

Résumé du cours M2OASC en assimilation de données

François Bouttier - contact: francois.bouttier@meteo.fr

Les transparents et annexes du cours sont disponibles sous <http://assim.chocolate.fr/>

Le présent texte fournit un survol simplifié des notions, pour les étudiants qui n'auraient pas la motivation ou la formation nécessaires pour étudier à fond les algorithmes.

Introduction

L'*analyse objective* est un ensemble de méthodes mathématiques appliquées, destinées à extraire de l'information à partir de mesures de systèmes physiques. En sciences de la terre, notamment celles enseignées dans ce Master, la motivation vient de ce que ces mesures - appelées observations - sont souvent de qualité variable, et disposées de manière irrégulière dans l'espace et/ou le temps. L'objectif est de convertir des mesures de systèmes physiques (océan, atmosphère, sols...) hétéroclites, en représentations de l'état de ces systèmes sur des grilles faciles à étudier (par exemple, champs tracés sous forme de cartes) et de qualité optimale.

L'*assimilation de données* est une classe particulière de techniques d'analyse objective, qui consiste à s'aider d'un modèle numérique du système physique étudié, pour mieux en interpréter les mesures. Le but est donc de présenter un ensemble de techniques utiles au chercheur et à l'ingénieur pour bien utiliser des mesures... ou pour bien comprendre comment des simulations numériques peuvent être effectuées à partir de telles mesures.

Exemples de systèmes physiques où l'assimilation de données est importante:

atmosphère météorologique: l'assimilation de données est indispensable pour simuler l'atmosphère au jour le jour, car son état évolue sans cesse. Les progrès des prévisions météorologiques de ces 20 dernières années sont dûs, au moins pour moitié, aux progrès effectués en assimilation de données (plutôt que des modèles météorologiques eux-mêmes). Tous les grands centres de prévision météorologiques ont une importante activité en assimilation de données. L'étude des techniques d'assimilation est une branche très dynamique de la science météorologique.

océanographie dynamique: l'océan évolue un peu plus lentement que l'atmosphère, mais l'assimilation de données y est aussi importante, surtout depuis que des observations abondantes des océans sont disponibles depuis quelques années (stations automatiques et satellites). Depuis peu, des prévisions quotidiennes de l'état des océans sont produites à partir d'assimilations de l'état des océans (cf. système MERCATOR); des prévisions saisonnières du temps reposent sur des assimilation à grande échelle de l'état des océans tropicaux (par ex. pour détecter les précurseurs d'un El Nino).

chimie atmosphérique: la chimie stratosphérique, ou la prévision de la qualité de l'air, par exemple, ont énormément progressé récemment grâce à la mise au point de techniques d'assimilation de nouvelles observations (d'ozone, méthane, aérosols etc...) dans des modèles chimiques.

étude du climat: la méthode la plus commune pour étudier les évolutions du climat passé est l'examen de séries temporelles d'observations anciennes. Cette technique ne fournit qu'une vision parcellaire du problème, et elle est vulnérable à des dérives des instruments de mesure (par exemple, l'urbanisation croissante autour d'une station météo tend à augmenter

localement la température mesurée, ce qui n'est pas forcément représentatif d'une tendance à grande échelle). L'assimilation de données atmosphérique anciennes, appelée *réanalyse*, permet de s'affranchir de ces problèmes, elle est une nouvelle manière d'étudier l'évolution du climat, bien qu'elle possède ses propres limitations.

état des surfaces: la présence de neige, de banquise, l'humidité du sol, la biomasse, etc, sont des paramètres qui évoluent rapidement et que l'on ne peut pas observer globalement par tous les temps. Le seul moyen d'en faire des cartographies fiables et à jour, est l'assimilation de données dans des modèles de surface.

biologie marine: les modèles numériques d'évolution de la biologie marine commencent à utiliser des systèmes d'assimilation de données pour la définition de leurs états initiaux, cgrâce à l'apparition de satellites appropriés (par ex. mesure de la couleur de l'eau).

observations satellitaires: l'assimilation de données est généralement très utile pour exploiter les mesures par satellites, qui sont généralement trop complexes pour être directement interprétées en termes de quantités physiques concrètes. L'assimilation de données effectue cette conversion à l'aide de modèles numériques du processus de mesure (par exemple, un logiciel de transfert radiatif pour inverser des radiances).

