

新領域融合PJから： 融合研究の苦勞と成果3

「データ同化」

(PD 統計数理研究所 中野純司)

2014年10月17日
情報・システム研究機構シンポジウム2014



大学共同利用機関法人 情報・システム研究機構

統計数理研究所

アウトライン

- *はじめに
- *データ同化とは
- *具体的な成果
- *研究基盤整備
- *融合研究の難しさとおもしろさ

アウトライン

- *はじめに
- *データ同化とは
- *具体的な成果
- *研究基盤整備
- *融合研究の難しさとおもしろさ

プロジェクトの変遷

新領域融合プロジェクト 統計数理基盤：
データ同化による複雑システムの定量的理解と計測デザイン



データ中心科学リサーチコモンズ モデリング・解析基盤整備：
データ同化・シミュレーション支援技術



本プロジェクトの目的

- * 広範囲な分野の複雑なシステムを対象とする研究の具体的な問題解決に向けた共同研究をととして、シミュレーションとデータ解析の作業を一体化する手法の高度化と一般化を実現する。あわせて、未適用分野の発掘を戦略的にすすめ、個別科学を横断的につなぐ新しい学問領域を創る。



- * データに基づいて学術研究や社会における課題解決を目指す研究コミュニティに対して、データ基盤整備事業によって提供するデータを高度活用し、課題の背後にある構造を可視化・モデリングし、シミュレーションなどを通じて、適切な課題解決を促進する研究環境を提供する



本プロジェクトの研究体制

- * 3チームで構成

- * 数理・計算チーム： 理論・ツールの開発
- * モデリングチーム： 既存の多量データのモデリング
- * データデザインチーム： データ取得からの解析
- * データ同化研究開発センター(統数研)が運営のとりまとめ
- * データ同化を用いる具体的明確な研究テーマ
 - * 地震による微気圧変動
 - * 感染症伝搬
 - * グローバルMHDシミュレーション
 - * 大気レーダーの3次元高分解能観測
 - * 動物細胞の分裂の動態モデル
 - * 「ほ乳類体節形成と精子形成」のモデル化に向けたデータ取得

アウトライン

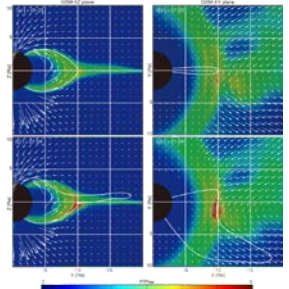
- *はじめに
- *データ同化とは
- *具体的な成果
- *研究基盤整備
- *融合研究の難しさとおもしろさ

データ同化とは

・科学研究推進には**4輪**が必要！

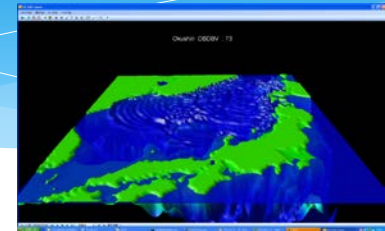
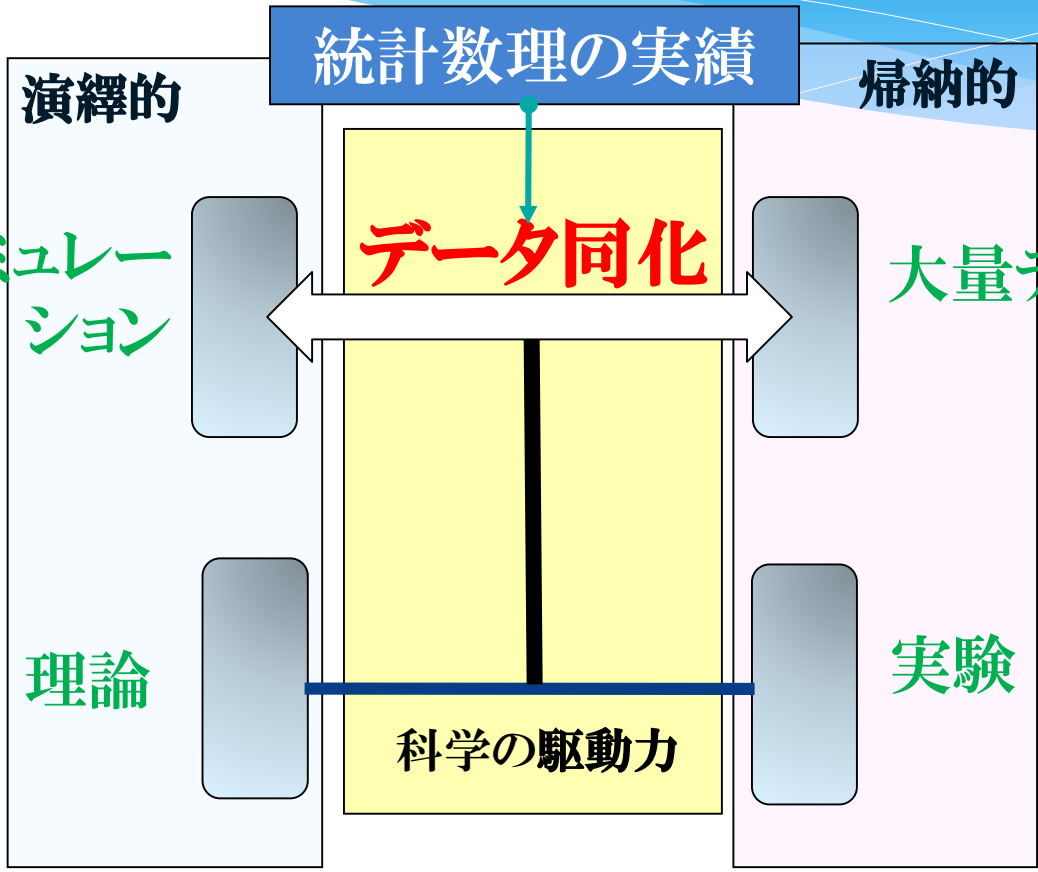
・前輪の軸が今までなかった

・**データ同化**がその軸となる

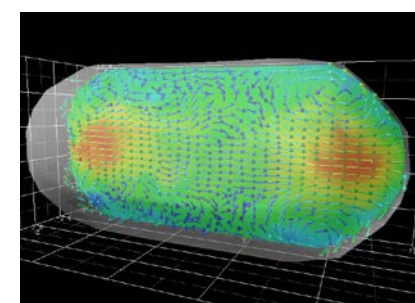
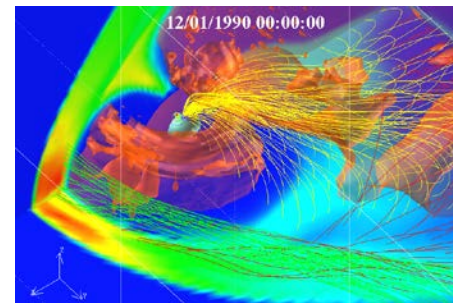
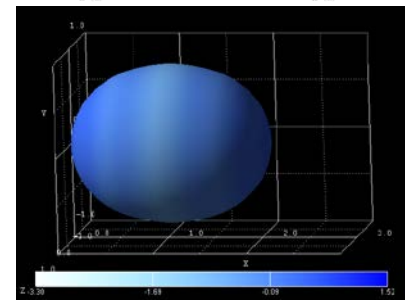
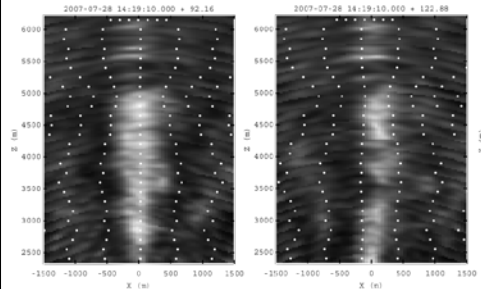


本プロジェクトで解明する現象

- ・オーロラ発生時の磁気圏の状況
- ・大気レーダーの3次元高分解能観測
- ・動物細胞の分裂の動態モデル
- ・ほ乳類体節形成と精子形成 など



大量データ処理



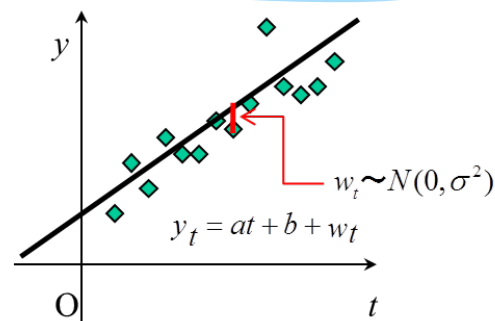
データ同化とは？(1)

要するに、
観測データにモデルをあてはめること

$$p(y_t | \theta = [a, b, \sigma^2])$$

$$E = \sum_{t=1}^T \{y_t - (a \cdot t + b)\}^2 / \sigma^2$$

データに直線をあてはめる



データにシミュレーションモデルをあてはめる

= 「シミュレーションモデルにデータを同化する」

データ同化

複雑なモデルを扱うことになるのが特徴

気象・海洋分野での“方言”

データ同化とは？(2)

- 気象学・海洋学の分野で発達（1990年代中頃から）
- 物理数値シミュレーションモデルと実際の観測を統合する手法
 - シミュレーションのみでは適切に現実の物理現象を再現できない
 - シミュレーションモデルには、モデルの不完全性や境界条件が正確にはわからないなどの不確かさが存在
 - （たとえば）正確な気象予測には適切な初期条件の構成が必要
 - 観測データは物理的・社会的制約により得られる情報に限界がある

 観測データを用い数値シミュレーション内の変数を修正
＝データ同化

シミュレーションモデルと誤差

実システムに対応した偏微分方程式
(連続時間・空間)

$$\frac{\partial x}{\partial t} = cx^2 + \dots$$

数値シミュレーションモデル
(離散時間・空間, 有限差分方程式)

$$x_t = f_t(x_{t-1})$$

境界条件・モデル化誤差
由来の不確かさ

非線形状態空間表現のシステムモデル
(離散時間・空間, 確率差分方程式)

$$x_t = f_t(x_{t-1}, v_t)$$



データ同化の定式化

- データ同化は非線形非ガウス状態空間表現での状態 x_n の推定:

