Analyse et assimilation de données

Introduction

L'analyse des observations de l'environnement est une étape indispensable pour garantir la cohérence des modèles numériques (ou statistiques, dans le cas de l'IA) avec la réalité. En sciences de l'atmosphère et du climat, la rareté et la complexité des observations nécessitent des techniques particulières: un soin particulier doit être apporté à la comparaison entre données mesurées et variables modélisées, car leur relation peut être complexe. De plus, les observations sont souvent trop rares pour initialise les modèles météorologiques ou climatiques avec des interpolations directes. On utilise donc des techniques mathématiques appelées "assimilation", combinant contraintes physiques et modèles statistiques pour trouver les états de modèle les plus probables compte tenu des observations. Elles permettent d'estimer des variables environnementales qui sont peu ou pas mesurées directement. L'assimilation reste un problème ouvert, crucial pour produire des prévisions de l'environnement et des jeux de données d'apprentissage pour les modèles d' intelligence artificielle.

Introduction : l'équilibre entre information et incertitudes

Ce chapitre explique des techniques de production d'information utile à partir de mesures du système climatique appelées **observations**, notamment pour la modélisation numérique : elles permettent de critiquer le comportement des modèles (en fournissant des références pour les valider), voire de les initialiser, en contraignant les simulations numériques à rester conformes au comportement de la nature.

Ce domaine mathématique s'est développé en météorologie vers le XIXe siècle, lorsqu'il qu'il est apparu que certains phénomènes, comme les tempêtes, pouvaient se prévoir (plus ou moins bien) par extrapolation des observations effectuées en amont du flux atmosphérique. Au XXe siècle, les progrès des connaissances sur les lois

d'évolution de l'atmosphère ont permis de perfectionner les modèles (conceptuels, puis numériques, et maintenant IA) au point de les rendre capables de simuler fidèlement la quasi totalité des phénomènes d'importance pratique en météo. Cependant, notre capacité à les prévoir a été - et reste aujourd'hui - limitée par la qualité de l'initialisation de ces modèles. En effet, la plupart de ces phénomènes sont des problèmes d'état initiaux, c'està dire que même si on sait modéliser précisément leur évolution avec des équations du type $dx/dt = M(x,t)$ (où t est le temps), la valeur de l'état futur x est souvent très sensible à la spécification de x à l'instant auquel la prévision démarre. Autrement dit, la principale raison pour laquelle on ne sait pas prévoir le temps dans une semaine est que l'on ne connait pas assez bien le temps qu'il fait aujourd'hui.

Ces problématiques existent aussi hors de la météorologie. En mécanique classique, de nombreux systèmes sont modélisables par des équations $dx/dt=M(x,t)$: chute des corps, équation du pendule, trajectoire d'une boule de billard, équations des marées astronomiques... Lorsque cette équation est précisément connue, on peut améliorer autant qu'on le souhaite la prévision en améliorant la précision de l'état initial. Cela s'est avéré faux en météorologie pour les raisons suivantes:

- les erreurs de prévision croissent significativement au fil du temps parce que les équations d'évolution (la fonction M) sont incomplètement connues (par exemple, l'effet de la convection nuageuse sous-maille, de processus de surface complexes, ou d'aérosols mal spécifiés), ou impossibles à calculer avec une précision suffisante (à cause du coût informatique excessif qu'impliqueraient des calculs à une haute résolution spatio-temporelle modélisant l'effet de tous les phénomènes influents). Ce sont les erreurs sur la fonction *M*, dites *de modélisation*.
- ces erreurs sont amplifiées de manière exponentielle dans certaines conditions, notamment en présence d'instabilité hydrodynamique: c'est l'"effet papillon". Ces conditions, dites "*chaotiques*", sont fréquentes dans l'atmosphère (et dans d'autres compartiments du système climatique, comme l'océan). La combinaison de cet effet avec les erreurs de modélisation implique que, quelle que soit la précision avec laquelle l'état initial d'une prévision est estimé, la prévision va presque toujours se dégrader au fil du temps, jusqu'à polluer la totalité de ses informations utiles, à l'exception des statistiques climatiques.
- par ailleurs, les modèles météorologiques (et, plus généralement, du système Terre) actuels comportent beaucoup plus de degrés de liberté qu'il n'existe d'observations pour les initialiser: vers 2020, les principaux modèles météorologiques globaux comportaient ~10^{10} variables historiques, alors qu'environ ~10^{8} variables pertinentes étaient observables en temps réel... et ces observations comportent des erreurs de mesures significatives par rapport à la précision requise pour décrire quantitativement les phénomènes les plus importants. Cela explique que *la rareté de l'information observée*

est aujourd'hui un frein majeur à la qualité des simulations numériques (et, probablement, des futurs modèles d'IA)

Estimer l'état initial d'un modèle métorologique est donc un problème mathématiquement mal posé. Le résoudre va nécessiter des hypothèses et astuces spécifiques pour utiliser au mieux les observations. Les 3 sources d'erreurs énumérées ci-dessus (erreurs de modélisation, amplification chaotique, rareté des observations) vont provoquer un équilibrage entre le taux de croissance des erreurs de prévision et leur contrôle par l'utilisation des observations, dont le but est de maintenir les états de modèle le plus près possible de la réalité. Cet équilibre va varier selon les variables, les régions et les phénomènes. Selon les cas on pourra avoir des simulations très informatives, ou au contraire quasiment inutilisables car polluées par des erreurs aléatoires. Les phénomènes climatiques les moins prévisibles en général comprennent les petits tourbillons (atmosphériques ou océaniques), les nuages convectifs (cumulus, éclaircies...), les phénomènes orageux (grêle, tornades...), et les brouillards par exemple.

Terminologie: analyse, assimilation, vecteur d'état

Ce chapitre décrit des techniques pour convertir des observations en représentations numériques d'un système physique (par ex. l'atmosphère météorologique, l'océan, une surface continentale, etc). Ces représentations appelées **analyses** seront le plus souvent des représentations numériques de variables physiques (par ex. température, vent, etc) sur une grille de discrétisation, le choix des variables et de la grille étant adapté *a priori* à l'utilisation que l'on veut en faire : s'il s'agit d'initialiser une simulation de modèle numérique, ce seront ses variables, mais on peut aussi effectuer des qanalysesl pour d'autres applications, comme la visualisation (une vidéo de pluies observées par radar météo est une représentation sous forme de pixels de pluie analysée à partir de mesures radar géométriquement et physiquement complexes). Les systèmes actuels de prévision météo par IA sont majoritairement entraînés avec des données analysées, et non des observations brutes, car la répartition géographique de ces dernières est souvent trop complexe pour être facilement traitée : on peut consdérer que l'analyse est activité de préparation de données pour l'IA.

En sciences de la Terre, le terme "*analyse*" est réservé au traitement d'observations à un instant donné; lorsque l'on reconstitue l'évolution d'un système physique à partir de séquences d'observations, on parle plutôt d'"*assimilation*".

Définitions:

Analyse : représentation numérique à un instant donné d'un système physique, à partir d'observations de ce système.

Assimilation de données : séquence d'analyses temporellement cohérentes avec l'évolution du système physique.

Les modes de représentations numériques les plus courants sont :

- en points de grille: discrétisation d'un fluide géophysique (atmosphère ou océan) sur une grille 3D (ou 2D pour une surface, 1D pour une colonne de surface continentale, etc). En chaque point de grille, l'analyse contient une approximation des quantités physiques voulues au voisinage de ce point: température, vecteur vent, courant, concentrations... Sa valeur, appelée **vecteur d'état**, est définie par la liste des variables en tous les points de grille.
- représentation spectrale: représentation approchée de champs physiques par combinaison linéaire d'un ensemble de fonctions spatialement distribuées. La plus connue est la transformée de Fourier, qui utilise des fonctions sinus et cosinus, l'analyse étant définie par leurs coefficients d'amplitude et de phase.
- pixels: si l'analyse n'est destinée qu'à de la visualisation, la grille cible est la matrice de pixels de l'image que l'on souhaite produire. C'est une représentation en points de grille indépendante de tout modèle de simulation numérique. Pour montrer une courbe, la grille sera la liste des abscisses x pour lesquelles on veut estimer y=f(x), etc.

