大規模アンサンブル気候データの効率的な解析に向けたコンテンツベース検索システム

中川　友進*・尾上　洋介**・荒木　文明*・小山田　耕二***

A Content-Based Database System for Large Ensemble Future Climate Data

Yujin Nakagawa*, Yosuke Onoue**, Fumiaki Araki*and Koji Koyamada***

Key words: Climate Data, Relational Database, Web Application

1. はじめに

コンピュータの性能の向上に伴い，将来気候の大規模なアンサンブルデータ（以後，大規模アンサンブル気候データ）のデータサイズは急激に増加している．例えば，「気候変動リスク情報創生プログラム」により作成された「地球温暖化対策に資するアンサンブル気候予測データベース」（the database for Policy Decision making for Future climate change; d4PDF）[1] の場合，データサイズは約 3 PB である．このような大規模アンサンブル気候データの系統的な解析は，確率的な気候変動の影響評価を行うために非常に重要である．しかし系統的な解析には，一般に高性能なコンピュータだけでなく，大容量のデータ記憶装置が必要である．そのため，個々の研究者による系統的な解析は困難になりつつある．

「気候変動適応技術社会実装プログラム」（Social Implementation Program on Climate Change Adaptation Technology; SI-CAT）は，将来の気候変動に対する適応策に資する技術の開発を目的とした国家プロジェクトである．SI-CAT では，地球の平均気温が産業革命以降 2℃上昇した将来気候のアンサンブルデータを作成した[2]．そのデータは d4PDF の一部として，「データ統合・解析システム」（Data Integration and Analysis System Program; DIAS; http://www.diasjp.net）で公開されている．

d4PDF のデータサイズは約 3 PB であり，過去のデータセットと比較して，とても大きい．DIAS において，d4PDF が保管されているデータサーバーと同じネットワークに，d4PDF のための解析サーバーがあれば，ユーザーはデータをダウンロードせずに解析を行うことができる．しかし，現状では d4PDF のための解析サーバーは無いため，ユーザーは d4PDF を自身のローカルコンピュータにダウンロードする必要がある．この場合，d4PDF のデータサイズが大きいことにより以下の問題が発生すると考えている．

1）d4PDF をダウンロードするユーザーのディスク容量が不足する．
2）DIAS のデータサーバーからユーザーのローカルコンピュータへのダウンロードに長い時間かかる．
3）データサーバーへ高い負荷がかかる．

これらの問題を解決するためには，ユーザーが要求を満たすデータを検索してダウンロードできる機能を有したユーザーフレンドリーなシステムが必要である．

DIAS や WDCC（https://cera-www.dkrz.de/WDCC/ui/cerasearch）など，気候データのための従来の Web ベースの検索システムは，気候の研究分野において検索や可視化に活用されている．これら全ての従来システムは，データファイルに格納されている物理量を使用してデータを検索するようには設計されておらず，データファイルそのものに関連付けられているメタデータを使用してデータファイルを検索するように設計されている．気候データのデータサイズが，ユーザーのローカルコンピュータにダウンロードすることができる程度であれば，これらの従来システムはとても有用である．

本研究では，ユーザーが必要とするデータを検索す

* 国立研究開発法人海洋研究開発機構
Japan Agency for Marine-Earth Science and Technology
** 日本大学
Nihon University
*** 京都大学
Kyoto University

るサービスを提供するために, SI-CAT の下で「SI-CAT 気候実験データベースシステム」(System for Efficient content-based retrieval to Analyze Large volume climate data; SEAL) を開発している．SEAL を使用すると，物理量など，データファイルの内容に関連付けられているメタデータを使用してユーザーがデータファイルを見つけることができる. そのため SEAL は, ユーザーがデータサーバーからローカルコンピュータにダウンロードするデータサイズを大幅に削減できる.

