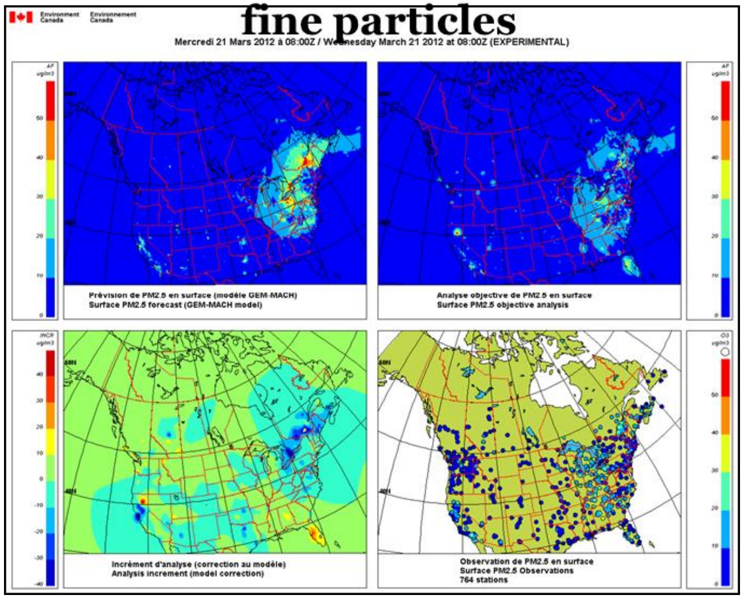
Analysis error statistics estimation
and a simple off-line ensemble AQ analysis

Richard Ménard, Martin Deshaies-Jacques,
Jean-François Cossette, James Abu

Air Quality Research, Environment and Climate Change Canada (ECCC)

Applications to multiyear surface analysis of air quality using surface observations



History

- O_3, $PM_{2.5}$ – using CHRONOS 2002-2009
- O_3, $PM_{2.5}$ – using GEM-MACH 2009-2015
- O_3, $PM_{2.5}$, NO_2, SO_2, PM_{10} since April 2015
- Multi-year data set (2002-2012) Robichaud and Ménard 2014, *ACP*)
- Operational surface analysis since 2013

- First part of this talk:
  Optimize the analysis, by improving the estimation of observation and background error covariances
- Second part :
  Minimal computation to rerun multiyear analyses
  off-line ensembles to provide error correlations

How do we know that our error covariances are the true error covariances ?

We need observations
We derive here conditions in observation space only

Assuming observation and background errors are **uncorrelated**
The error covariances are the true error covariances

$$i)\quad \tilde{\mathbf{R}} = \mathbf{R} \qquad\qquad ii)\quad \mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T = \mathbf{HBH}^T$$

Is equivalent to the respecting the following conditions

$$A)\quad \mathbf{H}\tilde{\mathbf{K}} = \mathbf{HK} \quad \textit{The gain in observation space is the optimal Kalman gain}$$

$$B)\quad \mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T + \tilde{\mathbf{R}} = \mathbf{HBH}^T + \mathbf{R} = \mathbb{E}[(O-B)(O-B)^T]$$

Innovation covariance matching / consistency

❑ Menard, R: Error covariance estimation methods based on analysis residuals: theoretical foundation and convergence properties derived from simplified observation networks. *Q. J. Roy. Meteorol. Soc.* **2016**, *142*, 257—273, doi:10.1002/qj2650

$$B)\quad \tilde{\mathbf{D}} = \mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T + \tilde{\mathbf{R}}$$

$$= \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R} = \mathbf{D} = \mathbb{E}[(O-B)(O-B)^T]$$

Innovation covariance matching / consistency

Global matching but indirect

$$\mathbb{E}[\chi^2] = \mathbb{E}[\mathbf{d}^T\,\tilde{\mathbf{D}}^{-1}\,\mathbf{d}] = \mathbb{E}[tr(\tilde{\mathbf{D}}^{-1}\mathbf{d}^T\mathbf{d})] = tr(\tilde{\mathbf{D}}^{-1}\mathbf{D}) = p$$

p is the number of observations

or

$$\mathbb{E}[J_{\min}] = \frac{p}{2}$$

Or ***variance matching***

$$diag(\tilde{\mathbf{D}}) = diag(\mathbf{D}) = \mathrm{var}(O-B)$$

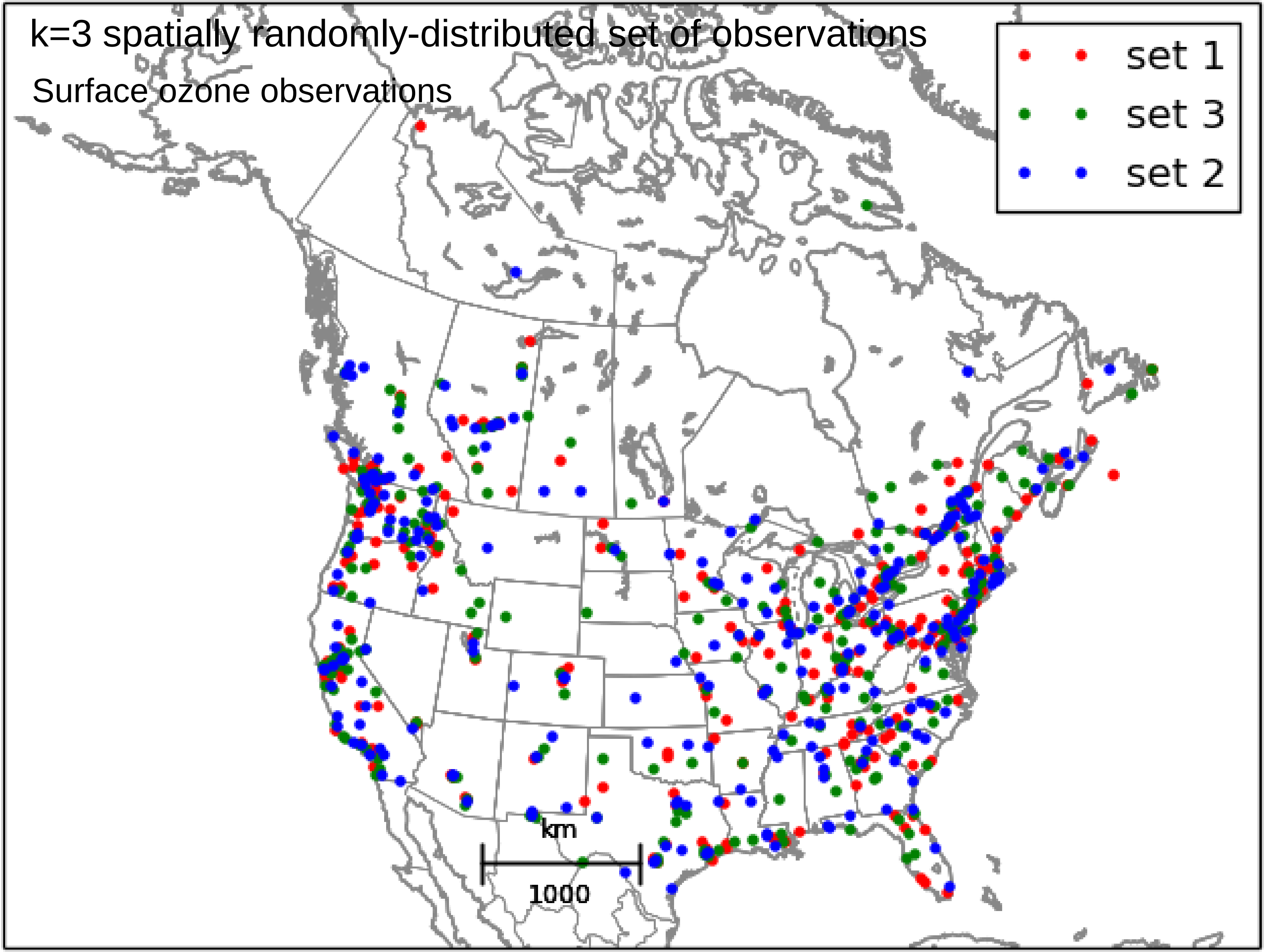
$A)$ $\mathbf{H\tilde{K}} = \mathbf{HK}$ *The Kalman gain condition*

