データ同化の考え方とその方法

[講演: 中野慎也 (統計数理研究所)]

電気通信大学 細川 敬祐

1 はじめに

近年, 地球惑星科学の研究においてデータ同化という手法が広く用いられるようになってきた. この文章は, MTI 研究領域において近年行われているデータ同化の試みを紹介するのではなく, データ同化の原理・方法論に関してより基礎的な知識を概観することを目的とする. 具体的には,

- データ同化とはそもそも何なのか?
- どのような問題を解くのか?
- 実際にどのようにして問題を解くのか?

といった基礎的な部分を解説する. 個々の適用事例に関しては, この文章で概説した基礎知識を元に個別に調べていただきたい.

2 データ同化とはなんなのか?

2.1 データ同化の概念

データ同化は, data assimilation の訳語である. Assimilation という言葉が, “ある民族が移民として他の民族に同化する” ことを表現する際に用いられることからも分かるように, データ同化は「数値シミュレーションに実測データを埋め込み, 馴染ませること」を意味する. 簡単に言えば数値シミュレーションに実測データを取り入れる手法のことである.

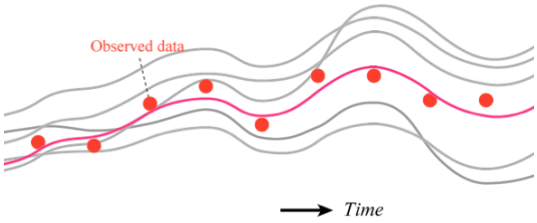

図 1: データ同化によるアプローチの模式図

数値シミュレーションを行うためには, 初期条件, 境界条件, パラメータなどを与える必要がある. しかし, それらの値は往々にして正確に決められない場合が多い. したがって, 初期条件・境界条件などをどのように与えるかによって, 数値シミュレーションは様々なシナリオを導きうる. その様子を図 1 に模式的に示す. 横軸は時間を示し, 同じ数値シミュレーションを用いても, 初期条件・境界条件の与え方によって様々なシナリオ (たくさんの灰色のライン) が結果として現れることが示されている. この図には, 赤い丸で観測データが示されているが, データ同化はこれらのデータを数値シミュレーションに埋め込み, 馴染ませていくことによって, 実際の観測データをうまく説明する, より尤もらしい推定 (ピンクのライン) を探しだすことを目指している.

平成 21 年度 MTI 研究会 サイエンスセッション
ⓒ Mesosphere Thermosphere Ionosphere (MTI) Research Group, Japan

2.2 データ同化の目的

データ同化には大きく分けて 2 つの目的がある. そのひとつ目は「実測データを用いて数値シミュレーションモデルの精度・性能を改善する」ことである. まず, シミュレーションにデータを同化させることによって, 適切な初期条件の設定が可能になる. 例えば, 過去のデータを用いてデータ同化を行い, 現在の状態に関して良い推定値を得ることによって, その初期条件に基づいて将来の予測を精度良く行うことできる. また, 過去からシミュレーションを走らせる場合にも, これまでのデータを全て同化させることによって, より良い初期条件を用いて計算をスタートさせることが可能になる. データ同化による尤もらしい初期条件の設定は, シミュレーションを現在からスタートさせる場合, 過去からスタートさせる場合の双方について有効である. また, 初期条件に限らず, 境界条件や, シミュレーションを走らせる際に “えいやっ”（地球物理学の分野で使われる常套句のひとつ）と与えられているパラメータに関しても, データ同化によってより尤もらしい値を推定し, シミュレーションに用いることができる. 加えて, シミュレーションと観測が合わない場合などにも, その原因となる箇所を明らかにするために用いることができると考えられる.

データ同化の目的のふたつ目は「物理法則を表現するシミュレーションモデルを用いることで, 観測の不足を補ったり観測誤差を修正したりする」ことである. 一般に, 地球物理学の分野においては, 時間的・空間的に均質な観測データを得ることは非常に困難である. 一方, 数値シミュレーションは, 時間的・空間的に均質なデータを得ることができる反面, そこから導き出される物理量は, 我々が把握している物理過程（物理法則）のみによって支配されているため, 実測値を完全に再現することはできない. データ同化は, データを数値シミュレーション埋め込み, 馴染ませることができるため, 観測の得られない時間・場所における物理量をより尤もらしく推定することができ, 時間的・空間的に均質なデータの生成を可能にする.