L'assimilation de données offre un cadre mathématique pour traiter de gros volumes de mesures, non seulement pour initialiser un modèle numérique d'un système physique, mais aussi pour fournir une estimation de la qualité de ces analyses, et pour évaluer la qualité de mesures (par comparaison avec un modèle assimilateur qui tient compte d'un grand nombre de mesures, donc peu vulnérable à un petit nombre de mesures erronées). En l'utilisant de manière astucieuse, l'assimilation permet aussi d'optimiser des réseaux de stations d'observations (par des expériences numérique d'impact), détecter des sources de pollution, ou l'origine d'erreurs de prévision (par des techniques de modèle adjoint), etc...

Vocabulaire:

Analyse: représentation de l'état d'un système physique (ex: océan, atmosphère...) sous forme numérique (en général, des valeurs sur une grille de discrétisation), aussi fidèle que possible à la réalité. Le concept d'analyse inclut celui d'interpolation (convertir des observations irrégulièrement disposées en champs sur des grilles) et d'inversion de mesures (convertir des paramètres observés par un instrument [radiances, rétrodiffusion radar...], en paramètres physiques faciles à interpréter: température, courant, etc...)

Ebauche: estimation de l'état d'un système physique avant d'en faire une analyse. En général, ce sera une prévision ancienne (avant de disposer de mesures récentes), ou une estimation climatologique (basée sur des historiques de mesures anciennes)

Assimilation: séquence d'analyses d'un même système physique, dont l'évolution fait intervenir un modèle numérique d'évolution de ce système (par ex un modèle de la dynamique de l'océan, de la neige, etc...)

Une technique fondamentale d'assimilation: l'assimilation séquentielle

Une *analyse* transforme un jeu de mesures, à un instant donné, en analyse notée x_a . Cette analyse peut éventuellement utiliser une ébauche (notée x_b), qui conceptuellement n'est autre qu'une observation un peu particulière de l'état du modèle. On peut donc traiter une séquence temporelle d'observations en alimentant chaque analyse par l'ébauche produite par les

observations précédentes (via un modèle numérique de prévision). Ce processus (voir transparent) appelée *cycle d'assimilation* séquentielle permet d'accumuler de l'information au fil du temps dans un modèle: généralement les erreurs de prévision d'un modèle numérique croissent dans le temps, c'est à dire que le modèle tend à "dériver" loin de la réalité, même si son état initial était une analyse réaliste. L'assimilation consiste à "rappeler" la trajectoire du modèle à proximité de la réalité. Sa qualité est fonction de la qualité des observations, et de la vitesse avec laquelle les erreurs de prévision croissent.

Dans un système géophysique fluide (océan, air, espèce chimique...) soumis à des phénomènes de transport, l'assimilation injecte de l'information à proximité des stations de mesure: grâce à l'interpolation spatiale des données, les analyses et les prévisions à courte échéance seront de bonne qualité à proximité de ces mesures. Les prévisions se dégradent en fonction de leur échéance (car les erreurs s'accumulent dans les modèles de prévision numérique), et l'information analysée va progressivement se déplacer avec le fluide: les meilleures prévisions auront lieu juste en aval des stations de mesure. En répétant assez souvent l'analyse dans une assimilation séquentielle, on pourra non seulement limiter en permanence la croissance des erreurs de prévision près des stations de mesure, on peut aussi espérer limiter ces erreurs en aval des stations. Dans un système où les erreurs croissent lentement, l'information mesurée va pouvoir se propager assez loin dans l'espace du modèle, d'une manière assez complexe qui dépend à la fois de la qualité des mesures, et de la manière dont les signaux se propagent dans le modèle de prévision.

En météorologie, cette notion de propagation de l'information est importante, car de nombreuses régions du monde ne sont pas observées: c'est la dynamique des modèles météo numérique qui permet de faire des prévisions correctes y compris dans les zones peu observées (comme les océans). Dans des systèmes non propagatifs comme le sol, il est impératif d'observer partout (par ex. via des images satellitaires), car on ne peut pas compter sur la dynamique du système pour propager l'information dans l'espace (l'humidité du sol se déplace peu horizontalement); en revanche le sol peut garder assez longtemps la mémoire d'une modification de son état (par ex. une humidification par la pluie, ou une chute de neige).

La notion d'analyse fait intervenir diverses préoccupations:

- *interpolation*: comment spatialiser un champ au voisinage d'une mesure ponctuelle;
- *conversion*: comment convertir une variable mesurée, en variable représentée par un modèle: ce processus n'est pas forcément linéaire, ni même univoque;
- *qualité*: comment prendre en compte la précision d'une observation, si elle est en compétition avec d'autres sources d'information (autres observations, ébauche, ou loi physique que le modèle doit respecter);
- *représentativité*: les modèles numériques n'ont qu'une résolution spatiale limitée. Par exemple, une station météorologique de montagne peut être située dans une vallée trop étroite pour être représentée dans un modèle dont la maille horizontale est de quelques kilomètres. Dans ce cas l'observation, même si elle est de bonne qualité intrinsèque, peut de pas avoir de sens pour le modèle. Comment prendre en compte ces observations ?
- *déconvolution*: certains instruments de mesure observent non pas un champ en un point (comme un thermomètre), mais l'intégrale de quantités physiques sur des volumes non négligeables (par exemple, le radiomètre SMOS mesurera la salinité de l'océan superficiel en moyenne sur des ellipses de ~30km de taille): comment utiliser ces mesures pour produire une analyse sur une grille ?