(システムモデル) $x_t = f_t(x_{t-1}, v_t \mid x_0)$

(観測モデル) $y_t = H_t x_t + w_t$ または $y_t = h_t(x_t) + w_t$

- * x_t : シミュレーションモデル内の全ての変数
- * y_t : 観測される全ての変数
- * v_t : モデルの不確かさ (境界条件, ...)
- * w_t : 観測誤差
- * v_t, w_t : (通常) ガウス分布, x_0 : 初期条件

超高次元

$$x_t : \begin{matrix} 10^4 \sim 10^6 \\ (1万 \sim 100万) \end{matrix} \quad y_t : \begin{matrix} 10^2 \sim 10^5 \\ (100 \sim 10万) \end{matrix} \quad \dim(x_t) \gg \dim(y_t)$$

分布の表現：モンテカルロ近似

$$p(x_t | y_{1:t-1}), p(x_t | y_{1:t}), p(x_t | y_{1:T})$$

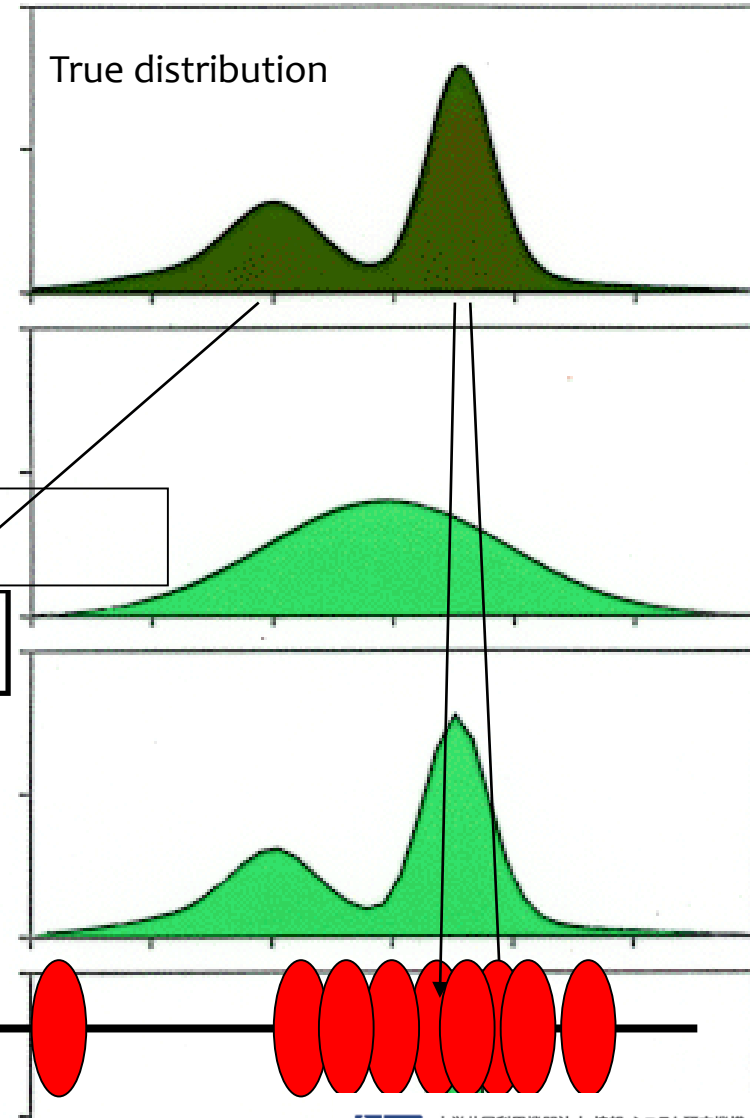
モンテカルロ近似

実現値の集合で分布を表現する

N : 粒子数

$$p(x_t | y_{1:t-1}) \cong X_{t|t-1} \equiv [x_{t|t-1}^{(1)}, x_{t|t-1}^{(2)}, \dots, x_{t|t-1}^{(N)}]$$

$$p(x_t | y_{1:t}) \cong X_{t|t} \equiv [x_{t|t}^{(1)}, x_{t|t}^{(2)}, \dots, x_{t|t}^{(N)}]$$



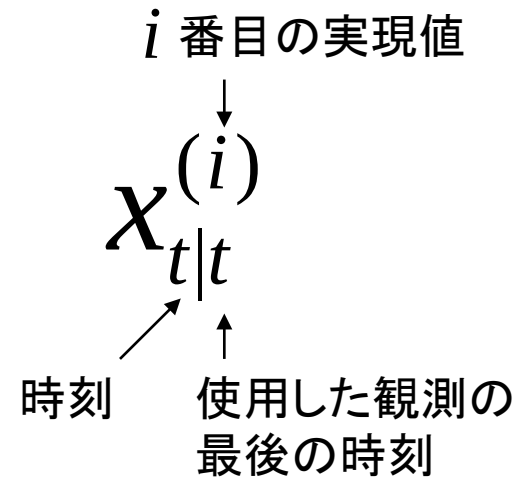
逐次データ同化の手法

- * 逐次データ同化においては、**アンサンブルカルマンフィルタ(EnKF)**がよく用いられる。
 - * 条件付分布を実現値の集合(アンサンブル)として保持
 - * フィルタリングにはカルマンフィルタの式を利用
- * われわれは**粒子フィルタ** (Particle Filter (**PF**))も応用する。
 - * 実現値集合を用いる点でEnKFと同様

$$p(x_t | y_{1:t-1}) \cong \frac{1}{N} \sum_{i=1}^N \delta(x_t - x_{t|t-1}^{(i)})$$

$$p(x_t | y_{1:t}) \cong \frac{1}{N} \sum_{i=1}^N \delta(x_t - x_{t|t}^{(i)})$$

$$\left\{ x_{t|t-1}^{(i)} \right\}_{i=1}^N$$
$$\left\{ x_{t|t}^{(i)} \right\}_{i=1}^N$$



アンサンブルカルマンフィルタ の手続き

1. $t \leftarrow 1$ とし, 初期状態の実現値集合 $\{x_{0|0}^{(i)}\}_{i=1}^N$ を作る.
2. 以下を行う
 - A) (一期先予測)
 - i. システムノイズの集合 $\{v_t^{(i)}\}_{i=1}^N$ を作る.
 - ii. 各 i に対して, $x_{t|t-1}^{(i)} = f_t(x_{t-1|t-1}^{(i)}, v_t^{(i)})$ を計算する.
 - B) (フィルタリング)
 - i. 観測ノイズの実現値集合 $\{w_t^{(i)}\}_{i=1}^N$ を作る.
 - ii. 分散共分散行列 $\hat{V}_{t|t-1}, \hat{R}_t$, カルマンゲイン $\hat{K}_t$ を計算する.
 - iii. 各 i に対して, $x_{t|t}^{(i)} = x_{t|t-1}^{(i)} + \hat{K}_t(y_t - H_t x_{t|t-1}^{(i)} + w_t^{(i)})$ を計算する.
3. $t=T$ ならば停止. それ以外は, $t \leftarrow t+1$ として2. に戻る.

アンサンブルカルマンフィルタ の手続き(続)

ただし

$$\hat{x}_{t|t-1} = \frac{1}{N} \sum_{i=1}^N x_{t|t-1}^{(i)} \quad \hat{x}_{t|t} = \frac{1}{N} \sum_{i=1}^N x_{t|t}^{(i)}$$
$$\hat{V}_{t|t-1} = \frac{1}{N-1} \sum_{i=1}^N (x_{t|t-1}^{(i)} - \hat{x}_{t|t-1})(x_{t|t-1}^{(i)} - \hat{x}_{t|t-1})'$$
$$\hat{R}_t = \frac{1}{N-1} \sum_{i=1}^N w_t^{(i)} w_t^{(i)'} \quad (\text{所与の } R_t \text{ で代替可能})$$

$$\hat{K}_t = \hat{V}_{t|t-1} H_t' (H_t \hat{V}_{t|t-1} H_t' + \hat{R}_t)^{-1}$$

EnKF vs PF

- * どちらも状態変数の分布を多数の粒子の集合(アンサンブル)で表すことは同じ
- * 異なるのは、時間の経過ごとの粒子の決め方
 - * アンサンブルカルマンフィルタ
 - * それぞれの粒子を観測に合うような方向に寄せる
 - * 粒子フィルタ
 - * 観測への当てはまりの悪い粒子を破棄し、減った分だけ当てはまりの良い粒子を複製して増やす
 - * なお、当てはまりの良さは**尤度**ではかる