Plus de détails sur les grilles des modèles de simulation numérique sont donnés dans le chapitre consacré aux méthodes numériques. Comme expliqué dans les exemples ci-dessus, une fois choisie une représentation numérique, convertir des observations en analyse revient à calculer le vecteur des variables ajustables dans cette représentation :

Définition: **vecteur d'état : liste des variables utilisées pour définir l'état d'une analyse.** Dans ce cours, ce vecteur sera noté x.

Dans la suite, pour simplifier, on ne donnera que des exemples à base de vecteurs d'état constitués de nombres réels en points de grille, mais on pourrait aussi inclure des variables binaires ou catégorielles. On s'intéressera aux représentations spatialement distribuées (grilles 1D, 2D ou 3D), sachant que dans certains domaines, comme l'analyse de surfaces continentales, il pourra être numériquement avantageux de calculer des analyses indépendamment point par point, ce qui revient à travailler sur une "grille" 0D.

Réseaux d'observations et analyses

Le but d'une analyse est de calculer un vecteur d'état x qui soit le plus réaliste possible (en un sens mathématiquement défini plus loin) étant données des observations. On va résumer ici leurs caractéristiques qui influent sur leur intérêt pour des analyses : types de mesures, distribution spatio-temporelle, latence, qualité, et relation avec le vecteur d'état.

RESTART HERE

Supposons que l'on dispose d'un ensemble d'observations pertinentes pour faire une analyse à l'instant où les mesures ont été faites. Ces observations seront définies par :

- les ***valeurs*** des mesures: "valeurs observées" notées *y* (vecteur)

- leurs ***positions*** dans l'espace et dans le temps (qui peut différer de l'instant d'analyse)
- la ***nature*** de la quantité observée, qui n'est pas forcément celle de l'analyse. Par exemple, on peut utiliser des mesures de température et d'humidité spécifique pour analyser un taux d'humidité relative.

Calculer une analyse nécessite de comparer mesures y et analyse x, puisqu'optimiser le réalisme de l'analyse suppose de définir une distance entre x et y. Ces derniers sont en général de nature et de dimension différentes, car il n'y a pas forcément le même nombre d'observations que de points de grille dans l'analyse). Dans la plupart des problèmes d'analyse du système climatique, on a beaucoup moins d'observations que de variables à analyser : le problème de transformer y en x est sous-déterminé. L'outil mathématique approprié pour comparer x et y sera donc une application H de l'espace des x vers celui des y. L'écart entre analyse et observations est mesuré par la différence $y-H(x)$. Pour chaque observation, H calcule la valeur que l'on aurait dû mesurer si le vecteur d'état x était parfaitement réaliste, autrement dit H simule des observations virtuelles à partir d'un état de modèle x. Par convention H s'appelle "opérateur d'observation" (ou *forward operator* en anglais)

définition. **Opérateur d'observation H : application de l'espace d'analyse x vers celui des observations, qui simule la valeur qu'aurait chacune d'elles si x était une représentation aussi fidèle que possible de la réalité**.

Si les observations sont des mesures ponctuelles de la même variable que celle que l'on veut analyser, il est naturel de définir H **comme une interpolation spatiale depuis la grille du modèle vers l'ensemble des points observés**. Dans le cas général H peut être plus compliqué:

- si les variables mesurées ne sont pas les mêmes que celles du modèle, alors H inclut la conversion physique des variables de x en celles de y.
- si les observations ne sont pas des mesures ponctuelles, mais des intégrales sur des volumes de l'espace, alors H doit reproduire cette opération d'intégration, numériquement sur la grille de x. C'est un cas fréquent en télédétection, où l'on mesure des pixels ou des voxels.

Une mesure de radiance satellitaire est un opérateur d'observation complexe : pour simuler une radiance observée par un radiomètre, il faut intégrer un modèle de transfert radiatif le long de sa ligne de visée et à l'échelle du pixel, ce qui peut faire intervenir de nombreuses variables modélisées : propriété radiatives des surfaces visées, température, humidité, nuages et chimie dans l'atmosphère... H joue le rôle d'un "simulateur" du processus instrumental, nécessaire pour comparer les mesures y en leur équivalent modèle $H(x)$. Cette opération ne sert pas qu'en analyse de données : la simulation d'observations est en soi une étape importante d'évaluation du réalisme physique des modèles de climat.

Le degré de réalisme des représentations numériques de l'environnement par x étant limité, il n'existe pas toujours un état de modèle parfaitement cohérent avec les observations. Par exemple, à résolution spatiale limitée, des observations affectées par

du relief, des petits nuages ou des détails de surface ne pourront pas être parfaitement approximées par une analyse x : il existera toujours un résidu irréductible $y\text{-}H(x)$, appelé **erreur de représentativité**, qui s'ajoutera aux **erreurs de mesures** (qui sont une propriété de y) et aux erreurs de calcul de l'analyse (qui sont une propriété de x). Par convention, dans le communauté d'analyse de données, on appelle ***"erreur d'observation"*** la somme des erreurs de mesure et de représentativité. Ce terme est un peu trompeur, car cette quantité ne provient pas seulement des observations elles-mêmes, elle dépend aussi de la représentation numérique utilisée (qui est indépendante des instruments de mesure). Elle est d'autant plus significative que la résolution spatiale de x est grossière par rapport aux processus physiques qui influent sur le processus de mesure. Si H^T simule des radiances, l'"erreur d'observation" intègrera les erreurs de calcul du code de transfert radiatif utilisé, et elle peut dépendre des conditions climatiques locales.

définition. **Erreur d'observation : différence $y - H(x_t)$ entre les mesures y et un état analysé x_t qui serait aussi réaliste que le permet la représentation numérique x.**

La définition ci-dessus est un peu abstraite (car en général on ne connaît pas x_t), elle servira dans ce chapitre pour représenter mathématiquement les incertitudes d'observation dans le calcul numérique de l'analyse.

D'autres propriétés des observations sont importantes pour la qualité des analyses. Il est avantageux d'utiliser le plus possible d'observations pour optimiser la qualité des analyses (dans les limites de leur qualité instrumentale et des budgets de calcul et stockage). Cela signifie que la ***densité spatiale*** des observations doit idéalement se rapprocher le plus possible de la résolution spatiale de la grille de x, et que leur couverture spatiale doit échantillonner au mieux le domaine modélisé par x. En pratique, c'est loin d'être le cas, même si le nombre de mesures se rapproche aujourd'hui du nombre de variables des modèles (grâce à la prolifération des instruments satellitaires à haute résolution saptiale et spectrale), les réseaux d'observation comportent en général des lacunes, avec des déficits en observations (qui limitent la précision des analyses) très importants pour certaines altitudes, certaines zones géographiques, et certaines variables. Pour combler ces "déserts d'observation", des techniques mathématiques d'extrapolation ont été développées, elles sont expliquées ci-dessous.

Les réseaux d'observation de l'environnement possèdent différentes caractéristiques liées à leurs objectifs, à leurs technologies et à leur modèles économiques. On peut distinguer 4 grands types de réseaux :

- les **réseaux opérationnels temps réel**, conçus pour offrir en permanence une densité spatio-temporelle régulière, avec de nombreux points de mesure, et une qualité instrumentale minimale garantie par une approche industrielle de la production et de la maintenance des instruments. C'est le type de réseau le plus ancien, qui a permis l'émergence des prévisions météorologiques opérationnelles entre le XIXe et le XXe siècle, sur des fonds publics gérés par des agences internationales pour les réseaux à vocation mondiale,

nationale ou régionale. L'Organisation Météorologique Mondiale (OMM) coordonne le réseau mondial de radiosondes et stations sol distribuées à tous les services météorologiques des états membres. Les grandes agences spatiales coordonnent les programmes de satellites d'observation de la Terre. De nombreux instituts déploient des réseaux régionaux à densité améliorée par rapport aux réseaux globaux. Un but important de ces réseaux est d'alimenter les systèmes d'assimilation des systèmes de prévision numérique, ainsi que les outils de visualisation temps réel de surveillance de l'environnement, notamment pour lancer des alertes en cas de détection d'événements dangereux.