La suite du cours explique comment traiter ces questions grâce à un cadre mathématique approprié pour l'assimilation de données.

Notions d'analyse objective

L'analyse objective consiste à synthétiser sous forme mathématique des observations d'un système réel, à partir de techniques automatiques, qui sont les seules utilisables lorsqu'il faut tenir compte de gros volumes de données ou de contraintes physiques complexes. La définition de ces techniques repose sur l'identification préalable

- des contraintes que l'on souhaite que l'analyse respecte (exemple de contrainte physique: l'humidité de l'air doit être positive, et en général inférieure ou égale à la saturation)
- des propriétés statistiques du système que l'on analyse: distribution climatologique, texture des champs, variations vraisemblables dans le temps et l'espace...
- du critère d'optimalité qui définit l'analyse en tant que "meilleure représentation possible de la réalité compte tenu des contraintes et des données connues"

Techniquement, il faut aussi définir comment on représente le système physique à analyser (typiquement, les valeurs d'un champ à un instant donné sur une grille de discrétisation), et les observations (typiquement, une distribution spatio-temporelle de mesures ponctuelles qui ont un rapport connu avec le champ à analyser). On veut calculer une analyse la plus optimale possible vis à vis des données disponibles, en respectant les contraintes identifiées. Pour analyser un système physique distribué dans l'espace et/ou le temps (ce qui est souvent le cas dans les matières étudiées dans ce Master), cela revient généralement à interpoler (c'est à dire à spatialiser) des données sur une grille.

Analyse par interpolation

L'interpolation est une technique d'analyse assez simple, qui privilégie l'objectif de spatialisation de données ponctuelles. D'un point de vue théorique, cela consiste à construire l'analyse à partir d'un ensemble de fonctions interpolantes, que l'on cherche à "faire passer" par morceaux au plus près des points de données. Les valeurs analysées sont les valeurs des fonctions interpolantes sur chaque point de la grille du modèle. Exemples de techniques d'interpolation:

- *plus proche voisin*: l'analyse est partout égale à la valeur de l'observation la plus proche.
- *interpolation linéaire*: l'analyse est définie par des fonctions linéaires qui passent entre les points d'observation avoisinants chaque point de grille.
- *interpolation fonctionnelle* (généralisation de l'interpolation linéaire): on prédéfinit une famille de fonctions interpolantes, ajustables via des paramètres (exemple: des polynômes, les paramètres sont leurs coefficients). En chaque point de grille, on sélectionne un ensemble de points d'observation pertinents pour analyser ce point (en général, le nombre de points avoisinants nécessaires pour définir une fonction interpolante de manière unique). Ce sont les *points d'appui* de l'interpolation. L'analyse est la valeur de la fonction interpolante au point de grille. Les paramètres de cette fonction obéissent à un critère d'optimalité à définir par l'utilisateur.

L'interpolation linéaire est un cas particulier d'interpolation fonctionnelle, où les fonctions sont des segments de droite. En 1D, si l'on choisit 2 points d'appui par point d'analyse, on a autant de contraintes que d'inconnues pour définir chaque segment. Si l'on veut tenir compte de plus de points d'appui, l'analyse ne peut plus passer par tous, et il faut définir un critère

d'optimalité: si ce sont les moindres carrés, cela revient à faire une *régression linéaire*. Si au contraire on avait plus de coefficients à optimiser que de points d'appui, il faudrait définir un critère d'optimalité supplémentaire pour avoir une solution unique.

Au sens mathématique, calculer une interpolation revient donc, en chaque point d'analyse, soit à résoudre un système d'équations s'il y a autant de coefficients à calculer que de points d'appui, soit à calculer le jeu de coefficients qui optimise le critère d'optimalité choisi lorsque le système d'équations qui définit les coefficients est sur- ou sous-déterminé.