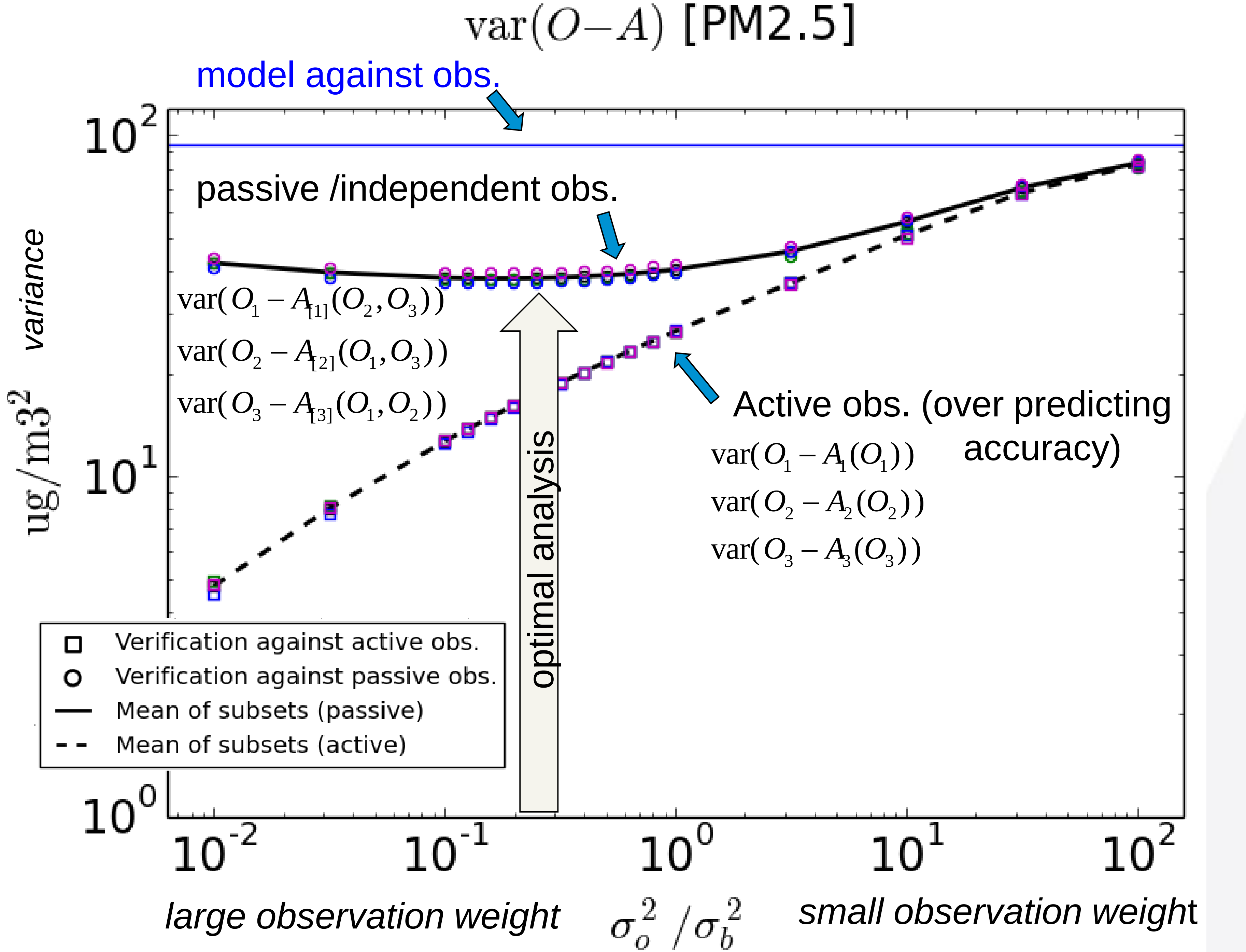
True analysis error in observation space is minimized

Cross-validation can give a way to evaluate analysis error variance (without using a forecast)

$$\mathrm{var}(O - A)_c = tr(\mathbf{R} + \mathbf{H}_c\mathbf{A}\mathbf{H}_c^T)$$

From an orderly set of station ID number, select each k^th station





$\mathrm{var}(O-A)$ [PM2.5]
model against obs.
passive /independent obs.
$\mathrm{var}(O_1 - A_{[1]}(O_2, O_3))$
$\mathrm{var}(O_2 - A_{[2]}(O_1, O_3))$
$\mathrm{var}(O_3 - A_{[3]}(O_1, O_2))$
optimal analysis
Active obs. (over predicting accuracy)
$\mathrm{var}(O_1 - A_1(O_1))$
$\mathrm{var}(O_2 - A_2(O_2))$
$\mathrm{var}(O_3 - A_3(O_3))$
variance
$\mathrm{ug/m3}^2$
Verification against active obs.
Verification against passive obs.
Mean of subsets (passive)
Mean of subsets (active)
large observation weight
σ_o^2 / σ_b^2
small observation weight

A geometric view of the analysis

Hilbert spaces of random variables

Define an inner product of two (zero-mean) random variables X , Y as

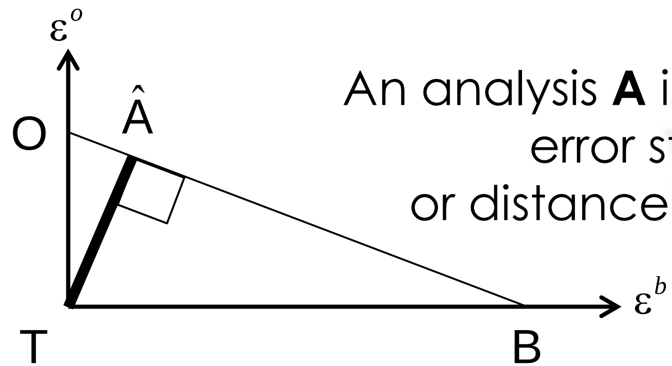
$$\langle X, Y \rangle = \mathbb{E}[X\,Y]$$

Uncorrelated random variables X , Y represented as orthogonal vectors

$$\langle X, Y \rangle = 0$$

Standard deviation is represented as the norm $\sqrt{\|X\|_2} = \sqrt{\mathbb{E}[X^2]}$

$$\mathbb{E}[\boldsymbol{\varepsilon}^o(\boldsymbol{\varepsilon}^b)^T] = \mathbf{0}$$



An analysis **A** is optimal when its error standard deviation or distance between **A** and **T** is minimized

Desrozier's et al (2005)

- Assumes the analysis is optimal, or that we know the background error covariance
- Assumes innovation covariance consistency
- Derive the observation error covariance

Cross-validation approach

- Assumes innovation covariance consistency
- Estimate covariance parameters to optimize the analysis error covariance
- Observations error covariance, background error covariance, and analysis error covariance

A geometric view of cross-validation

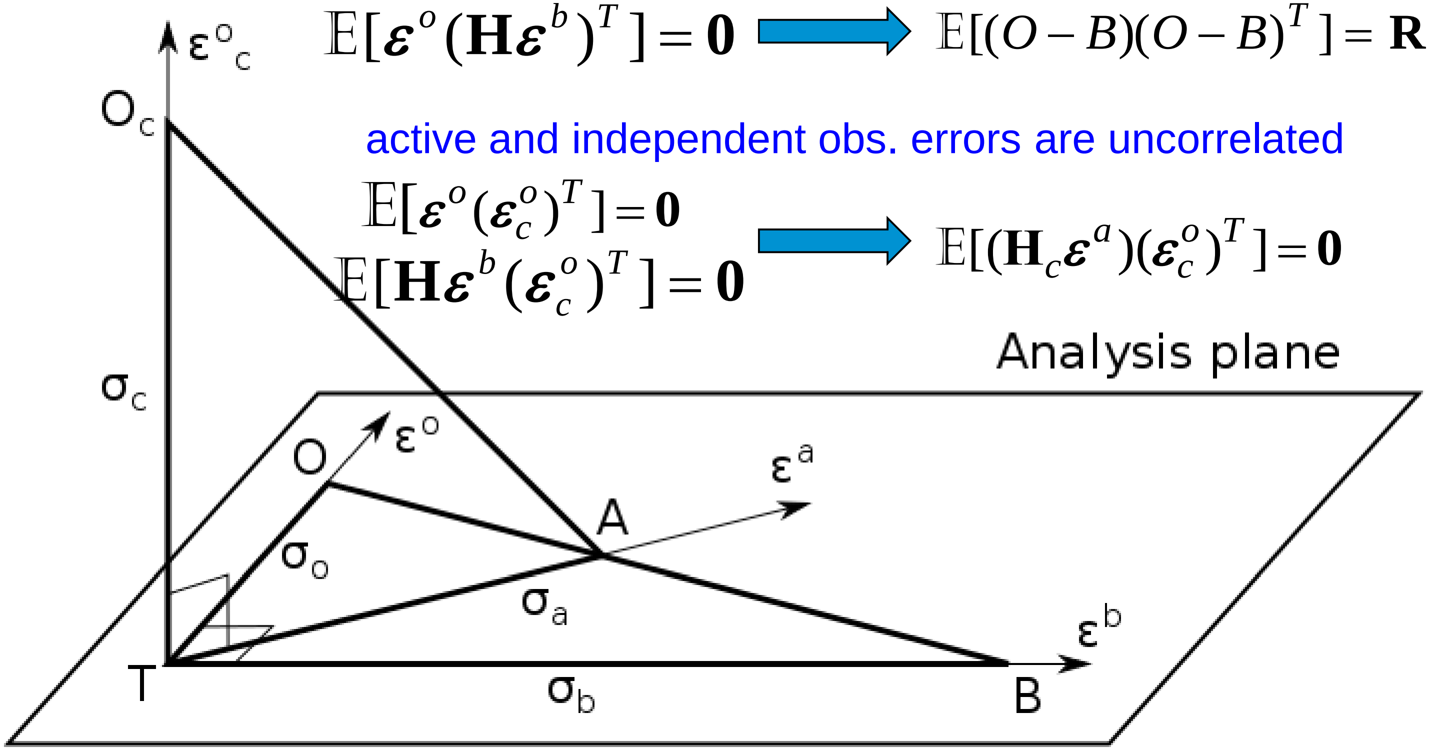
obs and background errors are uncorrelated

