3.2.2 ビッグデータの有効利用例①：衛星・観測データの有効利用

(1) 課題概要

大気、海洋、陸域の物理・化学および生物環境に関する観測データは、環境変動の監視・検出や影響予測のための基礎データとして重要な役割を果たしており、日々の天気予報から地球規模の気候変動まで幅広い環境問題に適切に対処することに貢献するものである。これらの地球環境観測データは GEOSS（Global Earth Observation System of Systems）などの国際的な枠組みとも関係しながら、データのアーカイブや相互流通が進められている。

地球観測データの中でも、気象、海洋物理に関連した物理環境データは最も基礎的なデータの一つであり、現場における直接観測に加え、人工衛星からのリモートセンシングなどさまざまな種類の観測が行われ、その利用分野も広い範囲に及んでいる。ただし、それぞれの観測データは異なる場所、時間で得られており、また観測される物理量や精度等もさまざまであるため、単に既存の観測データをまとめただけではそのまま実際に利用することは難しい。そこで、データ同化と呼ばれる異なった観測データを統合する手法を用いて、より使いやすいデータセットに加工されている。

データ同化は気象予測のための高精度な初期値を求める方法として発展してきたが、数値モデルをプラットフォームとして異なる種類の観測データを統合することができるため、さまざまな観測データをまとめた統合データセットの作成にも有効に利用されている。データ同化システムは、観測データを数値モデルによってつくられる位相空間に射影することによって観測データを統合しており、時空間的に均質なデータセットを作成することができる。一方で、観測データには非常に広いスペクトルの情報が含まれているのに対し、数値モデルによって再現される現象はモデルの分解能や再現可能なプロセスの制約によって観測データよりも情報量が少なくなっているため、数値モデルに射影することにより情報が欠落してしまう。そこで、限られた観測データをより有効に活用し、より現実的なデータセットを作成するためには、大規模並列計算機を用いた数値モデルの高分解能化などの改良によって再現可能なプロセスを増やすことは必須である。

また、観測技術の進歩により、観測データの高分解能化や観測される変数の多様化も進んでいる。例えば、2009 年に打ち上げられた人工衛星「いぶき」（GOSAT）による大気中の二酸化炭素濃度の観測や、海洋中を自動昇降しながら観測する ARGO フロートに近年クロロフィルセンサーが取り付けられる等、全球の炭素循環に関わる観測データも次々と得られるようになってきた。このような観測データと従来の物理データを統合するためには、データ同化システムも炭素循環プロセスを含むモデルを用いたものに改良する必要がある。更に、生物化学過程は強非線形であるため、データ同化手法についても、非線形かつ非ガウス型の、計算負荷が高いより高度なものに改良することが必要である。

加えて、計算機性能の向上によりシミュレーションが高分解能化し、精緻化すると、ゲリラ豪雨や竜巻といった小スケールの極端現象が再現できるようになる。これに合わせて、フェーズドアレイ気象レーダーや次期気象衛星「ひまわり」といった時空間分解能が桁違いに高い観測システムの展開も視野に入ってきている。これらにより、社会的インパクトの高い局所的な

顕著現象の予測に正面から取り組む材料がそろう。一方で、このようなシミュレーションと観測の双方の桁違いの大容量化に対応するには、これまでのデータ同化システムの延長ではリアルタイム処理を行うには困難であり、ビッグデータを扱うデータ同化技術のイノベーションが求められる。