2.3 データ同化の用途

データ同化の用途として最も実用的であるのは,「気象予報・予測」である. 気象システムは, 初期値鋭敏性を持つが, 初期値を精度良く決めることができれば決定論的に精度の良い推定値を得ることができる. 現在だけでなく過去のデータも活用して初期値をより尤もらしくすれば, 予測性能も上がっていくと考えられる. 一方, 気象の分野でなじみ深いものとして「再解析データ」がある. 再解析データの生成にはデータ同化が用いられており, シミュレーションに観測データを統合することで, 時間的・空間的に一貫性があり, 均質なデータセットが得られている. 生観測データには必然的に存在する時間的・空間的な偏りを気にすることなくデータを利用することができる. このように, データ同化によるアプローチは汎用性が高く, 気象学・海洋物理学・水文学など, 様々な分野で応用されている. 電離圏・磁気圏などの超高層大気分野においてもいくつかの応用例がある.

2.4 データ同化に必要なもの

データ同化を実際に行うためには以下の 4 つが必要となる.

- 数値シミュレーションモデル
- 観測データ
- 統計科学の知見
- 高性能な計算機

ここで, 数値シミュレーション（もしくはモデル）とデータの両方を使うということがデータ同化の本質であるということに注意したい. 物理法則に基づいたいわゆる数値シミュレーションではなく, 特に根拠もなく適当に作ったモデルで時間発展を記述してもデータ同化で使われる手法自体を適用することはできるが, 物理法則に基づいた数値シミュレーションを使うことで, 観測から得られる知見だけでなく, 物理学の知見も含めた色々な使える知識を投入することが可能になるのである.

2.5 データ同化手法のいろいろ

ひとくちにデータ同化といっても, 以下に挙げるように, 様々な手法がある. ここで挙げるうちの幾つかについては, 後にその詳細を述べる.

- 簡便な方法

  直接挿入

  ナッジング

- 3 次元データ同化

  最適内挿法 Optimal interpolation: OI

  3 次元変分法（3D-VAR）

- 4 次元データ同化

  逐次データ同化
  - カルマンフィルター
  - アンサンブルカルマンフィルター

  4 次元変分法（4D-VAR）

簡便な方法として挙げられている直接挿入と呼ばれるアプローチでは, 観測が得られている数値シミュレーショングリッドに観測データを直接挿入する. また, ナッジングという手法では, 観測が得られているグリッドにおいて, シミュレーションの値を観測値に少しだけ近づけるという処理を行う. 3 次元データ同化は, もう少し複雑な処理を行ってデータをシミュレーションに馴染ませるが, 過去に得られた情報を用いることはしない. それに対して 4 次元データ同化では, 過去からの様々な情報を履歴として加味し, データをシミュレーションに入れ込んでいく. 言うまでもないが, ここで挙げた手法は, 下にいくほど手間がかかる.

3 どのような問題を解くのか?

ここでは, データ同化のプロセスを構成する数値シミュレーションと観測データの間のリンクをどのように定式化し, どのように問題を解いていくのかについて述べる. まず, シミュレーション

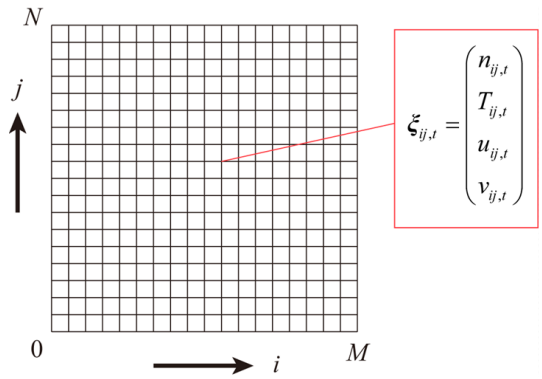

図 2: 数値シミュレーションの各空間グリッドにおける物理量のベクトル化の様子

コードで扱っている全ての変数の時刻 t における値を, 一つのベクトルにまとめる形で $\mathbf{x}_t$ とおく. 例えば, ある時刻 t, あるグリッド (i, j) における全変数の値をまとめたベクトルを $\xi_{ij,t}$ とするならば（図 2 参照）, 全グリッドのシミュレーション変数をまとめて, 以下のようにひとつのベクトルで書くことができる.

$$\mathbf{x}_t = \begin{pmatrix} \xi_{00,t} \\ \vdots \\ \xi_{M0,t} \\ \xi_{01,t} \\ \vdots \\ \xi_{M1,t} \\ \vdots \\ \vdots \\ \xi_{0N,t} \\ \vdots \\ \xi_{MN,t} \end{pmatrix} \tag{1}$$

