Environment and Climate Change Canada   Environnement et Changement climatique Canada

Canada

Atmospheric Composition Data Assimilation: Selected topics

Richard Ménard, Martin Deshaies-Jacques

Air Quality Research Division, Environment and Climate Change Canada

Simon Chabrillat, Quentin Errera, and Sergey Skachko

Belgium Institute for Space Aeronomy

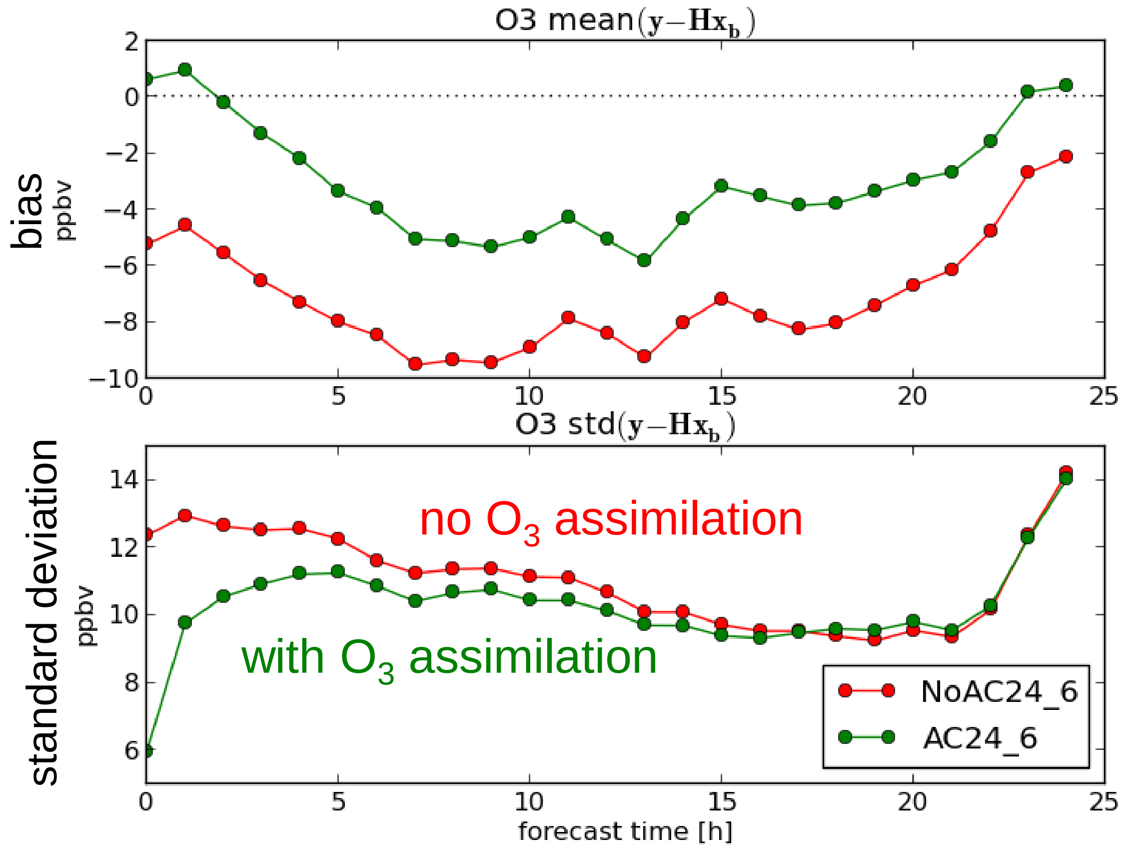
International Symposium on Data Assimilation, Munich , March 2018

Atmospheric composition : A strongly constrained system

- Coupling with the meteorology is in each 3D grid volume (winds, temperature, humidity, clouds/radiation)
- Chemical solution is also controlled by chemical sources and sinks

Boundary layer O_3

Forecast verification after O_3 assimilation



Atmospheric composition as a slaved /controlled system

- As coupled system have to consider the Information content:
  - Chemical model
  - Chemical observations
  - Meteorology
- Favors online chemistry-meteorology models
- No for Incremental formulation with a lower resolution chemistry
- Because the control of meteorology on chemistry is so strong, the feedback of chemistry on meteorology is small
- Except for long-lived species, chemical forecast is not the most useful product, but analyses are
- Forecast improvement is better addressed with parameter (e.g. sources) estimation

Notation and definitions

Continuity equation; ρ dry air density

$$\frac{1}{\rho}\frac{D\rho}{Dt}+\nabla\cdot\mathbf{V}=0 \quad \text{Lagrangian} \qquad ; \qquad \frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\mathbf{V})=0 \quad \text{Eulerian}$$

Mixing ratio c_i of constituent i is defined as $c_i=\dfrac{\rho_i}{\rho}$

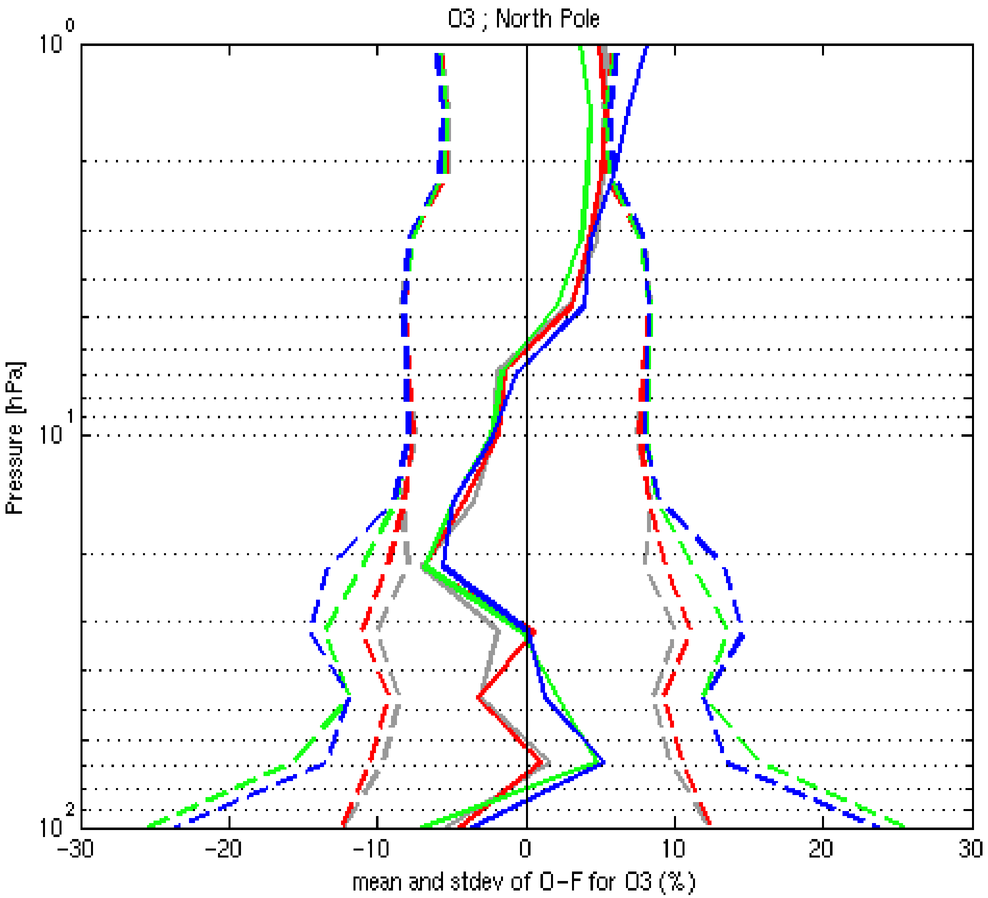
$$\frac{Dc_i}{Dt}=0 \quad \text{Lagrangian} \qquad ; \qquad \frac{\partial c_i}{\partial t}+\mathbf{V}\cdot\nabla c_i=0 \quad \text{Eulerian}$$

The consistency between the wind field $\mathbf{V}$ used for the continuity equation and the wind field $\mathbf{V}$ used for the transport of the constituent i is called the **mass consistency**

CTM (Chemical Transport Models) are offline (from meteorology) models that uses different $\mathbf{V}$ and often at different resolution: they were introduced to use met analyses

Online chemical models have mass consistency and same resolution as the meteorology

GEM-BACH online with meteo model / BASCOE CTM offline model driven by 6-hourly meteorological analyses



Validation with MIPAS OFL
20030825-20030904, 60°N-90°N

GEM-BACH, EC 4D-VAR (K4BCS304)

GEM-BACH, EC 3D-VAR (K3BCS304)

BASCOE CTM, EC 3D-VAR (v3s85)

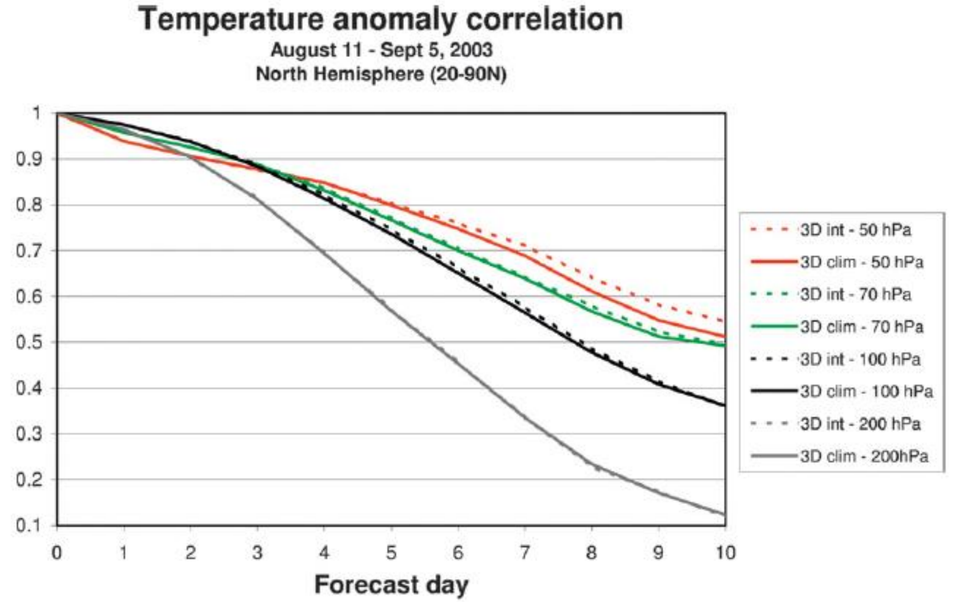
BASCOE CTM, ECMWF (d2003G)

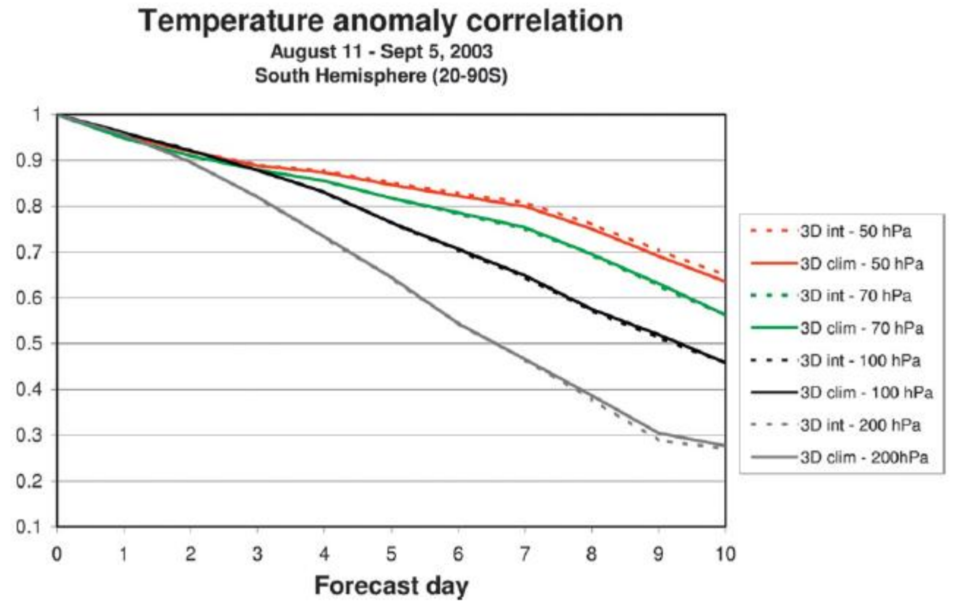
Coupled Assimilation O_3/meteorology

Assimilation of limb sounding MIPAS observations of O_3 with AMSU-A

MIPAS vertical resolution 1-3 km

solid lines
    no radiation feedback
dashed lines
    assim O_3 used in radiative scheme (k-correlated method)