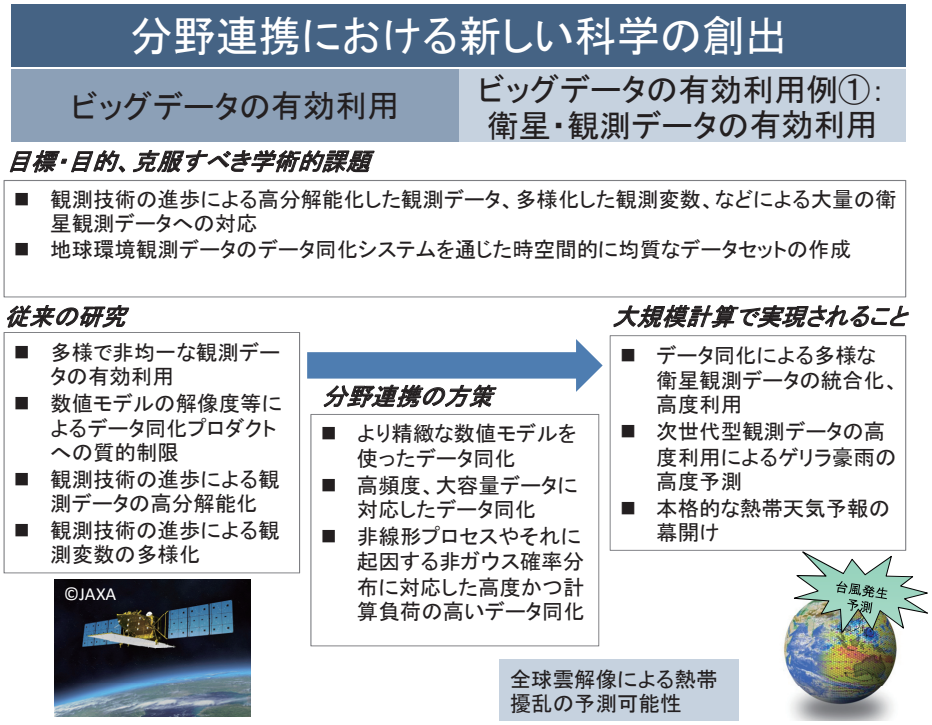

図 3.2.2-1　ビッグデータの有効利用例①：衛星・観測データの有効利用

(2)　サイエンスの質的変化と長期的目標

データ同化システムでは、観測データと整合的な数値モデルの結果を求める最適化問題を解くことによって統合プロダクトを作成しており、その解法は大きく分けて二つのアルゴリズムがある。一つは数値モデルとその随伴方程式（adjoint）モデルを繰り返し計算することによって最適解を求める 4 次元変分法（adjoint 法）と呼ばれる手法であり、もう一つは条件を少しずつ変えた数値モデルの計算を複数回実行（アンサンブル計算）し、その分散や共分散等の予報誤差情報を利用して最適解を求めるアンサンブルデータ同化と呼ばれる手法である。いずれの手法においても、必要となる計算機性能は、基本的に数値モデルの繰り返し計算を何回行うか、もしくは予報誤差情報を得るためにいくつのアンサンブル計算を行うか、によって見積もることができる。更に、データ同化システムの特徴として最適化問題を解く際に数値モデルのすべての結果が必要となることから、メモリおよび I/O に関しての要求が追加される。

4 次元変分法データ同化手法では、上述のとおり最適化問題を解くための adjoint モデルの計算において、前方積分モデルのすべての時間ステップでの結果が必要となるが、同化期間すべての結果をメモリ上に残すことは非常に困難であるため、チェックポイント法と呼ばれる方法を用いてこれを解決している（図 3.2.2-2 参照）。この方法は、（1）まず数値モデルの結果を定期的にディスクに保存しながら同化期間の前方積分を行う。（2）次に、メモリに保存でき

るだけの計算を改めて行い、その結果を用いて adjoint モデルの計算を行う。（3）これを繰り返すことにより、期間全体の adjoint 計算を行う。以上により繰り返し法の 1 往復計算が完了する。