Limitations des techniques d'interpolation: dans une géométrie multidimensionnelle (2D ou 3D), le choix des points d'appui n'est pas forcément trivial, surtout si les observations sont irrégulièrement réparties. Il existe des techniques mathématiques spécialisées pour résoudre ce problème (ex: triangulation de Delaunay). La manière d'interpoler un champ dans les zones dépourvues en observations, ou au contraire dans les zones où coexistent des observations contradictoires entre elles, n'est pas claire, et l'interpolation peut créer des artefacts dans cas-là (notamment des *overshoots*, c'est à dire que l'interpolation par ex. par des splines peut créer des maxima et minima locaux très au-delà de ce qui est observé). Enfin, on ne peut interpoler que des observations qui portent sur le même paramètre physique que celui que l'on veut analyser. Pour toutes ces raisons, les techniques conventionnelles d'interpolation sont plutôt adaptées à la visualisation de données, qu'à l'analyse de fluides géophysiques complexes, notamment s'il s'agit d'initialiser un modèle de prévision numérique.

Remarque: si l'on cherche à interpoler sur une grille beaucoup plus lâche que le réseau d'observations, il faut se méfier du choix des points d'appui. Une interpolation qui n'utilise que les plus proches voisins, implique que de nombreuses observations n'auront aucun impact sur l'analyse (celles qui ne sont jamais sélectionnées comme point d'appui car elle sont trop loin de tous les points de la grille d'analyse). Pour éviter ce syndrome, il faut utiliser des techniques d'interpolation particulières, dites "*conservatives*", qui sont formulées de manière à conserver au mieux l'intégrale des valeurs observées.

Analyse objective de type Cressman

Plusieurs techniques d'analyses empiriques, plus faciles à mettre en oeuvre que les interpolations ci-dessus, ont été mises au point au fil des années. On expose ici la technique dite de Cressman, qui est très simple et robuste. Il faut définir au préalable une fonction de poids notée w , qui est une fonction positive décroissante de la distance géométrique entre points d'analyse et points d'observations. Cette fonction définit dans quelle mesure un point d'analyse est influencé par les observations voisines; lorsqu'elle vaut zéro, les observations n'ont plus aucun poids. L'analyse de Cressman x_a est définie en chaque point i comme le barycentre des observations avoisinantes:

$$x_a(i) = \sum_j [w(d_{ij}) y_j] / \sum_j w(d_{ij})$$

où l'indice j parcourt l'ensemble des observations, et d_{ij} est la distance entre le point de grille i et l'observation j . L'analyse est indéfinie là où $\sum_j w(d_{ij})=0$, c'est à dire aux points trop éloignés des observations au sens de la fonction w .

Choix de la fonction w : on peut choisir w de manière empirique, par exemple $w(d)=\max(0,1-d/D)$ où D est une constante (*rayon d'influence*) à définir par l'utilisateur. Intuitivement, l'analyse sera plus réaliste si la manière dont les observations sont lissées correspond à la texture typique du champ que l'on cherche à analyser:

- champ habituellement très lisse -> lisser les observations sur de grandes distances
- champ comportant de petites structures -> chaque observation ne renseigne que sur un petit voisinage

Une manière objective de prendre en compte cette notion de texture consiste à définir w à partir du *variogramme* des données à analyser, puisque le variogramme est une mesure statistique de la manière dont des points voisins sur lesquels les valeurs du champ ont tendance à être corrélées entre elles.

Rappel: le *variogramme* $W(d)$ d'un jeu de données spatialement distribuées $a(i)$ est la fonction de covariance entre ces données, et ces mêmes données décalées d'une distance d :

$$W(d) = \text{cov}(a(i) - \text{moy}[a(i)] , a(j) - \text{moy}[a(j)]) \text{ avec } j \text{ tel que } d_{ij}=d$$

En pratique on peut calculer le variogramme en prenant un grand nombre d'observations, on construit l'ensemble de tous les couples d'observations que l'on classe en fonction de leur distance mutuelle. On calcule l'histogramme des covariances pour différentes classes de distance (en définissant l'histogramme afin qu'au moins quelques dizaines de couples peuplent chaque classe, afin que les covariances soient statistiquement bien échantillonnées), et on construit la fonction variogramme W empiriquement de manière à ce qu'elle passe près de toutes les covariances obtenues.

Ce type d'interpolation fondée sur le variogramme s'appelle un *krigeage*. Pour les champs dont la texture est très anisotrope, ou très inhomogène, on peut raffiner en faisant dépendre la fonction W de l'orientation, de l'emplacement, etc, plus cette fonction sera complexe et plus il sera compliqué de la calibrer de manière statistique.