- les **réseaux de surveillance du climat** sont analogues aux précédents, mais avec une gestion qui privilégie l'homogénéité des observations sur de longues périodes (au moins plusieurs décennies), par rapport à l'optimisation de la qualité des mesures. Leurs sites de mesures sont choisis de manière à maximiser la représentativité (vis à vis de l'environnement de grande échelle) et à minimiser la vulnérabilité à des évolutions locales sans rapport avec le climat de grande échelle, comme l'urbanisation au voisinage des villes. Ces réseaux aident à valider le comportement à long terme des simulations numérique du climat, à faire de la validation croisée avec des instruments à plus faible durée de vie (comme des satellites), et à mener des études d'observation du changement climatique.
- les ***réseaux de recherche*** visent à fournir des mesures aussi précises et denses que possibles sur une zone restreinte, souvent sur des périodes brèves (intensive observing periods (IOP) des campagnes de mesures), et/ou avec des instruments très précis mais coûteux, dont la disponibilité n'est pas garantie sur de longues périodes. Ces réseaux permettent d'étudier de manière intensive des phénomènes spécifiques (nuages, couche limite, aérosols, sites industriels, certains types de surfaces...) dans le cadre d'expériences scientifiques à durée limitée. Certains sites instrumentés observent la quasi totalité des paramètres environnementaux pour étudier des processus physiques et faire des validations croisées (par ex. les sites ARM).
- les ***réseaux d'opportunité*** sont apparus plus récemment grâce au développement d'instruments et de télécommunications à bas coût (l'internet des objets : *IOT, internet of things*) avec l'appui d'un secteur privé motivé par l'acquisition de données environnementales à valeur marchande croissante. On peut citer les stations météo d'amateurs, les smartphones, la domotique, les véhicules connectés dotés de capteurs (bateaux, aéronefs, voitures, engins agricoles...). Ce sont souvent des réseaux à couverture riche, mais avec des instruments peu fiables, et des données contrôlées par des opérateurs privés. Avec le développement de constellations de petits satellites, il est probable que cette tendance s'étendra bientôt aux données spatiales, et pourra nourrir une concurrence entre opérateurs privés et publics pour la fourniture de services environnementaux variés. L'évolution rapide de ces réseaux, leurs

volumes et les problèmes de qualité de mesure qui les accompagnent, en font de bons candidats pour une automatisation croissante du traitement des données par des techniques d'IA.

Pour une utilisation temps réel, la **rapidité d'accès aux données** est essentielle. Les réseaux modernes de communication par internet câblé, réseaux cellulaires et satellites offrent des transmission rapides, mais à un coût qui peut être un frein dans les zones reculées. L'accès aux données des satellites en orbite basse peut être ralenti lorsque ces derniers ne sont pas en vue de stations de télécommunications, ce qui peut occasionner des délais de transmission de plusieurs heures. La production de données peut être freinée par les techniques de mesure: produire un profil vertical de l'atmosphère par radiosondage nécessite d'attendre l'ascension d'une sonde à travers la troposphère, ce qui prend environ 30mn. Des délais d'acquisition de 5 à 15mn sont inhérents aux scans 3D de l'atmosphère par les radar à rotation mécanique, ainsi qu'aux images des caméras de satellites géostationnaires.

Ces délais d'accès à certaines données posent un dilemme aux services d'analyse et de de prévision temps réel : plus on attend les données sur un phénomène d'intérêt, et meilleures seront ses analyses et prévisions, mais ces dernières auront moins de valeur si cela les rend trop tardives pour en tirer parti, par exemple pour lancer des alertes, ou pour agir sur les marchés financiers d'activités météosensibles (cela concerne notamment la production d'électricité éolienne et solaire, ainsi que la consommation d'énergie en fonction des températures). Une solution consiste à fixer un délai appelé ***cutoff*** durant lequel un système d'analyse attend les données après l'heure physique visée. Chaque analyse - et la prévision qui va en découler - est effectuée en pluseurs versions, chacune ayant un cutoff différent. Par exemple, en 2024, les prévisions de Météo-France avec le modèle numérique ARPEGE étaient typiquement effectuées en 3 exemplaires :

- des prévisions accélérées avec un cutoff inférieur à 1h, disponibles avec un timing optimisé pour l'examen par des prévisionnistes de la situation météorologique à grande échelle dans leur planning de travail,
- des prévisions nominales avec un cutoff d'environ 1h, utilisées pour la majorité des applications,
- des prévisions à cutoff "long" (environ 6h), uniquement utilisées pour alimenter le calcul de l'analyse suivante avec la meilleure condition initiale possible.

Techniques d'analyse de base

L'analyse de données est un problème classique en mathématiques, pour lequel de nombreuses solutions existent. Cette section introduit les principales approches motivant les techniques de type BLUE plus particulièrement utilisées en sciences de l'environnement. Pour simplifier, on supposera ici que l'on produit, sur une grille

régulière x à 1 dimension (= un intervalle), une analyse x_a d'une variable physique réelle, observée par des mesures ponctuelles y effectuées au même instant.

La technique de base est l'**interpolation par plus proche voisin**:

1. en chaque point de grille $x(i)$, trouver l'observation $y(j)$ spatialement la plus proche
2. construire l'analyse en reportant cette valeur: $x(i)=y(j)$

Cela s'apparente à l'estimateur kNN (*k nearest neighbours*) bien connu en apprentissage machine. Cet algorithme ne coûte que la recherche du plus proche voisin, multipliée par le nombre de points, ce qui est trivial sur une grille 1D régulière, mais un peu plus complexe en 2D ou 3D si les observations ne sont pas régulièrement disposées. Des algorithmes de programmation géométrique (comme le KDtree) existent pour effectuer cette opération à cout numériquement minimal. Physiquement, l'interpolation par plus proche voisin a l'avantage de préserver les valeurs physiques mesurées. En particulier, si la grille de x est au moins aussi dense que le réseau d'observation, maxima et minima sont préservés, ce qui est crucial pour analyser des phénomènes extrêmes comme les pluies intenses. Si la grille n'est pas suffisamment dense, certaines observations peuvent ne pas être utilisés (celles qui ne sont les plus proches d'aucun point de grille), ce qui peut poser problème. Cette interpolation produit des analyses spatialement discontinues, ce qui peut être inesthétique d'un point de vue visualisation, voire physiquement inacceptable si la régularité du champ analysé est importante. Par exemple, une analyse discontinue d'un champ de vent sera inappropriée pour diagnostiquer une divergence ou un tourbillon.

Les discontinuités spatiales sont évitées par l'**interpolation linéaire** : en 1D,

1. pour chaque point d'analyse $x(i)$, identifier les observations $(y(j_1),y(j_2))$ qui encadrent spatialement le point i, avec les abscisses $j_2>j_1$
2. construire la ***fonction interpolante*** $f(i) = y(j_1) + (i-j_1)/(j_2-j_1) * ([y(j_2)-y(j_1)]$
3. l'analyse est définie en i par $x(i)=f(i)$

Les points j_1, j_2 s'appellent ***points d'appui*** de la fonction f. On peut généraliser ce concept à des fonctions plus complexes que les fonctions linéaires, notamment les fonctions **splines** qui sont plus lisses et fréquemment employées en visualisation. On peut utiliser des fonctions polynômes, sinus, ondelettes, etc, ce qui permet d'imposer (via le choix de l'espace de fonctions interpolantes) des contraintes plus ou moins complexes sur les caractéristiques physiques de l'analyse. Sinus et ondelettes sont utiles pour contrôler les échelles spatiales que l'on souhaite représenter dans l'analyse.

Dans l'exemple ci-dessus, la fonction interpolante était algébriquement définie parce qu'il y avait autant de points d'appui que de degrés de liberté dans l'espace de fonctions choisi (les 2 coefficients de la fonction affine). On peut varier le nombre de points d'appui en exprimant plus généralement l'algorithme comme un problème d'**interpolation fonctionnelle** :

1. définir un espace de fonctions interpolantes $f_a(i)$, où a est un vecteur de coefficients ajustables des fonctions f

2. pour chaque point d'analyse $x(i)$, définir un sous-ensemble de points d'appui J parmi les observations $y(j)$ qui serviront à calculer l'analyse en i
3. calculer les coefficients a qui produisent en i la "meilleure" analyse $x(i) = f_a(i)$ possible. On utilise surtout la distance quadratique entre l'analyse et les observations : $J(a) = || f_a(i) - y(i)||^2$.
4. une fois le vecteur a calculé, l'analyse est définie en évaluant $x(i)=f_a(i)$ en chaque point de grille.