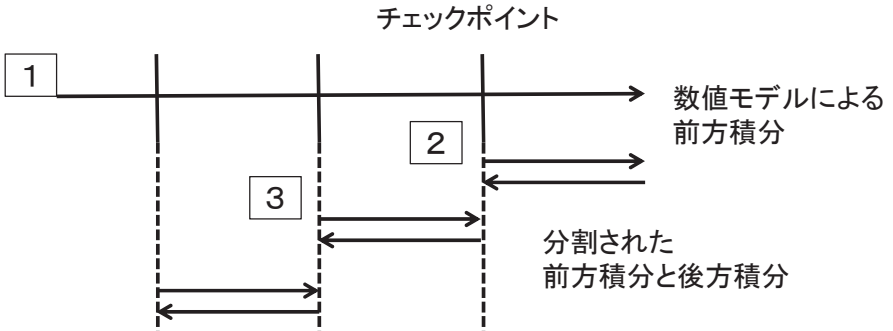

図 3.2.2-2　チェックポイント法の概略図

この手法では分割された計算ごとに計算をディスクに保存するための I/O が発生するため、高性能のディスクが必要となる。また、メモリサイズが大きくなれば、一度にメモリ上に保存できるステップ数が多くなり分割数が少なくてすむため、メモリ性能と I/O 性能の間にはトレードオフが存在する。

データ同化システムの計算には二つの運用形態が考えられる。一つは観測データが得られるたびに同化計算を行うもので、例えば 1 ヶ月に一度、3 ヶ月間分の同化計算を行うような形態を想定している。この場合には 3 ヶ月間の同化計算を 1 週間程度の実行時間で行うことができれば実用上問題はない。もう一つの形態は長期間の統合データセットを作成する場合である。例えば 1950 年からの約 60 年分の観測データを統合したデータセットを作成することを想定すると、システムを占有できる現実的な期間である半年で 60 年分の統合データセットを作成するためには、3 ヶ月の同化計算を 18 時間程度で行うこと必要となる。

(3)　コミュニティからの意見

2013 年 3 月 25 日東京海洋大学で行われた日本海洋学会春季大会において「ポスト『京』に向けた計算科学としての海洋学の展望」と題したシンポジウムを開催し、海洋モデル研究者およびデータ同化研究者を含めた議論を行った。データ同化の視点からのコメントとして、主なものを挙げる。

- データ同化システムのためにはメモリ、CPU に加えてディスク性能が非常に重要となる。データ同化プロダクトを擬似的な観測データとして解析に使いたいという希望もあり、大量データを保存する必要があり、大量かつ高速なストレージへの要求は高い。
- ディスク I/O についてそれほどの性能向上がない場合には、モデルの中に解析ツールを埋め込んでオンラインで行う必要があるかもしれない。
- データ同化システムとしては今後大気との結合、生態系との結合等他の分野と連携したデータ同化システムの発展が期待されている。

- データ同化システムの開発の効率化や人材育成のために、いろいろなグループが協力してコミュニティモデルとして開発する方向で動き始めている。
- トップ性能の計算機だけでなくその次のレベルの計算機には多様な特徴を持つ計算機がいくつかあってほしい。また、そのなかで海洋分野に使いやすい計算機をコミュニティマシンとして利用できる形態があれば望ましい。

2013 年 5 月 18 日に日本気象学会春季大会にて「ポスト『京』に向けた気象・気候シミュレーションの展望」と題した専門分科会が開催された。主な意見を以下に示す。

- 演算性能が向上し、モデルの解像度が上がることで、これまで有効利用できなかった高分解能な観測データが有効に活用できるようになる。
- 観測技術の向上により、次期気象衛星「ひまわり」やフェーズドアレイ気象レーダーなど、観測データの時空間分解能が桁違いに向上していく見込みである。このため、次世代の計算機を考えるにあたっては、演算性能だけでなく、それに見合った I/O 性能の向上が、データ同化には重要となる。
- 次世代型大容量観測データを用いたリアルタイムの天気予報には、少なくとも 1 日 1 エクサバイト程度のスループットが必要だと見込まれる。