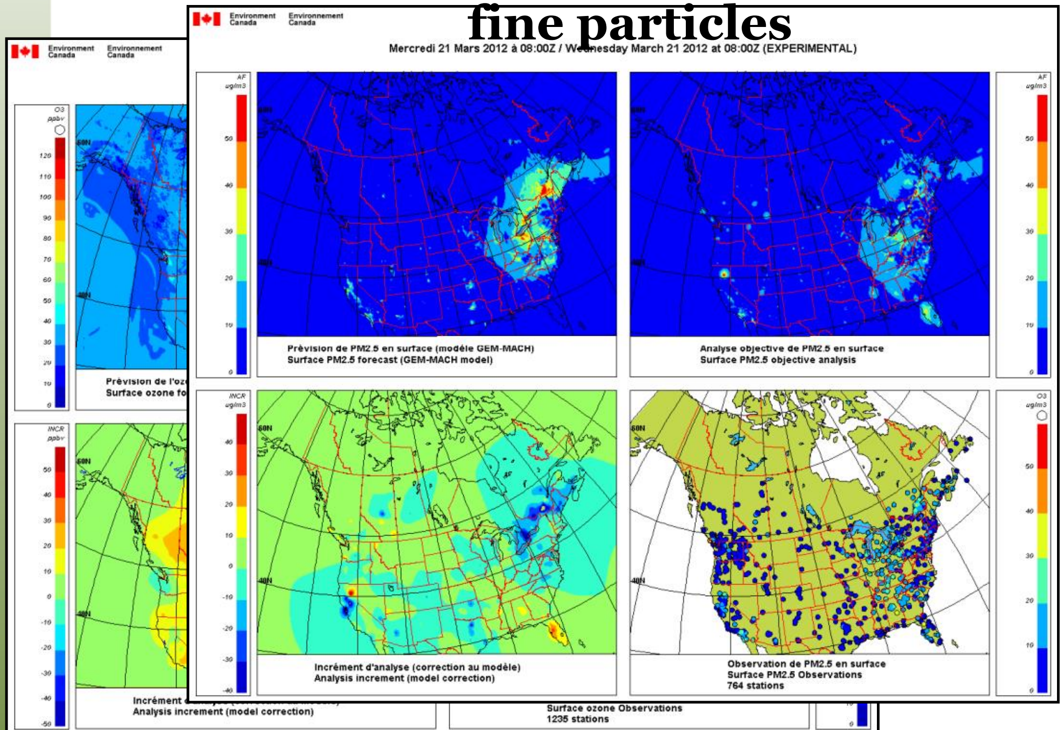


de Grandpré et al., 1997, MWR

Operational air quality analysis

Analysis of O_3, NO_2, SO_2, $PM_{2.5}$, PM_{10} each hour

experimental since 2002*, operational since Feb 2013*



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 (Robichaud et al. *2015*, *Air Qual Atmos Health)*
- Multi-year data set (2002-2012) Robichaud and Ménard 2014, *ACP*)

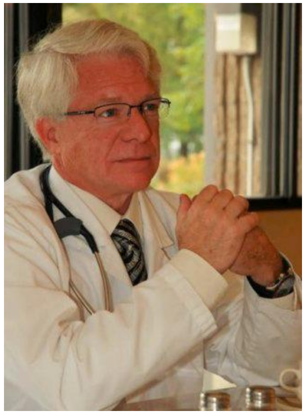
What does the medicine has to say about air pollution ?

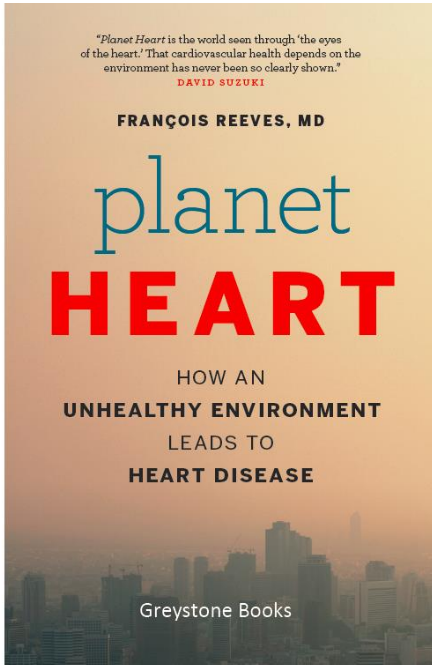
- Interact with many groups (e.g. Health Canada, Environmental Cardiology)

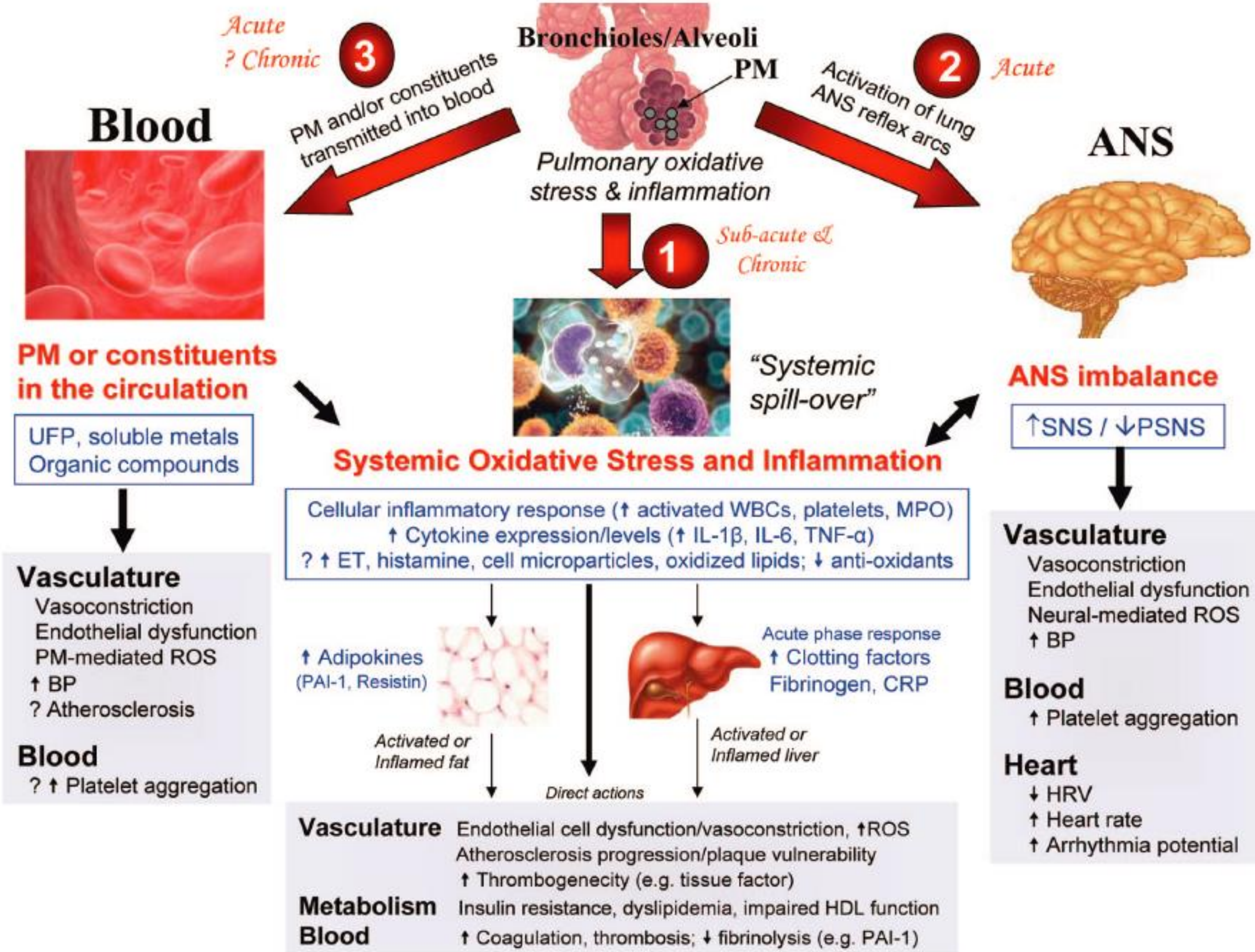
- *In terms of mass, we breathe-in per day about 20 kg of air,*
  2 kg of liquid
  and 1 kg of solid food

12-25 breath per minutes each is about 1 liter
≈ 20,000 liter of air/day

- *Pulmonology and cardiology are strongly linked*







Université de Montréal

Brook et al; Particulate Matter Air Pollution and CVD. Circulation 2010. 121; 2331-2378