このようにシミュレーション変数の定義を行うと, シミュレーションによる時間発展を以下のように抽象化して記述することが可能になる.

$$\mathbf{x}_t = f_t(\mathbf{x}_{t-1}) + \mathbf{v}_t \tag{2}$$

ここでは, 時刻 $t-1$ から時刻 t までシミュレーションを走らせたことによる変数の時間発展を演

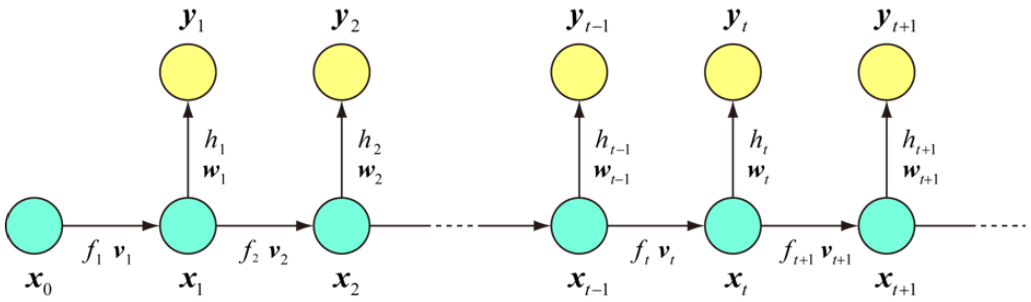

図 3: 数値シミュレーションの時間発展および観測データ取得プロセスの定式化の様子

算子 f_t で表している．数値シミュレーションは本来, 決定論的に時間発展を記述していくことができるが, ここで, $\mathbf{x}_t = f_t(\mathbf{x}_{t-1})$ とせず, $\mathbf{v}_t$ という遊びを含めてあるのは, 数値シミュレーション自体が系の時間発展を “正確に” 記述できない可能性が存在するためである．

ついで, 数値シミュレーションに同化させるべき観測データを取得する過程を定式化する． まず, 用いる観測データを一つのベクトルにまとめて $\mathbf{y}_t$ とする．ここで, 実際の観測を模倣して, シミュレーションの世界で仮想的にデータを取得する過程を演算子 h_t で表すと, シミュレーション中の変数と観測データとの関係は,

$$\mathbf{y}_t = h_t(\mathbf{x}_t) + \mathbf{w}_t \tag{3}$$

と書くことができる．ここで現れる観測演算子 h_t をどのように定義するか問題であるが, 例えば, シミュレーション変数 $\mathbf{x}_t$ のうち, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ に対応する物理量のみが直接観測できるとした場合,

$$H_t = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix} \tag{4}$$

のような行列を介して, $h_t(\mathbf{x}_t) = H_t\mathbf{x}_t$ というように記述することができる．ただし, シミュレーショングリッドの場所でエグザクトに観測データが得られるとは限らないため, 周囲のグリッドで重みつき平均を取って直接観測と比較するような場合もある．全電子数（Total Electron Content: TEC）のような積分量を用いる場合も, シミュレーションで扱う物理量と観測との関係が線型であるため, 基本的には行列の形で表現できるが, H_t はもっと複雑な形になる．ここで, 数値シミュレーション中の変数と観測データとの関係を, $\mathbf{y}_t = h_t(\mathbf{x}_t)$ としないのは, シミュレーションによる予測値と実測とを完璧に一致させるのがまず不可能だからである．その理由として, データには観測誤差が含まれていることや, 数値モデルが（空間・時間分解能や離散化近似などの影響で）現実の世界を完璧には表現できないことがあげられる．以上で述べた数値シミュレーションの時間発展および観測データ取得プロセスの定式化の様子を図 3 に模式的に示しておく．

これで, 式 (2), (3) を用いることで, 各ステップにおけるシミュレーション変数と観測データの間のリンクをとることができ, データ同化で登場する全ての量の間の関連を記述することが可能になった．式 (2), (3) のような関係が成り立ち, 更に各時刻の観測データ $\mathbf{y}_t$ が与えられたという条件のもとで, 各時刻の $\mathbf{x}_t$ を推定するのがデータ同化である．実際の推定プロセスでは, $||\mathbf{x}_t - f_t(\mathbf{x}_{t-1})||$ および $||\mathbf{y}_t - h_t(\mathbf{x}_t)||$ を小さくしていくように $\mathbf{x}_t$ を決定していく．このような問題を解くには色々なやり方があり, それぞれで, どのくらい頑張るか, どこでどのくらい手を抜くかが違ってくることになる．

4 どのようにして問題を解くのか?

