jh250074

高レイノルズ数乱流のデータ駆動科学プラットフォームの構築

石原　卓（岡山大学）

概要

「京」,「富岳」の性能を最大限に活用して構築してきたフーリエ・スペクトル法に基づく非圧縮性乱流，および 8 次精度コンパクト差分法に基づく圧縮性乱流の大規模直接数値計算（DNS）のデータベースを維持・管理し，共有することにより日本の乱流の計算科学とデータ駆動科学の発展に貢献するためのプラットフォームを構築することが本研究の目的である．本年度は mdx を用いて公開する乱流データを追加しデータ公開実験を行うとともに，名古屋大学不老 Type2 および九州大学玄界ノードグループ B を用いて，乱流 DNS データのリレーショナルデータベースサーバーを立ち上げ GPU を用いた高速データ活用環境の構築とデータ活用を実施した．

1. 共同研究に関する情報

⑴ 共同利用・共同研究を実施している拠点名

名古屋大学 情報基盤センター

九州大学 情報基盤研究開発センター

mdx I

(2) 課題分野

データ科学・データ利活用課題分野

(3) 参加研究者一覧と役割分担

石原卓(岡山大学)：総括，データ構築・管理

岡本直也(愛知工業大学)：プログラム開発，データ解析

横川三津夫(東北大学)　：コーディネーター

宇野篤也(防災科学技術研究所)：可視化

大島聡史(九州大学)：基盤センター活用技術指導

片桐孝洋(名古屋大学)　：HPC 技術指導

櫻井幹記(横浜国立大学)：DNS プログラム作成，DNS データ構築

金田行雄,芳松克則(名古屋大学)：データ解析

山本和寿（岡山大学）：データベース構築

金尾祐伸（岡山大学）：データ活用

坂井恵（アートライ）：データベース環境開発

2. 研究の目的と意義

【目的】我々のグループはこれまで地球シミュレータ，京，富岳の性能を最大限に活用して，乱流の大規模直接数値計算(DNS)データベースを構築してきた．これらの乱流データベースは，容易には構築できない貴重な乱流データであり，近年の AI 技術の発展により，教師データとしての価値が改めて高く評価されている．そこで本研究では，これらの一連の乱流データ，すなわち多様なレイノルズ数と解像度を有する乱流 DNS データを維持・管理し，共有するためのプラットフォームを構築する．これにより，日本における乱流の計算科学およびデータ駆動科学の発展に貢献することを目的とする．本年度は mdx を用いたデータ共有実験を行いつつ，GPU を用いた乱流データに対して関係データベース管理システムの活用実験を実施し，乱流 DNS データ駆動科学の新たな可能性を追求する．

【意義】我々のグループが実施してきた大規模乱流 DNS で得られた非圧縮性/非圧縮性の乱流データはいずれも国際的にも貴重なものとして知られており，そのデータ解析や可視化のみならずデータを利活用した数値実験やデータ駆動計算によって多くの知見および発見や現実的な乱流現象を

理解するためのヒントを得ることが可能である．

現在，大規模な乱流データを国際的に公開しているプラットホームとしては，Johns Hopkins 大学の Johns Hopkins Turbulence Databases（JHTDB）が有名であり，一様等方性乱流のみならず壁乱流や MHD 乱流のデータセットも含み，スナップショットのみならず時系列データの切り出しや解析環境の提供も行なっており，多くのユーザーに使用されている．しかしながら，その乱流データ構築の条件やデータの解像度に一貫性があるわけではないため乱流統計量のレイノルズ数依存性を見ているのか解像度依存性を見ているのかの切り分けが困難である．また，リレーショナルデータベース（RDB）ではないため SQL 言語を用いたデータ操作はできないものとなっている．