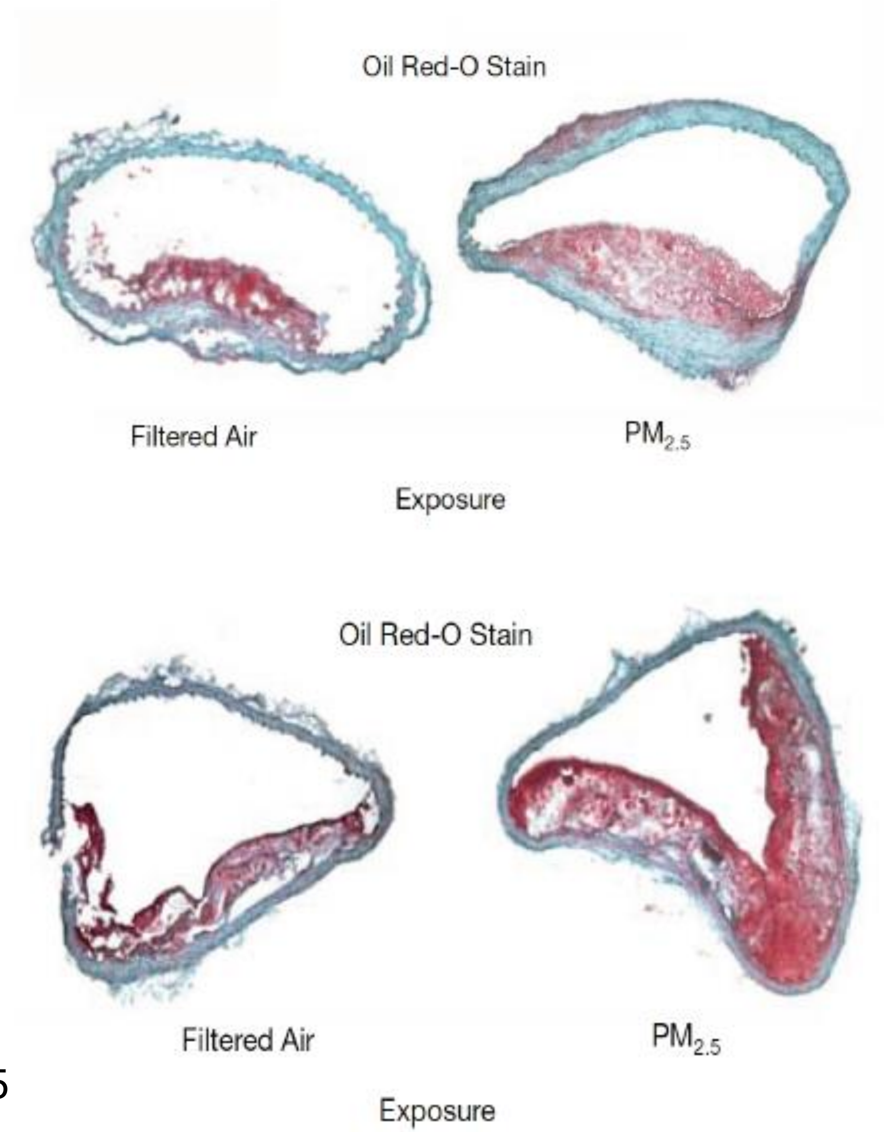
Filtered air

Polluted air

Mean 15 $\mu g/m^3$

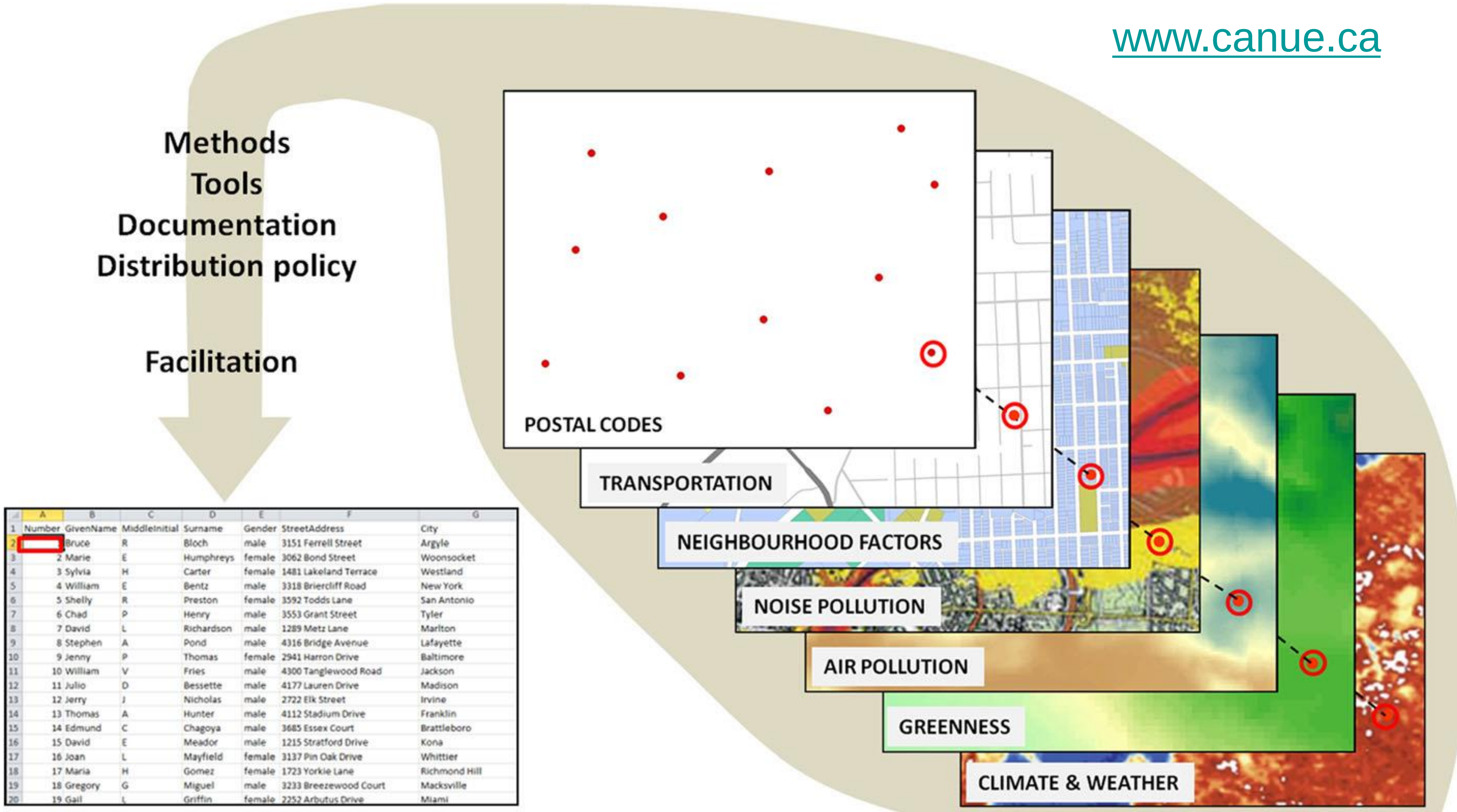
Normal Chow

High-Fat Chow



Sun et al., JAMA 2005

Canadian Census Health and Environmental Cohort (CanCHEC) and the Canadian Urban Environmental Health Research Consortium (CANUE)



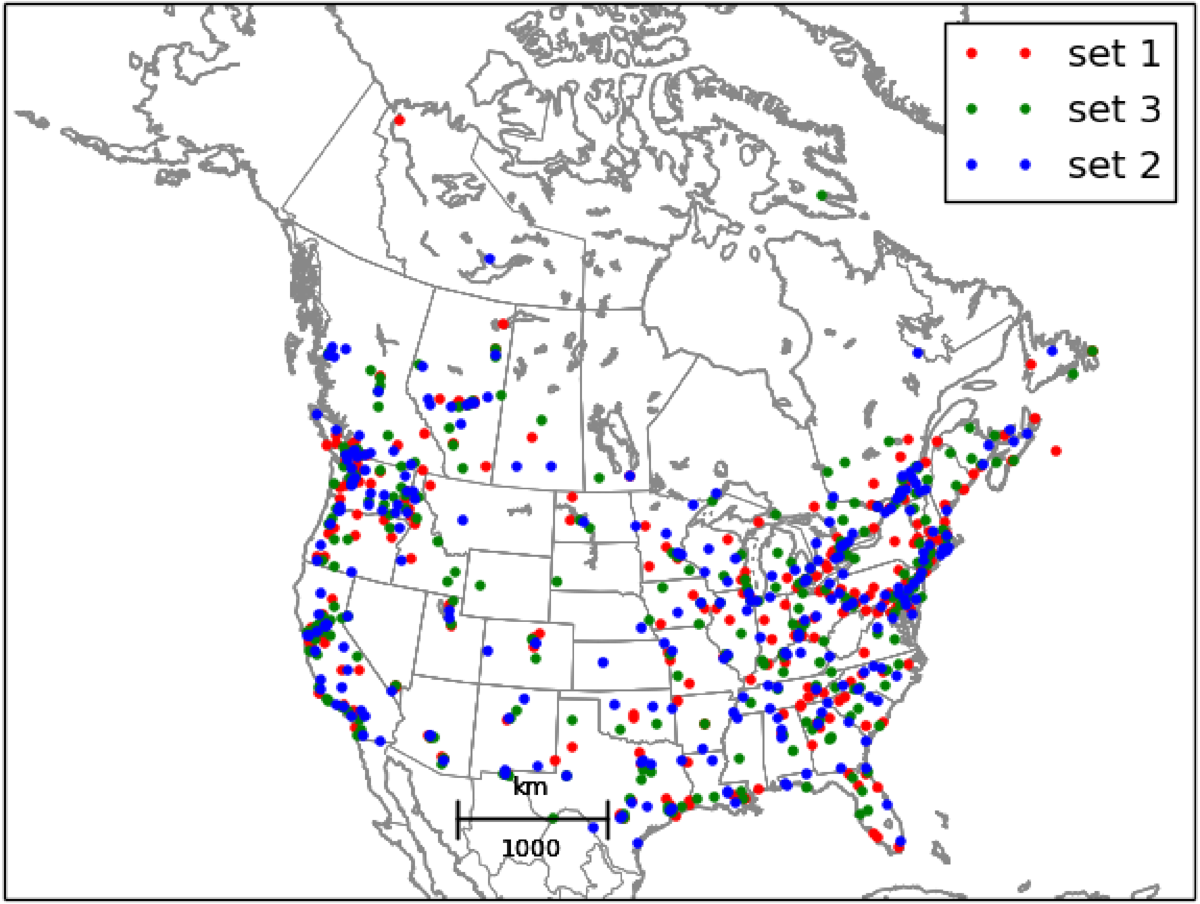
Brook et al. BMC Public Health (2018) 18:114

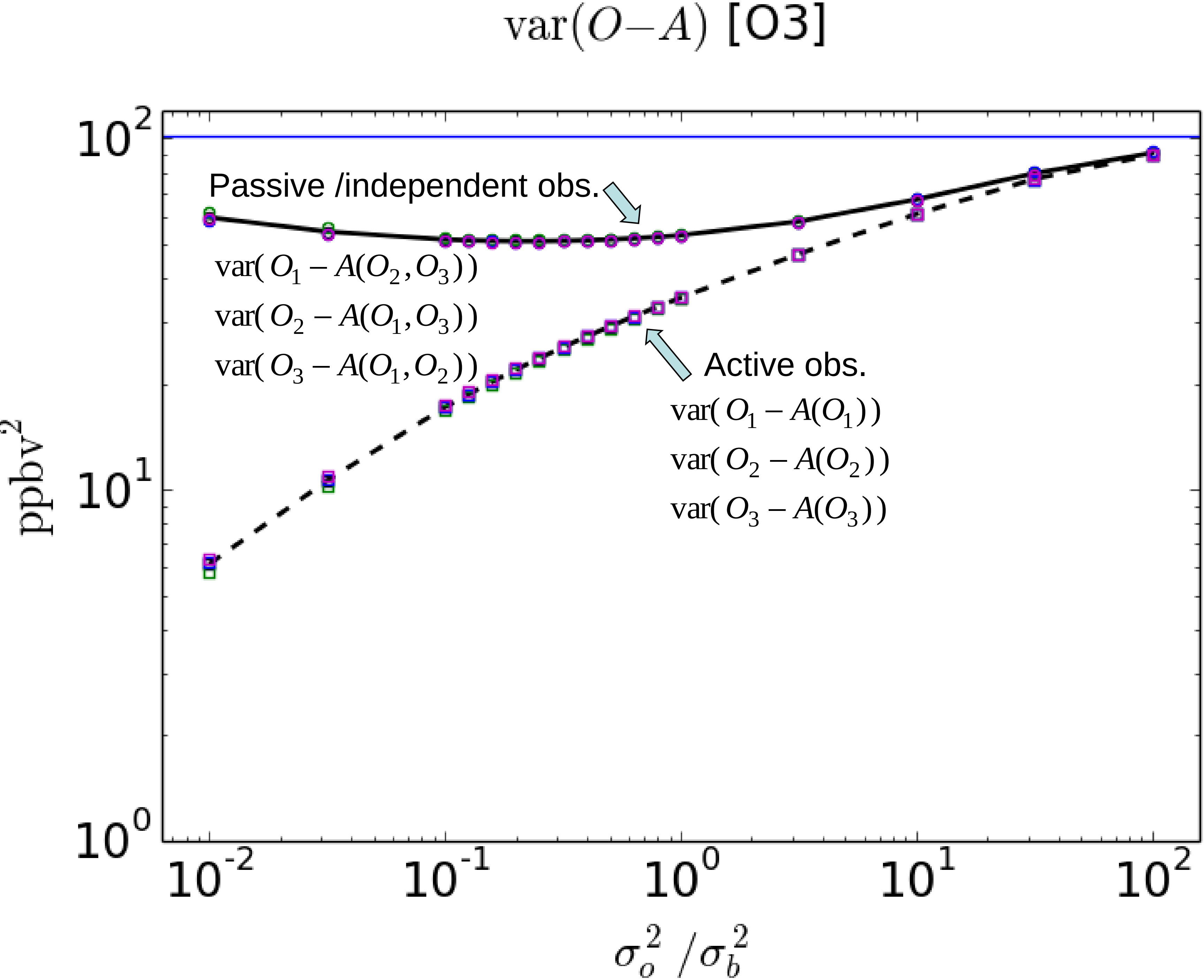
Evaluation of analysis by cross-validation (or leave-out observations)