簡便な方法として直接挿入やナッジングという手法があることを述べたが, これらの手法に関しては, 単にデータを数値シミュレーションに挿入するだけであるので解説は省略し, ここでは 3 次元データ同化からスタートし, 4 次元データ同化のあらましまでの説明を行う.

4.1 3 次元データ同化

式 (2), (3) に基づいてデータ同化を行っていく際に, より扱いが大変なのは $\mathbf{x}_t = f_t(\mathbf{x}_{t-1}) + \mathbf{v}_t$ の方である. つまり, $||\mathbf{x}_t - f_t(\mathbf{x}_{t-1})||$ を小さくするという要請を真面目に推定に取り入れるのは非常に難しい. なので, まずは $\mathbf{y}_t = h_t(\mathbf{x}_t) + \mathbf{w}_t$ のほうだけを考え, $||\mathbf{y}_t - h_t(\mathbf{x}_t)||$ を小さくするという方向性の元に推定に取り組んでいくのが 3 次元データ同化の基本的な考え方である. つまり, 異なるステップ間の繋がりはあまり重要視せず, 推定値と観測値の差が小さくなるように推定を行う.

4.1.1 最小二乗法による推定

推定値と観測値の間の差を小さくするように推定を行うための最もシンプルな方法は, 最小二乗法である. ここでは, 簡単のため, しばらくの間は, 観測演算子が線型の場合（$h_t(\mathbf{x}_t) = H_t\mathbf{x}_t$ と書ける場合）について考える. 最小二乗法を用いて, $||\mathbf{y}_t - H_t\mathbf{x}_t||^2$ を最小にするように数値シミュレーションの物理量 $\mathbf{x}_t$ を決定する. 観測が十分にある場合（$\dim \mathbf{y}_t > \dim \mathbf{x}_t$）であればこの問題は比較的簡単に解くことができ,

$$\mathbf{x}_{t,est} = (H_t^T H_t)^{-1} H_t^T \mathbf{y}_t \tag{5}$$

となる. しかし, このような簡単なアプローチでは, 観測が空間的に偏りなく得られていないとき, 全く見当外れの推定値が得られてしまう場合があり, うまくいかないことがある. 実際問題として, 観測が量的に十分かつ空間的に偏りなく得られることは滅多にない. このため, 簡単に式 (5) を用いて解を直接推定することは稀で, 大抵の場合は, 解が, 大体 $\mathbf{x}_{t,b}$ のあたりだろうと予想をつけ, 予想値に近い部分で解を探すことで偏りのある観測を用いたデータ同化を行うことになる. 実際には,

$$||\mathbf{x}_t - \mathbf{x}_{t,b}||^2 + \alpha^2||\mathbf{y}_t - H_t\mathbf{x}_t||^2 \tag{6}$$

が最小になるように推定を行う. ここで, $\mathbf{x}_{t,b}$ には前ステップからシミュレーションを走らせた結果を使ってもよいし, 過去のデータの平均値のような経験的な値を使うこともある. 観測演算子が線形の場合, 解は

$$\begin{aligned}\mathbf{x}_{t,est} &= \mathbf{x}_{t,b} \\ &\quad + H_t^T(H_t H_t^T + \alpha^2 I)^{-1}(\mathbf{y}_t - H_t\mathbf{x}_{t,b})\end{aligned} \tag{7}$$

で与えられる. ただし, この式で $\mathbf{y}_t$ で表されている観測データベクトルは, 様々な種類のデータをまとめてひとつのベクトルで表現したものであり, 各成分ごとにばらつきが異なる. よって, 適宜調整しながら推定を行っていく必要がある.

4.1.2 ベイズの定理を用いた一般化

上で述べた, この辺りだろうと予想をつけて実データで修正をかけるという考え方は, ベイズの定理を用いて表現しなおすことができる. 図 4 に模式的に示されているように, ベイズの定理は, 予想の分布をまず与えてやり（青い分布）, データが得られたときにデータの情報を使って推定をアップデートしてやる（ピンクの分布）というプロセスを表現したものである. 観測を用いて推定を更新した場合, 観測がない時点の予想よりも確率分布の分散が小さくなり, 予想が絞られていることが分かる. ベイズの定理は以下のように表すことができる.

$$p(x_t|y_t) = \frac{p(y_t|x_t)p(x_t)}{\int p(y_t|x_t)p(x_t)dx_t} \tag{8}$$

ベイズの定理の左辺 $p(x_t|y_t)$ は「観測結果が y_t だったとしたら, x_t の値はこのあたりだろう」と

$$p(x_t \mid y_t) = \frac{p(y_t \mid x_t)p(x_t)}{\int p(y_t \mid x_t)p(x_t)\,dx_t}$$