$$\mathbb{E}[\boldsymbol{\varepsilon}^o(\mathbf{H}\boldsymbol{\varepsilon}^b)^T] = \mathbf{0} \Longrightarrow \mathbb{E}[(O-B)(O-B)^T] = \mathbf{R} + \mathbf{HBH}^T$$

active and independent obs. errors are uncorrelated

$$\mathbb{E}[\boldsymbol{\varepsilon}^o(\boldsymbol{\varepsilon}_c^o)^T] = \mathbf{0}$$

$$\mathbb{E}[\mathbf{H}\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}_c^o)^T] = \mathbf{0} \Longrightarrow \mathbb{E}[(\mathbf{H}_c\boldsymbol{\varepsilon}^a)(\boldsymbol{\varepsilon}_c^o)^T] = \mathbf{0}$$

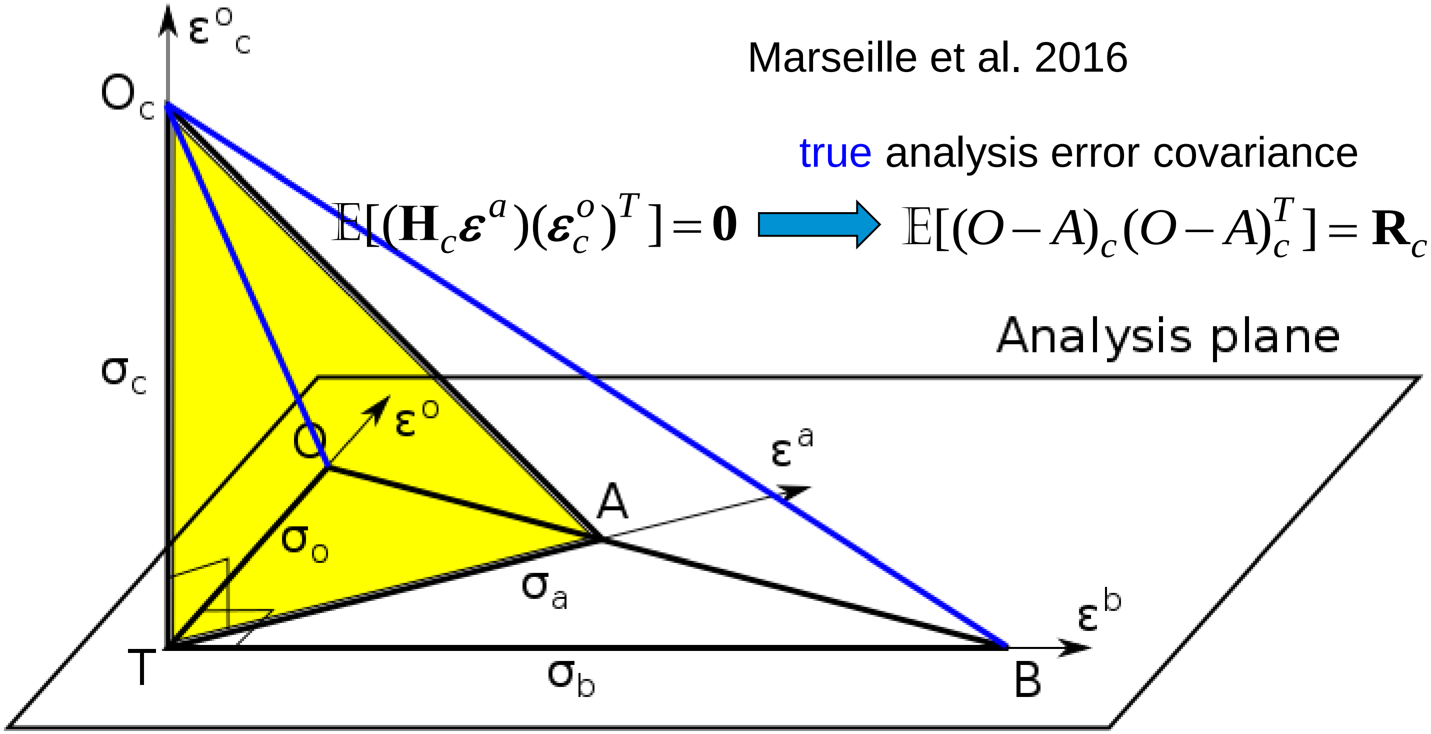


A geometric view of cross-validation

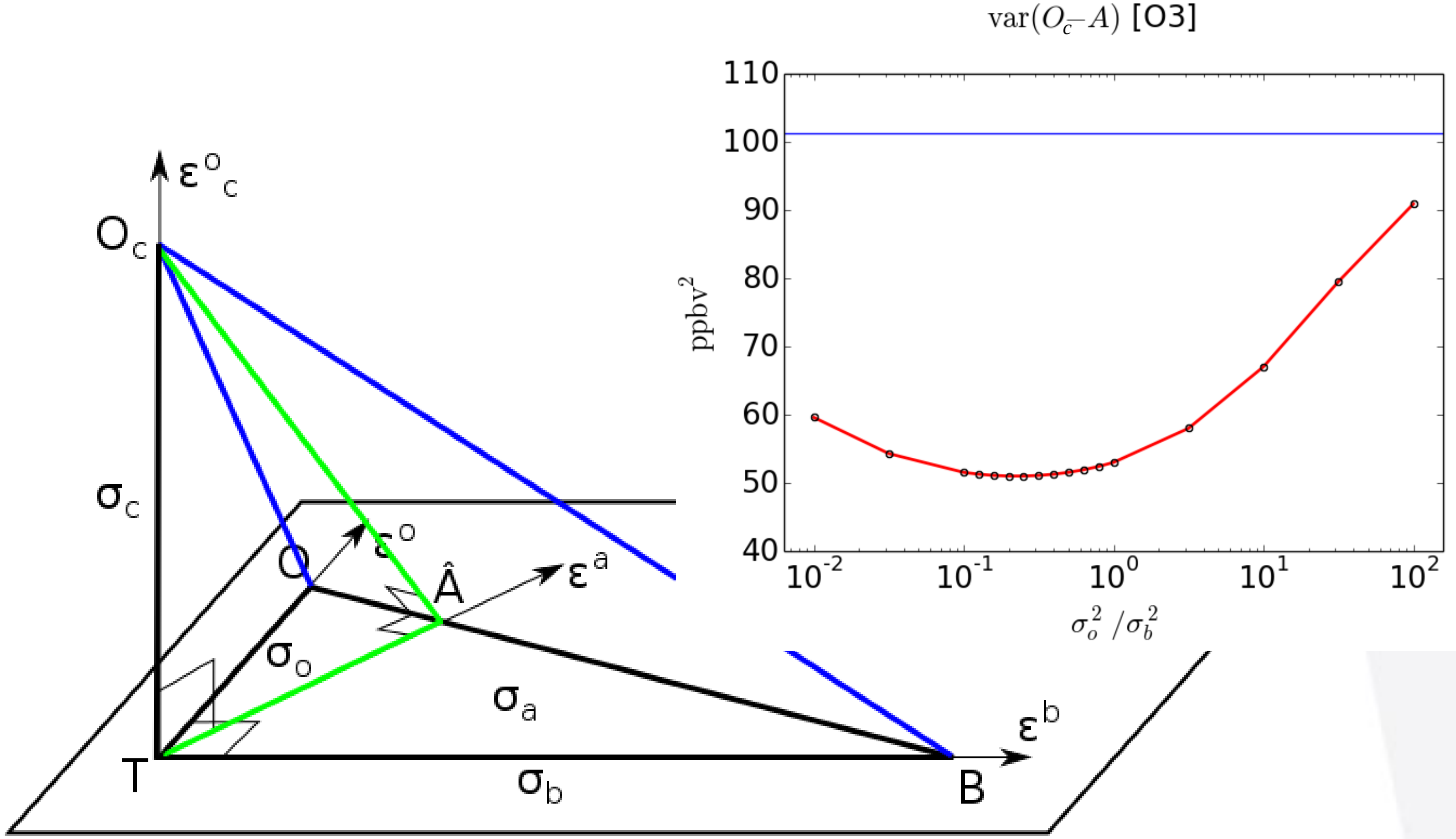
Marseille et al. 2016

true analysis error covariance

$$\mathbb{E}[(\mathbf{H}_c\boldsymbol{\varepsilon}^a)(\boldsymbol{\varepsilon}_c^o)^T] = \mathbf{0} \Rightarrow \mathbb{E}[(O-A)_c(O-A)_c^T] = \mathbf{R}_c + \mathbf{H}_c\mathbf{A}\mathbf{H}_c^T$$



