jh240062

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 上にデータベースサーバーと web サーバーを立ち上げ、web 上のユーザーインターフェースを通じて DNS データの切り出しとダウンロードや情報抽出と可視化を実施する環境の開発を行った.

1.　共同研究に関する情報

(1) 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

(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

横川三津夫(神戸大学)：コーディネータ

畑中裕翔(神戸大学)：mdx 上の webpage 作成

大島聡史(九州大学)：mdx 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

宇野篤也(防災科学技術研究所)：可視化

今村俊幸(理化学研究所)、福井匠(神戸大学)：プログラム開発

金田行雄、芳松克則(名古屋大学):データ解析

岡本直也(愛知工業大学)、関本敦(岡山大学):プログラム開発、データ解析

2.　研究の目的と意義

【目的】我々のグループが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構築してきた、乱流の大規模直接数値計算(DNS)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。

【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。実際、発達した乱流場のデータと乱流 DNS コードを活用した数値実験が天文分野や気象分野で活用され成果が得られている。そこで日本を中心とした研究グループで乱流の大規模 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

本研究では乱流の大規模 DNS の結果を活用するためのプラットフォームの構築を mdx（様々なデータやソフトウエアなどを迅速かつ効率的に連携させたデータ活用の実現を目的として導入されたプラットフォーム）を用いて試験的に実施してみることをメインな課題の一つとした。実際に mdx を利用するにあたっては mdx の仕組みや詳細な設定に詳しい研究者との共同研究が不可欠であったので、公募型共同研究として実施した意義は十分にあった。

4. 前年度までに得られた研究成果の概要

該当なし

5. 今年度の研究成果の詳細

(1) 京や富岳を用いた構築した乱流の大規模データに対するデータ処理・解析・可視化ツールの整備を行った。その結果、乱流の大規模 DNS による、エネルギー散逸率とエンストロフィーの実空間データの指定した任意の領域が容易に取り出せ、可視化ソフト paraview を用いて可視化できるようになった。また、省メモリ可視化ソフトを自前で開発し、複数のスカラー場の同時表示など種々の目的を満たす可視化が実現できることを確認した。

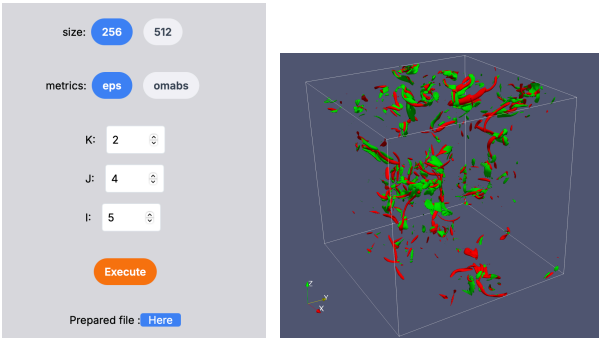

図 1.(左)mdx の Web サーバー上の DNS データ領域指定のためのインターフェース、(右)Download したデータの Paraview による可視化結果(赤：高渦度領域、緑：高エネルギー散逸領域

(2) mdx 上に乱流 DNS のデータサーバーを構築し、データ公開用の試験用 Web サーバーを立ち上げた。データサーバー上に乱流 DNS で作成した格子点数 4096^3 のエネルギー散逸率とエンストロフィーの実空間データの全領域分、および格子点数 8192^3 のエネルギー散逸率の実空間データの部分領域分を試験的に upload し、Web ブラウザのインターフェースに入力した情報によって、指定した任意の領域の paraview 可視化用データが download できるような実装に成功した（図 1)。

このデータ管理の方法は web サーバーの GUI に入力した内容に応じて、あらかじめ用意したプログラムが実行され、大規模 DNS で書き出されたデータを直接扱うため、必要なストレージのサイズはオリジナルデータサイズと一致し、データ切り出しも比較的高速であったが汎用性が乏しいと感じられた。

(3) データ活用に汎用性を持たせるためのデータベースソフトウェアの活用が考えられた。本研究では PostgreSQL の使用と Zarr の使用を試み、両者の比較を行なった。PostgreSQL の多次元データを管理するための拡張機能として PostGIS を使用する場合は各物理量の位置情報も格納する必要が生じる。時間 1 ステップ分の単精度乱流データ(各方向の速度，圧力，渦度の絶対値)の合計 20GiB を PostgreSQL のテーブルとし smallint 型の位置座標データと共に格納した場合、そのテーブルサイズは約 3 倍の 60GiB となった。一方、圧縮アルゴリズムを LZ4 として Zarr ストレージ形式でデータを保存した場合、そのサイズは約 0.85 倍の 17 GiB となった。この結果からはストレージの節約という観点からは Zarr のストレージ形式が優れていることが分かった。

