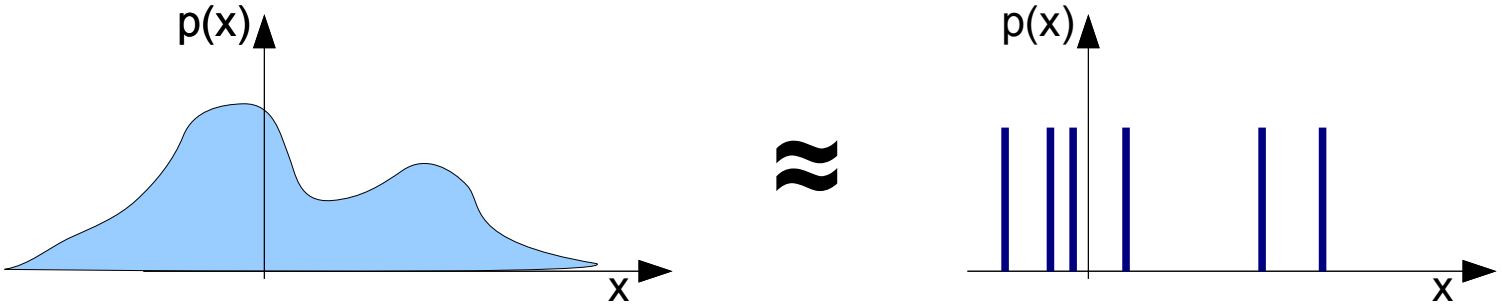
Assimilation de données

méthodes stochastiques

stochastique = qui utilise des nombres aléatoires pour modéliser des incertitudes.

- Tirages aléatoires, test statistique
- **Prévision d'ensemble**
  - technique
  - applications
  - validation
- Assimilation par filtre de Kalman; KF d'ensemble
- Filtres particulaires

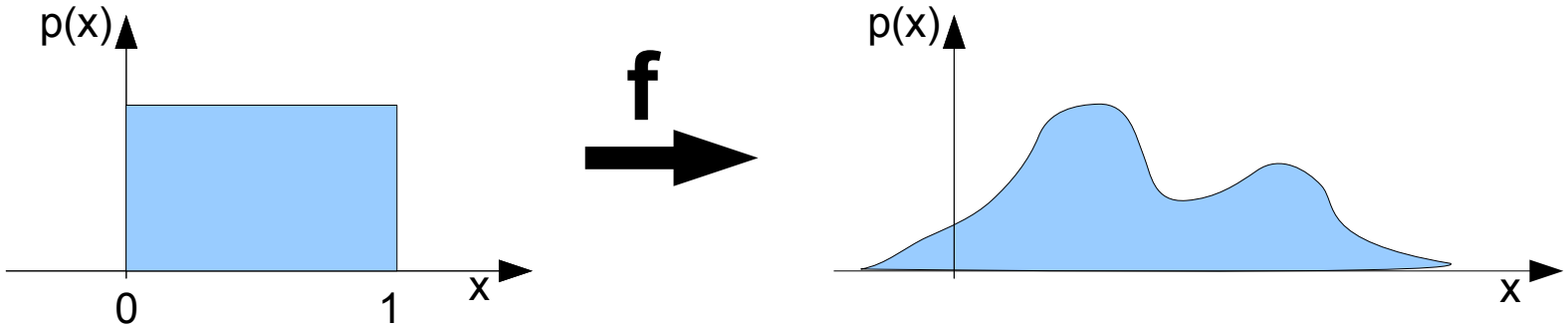
Tirages aléatoires 'Monte Carlo'



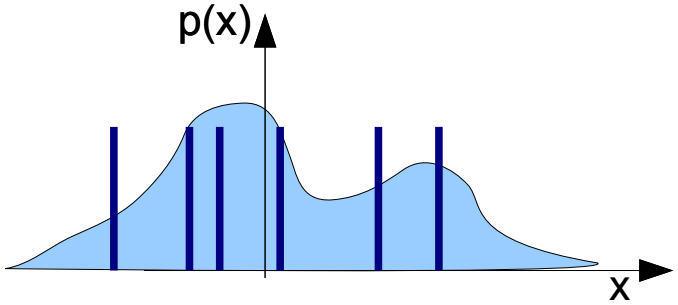
But: représenter une loi de probabilité avec un échantillonnage discret.

Technique:

- 1 générateur de nombre aléatoires: rand(n_{seed}) produit une séquence à distribution uniforme dans [0,1] déterminée par n_{seed}
- transformer $a_i \rightarrow b_i = f(a_i)$ tels que (b_i) ait la distribution souhaitée



représentation de probabilités par des échantillons discrets



Fonction p représentée par un échantillon discret {xi} :

- càd mathématiquement : $p(x) \sim 1/N \sum \delta(x-x_i)$ où δ=fonction de Dirac
- moyenne : $E(p) = \int x\, p(x)\, dx \sim \bar{x}_i = 1/N \sum x_i$

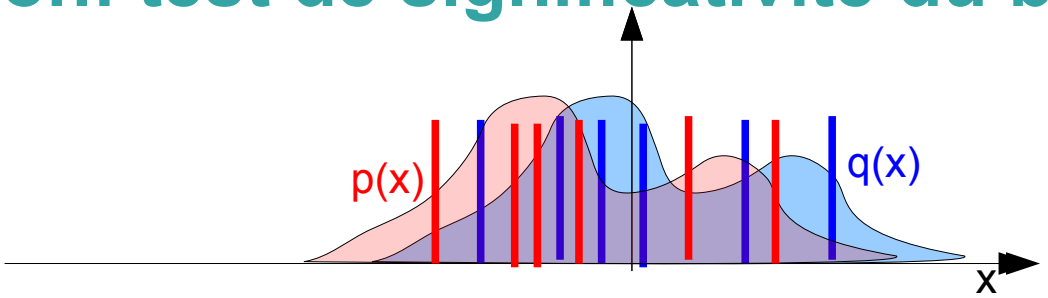
C'est comme une intégration numérique de x p(x), mais avec des points disposés aléatoirement:

- le bruit d'échantillonnage sur les covariances a un écart-type de l'ordre de $1/\sqrt{N}$

Méthode généralisable à un échantillon non-équiprobable:

$p(x) \sim 1/N \sum w_i\, \delta(x-x_i)$ où w_i = probabilité de chaque x_i (avec $\sum w_i=1$)

Application: test de significativité du bootstrap



Exemple: est-ce que $\overline{p} < \overline{q}$ dans la figure ci-dessus ? Ou effet d'échantillon ?

But: Dans un jeu de données x sur lequel on a calculé une statistique S(x), tester si une hypothèse sur S est vraie. (S peut être une moyenne, une corrélation, etc...)

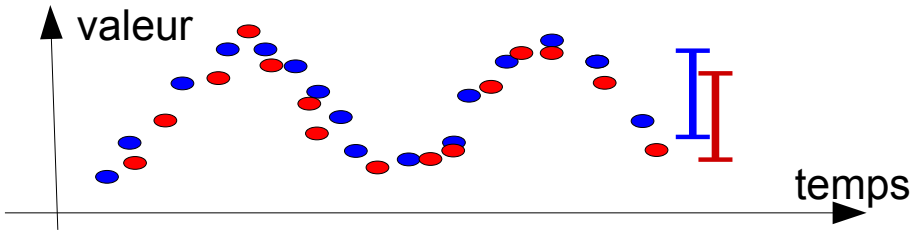
Technique:

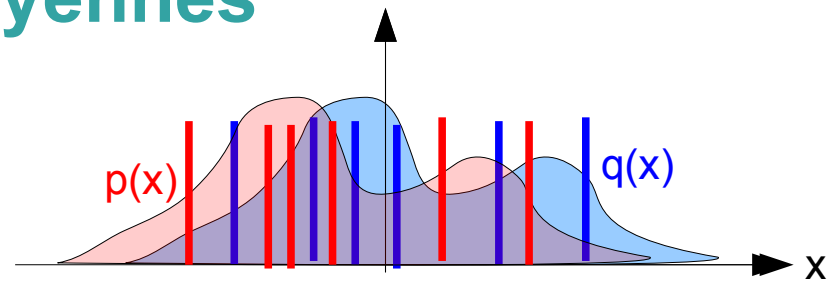
- à partir de (x), fabriquer p jeux de données perturbés (x_k) de même dimension, tirés aléatoirement parmi les éléments de (x). On peut tirer plusieurs fois le même x_k.
- alors l'ensemble de statistiques perturbées $S_k = S(x_k)$ suit approximativement la loi de probabilité de S
- (strictement vrai uniquement si les éléments de x sont indépendants et suivent la même loi de probabilité)

Application:

- on peut mettre une barre d'erreur sur S(x) en regardant les quantiles 5% et 95% de l'ensemble des S_k

Significativité de la différence entre 2 moyennes





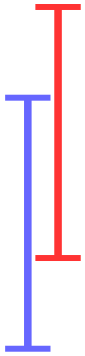
Problème: on a 2 jeux de données qui mesurent un même paramètre: le jeu rouge, et le jeu bleu. Leurs moyennes sont-elles différentes ?

Exemples:

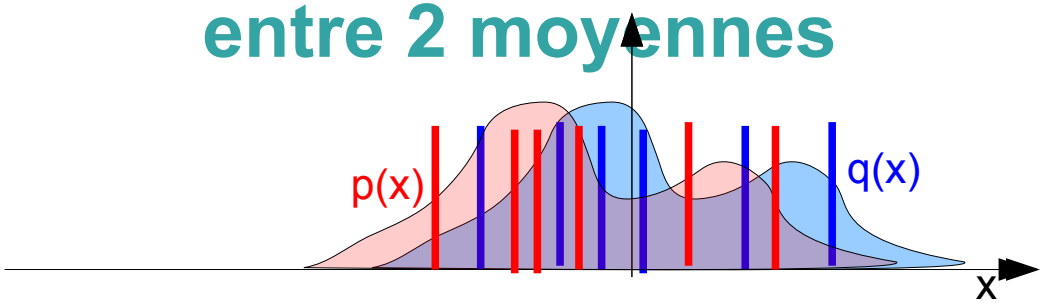
- Températures en 1900-1950 et en 1960-2010: le climat se réchauffe-t-il ?
- Mesures d'un paramètre océanique regroupées suivant 2 phases différentes d'El Nino: El Nino a-t-il un impact significatif ?

Technique naïve:

- estimer une barre d'erreur sur la moyenne rouge. (par ex. par bootstrap)
- idem sur les données bleues
- (fausse) conclusion: les moyennes rouge et bleue ne sont significativement différentes que si les barres d'erreur sont séparées.



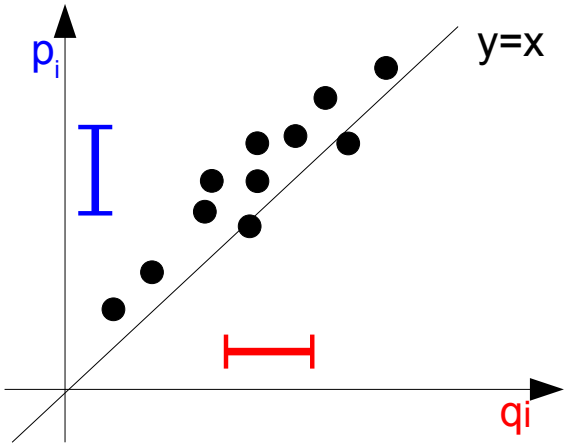
Significativité de la différence entre 2 moyennes

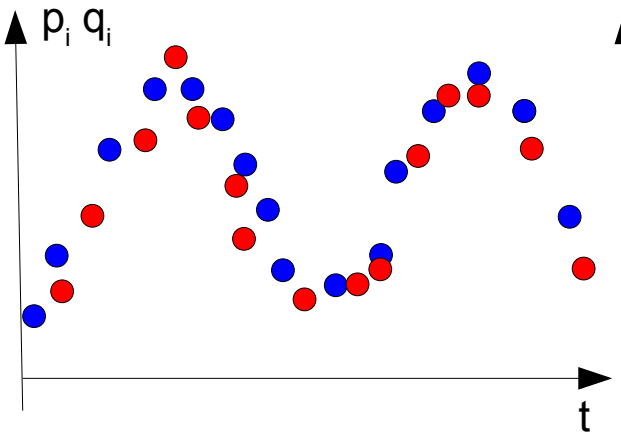


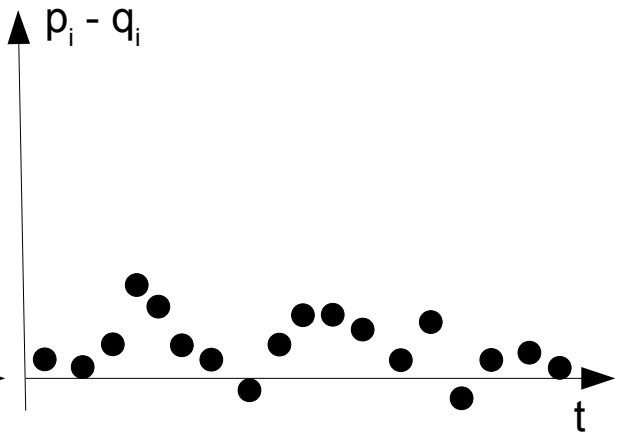
La bonne méthode: faire le test sur question posée, pas sur autre chose

- estimer une barre d'erreur les *différences* entre données rouges et données bleues. Cela donne une barre d'erreur sur la *moyenne* m(rouge-bleu).
- si la barre d'erreur n'intercepte pas la valeur zéro, alors la différence des moyennes est significative.