A geometric view of cross-validation

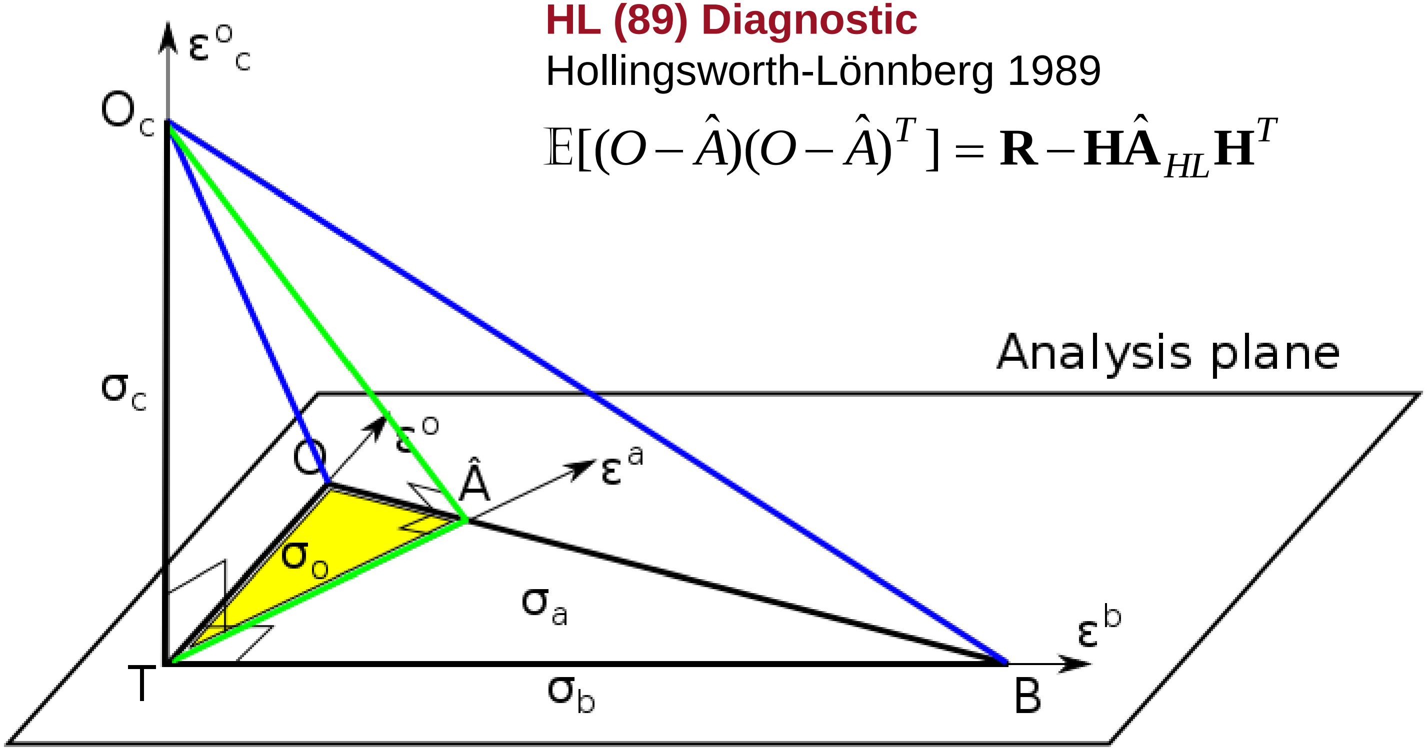


Diagnostics: A geometric view: Active observation space

HL (89) Diagnostic

Hollingsworth-Lönnberg 1989

$$\mathbb{E}[(O-\hat{A})(O-\hat{A})^T] = \mathbf{R} - \mathbf{H}\hat{\mathbf{A}}_{HL}\mathbf{H}^T$$



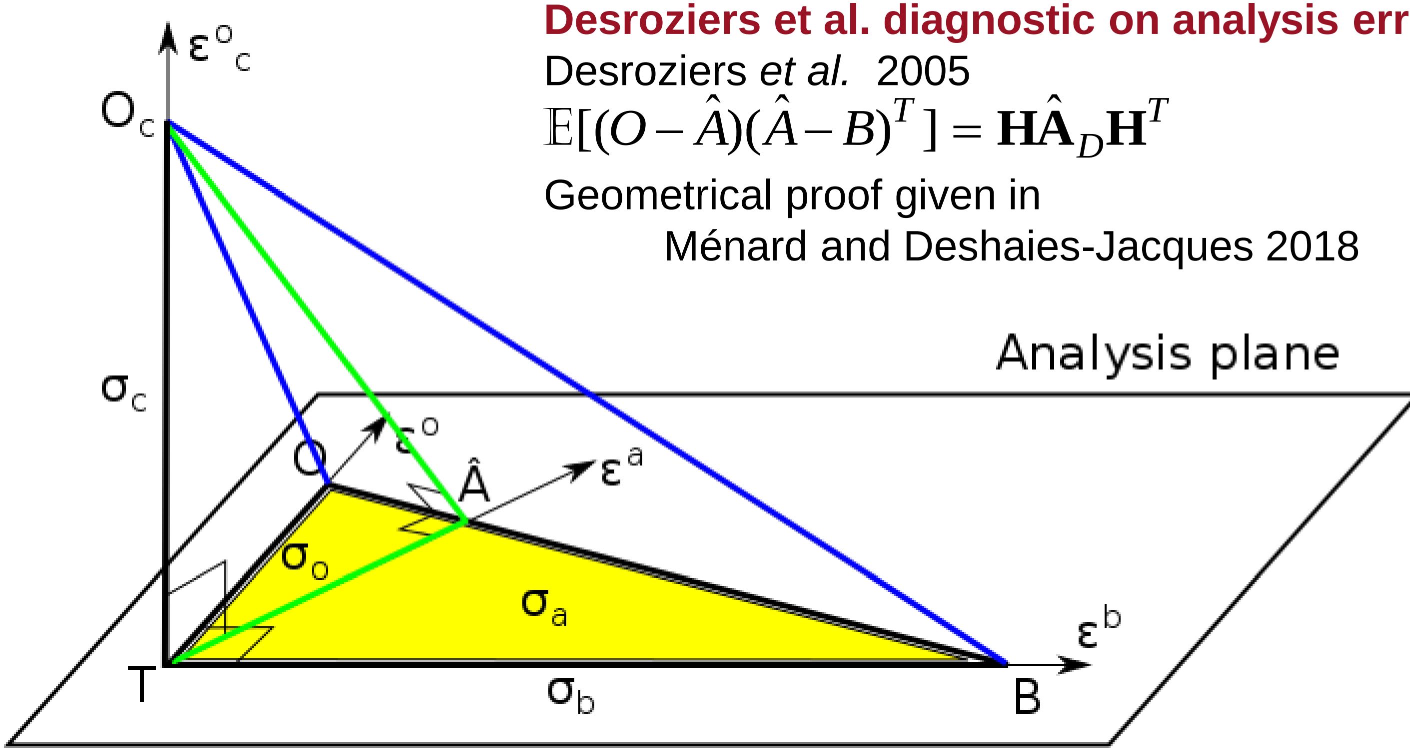
Diagnostics: A geometric view: Active observation space

Desroziers et al. diagnostic on analysis error

Desroziers *et al.* 2005

$$\mathbb{E}[(O-\hat{A})(\hat{A}-B)^T] = \mathbf{H}\hat{\mathbf{A}}_D\mathbf{H}^T$$