(4) 以上の mdx 活用試験と事前調査の結果を踏まえ、Web 上で可視化が行え、小規模 DNS データを提供するのみならず、大規模データへの拡張性も有する、乱流 DNS データ活用システムの構築を行なった。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。mdz 上に構築したシステムはリバースプロキシと Web サーバーで構成されている。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/）を採用した。また、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 サーバの構築は Next.js（https://nextjs.org/）を使用して行なった。

(5) 乱流データ活用システムを構築した結果、https://www.turbulencebox.jp/ にて「Data」ページでは、図 2 のように本 Web アプリケーションで公開しているデータの種類を確認でき、「Download」ページでは、図 3 のようにフォームと進捗状況が表示されてい る.フォームを入力した後、ボタン「Download」をクリックすることで，フォー ム（データベースの種類、時間、変数、取得領域の原点、取得領域のサイズ、ファイル形式を選択できる）で指定した条件の DNS データをダウンロードでき、「Visualization」ページでは、図 4 のようにフォームで条件を指定したのち[Visualize]ボタンをクリックすると、フォームで指定した条件の DNS データが Web ブラウザ上で可視化できるようになった。

6. 進捗状況の自己評価と今後の展望

2024 年度は mdx を活用して、乱流 DNS データのデータ駆動科学プラットフォーム構築を目標とした共同研究を推進した。データ公開用の Web ページとしては完成形に近いものが得られた。より大規模のデータの公開についてはデータの管理や運用の仕方も含め、今後の課題となる。

なお、PostgreSQL においても Apache Arrow 形式のファイルを用いることで位置情報のテーブルが不要となることや PG-Strom による GPU の活用で高速にデータ処理が可能であることが確認できた。今後は PostgreSQL や PG-Strom の活用による情報の抽出や発見を行うデータ活用実験を行い、乱流 DNS データベース活用の新しい可能性についても追求していきたいと考えている。

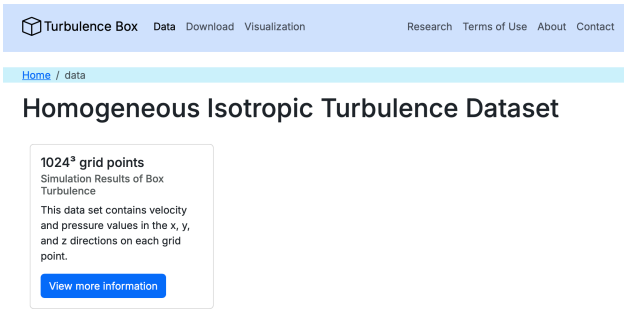

図 2 「Data」ページ

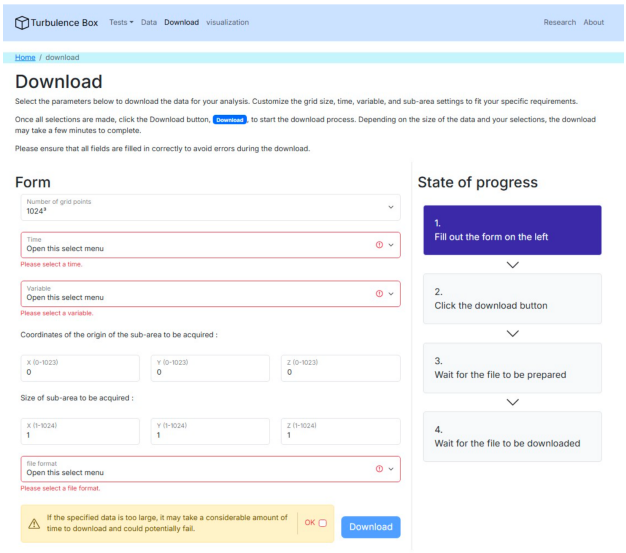

図 3 「Download」ページ

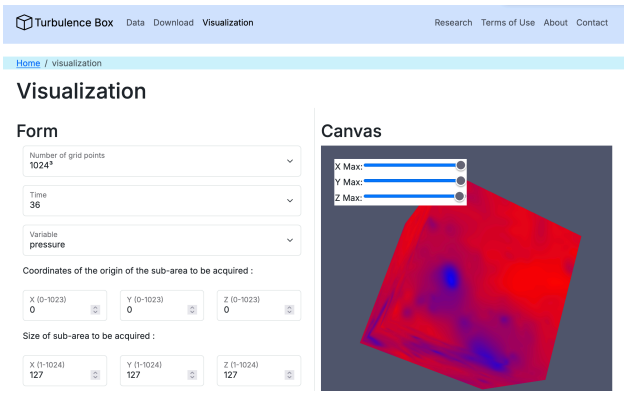

図 4 「Visualization」ページ