一方，我々の一様等方性乱流 DNS データは，表 1 で示すように，格子点数が最大で16384^3，解像度は$k_{max}\eta = 1, 2, 4, 8$,テイラー長に基づくレイノルズ数は最大で$R_\lambda = 2300$となっており，ほぼ同一のR_λの値に対して複数の解像度，および,ほぼ同一の解像度に対して複数のR_λのデータセットとなっており，組織的な解析が可能である．また，大小の渦の振幅比が 10^8 を超えるような，最大格子点数の乱流 DNS も全て演算は倍精度で行なって得られたデータあるため丸め誤差の影響は小さいものとなっている．

Rλ	Run1	Run2	Run4	Run8
700	2048-1(4.7)	4096-2(1.2)	8192-4(0.4)	**16384-8(0.005)**
1100	4096-1(2.4)	8192-2(0.8)	**16384-4(0.008)**	
1400	6144-1(1.7)	**12288-2(0.1)**	**24576-4(0.002)**	
1750	8192-1(1.2)	**16384-2(0.016)**		
2300	12288-1(0.8)			
2800	**16384-1**			

表 1:乱流 DNS データベースの格子点数Nと解像度の$k_{max}\eta$組み合わせ:N-$k_{max}\eta$とテイラー長に基づくレイノルズ数R_λの値．() 内の数字はエディターンオーバータイムTで規格化した積分時間．

そこで日本を中心とした研究グループで信頼性の高い貴重な乱流の大規模 DNS データベースを維持・管理し，データと解析ツールを共有することで更なる大規模計算を目指す計算科学や多様な知見を得るためのデータ駆動科学を推進するプラットホームを構築することは意義深いことであると考えられる．

3. 当拠点の公募型共同研究として実施した意義

本研究では最大格子点数 16384^3 の大規模な乱流 DNS データを扱う．乱流 DNS ではフーリエスペクトル法を用いており，DNS の結果の保存は波数空間データで行なっている．しかし波数空間データを扱うには大規模な並列計算でフーリエ変換が必要となるため，データの共有と活用に適さない．従って，共有するデータは乱流場の速度や圧力，エネルギー散逸率などの多様な実空間データである．そのため必要なディスクスペースは膨大となる．これらの大規模かつ膨大なデータベースの保存・維持・管理を実施するためには大規模ストレージを有する名古屋大学の情報基盤センターと九州大学の情報基盤研究開発センターの特別なサポートが必要・不可欠である．

本課題では名古屋大学 Type2 や九州大学ノードグループ B の GPU を用いて高速に SQL クエリ実行ができる乱流データベースサーバーの開発を進めている．国産の新しい GPU 活用技術（PG-Strom）を用いているところが特徴である．データベースサーバーの開発とその活用の試行錯誤を繰り返し，実用的な乱流データ活用基盤の共有を行うためには，名古屋大学の情報基盤センターや情報基盤研究開発センターの計算資源や技術協力や情報提供等が必要不可欠である．

4. 前年度までに得られた研究成果の概要

(1) フーリエ･スペクトル法に基づく乱流 DNS おける丸め誤差の影響を解析し，$R_\lambda = 263$の乱流 DNS では単精度演算で 1.6T 以上時間積分すると倍精度演算と結果が異なることを明らかにし，この差はレイノルズ数が大きいほど顕著になることを明らかにし，論文として発表した．

(2) 最大16384^3の乱流 DNS データを処理するプログラムを整備し，大規模データの部分領域

の paraview を用いた可視化を実施した．そして$R_\lambda = 700$の乱流場中のエネルギー散逸率の値が最大となる領域が$k_{max}\eta = 8$の DNS で十分に解像できていることを明らかにした．

(3) 省メモリ可視化ソフトを整備し，格子点数2048^3の圧縮性乱流の計算領域全域における可視化を実施した．その結果，圧縮性乱流において等温を仮定した場合と仮定しない場合で速度発散が負で絶対値が大きくなるショックレット領域の分布が異なることを明らかにした．

(4) mdx を用いてデータベースサーバーのプロトタイプを立ち上げ，格子点数4096^3の乱流場のエネルギー散逸率の部分領域を切り出して download できるようにした．