Ménard, R and M. Deshaies-Jacques Part I: Using verification metrics. *Atmosphere* **2018**, *9*(3), 86, doi:10.3390/atmos9030086

Ménard, R and M. Deshaies-Jacques Part II: Diagnostic and optimization of analysis error covariance. *Atmosphere* **2018**, *9*(2), 70; doi:10.3390/atmos9020070

3 spatially random distributed set of observations





var(O−A) [O3]
Passive /independent obs.
var($O_1 - A(O_2, O_3)$)
var($O_2 - A(O_1, O_3)$)
var($O_3 - A(O_1, O_2)$)
Active obs.
var($O_1 - A(O_1)$)
var($O_2 - A(O_2)$)
var($O_3 - A(O_3)$)
ppbv^2
10^2
10^1
10^0
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
σ_o^2 / σ_b^2

Estimation of error covariances in observation space ($\mathbf{HBH}^T$, R)

Theorem on estimation of error covariances in observation space

Assuming that observation and background errors are uncorrelated, the *necessary* and *sufficent* conditions for error covariance estimates to be equal to the *true* observation and background error covariances are:

1) $\mathbb{E}[(O-B)(O-B)^T] = \mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T + \tilde{\mathbf{R}}$ *Innovation covariance consistency*

A scalar version of this condition is

$$tr\left\{\mathbb{E}[\mathbf{d}\mathbf{d}^T]\,(\mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T + \tilde{\mathbf{R}})^{-1}\right\} = \mathbb{E}\left\{tr[\mathbf{d}^T(\mathbf{H}\tilde{\mathbf{B}}\mathbf{H} + \tilde{\mathbf{R}})^{-1}\mathbf{d}]\right\} = \mathbb{E}[\chi^2] = N_p$$

or

$$J_{\min} = N_p / 2$$

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

Ménard, **2016**, QJRMS

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

Diagnostic of the Kalman gain condition

- Daley 1992 (MWR) suggested that the time lag-innovation covariance to be equal zero (assuming observation errors are serially uncorrelated)
- Here we suggest to use cross-validation

Simple interpretation with a scalar problem

1) Innovation consistency says that the **sum** of observation and background error variances is the sum of the true error variances

2) Kalman gain condition says that the **ratio** of observation to background error variance is the ratio of the true error variances

Verification of analysis by cross-validation: A geometric view

Hilbert spaces of random variables

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

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

Can defined a metric as

$$\|X\|_2 = \mathbb{E}\left[X^2\right]$$

Uncorrelated random variables X , Y would be orthogonal $\langle X, Y \rangle = 0$

Verification of analysis by cross-validation: A geometric view

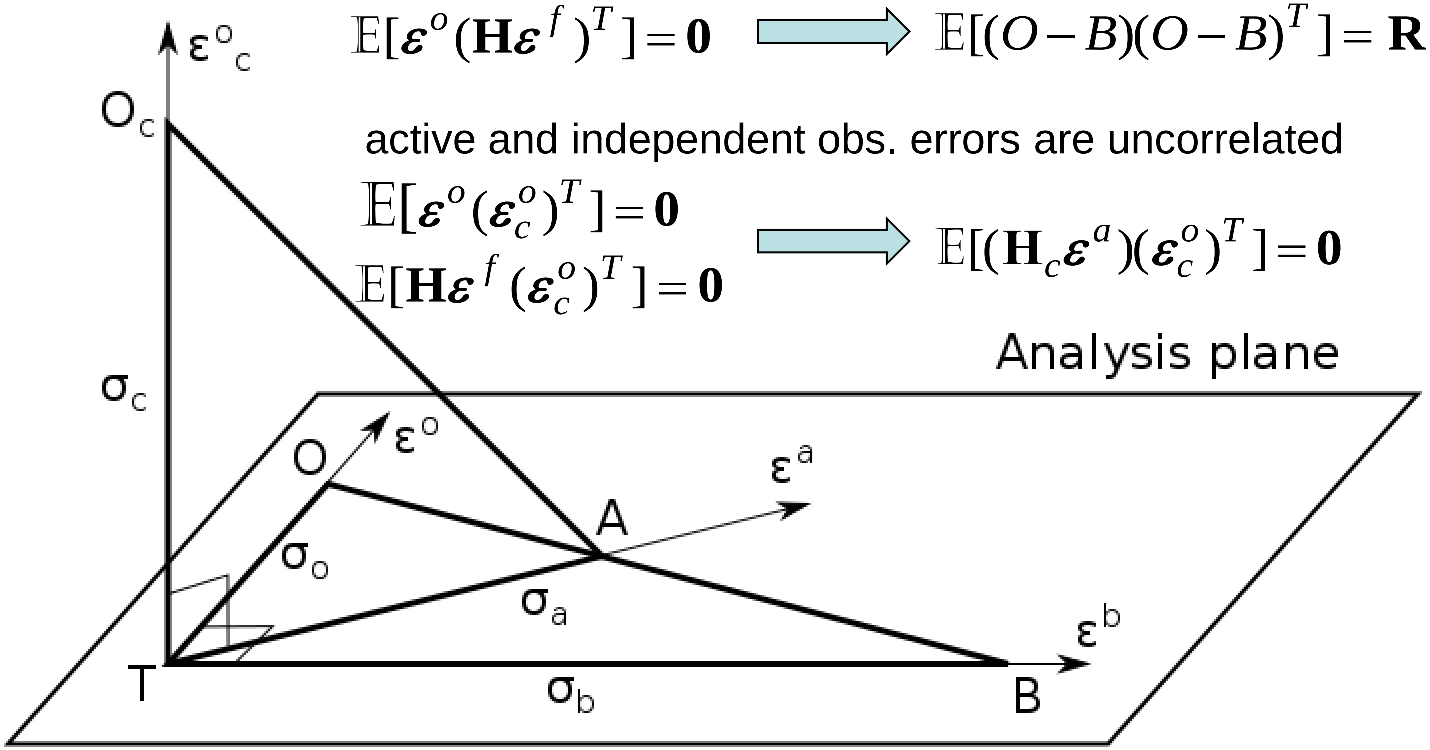
obs and background errors are uncorrelated

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

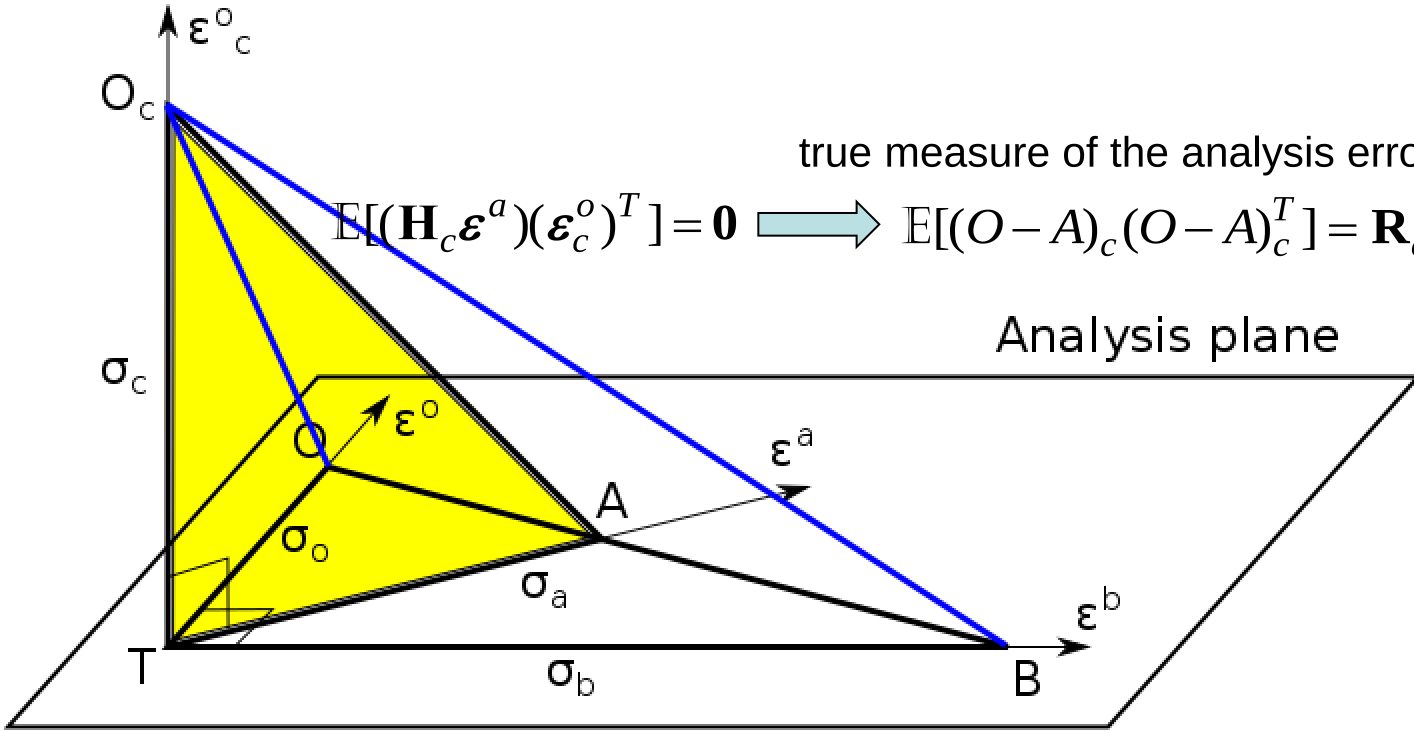
active and independent obs. errors are uncorrelated