L'interpolation devient alors un problème d'optimisation:

$x = f_a$ où $a = Argmin\ J$, avec J la **fonction-coût** $J(a) = || f_a(i) - y(i)||^2$

Si le problème est bien posé (avec suffisamment de points d'appui, et une dépendance de f à a assez régulière), alors J est une fonction strictement convexe, son minimum a détermine la fonction interpolante f_a qui passe au plus près des observations. S'il n'y a pas suffisamment de points d'appui, alors le problème est sous-déterminé. Il faut alors ajouter une contrainte supplémentaire qui rende J strictement convexe, par exemple avec un terme de régularisation :

$$J(a) = || f_a(i) - y(i) ||^2 + \beta ||a||^2$$

où β est un coefficient réglable (un "hyperparamètre") pour obtenir le comportement voulu. Cet algorithme est lié à de nombreux algorithmes célèbres:

- si toutes les observations y sont des points d'appui pour tous les points d'analyse, cela revient à une régression (linéaire si $f_a(i)$ est une fonction affine de i) : on l'analyse est la fonction linéaire de la coordonnée d'espace qui passe au plus près de toutes les observations.
- en IA, l'***apprentissage supervisé*** revient à faire de l'interpolation fonctionnelle, seul le jargon change:
  - les observations y sont les *données d'apprentissage*
  - la famille de fonction f_a s'appelle le *modèle*, c'est une composée de fonctions élémentaires appelées *neurones*, dont les coefficients d'activation forment le vecteur de poids a
  - $J(a)$ est la *fonction perte* qui mesure la distance entre le modèle et les données y
  - la minimisation de J s'appelle *apprentissage* du modèle f_a
  - l'évaluation de f_a en chaque point i s'appelle *inférence*

  Alors que l'analyse de données conventionnelle vise à produire une analyse x, l'IA produit d'abord l'analyse sous forme d'une *fonction apprise* f_a, qui sera ensuite évaluée en chaque point désiré. Les logiciels modernes d'IA intègrent des outils pour construire des fonctions f_a complexe et des outils de minimisation qui effectuent l'apprentissage de manière numériquement efficace. L'aspect spectaculaire de l'IA, pour le grand public, est que f_a peut être évaluée en n'importe quel point, pas seulement sur une grille, ce qui peut laisser imaginer que f_a l'apprentissage a fabriqué une connaissance physique du problème. Il ne s'agit évidemment que d'un modèle statistique qui suit des

hypothèses préalables (plus ou moins licites) lors du choix de la famille de fonctions f_a, de la sélection des données et de la minimisation.

D'autres méthodes intéressantes sont les modèles additifs, dont l'ancêtre en météorologie est la méthode de Cressman :

1. prédéfinir une fonction poids $w(d_{ij})$ positive décroissante vers zéro, en fonction de la distance géométrique d_{ij} entre les points de grille i et d'observation j. Par exemple, $w = max(0,1-d_{ij}/R)$ ou $w=exp(- d_{ij}^2 / R^2)$. R est un hyperparamètre caractérisant la distance à partir de laquelle w tend vers zéro.
2. en chaque point de grille i, on définit l'analyse comme le barycentre de toutes les observations $y(j)$ pondérées par la fonction poids de leurs distances à i :
   $x(i) = \sum_{i,j} y(j)\, w(d_{ij}) / \Sigma_j w(d_{ij})$

Physiquement, cette formule signifie que l'analyse est la moyenne de toutes les observations situées à proximité, les observations les plus proches ayant un poids maximal. Les observations situées beaucoup plus loin que R ont un poids négligeable. Numériquement, il faut la compléter si où aucune observation n'est disponible à proximité de i, auquel cas le dénominateur $\Sigma_j w(d_{ij})$ est petit (voire nul) : il faut alors déclarer que l'analyse est indéfinie en i. Cela traduit l'évidence que, si aucune observation n'est disponible, alors l'analyse ne peut pas être fondée uniquement sur des observations ; une autre source d'information est nécessaire.

Cette analyse "de Cressman" a de noms dans les diverses chapelles mathématiques. Son intérêt est sa mise en oeuvre est simple et son sens physique clair. Elle s'apparente à l'analyse de Shepard et aux modèles additifs généralisés (*generalized linear models*), notamment les modèles gaussiens (*gaussian models*), qui peuvent être vus comme des réseaux de neurones (les fonctions appliquées à $y(j)$ sont les neurones qui s'activent à proximité de chaque observation, leur somme pondérée en fait un réseau de neurones rudimentaire). On peut chercher à objectiviser la définition de la fonction w en la faisant dépendre des autocovariances spatiales des observations, sous l'hypothèse qu'il est souhaitable que la variabilité spatiale du champ analysé $x(i)$ soit proche de la variabilité spatiale des observations $y(j)$: c'est l'idée de départ de l'algorithme de ***krigeage*** utilisé en géostatistique. L'important est de noter que le choix de la fonction w influe sur le lissage des points d'observation pour construire l'analyse.

Les exemples ci-dessus montrent qu'une analyse est avant tout un lissage de points d'observation. La spécification de ce lissage (par le choix de fonctions interpolantes ou du poids w) a une influence cruciale sur le résultat. Pour des problèmes d'analyse complexes il est donc important de pouvoir s'appuyer sur une méthode qui optimise automatiquement l'opérateur de spatialisation : c'est l'idée des méthodes BLUE exposées dans la suite.

Over/underfitting et assimilation séquentielle

Deux pathologies sont possibles lorsqu'on considère le degré de lissage apporté par une analyse :

- le **surapprentissage** (*overfitting*) : une analyse qui cherche à passer au plus près des valeurs observées peut engendrer des structures non-physiques, si deux observations proches l'une de l'autre ont des valeurs différentes, ce qui peut arriver si elles sont affectées par un phénomène de petite échelle, ou si elles sont entachées d'erreurs indépendantes. L'overfitting peut se manifester par une extrapolation fantaisiste depuis une zone dense en observations vers une zone qui en est dépourvue.
- le **sous-apprentissage** (*underfitting*) : c'est l'inverse du problème précédent : une analyse dans laquelle la spatialisation implique un lissage trop fort peut avoir tendance à "gommer" des structures de fine échelle importantes pour prédire l'évolution future du système physique analysé: présence d'un front, d'un tourbillon instable, d'un nuage convectif en développement, d'une dépression, etc...

L'importance pratique de ces problèmes justifie le soin apporté à la définition des opérateurs de spatialisation dans les systèmes d'analyse opérationnels.

Dans un système qui produit des analyses à différents instants, une solution évidente apparaît pour résoudre les problèmes d'overfitting (et d'indétermination de l'analyse dans les zones pauvres en observation, dans le cadre de l'analyse Cressman). En effet, si l'évolution du système physique analysé est suffisamment prévisible entre deux analyses successives, l'extrapolation de la première à l'instant de la seconde peut être une source d'information intéressant pour compléter les observations. Supposons deux instants d'analyse consécutifs t_1 et t_2 :

1. à l'instant t_1, effectuer une analyse $x_a(t_1)$ en utilisant les observations $y(t_1)$
2. extrapoler cette analyse à t_2 de manière à produire une ébauche $x_b(t_2)$, par persistance (hypothèse $x(t_2) \sim x_a(t_1)$) ou par application d'un modèle de prévision $x_b(t_2) = M(x_a(t_1))$
3. effectuer l'analyse $x_a(t_2)$ en utilisant les observations $y(t_2)$ et l'état $x_b(t_2)$ comme données d'entrée. $x_b(t_2)$ peut être considéré comme un groupe de pseudo-observations disponibles en tout point de la grille x.