Utilisation d'une ébauche

On a mentionné ci-dessus la présence de "trous" dans l'analyse lorsqu'on est loin de toute observation. On peut contourner ce problème en utilisant une ébauche, c'est à dire une estimation préalable x_b du champ à analyser. En général on construit l'ébauche à partir d'une prévision ancienne, une climatologie, ou une analyse déjà faite à plus grande échelle. Le problème de calcul d'une analyse par interpolation (Cressman ou autre) des observations, se ramène à celui de l'interpolation des écarts à l'ébauche. Schématiquement, si on note A l'opération d'analyse, cela signifie qu'au lieu de calculer

$$x_a = A(y)$$

(y est le vecteur formé par les valeurs observées à l'instant d'analyse),

on veut calculer

$$x_a = x_b + A(y - H[x_b])$$

de telle sorte que $x_a(i) = x_b(i)$ aux points i éloignés de toute observation. L'opérateur H signifie que l'on interpole le champ d'ébauche en chaque point d'observation, ce qui est nécessaire pour comparer les observations à l'ébauche: là où les deux sont égaux, on laissera l'analyse égale à l'ébauche. Si les observations contiennent des valeurs plus élevées que l'ébauche, on souhaitera que l'analyse soit elle aussi plus élevée, etc... En jargon d'assimilation, on appelle:

- *innovation* la quantité $y - H[x_b]$
- *incrément* la quantité $x_a - x_b$
- *opérateur d'observation* la fonction H qui interpole l'ébauche aux points d'observation.
- *fonction de structure* l'incrément obtenu en présence d'une seule observation scalaire.

En résumé, la présence d'une ébauche permet de transformer le problème d'analyse d'observations, en problème d'interpolation d'écarts à l'ébauche. Cela résout le problème des "trous", au prix de la définition de l'opérateur H .

Opérateur d'observation

Dans la section précédente, l'opérateur d'observation H est introduit dans le cas où l'on observe en divers points les mêmes variables que celles que l'on analyse: c'est simplement une interpolation spatiale. En pratique, on peut avoir à traiter des observations beaucoup plus complexes:

- observations de paramètres physiquement différents de ceux représentés par le modèle
- observations découlant d'un processus d'intégration spatiale complexe (ex. pixel des mesures satellitaires, volumes d'échantillonnage radar)
- observations découlant d'un processus physique complexe (ex. transfert radiatif: on observe des radiances, qui dépendent à la fois de la température, de l'humidité, des nuages, des surfaces...)

Dans chaque cas on peut continuer à définir H comme une application de l'espace du modèle vers celui des observations: H simule la valeur qu'auraient les observations si l'ébauche était parfaite. En revanche, la technique d'analyse devra convertir des innovations depuis l'espace des observations y vers la grille d'analyse x . C'est l'objet des algorithmes de la famille du BLUE, qui sont exposées dans la suite.

Analyse statistique aux moindres carrés BLUE

Pour remédier aux faiblesses des techniques présentées ci-dessus, un ensemble d'algorithmes d'analyse et d'assimilation de données a été mis au point. C'est maintenant la norme dans la quasi totalité des centres de modélisation numérique de l'atmosphère et de l'océan (à quelques exceptions près, très expérimentales ou très anciennes). La manière la plus simple de comprendre ces algorithmes est de partir du cadre mathématique (le BLUE), qui est commun, les différents types d'applications pratique (OI, 3D-Var, 4D-Var, filtre de Kalman, etc) s'en déduisent comme différentes approximations du même BLUE.

Le BLUE est explicité sur les transparents (section 7 du cours). On retiendra l'équation d'analyse 'directe':

$$x_a = x_b + BH^T(HBH^T + R)^{-1} (y - H[x_b])$$

et sa *forme variationnelle*:

$$x_a = \text{Arg min } J \quad \text{avec } J(x) = (x - x_b)^T B^{-1} (x - x_b) + (y - H[x])^T R^{-1} (y - H[x])$$

Il est important de bien comprendre le sens des symboles utilisés dans les équations ci-dessus.

Méthodes de calcul du BLUE

Dans les problèmes de petite dimension on peut calculer directement l'équation d'analyse directe, qui est en fait un système linéaire d'équations dont le vecteur des inconnues est $(x_a - x_b)$. Pour des dimensions (de x ou y) plus grandes on peut rendre creux le système d'équations en supposant que certaines covariances d'erreur sont nulles; c'est l'ancien algorithme dit "d'interpolation optimale".