Geometrical proof given in
Ménard and Deshaies-Jacques 2018



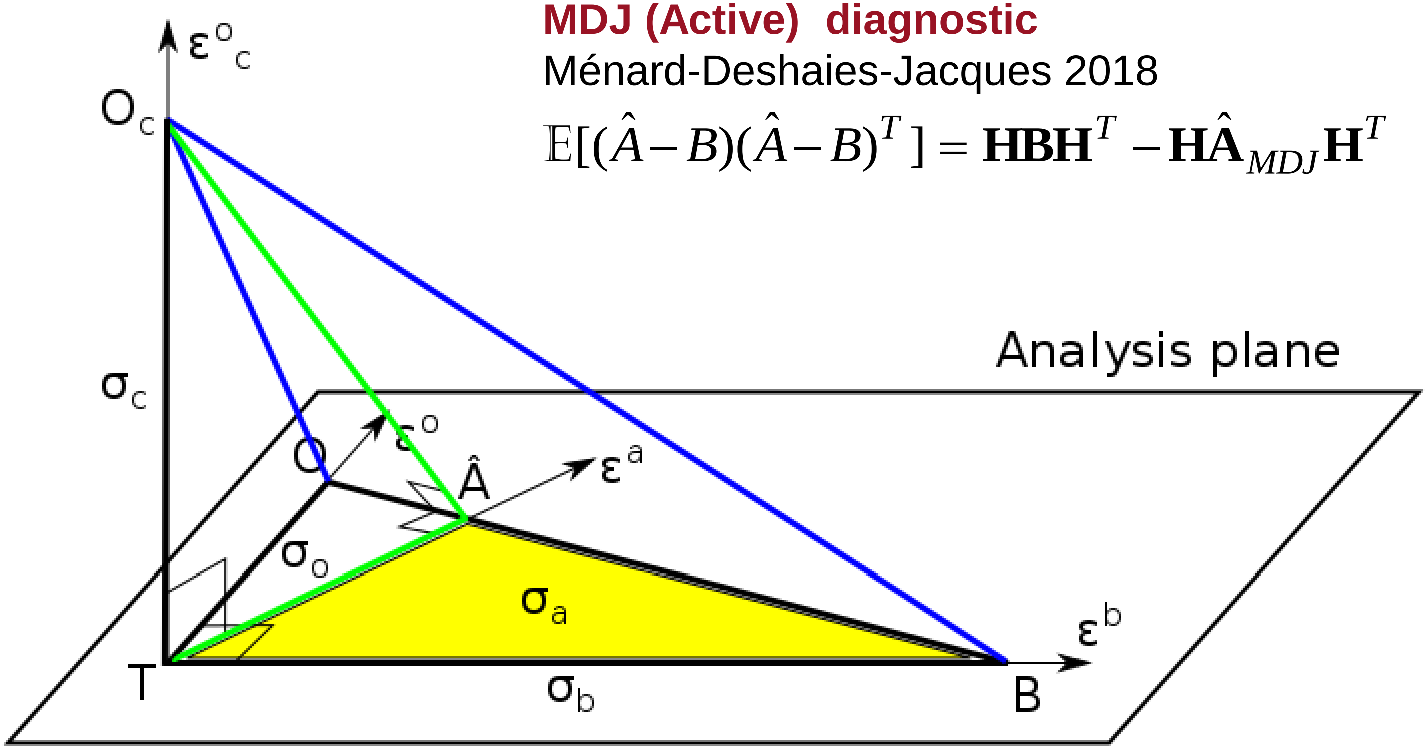
- Ménard, R and M. Deshaies-Jacques. Evaluation of analysis by cross-validation. Part I: Using verification metrics. *Atmosphere* **2018**, 9(3), 86, doi:10.3390/atmos9030086
- Ménard, R and M. Deshaies-Jacques. Evaluation of analysis by cross-validation. Part II: Diagnostic and optimization of analysis error covariance. *Atmosphere* **2018**, 9(2), 70; doi:10.3390/atmos9020070

Diagnostics: A geometric view: Active observation space

MDJ (Active) diagnostic

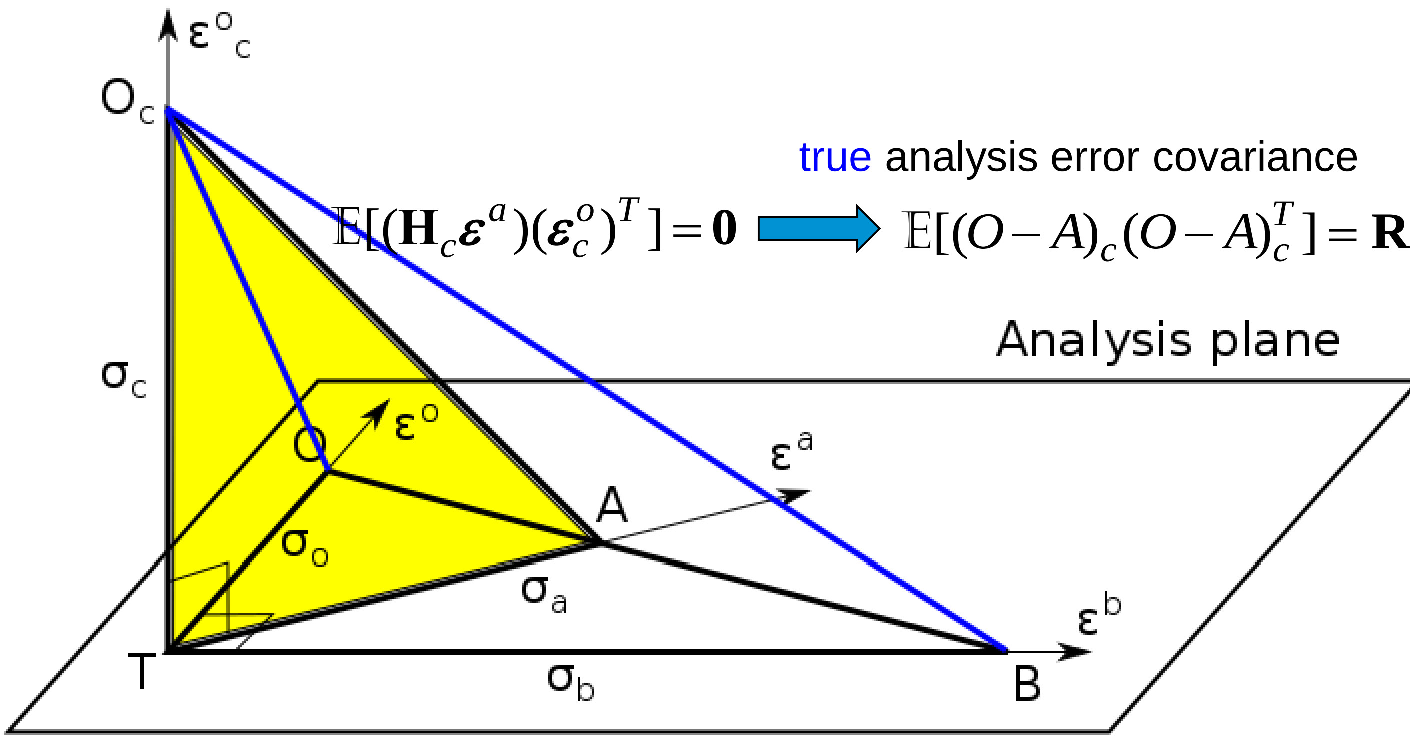
Ménard-Deshaies-Jacques 2018

$$\mathbb{E}[(\hat{A}-B)(\hat{A}-B)^T] = \mathbf{HBH}^T - \mathbf{H\hat{A}}_{MDJ}\mathbf{H}^T$$



Diagnostics: A geometric view: Passive CV space

Marseille et al. 2016 diagnostic



true analysis error covariance

$$\mathbb{E}[(\mathbf{H}_c\boldsymbol{\varepsilon}^a)(\boldsymbol{\varepsilon}_c^o)^T] = \mathbf{0} \Longrightarrow \mathbb{E}[(O-A)_c(O-A)_c^T] = \mathbf{R}_c + \mathbf{H}_c\mathbf{A}\mathbf{H}_c^T$$

