# 気候変動予測データ公開サーバのアクセスログの可視化

根本 利弘† 喜連川 優‡

†東京大学地球観測データ統融合連携研究機構 〒153-8505 東京都目黒区駒場 4-6-1
‡東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1
E-mail: †nemoto@tkl.iis.u-tokyo.ac.jp, ‡kitsure@tkl.iis.u-tokyo.ac.jp

**あらまし** 我々の研究所では、現在 300TB 以上の気候変動予測データの公開を行っている。本稿では、この気候変動予測データおよびデータ公開システムについて説明するとともに、公開サーバにおけるアクセスログを可視化するシステムについて述べる。

**キーワード** 気候変動予測データ，アクセスログ，可視化

# Visualization of Access Log of Climate Change Prediction Data Server

Toshihiro NEMOTO† and Masaru KITSUREGAWA‡

†EDITORIA, The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
‡Institute of Industrial Science, The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
E-mail: †nemoto@tkl.iis.u-tokyo.ac.jp, ‡kitsure@tkl.iis.u-tokyo.ac.jp

## 1. はじめに

地球温暖化は極めて重大な問題となっている。IPCC（Intergovernmental Panel on Climate Change：気候変動に関する政府間パネル）は、2007 年に第 4 次評価報告書を公開し、地球温暖化は明白であり、その原因が人間活動による温室効果ガスによる可能性が高く、地球温暖化に対する適応・緩和に対する努力が不可欠であることを著した[1]。IPCC は、2014 年の完成を目指し第 5 次報告書を作成しており、このために作成されている気候変動予測データが、WCRP（World Climate Research Programme：世界気候研究計画）による CMIP5（Coupled Model Intercomparison Project Phase 5：第 5 次結合モデル比較実験）データである[2]。CMIP5 では、世界の 20 以上の研究機関・モデルによる出力データが ESG（Earth System Grid）と呼ばれる分散システムによりアーカイブ・公開されつつあり[3]、日本より提供される CMIP5 データは、データ統合・解析システム（Data Integration and Analysis System：DIAS）から公開されている。DIAS から公開している CMIP5 データは約 670 万ファイル、350TB に達し、1 日当たり数千ファイルがダウンロードされている。大規模なアーカイブシステムでは、ファイルへのアクセスパターンの把握はシステムの性能へ大きな影響を及ぼす。アクセスパターンを解析することにより、アクセスが予測されるファイルのプリフェッチ、高アクセス頻度フィアルの高性能領域への配置による応答性能の向上、高アクセス頻度ファイルと低アクセス頻度ファイルを分離配置し、低アクセス頻度領域のディスク停止時間を増進させることによる消費電力の削減等が期待される。しかしながら、CMIP5 データは規模が大きく、そのアクセスパターンの把握は容易ではない。そこで、我々は複数の属性を有し、多次元構造を持つ CMIP5 データへのアクセスを三次元可視化し、アクセスパターンを俯瞰的に見ることができるシステムを作成している。本稿では、CMIP5 データおよびデータ公開システムについて簡単に説明するとともに、我々が開発中の DIAS における CMIP5 データに対するアクセスログを可視化するシステムについて述べる。

## 2. 気候変動予測データ（CMIP5 データ）

CMIP5 データは、IPCC の第 5 次報告書での使用を目的に作成されている気候変動予測シミュレーションの出力データである。CMIP5 データは、第 4 次評価報告書をまとめる際に生じた科学的な争点に迫り、さらなる気候変動の把握、予測精度の向上のために、第 4 次評価報告書に用いられた CMIP3 データに対して、高精細化、大規模化がなされている。実験について厳格に規定されているばかりでなく、出力されるデータ自体についてもデータフォーマットや名前空間が詳細に規定されている[4][5][6]。特に Data Reference Syntax (DRS)によって規定されている属性がデータファイルおよび複数のデータファイルで構成されるデータセットに与えられ、データファイル・データセットはこの属性値によって同定される。データセットは、Activity、