2. 設計と実装

2.1 データ

SEAL の設計に使用した d4PDF は，全球シミュレーションデータ，および領域シミュレーションデータで構成される．全球シミュレーションデータは，気象研究所が開発した水平格子間隔 60 km の全球大気モデル (MRI-AGCM[3]) を用いて作成された．領域シミュレーションデータは，気象研究所によって開発された水平格子間隔 20 km の非静力学地域気候モデル (MRI-NHRCM[4,5]) を用いて作成されており，将来 4℃昇温実験，将来 2℃昇温実験，過去実験で構成されている．これらのデータのうち，領域シミュレーションデータの地上大気データ[1,2] を使用した．地上大気データは「GRIdded Binary」(GRIB) 形式で保管されており，時間分解能は 1 時間である.

2.2 基本コンセプト

SEAL の設計は，以下の 3 つの基本コンセプトに基づいて行なった. 第 1 のコンセプトは, ユーザーのニーズを満たすための実用性を持たせることである．ユーザーとして想定される SI-CAT に所属する影響評価の研究者に対してニーズ調査を行なったところ，物理量としては降水量と気温が重要であることが分かった. また検索条件としては，「ある行政区域の指定期間における，3 日間積算降水量の上位から順番に，指定したケース数だけ表示する.」などが重要であることが分かった．第 2 のコンセプトは，生データ (GRIB 形式のデータ) に格納されている物理量を出来るだけ加工しないことである．そのため，バイアス補正など，ユーザーに依存する加工は行わないこととした．第 3 のコンセプトは，SEAL のために開発された技術を様々なデータセットに適用するための汎用性である.

図 1 に SEAL の概念図を示す．SEAL は, リレーショナルデータベース，データ提供機能，およびユーザーインターフェースで構成されている．これら 3 つ機能のうち，時間的，空間的に圧縮されたデータを登録するように設計されている PostgreSQL を使用したリレーショナルデータベースが重要な役割を担っている．データ提供機能は，ユーザーが検索結果に基づいて時間的，空間的に切り出されたデータをダウンロードする機能を提供する. さらに, Web ベースのユーザーインターフェースにより，ユーザーは PostgreSQL の知識がなくてもリレーショナルデータベースを簡単に使用できる.

2.3 リレーショナルデータベースの設計

ユーザーは，自身が研究の対象としている特定の領域に関心があるため，単一の格子点ではなく，複数の格子点の組み合わせによって定義される領域 (行政区域や流域など) における物理量を必要としている．さらに，ユーザーは，気温のような物理量については，時別値ではなく，日別値，月別値または年別値を必要とする．そこで，d4PDF の物理量に対して，空間的，時間的な圧縮を行い，それらの圧縮された物理量をリレーショナルデータベースへ登録することとした．ここで，空間的圧縮とは，複数の格子点における物理量

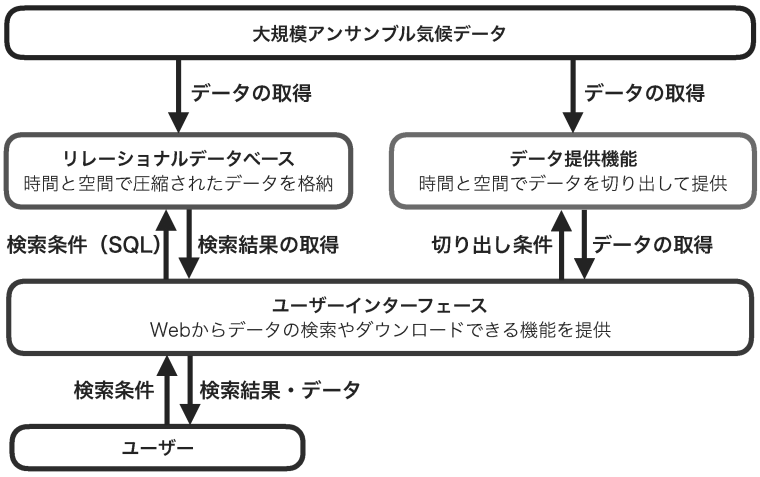

図 1 SEAL の概念図.