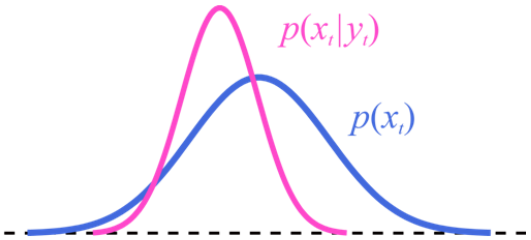

図 4: ベイズの定理

いう条件付き確率を示す．また, 右辺に出てくる $p(x_t)$ は,「観測のない時点での x_t はこのあたりだろう」という確率を表す．これらの分布は図 4 に模式的に示されている．

ベイズの定理の理解のために例を挙げる．検出率 70 % のインフルエンザ検査で, 結果が陰性であったとして, インフルエンザにかかっている確率をベイズの定理を用いて推定してみる．ここでは, インフルエンザにかかっている事象を $x_t = 1$, かかっていない事象を $x_t = 0$ とする．また, 検査で陽性であるという事象を $y_t = 1$, 陰性であるという事象を $y_t = 0$ とする．まず, ベイズの定理の右辺分子について考える．検査前の確率を五分五分と設定すると, 検査前にインフルエンザにかかっている確率 $p(x_t = 1)$ は 0.5 となる．またインフルエンザにかかっているとして検査でそれが陰性であると検出される確率 $p(y_t = 0|x_t = 1)$ は 0.3 となる．次いで, 分母について考えると, 全ての場合（インフルエンザにかかっている場合とかかっていない場合）について積分するため分母は, $p(y_t = 0|x_t = 0)p(x_t = 0) = 1.0 \times 0.5 = 0.5$（インフルエンザにかかっていない場合）と $p(y_t = 0|x_t = 1)p(x_t = 1) = 0.3 \times 0.5 = 0.15$（インフルエンザにかかっている場合）を足し合わせたものになる． 最終的に, 検査結果が陰性だった場合にインフルエンザにかかっている確率 $p(y_t = 0|x_t = 1)$ は $0.3 \times 0.5/(0.5 + 0.15) = 0.23$ と推定することができる．勿論, 検査前の確率をどう設定するかによって, 推定値は変わる．

3 次元データ同化に話を戻し, $p(\mathbf{x}_t)$, $p(\mathbf{y}_t|\mathbf{x}_t)$ が以下のような正規分布に従うと仮定すると,

$$p(\mathbf{x}_t) = \frac{1}{\sqrt{(2\pi)^k|V|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_t - \mathbf{x}_{t,b})^T V^{-1}(\mathbf{x}_t - \mathbf{x}_{t,b})\Big) \tag{9}$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = \frac{1}{\sqrt{(2\pi)^k|R|}} \exp\Big(-\frac{1}{2}(\mathbf{y}_t - H_t\mathbf{x}_t)^T R^{-1}(\mathbf{y}_t - H_t\mathbf{x}_t)\Big) \tag{10}$$

ベイズの定理より,

$$\frac{1}{2}(\mathbf{x}_t - \mathbf{x}_{t,b})^T V^{-1} \ (\mathbf{x}_t - \mathbf{x}_{t,b}) + \frac{1}{2}(\mathbf{y}_t - H_t\mathbf{x}_t)^T R^{-1}(\mathbf{y}_t - H_t\mathbf{x}_t) \tag{11}$$

を最小にする $\mathbf{x}_t$ が, $p(\mathbf{x}_t|\mathbf{y}_t)$ を最大化する $\mathbf{x}_t$ となる．このようにして, ベイズの定理を用いることで 3 次元データ同化のより一般的な形を得ることができる．この解を計算すると,

$$\bar{\mathbf{x}}_{t,est} = \mathbf{x}_{t,b} + VH_t^T(H_tVH_t^T + R)^{-1}(\mathbf{y}_t - H_t\mathbf{x}_{t,b}) \tag{12}$$

のようになる．最適内挿法（Optimal Interpolation: OI）と呼ばれる方法では, この式に基づいてデータを数値シミュレーションに埋め込んでいく．

一方, 式 (11) の最小化を, 共役勾配法や準ニュートン法のような反復法で行うこともできる．これが 3 次元変分法（3D-VAR）と呼ばれる方法に対応する．この方法は, 観測演算子が非線形な場合にも適用が可能であるという特徴がある．再解析データの生成などには, 今でも使われている方法である．