Cette technique revient à transformer le problème d'analyse statique à partir d'observations

$$x_a = F(y)$$

où F est un algorithme d'analyse, en un problème d'analyse incrémental

$$x_a = x_b + F(y - x_b)$$

en effet il est naturel de ne modifier x_b que dans la mesure où il est incohérent avec les observations (donc $F(0)=0$). Ce formalisme présente l'avantage de permettre aux observations de porter sur d'autres variables que celles qui sont analysées, puisque l'opérateur d'observation H vu ci-dessus permet de quantifier cette incohérence :

$$x_a = x_b + F(y - H(x_b))$$

L'équation ci-dessus, couplée avec l'équation de prévision $x_b = M(x_a)$ définit la production d'une séquence d'analyses comme une suite d'analyses incrémentales : cela s'appelle l'**assimilation séquentielle**. Elle fournit une technique de traitement en temps réel des observations, puisque les analyses peuvent être calculées dès que les observations ont été collectées (l'ébauche ayant été calculée par prévision à partir de l'analyse précédente). Notons un peu de jargon de l'assimilation de données pour la suite:

- x_b : **ébauche** (*background, first guess,* ou *prior*) : estimation de l'état de modèle, disponible avant que les observations de cet instant ne soient connues.
- $(x_a - x_b)$: vecteur **incrément d'analyse** (calculé en combinant ébauche et observations)
- $(y - H(x_b))$: vecteur **innovation**
- le vecteur incrément obtenu lorsqu'il y a un seul point d'observation (de valeur différente de l'ébauche) s'appelle **fonction de structure**.

Théorème d'analyse BLUE

Le théorème du BLUE indique comment calculer des analyses de manière quasiment optimale, robuste, sous des hypothèses raisonnables, avec des coûts de calcul maîtrisables (moyennant quelques approximations), y compris pour des problèmes de très grande dimension. Il est à la base des algorithmes 3DVar, 4DVar et ETKF aujourd'hui universellement utilisés dans les grands centres d'analyse environnementaux, pour produire des réanalyses et des prévisions en temps réel. La présentation ci-dessous a été simplifiée en évitant les détails mathématiques inutiles pour une utilisation à des fins de modélisation physique.

Notations :

- x_a, x_b : états de modèle analysé et ébauche
- y : observations utilisées pour convertir x_b en x_a
- y_t, x_t : observations qui obtenues en l'absence de toute erreur de mesure, et état "idéal" de modèle représentant parfaitement la réalité physique dans les limites autorisées par la définition de x (notamment la résolution finie de la grille modèle). y_t et x_t sont des objets mathématiques inaccessibles en pratique, ils ne servent qu'à définir les distributions d'erreurs B, R et A ci-dessous.
- $B = E((x_b-x_t)(x_b-x_t)^T)$: matrice d'autocovariance des erreurs d'ébauche x_b (E est l'opérateur de moyenne statistique, ou espérance, sur un grand nombre de réalisations des erreurs $x_b - x_t$).
- $R = E((y-H(x_t))(y-H(x_t)^T)$: matrice d'autocovariance des "erreurs d'observation" définies précédemment. On rappelle que cela inclut les erreurs de simulation des observations par H.

- $A = E((x_a-x_t)(x_a-x_t)^T)$: matrice d'autocovariance des erreurs d'analyse x_a.

Par abus de langage, *B, R* et *A* sont appelées matrices d'erreur d'ébauche, d'observation, d'analyse. L'analyse BLUE prend comme données d'entrée : x_b, *y, H, B, R*, et produit en sortie x_a et *A*. Son calcul nécessite donc que les matrices *B* et *R* aient été calculées au préalable, comme indiqué plus loin.

Le théorème du BLUE s'écrit sous 2 formes équivalentes. Toutes deux garantissent que l'analyse x_a produite est optimale, au sens où sa distance quadratique à l'état idéal x_t, $|| x_a - x_t ||$, est minimale en moyenne.

Théorème du BLUE - calcul matriciel : L'analyse x_a calculée comme $x_a = x_b + K(y - H(x_b))$, avec *K* opérateur linéaire, est optimale pour $K = BH^T (HBH^T+R)^{-1}$ et la qualité de l'analyse est donnée par $A = (I-KH)B$.

Théorème du BLUE - calcul variationnel : L'état x_a qui minimise la fonction-coût $J(x) = (x-x_b)B^{-1}(x-x_b)^T + (y-H(x)) R (y-H(x))^T$ est optimal, la qualité de l'analyse est donnée par $A = (J'')^{-1}$

Ces 2 théorèmes sont vrais sous les conditions suivantes:

- les erreurs doivent avoir des espérances nulles: $E(x_b-x_t)=0$ et $E(y-H(x_t))=0$ autrement dit ébauche et observation doivent être sans biais.
- les erreurs d'ébauche et d'observation doivent être mutuellement décorrélées: $E((x_b-x_t)(y-H(x_t)))^T)=0$ ce qui est réalisé si les observations ont été élaborées sans utiliser x_b ni les observations précédentes (qui ont pu participer au calcul de x_b): c'est automatiquement vrai pour de "vraies" observations (produites par des instruments de mesure), mais pas forcément si elles ont été traitées par des contrôles qualité faisant référence à x_b, ou des corrections de biais utilisant un historique d'observations anciennes.
- l'opérateur *H* doit être linéaire, ou au moins linéarisable au voisinage de x_b : $H(x_a) - H(x_b) \sim H(x_a - x_b)$

Ces conditions ne sont jamais parfaitement réalisées en pratique, mais elles le sont approximativement si l'assimilation a été implémentée de manière physiquement raisonnable, cela implique alors que l'analyse x_a sera approximativement optimale. Pour décider si cette approximation est acceptable ou pas, il faut la mettre en balance avec les autres sources d'erreurs inévitablement présentes dans le BLUE, notamment le fait que l'on ne sait jamais fournir des matrices *B* et *R* parfaites.

Démonstration (simplifiée) du théorème du BLUE: l'optimalité de *K* se montre en minimisant l'espérance des erreurs d'analyse. L'erreur quadratique de x_a est la trace $Tr(A) = Tr(E((x_a-x_t)(x_a-x_t)^T)$

Réarrangeons l'expression des erreurs du BLUE :

$(x_a-x_t) = (x_b-x_t) + K (y - H(x_t) + H(x_t) - H(x_b)) = (I-KH) (x_b-x_t) + K (y - H(x_t))$

On multiplie cette dernière par sa transposée, et on en prend l'espérance : les produits croisés entre (x_b-x_t) et $(y-Hx_t)$ sont nuls par hypothèse de décorrélation entre les erreurs d'ébauche et d'observation.

$E((x_a-x_t) (x_a-x_t)^T) = (I-KH) B (I-KH)^T + K R K^T$

La différentielle de cette expression par rapport à la matrice K prend la forme suivante (on utilise le fait que la fonction trace permute avec l'espérance, qu'elle est linéaire par rapport à son argument, et qu'elle est inchangée quand on transpose son argument):

$dTr(A) = 2\ dK\ H\ B\ (I-KH)^T + 2\ dK\ R\ K^T = 2dK\ (\ HB\ (I-KH)^T + RK^T\ )$

A l'optimum, le facteur de dK est nul donc le K optimal vérifie

$0 = HB(I-KH)^T + RK^T = HB - (HBH^T + R)K^T$ donc

$K = BH^T(HBH^T+R)^{-1}$ ce qui prouve l'expression matricielle de K.

Annulons maintenant le gradient de J en x_a :

$1/2\ dJ(x_a) = 0 = B^{-1}\ (x_a-x_b) - H^T\ R^{-1}\ (\ y - H(x_b) + H(x_b) - H(x_a)\ )$

$= (B^{-1} + H^TR^{-1}H)(x_a-x_b) - H^T\ R^{-1}\ (y-H(x_b))$

et donc

$x_a - x_b = (B^{-1} + H^TR^{-1}H)^{-1}\ H^TR^{-1}(y-Hx_b\ )$

Prouver l'identité matricielle $BH^T(HBH^T+R)^{-1} = (B^{-1} + H^TR^{-1}H)^{-1}\ H^TR^{-1}$ est laissé comme un exercice (c'est un cas particulier de l'identité de Sherman-Morrison-Woodbury). Cela montre l'équivalence entre forme matricielle et variationnelle du BLUE.