Product、Institute、Model、Experiment、Frequency、Modeling realm、MIP table、Ensemble member の属性値により同定され、ID が付与される。さらに公開されたデータセットに関しては、訂正されたデータセットなどを区別するために Version number が付けられる。データセットは変数毎に独立したデータファイルによって構成され、データファイルはデータセットの DRS 属性値に Variable name 属性が付与されて同定されるが、さらに、データセット内の 1 つの変数データが時間的に複数のファイルに分割されている場合には、Temporal Subsets 属性が付与される。これらの属性は必ずしも独立しているわけではない。各実験について、提供すべき変数、期間などが異なり、要求度が定められている。データ提供機関は、要求度、使用するモデルの特性、計算能力などに従って、必ずしも全ての変数を提供する必要はない。このため、提供されている変数は、Model、Experiment、Frequency などによって大きく異なる。

**表 1　CMIP5 データの属性**

| 属性 | 説明 |
|---|---|
| Activity | データの収集アクティビティ名。CMIP5 で規定されるデータは属性値として`CMIP5'のみであるが、CMIP5 では規定されていない関連実験データが別のアクティビティとして公開されている。 |
| Product | プロダクト名。要求度に応じて`output1'、`output2'等となる。 |
| Institute | データに対する責任機関名。 |
| Model | 使用したモデル名。モデルのバージョンナンバーを含む。 |
| Experiment | 実験名。過去の再現実験や定められたシナリオによる長期予測などを示す。 |
| Frequency | データの出力間隔。`yr'（年データ）、`mon'（月データ）、`day'（日データ）等。 |
| Modeling realm | 出力変数の上位レベル名。`atmos'（大気）、`ocean'（海洋）等。 |
| Variable name | 物理量を表す変数名。 |
| MIP table | 実験グループ、出力間隔に対して要求される変数、要求度を記載したテーブル名。 |
| Ensemble member | アンサンブルメンバー。同一実験における初期値の違いやモデルのパラメータの違いにより付与される数値の組み合わせで表現される。 |
| Version number | バージョンナンバー。訂正されたデータなどを区別するために付与される。 |

## 3. Earth System Grid （ESG）

### 3.1. アーキテクチャ

CMIP5 データは Earth System Grid (ESG)と呼ばれる分散システムにより公開される[7][8]。ESG は、米国エネルギー省の研究費により、ローレンス・リバモア国立研究所（Lawrence Livermore National Laboratory）、アメリカ大気研究センター（National Center for Atmospheric Research）などにより構成される ESG-CET（Earth System Grid Center for Enabling Technologies）により開発がなされている。MyProxy、THREDDS Data Server[9]など多くの部分でコンポーネントとして既存技術を用いている。

ESG はローカルにデータをアーカイブし、それらのデータに対する基本的機能を提供する ESG node と、複数の ESG node と協調し分散する資源へのアクセスを提供する ESG gateway により構成される。図 1 は ESG におけるサービスのアーキテクチャであり、3 つの階層により構成される。Tier 1 は、ESG gateway 群と ESG node 群によって構成されるフェデレーションにより提供されるサービスであり、登録されたユーザに対して全システムへのアクセスを可能とするシングルサインオン機能、アクセスしている ESG フェデレーションのポータルサイトによらないデータ検索機能を提供する。Tier 2 は、ESG gateway における、データリクエストブローカーとしてのサービスである。インタフェース機能を含めた、データ、メタデータの検索、ブラウジング機能の提供を行う。Tier 3 は、実際にデータをアーカイブする ESG node による、データ、メタデータへのアクセス機能である。データ公開時の ESG gateway へのメタデータの発行（publish）、ESG gateway を通じたデータ要求の処理を行う。

ESG gateway は現在、ローレンス・リバモア国立研究所、アメリカ大気研究センター、ドイツ気象センター（DKRZ）など 7 か所で、ESG node は 15 か所で運用