(5) mdx を用いて，大規模データへの拡張性のためデータの保存形式として Zarr ストレージ形式を採用した．また，クライアントからのリクエストを受け取り，それを適切 なバックエンド（Web サーバ）に転送する役割を担うリバースプロキシとして Nginx（https://nginx.org/） を採用，リバースプロキシから転送されたリクエストを処理し，静的コンテンツ及び動的コンテンツをクライアントに提供する役割を果たす Web サーバを Next.js（https://nextjs.org/）を使用して構築した．

(6) 乱流 DNS データ活用のための web ページ（https://www.turbulencebox.jp/）を立ち上げ，その web ページ上でデータの切り出しと可視化を実現した．

5. 今年度の研究成果の詳細

(1) 【データ整備】名古屋大学情報基盤センターおよび九州大学情報基盤研究開発センターのストレージを活用して，京や富岳で構築したスペクトル法に基づく非圧縮性乱流 DNS データ，および高精度・ 高解像度なコンパクト差分法に基づく等温/非等温の圧縮性乱流 DNS データを整備し，データ活用基盤を整備した．名古屋大学情報基盤センターのスーパーコンピュータのリプレイスに伴い，重要な乱流 DNS データ（表 1 および図 1 参照）を選別し，共用ストレージや九州大学情報基盤研究開発センターのストレージ等を用いてデータの大規模な転送を行った．こうして乱流基盤データの維持・管理を実現した．

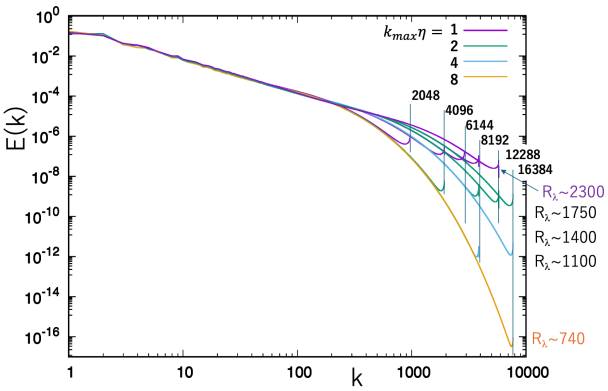

図 1 本研究で整備とデータ活用を進めている乱流データのエネルギースペクトル．格子点数16384^3, $k_{max}\eta \geq 4$のデータでは大小の渦の振幅比が10^8を超えることが確認できる．

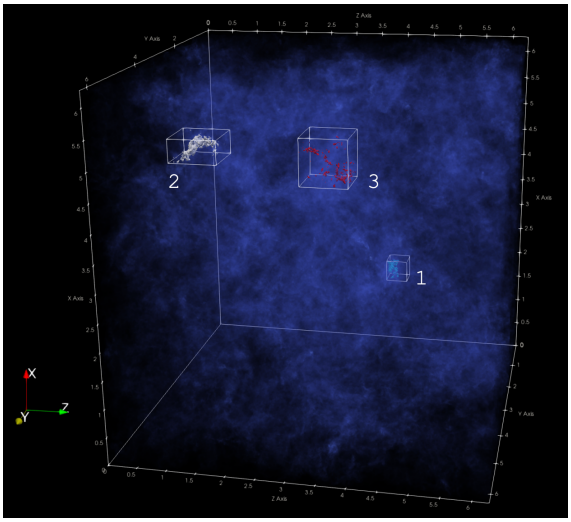

図 2 高レイノルズ数乱流場（格子点数12288^3, $R_\lambda = 2300$）のエネルギー散逸率のサマリーデータの可視化. 16^3個に分割した小領域中のエネルギー散逸率の局所平均 top3 を含む領域における特徴的な散逸構造の抽出

(2) 【データ活用】高レイノルズ数乱流場（格子点数12288^3, $R_\lambda = 2300$）のエネルギー散逸率と

エンストロフィーのサマリーデータ(図 2)を作成し，特徴的な領域の抽出を行い，その領域の解析を行った（図 3）.