アウトライン

- *はじめに
- *データ同化とは
- *具体的な成果
- *研究基盤整備
- *融合研究の難しさとおもしろさ

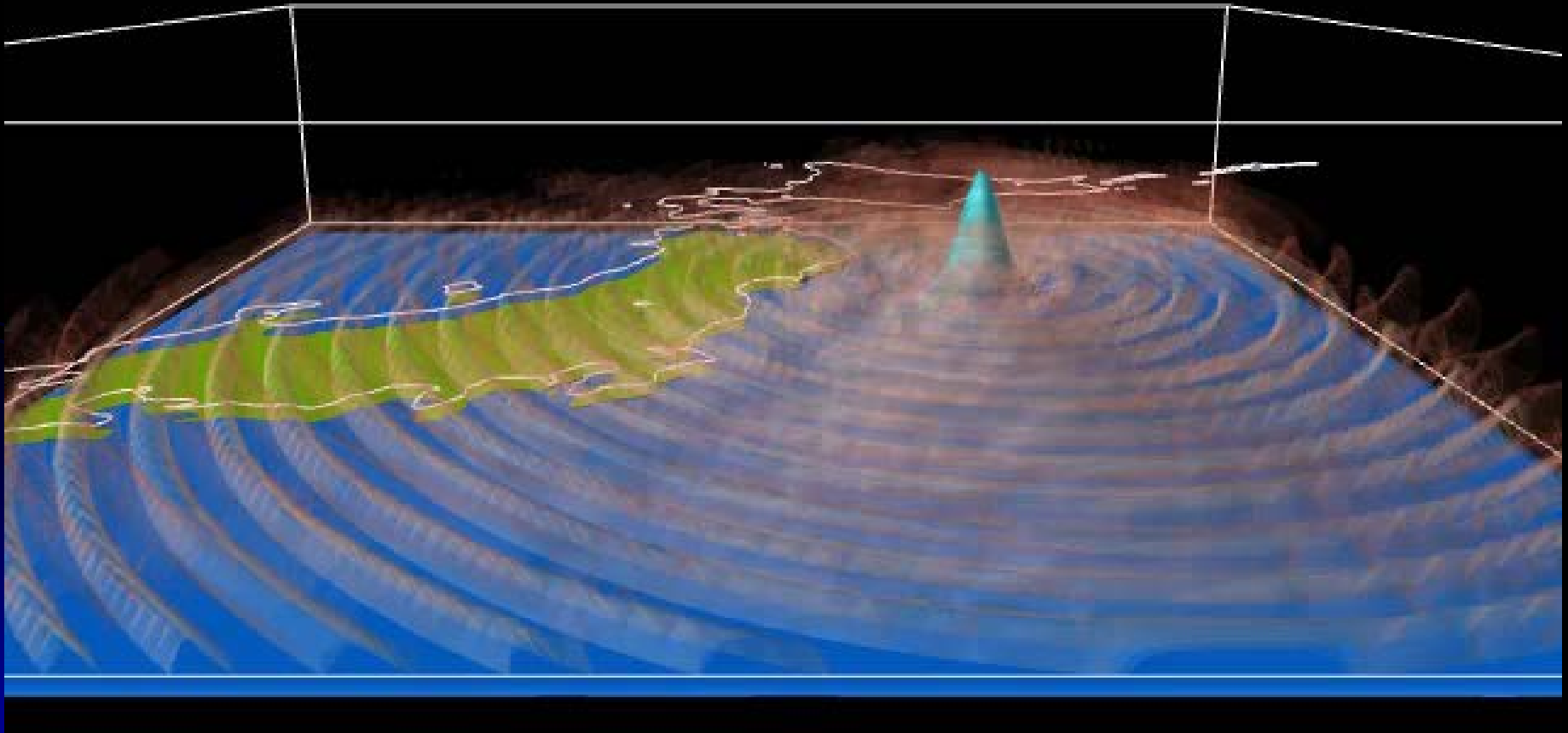
地震音波検出を目的とした微気圧観測計画

(京都大学・極地研究所・電気通信大学との共同研究)



東日本大震災に伴う地震波・地震音波伝搬シミュレーション

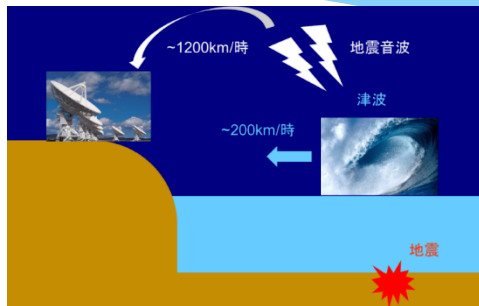
0000



東日本大震災による地震音波強度や伝搬特性を評価した

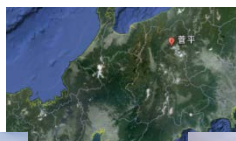
地震音波を利用した津波早期警戒システム

概念図



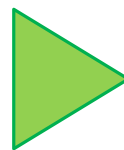
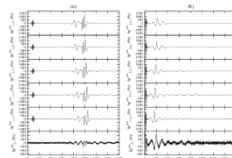
- 地震音波が津波よりもずっと速い伝播速度で沿岸に到達することを利用した津波早期警戒システムを提案。日本経済新聞2012年1月26日夕刊第一面で取り上げられた。
- 菅平宇宙電波観測所と潮岬風力実験所に微気圧観測装置を設置し、それぞれ2011年12月および2013年3月から測定を開始。
- 津波早期警戒システムの基盤となる地震音波データ同化システムを開発。震源に関するモデルパラメータと各時刻の地震音波を高速に推定できることを検証した。その研究結果は、日本地球惑星科学関連学会2012年大会のハイライト論文に選出された。

微気圧観測拠点

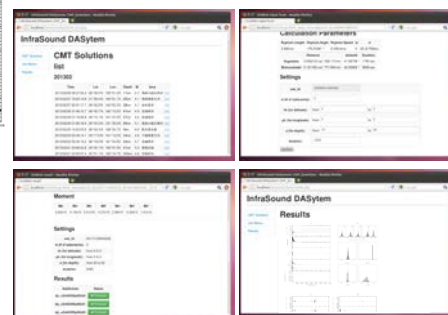


京都大学 潮岬風力実験所

電気通信大学 宇宙電波観測所



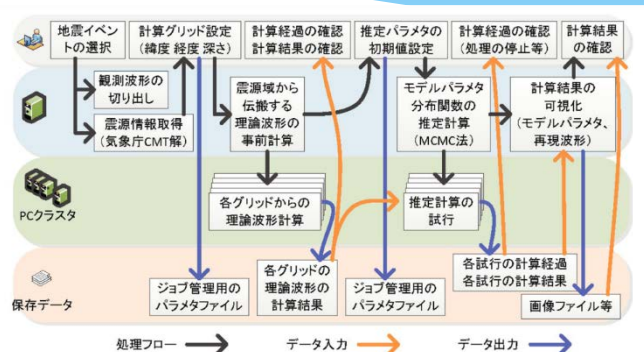
地震音波データ同化計算



統計数理研究所設置PCクラスター

システムの詳細と動作検証実験

津波早期警戒システム



モデルパラメータ

震源位置 断層破壊伝搬速度 $\theta = \left(\begin{matrix} C \\ V \end{matrix} \right)$ マグニチュード $m_{1:N}$ 風による時間遅延効果 t_0 観測ノイズ $\sigma_{1:K}^2$

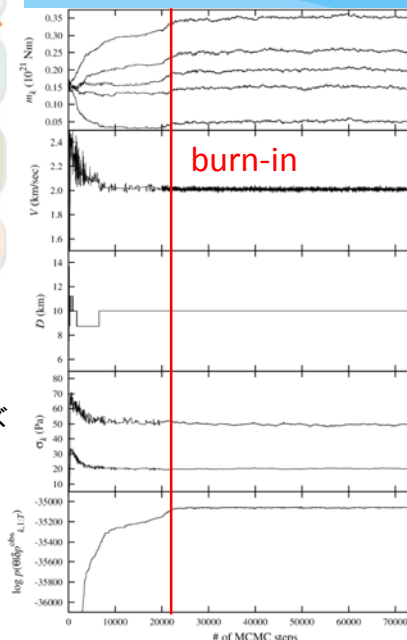
事後分布

$$p(\theta | \delta p_{1:K,1:T}^{\text{obs}}) \propto$$

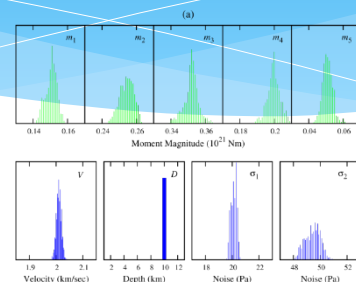
$$p(\theta) \prod_{t=1}^T \prod_{k=1}^K \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left[- \left(\delta p_{k,t}^{\text{obs}} - \sum_{n=1}^N m_n \delta p_{n,k,t-t_0}^{\text{sim}} \frac{nL}{(N-1)V} \right)^2 / 2\sigma_k^2 \right]$$

MCMCによる推定

パラメータの収束過程

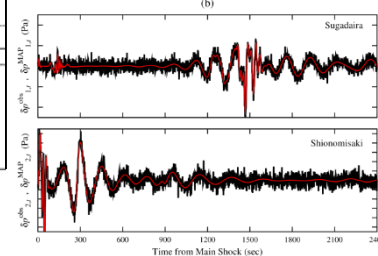


推定されたパラメータ分布



地震音波の状態推定

赤: 推定値
黒: 疑似観測データ(真値+ノイズ)

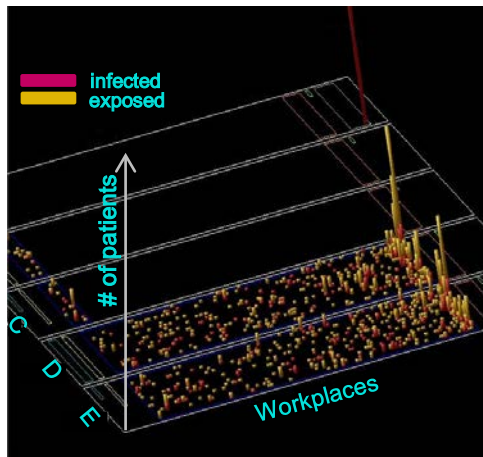


感染症数理モデルによる感染症介入政策の評価



シミュレータの模式図。鉄道で結ばれた複数の地区からなる都市を扱う。それぞれの住民は各自のスケジュールによって行動し、その結果として形成される集団内でのインフルエンザの確率的伝播をシミュレーションする。

- 都市における個人の行動を考慮に入れた大規模感染症シミュレーション(人口:122万人)を多数の計算機を接続して実行する、並列計算によって実現した。
- ワクチンを会社員などの活動的な集団に優先的に接種することで集団免疫を高め、高リスク者を含む住人全体の感染割合を低下させ得ることを示唆する結果を得た。
- 本研究の成果は、PLOS ONEに掲載され、毎日新聞平成25年9月19日夕刊に取り上げられた。



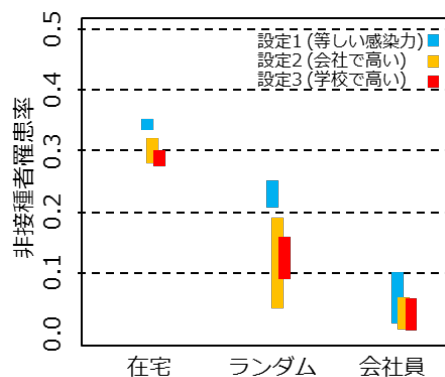
シミュレータの可視化。エージェント・シミュレーションの内部で起っている状況を2次元グラフで捉えるのは困難である。また、感染の進行を非専門家にもわかりやすく伝えることが感染症予防に役立てるために重要である。このような背景からシミュレータの開発と並行して、可視化システムの開発にも取り組んできた。