L'interprétation physique du BLUE peut se comprendre avec des exemples simples :

- si modèle et observation sont de dimension 1, et que l'on observe la variable que l'on veut analyser), alors les matrices de covariances se réduisent à des variances: $B=\sigma_b^2$ et $R=\sigma_o^2$ (écart-types d'erreurs d'ébauche et d'observation), $H=I$, et $x_a = x_b + (y-x_b)/(1+\sigma_o^2/\sigma_b^2)$: x_a est une moyenne pondérée entre x_b et y, d'autant plus proche de l'observation y que les erreurs d'observation sont faibles devant celles de l'ébauche. Autrement dit, l'analyse est un barycentre de l'ébauche et de l'observation, pondérées par leurs confiances respectives.
- si l'observation est mesure une variable de x sur un point de grille i, alors H est la forme linéaire correspondant au vecteur de base e_i, l'incrément d'analyse est $x_a - x_b = (B\ e_i)\ w\ (y-x_{bi})$ où $(\ Be_i\ )$ est la i-ième colonne de B et $w = 1/(B_{ii} + \sigma_o^2)$ est un terme de pondération entre observation et ébauche comme dans l'exemple précédent. Autrement dit, l'incrément est un champ proportionnel à une colonne de B, qui décrit les covariances d'ébauche entre la variable observée et toutes les variables du modèle. Ce résultat crucial montre que la structure (spatiale et multivariée) de l'analyse est régie par le contenu de la matrice B, qui est "responsable" de la spatialisation des observations. Cela explique l'importance donnée aux améliorations de B en assimilation de données, avec des ressources numériques considérables investies pour le raffiner avec les techniques indiquées dans la suite: algorithmes de filtre de Kalman, 4DVar, 3DEnVar... qui sont pour la plupart fondées sur une modélisation ensembliste de la distribution des erreurs d'ébauche.

Techniques numériques d'analyse

On explique ici les principales approches utilisées pour implémenter des analyses. La plupart sont directement dérivées des équations du BLUE, leurs différences résident dans les compromis effectués sur la précision de la spécification de B et le coût calcul de l'analyse.

Calcul matriciel direct: "interpolation optimale" : la matrice à inverser dans la forme matricielle ayant la taille du nombre d'observations, on peut calculer le BLUE directement s'il n'y en a pas plus de quelques centaines (au-delà, les coûts et les risques d'instabilité numérique augmentent). Il ne faut pas chercher à inverser la matrice (HBH^T+R), sauf si B, R et H sont constants dans le temps, auquel cas on peut précalculer K. Si K est constant ou lentement évolutif, on peut l'apprendre en ligne avec des méthodes d'IA, par ex. des réseaux de neurones récurrents, qui optimisent itérativement l'erreur d'analyse avec une architecture réminiscente du BLUE.

Dans le cas général il est recommandé de décomposer l'équation du BLUE comme suit: $d=(HBH^T+R)^{-1}(y-H(x_b))$ puis $x_a - x_b = BH^T d$. La première étape est le calcul du vecteur d comme solution d'un système d'équations à matrice symétrique définie positive : $(HBH^T+R)d = y-H(x_b)$; des solveurs efficaces pour cela sont disponibles dans les librairies scientifiques. La seconde étape calcule simplement l'analyse par $x_a = x_b + BH^Td$. C'est ainsi que l'on calculait les analyses en météorologie dans les années 80. Pour réduire les coûts calcul on peut exploiter le fait que, pour des choix raisonnables de B et de R, les observations n'ont d'influence que dans un voisinage limité, on peut donc analyser chaque région de l'espace en n'utilisant que les observations à proximité (cela revient à supposer que (HBH^T+R) est une matrice diagonale par bloc). Ces approches sont encore utilisées dans les systèmes d'analyse temps réel pour lesquels les temps de réponse sont critiques (par ex. dans le spatial et dans les systèmes embarqués).

Analyse variationnelle: "3DVar" : dans les problèmes de plus grande dimension, le choix recommandé est souvent l'analyse variationnelle en raison de sa simplicité d'implémentation et sa robustesse. Elle est appelée "3DVar" en météorologie et océanographie, car alors x est l'état de modèle sur une grille 3D, mais elle s'applique aussi aux problèmes 2D (analyses de surfaces) ou 1D (analyses colonne par colonne). Le principe du 3DVar est d'approximer le minimum de la fonction J en effectuant un nombre limité d'itérations par descente de gradient. Pour cela on effectue la récurrence :

- initialiser l'état du miminiseur par l'ébauche : $x_0 = x_b$
- à l'itération i, l'état est mis à jour par un algorithme de minimisation : $x_{i+1} = x_i + F[J(x_i), \nabla J(x_i)]$ où ∇J est la dérivée (ou *gradient*) de J au point x_i, de telle sorte que $J(x_{i+1}) < J(x_i)$
- on arrête la minimisation, soit lorsqu'un nombre limite d'itérations a été effectué, soit parce qu'une approximation satisfaisante de l'optimum a été atteinte : $|| \nabla J (x_i) || / || \nabla J(x_b)|| < \varepsilon$ où ε est un ratio prédéfini (par

exemple $\varepsilon = 10^{-2}$ signifie que x_i est 100 fois plus proche de l'optimum que x_b, au sens de la métrique impliquée par J)

- on *définit* alors l'analyse par $x_a = x_i$

L'opérateur F le plus simple est une descente de gradient : calcul du α qui minimise $J(x_i + \alpha \nabla J(x_i))$, qui est un polynôme de degré 2 en α. En pratique on utilise des algorithmes plus efficaces qui réutilisent tous les $J(x_i)$ et $\nabla J(x_i)$ calculés depuis le début de la minimisation, et numériquement robustes. Les algorithmes les plus classiques sont le gradient conjugué (optimal si H est parfaitement linéaire), quasi-Newton et BFGS (robustes à la présence de faibles non-linéarités dans H). Plus récemment, les librairies d'IA utilisent notamment l'optimiseur Adam ainsi que le gradient stochastique, adaptés aux fonctions-coût à topologie complexe et aux contraintes mémoire pour des observations nombreuses (pour éviter de les charger en mémoire centrale elles sont traitées par paquets ou *batches*).

D'un point de vue codage, le point clé du 3DVar est le calcul du gradient : $1/2 \nabla J(x) = B^{-1}(x-x_b) - H^T R^{-1} (y -H(x))$. Son avantage est de ne faire intervenir que des produits matrice x vecteur, ce qui permet d'éviter de construire et stocker de grosses matrices, à la différence du calcul matriciel de K, notamment si B^{-1} et R^{-1} sont codés comme des opérateurs linéaires (sans allouer les matrices correspondantes). La contepartie est que le 3DVar nécessite le codage du produit par H^T, appellé **opérateur adjoint** de H (transposée de la différentielle de H). Cela a longtemps été vécu comme un repoussoir pour certaines communautés, puisque cela impliquait de développer à la main les adjoints de codes parfois complexes. Ce problème est en train de disparaître grâce à la popularisation progressive d'outils de différentiation et minimisation automatique via les librairies modernes d'IA (par ex. *pytorch* et *tensorflow*).

Analyse variationnelle "4DVar": il s'agit d'une généralisation du 3DVar au traitement d'un flux quasi-continu d'observations. Jusqu'ici on a considéré l'analyse au sein d'une assimilation séquentielle: l'état x est propagé temporellement par un modèle d'évolution $x_{t2} = M_{t1->t2} (x_{t1})$, et mis à jour par intermittence à chaque fois qu'un paquet d'observations à l'instant t doit être traité : $x_{a,t} = x_{b,t} + K_t (y(t) - Ht(x_t))$. Le 4Dvar consiste à analyser simultanément toutes les observations disponibles durant une fenêtre temporelle $ti \in [t_1,t_2]$, en généralisant le terme de distance aux observations dans la fonction coût J du BLUE :

$$J(x) = (x - x_{b,t}) B_t^{-1} (x-x_{b,t}) + \sum_{t \in t1... t2} (y(t) - H_t (M_{t1->t} (x_t))) R_t^{-1} (y(t) -Ht(M_{t1->t}(x_t))$$

On peut montrer qu'optimiser ce J produit une analyse optimale x_a à t_1 si le modèle M est linéaire (càd linéarisable au voisinage de $M(x_{b,t})$), et la prévision $x(t)=M_{t1->t}(x_a)$ est optimale sur la totalité de l'intervalle $t_1...t_2$. Cette propriété provient du fait que minimiser le J du 4D-Var permet de prendre implicitement en compte les contraintes physiques sur l'évolution de x via l'intégration de l'opérateur M dans J. Pour s'en rendre compte, on peut reprendre la 2e interprétation physique du BLUE de la section précédente, pour une unique observation située à l'instant t_2: dans ce cas, $x_a - x_b = B M^T H^T (H M B M^T H^T + R)^{-1} (y-H(x_b))$, puisque le 4Dvar revient à remplacer H par

HM dans les équations du BLUE. Si on n'a qu'une observation, alors la structure physique de l'incrément est donnée par BM^TH^T (contre BH^T en 3DVar): le point d'observation défini par H^T est traité par le ***modèle adjoint*** M^T, qui représente la propagation d'information dans le fluide, puis lissé par les covariances d'ébauches *B*. Ce processus permet de moduler poids et spatialisation de l'observation de manière physiquement cohérente avec la dynamique du système analysé.