Penser à visualiser les données:







Simulation stochastique de covariances

Idée: *et si au lieu de modéliser des covariances on les remplaçait par un ensemble de x_i dont les covariances ont les propriétés voulues ?*

But: fabriquer un ensemble (x_k) de covariance cov(x,x)=M (ex: M=B)

Technique:

- fabriquer des vecteurs (v_k) dont les composantes suivent la loi normale $\mathcal{N}(0,1)$: $\mathrm{cov}(v,v)=m(vv^T)=I$
- factoriser $M=NN^T$ (avec une lib scientifique, ex. méthode Cholesky)
- alors les vecteurs ($x_k = Nv_k$) vérifient:

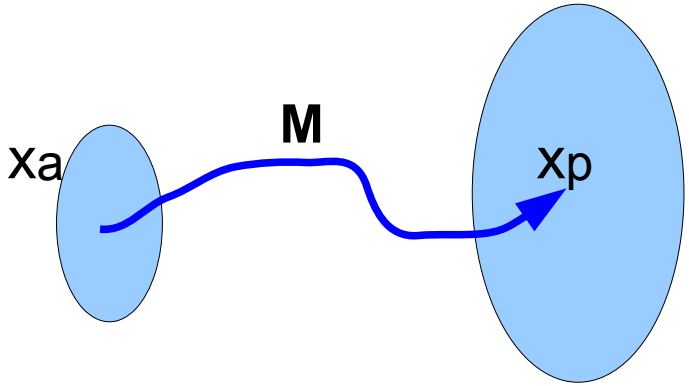
$$\mathrm{cov}(x,x)=\mathrm{moy}(xx^T) = \mathrm{moy}(Nvv^TN^T) = N\ \mathrm{moy}(vv^T)N^T = NIN^T = M$$

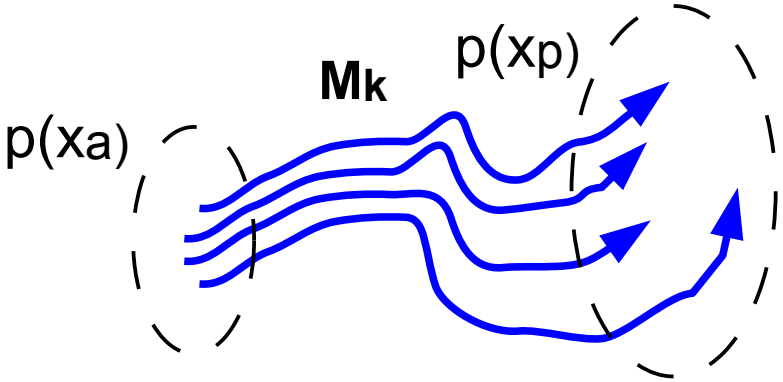
Prévision d'ensemble (ou: prévision stochastique)

But: tirer un ensemble de prévisions numériques x_{pk} qui approxime la loi de probabilité d'une prévision *déterministe* $x_p=M(x_a)$

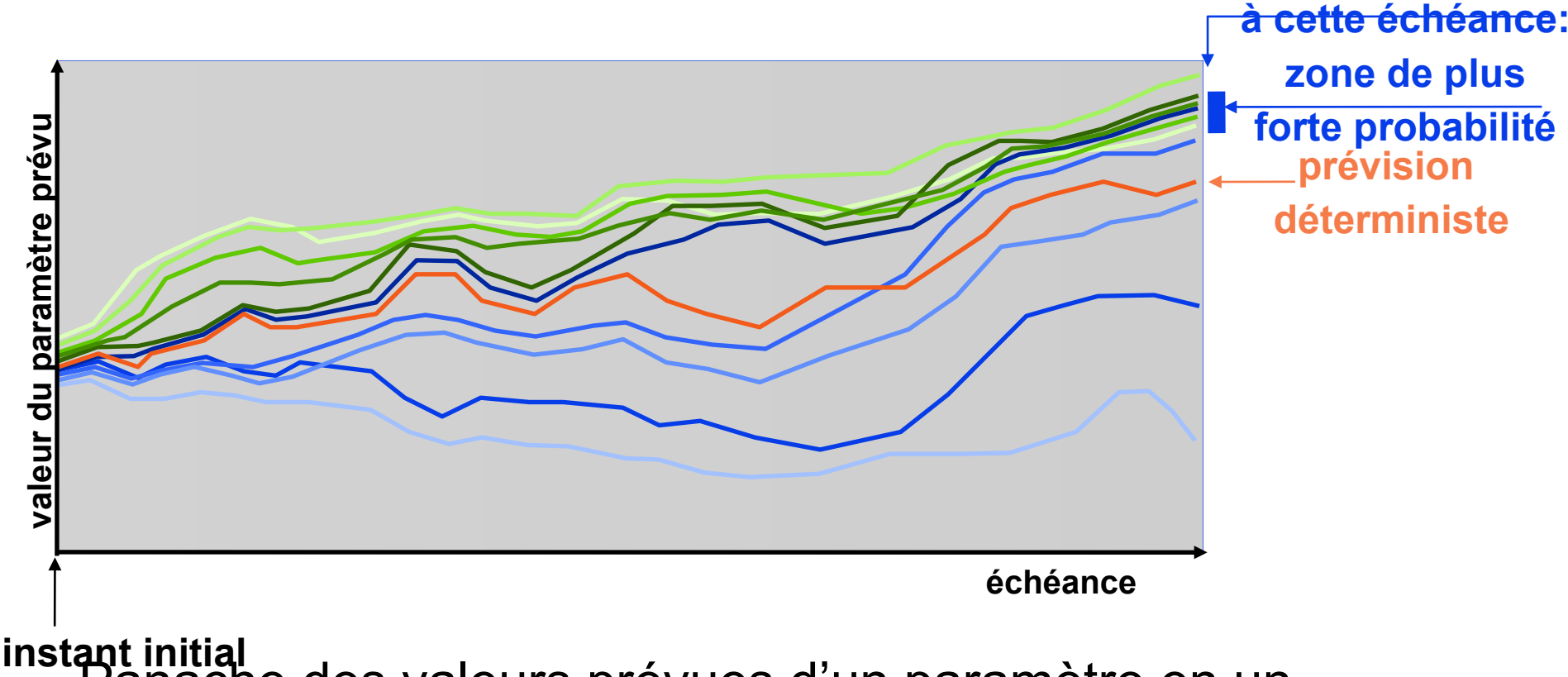
Technique:

- tirer k états initiaux x_{ak} qui suivent la loi de probabilité de x_a
- perturber k fois les équations du modèle numérique, d'une manière qui représente l'erreur de modélisation (erreur de discrétisation, erreurs des tendances physiques, etc)
- l'ensemble est fourni par k prévisions perturbées: $x_{pk}=M_k(x_{ak})$



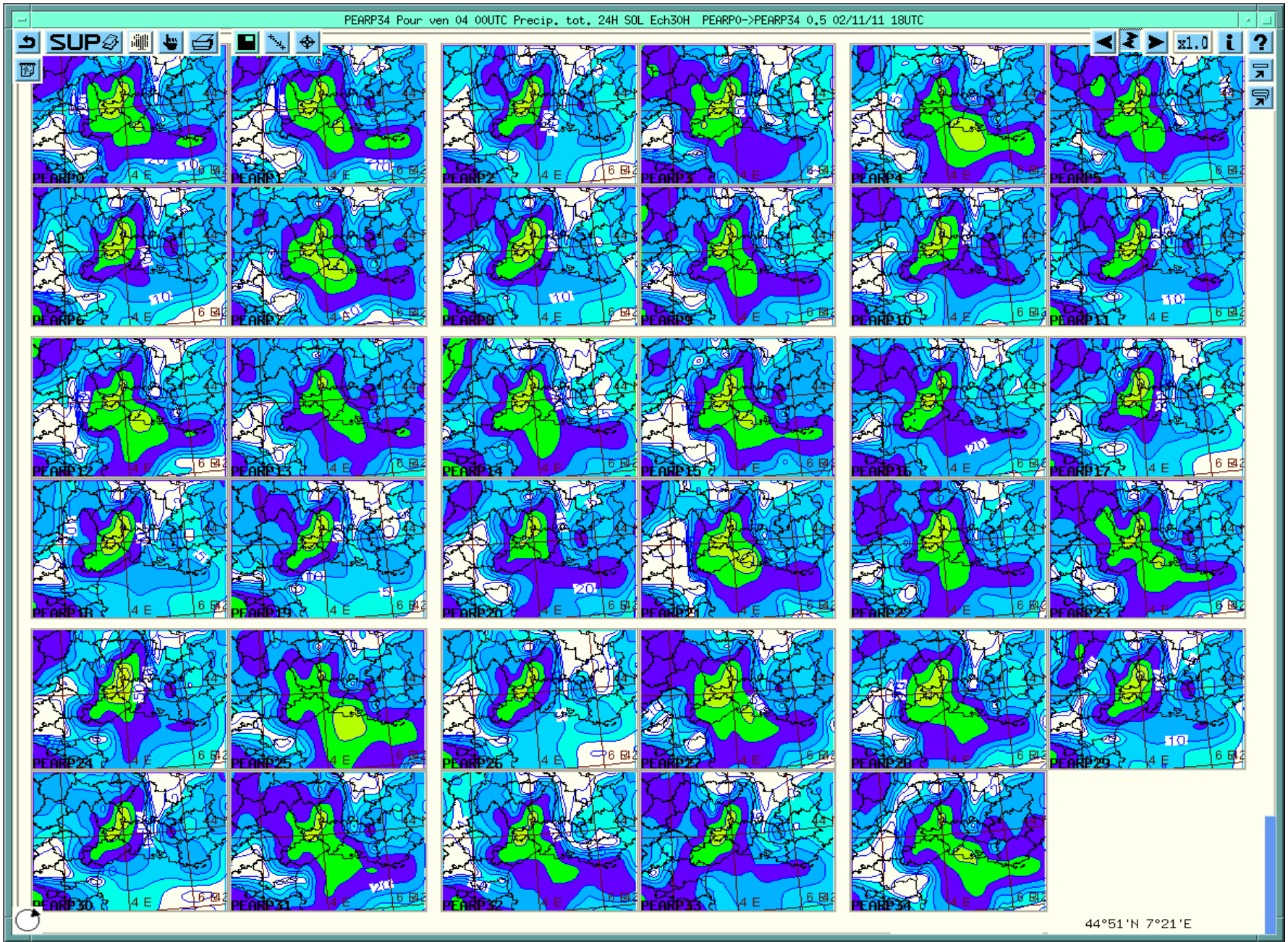


Visualisation de l'incertitude d'une prévision: le panache d'un ensemble $M(x_k)$