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

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



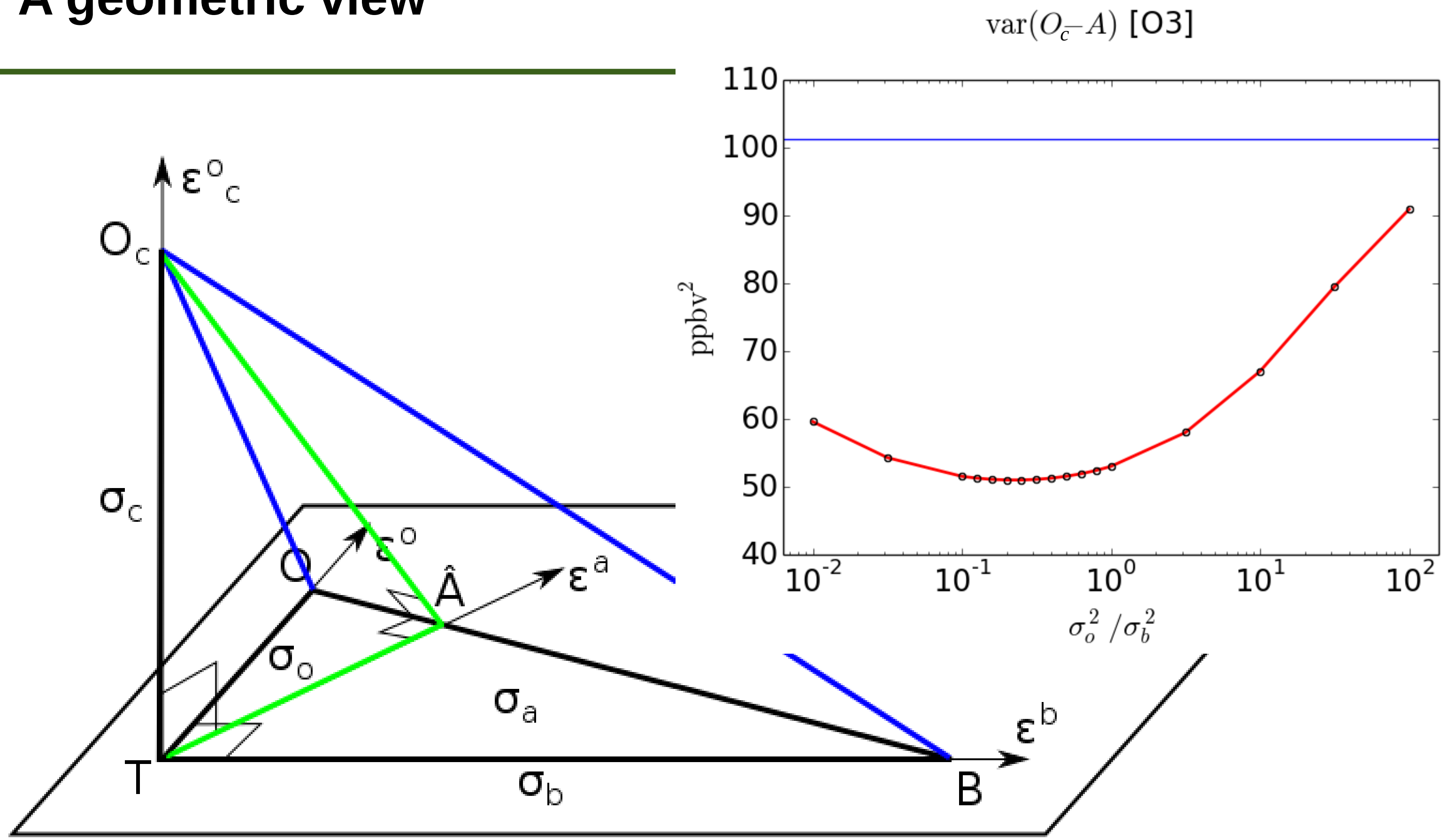
Verification of analysis by cross-validation: A geometric view



true measure of the analysis error covariance

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

Verification of analysis by cross-validation:
A geometric view

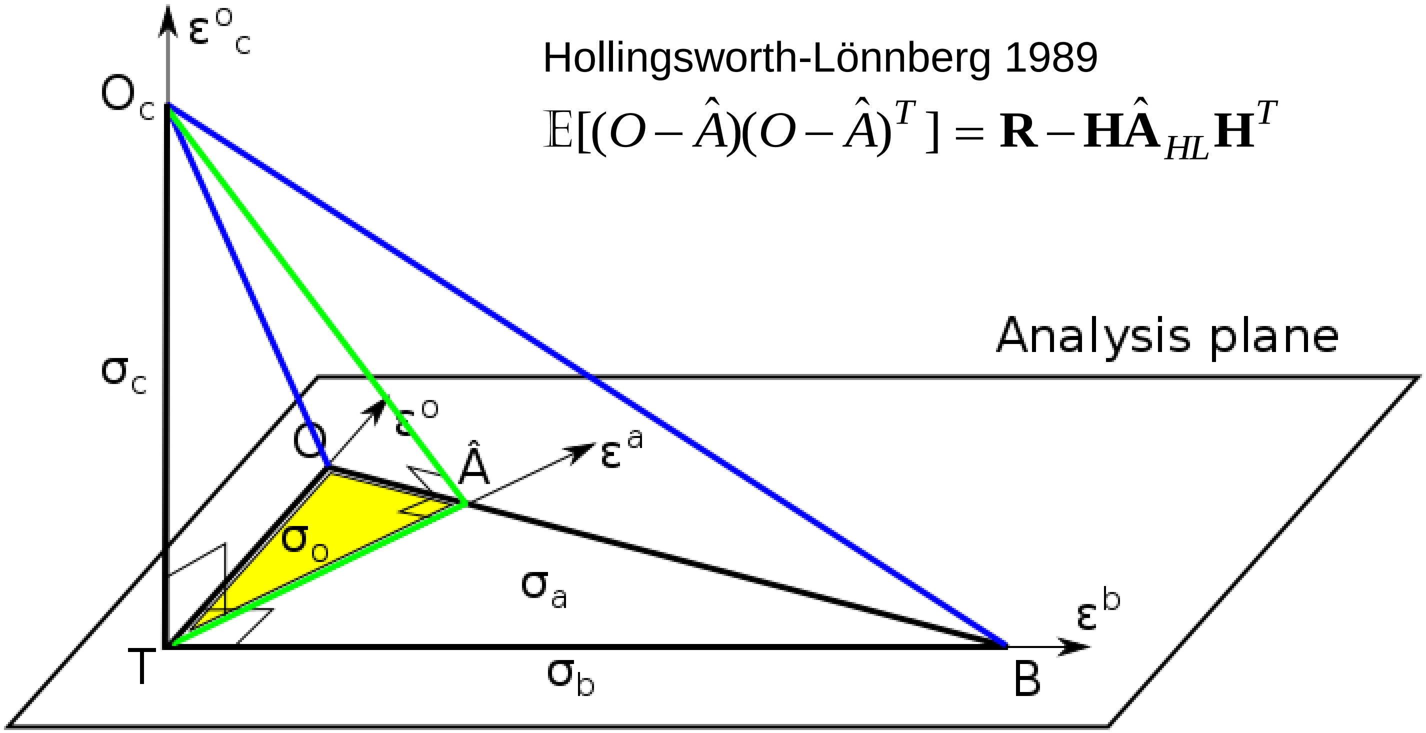


by varying the observation weight while $\mathbb{E}[(O-B)(O-B)^T] = \mathbf{R} + \mathbf{HBH}^T$ we can find a true optimal analysis

Verification of analysis by cross-validation:
A geometric view

Hollingsworth-Lönnberg 1989

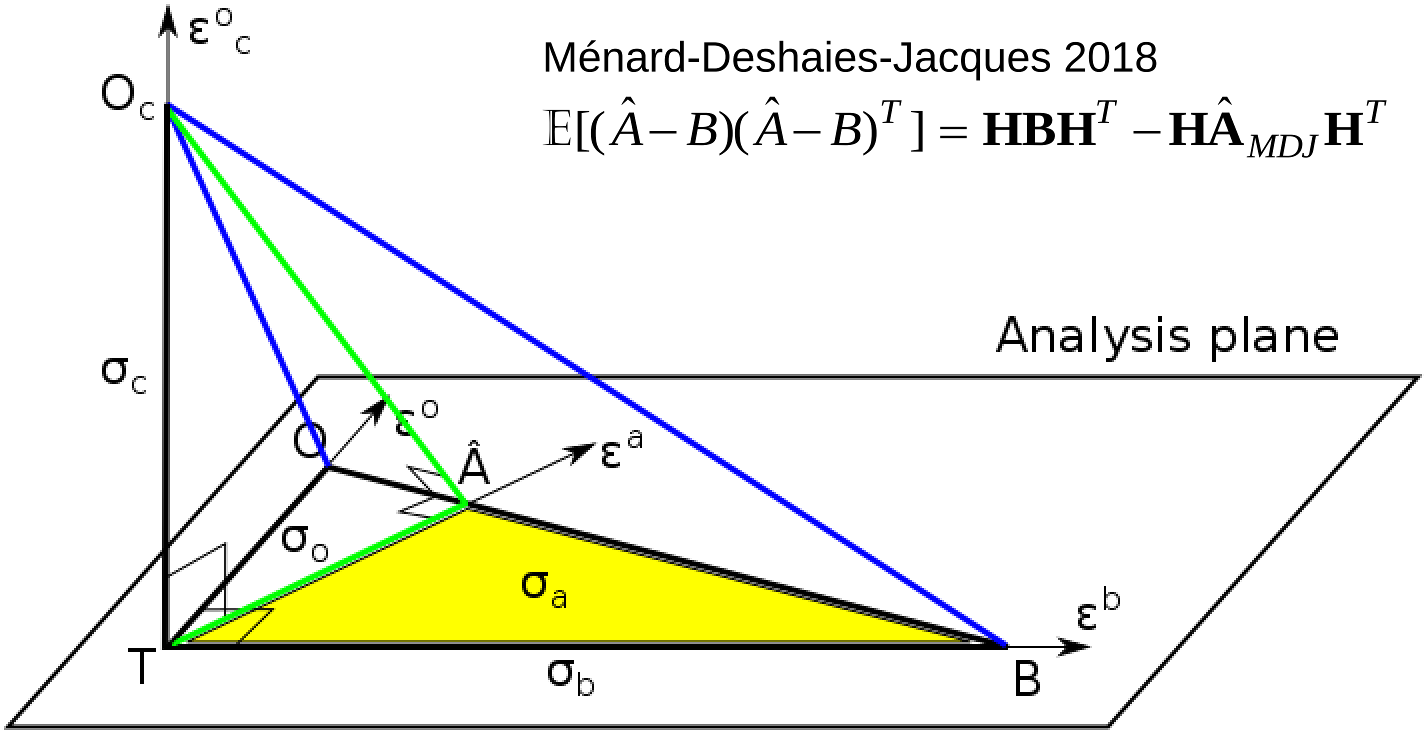
$$\mathbb{E}[(O-\hat{A})(O-\hat{A})^T] = \mathbf{R} - \mathbf{H}\hat{\mathbf{A}}_{HL}\mathbf{H}^T$$