La *technique variationnelle* a maintenant supplanté ces approches, elle consiste à minimiser approximativement la fonction J de la forme variationnelle, en démarrant de $x = x_b$ et en améliorant itérativement la valeur de x en fonction des dérivées partielles de J (appelées *gradient de J en x* , noté $\nabla J(x)$), pour que chaque nouveau x réduise un peu plus la valeur de J (ou de la norme de son gradient, ce qui revient au même si J est quadratique, donc convexe). La qualité de l'approximation du minimum de J se mesure en regardant la norme du gradient: pour la plupart des problèmes pratiques on utilise des algorithmes de minimisation standards

disponible dans des bibliothèques de logiciels scientifiques (par ex. le gradient conjugué), et on effectue quelques dizaines d'itérations pour diminuer la norme du gradient de J d'un facteur de 10 à 100 environ par rapport à sa valeur initiale.

Illustrations du BLUE: les transparents montrent le comportement de l'analyse x_a fournie par la technique du BLUE, sur des problèmes idéalisés (0-D et 1-D avec des covariances gaussiennes). On retiendra que le BLUE:

- donne d'autant plus de poids aux observations par rapport à l'ébauche, que les variances de R sont petites par rapport à celles de B
- effectue une spatialisation des observations avec des fonctions de structure proportionnelles aux covariances contenues dans B , si R est diagonale (ce qui est le cas le plus fréquent). Le BLUE se comporte donc de manière analogue au krigeage.
- la structure de H influe aussi sur les fonctions de structure (qui valent BH^T lorsqu'on analyse une seule observation)
- effectue un lissage entre observations voisines, lorsqu'elles sont plus proches les unes des autres que la distance de corrélation impliquée dans B .
- accorde plus de poids à l'analyse du gradient entre observations voisines, plutôt qu'à leur valeur absolue, lorsque les erreurs d'observations sont corrélées entre elles.

Construction des matrices B et R

Ce sont des matrices de *covariances d'erreur*; se reporter aux transparents pour leur définition exacte. On peut les calculer:

- en utilisant la connaissance de la qualité des instruments de mesure pour spécifier les variances de R , et celle de la qualité des prévisions pour les variances de B
- de manière statistique en calculant des *histogrammes de covariances* sur un grand nombre d'innovations et en supposant que les erreurs d'observations sont décorréélées entre elles (donc que R est diagonale)
- de manière physique en supposant certaines relations physiques entre les erreurs (par exemple, les erreurs sur le vent et sur le champ de pression atmosphériques sont approximativement corrélées entre elles selon la balance géostrophique). On appelle *balance* ces relations physiques.
- directement par la technique du filtre de Kalman (très coûteuse), que l'on peut approximer en pratique par des techniques de prévision d'ensemble. Il faut alors supposer que les différences entre membres de l'ensemble ont les mêmes statistiques que les erreurs de prévision, ce qui permet de calculer B .

Ces calculs sont en pratique complexes et très approximatifs. Les méthodes d'assimilation 4D-Var et de filtre de Kalman d'ensemble permettent sous certaines hypothèses de contourner le problème de la construction de B , en utilisant un modèle de prévision numérique pour fabriquer automatiquement des fonctions de structures optimales. Ces techniques sont souvent utilisées en météorologie et océanographie opérationnelle, mais elles sont très coûteuses à calculer, et complexes à coder.

Assimilation variationnelle 4D-Var et technique adjointe

La forme variationnelle du BLUE:

$$x_a = \text{Arg min } J \quad \text{avec } J(x) = (x - x_b)^T B^{-1} (x - x_b) + (y - H[x]) R^{-1} (y - H[x])$$

se généralise à des observations distribuées dans le temps, sur une fenêtre temporelle donnée (de quelques heures à quelques jours). Pour cela il faut

- considérer que x et x_b sont définis au début de la fenêtre temporelle
- remplacer H par un opérateur composé (1) d'une prévision numérique partant de x , jusqu'à l'instant où chaque observation est définie, et (2) de l'opérateur d'observation à chacun de ces instants. Autrement dit, H intègre maintenant un modèle de prévision numérique qui compare une trajectoire prévue à un jeu d'observations. Minimiser J , c'est trouver le x_a tel que la prévision numérique initialisée par x_a passe au plus près de toutes les observations disponibles sur une fenêtre temporelle. L'ébauche x_b joue le rôle d'une observation un peu particulière, au début de la fenêtre.
- minimiser J , ce qui nécessite en pratique de calculer le gradient de J pour différentes valeurs de x . Ce gradient nécessite (1) la dérivation du modèle de prévision numérique par rapport à sa condition initiale, qui s'appelle *modèle linéaire tangent*, et (2) la transposition de cet opérateur dérivé, qui s'appelle *modèle adjoint*.
- une factorisation de l'expression de la dérivée de J permet, pour chaque valeur de x testée au cours de la minimisation, de calculer la valeur $J(x)$ et le gradient $\nabla J(x)$ au prix d'une intégration du modèle numérique, et d'une intégration de son modèle adjoint (forcé par un terme qui dépend des observations)

