2. Assimilation de données

2.1. **Introduction.** L'assimilation de données est l'ensemble des techniques qui permettent de combiner, de façon optimale (en un sens à définir), l'information mathématique contenue dans les équations modélisant le phénomène, et l'information physique provenant des observations, en vue de reconstituer l'état du système.

Dans un cadre géophysique (e.g. météorologie ou océanographie), grâce notamment aux puissants moyens de calcul maintenant disponibles, la modélisation en vue de la prévision a connu d'importants développements ces dernières décennies. Les fluides géophysiques : l'air,l'eau atmosphérique, océanique ou terrestre sont régis par les équations générales de la mécanique des fluides : conservation de masse, de l'énergie, loi de comportement, toutefois certaines spécificités doivent être prises en compte.

2.2. **Spécificités en géophysique.** Les processus géophysiques sont fondamentalement non-linéaires d'abord en raison de leur l'aspect fluide et aussi de certains processus physiques propres comme les transferts radiatifs. Il y a donc des interactions entre les différentes échelles en temps et en espace. La résolution numérique des équations impose des discrétisations donc des troncatures dans les échelles. Cependant les phénomènes de taille inférieure à la troncature peuvent correspondre à de très importants flux d'énergie dont il faudra tenir compte dans la modélisation. A titre d'exemple un nuage de type cumulo-nimbus a une taille caractéristique de l'ordre de 10km. dans toutes les directions, un modèle de circulation générale a des mailles de l'ordre de 50 à 100 km. Or l'énergie thermique (chaleur latente) d'un tel nuage est considérable, de même les vitesses verticales caractéristiques d'un modèle de circulation générale sont de l'ordre du centimètre ou du décimètre par seconde. Dans un nuage on a pu mesurer des vitesses verticales de l'ordre de 100 mètres par seconde. Il convient donc de représenter ces flux d'énergie dans les équations discrétisées par l'adjonction de termes supplémentaires. Nécessairement ces termes, dits de paramétrisation, inclueront des cœfficients empiriques non accessibles à la mesure expérimentale. Néanmoins il faudra estimer ces grandeurs à partir de données d'observation.

Mais les seules équations de la dynamique des fluides ne sont pas suffisantes pour faire une prévision, il faut en outre une condition initiale et des onditions aux limites. Dans la plupart des cas les fluides géophysiques n'ont pas de frontières naturelles, pas plus qu'une condition initiale, comme une solution stationnaire, ne s'impose naturellement. Là aussi ces termes de bord devront être estimés à partir de donnés d'observation.

On voit que la modélisation devra tenir compte des données d'observation. Or, a priori, données et modèles ne sont pas nécessairement compatibles : une même donnée de vent ou de température pourra être utilisée dans un modèle de circulation générale tout aussi bien que dans un modèle local d'écoulement. Selon le contexte (le modèle) la mesure recevra une interprètation différente.

2.3. **Analyse.** L'analyse est le résultat de l'assimilation de données. Si le système est sur-déterminé par les observations, alors l'étape d'analyse se résume essentiellement à un problème d'interpolation. Mais dans la plupart des cas, le système est sous-déterminé, car les données sont éparses et pas forcément directement reliées aux variables du modèle (ex : météo - données satellites).

De manière à rendre le problème bien posé, il est nécessaire de rajouter une information supplémentaire, par exemple sur une estimation a priori de l'état du système. Dans le cadre météo, il peut s'agir d'une information de climatologie, mais il peut aussi s'agir d'un état trivial (constant, ou nul) d'un point de vue mathématique.

Il existe essentiellement deux grandes catégories d'assimilation de données. Les méthodes séquentielles ne tiennent compte que des observations disponibles avant l'instant d'analyse. C'est typiquement le cas pour de l'assimilation en temps réel. Dans les méthodes variationnelles, on cherchera à identifier l'état du système à un instant, en utilisant des observations passées, présentes ou futures.

2.4. **Estimation de paramètres : vecteur d'état, contrôle, et observations.**

2.4.1. *Vecteur d'état.* La première étape de la formulation mathématique du problème inverse est de définir le cadre de travail. On suppose que le système physique considéré peut être représenté par un vecteur x, qu'on appelle le vecteur d'état. L'espace dans lequel vit ce vecteur, l'espace des états, peut être de dimension infinie ou finie, suivant que le problème a été discrétisé ou non (ou projeté sur une base).

Nous aurons besoin dans la suite de l'état réel (ou état vrai) du système, x_t, qui est la meilleure représentation possible de la réalité dans l'espace des états. On notera x_b l'ébauche de l'état du système, qui pourra servir pour régulariser le problème et le rendre bien posé. Et enfin, l'analyse x_a, qui est l'état obtenu après assimilation.