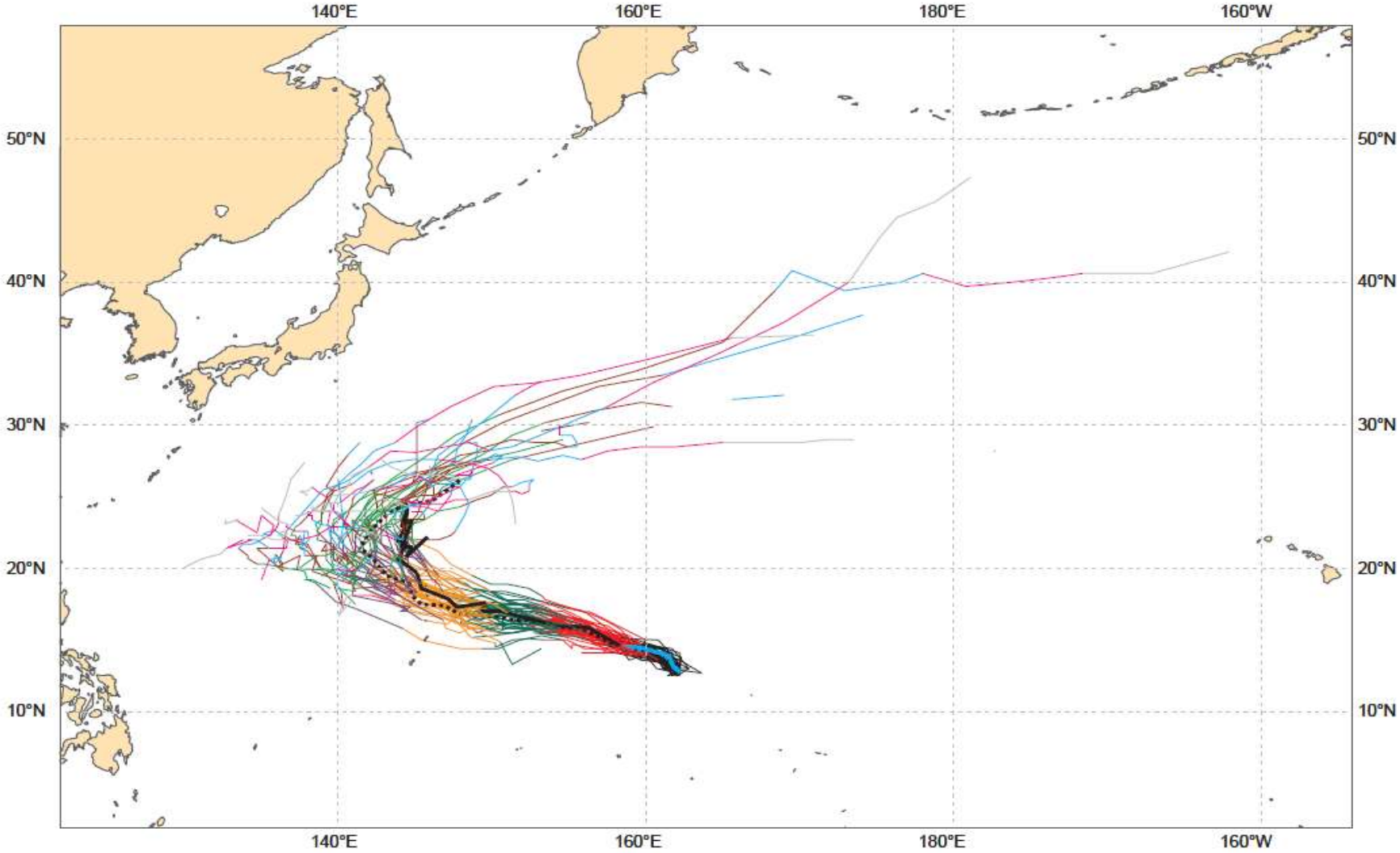
Panache des valeurs prévues d'un paramètre en un point pour différentes échéances

Prévision d'ensemble (pluies en 24h sur SE France)



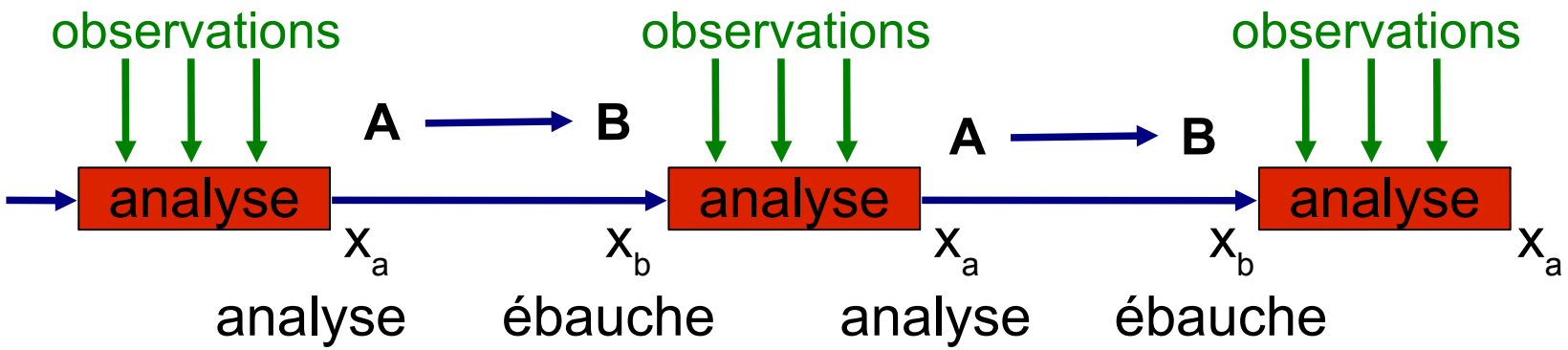
prévision cyclonique par un ensemble

Date 20151013 00 UTC @**ECMWF**
Individual trajectories for **25W** during the next **240** hours
tracks: **thick solid**=HRES; **thick dot**=CTRL; **thin solid**=EPS members [coloured]
0-24h **24-48h** **48-72h** **72-96h** **96-120h** **120-144h** **144-168h** **168-192h** **192-216h** 216-240h



Prévision d'ensemble: applications

- on peut calculer des **probabilités de dépassement de seuil** (ex: pluies > 100mm en 6h → risque d'inondation)
- les prévisions se dispersent d'autant plus vite que le fluide est instable: la dispersion des prévisions indique le **degré de confiance** dans la prévision
- l'ensemble fournit des échantillons rigoureux pour évaluer le B du BLUE: c'est l'**assimilation d'ensemble** ou le **filtre de Kalman**



Prévision d'ensemble: techniques de perturbation

Etat initial:

- 'breeding': utiliser les différences entre prévisions produites hier
- vecteur singuliers (singular vectors): calculer les n perturbations de l'analyse dont l'énergie va croître le plus vite (ce sont les plus grands vecteurs propres de MAM^T, calculables par méthode de Lanczos)
- assimilation d'ensemble (cf. plus loin)
- filtre de Kalman d'ensemble (cf. plus loin)

Modèle:

- perturber les paramètres du modèle
- utiliser plusieurs physiques
- injecter du bruit dans l'équation d'évolution du modèle (physique stochastique)

Validation des ensembles

Comment mesurer la qualité d'une prévision de probabilités ?

3 concepts:

- **fiabilité**
- **résolution**
- **utilité**

Validation des ensembles

Objectif d'une prévision d'ensemble: la PDF de l'ensemble doit être indiscernable de la PDF des observations (on veut 'simuler les erreurs')

Mais on ne peut pas observer la PDF des observations, uniquement des réalisations (les "erreurs de prévision"). Donc on utilise des critères de qualité indirects et partiels:

- **fiabilité**: les statistiques des *erreurs de prévision* et des *membres de l'ensemble* sont-elles indiscernables ?
- **résolution:** les PDF prévues sont-elles différentes de la climatologie ? (probabilités proches de 0 ou de 1)
- **utilité*:*** l'ensemble permet-il de prendre de meilleures décisions qu'un système de prévision moins coûteux ?

Ces performances dépendent:

- de la prévisibilité du système (limite naturelle)
- de la qualité de l'état initial et du modèle (limite de modélisation)
- de la qualité des perturbations (cohérence entre PDF de l'ensemble et PDF réelle)
- des erreurs de mesure (erreur d'observation)

Mesure de la *fiabilité* d'un ensemble

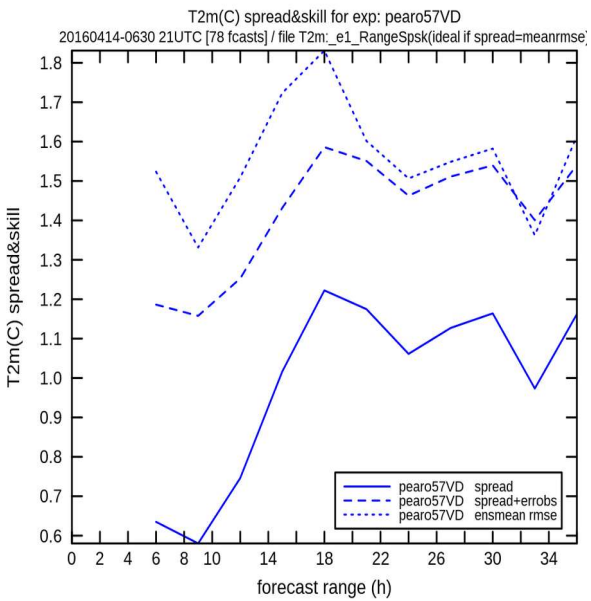
"PDFs des prévisions et des observations doivent avoir les mêmes statistiques"

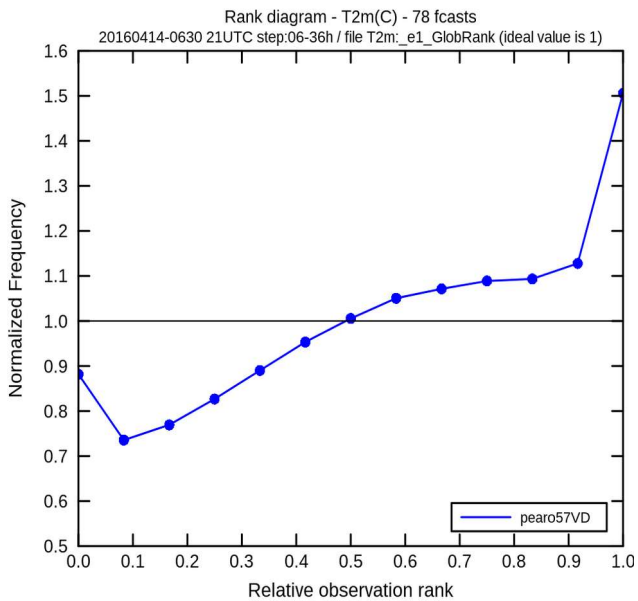
- condition nécessaire mais non suffisante d'utilité de l'ensemble (par ex., une climatologie est fiable mais inutile)
- se mesure en moyenne sur un grand nombre de points de mesure et de situations

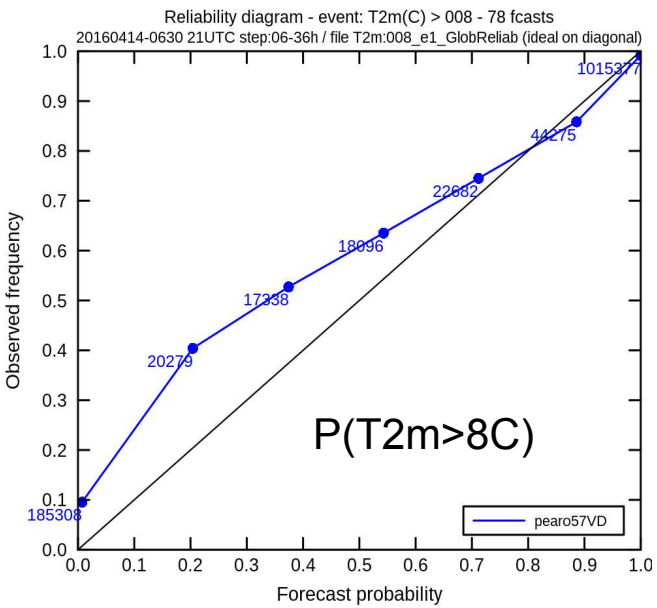
spread/skill ratio:
cohérence entre dispersion et erreur de prévision
(pour paramètre scalaire)

rank diagram:
cohérence entre rangs des obs et des prévisions
(pour paramètre scalaire)

reliability diagram:
cohérence entre probabilités prévues et fréquences observées
(pour paramètre binaire)