シミュレーション例

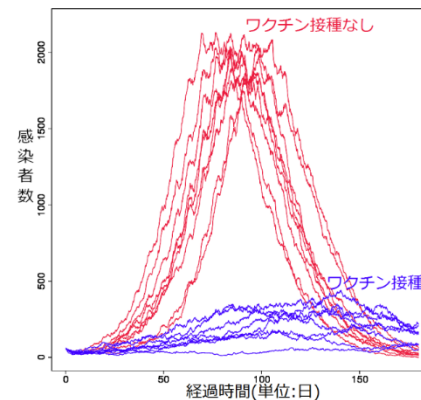


利用事例: ワクチン優先接種対象群の選択

- * 通常ウィルス上陸前に当該地域の全住民に接種することは不可能である。そのため、どのような集団に優先接種すれば、ワクチンを受けられなかった集団内での感染を最小化できるかが焦点になる。活動性の高い会社員に集中投与すると、非接種者感染割合を最小化できるというシミュレーション結果を得た。
- * 最適な優先接種のもとでは、感染者数のするどいピークが消滅し、シーズン全体に分散していることが確認できる。これは、感染路が消滅しない程度に伝達確率が低く抑えられたことによる。



優先接種対象群と感染割合

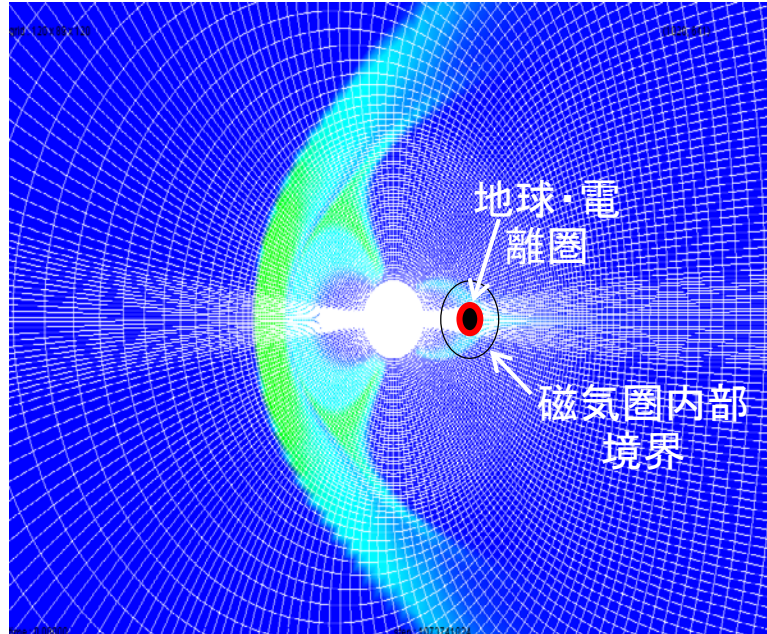


最適な選択の下での感染者数の推移

地球磁気圏グローバルMHDシミュレーションモデルによるデータ同化の研究

才田聡子(ROIS)、門倉昭(極地研)、藤田茂(気象大)、田中高史(九州大学)、他

地球磁気圏グローバルMHDシミュレーションモデル



MHD (MagneToHydroDynamics): 電磁流体力学

磁気圏モデル内部境界～接続(任意性)～電離圏モデル
(3Re(地球半径)) ↑ (1Re)

データ同化対象パラメータ

同化対象データ(候補)

1. 電離圏電位分布: SuperDARNレーダー、衛星観測
2. 電離圏電流: 地上磁場、衛星観測
3. 電離圏電気伝導度: EISCATレーダー、オーロラ観測

<全体計画>

平成22年度

非同化モデルの検証

データ獲得戦略立案

平成23年度

データ同化初期実験

データ獲得・収集

平成24年度

初期実験の解析と評価

データ獲得・適用

平成25年度

最適値推定・評価(続き)

データ獲得・適用・結果検討

平成26年度

同化手法の実装・改善

データ獲得・適用・結果検討

平成27年度

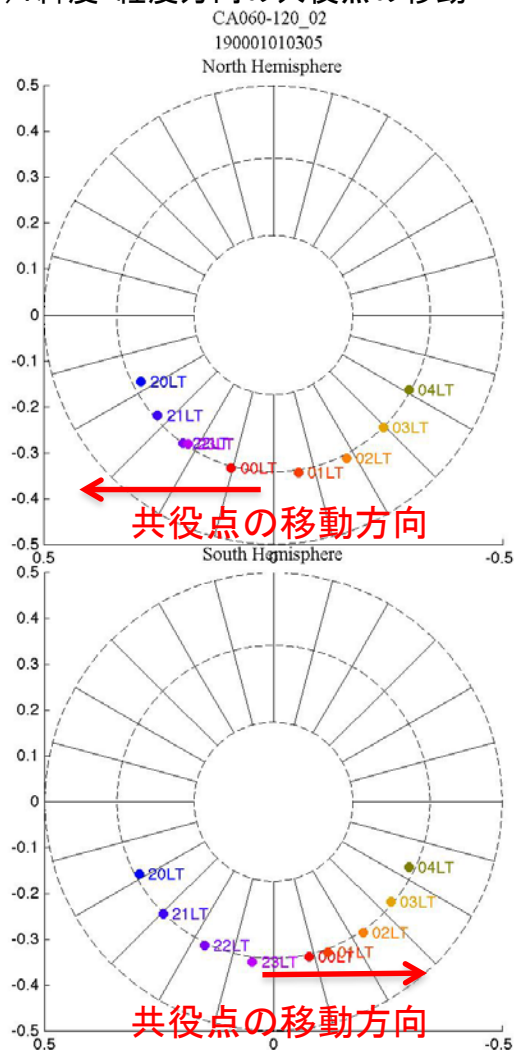
同化手法の実装・改善・運用

成果公開

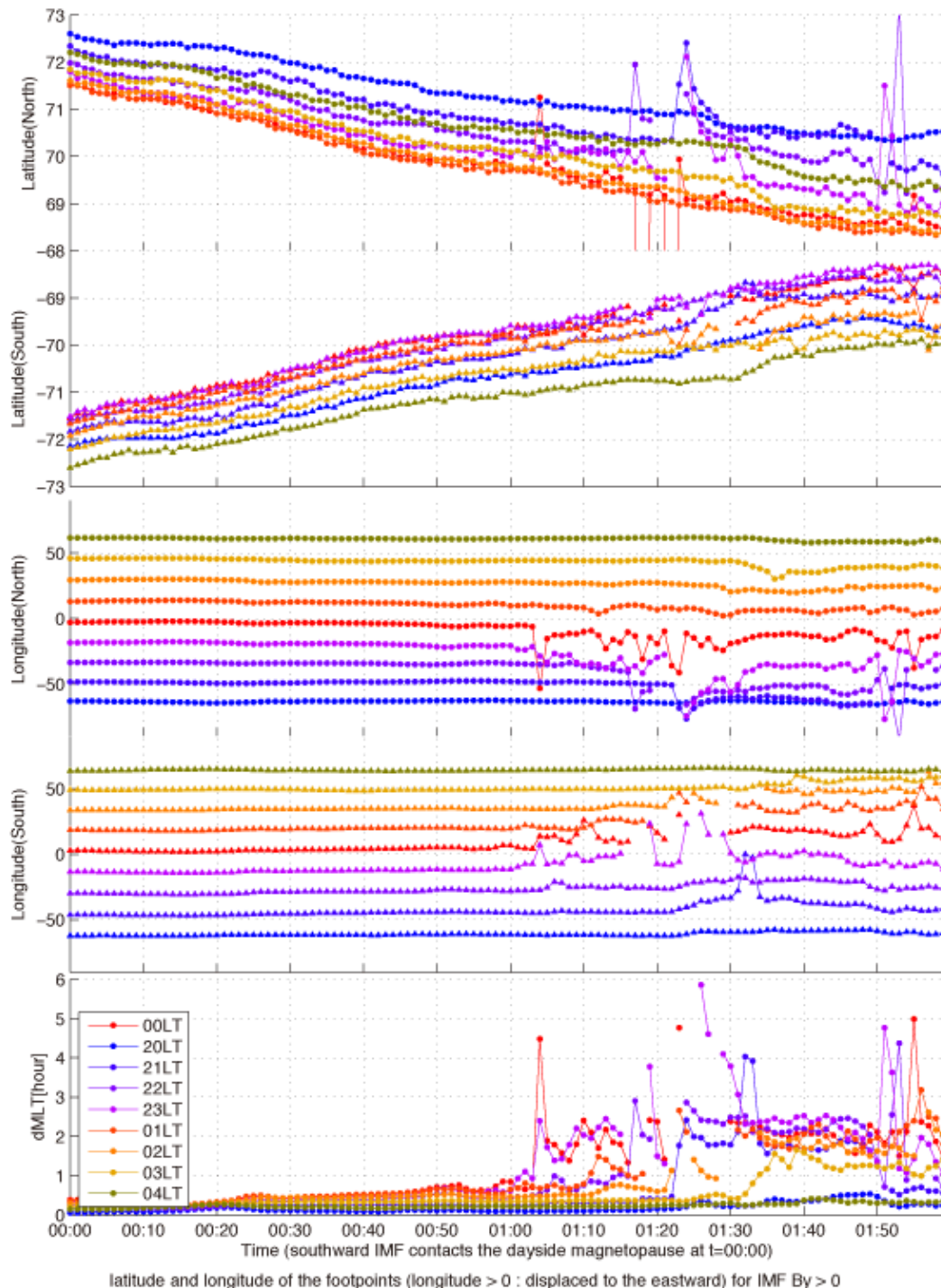
オーロラ発生時の共役点の移動データの取得について

図1(下): 共役点移動の様子

図2(右): 緯度・経度方向の共役点の移動



北半球(上)と南半球(下)で
反対方向に共役点が移動する



地球磁気圏グローバルMHDシミュレーションモデルによるデータ同化の研究

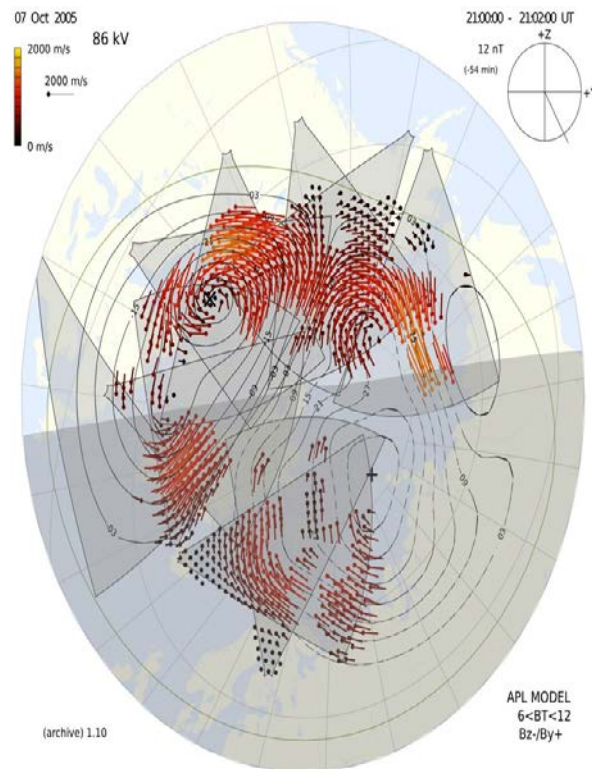
才田聡子(ROIS)、門倉昭(極地研)、藤田茂(気象大)、田中高史(九州大学)、他