Verification of analysis by cross-validation: A geometric view

Ménard-Deshaies-Jacques 2018

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

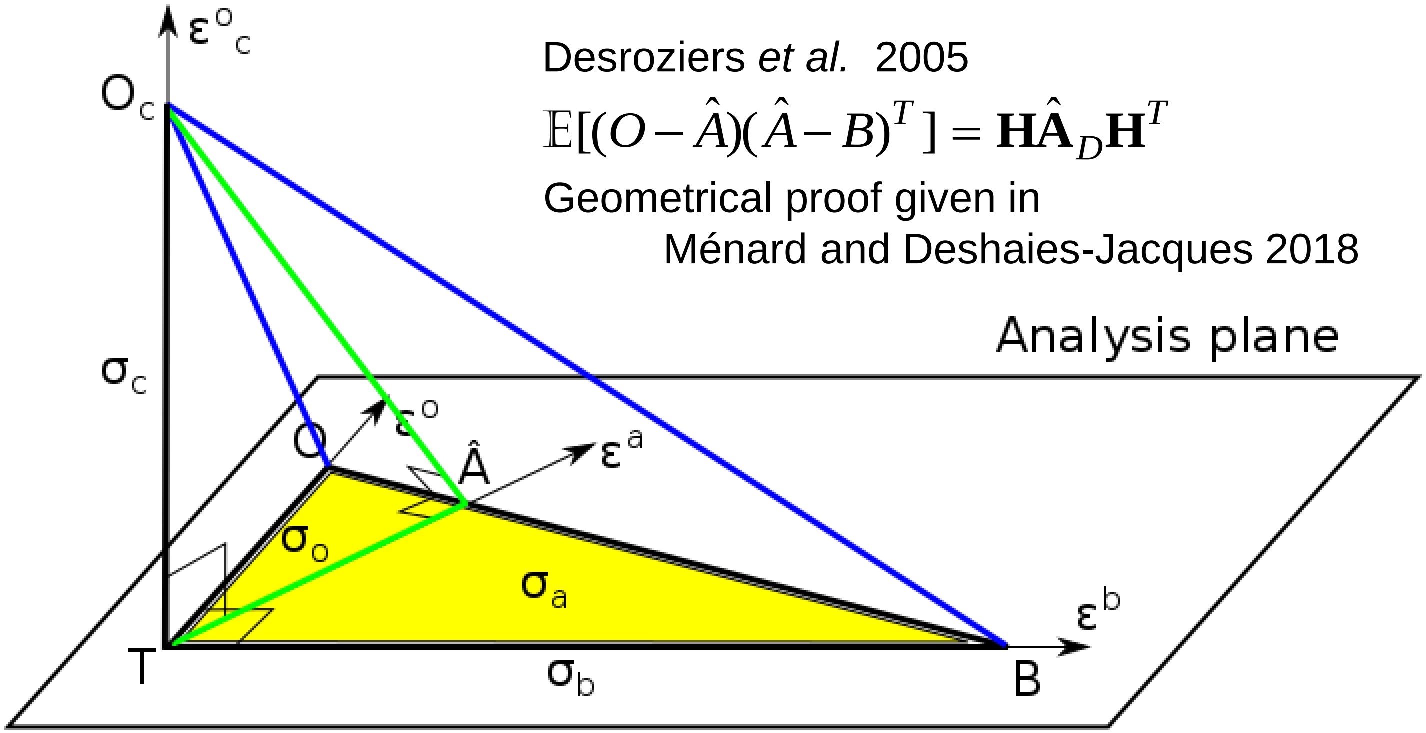


Verification of analysis by cross-validation: A geometric view

Desroziers *et al.* 2005

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

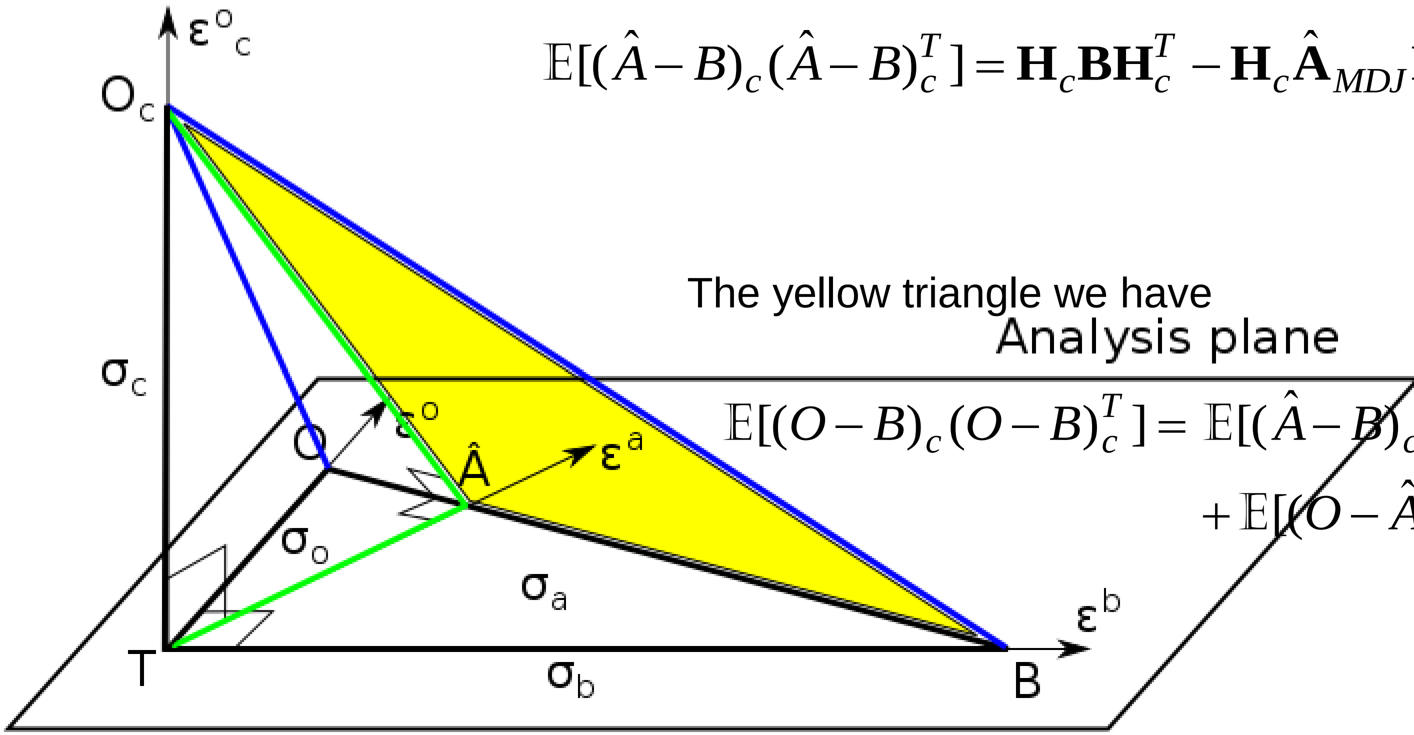
Geometrical proof given in
Ménard and Deshaies-Jacques 2018



Verification of analysis by cross-validation: A geometric view

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

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

Application to O_3 surface analysis

iter 0 (first guess error correlation)
Iter 1 (Maximium Likelihood estimation of correlation length)

Input error statistics

Experiment	L_c (km)	$\langle (O-B)^2 \rangle$	$\hat{\gamma} = \hat{\sigma}_o^2 / \hat{\sigma}_b^2$	$\hat{\sigma}_o^2$	$\hat{\sigma}_b^2$	χ^2 / N_p
O_3 iter 0	124	101.25	0.22	18.3	83	2.23
O_3 iter 1	45	101.25	0.25	20.2	81	1.36

Estimate of analysis error variance at passive observation sites

Experiment	**Passive** $diag(\mathbf{H}_c \hat{\mathbf{A}}_{MDJ} \mathbf{H}_c^T)$	**Passive** $\mathrm{var}[(O-\hat{A})_c] - \sigma_{oc}^2$
O_3 iter 0	26.03	32.72
O_3 iter 1	28.95	28.75

Estimation of active analysis error variance

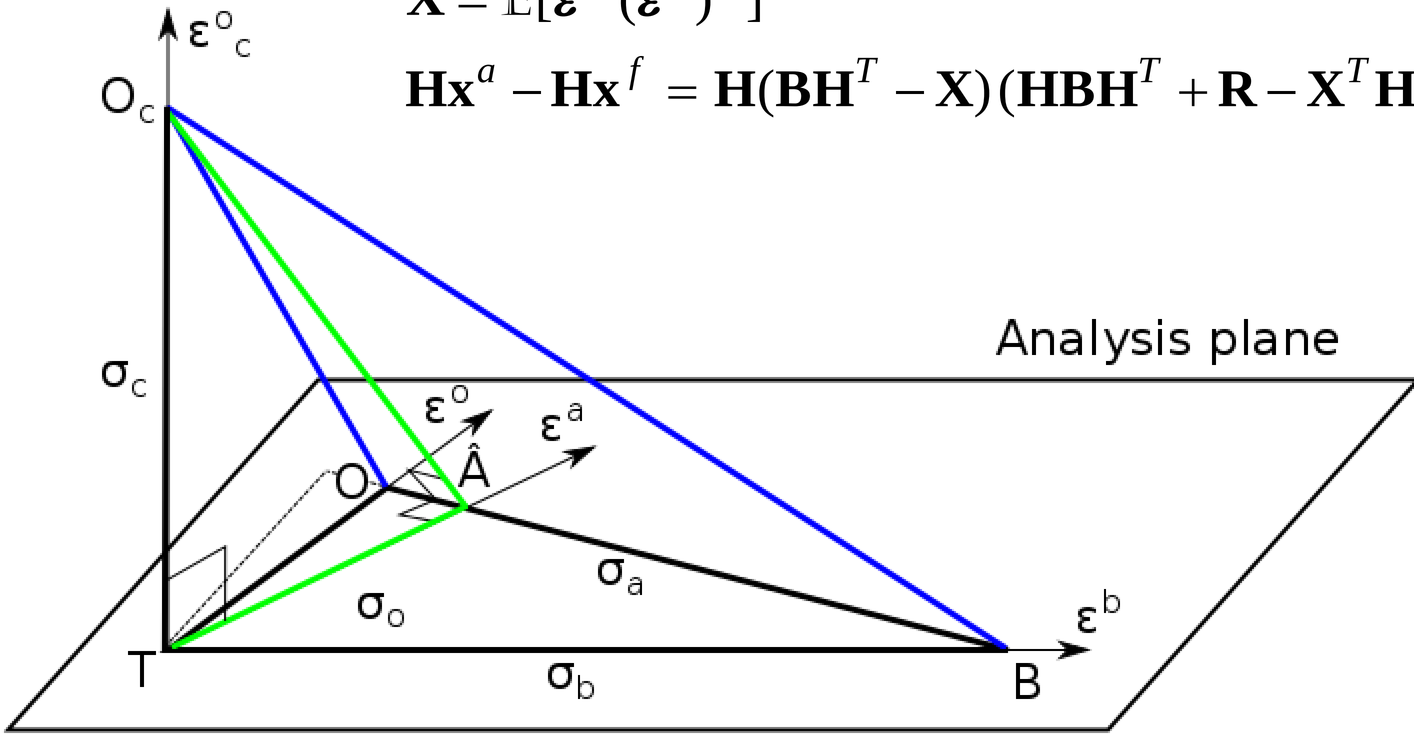
Experiment	**Active** $diag(\mathbf{H}\hat{\mathbf{A}}_{MDJ}\mathbf{H}^T)$	**Active** $diag(\mathbf{H}\hat{\mathbf{A}}_{D}\mathbf{H}^T)$	**Active** $diag(\mathbf{H}\hat{\mathbf{A}}_{HL}\mathbf{H}^T)$	**Active** $diag(\mathbf{H}\hat{\mathbf{A}}_{P}\mathbf{H}^T)$
O_3 iter 0	22.69	9.61	-6.03	5.77
O_3 iter 1	13.32	13.68	8.94	11.60

Verification of analysis by cross-validation:

when active observation and background errors are correlated

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

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



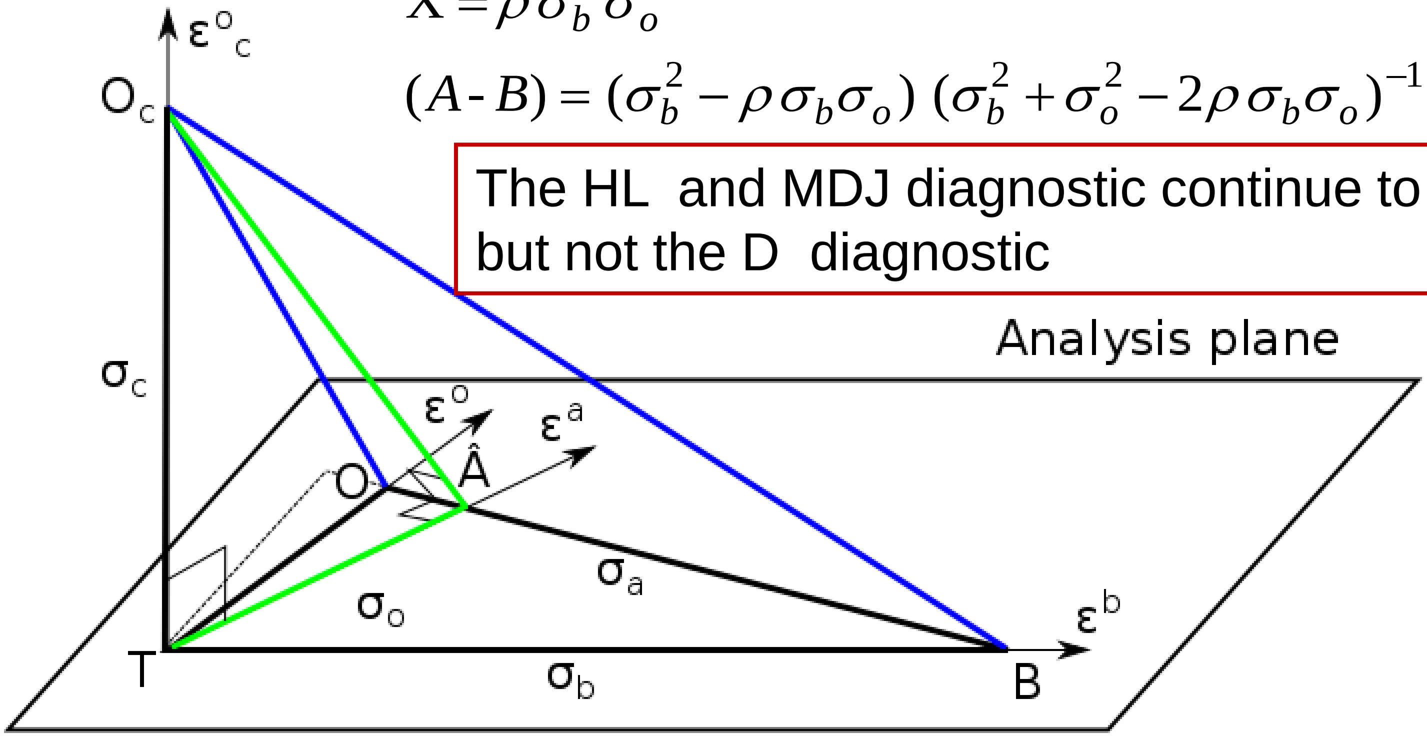
Verification of analysis by cross-validation:

when active observation and background errors are correlated

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

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

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



Verification of analysis by cross-validation:

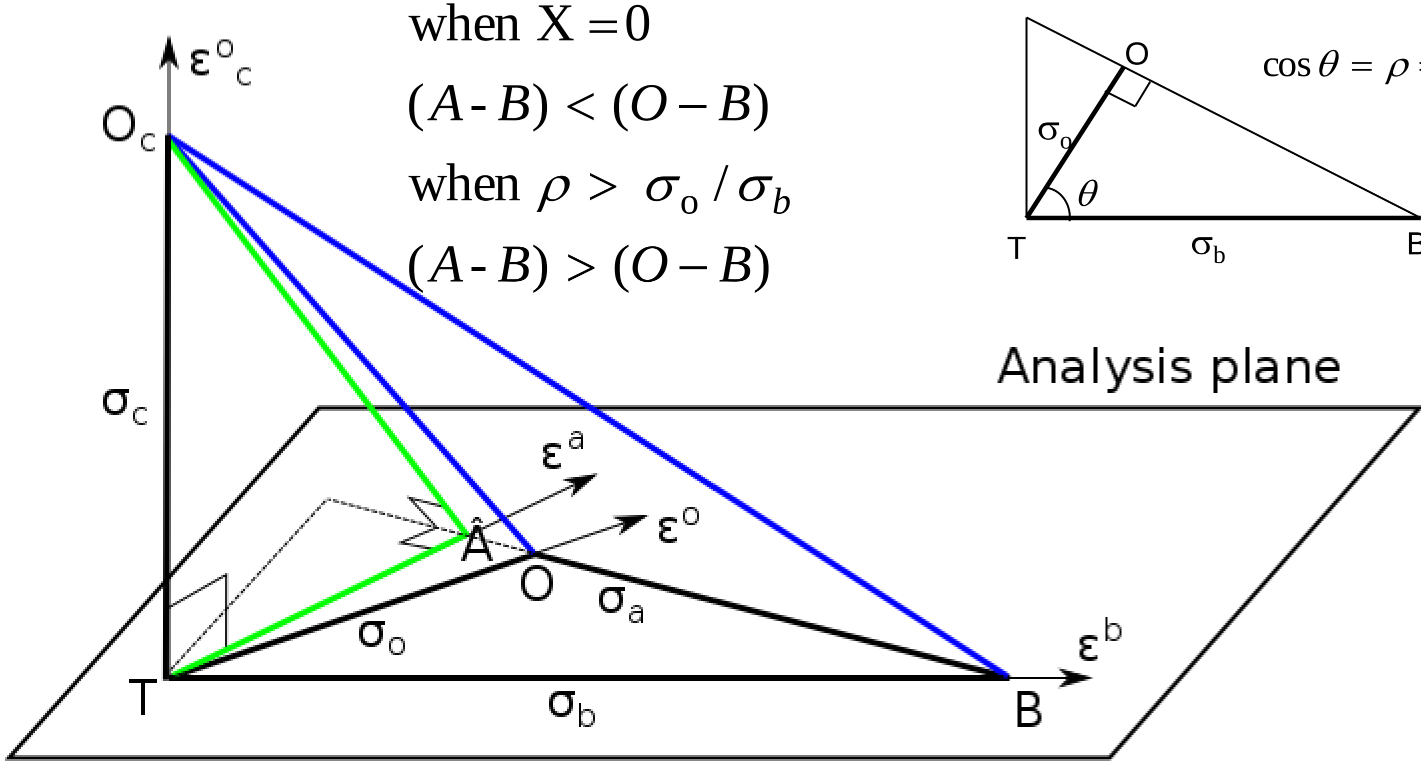
when active observation and background errors are correlated

$$\text{when } X = 0$$

$$(A - B) < (O - B)$$

$$\text{when } \rho > \sigma_o / \sigma_b$$

$$(A - B) > (O - B)$$



$$\cos\theta = \rho = \frac{\sigma_o}{\sigma_b}$$

Lognormal

Simple model: **Random proportional tendencies**

$$\frac{c_i - c_{i-1}}{\Delta t} = \varepsilon_i \, c_{i-1}$$

where ε is random parameter. Suppose we divide a time inverval [0,1] into K Δt intervals

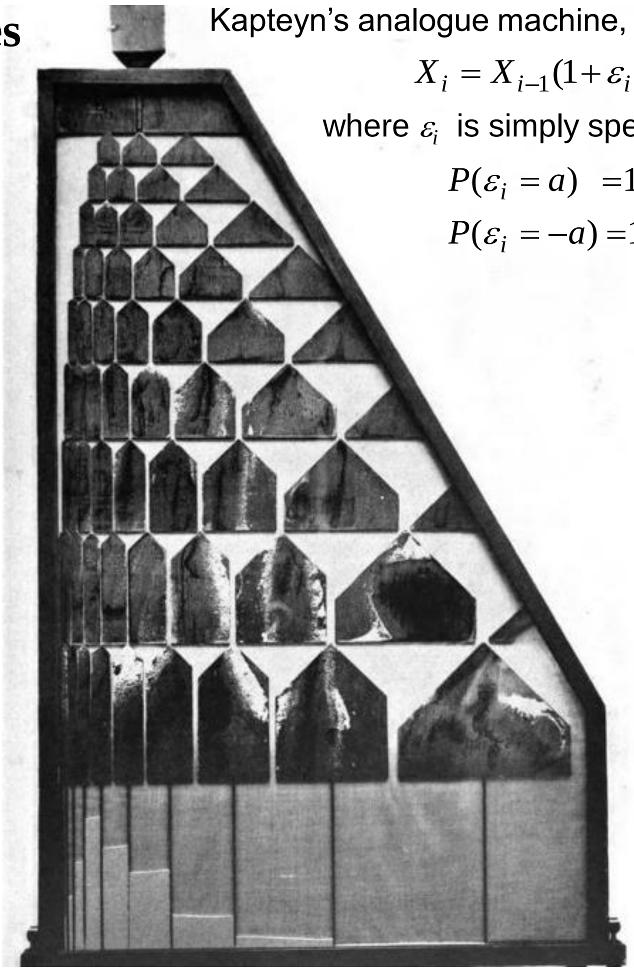
$$\sum_{i=1}^{K}\left(\frac{c_i - c_{i-1}}{c_{i-1}}\right) = \Delta t \sum_{i=1}^{K}\varepsilon_i$$

we can approximate

$$\sum_{i=1}^{K}\left(\frac{c_i - c_{i-1}}{c_{i-1}}\right) \approx \int_{c(t_0)}^{c(t_K)} \frac{dc}{c} = \log c(t_K) - \log c(t_0)$$

and by central limit theorem

$$\Delta t \sum_{i=1}^{K}\varepsilon_i = \Delta t\, K\left(\frac{1}{K}\sum_{i=1}^{K}\varepsilon_i\right) \sim N(\mu,\sigma^2)$$



Kapteyn's analogue machine, replicate

$$X_i = X_{i-1}(1+\varepsilon_i)$$

where ε_i is simply specified by

$$P(\varepsilon_i = a) = 1/2$$
$$P(\varepsilon_i = -a) = 1/2$$

Huize de Wolf laboratory University of Groningen

$$\log \frac{c(t_K)}{c(t_0)} \sim N(\mu, \sigma^2)$$

then

$$\frac{c(t_K)}{c(t_0)} \sim LN(\mu, \sigma^2)$$

is lognormally-distributed.

Analysis step

Positive analysis (lognormal formulation)

(Cohn 1997, Fletcher and Zupanski 2006)

Letting **w** = log **c** , the analysis equation is then of the form

$$\mathbf{w}^a = \mathbf{w}^f + \overline{\mathbf{K}}(\mathbf{w}^o - \mathbf{H}\mathbf{w}^f - \mathbf{b})$$

$$\overline{\mathbf{K}} = \overline{\mathbf{B}}^f \mathbf{H}^T \left(\mathbf{H}\overline{\mathbf{B}}^f \mathbf{H}^T + \overline{\mathbf{B}}^o\right)^{-1}$$

where the $\overline{\mathbf{B}}$ s are the error covariances defined in log space.

b is an observational correction, and has different form whether we consider that:

a) the measurement in log space that is unbiased (in which case **b** = 0)

b) the measurement in physical space is unbiased $b_i = -\overline{B}_{ii}/2$

Transformation of mean and covariance between log and physical quantities

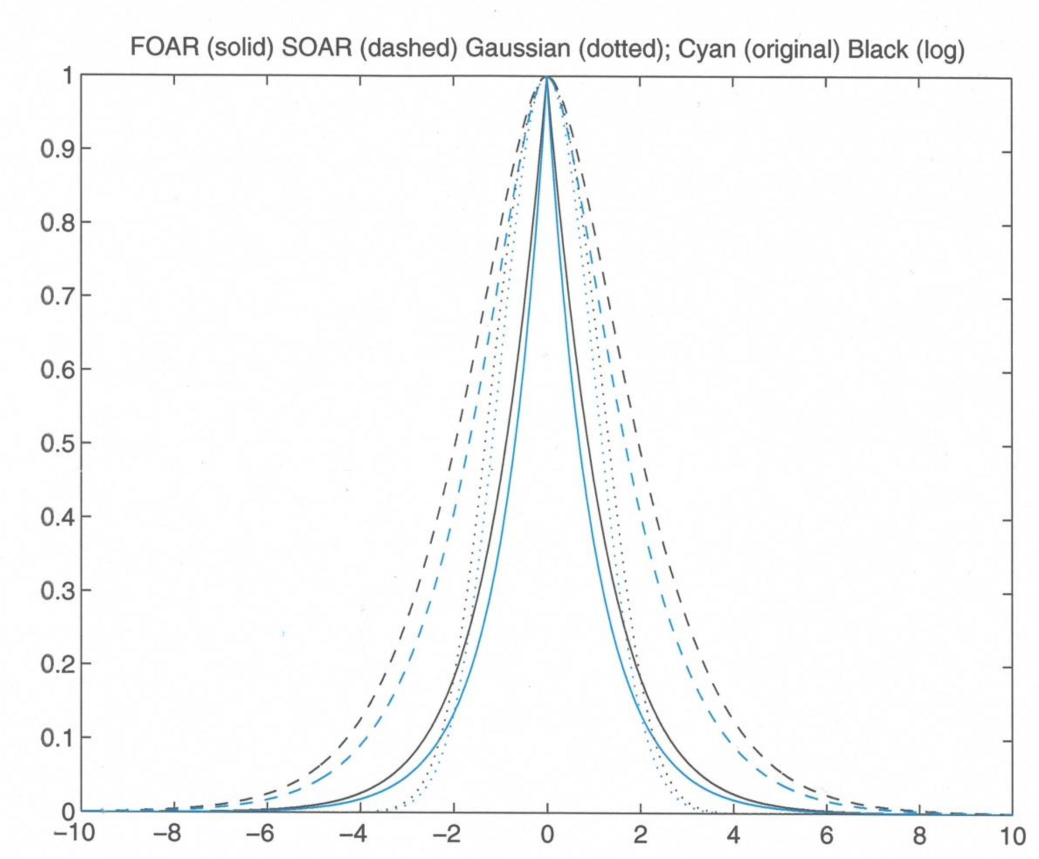
$$\langle c_i \rangle = \exp\left[\langle w_i \rangle + \frac{1}{2}\overline{B}_{ii}\right]$$

$$P_{ij} = \langle c_i \rangle \langle c_j \rangle \left[\exp(\overline{B}_{ij}) - 1\right]$$