Le prix à payer pour cette sophistication est numérique : outre la nécessité de coder l'adjoint du modèle de prévision M^T, le calcul du 4Dvar implique, à chaque étape de la minimisation, de calculer l'équivalent de 2 prévisions du modèle *M* (une fois pour calculer *HM*, une autre fois pour calculer M^TH^T) pour évaluer J et son gradient.

Analyse ETKF "ensemble transform Kalman Filter" : la construction de la matrice *B* est un problème majeur en analyse. Une classe importante de techniques, appelées "filtres de Kalman d'ensemble", suppose que l'on dispose d'une approximation de *B* de la forme $B = \sum_{i=1...p} dx_{bi}\, dx_{bi}^T$ où les dx_{bi} sont des perturbations de l'état x_b appelées ***ensemble*** de perturbations (on verra en section suivante comment construire un tel ensemble de prévisions). Typiquement, *p* est de l'ordre de 10 à 100. Cette approximation revient à supposer que les erreurs possibles d'ébauche, (x_b-x_t) appartiennent à un sous-espace de dimension *p* engendré par la famille de vecteurs $\{dx_{bi}\}$, tel que $p \prec\prec dim\ x$. Les techniques d'ETKF utilisent la petite dimension de ce sous-espace pour calculer l'équation du BLUE à coût numérique réduit. On part de l'identité $K = BH^T(HBH^T+R)^{-1} = (B^{-1} + H^TR^{-1}H)^{-1}\, H^TR^{-1}$ en remarquant que $A = (B^{-1} + H^TR^{-1}\, H)^{-1}$, donc que $K = A\, H^TR^{-1}$: si l'on dispose d'une approximation de *A*, cela donne une approximation de *K* et donc de x_a. Le coeur de l'ETKF est le calcul de A à partir du B ci-dessous, sous la forme $A = \sum_{i=1...p} dx_{ai}\, dx_{ai}^T$ où les x_{ai} sont calculés en transformant linéairement l'ensemble dx_{bi}, d'où le nom *ensemble transform*. Le calcul de la transformation étant un peu ardu, on n'en donne que les grandes lignes:

- reporter $B = \sum_{i=1...p} dx_{bi}\, dx_{bi}^T$ dans $A = (B^{-1} + H^TR^{-1}\, H)^{-1}$
- factoriser *B* sous la forme $B = C\, C^T$, où *C* est une matrice facile à inverser, et donc $A = C^{-T}\, (I + C^T\, H^T\, R^{-1}\, H\, C)^{-1}\, C^{-1}$
- ne tenir compte dans le terme de droite que d'un nombre limité d'observations pour analyser chaque point de grille du modèle (en supposant, comme dans la techique d'interpolation optimale évoquée précédemment, que les observations n'ont qu'une infuence locale sur les champs analysés). Cela implique que ce terme est une matrice de rang (=dimension) faible et donc que l'on peut la diagonaliser à coût numérique réduit.
- utiliser cette diagonalisation pour former une approximation de *A* de la forme $A = \sum_{i=1...p} x_{ai}\, x_{ai}^T$, ce qui donne *K* puis $x_a = x_b + A\, H^TR^{-1}\, (y - H(x_b))$. Cette équation peut être appliquée à chacune des perturbations dx_{bi} pour calculer un ensemble de perturbations d'analyses $\{dx_a\}$, leur moyenne qui constitue un estimateur optimal de x_a .

L'ensemble $\{dx_a\}$ peut être interprété comme un ensemble d'erreurs possibles d'analyses, utilisé pour calculer l'ensemble $\{dx_b\}$: c'est l'idée du filtre de Kalman d'ensemble, développée dans la section suivante.

Assimilation séquentielle: ensembles et filtres de Kalman

Les section précédentes ont expliqué comment, à partir des données (x_b, B, R, H), on peut traiter un paquet d'observations y pour en déduire l'analyse x_a et son incertitude A. Pour assimiler un long historique d'observations, et en particulier dans un système temps réel où l'on souhaite mettre à jour l'analyse au fur et à mesure de l'arrivée de nouvelles données y, le traitement en bloc de toutes les observations n'est plus possible (cela coûte cher en calcul et stockage, et pose des problèmes de stabilité numérique). On a vu que la qualité de l'analyse dépend crucialement de celle de B, on a donc besoin d'une méthode qui calcule les meilleurs B possibles. Cela nécessite de réfléchir à la manière dont l'opérateur d'analyse s'insère dans un système de prévision. L'approche la plus courante consiste, comme en 4DVar, à s'appuyer sur la disponibilité d'un modèle de simulation numérique (ou de prévision par IA), fournissant à chaque instant ti la meilleure ébauche possible comme prévision issue d'une analyse récente à $ti-1$:

$x_{b,ti} = M_{ti-1->ti}(x_{a,ti-1})$

en alternant cette opération avec celle de l'analyse BLUE :

$x_{a,ti} = x_{b,ti} + K_{ti}\,(\,y(ti) - H_{ti}(x_{b,ti}))$

cela définit par récurrence un filtre digital qui à partir d'une suite d'observations $y(ti)$ produit une séquence d'analyses x_a : c'est ce que l'on appelle l'*assimilation séquentielle*. L'opérateur K_{ti} peut utiliser n'importe laquelle des techniques décrites précédemment, s'il s'agit d'un 4DVar cela signifie que l'on tient compte de la distribution temporelle des observations à l'intérieur de chaque paquet d'observations analysées. Le calcul de K_{ti} utilise les opérateurs H_{ti} et R_{ti}, qui ne dépendent que des observations (et de $x_{b,ti}$ si H est non-linéaire), et la matrice B_{ti}, qui doit être calculée séparément.

L'approche la plus simple suppose que B_{ti} est temporellement constante, on peut alors la modéliser à partir d'informations climatologiques sur les erreurs passées. Les informations les plus importantes sont:

- l'amplitude typique des erreurs d'ébauche. Cela peut s'estimer par exemple en effectuant des **validations croisées** (statistiques des distances entre prévisions et observations non utilisées dans l'assimilation). Pragmatiquement, comme seul le ratio entre variances d'erreur d'ébauche et d'observations compte dans le calcul du BLUE (cf les exemples donnés ci-dessus), si l'on peut supposer qu'une assimilation produit, en régime de croisière, des ébauches de qualité analogue à celle des observations utilisées, alors les variances d'erreurs de B doivent être du même ordre de grandeur que

celles des observations des mêmes paramètres, dans les zones pourvues en observations denses.

- l'échelle spatiale typique des erreurs d'ébauche. Cela peut s'étudier avec des variogrammes de différences, soit entre ébauches et observations, soit entre différentes prévisions. Bien que la géométrie de ces erreurs puisse être localement complexe (notamment en présence de fronts ou d'objets cohérents: tourbillons, nuages convectifs), un opérateur de lissage simple comme la gaussienne peut fournir une bonne base de construction des coefficients de B.
- les relations statistiquement bien vérifiées entre les différentes variables (et donc entre leurs erreurs). Par exemple, les équilibres géostrophiques et hydrostatiques sont approximativement respectés presque partout dans l'atmosphère et l'océan, il est donc intéressant de les inclure comme contraintes statistiques (corrélation inter-variables) des coefficients de *B*.

Dans cette approche, *B* est un modèle statistique d'erreurs qui peut devenir très sophistiqué : cela finit par représenter un travail conséquent, et il faut s'appuyer sur des archives de prévisions passées qui deviennent de plus en plus lourdes lorsque l'on veut complexifier *B*.