を平均することである．また，時間的圧縮とは，時別値を積算（例えば降水量の場合）または平均（例えば気温の場合）することにより，日別値，月別値，年別値を計算することである．

2.4 データ提供機能の設計

データ提供機能により，ユーザーは，リレーショナルデータベースによる検索結果に基づいて，時間的，空間的に切り出された生データをダウンロードすることができる．DIAS は，ユーザーが「Grid Analysis and Display System」（GrADS）のバイナリデータ（ビッグエンディアン，ヘッダーなしの 4 バイト浮動小数点）をダウンロードできる機能を提供している．しかし，ほとんどのユーザーは，テキスト形式や CSV 形式などの人間が読める形式へ変換したデータを必要としている．そこで，GrADS バイナリデータをテキスト形式または CSV 形式へ変換する機能を Python スクリプトで開発した．

2.5 ユーザーインターフェースの設計

ユーザーからのニーズを考えると，ユーザーが GRIB 形式や PostgreSQL に関する知識がなくても検索を行えるシステムであることが望ましい．また，コマンドラインインターフェースの知識がなくても GrADS バイナリデータをテキスト形式または CSV 形式へ変換できることが望ましい．そこで我々は，これらの知識がなくてもリレーショナルデータベースとデータ提供機能を利用できる Web ベースのユーザーインターフェース（Web UI）を開発した．Web UI は，条件に合うデータを見付けることを主な機能とした「SEAL-Finder」（SEAL-F），データを分析および可視化することを主な機能とした「SEAL-Visualizer」（SEAL-V）で構成されている．そのため，ユーザーは，SEAL-F を使用すると検索の返り値を数値として取得することができ，また SEAL-V を使用すると検索の返り値をヒストグラムなどの画像として取得することができる．

2.6 リレーショナルデータベースの実装

検索の高速化と高い実用性の両方を満たすために，降水量は，時別値と日別値の 2 つの異なる時間分解能のデータをリレーショナルデータベースへ登録することとした．また，気温は，日平均値，日最大値，日最小値を登録することとした．ほとんどの場合，ユーザーは自身が研究の対象とする行政区域の物理量を必要とする．本研究で使用している地上大気データの空間分解能が 20 km であることを考慮して，47 都道府県それぞれについて，行政区域と重なる格子点の物理量を積算して，リレーショナルデータベースへ登録することとした．

2.7 ケーススタディ

2.7.1 空間的圧縮，時間的圧縮による圧縮率

空間的圧縮，時間的圧縮を適用した際に，生データと比べて，どのくらいデータサイズが縮小されるかを定量的に調べるため，それらの圧縮率を計算した．空間的圧縮の圧縮率は，北海道（302 個の格子点で構成）では 1/302 ≒ 0.33%，東京都と大阪府（それぞれ 13 個の格子点で構成）では 1/13 ≒ 7.7%となる．また，時間的圧縮の圧縮率は，時別値から日別値へ圧縮した場合は 1/24 ≒ 4.2%となる．そのため，空間的圧縮と時間的圧縮の両方を適用した場合，日別値のデータサイズは，最大で 0.01%，最小で 0.3%へ縮小する．データサイズの縮小は，従来システムを使用する方法と比較して，必要なデータファイルを取得するまでの所要時間の短縮に繋がる．

2.7.2 必要なデータファイルの取得までの所要時間

SEAL の利点を明確にするために，従来システムを使用した方法と国立研究開発法人海洋研究開発機構（Japan Agency for Marine-Earth Science and Technology; JAMSTEC）のローカルサーバー上の SEAL を使用した方法について，必要なデータファイルを取得するまでの所要時間を定量的に調べた．ローカルサーバーは，Intel Xeon E7-4820（CPU; 40 コア）と 512 GB の物理メモリを搭載している．すべての解析において 1 つの CPU コアを使用した．**図 2** は SEAL の利点の概略図であり，北海道で降水量の日別値が 100 mm を超える日について，将来 4℃昇温実験の降水量と気温の時別値のデータファイルをダウンロードする状況を想定している．従来システムを使用した方法では，北海道周辺の降水量の時別値のデータファイルのダウンロードに 3〜90 時間，条件を満たす日にちの探索に約 31 時間，条件を満たす日にちにおける気温の時別値のデータファイルのダウンロードに 1〜30 分を要する．一方，SEAL を使用した方法では，条件を満たす日にちを約 5 秒で見付けることができ，条件を満たす日にちの降水量と気温の時別値のデータファイルを 2〜60 分でダウンロードできる．ダウンロードする時別値のデータファイルのデータサイズは，生データと比較して約 0.5%へ減少する．したがって，従来のシステムを使用した方法と比較して，SEAL は必要なデータファイルを取得するまでの所要時間を 1%未満へ減らすことができる．