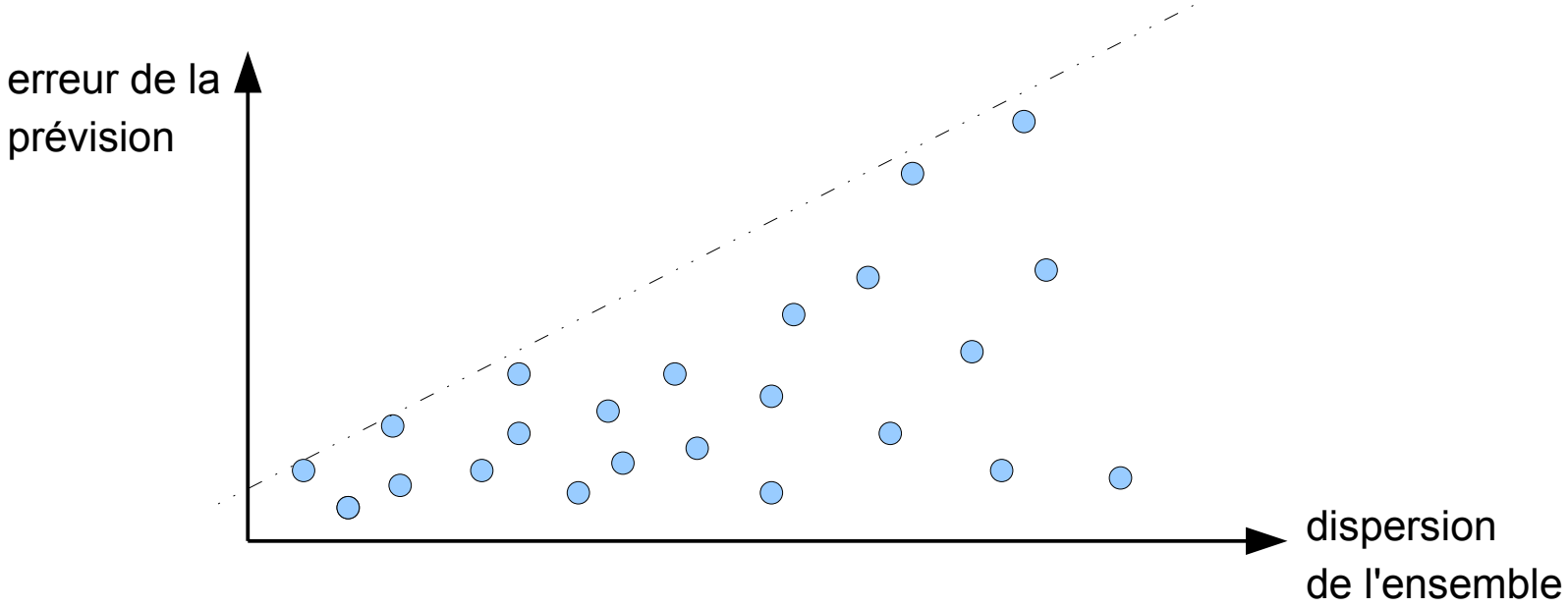




fiabilité de la dispersion d‘un ensemble

relation entre dispersion et erreur de prévision:

- la dispersion (=variance) de l'ensemble doit être plus grande lorsque l'erreur de prévision est plus grande (que la moyenne)
- mais l'inverse n'est pas vrai: on peut avoir une grande variance avec une réalisation proche de la moyenne.
- var(x_{pk}) doit être égale à var(y - moy(x_{pk})) aux erreurs d’obs près



Mesures de la *résolution* d'un ensemble

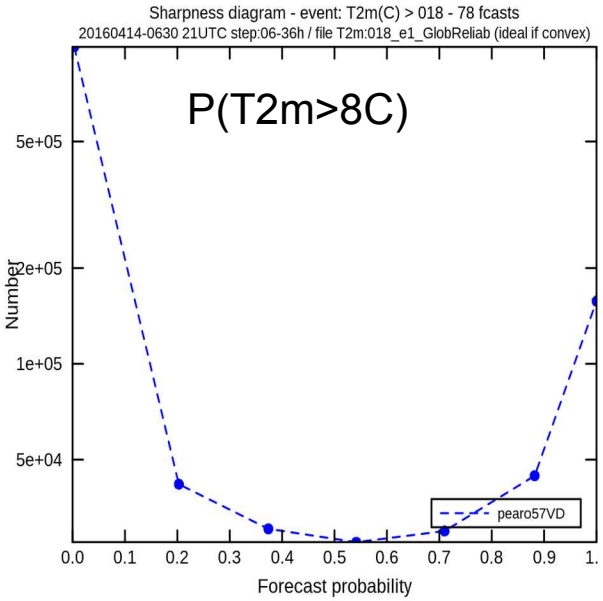
''les PDFs des prévisions doivent être différentes de la climatologie''

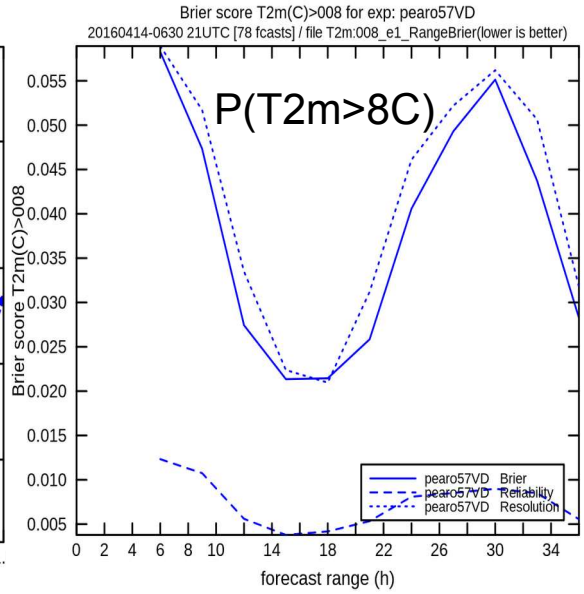
- condition nécessaire mais non suffisante (ne dit pas si on avait prévu juste)
- on préfère mesurer la *justesse* des PDFs prévues (mélange fiabilité et résolution)

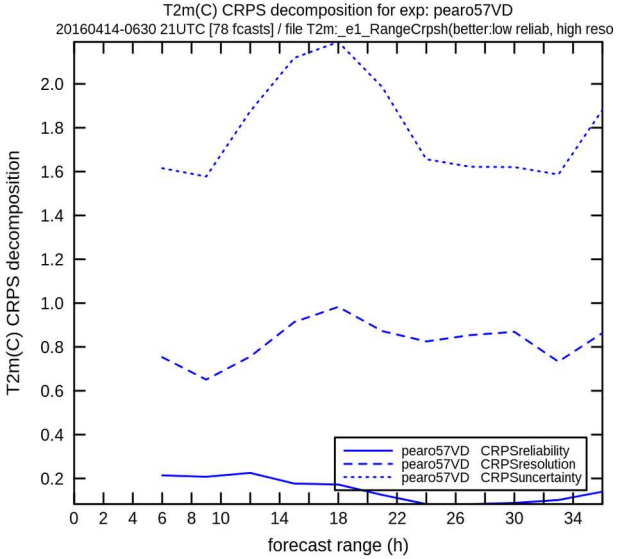
Sharpness diagram: fréquences des probabilités prévues (pour paramètre binaire)

Brier score: distance quadratique entre PDF prévue et obs (pour paramètre binaire)

CRPS score: distance quadratique entre PDF prévue et obs (pour paramètre réel)







Mesures de l'*utilité* d'un ensemble

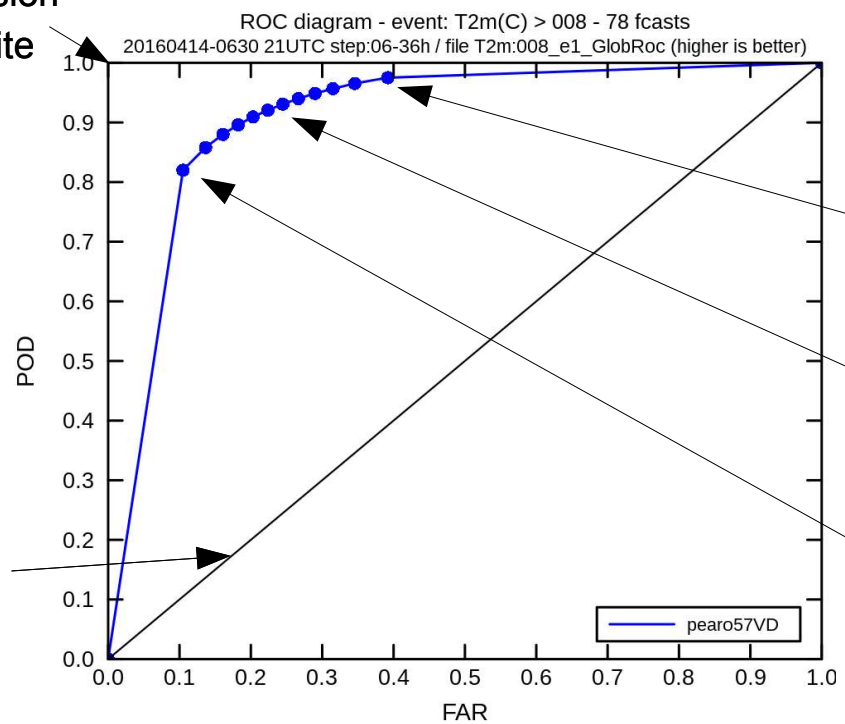
''l'ensemble doit permettre de faire de bonnes prévisions''

- définir l'''évènement'' qu'on prévoit sous forme binaire, ex: ''T2m>8C ou non''
- sur un grand nombre de mesures on regarde 2 types de mauvaises prévisions:
  - les *non-détections* (évènement observé mais non prévu), taux=POD
  - les *fausses alarmes* (évènement prévu mais non observé), taux=FAR

diagramme ROC: taux de succès de **différentes méthodes de prévision** utilisant le même ensemble

ROC: relative operating characteristic
POD: probability of detection
FAR: false alarm rate

prévision parfaite



ligne des prévisions triviales

on prévoit T2m>8C si l'ensemble le prévoit avec probabilité>8%

on prévoit T2m avec la médiane de l'ensemble

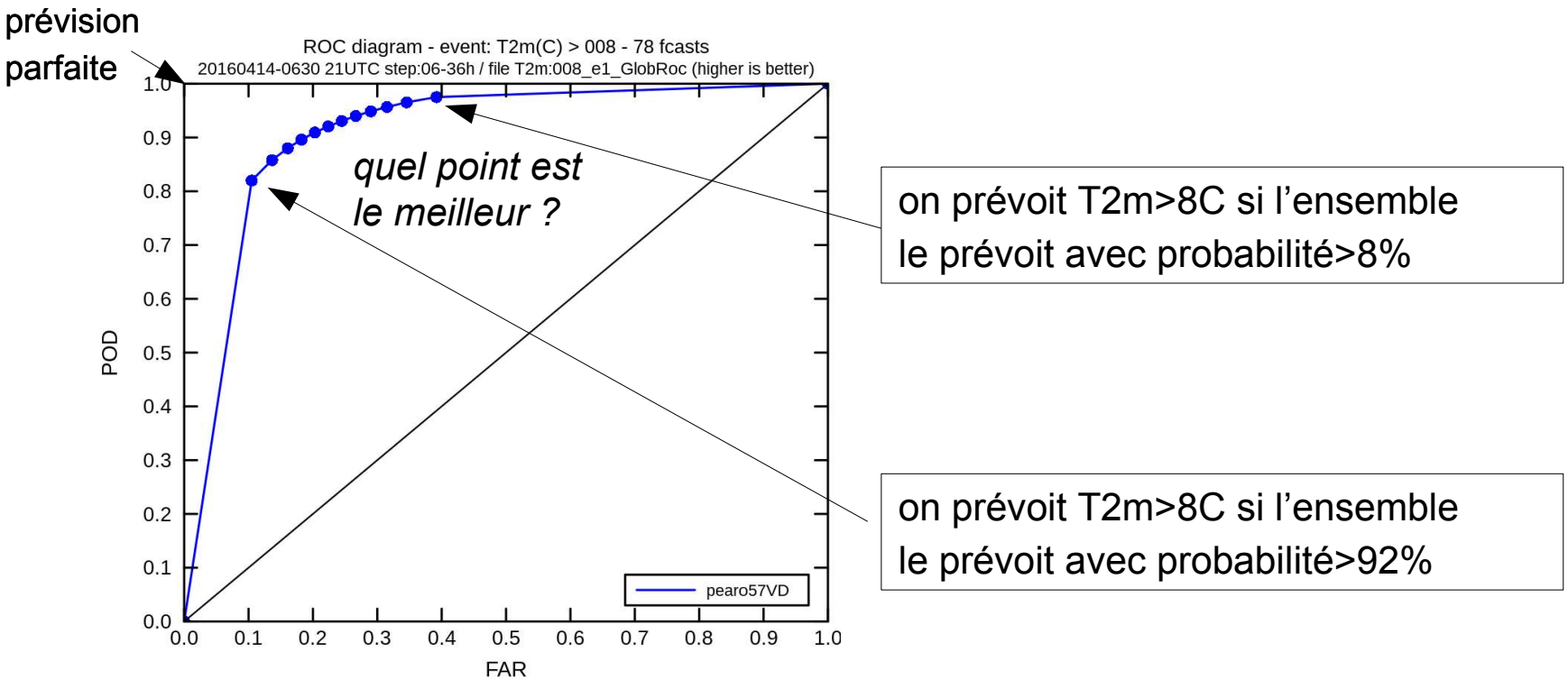
on prévoit T2m>8C si l'ensemble le prévoit avec probabilité>92%