L'assimilation 4D-Var produit des analyses optimales si le modèle numérique est linéaire par rapport aux perturbations de sa condition initiale x ; si le modèle est faiblement non-linéaire, elle produit des analyses approximativement optimales, dont l'intérêt est que les fonctions de structure dépendent non seulement de B , mais aussi de la dynamique du modèle numérique. Cela permet de prendre en compte les effets de transport, diffusion, instabilités, ondes, etc. dans le processus d'assimilation. La contrainte principale du 4D-Var réside dans la nécessité de coder le modèle adjoint, qui est une transformation du modèle numérique original.

Autres applications du modèle adjoint

Le modèle adjoint présente d'autres applications importantes, au-delà de l'assimilation:

- pour calculer la *sensibilité des prévisions* du modèle numérique par rapport à des perturbations de sa condition initiale, ce qui permet de mieux comprendre comment l'information s'y propage, et aide à positionner des observations pour optimiser la prévision numérique de certains phénomènes comme les tempêtes (technique de *ciblage d'observations*)
- pour calculer des formes approchées du filtre de Kalman, cf section suivante
- pour calculer les perturbations de l'état du modèle qui risquent le plus de s'amplifier fortement au cours de la prévision: ce sont les *vecteurs singuliers* du modèle numérique.

Les vecteurs singuliers présentent des propriétés intéressantes pour construire des systèmes de prévision d'ensemble (cf ci-dessous); pour un opérateur de prévision M donné (=intégration d'un modèle numérique) ils sont définis comme les perturbations u de son état initial x , de norme fixée, dont la prévision est une perturbation de norme maximale:

$$u = \text{Arg max } \| M(x+du) - M(x) \| / \| du \|^2$$

On peut montrer que les vecteurs singuliers sont les vecteurs propres de la matrice $M^T N M$ où N est la matrice de la norme de la perturbation de la prévision, et M' est le modèle linéaire tangent, M'^T est le modèle adjoint. Il existe des techniques numériques efficaces (méthode de

Lanczos) pour calculer approximativement les vecteurs singuliers d'un modèle à partir du modèle adjoint.

Filtre de Kalman et prévision d'ensemble

Le filtre de Kalman est une généralisation de l'équation du BLUE qui définit un système d'assimilation séquentiel complet optimal. Il consiste à ajouter à l'équation du BLUE

- une équation de calcul des covariances d'erreurs d'analyse $A(t)$ à l'instant d'analyse t (montrée dans les transparents de présentation du BLUE: $A=(I-KH)B$ ou de manière équivalente $A=2(J'')^{-1}$)
- 2 équations de prévision: l'une pour l'état de modèle, qui est simplement une intégration d'un modèle numérique, et une pour construire les covariances d'erreur d'ébauche B de l'analyse suivante à $t+dt$ (ce sont les covariances d'erreur de prévision numérique)

Cette dernière équation a la forme

$$B(t+dt) = M' A(t) M'^T + Q$$

où Q est une matrice de covariances d'erreur de modélisation (en pratique spécifiée de manière très empirique). Cette équation est très coûteuse à calculer pour des modèles de grande dimension, puisqu'elle nécessite de l'ordre de $\dim(x)$ intégrations du modèle numérique. Le filtre de Kalman est donc réservé à des études théoriques, ou à des modèles de petite dimension (par ex des modèles de l'état d'une surface en un point).

L'intérêt du filtre de Kalman est

- de résoudre le problème de la spécification de la matrice B : même avec un terme Q très approximatif, l'équation ci-dessous injecte dans B de l'information sur l'évolution dynamique des erreurs de prévision, grâce à l'emploi du modèle M .
- de fournir une estimation précise des erreurs d'analyse grâce au calcul de A

L'assimilation 4D-Var est, pour des modèles quasi-linéaires (cas fréquent en météo et océanographie), plus intéressante que le filtre de Kalman pour calculer en pratique des analyses. En revanche, elle prend mal en compte les non-linéarités, et surtout le filtre de Kalman semble mieux adapté pour calculer les erreurs d'analyse.