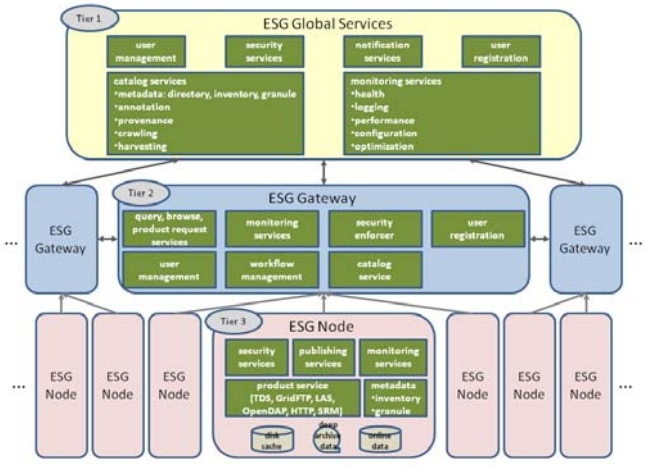


**図 1　ESG のアーキテクチャ（[8]を一部修正）**

されている。

### 3.2. データの公開

データの公開（publish）処理は ESG node から起動され、以下の手順で行われる。

1. データファイルを ESG node 上に格納。
2. publish 用証明書の取得。
3. データファイルのスキャン。メタデータの ESG node 内データベースへの格納。
4. THREDDS カタログの作成。
5. ウェブサービスによる ESG gateway への接続。
6. ESG gateway による THREDDS カタログの取得。メタデータの ESG gateway 内データベースへの格納と ESG gateway 間での共有。

### 3.3. データの取得

ユーザは ESG gateway のポータルサイトから検索を行い、データファイルを取得する。データの検索は ESG gateway にて行われ、データファイルの取得は、当該データファイルが存在する ESG node からダウンロードされる。ESG node におけるデータ提供手法としては、THREDDS Data Server、GridFTP などが選択可能であるが、多くの ESG node では THREDDS Data Server のみで提供している。ESG のポータルサイトでは、データ検索の結果として、ブラウザを通じた個々のファイルへのアクセスのためのリンクの表示のほか、検索されたデータセット内の複数のファイルを一括してダウンロード可能とするスクリプトを提供している。

### 3.4. アクセスログ

THREDDS Data Server のログは、トランザクションの開始時、および終了時にそれぞれ記録される。開始時には、タイムスタンプ、リモートホスト IP アドレス、リクエスト ID、リクエスト内容等が記録され、終了時には、タイムスタンプ、ステータスコード、クライアントへの送信サイズ等が記録される。これらのログは、エラー情報などとともに同一のファイルに記述される。従って、1 つのトランザクションの開始時のログと終了時のログの間に他のログ情報が記録されることとなるが、リクエスト ID を元にそれらの対応をとることが可能である。また、開始時のログのみが記録されていることより、そのトランザクションが継続中であることが分かる。さらに、ESG の管理データベースにも、リクエスト情報が記録される。しかしながら、ログに記録される項目と管理データベースに記録される項目は異なり、一方のみに記録される項目もある。例えば、ユーザ ID は、ESG の管理データベースのみに記録される。このため、本アクセスログ可視化システムでは、ログの記録と ESG の管理データベースの記録を、時刻、リモートホスト IP アドレス、ターゲットファイル名によってマッチングをとり、本システム用のデータベースを作成している。

## 4. データ統合・解析サーバによるデータの公開

我々が運用を行っているデータ統合・解析システム（DIAS）では、ESG node として日本で作成された CMIP5 データの公開を行っている。データの作成は 21 世紀気候変動予測革新プログラムに基づき、海洋研究開発機構、東京大学、気象庁気象研究所などが行い、作成されたデータがデータ統合・解析システムにアーカイブされ、THREDDS Data Server を通じて公開される。日本からは、2 種類の Institute 属性、7 種類の Model 属性の CMIP5 データが公開され、この他に TAMIP、LUCID という関連プロジェクトのデータも公開されている。2011 年 9 月より公開を開始し、2012 年 2 月現在、ファイル数約 670000、データセット数約 8500、容量 350TB 以上のデータが公開されている。ESG フェデレーションにより公開されている全 CMIP5 データは、レプリカを含めて約 900TB であり、データ統合・解析システムは、公開しているデータ量が最大の ESG node となっている。現在もデータの準備が整い次第、逐次データの追加・更新を行っており、今後、データ統合・公開サーバより公開される CMIP5 データは 900TB 程度となる予定である。