Mesures de l'*utilité* d'un ensemble

La qualité d'un ensemble dépend beaucoup de l'importance relative donnée aux non-détections et aux fausses alarmes. Exemples:

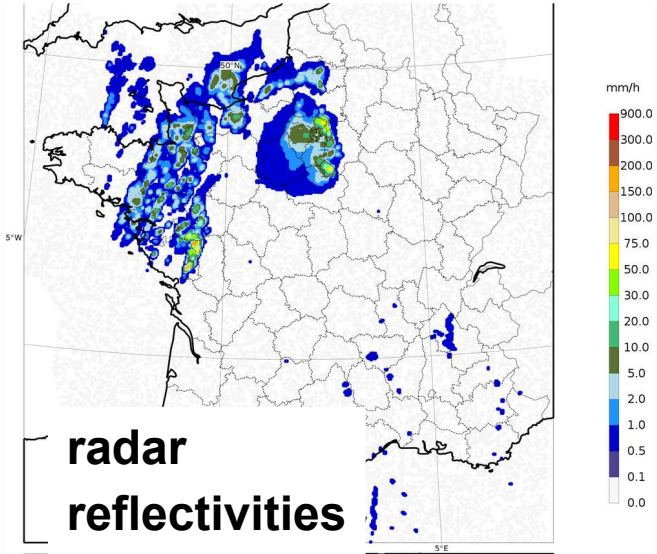
- en période de sécheresse on ne renonce à irriger que si l'on est sûr qu'il pleuvra.
- en période d'orages violents on alerte les populations même si le risque est localement faible.

Pour résumer l'utilité il faut modéliser l'utilisation qui sera faite de l'ensemble (scores de **''valeur économique''**)

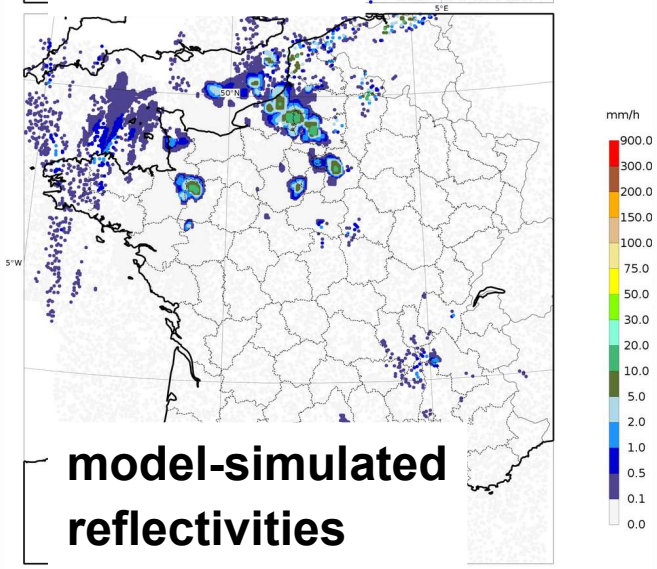


Taux de succès d'une prévision d'ensemble

case study



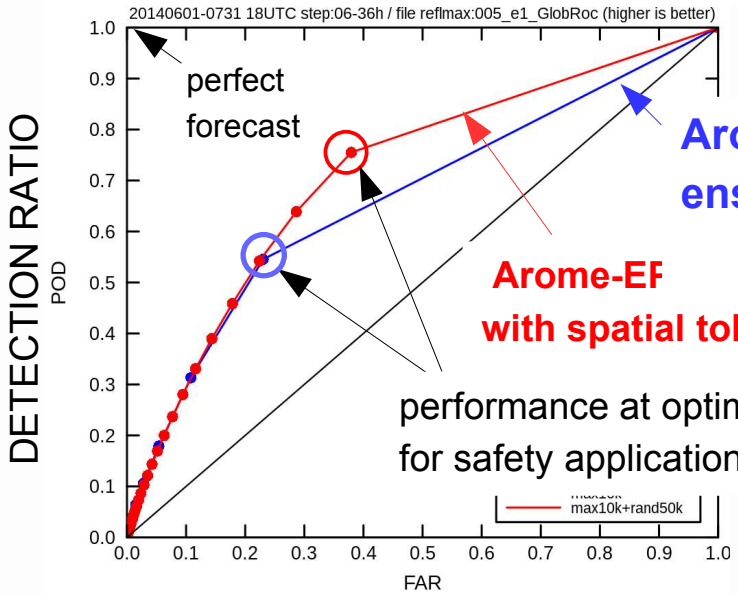
radar reflectivities



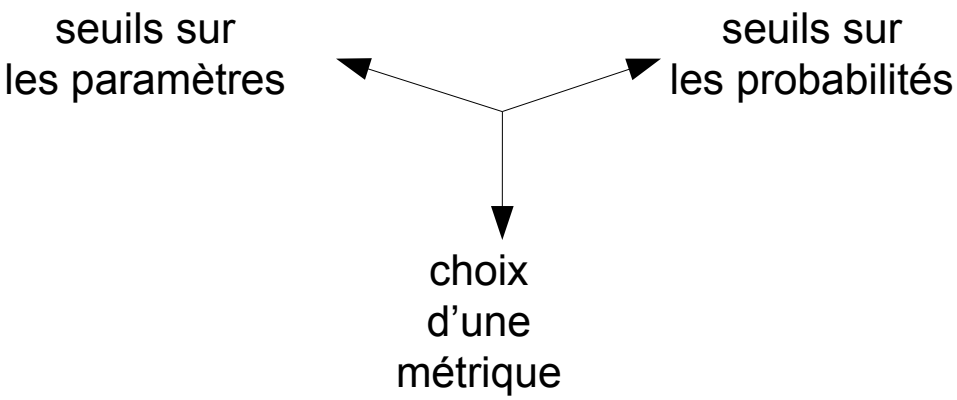
model-simulated reflectivities

Arome predicts thunderstorms, but with high uncertainties on timing and location : complex postprocessing is needed for end-user 'thunderstorm warnings'

ROC diagram (2-month average) of radar reflectivities



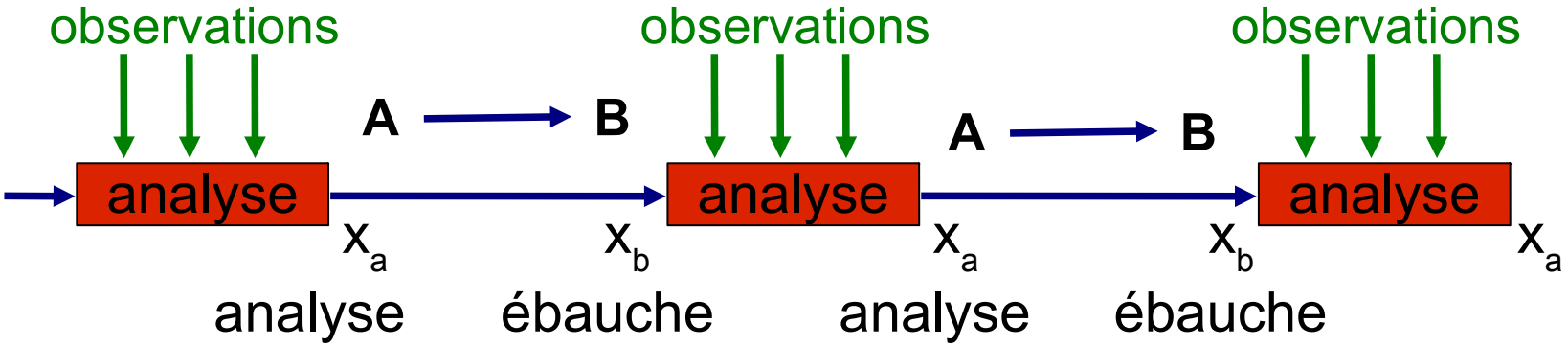
A retenir sur la validation des ensembles



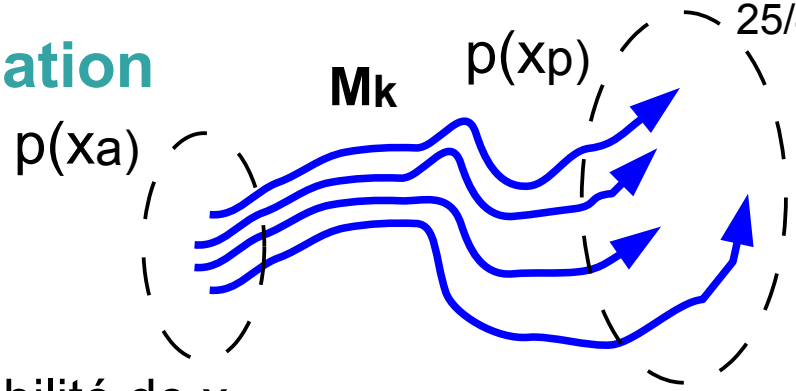
- beaucoup de scores, car beaucoup de manières de regarder un ensemble.
- certains scores mesurent les performances absolues (limitées par la qualité du système de prévision), d'autres la cohérence entre erreurs prévues et erreurs observées.
- il faut beaucoup de données pour mesurer la qualité d'un ensemble (d'autant plus pour les évènements rares...)

Retour à l'assimilation de données

- un ensemble fournit des échantillons rigoureux pour évaluer le B du BLUE: c'est l'**assimilation d'ensemble** ou le **filtre de Kalman**



Prévision d'ensemble et assimilation



Pour estimer la matrice B:

- tirer k états x_{ak} qui suivent la loi de probabilité de xa
- la PDF des prévisions $x_p=M(x_a)$ est approximée par k prévisions perturbées: $x_{pk}=M_k(x_{ak})$
- alors B ~ cov(x_{bk}) ~ cov(M (x_{ak}))→ *fournit les fonctions de structure*

Pour estimer les erreurs d'analyse:

- on tire k états x_{bk} y_k qui ont pour covariances B et R
- le BLUE produit un ensemble d'analyses: $x_{ak} = x_{bk} + K(y_k-H[x_{bk}])$
- alors l'erreur d'analyse A ~ cov(xa_k)

Ensemble d'assimilation cyclées (EDA= ensemble data assimilation):

- on combine $x_{pk}=M_k(x_{ak})$ et $x_{ak} = x_{bk} + K(y_k-H[x_{bk}])$
- ces k systèmes d'assimilation perturbés modélisent la *propagation temporelle des erreurs*

Equations du filtre de Kalman

$K = B\, H^T\, (H\, B\, H^T + R\,)^{-1}$

$x_a = x_b + K\,(y - H[x_b])$ analyse OI

$A = (I - K\,H\,)\,B$ ou $A=2(J'')^{-1}$

$x_b(t) = M_{t-1\rightarrow t}\; x_a(t-1)$ prévision

$B(t) = M'_{t-1\rightarrow t}\; A(t-1)\; M'^{T}_{t-1\rightarrow t} + Q(t)$

avec Q = covariances de "l'erreur modèle": $x_t(t) - M_{t-1\rightarrow t} x_{t-1}$

Le filtre de Kalman est statistiquement optimal si M est linéaire. On peut aussi l'utiliser avec un modèle non-linéaire (= filtre de Kalman étendu EKF)

