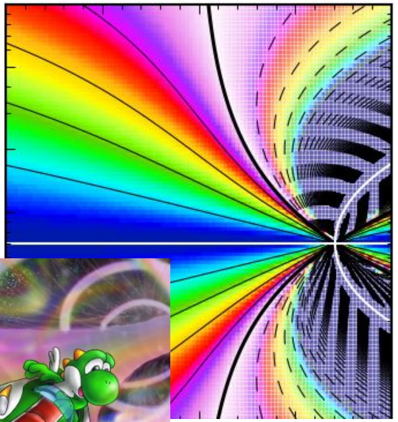
# SUMMARY / CONCLUSIONS




- Error estimation:
  - Multiple collocated datasets enable estimation of optimal error statistics, incl. cross-correlations
  - "Datasets" may include multiple forecasts, observations, …
  - Number of estimated error statistics increases with number of datasets
  - Some assumptions and conditions remain

Manuscript: Vogel & Ménard, "***How far can the statistical error estimation problem be closed by collocated data?***"
under review @ NPG: *egusphere-2022-996*

- Assimilation:
  - Improved analysis of multiple datasets
  - Direct assimilation of uncorrelated datasets equivalent to sequential form
  - Assimilation of correlated datasets becomes expensive
  - Critical sensitivities for high correlations and large differences in errors



https://www.pinterest.ca/pin/469711436108621063/

*"We cannot have too many datasets! (If we have appropriate assimilation algorithms…)"*

Contact: annika.vogel@ec.gc.ca