(4) 必要な計算機資源

（2）で述べた運用形態を踏まえ、4 次元変分法についての要求条件を JAMSTEC（独立行政法人海洋研究開発機構）において開発された大気海洋結合データ同化システム（K7-CDA）の解析から見積もる。K7-CDA は数値モデルとして大気海洋結合大循環モデルを用いており、次世代のシステムとしては以下の問題規模を想定する。

- 格子数　大気：T213L150（水平約 60km、鉛直 150 層）　海洋：3600×1800×150
- 積分時間と実行時間　60 年分のデータセットを半年で作成

（3 ヶ月積分×100 回繰り返しを 18 時間程度）

これに対し、既存の計算で得られた演算量は以下のとおりである（3 ヶ月積分×100 回繰り返し）。

- 格子数　大気 T42L24（水平約 280km、鉛直 24 層）　海洋：360×180×45
- 計算量 32PFLOP
- メモリ量：85GB
- I/O 量：95GB
- 実行効率：4%（地球シミュレータ 2　4node　3TFLOPS）

なお、この計算におけるチェックポイント法では、3 ヶ月積分に対し adjoint 法は 720 回に分割して行っている。

これから次世代システムにおける計算リソースは次のように見積もられる。ただし、上述のとおり、メモリ量と I/O 量はトレードオフの関係にあるため正確に見積もるのは難しいが、既存の計算と同程度の分割数で行うと仮定した。(1 ケース=3 ヶ月積分×100 回繰り返し)

- 計算量：48000EFLOP (200EFLOP/1 ケース)

（1 ケースの計算を 18 時間で終わらせるためには実行効率 0.5%を仮定すると 0.6EFLOPS のシステムが必要）

- メモリ量：18TB
- I/O 量：22TB/1 ケース

以上が 4 次元変分法データ同化システムを用いた大気海洋統合データセットの作成に必要な計算機性能であるが、4 次元変分法ではなくアンサンブルデータ同化手法を用いた場合についても見積もる。数値天気予報システムの場合、次世代のシステムとしては以下の問題規模が想定される。

- 実験形態：大気モデルのシミュレーションを 4.5 時間分行ない、3 時間積分した時刻の前後 1.5 時間のアンサンブル出力値(30 分毎)を用いてデータ同化を実行する。これを 480 サイクル繰り返し 2 ヶ月間のシミュレーションを行う
- ケース数：2
- アンサンブルメンバー数：1000
- 大気モデル設定
  - 格子数：43 億グリッド（NICAM G-level 11,水平解像度 3.5km 相当, 鉛直 100 層）
- アンサンブルデータ同化設定
  - 入力する観測データ数：80 万
  - 3 つの異なる空間スケールでの解析を行う

全ケースの実験に必要なリソースは京コンピュータにおける NICAM および LETKF の実行実績より、

- 総演算量：896000EFLOP
- 総メモリ使用量：10 万ノードの実行で 0.7PB
- 総ストレージ使用量：5PB

と見積もられる。演算量は大気モデルが解像度 n 倍に対して n^3 倍に増える一方、アンサンブルデータ同化は解像度に対して n^2 倍、アンサンブル数に対して m^2 倍で増加する。この実験を 1 ケースあたり 24 日で行うとする場合、必要となる計算機性能は演算性能で 220PFLOPS、メモリバンド幅性能で 270PB/s である。このうち、アンサンブルデータ同化は行列演算が大きな部分を占めるので、大気モデルと比較してより高い演算性能を要求する傾向にある。ネットワーク通信については、大気モデルでは主に隣接通信を行い、アンサンブルデータ同化では数回の大域通信を行う。どちらも現在のネットワーク速度で十分と見積もられる。ファイル I/O では、大気モデルシミュレーションとアンサンブルデータ同化にかかる時間が 1:1 であるとすると、大気モデルでは実時間で 20-30 秒ごとにノードあたり 1GB 程度のファイル出力を行うと見積もられ、100MB/s 程度のファイル I/O 性能が必要である。これに対しアンサンブルデータ同化は書き出された大気モデルの結果を一気に読み込んで処理するため、ストレージ帯域全体で 80TB/s 程度の性能があると望ましい。また、アンサンブルデータ同化手法では、アンサンブルメンバー数の増加が性能向上に寄与することが知られている。これまでは計算機性能の制約によってメンバー数はたかだか 100 程度に限られてきたが、次世代システムとしては、モデルの