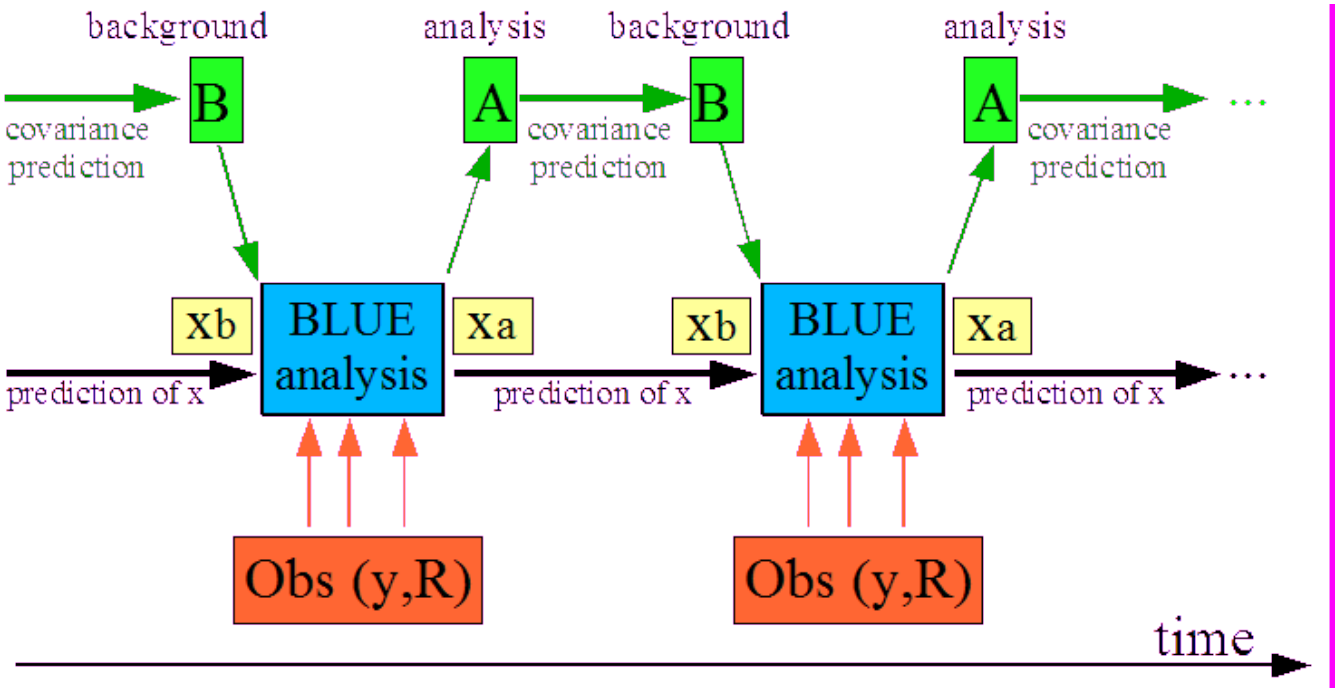
schéma du Filtre de Kalman

Le filtre de Kalman revient à compléter les équations d'analyse & de prévision de x par une récurrence sur B:

- **A** = covariances d'erreur d'analyse de l'OI (déjà vue)
- plus l'équation de **prévision des covariances d'erreur**:

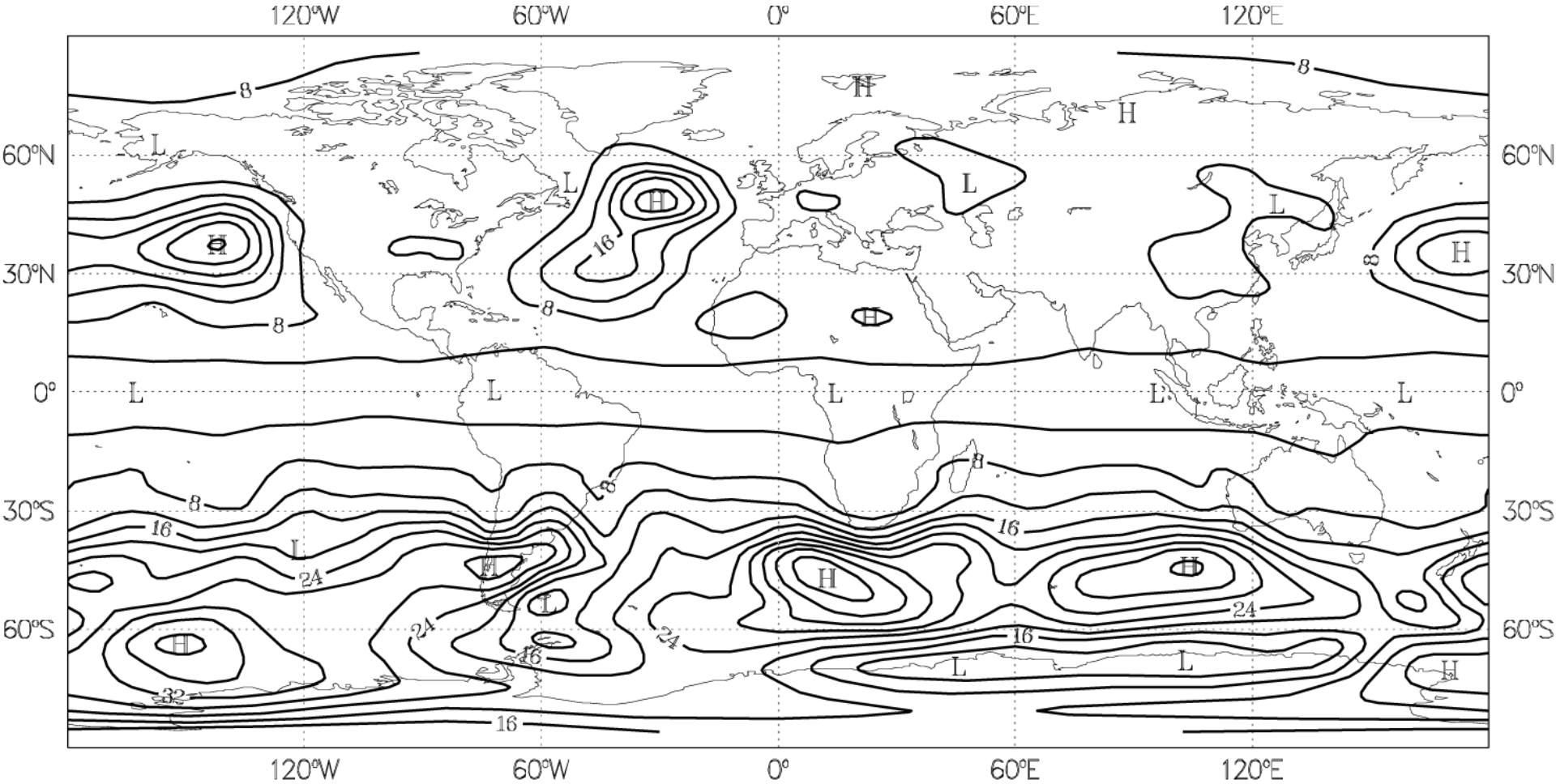
$$\mathbf{B = M\,A\,M^T + Q} \qquad \text{(Q erreur de modélisation)}$$

qui met à jour A avec le modèle **M**: **calcul physique de B**

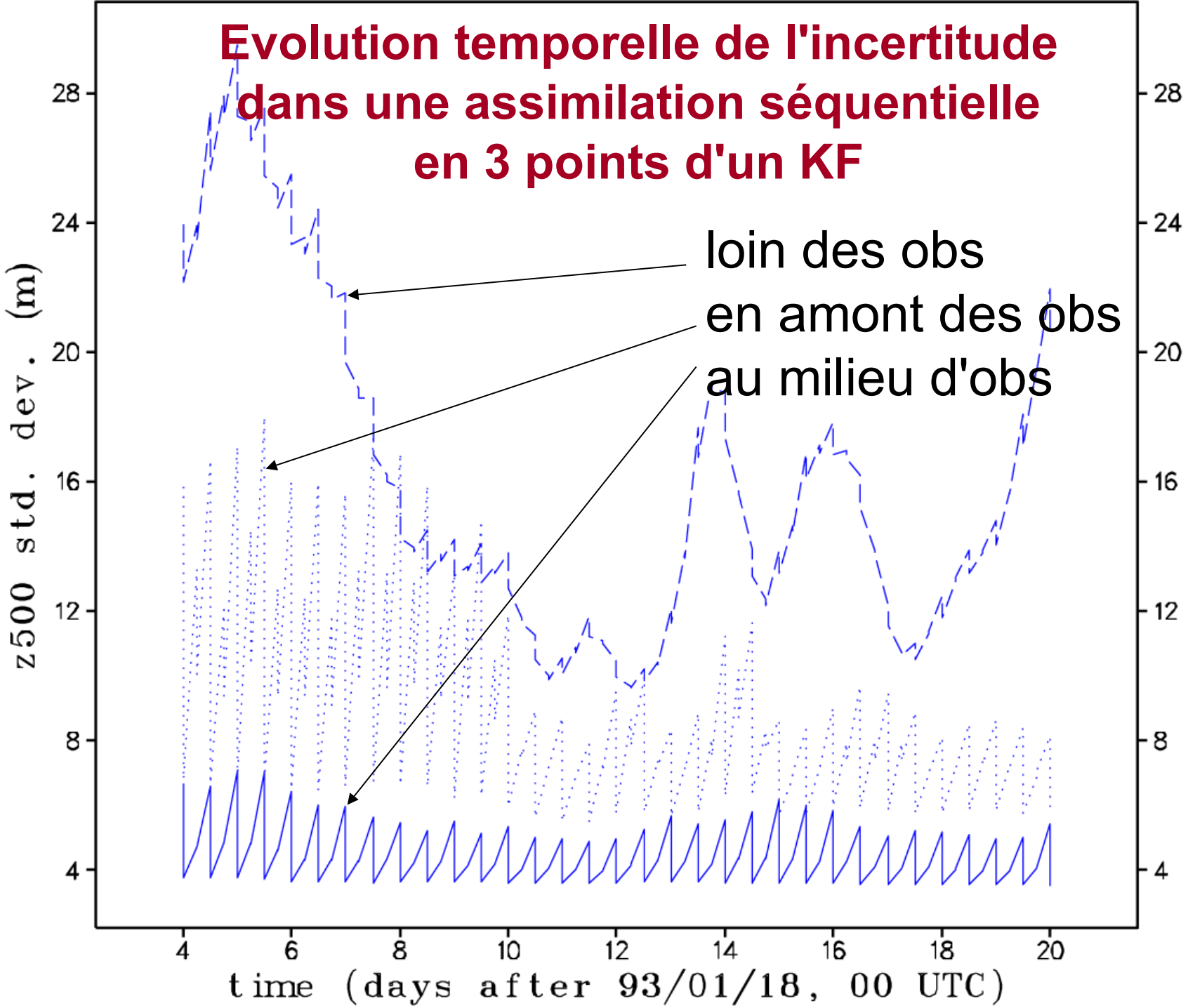


Information sur B fournie par le KF

Variances dynamiques dans B
(écart-type de z 500hPa dans l'atmosphère)



Evolution temporelle de l'incertitude dans une assimilation séquentielle en 3 points d'un KF