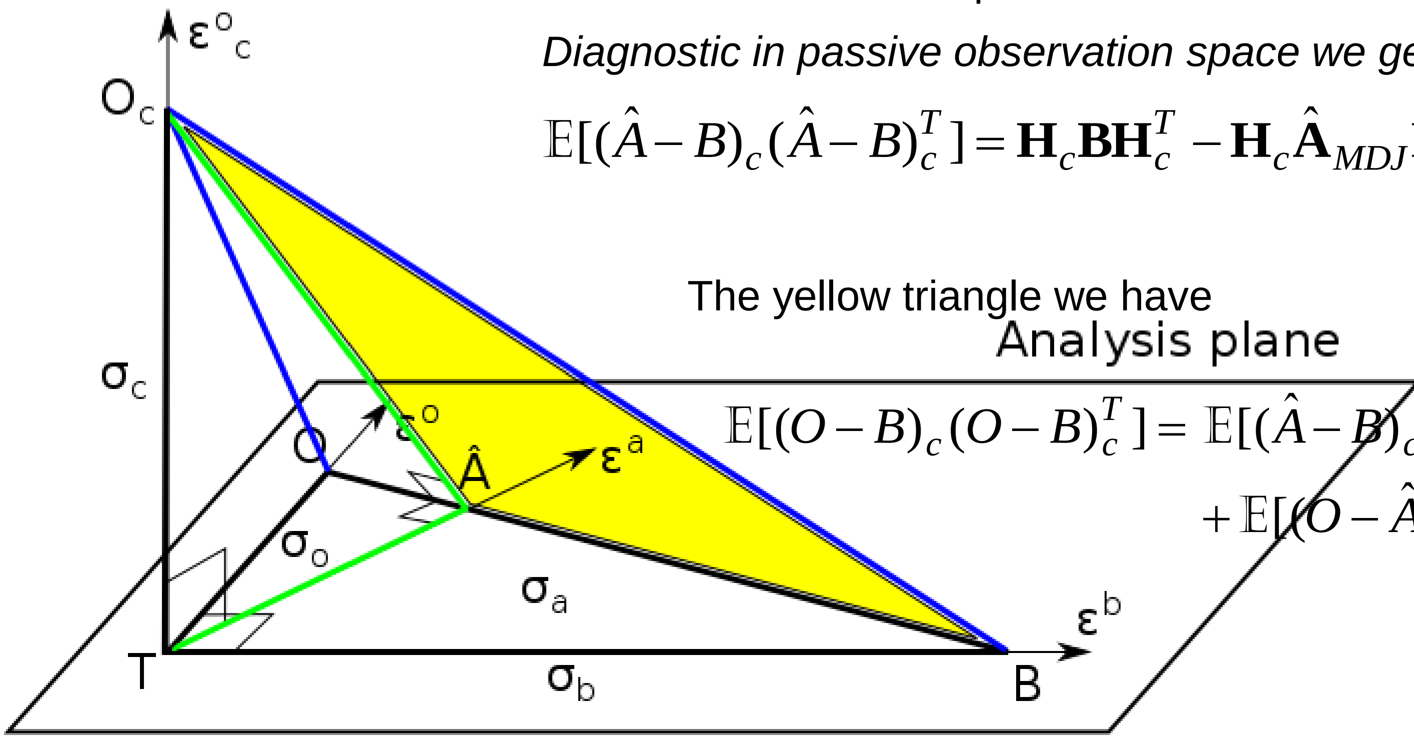
Diagnostics: A geometric view: Passive CV space

MDJ (Passive) diagnostic

Ménard-Deshaies-Jacques 2018

Diagnostic in passive observation space we get

$$\mathbb{E}[(\hat{A}-B)_c(\hat{A}-B)_c^T] = \mathbf{H}_c\mathbf{B}\mathbf{H}_c^T - \mathbf{H}_c\hat{\mathbf{A}}_{MDJ}\mathbf{H}_c^T$$



The yellow triangle we have

$$\mathbb{E}[(O-B)_c(O-B)_c^T] = \mathbb{E}[(\hat{A}-B)_c(\hat{A}-B)_c^T] + \mathbb{E}[(O-\hat{A})_c(O-\hat{A})_c^T]$$

and using $\mathbb{E}[(O-\hat{A})_c(O-\hat{A})_c^T] = \mathbf{H}_c\hat{\mathbf{A}}\mathbf{H}_c^T + \mathbf{R}_c$ we get the relation above

$$\mathbb{E}[(O-B)_c(O-B)_c^T] = \mathbf{H}_c\mathbf{B}\mathbf{H}_c^T + \mathbf{R}_c$$