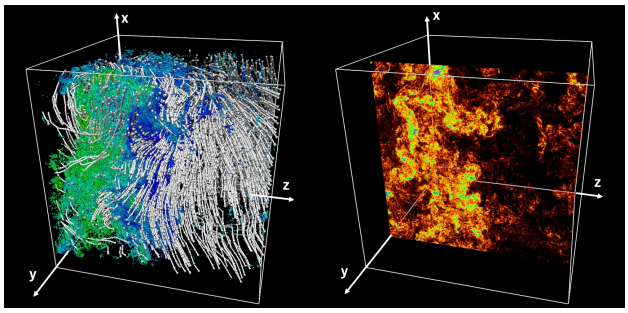

図 3 高レイノルズ数乱流場（格子点数12288^3, $R_\lambda = 2300$）中のエネルギー散逸率の局所平均が最大となる分割小領域（図 2 参照）の可視化．(左)エネルギー散逸率の等値面を速度の y 成分で色付けしたものと流線の可視化．大規模な剪断が形成されていることが確認できる．(右）断面におけるエネルギー散逸率のコンター図．大規模な層構造の中に小規模な層構造があることが確認できる．

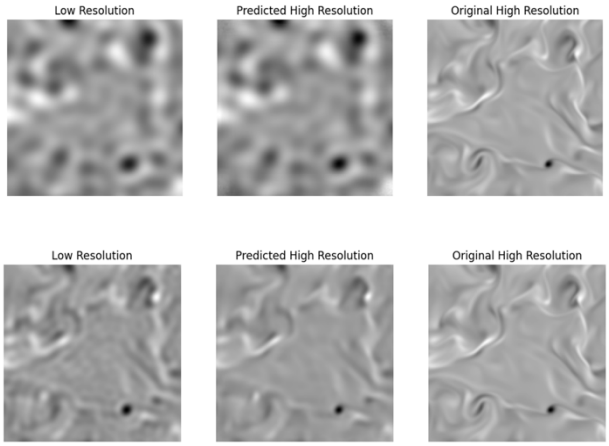

図 4 乱流 DNS の渦度データを用いた超解像の数値実験．上は$k_c\eta = 0.3$,下は$k_c\eta = 0.6$の低解像画像(左)から生成した超解像画像(中)．上と比べ下は超解像により高解像に近い結果が得られている．

(3) 【データ活用】 高精度・ 高解像度なコンパクト差分法を用いた等温/非等温の圧縮性乱流 DNS で作成したデータから，ショックレット領域（速度の発散が負で絶対値の大きくつながった領域）を抽出し，その大きさと数について等温/非等温の違いやレイノルズ数依存性についての統計解析を行った．その結果，強いショックレット領域の形状は基本的に 1 方向に薄く，平面状から棒状に分布することが定量的に確認できた．

(4) 【データ活用】非圧縮性乱流 DNS データから渦度成分や速度成分の 2 次元データを抽出し，超解像の機械学習実験を実施した．その結果，粗視化の特徴的な波数k_cが$k_c\eta \geq 0.5$を満たす場合には超解像モデルが良好に渦度の低解像画像から高解像画像を再構築できることが確認できた（図 4）.

(5) 【データ活用】非圧縮性乱流 DNS データ（格子点数 512^3～8192^3）を用いて，粗視化したスケールの流体の変形とそのスケールで平均したエネルギー散逸率の相関のデータ解析を行った．その結果，粗視化したスケールにおける剪断が強い領域でエネルギー散逸率の局所平均値が高い傾向があることが確認できた．

(6) 【データ公開】mdx を用いたデータ公開用 web サーバー（https://www.turbulencebox.jp/）で，格子点数 4096^3の乱流 DNS によって得られたエネルギー散逸率と渦度の大きさの場のデータを Zarr 形式で保存し，web のインターフェースで指定した領域のデータを download および可視化できるようにし，データの公開実験を行った．

(7) 【データ活用基盤構築】(1)で整備したデータを PostgreSQL のデータベースサーバーで取り扱うためのデータ活用基盤構築を名古屋大学「不老」の Type2 および九州大学「玄界」のノードグループ B を用いて進めた．