Une autre approche consiste à faire calculer la matrice *B* automatiquement en simulant l'évolution des ereurs dans x à partir du modèle de prévision *M*. L'idée est que, si *M* est physiquement réaliste, il sera capable de représenter l'évolution de ces erreurs au fil du temps. Cette hypothèse fonctionne bien pour les phénomènes dont les erreurs de prévision sont dominées par la croissance chaotique des erreurs, du fait d'instabilités hydrodynamiques que *M* sait simuler : instabilités barotrope, barocline, convective, etc... Elle peut s'avérer médiocre si *M* contient des biais importants, ou échoue à simuler des phénomènes importants pour expliquer les erreurs d'ébauche. Elle a fait ses preuves en prévision météorologique et océanographique, au moins jusqu'à des résolutions d'une dizaine de kilomètres.

Une équation classique pour calculer *B* est celle du filtre de Kalman étendu (EKF) :

$B_{ti} = M' A_{ti} M'^{T} + Q$

où M' est le modèle de prévision *M* linéarisé au voisinage de la prévision $x_{b,ti} = M(x_{a,ti-1})$, et *Q* est un terme correctif d'"erreur de modélisation" destiné à représenter l'effet des erreurs de *M* (phénomènes physiques non simulés par *M*). Avec le BLUE, cette équation constitue l'algorithme du filtre de Kalman, qui est un système complet d'assimilation séquentiel. Il ne présente qu'un intérêt historique car con coût est très élevé (il nécessite de calculer de l'ordre de *dim x* prévisions de *M* par cycle de prévision) et il est numériquement instable (*K* peut tendre vers zéro ou vers l'infini si les termes d'erreur ne sont pas correctement spécifiés).

Une technique plus commode pour produire *B* est la **prévision d'ensemble**, qui échantillonne les lois de probabilités impliquées par *A* et *B*. Exemple de mise en pratique :

- ayant calculé A avec l'équation du BLUE, tirer aléatoirement p vecteurs $dx_{a,i}$ dont l'autocovariance est A. Par exemple, si on factorise $A=LL^T$, en tirant aléatoirement une famille de vecteurs $v_i \sim N(0,1)$, la famille de vecteur $w_i = L v_i$ a cette propriété. On peut utiliser des méthodes plus astucieuses, comme l'algorithme de Lanczos, pour échantillonner le mieux possible les aspects les plus importants de A avec un nombre de membres p limité.
- calculer l'évolution de cet ensemble $\{dx_a\}$: $dx_{b,i}= M(x_a + dx_a) - M(x_a)$
- construire comme approximation de $B_p = \sum_{i=1} dx_{b,i}\ dx_{b,i}^T$; cette approximation n'étant pas définie positive si $p < dim x$, il faut soit la mélanger avec un B climatologique B_{clim} pour calculer le BLUE: $B = \alpha B_{clim} + (1-\alpha) B_p$ (où α est un coefficient d'hybridation empirique), ce qui conduit à l'algorithme du **filtre de Kalman d'ensemble** (EnKF), soit utiliser une approximation du BLUE adaptée à un B ensembliste, comme le LETKF décrit précédemment. Le LETKF produit directement un ensemble de $\{dx_a\}$ ce qui est un avantage pour initialiser des prévisions d'ensemble.

Une autre approche consiste à court-circuiter l'étape d'échantillonnage de A en calculant non plus un seul BLUE par cycle, mais un ensemble complet de systèmes d'assimilation parallèles, comme en *ensemble data assimilation* (EDA) :

- tirer aléatoirement p jeux de valeurs observées en ajoutant à y des perturbations qui suivent la loi de probabilité des observations $N(0,R)$
- former p analyses $x_{a,i}$ en calculant p analyses BLUE utilisant chacune un de ces jeux d'observations: l'ensemble $\{x_a\}$ échantillonne alors la distribution de probabilités de x_a .
- effectuer la prévision d'ensemble de p membres $x_{b,i} = M(x_{a,i}\ i)$
- la matrice B est approximée par la covariance empirique $\sum_i dx_{b,i}\ dx_{b,i}^T$ où $dx_{b,i} = x_{b,i} - E(x_b)$, où E étant l'opérateur de moyenne de l'ensemble d'ébauches.

Cette technique, numériquement coûteuse, présente l'intérêt de tenir compte des non-linéarités de M, et de pouvoir se coder modulairement au-dessus de n'importe quel algorithme d'analyse. En particulier, son interfaçage avec les algorithmes 3DVar et 4DVar (via des techniques d'hybridation de B) conduit aux algorithmes **3DenVar** et **4DenVar**, qui permettent d'optimiser la prise en compte des non-linéarités et des phénomènes physiques dans le calcul de l'analyse.

Conclusions

La richesse des options de calcul du BLUE de B a donné naissance à des nombreux algorithmes, parce qu'identifier le meilleur n'est pas toujours évident, mais aussi pour des raisons historiques propres à chaque centre de calcul. Aujourd'hui existent plusieurs "chapelles" d'assimilation de données, la distinction la plus importante étant entre ceux qui privilégient l'approche ETKF, et ceux qui, disposant de codes adjoints, ont investi dans le 3DVar et ses dérivés. Le caractère plus ou moins non-linéaire des

phénomènes visés, ainsi que l'importance donnée à l'assimilation des données satellitaires, influent aussi sur ces choix algorithmiques.

Bien qu'il y ait de grandes similitudes entre les techniques d'IA et d'assimilation de données, il n'existe pas à ce jour (en 2024) de modèle IA capable de rivaliser avec les approches de type BLUE évoquées ici. Il est fort possible que cela change à l'avenir, tant les enjeux sont importants sur les coûts calcul et sur l'influence de l'assimilation sur la qualité des prévisions. Il est probable que des modèles de prévision par IA vont progressivement remplacer les modèles à base physique (comme l'opérateur M) dans les algorithmes de type 4Dvar et EDA, cela en réduira considérablement les coûts. Il reste à voir dans quelle mesure cela permettra d'analyser et de prévoir des phénomènes qui n'auront pas été présents dans leurs données d'apprentissage.

A l'origine, l'assimilation de données s'est développée pour initialiser les modèles de prévision en temps réel de l'atmosphère (et, plus récemment, de l'océan, l'hydrologie, les sols, etc). Mais son application pour le suivi du climat prend de plus en plus d'importance : les **réanalyses** des dernières décennies (voire siècles) à partir d'algorithmes modernes, et d'observations retraitées, ont permis de produire des archives historiques de variables d'intérêt climatique, cruciales à la fois par leur précision (grâce à la fusion de nombreux types d'observations, notamment satellitaires) et par leur facilité d'accès (qui a rendu ces informations accessibles pour des études interdisciplinaires, notamment pour comprendre l'effet du changement climatique très au-delà de la communauté météo-océano). Les réanalyses des grands centres comme le NCEP des USA et le programme Copernicus de l'Union Européenne jouent un rôle clé dans ce domaine. Outre leur rôle de rétrospective historique, ils mettent à jour en temps quasi-réel leurs réanalyses, ce qui fournit un suivi en continu de l'évolution du climat.

L'assimilation de données comporte cependant des problèmes non résolus. D'une part, de nombreuses variables essentielles pour le suivi de l'environnement sont insuffisamment observées pour que l'on sache en produire des analyses précises. Par exemple, les nuages, les flux de rayonnement et précipitations issus des analyses atmosphériques comportent des biais qui reflètent ceux des modèles assimilateurs, car les systèmes d'analyses ne parviennent pas à les contraindre précisément par les observations.

Il n'existe pas à ce jour de technique satisfaisante pour optimiser le déploiement de nouvelles observations capables de réduire ces problèmes. Lorsqu'une prévision échoue - par exemple sur des cas de cyclone, de tempête ou d'inondation - il est souvent impossible de comprendre si cet échec provient de déficiences du modèle de prévision, ou d'imprécisions sur son analyse de départ.

Enfin, les prévisions de phénomènes très chaotiques, comme les nuages, les orages ou les précipitations, sont encore notoirement imprécises même à très courte échéance, malgré le disponibilité d'abondantes observations, ce qui indique que l'on ne sait pas encore interpréter l'information observée de manière cohérente avec la physique atmosphérique.