4.2 4 次元データ同化

3 次元データ同化では時間の繋がりをあまり考えなかったが, 4 次元データ同化では時間の繋が

りをきちんと含めた形でデータ同化を行う．何故，時間の繋がりを考慮するかというと，予測の分布，ベイズの定理でいうところの $p(x_t)$，をきちんと把握したいからである．式 (11) に出てくるパラメータのうち，$\mathbf{x}_{t,b}$ は前のステップからのシミュレーションから一応出せるが，V，R を決めるのは難しい．特に V には，本来，前のステップまでのデータの性質や扱っている系の性質などを踏まえ，$\mathbf{x}_{t,b}$ の各成分にそれぞれどのくらい自信があるか，$\mathbf{x}_t$ の異なる成分間にどのくらいの関連があるかといった情報が入っていてよい．3 次元データ同化では，$p(\mathbf{x}_t)$ の分布の形を正規分布として仮定し，V は適当に与えていた．4 次元データ同化では，前のステップまでの情報とシミュレーションモデルを用いて V を見積もり，前のステップまでの情報を踏まえた $\mathbf{x}_t$ の分布 $p(\mathbf{x}_t|\mathbf{y}_0, \ldots \mathbf{y}_{t-1})$ を計算し，これを $p(\mathbf{x}_t)$ の代わりに使う．そうすることで，その時刻の観測を取り入れた $\mathbf{x}_t$ の分布も改善し，$\mathbf{x}_t$ の推定精度もよくなることが期待される．改善された $\mathbf{x}_t$ の推定結果を次のステップの予測，推定にも反映させるというように連鎖させることで，全時間ステップでの $\mathbf{x}_t$ の推定の改善を図ることが可能になる．

ここでカルマンフィルタというアルゴリズムを用いることになる．カルマンフィルタは $\mathbf{x}_t = F_t\mathbf{x}_{t-1} + \mathbf{v}_t$ という線型システムのもとで，1 つ前のステップの $\mathbf{x}_{t-1}$ の平均，および分散共分散行列が与えられていたときの $p(\mathbf{x}_t|\mathbf{y}_t)$ を計算するアルゴリズムである．但し，ここでは，$p(\mathbf{x}_t|\mathbf{y}_t)$ は正規分布に従うと仮定するので，実際に求めるのは平均と分散共分散行列である．$p(\mathbf{x}_t)$ の平均 $\mathbf{x}_{t|t-1}$，分散共分散行列 $V_{t|t-1}$ は，

$$\mathbf{x}_{t|t-1} = F_t\mathbf{x}_{t-1|t-1} \tag{13}$$
$$V_{t|t-1} = F_tV_{t-1|t-1}F_t^T + Q_t \tag{14}$$

のように表される．また，$p(\mathbf{x}_t|\mathbf{y}_t)$ の平均 $\mathbf{x}_{t|t}$，分散共分散行列 $V_{t|t}$ は，

$$\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1} + V_{t|t-1}H_t^T(H_tV_{t|t-1}H_t^T + R_t)^{-1}(\mathbf{y}_t - H_t\mathbf{x}_{t|t-1}) \tag{15}$$

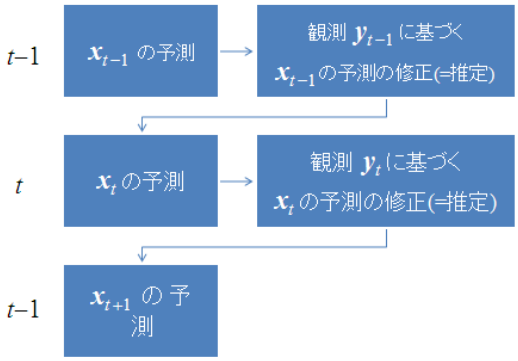

図 5: カルマンフィルタの流れ

$$V_{t|t} = [I - V_{t|t-1}H_t^T(H_tV_{t|t-1}H_t^T + R_t)^{-1}H_t]V_{t|t-1} \tag{16}$$

のように表現される．前のステップまでの結果を踏まえて $p(\mathbf{x}_t)$ (厳密にいうと $p(\mathbf{x}_t|\mathbf{y}_0, \ldots, \mathbf{y}_{t-1})$) を出しているのがポイントである．但し，シミュレーションモデル $f_t(\mathbf{x}_{t-1})$ を線型化しなくてはならないのが難点である．カルマンフィルタの流れを図 5 に示す．式 (15), (16) を用いて，ある時刻 $t-1$ における予測値 $p(\mathbf{x}_{t-1})$ を観測データ $\mathbf{y}_{t-1}$ に基づいて修正（= 推定）してやり，今度は式 (13), (14) を用いて次のステップ t における予測値 $p(\mathbf{x}_t)$ を推定する．このサイクルを続けることで，過去の情報を含んだ予測値に基づいた推定を行っていくことができる．