## 5. アクセスログ可視化システム

### 5.1. アクセスログの三次元可視化

データ統合・解析システムにはこれまで、正常に終了した CMIP5 データファイルに対するアクセスが 60

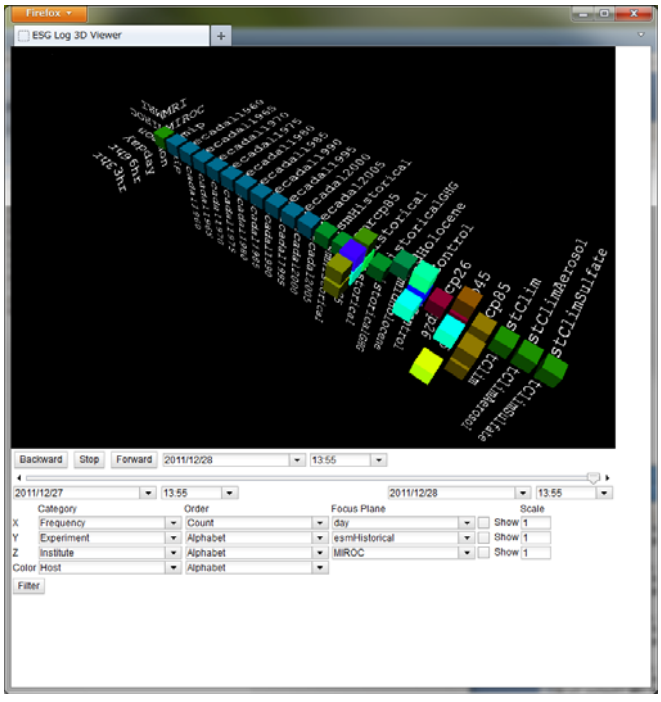


図 2　アクセスログ可視化システム画面

万件以上あり、また、公開している CMIP5 データは多種多様である。これら CMIP5 データファイルへのアクセス分布を俯瞰的にみることができるシステムを開発した。

本システムは WWW を用いて実装した。サーバ側ではログ情報をデータベース化し、クライアントからのリクエストに応じて、クライアント側で設定された可視化対象期間内のアクセス情報を返送する。クライアントは JavaScript、Dojo Toolkit、WebGL を用いて記述しており、多くのウェブブラウザで特別なプラグインなしで利用できる。

図 2 はアクセスログ可視化システムの画面である。X 軸、Y 軸、Z 軸に CMIP5 データファイルの属性、リモートホスト、アクセス時刻をそれぞれ設定し、設定された可視化対象期間内におけるアクセスを、ぞれぞれの軸に従ってキューブを配置して表示する。各軸について、属性、リモートホストを設定した場合には、その並び順にアルファベット順、アクセス数順、転送データ量順を選択することが可能である。キューブの色についても、各軸と同じように設定可能であり、設定された属性、順序に従い色がつけられる。また、時刻に関しては、対象時刻が設定されており、その時刻にファイルがアクセスされている場合には、キューブを高輝度で表示する。対象時刻は直接入力、カレンダーからの選択の他に、スライドバーで変更することも可能である。また、各軸の任意の位置に軸に垂直な面を表示することも可能であり、視認性を高めている。

### 5.2. フィルタリング機能

図 3 に示すように、各属性、リモートホストの一覧から選択することで、可視化の対象とするアクセスを選択・除外することが可能である。データ統合・解析システムから公開している CMIP5 データは、変数が 300 種類以上にもなり、これまで 600 以上のユーザ、1000 以上のホストからリクエストを受けている。対象とするアクセスをフィルタリングすることにより、CMIP5 データへのアクセスを適切に把握することが可能となる。

### 5.3. トレース機能

対象時刻を順方向、逆方向に変化させ、アクセスの変化の様子をアニメーション表示することが可能なトレース機能を有する。図 4 は、トレース機能実行時のアクセスログ表示システムの画面である。X 軸に時刻、Y 軸にリモートホスト、Z 軸に変数を設定している。キューブの色はリモートホストにより設定されている。また、対象時刻を示す X 軸に垂直な面を表示している。トレース機能を動作させることにより、対象時刻の変化に従って、その時刻にアクセスが行われていることを示す高輝度のキューブが移動し、アクセス対象の移動を見て取ることができる。

## 6. おわりに

本稿では、気候変動予測データの公開サーバにおけるアクセスログを可視化するシステムについて述べた。気候変動予測データについて説明をするとともに、そのデータの公開システムについて述べた。さらに、我々が公開しているデータに対するアクセスログを三次元