2.7.3 大雨の検索

SEAL の実用性を明確にするために，河川工学を専門とするユーザーが利用することが多いと考えられる

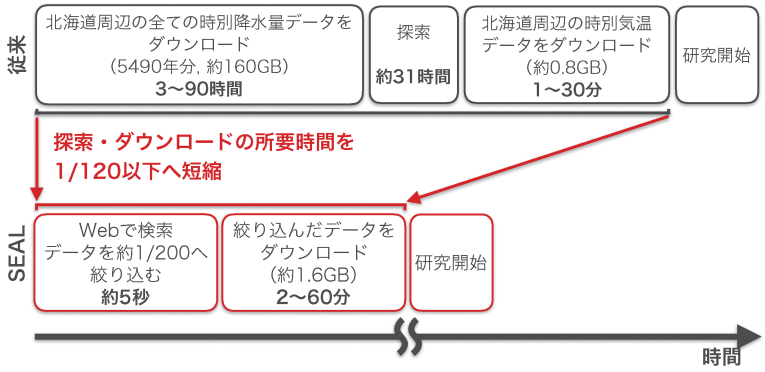

図 2 SEAL の予想される利点の概略図．図に示されている経過時間は，DIAS から JAMSTEC へのダウンロード速度(4〜120 Mbps)に基づく，将来 4 度昇温実験のデータを使用した際の実測値である．

検索条件を想定し，ローカルサーバー上で動作する SEAL を用いて検索を行った．将来 4℃昇温実験の東京都のデータのうち，連続降水日数が 20 日を超え，かつ累積降水量が 800 mm を超えるという検索条件を満たすイベントを抽出した．ここで，降水量の日別値が 0.1 mm を超える日を降水日と定義している．その結果，検索に約 30 秒かかり，条件を満たす 6 つのイベントが見つかった．その中でも，累積降水量が最大のイベントは, 24 日間に約 1,109 mm を示した（**図 3**）．累積降水量の等高線図を調べたところ，このイベントは，静岡県を中心とする大雨であった．従来の方法で上記の現象を調べた場合，東京都に対応する降水量の時別値のデータファイルのダウンロードに 0.13〜3.9 時間かかる．さらに，ユーザーは自身で連続降水日数と累積降水量を計算する必要があるため，さらに時間がかかる．したがって，SEAL は必要なデータファイルを取得するまでの所要時間を短縮する十分な能力を持っている．

3. 考察

従来のフレームワークの 1 つである OPeNDAP（https://www.opendap.org）は，単一の時系列グリッドデータ（すなわち，単一のデータファイル）から値を検索する機能を提供している．しかし，データファイル内のすべての値をスキャンするため，検索速度は速くは無い．一方，SEAL は，リレーショナルデータベースを採用することにより, 複数のアンサンブル(すなわち全てのデータファイル）から基準を満たす値を高速に検索できる．これは，データベースインデックスの影響を強く受ける検索については，リレーショナルデータベースは検索の速度が大幅に向上するためである．さらにリレーショナルデータベースは，複数のアンサンブルを一度にスキャンできるという利点を有する．

第 1 章で述べたように，大規模アンサンブル気候データから必要なデータファイルを取得するには 3 つの問題が予想される．すなわち，ユーザーのディスク容量の不足，長時間のダウンロード，データサーバーへの高い負荷，である．第 2.7 節のケーススタディにおいて，ユーザーがローカルコンピュータにデータファイルをダウンロードするデータサイズおよび所要時間は，従来の方法と比較して，それぞれ約 0.5% および約 1% に縮小された．その結果，1 番目と 2 番目の問題が解決された．さらに，データサイズと必要なデータを取得するまでの所要時間が大幅に削減されるため，データサーバーへの負荷が大幅に減少した．したがって，3 番目の問題も解決された．その結果，SEAL は予想どおりに機能し，大規模アンサンブル気候データから必要なデータを取得する際に予想される問題をすべて解決した．