Adding physical/chemical content in the error statistics

Spatial correlations based on (passive) ensemble of AQ forecasts

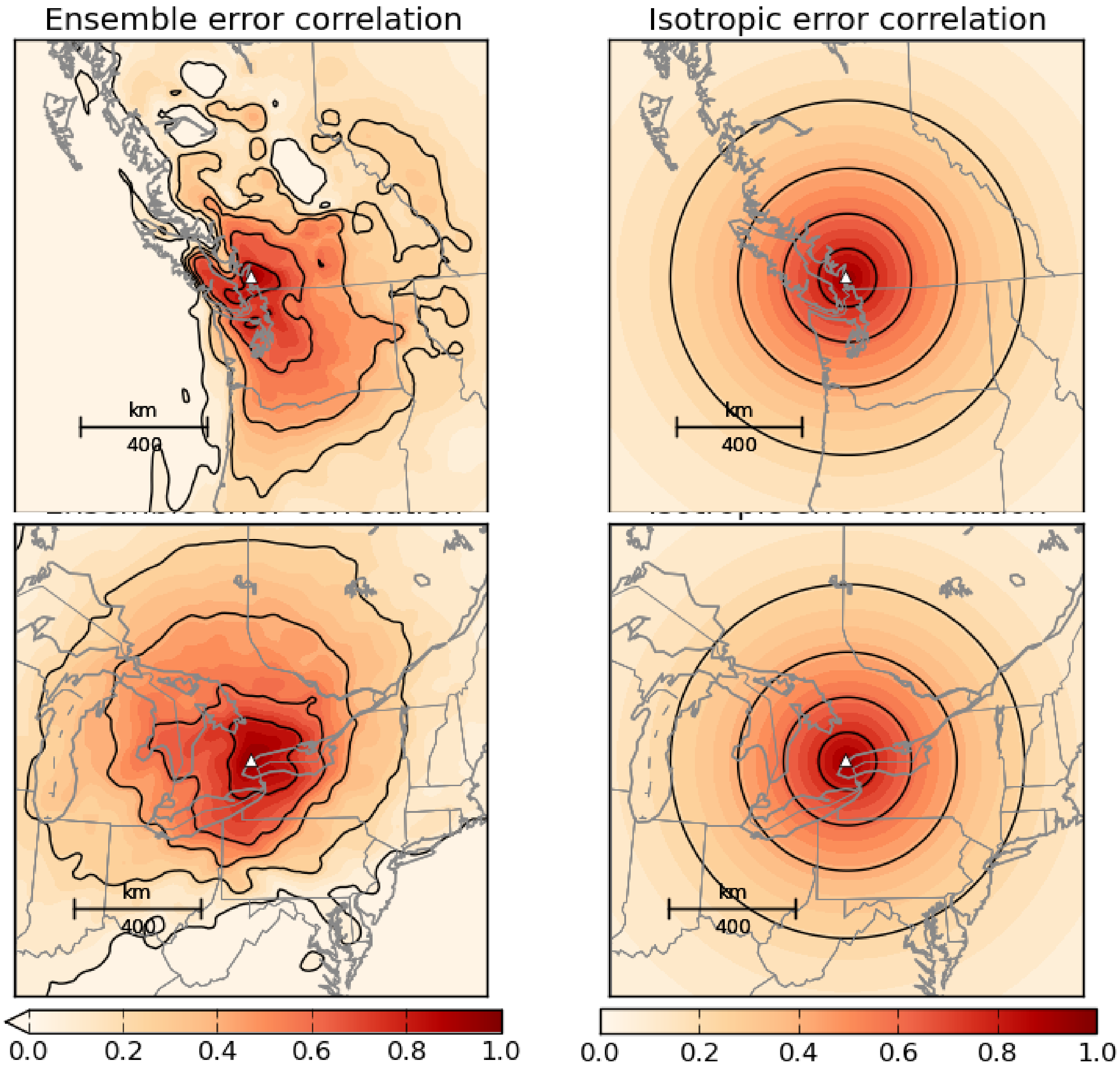
representors

60 days of GEM-MACH output at 21 UTC
And using localization

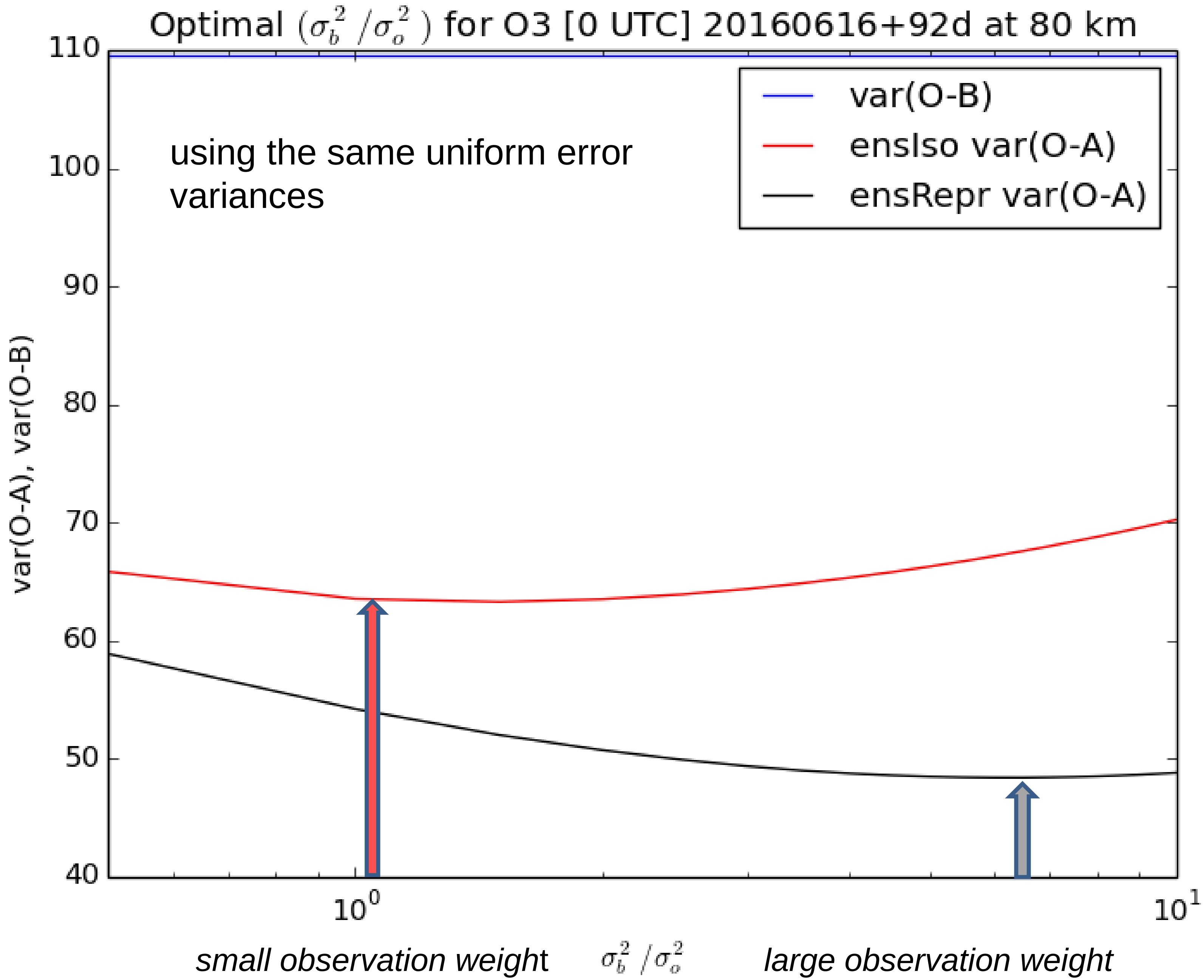
Vancouver

Ozone
21 UTC

Toronto



non-isotropic, non-homogeneous error correlations depends on stationary features

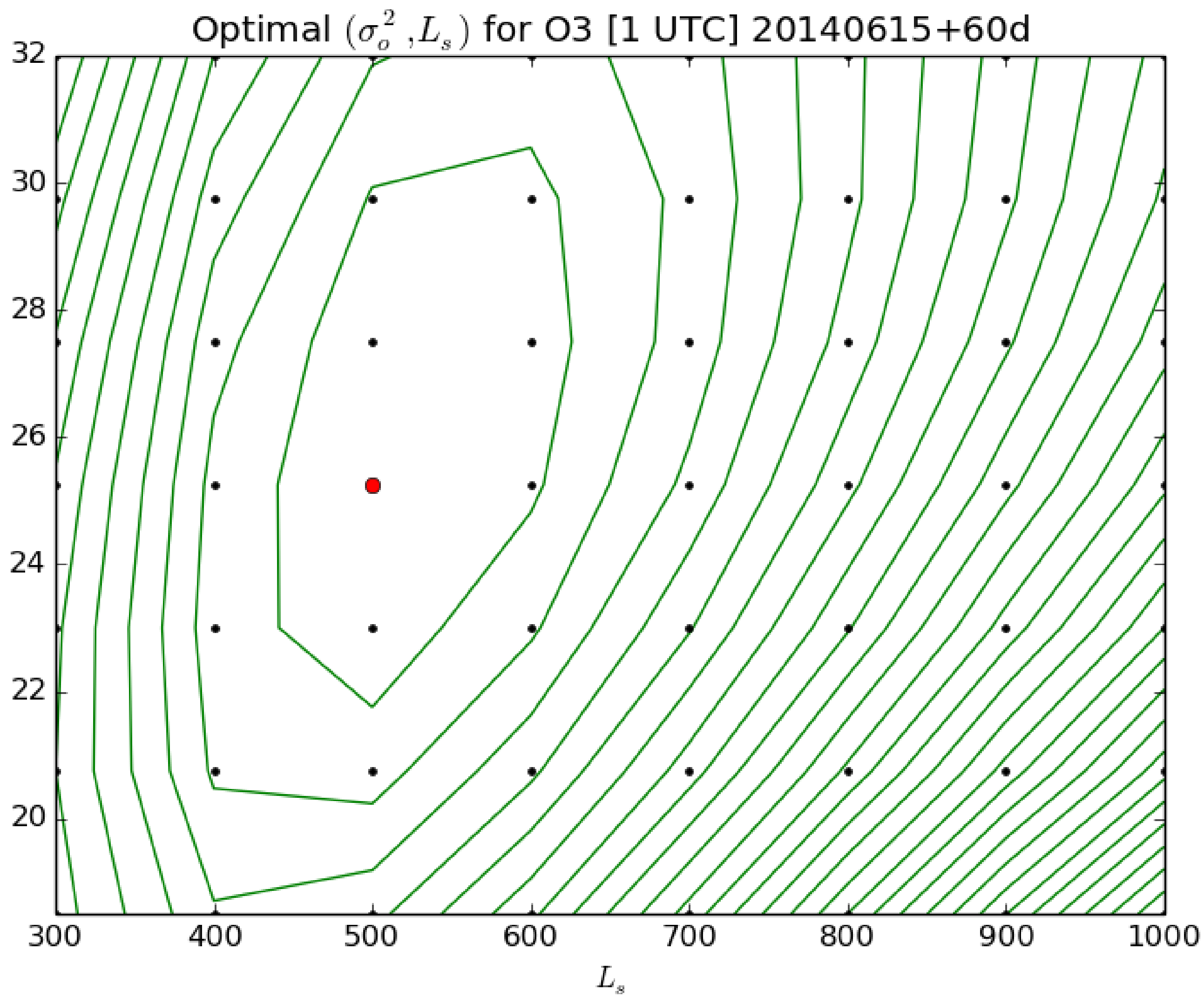


Optimal (σ_b^2/σ_o^2) for O3 [0 UTC] 20160616+92d at 80 km
var(O-B)
ensIso var(O-A)
ensRepr var(O-A)
using the same uniform error variances
var(O-A), var(O-B)
110
100
90
80
70
60
50
40
10^0
10^1
small observation weight
σ_b^2/σ_o^2
large observation weight

Adding physical/chemical content in the error statistics

2.2 Error variances by optimizing (minimizing) the analysis error

contours of var(O-A)$_{cv}$



obs error variance

background error from

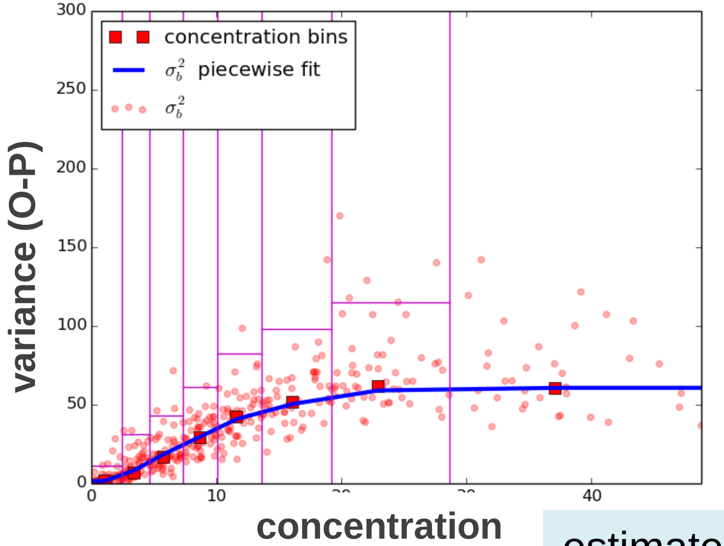
$$\left\langle (O-P)^2 \right\rangle = \sigma_O^2 + \sigma_B^2$$

length-scale used for localization

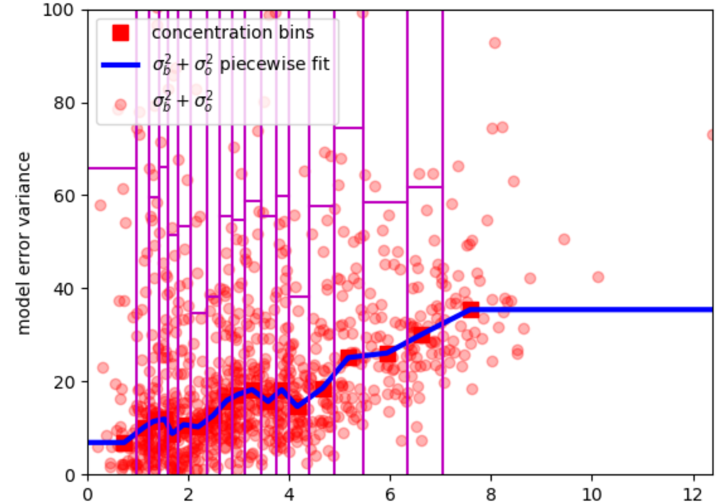
Adding physical/chemical content in the error statistics