解像度を上げるのがよいのか、アンサンブルメンバー数を増やすのがよいのか、というトレードオフをますます考慮する必要が出てくるだろう。このための研究開発に、アンサンブル数を大幅に増やした実験などが必要となるため、以上の見積もりの数倍程度の性能が必要となるだろう。

なお、今回の見積もりは現行のデータ同化システムに対して単純に分解能を上げた場合のみを考えているが、より高度で計算負荷の高いデータ同化手法を用いることは想定していない。例えば、強非線形システムに対応するために非ガウス型の確率過程を考えることが可能な粒子フィルターと呼ばれる手法を用いた場合、従来のアンサンブルカルマンフィルターに比べ 10 倍以上のアンサンブル数が必要となる。数値モデルが高分解能化、複雑化した場合にはスケールに応じてシステムの非線形性も強くなり、粒子フィルターなどの高負荷な手法を用いる必要性が高まることが予想され、その場合には今回の想定以上の計算機性能が必要となるだろう。この他、観測データの高度利用、特に増え続ける衛星データを更に高度に利用していくことが、更なる計算性能を要求することへつながる。昨今の欧米の経済事情により各国の宇宙開発計画が見直され、これまでのように増大の一途をたどるかどうかは不確定要素も大きいが、近年、衛星による地球観測データは指数関数的に増大の一途をたどってきている。一方、現時点では、地球環境シミュレーションにおけるデータ同化での利用は限られており、得られているデータ量に比べて、実際に有効に利用されているデータ量はほんの一握りにすぎない。これまでも、衛星観測データの更なる高度利用に向けた研究開発は、欧米およびわが国の宇宙開発研究機関をはじめ現業天気予報機関や関連する研究機関等で精力的に進められてきており、ハイパースペクトラルサウンダといった 1 つの測器で 2000 チャンネルを超えるような大量のデータを扱えるようになってきた。今後もより高度な衛星観測センサーが開発され、これらを高度に利用する研究が続けられるとすると、データ同化システムで扱う観測データ量は膨大に増加していくことが想定される。これに対処するための I/O、メモリ、および演算速度のすべての面において、計算機性能が要求されるようになり、特に、データ同化システムではメモリ量／速度、I/O スピードに対する要求は非常に高くなることが予想される。

課題	要求性能 (PFLOPS)	要求メモリバンド幅 (PB/s)	メモリ量/ケース (PB)	ストレージ量/ケース (PB)	計算時間/ケース (hour)	ケース数	総演算量 (EFLOP)	概要と計算手法	問題規模	備考
局所的・集中的大雨、熱帯気象の高度予測	220	270	0.7	5	580	2	900000	大気モデル:NICAM(有限体積法), アンサンブルデータ同化:LETKF	水平解像度3.5km、鉛直100層、1000アンサンブルメンバー、3時間おきの同化サイクル 2ヶ月積分	10万ノードを仮定(大気モデルのノードあたり隣接通信1GB/s) 演算量、メモリ転送量、メモリ使用量は、京でのプロファイルを元に外挿
統合地球環境再解析	3.1	13	0.018	0.022	18	240	480004	次元変分法	格子点: 大気 640x320x150, 海洋 3600x1800x150 Δt: 大気1min, 海洋 30sec, 結合 10min 100イタレーション 積分時間: 3month	B/F値: 大気4.66, 海洋4.24 演算量、メモリ使用量は、ES2のプロファイルを元に精査、メモリ転送量は、ソースから見積もったB/F値をもとに、演算量から算出(キャッシュは考慮していない)。