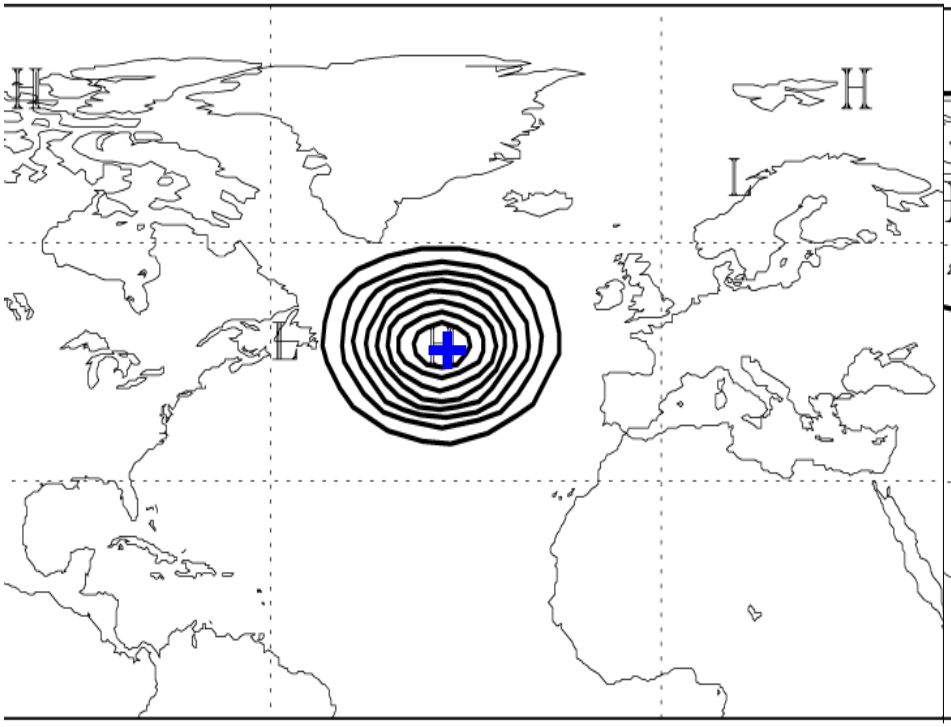


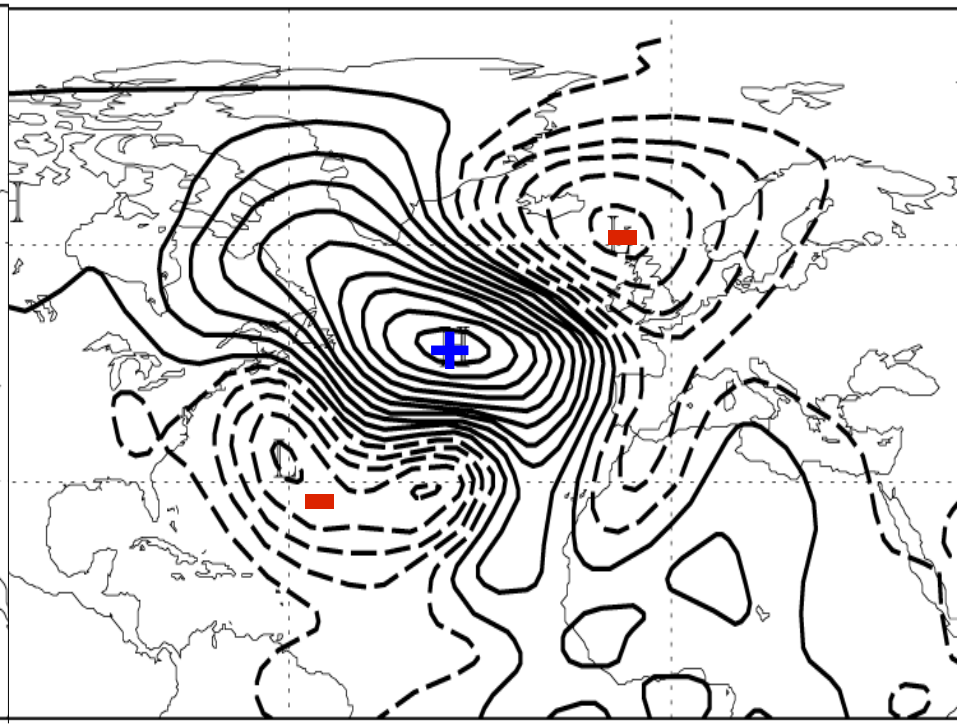
Exemples de fonctions de structure dynamiques

(théorème: 4Dvar et KF produisent le même résultat)

3DVar: B statique modélisé avec corrélations gaussiennes, structure=BH^T

filtre de Kalman et 4DVar: B produit par le modèle structure=BM^TH^T





Evolution dynamique des covariances d'erreur en filtre de Kalman

$$\mathbf{B = M\,A\,M^T + Q}$$

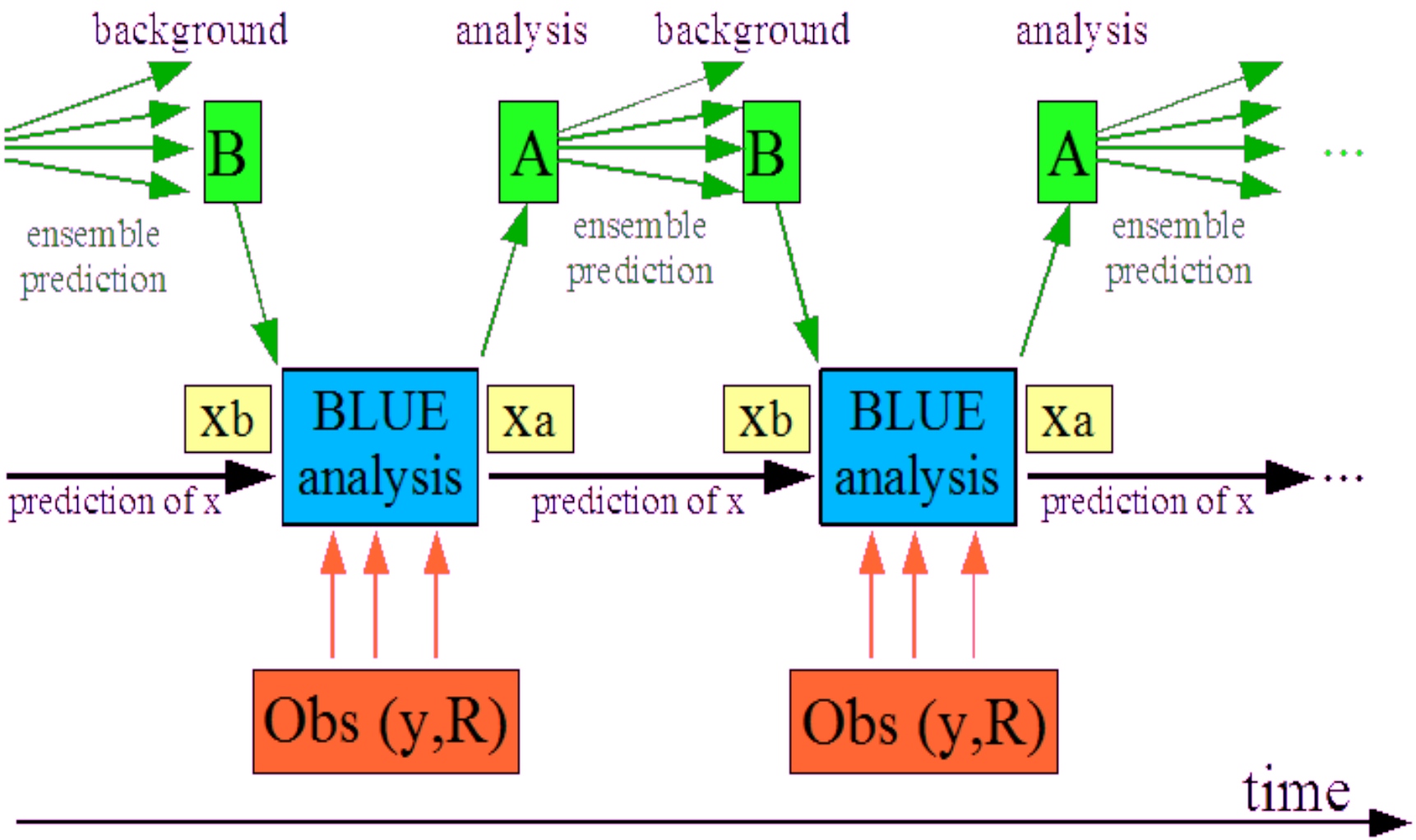
- **amplification** des variances par le modèle dans les zones dynamiquement instables (=peu prévisibles)
- **réduction des variances par l'analyse** au voisinage des observations
- **transport** des erreurs le long du flux

'Divergence du filtre de Kalman': si les variances deviennent trop petites, l'analyse n'utilise plus les obs.

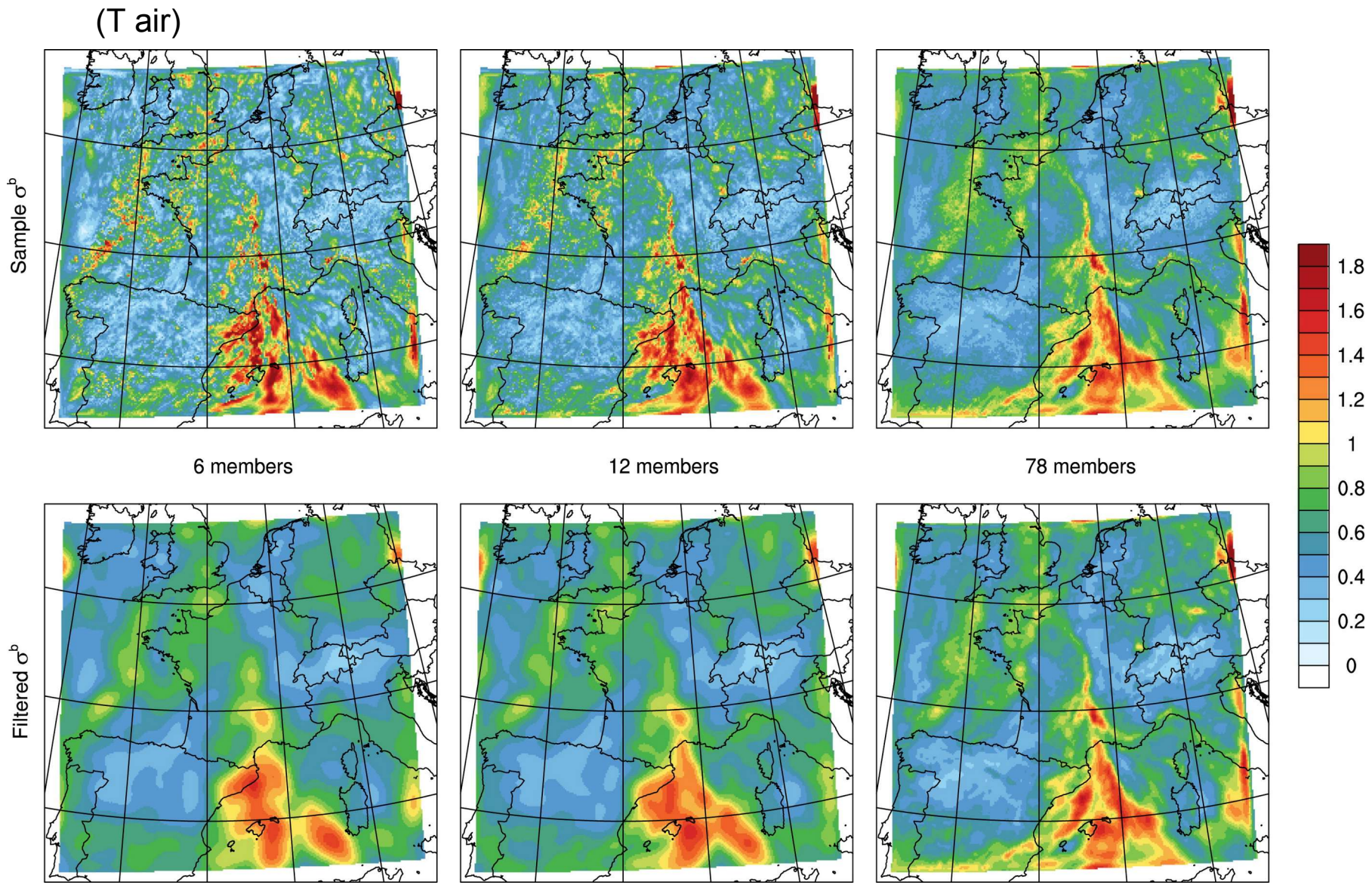
- dispersion des erreurs = étalement des corrélations
- activité ondulatoire = corrélations négatives

Filtre de Kalman d'ensemble:

remplacer l'équation $B=MAM^T+Q$ qui est très chère,
par des prévisions d'ensembles



Exemples de variances d'ébauche ensemblistes



Trois types de Filtre de Kalman d'ensemble:

EnsKF 'square root ensemble filter':

•ensemble de x_b →B → BLUE → A → ensemble d'analyses x_a

•on perturbe aléatoirement les observations avec R

ETKF 'ensemble transform Kalman filter':

•comme EnsKF, plus une *restriction du calcul du gain K au sous-espace engendré par les* x_{bk}

•on ne perturbe pas les obs (=filtre 'déterministe')

•peut être combiné avec une stratégie de sélection locale des observations ('LETKF' = local ETKF)

•plus compliqué mais moins coûteux que le EnsKF

EDA ’ensemble data assimilation’:

- comme EnsKF, mais *on fait autant d’analyses que de membres*
- évite de passer par A
- plus coûteux que EnsKF, mais plus simple, et performant

le futur: **EnVar** : hybridation d’EDA et de 4DVar.

En résumé:

Le filtre de Kalman est une méthode pour produire 1 assimilation

- avec une équation d'évolution de B
- on peut utiliser des méthodes stochastiques (=des prévisions ou des assimilations d'ensemble) pour produire B
- le but reste de produire un seul x_a optimal à chaque instant

L'assimilation d'ensemble est une méthode pour produire des PDFs de x_a et de x_b

- on calcule en parallèle k instances perturbées d'une assimilation
- cette assimilation peut être un filtre de Kalman ou une autre méthode
- elle peut utiliser ou non les x_{bk} pour contruire son B

Ces 2 approches peuvent être combinées ou non.