4. 教訓

第 2.2 節で述べたように，SEAL の開発において，ユーザーのニーズを満たす実用性が重要なコンセプトの一つである．そこで，SEAL の開発を始めるにあたり，ユーザーとして想定される研究者へのニーズ調査を行なった．ニーズ調査に際して，回答の自由度が高すぎると，SEAL では実現できない要求が出てくる可能性があり，一方で制限を付けすぎると回答を誘導する可能性がある．そのため，正しくニーズを引き出すためには，アンケートへどのように設問を記載するかという点が重要となる．SEAL の検索において，検索条件，および「どのような戻り値が必要となるか？」という検索結果を問う設問は重要であり，特に慎重に議論を行った．その結果，こちらが想定している具体的な回答例を併記することにより，知りたいニーズを

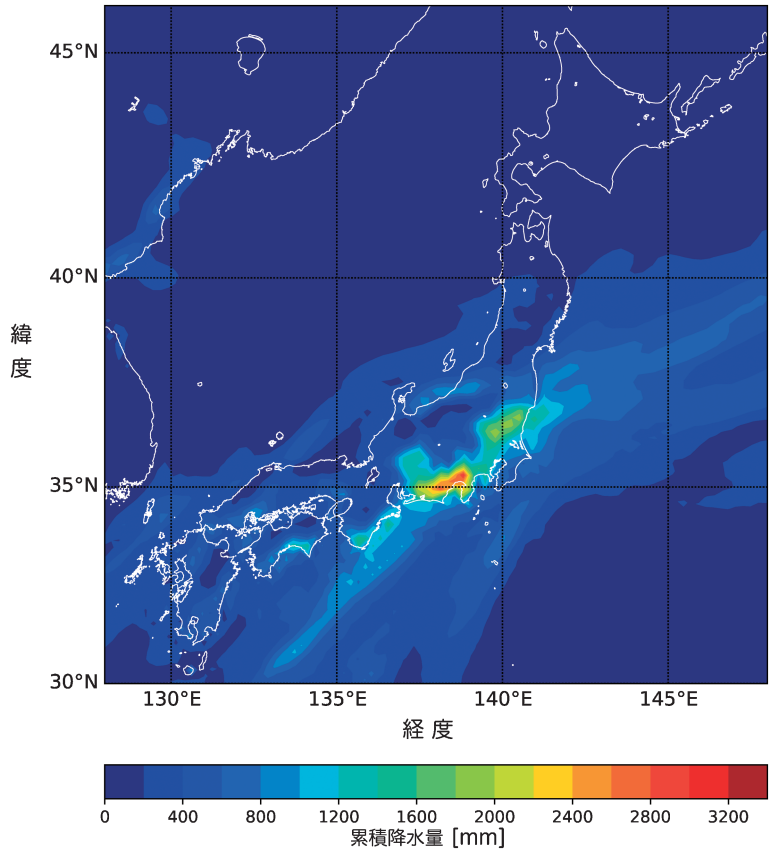

図 3　SEAL を使用して見付けた大雨の累積降水量の等高線図.

得ることができた．ただし戻り値については，d4PDF に格納されていないデータへの要望があったため，戻り値として可能な範囲を示すべきだったと考えている．すなわち，d4PDF の生データに対して基本的に加工は行わないこと（第 2.2 節を参照）を明確にしておくべきだったと考えている．SEAL の開発のチームメンバーだけでなく，他の SI-CAT のメンバーの意見を聞くことにより，設問の問題点を洗い出すことができ，ユーザーのニーズを正しく引き出し，ニーズを満たすシステムを開発できたと考えている．