平成25年度の成果

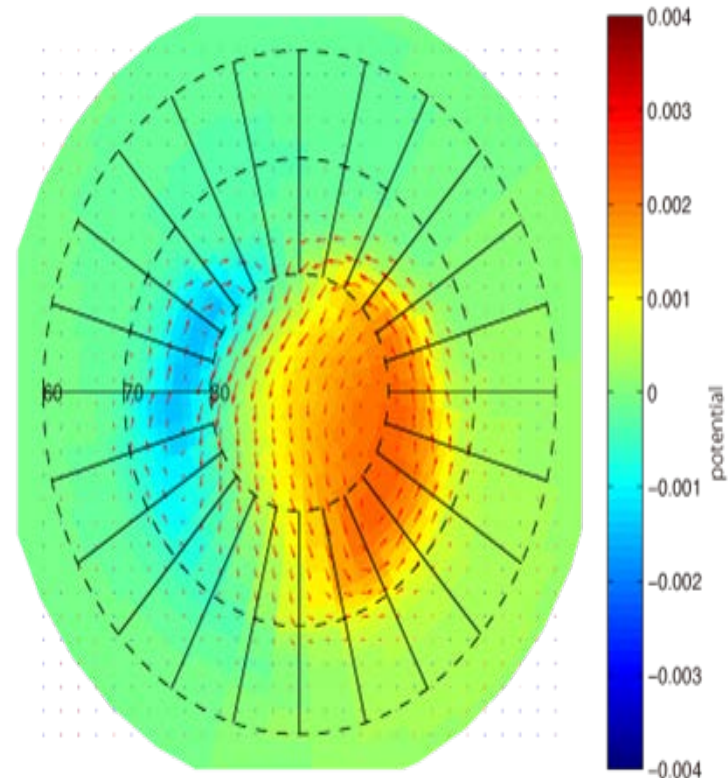
SuperDARNレーダーネットワーク観測データとシミュレーション結果の比較

電離圏電位分布とプラズマ対流分布

SuperDARNレーダー観測データ



シミュレーション結果



大気レーダーの3次元高分解能観測と同化の研究

西村耕司, 中村卓司 (極地研), 佐藤亨 (京大), 佐藤薫 (東大), 堤雅基, 富川喜弘, 山内恭 (極地研)

1. 新アルゴリズムによる鉛直風速推定 精度向上を実証

- 同化データのための物理量として精度の高い鉛直風速を得るため、ノルム拘束付DCMP法を用いる手法を開発。これにより推定精度が大幅に向上することを京都大学MULレーダーによる実験で実証した。



MULレーダー

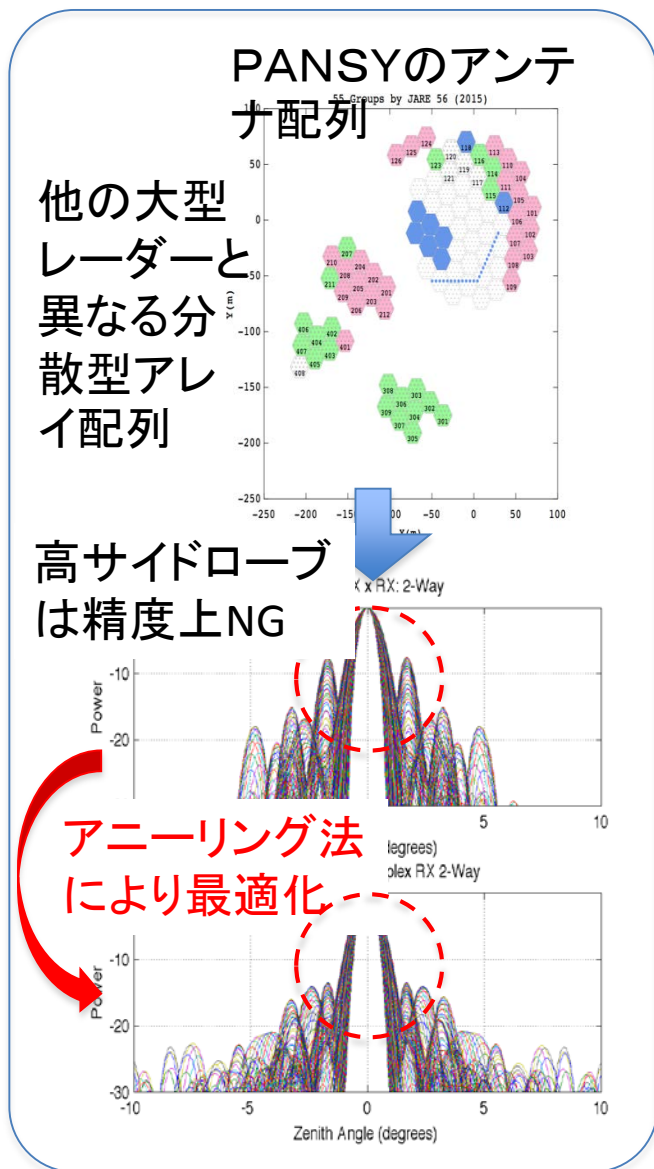


南極昭和基地大型大気 (PANSY) レーダー

2. PANSYレーダーとの組み合わせによる超高精度風速推定

- PANSYでは積雪対策のためにアンテナを分散化しており、細いメインビームと高いサイドローブを有する。数値最適化により静的にサイドローブを抑圧した上で、上記手法と組み合わせることにより、mm/sオーダーの理論精度が得られた。これは地球大循環における極域での鉛直速度と同程度であり全球モデルの同化への利用が可能性が見えてきた。

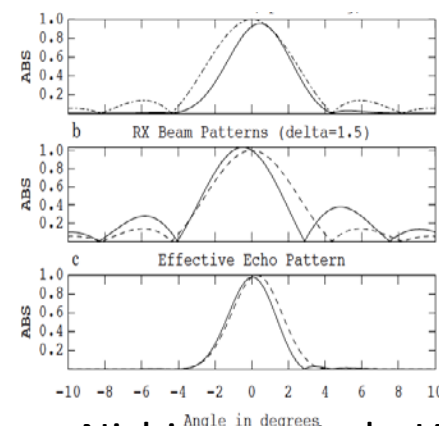
高精度推定アルゴリズムと分散アレイによる超高精度風速推定



さらに

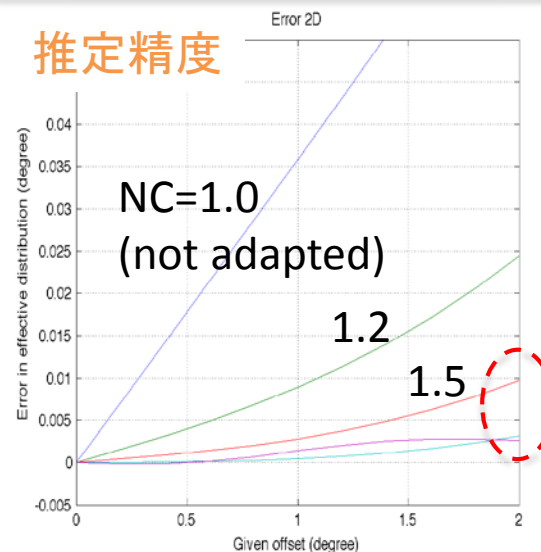


NC-DCMPアルゴリズムにより
適応最適ビーム制御



Nishimura et al., JAOT, 2012.

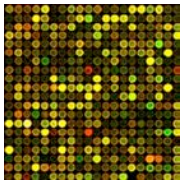
推定精度



2つの技術の相乗効果によりmm/sオーダーの超高精度推定が可能に

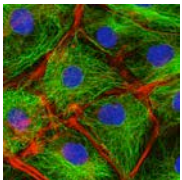
データデザインチーム（遺伝研中心）

目標: データ同化手法を発生細胞生物学分野の問題に適用し、同分野における有力な手法として確立する



これまで、バイオインフォマティクスの分野で、データ同化の手法が適用されてきた。
例えば、遺伝子発現量の網羅的計測の時間的変化の背後にある遺伝子発現制御ネットワークの推定など。
Yoshida R, Nagasaki M, Yamaguchi R, Imoto S, Miyano S, Higuchi T. Bayesian learning of biological pathways on genomic data assimilation. Bioinformatics. 24 (22),2592-2601 (2008)

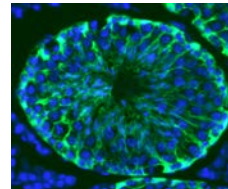
データデザインチームは、発生生物学、細胞生物学の分野においてデータ同化法、あるいは関連手法を適用し、生物学と統計学の融合による、同分野でのブレークスルーを目指す。



動物細胞の分裂、細胞質流動



ほ乳類の体節形成、精子形成



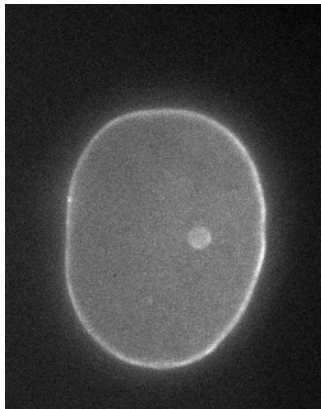
細胞分裂(線虫): 力学的変化

主な参加研究者: 木村(NIG)、樋口(ISM)、
小山(元TRIC研究員)