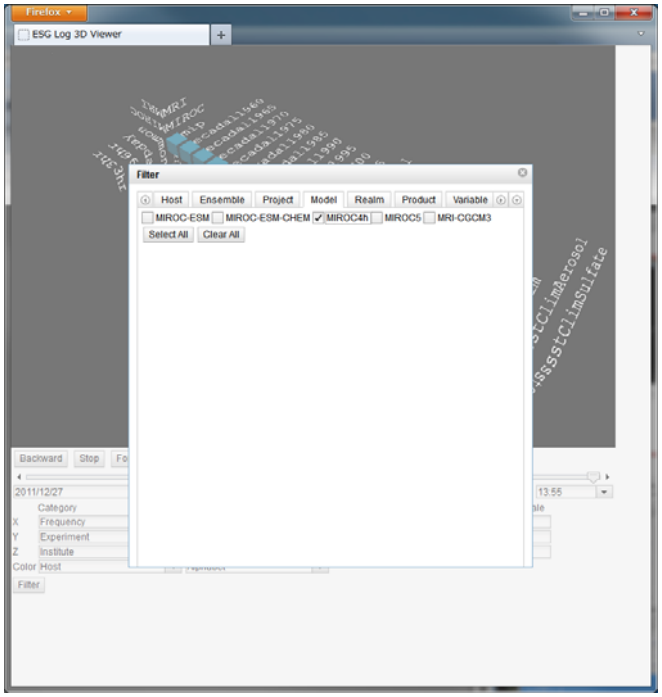


図 3　フィルタリング機能

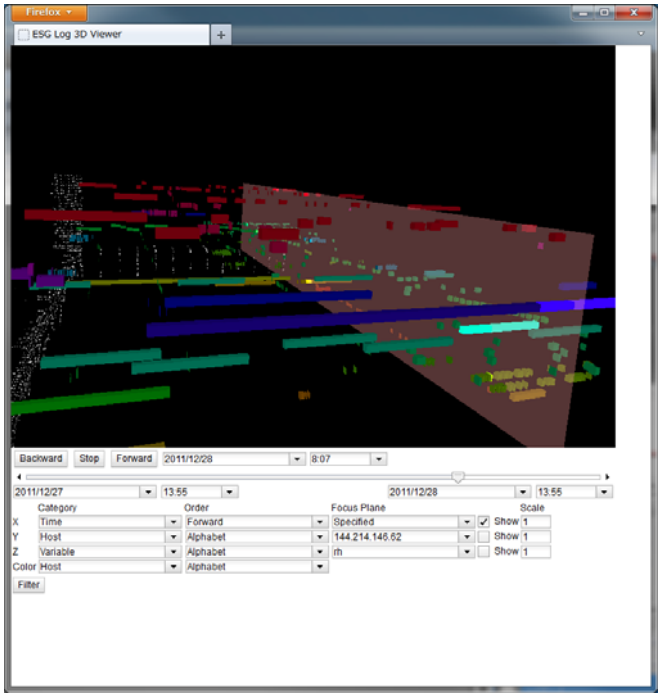


図 4　トレース機能実行画面

可視化するシステムについて、その機能を述べた。本システムを利用することにより、多数の変数を横断的にダウンロードする様子や、少数の変数を多数のストリームによりダウンロードする様子などを視覚的に見て取ることが可能となり、ダウンロード形態が多様であることが明らかとなった。

本システムは、アクセスパターンを俯瞰的に把握することを目的として実装を行ったが、今後、より詳細にアクセス状況をみることが可能なよう、ドリルダウン機能等を実装していく予定である。

## 参 考 文 献


[1] “気候変動 2007:統合報告書 政策決定者向け要約”, http://www.env.go.jp/earth/ipcc/4th/syr_spm.pdf

[2] “CMIP5 – Coupled Model Intercomparison Project Phase 5 – Overview”, http://cmip-pcmdi.llnl.gov/cmip5/

[3] “About the Earth System Grid”, http://www.earthsystemgrid.org/about/overview.htm

[4] K. E. Taylor, R. J. Stouffer and G. A. Meehl, “A Summary of the CMIP5 Experiment Design”, http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf

[5] K. E. Taylor and C. Doutriaux, “CMIP5 Model Output Requirements: File Contents and Format, Data Structure and Metadata”, http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_output_metadata_requirements.pdf

[6] K. E. Taylor, V. Balaji, S. Hankin, M. Juckes, B. Lowrence and S. Pascoe, “CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies”, http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

[7] D. N. Williams et al., “The Earth System Grid: Enabling Access to Multimodel Climate Simulation Data”, in the Bulletin of the American Meteorological Society, February, 2009.

[8] D. N. Williams et al., “Data management and analysis for the Earth System Grid”, in the Journal of Physics: Conference Series, SciDAC’08 conference proceedings, vol. 125

[9] “THREDDS (Thematic Realtime Environmental Distributed Data Services)”, http://www.unidata.ucar.edu/projects/THREDDS/