Filtres particulaires: introduction très simplifiée

t

Filtres particulaires

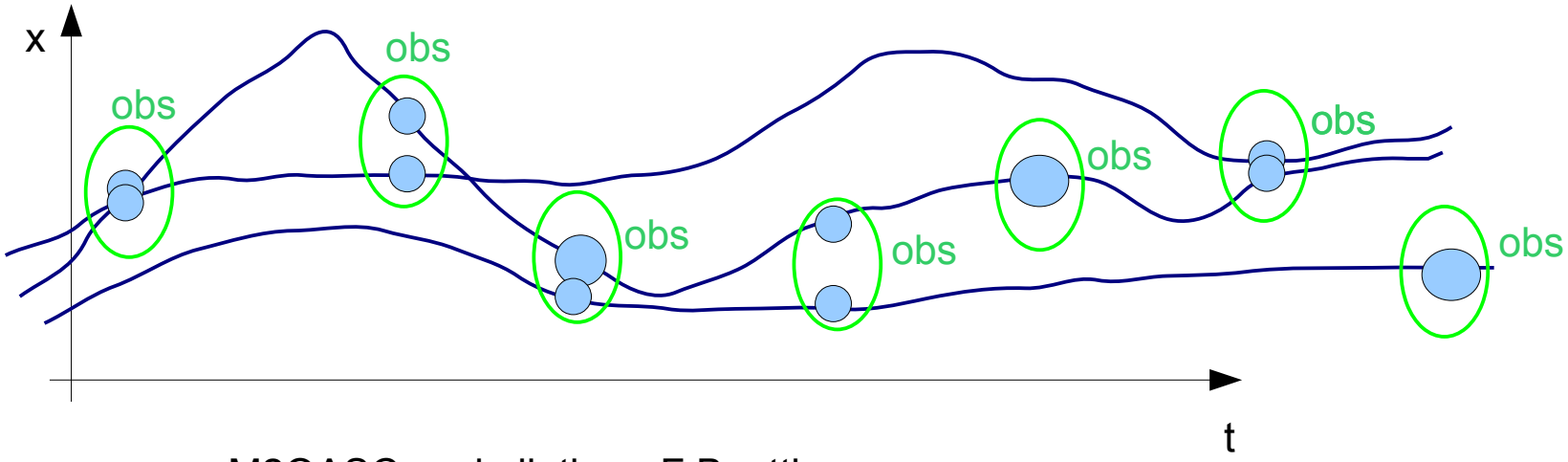
- les filtres de Kalman résolvent (en principe) tous les problèmes de l'assimilation, mais
  - compliqués à mettre en oeuvre
  - font des hypothèses pas toujours acceptables (linéarisation des erreurs de prévision et d'observation)
- les filtres particulaires sont plus simples et plus généraux
  - idée: construire l'analyse en mélangeant plusieurs prévisions d'un ensemble. Les prévisions plus proches des obs ont plus de poids.
  - bien adapté aux problèmes très non-linéaires
  - ne marche bien que dans les petits modèles (n<<100)
  - encore du domaine de la recherche
  - théorie très riche (théorème de Feynman-Kac, Monte-Carlo par chaînes de Markov...)

t

Filtres particulaires

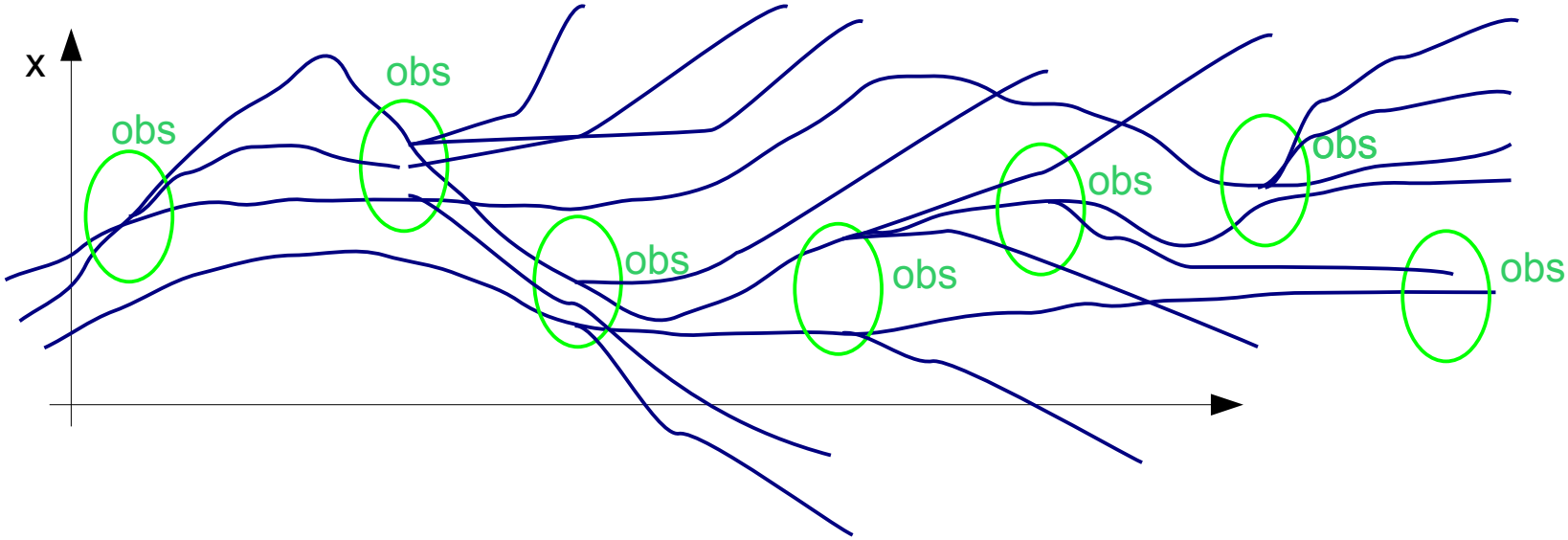
- **Principe:**
  - on effectue des prévisions d'ensemble x_i au cours de l'assimilation
  - **on calcule l'analyse comme étant la moyenne des prévisions x_i les plus proches des obs:**

    $$x=\Sigma w_i x_i \quad \text{avec} \quad w_i= \exp[- (y-H(x_i))^T R^{-1} (y-H(x_i))]$$
  - (plus d'analyse à calculer ! seulement des poids)
  - les membres les plus proches des obs sont les plus probables

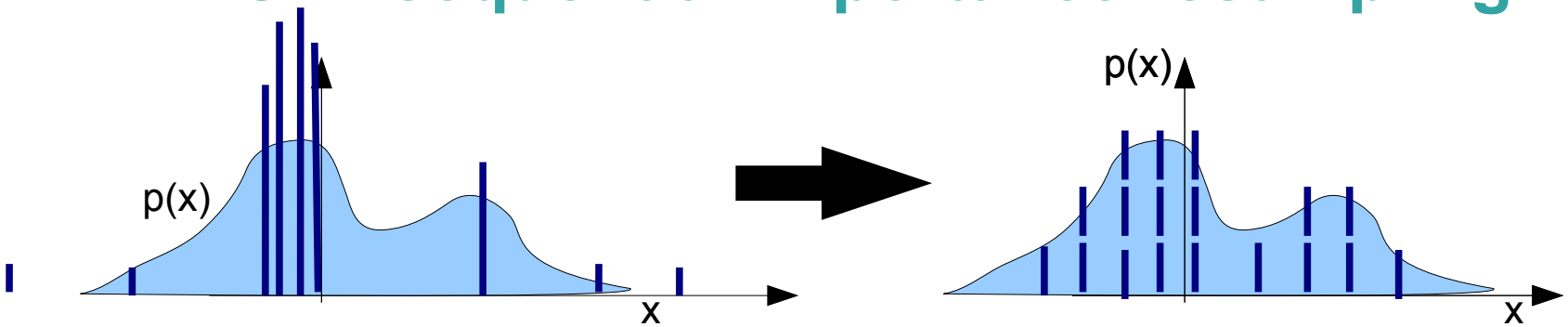


Filtres particulaires

- **Problèmes:**
  - des ensembles énormes sont nécessaires dès que dimx augmente
  - **'dégénérescence'** de l'ensemble: quelques membres deviennent infiniment plus probables que tous les autres
  - solution-type: dupliquer les bons membres, jeter les mauvais (~méthode darwinienne, ou "génétique")



Filtres particulaires à rééchantillonnage: SIR 'sequential importance resampling'

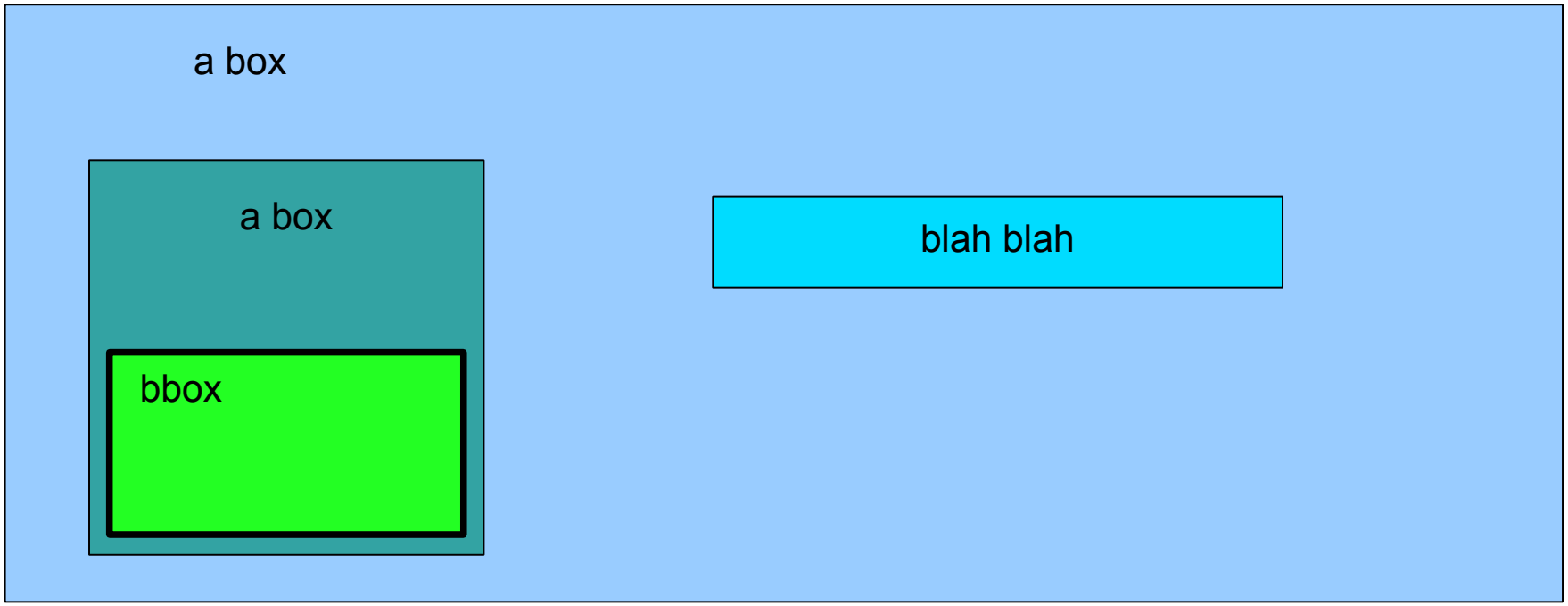


- La **dégénérescence dégrade l'échantillonnage:**
  - beaucoup de membres inutiles (invraisemblables)
  - dispersion mal échantillonnée = le filtre va diverger de la réalité
- Mots-clés des méthodes de **rééchantillonnage**:
  - re-tirage aléatoire dans la densité {wixi} ('**proposal density**')
  - autres tirages plus astucieux (**Metropolis-Hastings**, **Chaines de Markov**)
- Problèmes de **régularisation**: il faut 'secouer' le modèle pour permettre à l'analyse d'être différente de ce que le modèle sait faire. Ex:
  - perturbation aléatoire des éléments dupliqués
  - habillage de la loi de probabilité discrète (**Kernel density estimation**)

Merci pour votre attention

a slide

some text



more text

-

a text

A

- c