Singularity を用いて PostgreSQL の RDB サーバーを計算ノードで立ち上げ，Apache Arrow 形式に変換した乱流データ（格子点数4096^3の速度 3 成分，エネルギー散逸率および渦度の大きさの場の実空間データ）を PostgreSQL のテーブルとして登録することに成功し，PG-Strom (https://www.heterodb.com)を用いて，GPU による SQL クエリの高速実行を実現した．その結果，乱流場中のエネルギー散逸率や渦度の大きさの平均値や分散値が大きいなど，特

徴的な領域の情報を見ることが可能なサマリーテーブルを効率的に作成でき，それを参照しながら興味のある領域のデータ抽出ができるようになった．また，速度場データのみからSQL クエリでエネルギー散逸や渦度などの近似値も計算可能であることが確認できた．

図 5 開発した可視化システムの GUI．サマリーデータを参考にして指定した領域の可視化が可能．

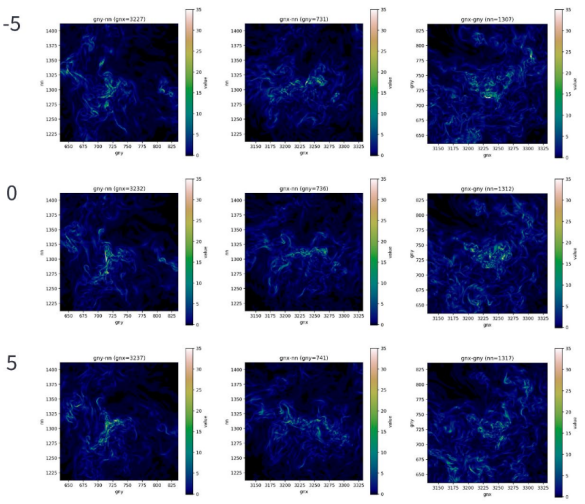

図 6 開発した可視化システムのブラウザ上に表示された可視化結果の例

(8) 【データ活用基盤構築】玄界ノードクラス B を用いて，乱流データベースから極端現象を発見・抽出するためのデータ探索・可視化システムを構築した．開発したシステムでは乱流 DNS の波数空間データから生成された，速度 3 成分，エネルギー散逸率，エンストロフィーの実空間データを Apache Arrow 形式のファイルに変換することで PG-Strom から PostgreSQL の RDB テーブルとして認識させ，SQL クエリの高速実行可能な状態にする．そして，別ノードから Streamlit で構築したシステムが PG-Strom にアクセスしてデータ取得を行い，手元のブラウザから Streamlit にアクセスして可視化を行う（図 5 および図 6）．

6. 進捗状況の自己評価と今後の展望

昨年度 mdx 上で開発したシステムでは，乱流データを Zarr 形式で保存し，Web ブラウザから任意領域の切り出しと可視化を行う機能を実現した．これに対し，今年度開発した 5 の(7)および(8)のシステムでは，データを Apache Arrow 形式で保存し，PG-Strom からテーブルとして認識させることで，SQL の高速実行と，Streamlit を介した Web ブラウザ上での可視化を実現している．

両システムは，いずれもデータの切り出しと可視化という基本機能を有している．一方で，今年度のシステムは，(1) Apache Arrow 形式のファイルを高速に作成できること，(2) PG-Strom においてテーブルとして認識されたデータに対して，GPU を用いた SQL クエリの高速実行が可能であり，データ探索機能を備えていること，という特徴がある．さらに，Apache Arrow 形式の使用により，RDB として登録する際に生じるデータサイズの増加を回避できている．以上より，SQL 機能を有する乱流データの RDB サーバーを構築するという当初計画を達成できており，進捗は 100%であると判断できる．

今後は，新たに開発した機能を活用して，データ駆動科学に基づく乱流データの利活用を実践していくとともに，mdx 上で PG-Strom を活用したシステムの構築を試みる予定である．

※7. 研究業績はウェブ入力です