カルマンフィルタの大きな問題のひとつは，シミュレーションモデルを線形化しなければ適用できないという点であるが，シミュレーションモデルを線形化せずに，カルマンフィルタと同じように時系列に沿って予測を更新していくアンサンブルカルマンフィルタという手法がある．アンサンブルカルマンフィルタは，シミュレーションモデル $\mathbf{x}_t = f_t(\mathbf{x}_{t-1}) + \mathbf{v}_t$ を線型化せずに，$p(\mathbf{x}_t)$ の平均 $\mathbf{x}_{t|t-1}$，分散共分散行列 $V_{t|t-1}$ を求める．線形化を行わない代わりに，前ステップの推定値 $\mathbf{x}_{t-1|t-1}$ の周りに $V_{t-1|t-1}$ の揺らぎを持った多数の $\mathbf{x}_{t-1}$ の値を用意し，多数回シミュレーションを走らせる．多数のシミュレーション（アンサン

ブル）を走らせた個々の結果を, カルマンフィルタと同様の更新式で更新することからこの名が付いている. 分散共分散行列を求めるのに, 多数回シミュレーションを走らせるというのは, 気象予報などにおいて, 初期値鋭敏性の影響などを把握するために, 元々よく用いられている（アンサンブル予報と呼ばれる）. 並列計算には向いている方法と言える.

これまでは, カルマンフィルタを用いて, $||\mathbf{x}_t - f_t(\mathbf{x}_{t-1})||$ および $||\mathbf{y}_t - h_t(\mathbf{x}_t)||$ を小さくするように $\mathbf{x}_t$ を決定していくプロセスを述べてきたが, 同様の作業を 4 次元変分法（4D-VAR, アジョイント法とも呼ばれる）という手法によって行うこともできる. 気象予報などの分野においてよく使われている手法である. $||\mathbf{x}_t - f_t(\mathbf{x}_{t-1})||$ および $||\mathbf{y}_t - h_t(\mathbf{x}_t)||$ を小さくするという問題をまともに解いていく方法である. 実際には, 4 次元変分法では, $\mathbf{x}_t = f_t(\mathbf{x}_{t-1})$ （差が小さくなるではなく, エグザクトに一致する）を拘束条件として, 全時間ステップにおいて「$||\mathbf{y}_t - h_t(\mathbf{x}_t)||$ を小さく」を実現するという方向性で推定を進めていく. 具体的には, 式 (11) の最小化を, 全時間ステップで $\mathbf{x}_t = f_t(\mathbf{x}_{t-1})$ が成り立つという拘束条件の下で解く. 4 次元変分法の流れを図 6 に示す. まず, 初期値を適当に決めたあと, モデルを用いて時間発展を解き, データとの差を最小にするように推定を行う. この結果に基づいて, 元のモデルの逆変換モデル（アジョイントモデル）を用いて時間をさかのぼってやることで, 初期値の修正を行い, それをもとにまた時間発展を解く ... というサイクルを繰り返す. 4 次元変分法のほうがカルマンフィルタに比べて計算量が少なくて済むことが知られているが, アジョイントモデルの作成が面倒であるなど, 実装が難しいという側面がある.

これまで, アンサンブルカルマンフィルタと 4 次元変分法を用いた 4 次元データ同化について述べてきたが, これら 2 つの手法の優劣は決めづらい. ただ, 4 次元変分法の方が計算量が少なくて済み, 気象予報の分野などで実用化も先行して行われてきている. また, アンサンブルカルマンフィルタは, 十分な回数のシミュレーションを使えば, 4 次元変分法と同等の精度が出るはずであるが, 実際には十分な計算回数を取ることが困難なため, それに起因する問題が起こることが指摘されている. しかし, 分散共分散行列の計算に多少の工夫を施すなどすることで, アンサンブルカルマンフィルタでも, 4 次元変分法に劣らぬ結果が出るようになってきている. 加えて, アンサンブルカルマンフィルタの方が, 実装がはるかに容易で, 並列計算機の発達によってシミュレーション回数の問題も克服できる可能性があることから, 将来性を見込む研究者も多い.