Dans le cadre de l'assimilation de données en géophysique, on se placera en dimension finie en supposant que l'ensemble du problème a été discrétisé en espace : le phénomène physique est discrétisé sur une grille de points en espace. On notera n la dimension de l'espace des états. Autrement dit, $x \in \mathbb{R}^n$.

2.4.2. *Vecteur de contrôle.* Le vecteur de contrôle correspond aux variables que l'on souhaite identifier, ou de façon équivalente sur lesquelles on peut influer (ou modifier les valeurs), afin de faire coller la sortie du modèle avec les observations.

Dans le cas d'un système météo ou océano, il s'agit généralement de l'état complet du système à un instant donné. Dans certains cas, il peut s'avérer suffisant d'identifier l'état seulement dans une partie du domaine, afin de réduire la dimension de l'espace de contrôle (et donc du problème).

On cherche généralement l'état analysé sous la forme suivante :

$$x_a = x_b + \delta x,$$

où la correction (ou l'incrément) δx est telle que l'état analysé x_a est aussi proche que possible de l'état réel du système x_t. Au lieu de chercher l'analyse, on peut chercher de façon équivalente la correction δx par un simple changement de variable.

Sauf mention contraire, on cherchera donc le contrôle (x_a ou δx) dans l'espace de contrôle $\mathbb{R}^n$.

2.4.3. *Observations.* Comme pour le vecteur d'état ou de contrôle, on suppose que les différentes observations sont regroupées dans un vecteur d'observations y. Ce vecteur vit dans un espace dit d'observations, qui est généralement distinct de l'espace des états (ou de celui du contrôle). En effet, d'un point de vue physique, il est généralement impossible d'observer complètement l'état du système. Et souvent, les mesures portent sur des quantités physiques différentes de celles de l'état.

Afin de pouvoir malgré tout comparer les observations avec l'état du système, il est nécessaire de disposer d'un opérateur H, que l'on appelle opérateur d'observation, allant de l'espace des états dans l'espace d'observations. Pour un état x donné, $H(x)$ appartient à l'espace des observations.

En dimension finie, dans un cadre linéaire, on peut imaginer que chaque ligne de la matrice H représente un opérateur d'interpolation entre les points de grille du modèle discret et les points d'observation.

On notera p la dimension de l'espace des observations. On considèrera que $y \in \mathbb{R}^p$. H est donc un opérateur (pas forcément linéaire) de $\mathbb{R}^n$ dans $\mathbb{R}^p$.

2.4.4. *Modélisation des erreurs.* Il existe un certain nombre d'incertitudes, que ce soit dans l'ébauche de l'état du système, dans le processus d'observation, ou dans l'étape d'analyse. Il faudrait d'un point de vue stochastique représenter tous ces phénomènes par des variables aléatoires suivant des lois, et considérer leur fonction de densité de probabilité (ou fonction de répartition).

Dans la suite, on pourra considérer les trois erreurs suivantes :

- erreur d'ébauche : c'est $\varepsilon_b = x_b - x_t$, l'écart entre l'ébauche et l'état réel du système. Si cette erreur est nulle, alors l'analyse est triviale : on conserve l'ébauche, sans tenir compte des observations.
- erreurs d'observation : $\varepsilon_o = y - H(x_t)$, c'est l'écart entre les observations et l'état correspondant dans l'espace des observations à la réalité. Si cette erreur est nulle, alors les observations sont un reflet exact de la réalité.
- erreur d'analyse : c'est $\varepsilon_a = x_a - x_t$, l'écart entre l'analyse et l'état réel du système. On espère qu'après assimilation, l'erreur d'analyse est plus petite que l'erreur sur l'ébauche.

Pour chacune de ces erreurs, la valeur moyenne $\bar{\varepsilon}$ est appelé biais. C'est le signe d'une dérive dans le modèle (mauvaise modélisation), ou d'un biais systématique dans les observations (problème dans l'appareil de mesure).

On peut représenter les erreurs à l'aide de leurs covariances. La matrice de covariance d'une erreur ε est l'espérance mathématique de la matrice $(\varepsilon - \bar{\varepsilon})(\varepsilon - \bar{\varepsilon})^T$. On notera $B \in \mathcal{M}_n(\mathbb{R})$ la matrice de covariance d'erreurs sur l'ébauche, et $R \in \mathcal{M}_p(\mathbb{R})$ celle des erreurs d'observation.