5. 結論

大規模アンサンブル気候データのデータサイズが増加するにつれて，従来の Web ベースの検索システムを使用した検索に際して，3 つの問題が予想される．すなわち，ユーザーのディスク容量の不足，長時間のダウンロード，データサーバーへの高い負荷，である．これらの問題を解決するために，物理量などのデータファイルの内容に関連付けられたメタデータを使用してユーザーがデータファイルを検索できるようにするシステムである SEAL を開発した．SEAL の使用に際して，ユーザーは GRIB 形式と PostgreSQL の知識は必要とせず，ユーザーは検索結果に基づいて GrADS バイナリ形式，テキスト形式，CSV 形式のデータをダウンロードすることができる．さらに，ユーザーは Web ベースのユーザーインターフェースを介してすべての作業を行うことができる．SEAL のリレーショナルデータベースに格納されるデータサイズは，空間的圧縮および時間的圧縮を採用することにより，生データと比較して最大で 0.01％，最小で 0.3％に縮小される．リレーショナルデータベースは，データベースインデックスの影響を大きく受ける検索に対しては，検索速度を大幅に改善すると共に，複数のアンサンブルを一度にスキャンすることができる．さらに必要なデータの取得に際して，データサイズと所要時間をそれぞれ約 0.5％と約 1％へ減らすことができる．これらの利点により，データサーバーへの負荷が削減される．その結果，SEAL は期待通りに動作し，すべての問題を解決することができた．

SEAL は現在，JAMSTEC のローカルサーバーにおいて運用試験中であり，2019 年度末までに DIAS から公開される予定である．SEAL のために開発された技術は，他の研究分野におけるシミュレーションや観測のデータに対しても，非常に有用だと考えている．

謝辞

本研究は，気候変動適応技術社会実装プログラム（SI-CAT）の一環として行われている．また，本研究は SOUSEI と SI-CAT の下で作成された，地球温暖化対策に資するアンサンブル気候予測データベース（d4PDF）を使用した．

参　考　文　献

1）R. Mizuta, A. Murata, M. Ishii, H. Shiogama, K. Hibino, N. Mori, O. Arakawa, Y. Imada, K. Yoshida, T. Aoyagi, H. Kawase, M. Mori, Y. Okada, T. Shimura, T. Nagatomo, M. Ikeda, H. Endo, M. Nosaka, M. Arai, C. Takahashi, K. Tanaka, T. Takemi, Y. Tachikawa, K. Temur, Y. Kamae, M. Watanabe, H. Sasaki, A. Kitoh, I. Takayabu, E. Nakakita, and M. Kimoto: Over 5,000 years of ensemble future climate simulations by 60-km global and 20-km regional atmospheric models, Bulletin of the American Meteorological Society, vol. 98, no. 7, pp. 1383–1398（2017）.

2）M. Fujita, R. Mizuta, M. Ishii, H. Endo, T. Sato, Y. Okada, S. Kawazoe, S. Sugimoto, K. Ishihara, and S. Watanabe: Precipitation changes in a climate with 2-k surface warming from large ensemble simulations using 60-km global and 20-km regional atmospheric models, Geophysical Research Letters, vol. 46, no. 1, pp. 435–442（2019）.

3）R. Mizuta, H. Yoshimura, H. Murakami, M. Matsueda, H. Endo, T. Ose, K. Kamiguchi, M. Hosaka, M. Sugi, S. Yukimoto, S. Kusunoki, and A. Kitoh: Climate simulations using mri-agcm3.2 with 20-km grid, Journal of the Meteorological Society of Japan. Ser. II, vol. 90A, pp. 233–258（2012）.

4）H. Sasaki, A. Murata, M. Hanafusa, M. Oh'izumi, and K. Kurihara: Reproducibility of present climate in a non-hydrostatic regional climate model nested within an atmosphere general circulation model, SOLA, vol. 7, pp. 173–176（2011）.

5）A. Murata, H. Sasaki, M. Hanafusa, and K. Kurihara: Estimation of urban heat island intensity using biases in surface air temperature simulated by a nonhydrostatic regional climate model, Theoretical and Applied Climatology, vol. 112, pp. 351–361（2013）.