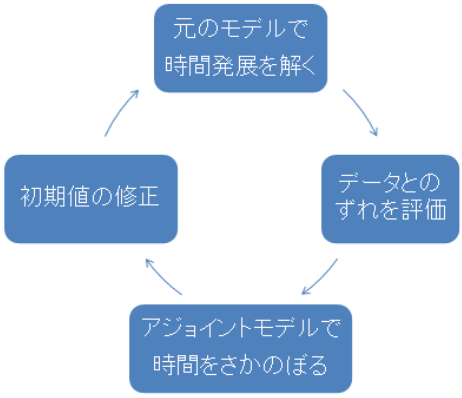

図 6: 4 次元変分法（アジョイント法）の流れ

5 まとめ

データ同化についてのあらましをまとめる.

- データ同化とは, 数値シミュレーションと観測データとを統合する手法である.
- シミュレーションの出力として得られる観測のモデル値と実際の観測データとの差が小さくなるような値を求める.
- 前のステップまでの情報を用いて予測を立て, 観測データを使った値の絞り込みも行う.

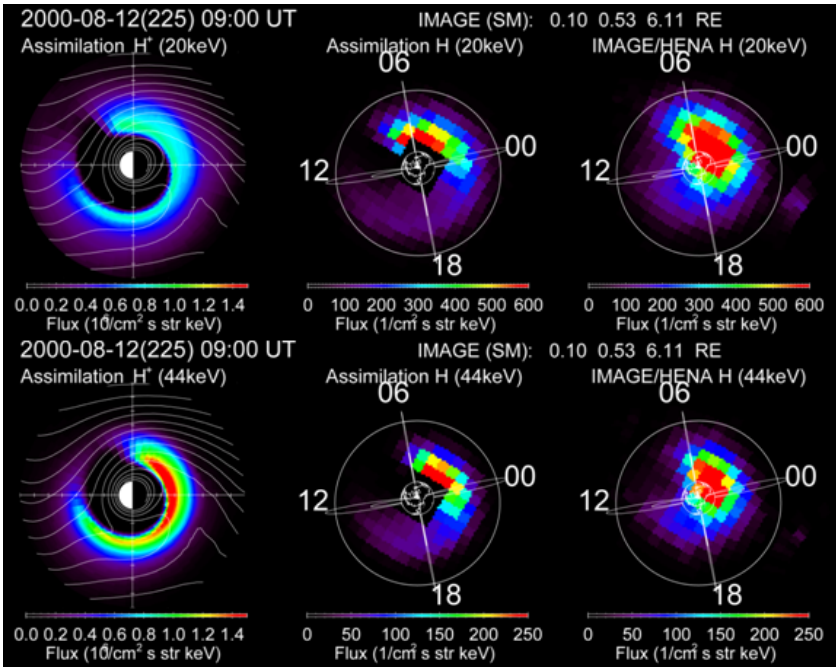

図 7: 磁気圏のポテンシャル分布を 20 個程度のパラメータで表現して, IMAGE 衛星の中性粒子観測と同化させた事例（Nakano et al., 2008）.

- それを実現するために様々な方法が提案されており, それぞれ頑張り具合が異なってくる.

宇宙科学（space science）への応用という観点から考えると, 気象分野と違い, space science 分野では, 境界条件をどう設定すべきかなど, よくわからないことが多い（事前知識が乏しい）ことが問題となる. 観測データが少ないことも難点で, あまり細かいところまでデータで抑えるのは難しい. また, 初期値を与えれば自律的に動く気象システムと異なり, 外部からの駆動の寄与が大きい space science 分野のシステムでは, 外的要因の取り扱いも問題となるため注意を要する. 最後に磁気圏のポテンシャル分布を 20 個程度のパラメータで表現して, IMAGE 衛星の中性粒子観測と同化させた事例を図 7 に示す [Nakano et al., 2008]. 左の 2 つのパネルはデータ同化によって得られた 16-27 keV と 39-50 eV のプロトンのフラックスを示し, 中央のパネルはそこから計算された中性水素フラックスの分布を示している. 右のパネルは, データ同化に用いられた IMAGE 衛星の中性水素のイメージング観測によるデータである. このデータを同化させることで, 左に示されているプロトンフラックスの分布（リングカレントの分布）がより現実に近いものになっていると考えられる.

参考文献

Nakano, S., G. Ueno, Y. Ebihara, M. - C. Fok, S. Ohtani, P. C. Brandt, D. G. Mitchell, K. Keika, and T. Higuchi, A method for estimating the ring current structure and the electric potential distribution using energetic neutral atom data assimilation, J. Geophys. Res., **113**, doi:10.1029/2006JA011853, 2008.