Filtre de Kalman d'ensemble: une approximation intéressante du filtre de Kalman consiste à remplacer l'équation ci-dessus par une prévision d'ensemble (typiquement avec quelques centaines de prévisions, qui effectue un échantillonnage des erreurs de prévision par les étapes suivantes:

1. tirer aléatoirement (ou avec une technique plus astucieuse, comme des vecteurs singuliers ou les *vecteurs de Lyapunov*) un nombre n de perturbations u les plus probables de l'analyse, telles qu'elles échantillonnent la distribution statistique impliquée par A
2. effectuer n prévisions perturbées $M(x_a + u)$ en plus de la prévision optimale $M(x_a)$
3. calculer Q plus la matrice de covariances des perturbations prévues $[M(x_a + u) - M(x_a)]$: c'est une approximation de B (en pratique il faut la modifier pour la rendre définie positive)
4. appliquer les équations d'analyse du filtre de Kalman pour calculer x_a et A pour le cycle suivant.

Cela réduit considérablement le coût de l'assimilation (qui devient de l'ordre de n au lieu de $\dim(x)$ pour le filtre complet), au prix d'approximations sur l'échantillonnage ensembliste de

A, et sur la reconstruction de B à partir de l'information ensembliste (qui est incomplète puisqu'elle ne gère que n structures d'erreur possibles, au lieu des $\dim(x)$ présentes en théorie).

Filtre de Kalman par transformée d'ensemble (ETKF): un intérêt du filtre de Kalman d'ensemble est la prise en compte des non-linéarités du modèle au cours des prévisions de l'ensemble; on peut conserver cette information en effectuant n analyses supplémentaires à chaque cycle, chacune utilisant une ébauche perturbée $M(x_a + u)$: cela fournit directement n analyses perturbées, au prix du calcul de ces analyses supplémentaires. L'ETKF est surtout utile pour initialiser des prévisions d'ensemble, puisqu'il modélise avec précision l'évolution des erreurs sur x tout au long du cycle d'assimilation.

Actuellement l'ETKF et le 4D-Var font l'objet de recherches intenses pour les améliorer, tant du point de vue du calcul des analyses, que de celui de la prévision d'ensemble. On note deux spécificités de leurs implémentations actuelles:

- le 4DVar est généralement calculé de manière *incrémentale*, c'est à dire à résolution réduite par rapport à celle du modèle de prévision (on ajoute un incrément basse résolution à une ébauche haute résolution pour calculer une analyse haute résolution) et avec des opérateurs d'observation linéarisés (pour réduire le coût de la minimisation et garantir sa convergence).
- l'ETKF utilise généralement des hypothèses simplificatrices fortes pour le calcul de l'analyse et de ses covariances d'erreur A, en projetant l'équation d'analyse dans le sous-espace engendré par l'ensemble d'ébauches. En pratique on modifie artificiellement aussi la dispersion des ensembles pour la rendre cohérente avec les erreurs réelles (mesurable par de statistiques sur les innovations), notamment pour éviter l'"effrondement" de l'ensemble (décroissance irréaliste des covariances d'ébauche jusqu'au point où les analyses n'utilisent plus les observations). Les implications sur l'optimalité de l'assimilation sont encore mal comprises.

La **prévision d'ensemble** proprement dite vise à accompagner une prévision numérique d'une estimation de son incertitude, en échantillonnant sa loi de probabilité par un petit ensemble de prévisions (quelques dizaines). L'objectif est d'estimer précisément des lois de probabilités significatives pour les utilisateurs (probabilité d'une tempête, d'un El Nino, de pluie, de dissipation d'un brouillard...) pour un coût de prévision minimal. Cela nécessite de modéliser précisément les sources d'incertitudes:

- les erreurs de modélisation (terme Q dans l'équation du filtre de Kalman) peuvent être représentées en ajoutant un bruit artificiel dans les équations d'évolution du modèle, en perturbant ses paramètres de réglage, ou ses conditions aux limites, ou enfin en dilatant artificiellement la dispersion de l'ensemble
- l'évolution des erreurs de prévisions à courte échéance, dépend de leur évolution au cours du cycle d'assimilation passé, on utilise pour cela un ensemble d'assimilations perturbées les unes par rapport aux autres, par exemple avec la technique de l'ETKF; il existe des techniques voisines d'ensemble d'assimilation, plus rudimentaires (*breeding method*, apparentée au calcul de vecteurs de Lyapunov)
- les erreurs d'observation (terme R) dans l'assimilation peuvent être modélisées en perturbant aléatoirement les valeurs observées dans un ensemble d'assimilations;
- les erreurs de prévision à croissance rapide (trop rapide pour être déduites des erreurs d'analyse) sont en pratique bien modélisées en ajoutant aux prévisions des perturbations proportionnelles aux vecteurs singuliers (cf ci-dessus) du modèle de prévision linéarisé.

La tendance actuelle est une imbrication de plus en plus étroite entre algorithmes d'assimilation et de prévision d'ensemble, pour prendre en compte les aspects non-linéaires de l'évolution des erreurs.