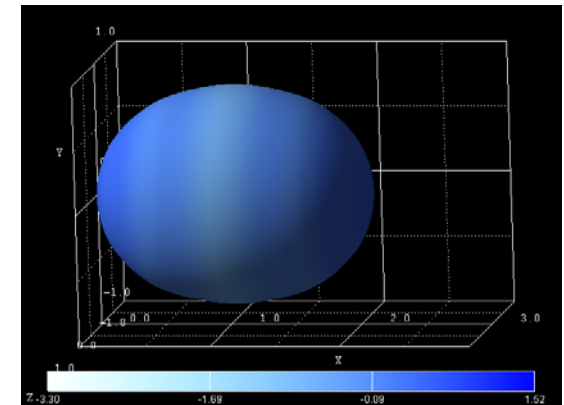
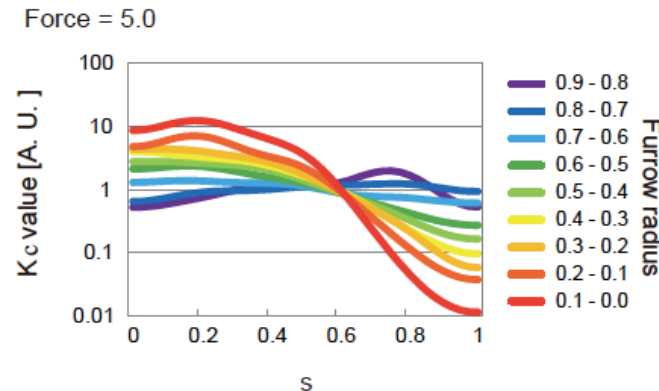
【目的】 細胞分裂時の細胞形状の時系列データから細胞表層の硬さを推定する。

【結果】

データの可視化ソフトウェア開発:
H22融合研究シーズ探索(長尾代表)



細胞分裂時の細胞形状
(座標、曲率)



表層
硬い

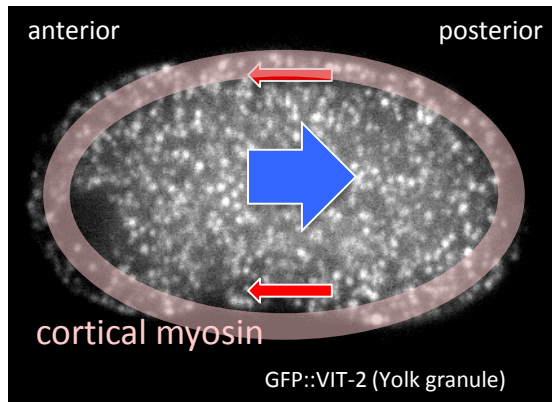
表層
硬い

- ① 細胞分裂時の細胞表層の硬さを推定した→分裂面が柔らかくなっていることを支持する
(従来、分裂面の軟化は提案されてきたが、定量的な証拠はなかった)
- ② 細胞分裂に異常をきたす遺伝子変異体では、表層の軟化が阻害されていた
→推定の信頼度を支持する、表層の軟化の機能的意義を支持する

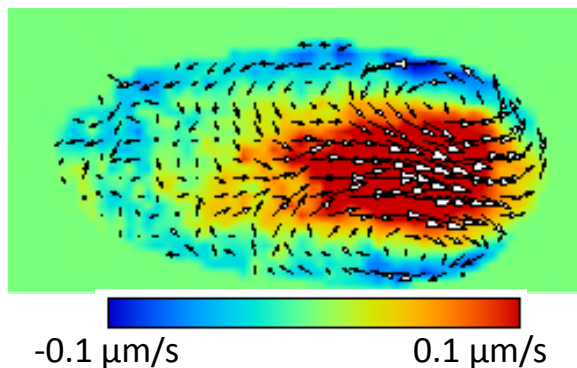
Koyama H., Umeda T., Nakamura K., Higuchi T., and Kimura A. A high-resolution shape fitting demonstrated equatorial cell surface softening during cytokinesis. *Submitted*.

【目的】 細胞質流動の流速分布から、流動の原動力の分布を推定する

細胞質流動とその定量化

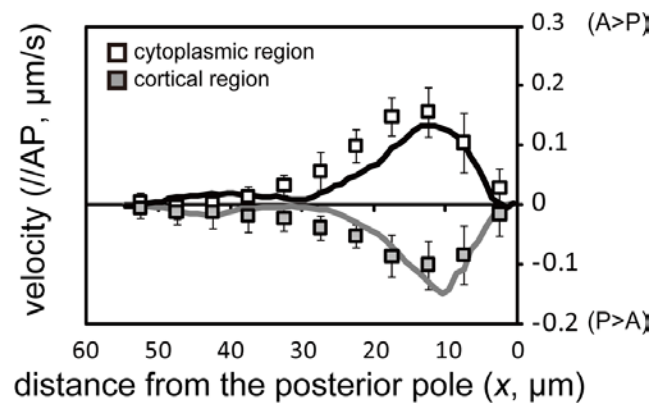


PIV methodによる流速の定量化



従来: 手作業(試行錯誤)でのパラメータ推定

Niwayama, Shinohara, Kimura A, *PNAS*, 2011



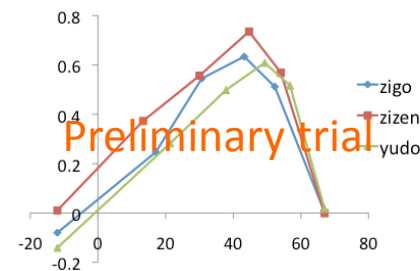
流体シミュレーション(MPS法)



データ同化を用いたパラメータ推定

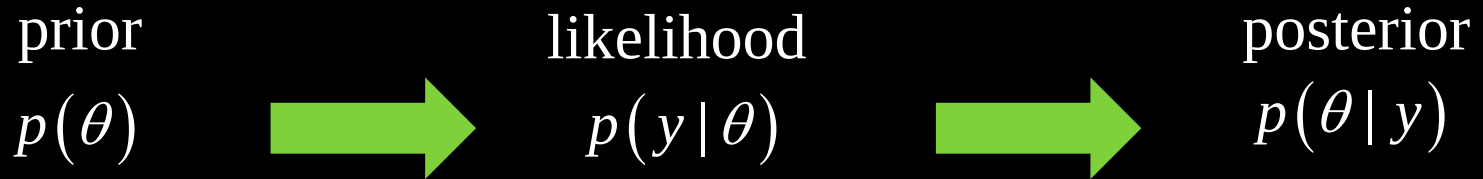
→力の分布が高い精度で推定できる

→力を生み出す機構についての考察が可能に



線虫初期胚における細胞内流動データ同化

(遺伝学研究所との共同研究)



Model parameters

- **velocity of the cell wall**
how to design $u_{wall}(L, t)$
- viscosity coefficient
- density

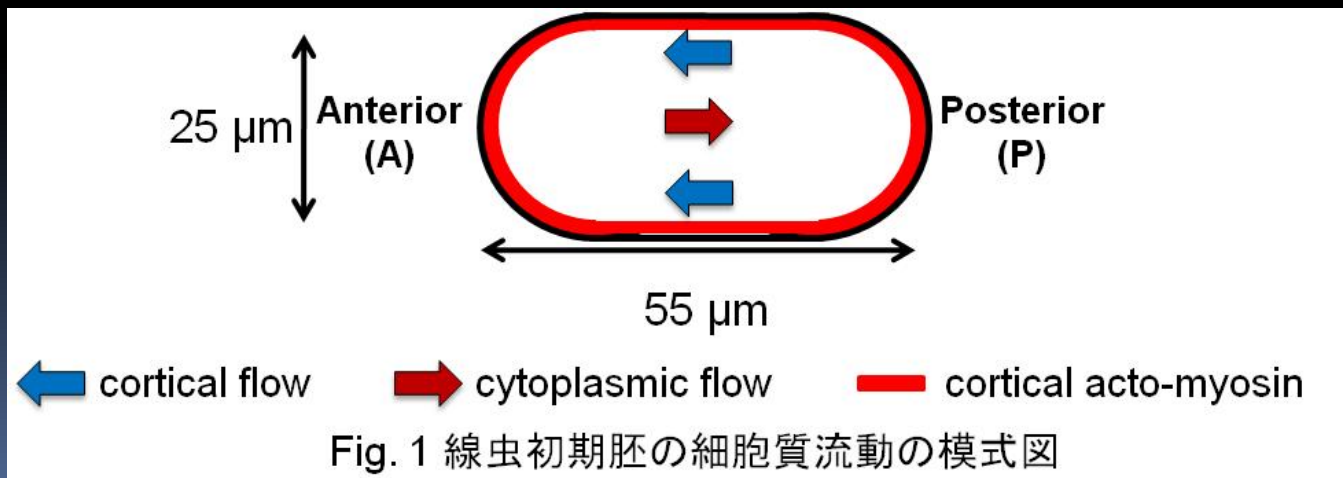
Particle filter

- search model parameters that maximize the likelihood
- **observed velocity field (2D)**