Background error variance functional mapping onto the grid

NO_2 night time

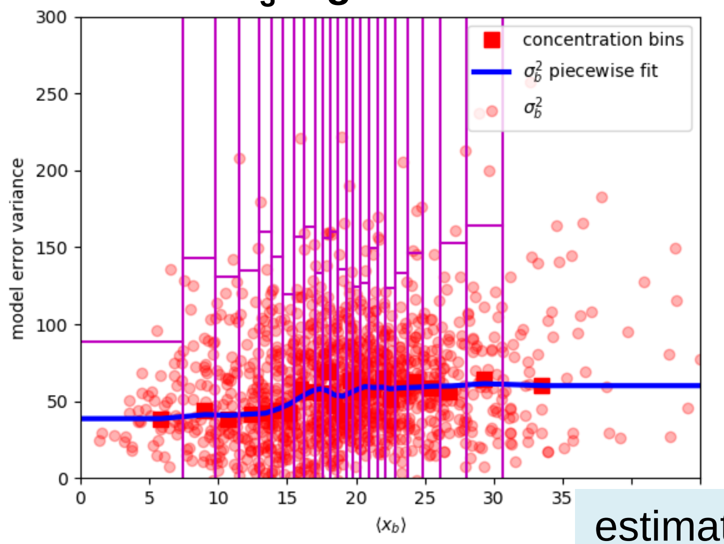


PM2.5

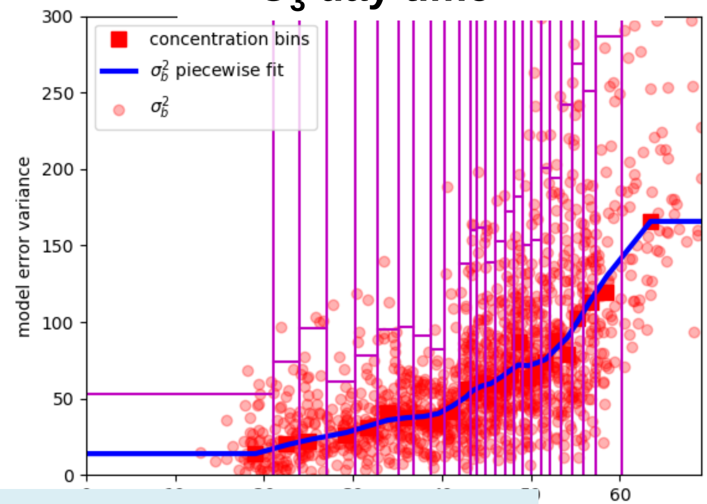


estimate the ratio σ_o^2 / σ_b^2 the observation weight

O_3 night time



O_3 day time



estimate the observation error variance

Conclusion/Summary

- Cross-validation can be used in conjunction
  with innovation covariance consistency
  Estimate an optimal analysis error variance,
  true observation error variance and true
  background error variance

- Geometric interpretation offer a way to
  Understand and synthesize
  the statistical diagnostics and their assumptions

- Next step is to develop an online estimation of covariance parameters

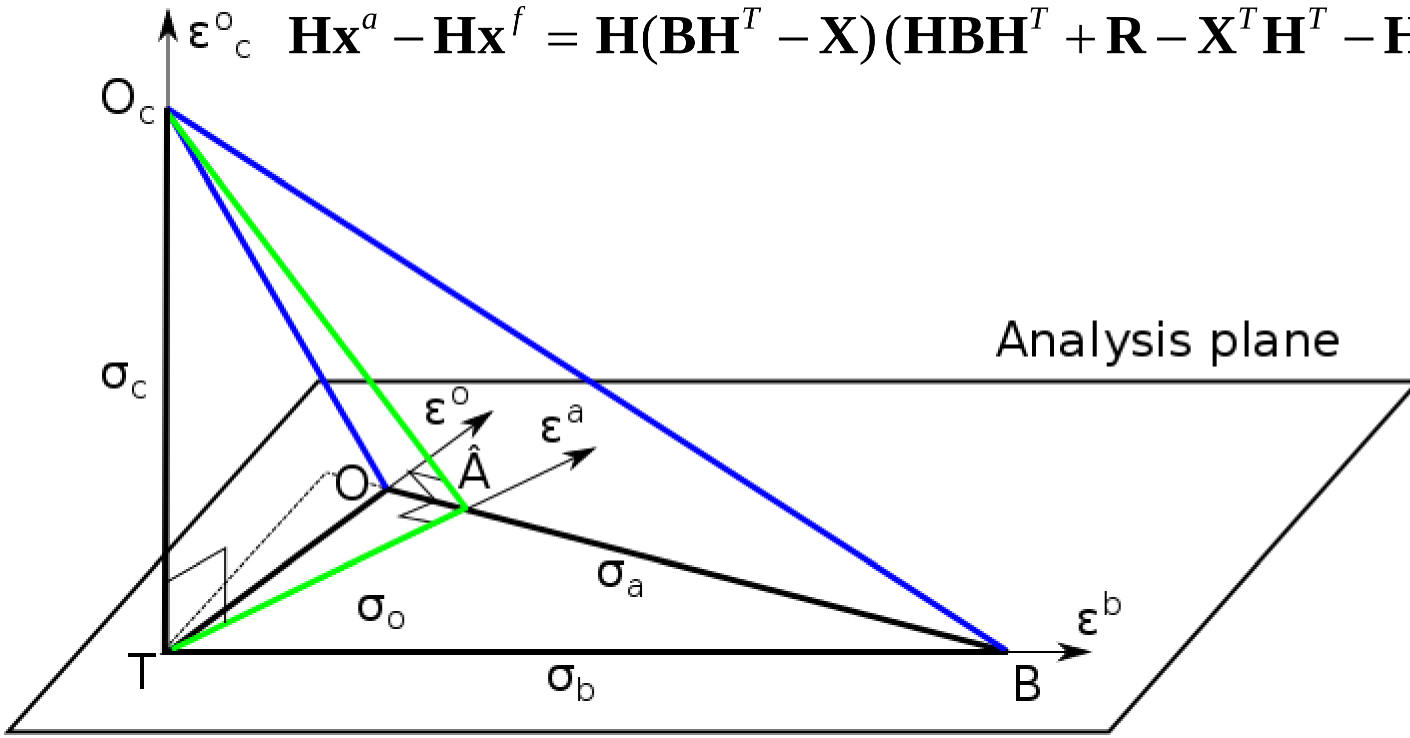
- Sina Voshtani presentation indicated that covariance parameter estimates still holds for satellite observations (spatial correlation of errors). Explain why this cross-validation method seems still work in th is context

End of presentation

Correlated observation and background errors

$$\mathbf{X} = \mathbb{E}[\boldsymbol{\varepsilon}^b (\boldsymbol{\varepsilon}^o)^T]$$

$$\mathbf{Hx}^a - \mathbf{Hx}^f = \mathbf{H}(\mathbf{BH}^T - \mathbf{X})(\mathbf{HBH}^T + \mathbf{R} - \mathbf{X}^T\mathbf{H}^T - \mathbf{HX})^{-1}$$



Correlated observation and background errors

$$X = \rho \sigma_b \sigma_o$$

$$(A - B) = (\sigma_b^2 - \rho \sigma_b \sigma_o)\,(\sigma_b^2 + \sigma_o^2 - 2\rho \sigma_b \sigma_o)^{-1}$$

The HL and MDJ diagnostic continue to be valid but not the D diagnostic