Inverse transformation $\overline{B}_{ij} = \log\left(1+\frac{P_{ij}^f}{\langle c_i \rangle \langle c_j \rangle}\right)$ and with $P_{ij} = \sigma_i^f \sigma_j^f C_{ij}^f$

and a *relative standard deviation* $\delta = \frac{\sigma}{\langle c \rangle}$ then

$$\overline{B}_{ij} = \log\left(1+\delta_i^f \delta_j^f C_{ij}^f\right) \approx \delta_i^f \delta_j^f C_{ij}^f \quad \text{for} \quad \delta \le 0.1$$



Data assimilation with a relative error formulation

If the state-dependent observation error is dependent on the forecast instead of the truth, i.e.

$$\boldsymbol{\mu}_k^o = \mathbf{H}_k \boldsymbol{\mu}_k^t + \mathbf{g}_k(\mathbf{H}_k \boldsymbol{\mu}_k^f) \circ \boldsymbol{\epsilon}_k.$$

and that the state-dependent model error is dependent on the analysis instead of the truth

$$\boldsymbol{\mu}_{k+1}^t = \mathbf{M}_k \boldsymbol{\mu}_k^t + \mathbf{f}_k(\boldsymbol{\mu}_k^a) \circ \boldsymbol{\epsilon}_k^q,$$

Then the Bayesian update, and KF propagation using the proper conditional expectation can be derived and give the standard KF with

$$\mathbf{R}_k \circ [\mathbf{g}_k(\mathbf{H}_k \boldsymbol{\mu}_k^f) \mathbf{g}_k^{\mathrm{T}}(\mathbf{H}_k \boldsymbol{\mu}_k^f)] \equiv \mathbf{R}_k^o.$$

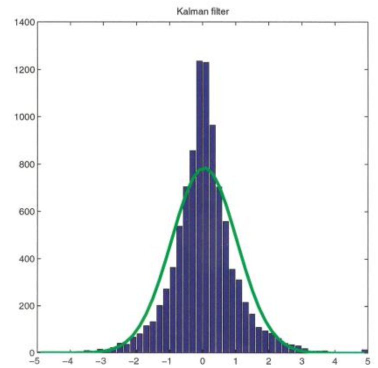
$$\mathbf{Q}_k \equiv \mathbf{f}_k(\boldsymbol{\mu}_k^a) \mathbf{f}_k^{\mathrm{T}}(\boldsymbol{\mu}_k^a) \circ \mathbf{C}_k^q.$$

Ménard et al., 2000. MWR

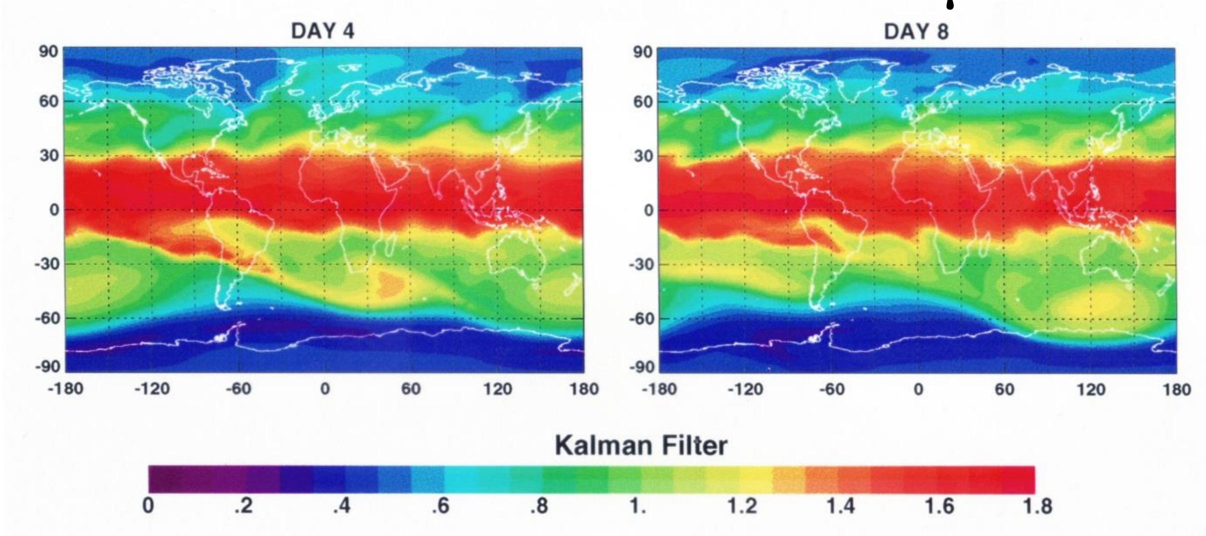
4.4 Lognormal KF

$$\boldsymbol{\varepsilon}^o = \boldsymbol{\mu}^f \circ \tilde{\varepsilon}^o$$

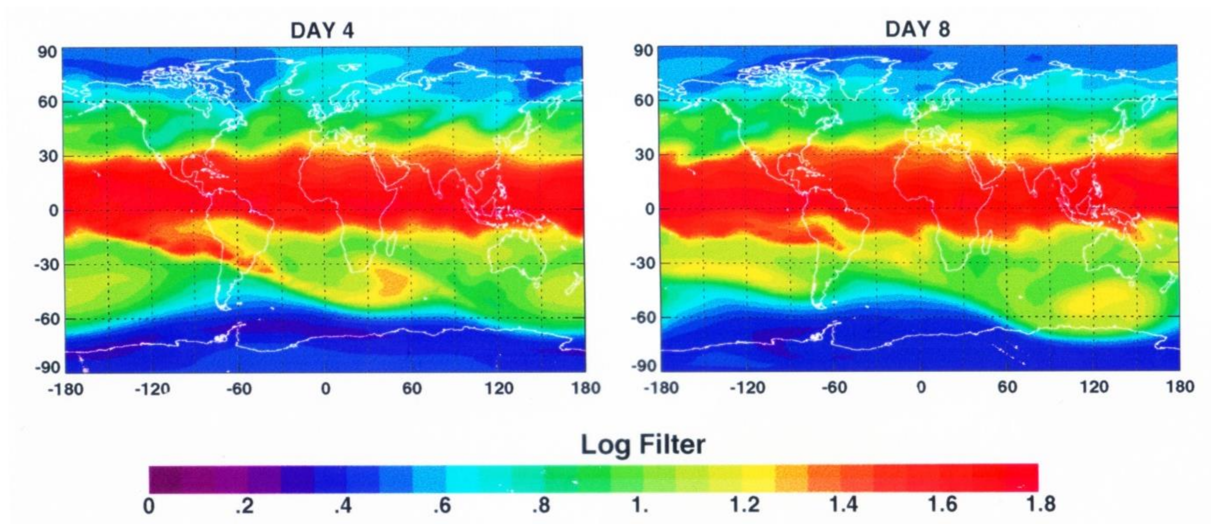
$$\boldsymbol{\varepsilon}^q = \boldsymbol{\mu}^a \circ \tilde{\varepsilon}^q$$



- KF relative error formulation



- KF lognormal formulation

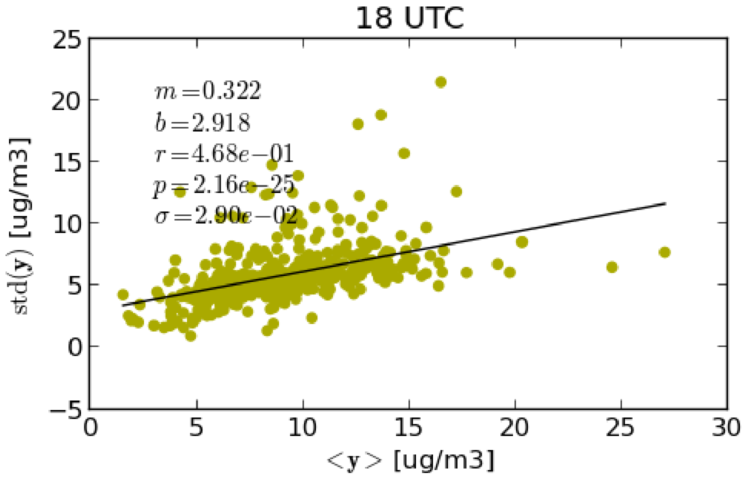


Lognormal distributions: $\sigma_c^2 = \langle c \rangle^2 \exp(\sigma_w^2 - 1)$

Observation errors is not lognormally distributed

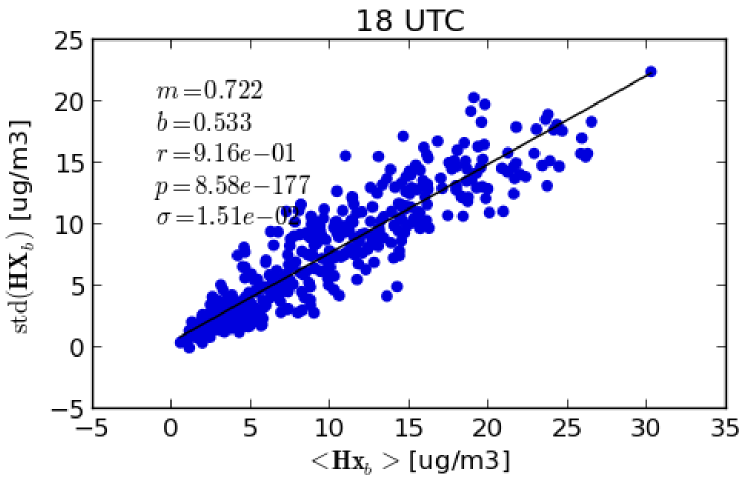
Observations at low concentrations can have negative values, because there are detection limits, or because of the retrieval

observations



$$\sigma_c = \langle c \rangle \cdot cte + b$$

model values at observation sites



$$\sigma_c = \langle c \rangle \cdot cte$$

Lognormal or something else ?

Process of random dilution

n is the number density (number of molecules per unit volume), v is the volume, and q is total number of molecules

$$q = nv$$

If all these molecules found themselves into a larger volume V, then the new number density is

$$q = NV$$

thus

$$N = n\left(\frac{v^{\alpha_i}}{V}\right) = n\alpha$$

where α is a dilution factor $\boxed{0 < \alpha < 1}$

And after k dilutions,

$$n_k = n_0 \prod_{i=1}^{k} \alpha_i \quad \text{or} \quad \ln n_k = \ln n_0 + \sum_{i=1}^{k} \ln \alpha_i$$

If $\alpha_i \sim U(0,1)$ then $-\log \alpha_i \sim EXP(1)$ and $-\sum_{i=1}^{K} \log \alpha_i \sim Gamma(k,1)$

so $n_k \sim$ *Log-Gamma* distribution

Thanks