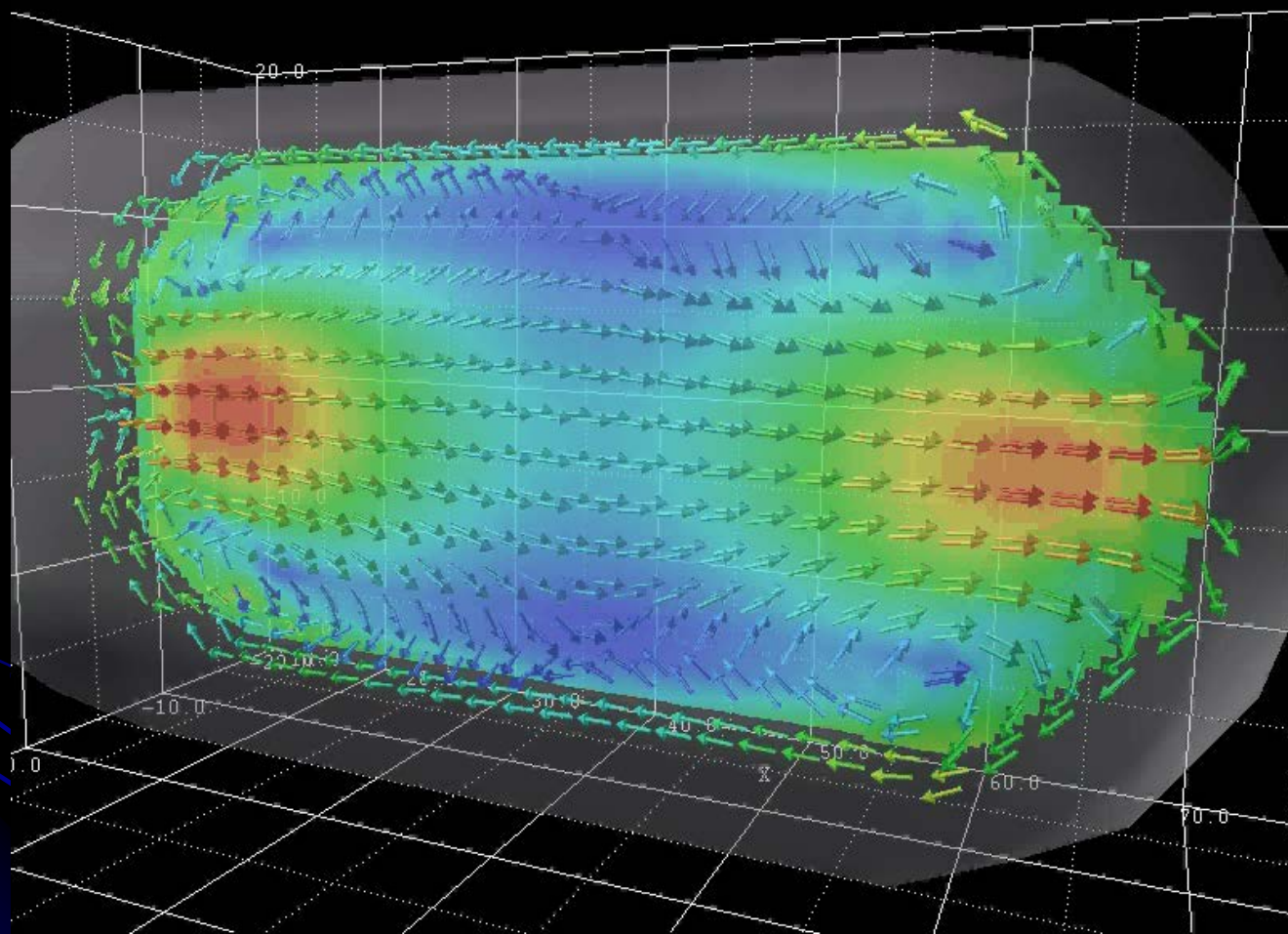
MCMC

- estimate posterior distribution

Niwayama et al. [PNAS 2011]



データ同化によって推定された細胞内流動の可視化



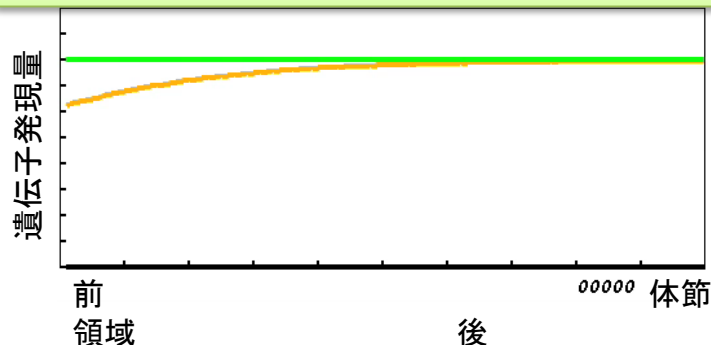
粒子フィルタ法によるデータ同化アルゴリズムを実装し、
理研RICCや京速コンピュータ「京」上で、ジョブ投入実験を実施した。

体節形成(マウス): 体節時計の同調機構

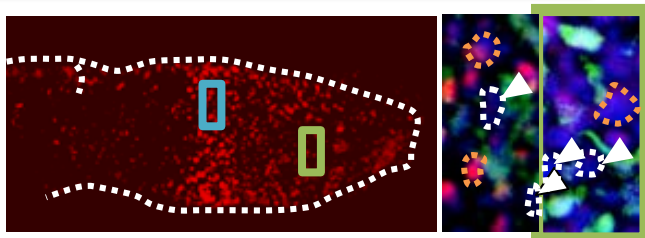
主な参加研究者: 相賀、木村、菅原(NIG)、
大久保(元TRIC研究員)

- 周期的な体節の形成機構に関する研究では、理論的解析もすすんでおり、数値シミュレーションも多く行われている。→データ同化の適用に適したモデル系?
- 体節は、各細胞にある「体節時計」の時間的振動が、空間的な縞模様に翻訳されることによって形成される。
- きちんと「縞模様」ができるためには、多数の細胞の「体節時計」が同調していなければならない。
- 同調機構の理解を進めるために、正常細胞と遺伝子機能欠損細胞をランダムに混合した「キメラ」胚を作製した。複雑なキメラ胚実験を解釈するには、データ同化を利用した数理モデルでの実験観察結果の再現が効果的である可能性を考えた。
- 現時点では、データ同化(パラメータ推定)は行っていない。データ同化に向けてシミュレーションモデルを作成したところ、キメラ胚実験の解釈はパラメータ値に左右されない定性的な性質で可能であることがわかってきた。

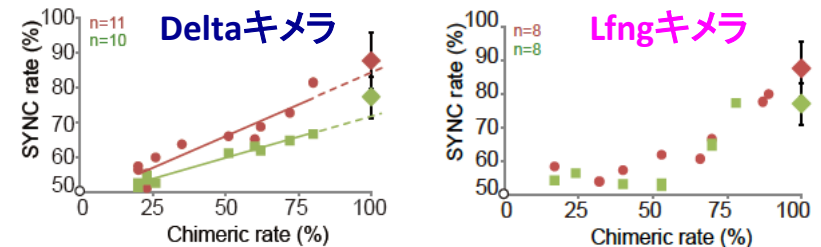
体節時計の振動と縞模様の形成 (Oginuma...Saga, 2010)



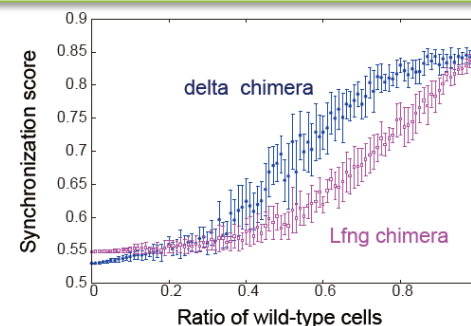
複雑なキメラ個体の結果を解釈したい



実験データ: キメラ率vs同調度



同調度の違いを説明するモデル

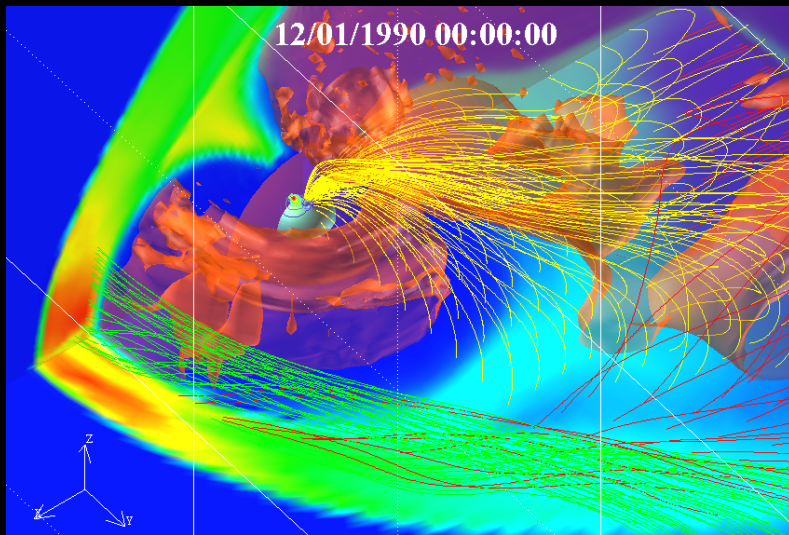


アウトライン

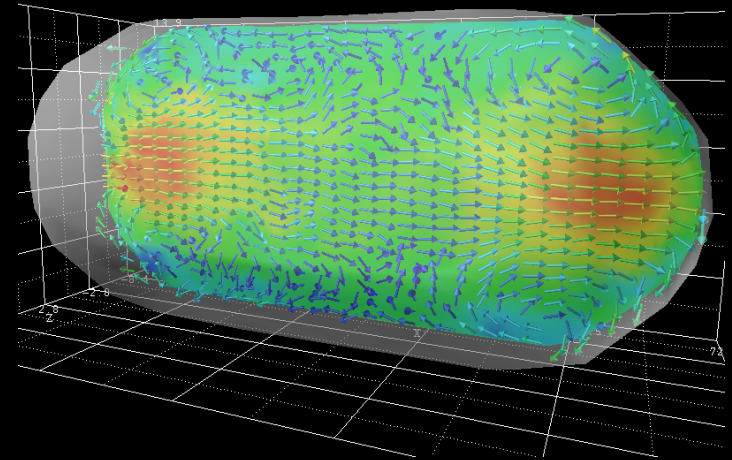
- *はじめに
- *データ同化とは
- *具体的な成果
- ***研究基盤整備**
- *融合研究の難しさとおもしろさ

データ同化のための4次元可視化ソフトウェア

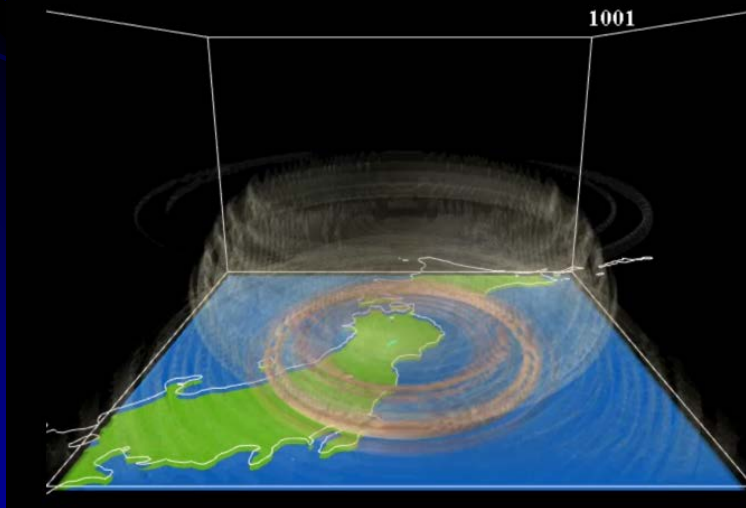
バーチャルオーロラ



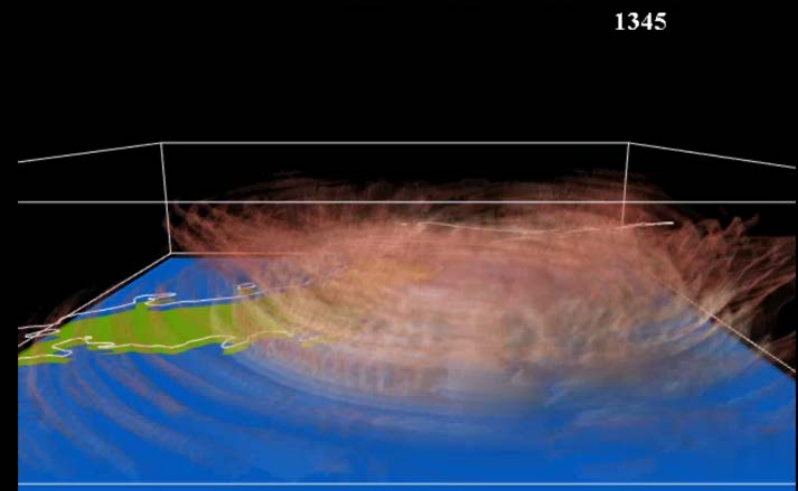
細胞内流動



地震音波伝搬(岩手・宮城内陸地震)

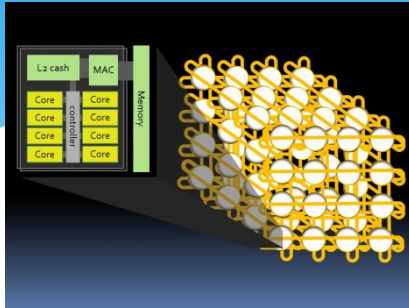


地震音波伝搬(東日本大震災)



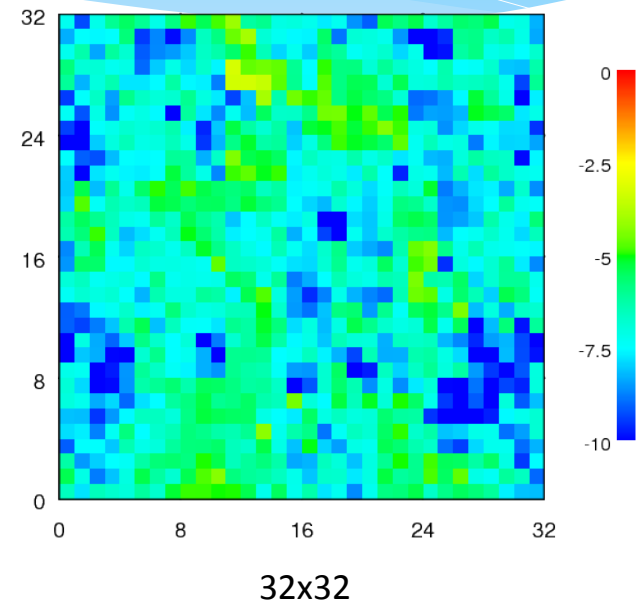
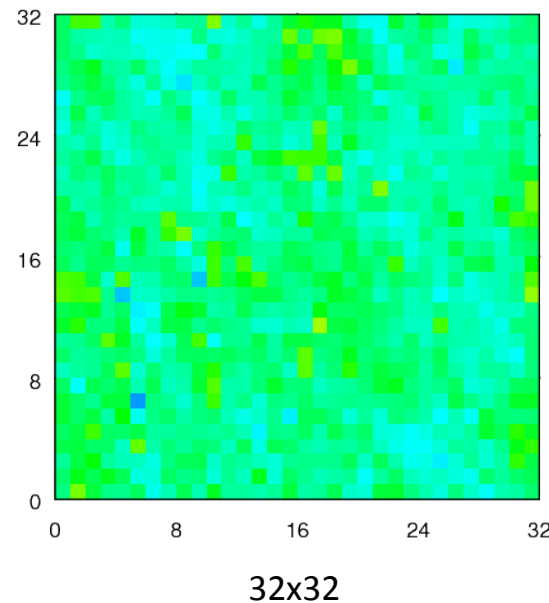
超高並列計算機のためのデータ同化アルゴリズムの高度化

～トーラス型ネットワークポロジに適した粒子交換アルゴリズム～



京速コンピュータ「京」は3Dトーラス型のネットワークポロジを持つため、粒子フィルタを適用するためには、効率的に粒子交換を実施するためのアルゴリズムが必要。

Lorenz 96モデルに対し、統計数理研究所が所有するCray XT6m (2Dトーラス) で実験



- The total number of particles is 4194304 in the left panel and 262144 in the right panel.
- If the number of particles for each node is sufficient, local degeneracy problem can be generally avoided.

数値実験結果

本年度上半期は、各ノードあたりの粒子数を一定としてノード数を増加させることで、計算速度を落とすことなく推定性能を向上できることを実験的に確認した。

PEs	Particles	RMSE	Elapsed times	Effective sample size	Effective number of nodes	Effective sample size for a node
4x4	65536	1.61	00m54s	22467.1	15.6	1440.1
8x8	262144	1.18	01m00s	90762.2	59.6	1523.3
16x16	1048576	1.11	01m04s	315334.6	203.8	1547.1
32x32	4194304	1.09	01m07s	993516.8	638.6	1548.5

- If the number of particles for each node is fixed, the increase of the number of nodes would result in the increase of the effective number of particles.
- The increase of the number of nodes would make sense if the aim is to improve the accuracy or to address high-dimensional problems.

$$S = -\sum_{\mu=1}^{\Lambda} \Omega_k^{(\mu)} \log \Omega_k^{(\mu)} + \sum_{\mu=1}^{\Lambda} \Omega_k^{(\mu)} \left(-\sum_{\nu=1}^{\lambda} \omega_k^{(\mu,\nu)} \log \omega_k^{(\mu,\nu)} \right)$$

並列計算ソフトウェア

- * CloCK-TiME (Cloud Computing Kernel for Time-series Modeling Engine)
 - * <http://sheep.ism.ac.jp/CloCK-TiME/>
 - * Web 上で多変量時系列データを粒子フィルタ法によって分析するシステム
 - * (現在, 統計解析システムRのパッケージに移植中)
- * Rhpc (An R package for High-Performance Computing)
 - * 統計解析システムRでMPIを使って並列計算を効率的に行うためのパッケージ
 - * スーパーコンピュータ上でのRの利用, データ同化の実行を念頭に置いて開発
 - * “The Comprehensive R Archive Network”の世界中のミラーサイトより入手できる
 - * 日本では例えば <http://cran.ism.ac.jp>



データ同化の普及活動

- * 樋口知之著,「予測にいかす統計モデリングの基本—ベイズ統計入門から応用まで」, 講談社, 2011.
- * 樋口知之編著, 上野玄太, 中野慎也, 中村和幸, 吉田亮, シリーズ<予測と発見の科学>6
「データ同化入門—次世代のシミュレーション技術—」, 朝倉書店, 2011.



新分野への適用など

- * データ同化の多くの新分野への適応の試み
 - * 地球・宇宙科学
 - * 多数(特にこれまであまりデータ同化が使われていない領域)
 - * 生命科学
 - * 細胞
 - * 個体形成
 - * ゲノム科学
 - * 医学
 - * 感染症の予防政策
- * 若手人材の育成
 - * 融合研究における
 - * 統計数理研と極地研, 遺伝研との密な人材交流
 - * 育った若手人材の外部への供給

アウトライン

- *はじめに
- *データ同化とは
- *具体的な成果
- *研究基盤整備
- *融合研究の難しさとおもしろさ

融合研究の難しさ

* 目的意識の違い

- * (データ同化を含む)統計数理は「データ」を扱うための学問
 - * 不確定性を含むデータから有用な情報を取り出すことがその目的
 - * ただし、特定のデータだけではなく汎用性があることが必要
- * 各分野の研究者は自分の問題が解決すればよい
 - * すでに結論を(自分の中で)決めている場合がある
 - * 統計的な推論法がまわりくどいと感じることが多い

* 統計的な考え方への理解不足

- * 日本では統計学がきちんと教えられていない
 - * データを軽んじる
 - * 白黒をはっきりとつけたがる
- * 日本には統計学科がない(先進国では特異な状況)

* 分野が異なると(専門)用語が異なる

- * 意思疎通に苦勞する＝時間がかかる



融合研究のおもしろさ

- * 学問の幅が広がる
 - * 統計数理側：データの重要性・価値の再認識
 - * 個別分野側：統計的な考え方・ツールを知る
- * コミュニケーション能力の向上
 - * 「井の中の蛙」「専門馬鹿」からの脱却
 - * 社会へのアピール能力も向上
- * 人間的な成長ができる(?)
 - * お互いを尊重することが必要
 - * 単なる「下請け」「丸投げ」「いいとこ取り」ではうまくいかない

今後の課題と期待

- * 融合研究は学問的には重要、でも参入障壁がある
 - * それを突破するためには内的・**外的**な助けが必要
 - * 本機構の融合プロジェクトは非常に有効な外的補助だった
 - * このような補助がなんらかの形で続くことを期待する
- * データ同化は応用の広い統計手法、でも参入障壁がある
 - * われわれでその参入障壁を低くしたい
 - * 多くの分野への応用
 - * スパコンやクラウドを有効利用できる使い易いソフトウェアの開発と配布
- * データ中心科学の有用なツールとしてのデータ同化がより広く使えることでビッグデータの**科学的な**有効活用が可能となると期待している

