Progress in Earth and Planetary Science

METHODOLOGY

Open Access

Development of a system for efficient content-based retrieval to analyze large volumes of climate data

Yujin Nakagawa[1*], Yosuke Onoue[2], Shitnaro Kawahara[1], Fumiaki Araki[1], Koji Koyamada[3], Daisuke Matsuoka[1], Yoichi Ishikawa[1], Mikiko Fujita[4], Shiori Sugimoto[4], Yasuko Okada[4], Sho Kawazoe[5], Shingo Watanabe[4], Masayoshi Ishii[6], Ryo Mizuta[6], Akihiko Murata[6] and Hiroaki Kawase[6]

Abstract

Analyses of large ensemble data on future climate are significantly useful for the probabilistic future projection of climate change in various interdisciplinary fields. However, the data volume of the Database for Policy Decision making for Future climate change or d4PDF, which is a mega-ensemble dataset, exceeds ~ 3 PB, which is too large to download to local computers. To allow users for retrieve and downloading necessary data, we developed a user-friendly system called "System for Efficient content-based retrieval to Analyze Large volume climate data" (SEAL) under the Social Implementation Program on Climate Change Adaptation Technology (SI-CAT). Conventional web-based retrieval systems allow retrievals using metadata associated with a data file itself. In contrast, SEAL allows the users to retrieve the necessary data by using metadata associated with contents, such as physical values, of a data file. We confirmed that SEAL can reduce data sizes and total time required for obtaining necessary data to less than 0.5% and 1%, respectively, compared to conventional web-based retrieval systems.

Keywords: Climate data, Relational database, Web application

Introduction

In the field of climate research, improvements in computer performances have led to significant growths in the volumes of large ensemble simulation data. For instance, the volumes are estimated to exceed ~ 3 PB in the case of the database for Policy Decision making for Future climate change (d4PDF; Mizuta et al. 2017), which is produced by the Program for Risk Information on Climate Change. Systematic analyses of such large ensemble simulation data are relatively useful for the projection of probabilistic effects of climate change for extreme weather events. However, such systematic analyses generally require large data storage as well as high-performance computers and are thus becoming increasingly complicated for individual researchers to work with.

The Social Implementation Program on Climate Change Adaptation Technology (SI-CAT) is a national-level Japanese project, which is intended to construct adaptation measures and technologies for near-future climate changes. To ensure the security of residents and protect their property from threats of near-future climate changes, SI-CAT establishes cooperative relationships among researchers of earth science, social science, and humanities, as well as office staff of local governments. In addition, the SI-CAT project is intended to help local governments by promoting developments in adaptation plans and by assisting companies to create new businesses based on climate change adaptation needs.

As part of the d4PDF dataset, SI-CAT has produced ensemble simulation data of near-future climate, where the global average temperature increases by 2°C after the industrial revolution (Fujita et al. 2019) as a part of the d4PDF dataset. The data produced by SI-CAT have been released on the Data Integration and Analysis System

*Correspondence: nakagawa.yujin@jamstec.go.jp
[1]Research Institute for Value-Added-Information Generation, Japan Agency for Marine-Earth Science and Technology, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan
Full list of author information is available at the end of the article

Program (DIAS; http://www.diasjp.net). In addition, SI-CAT produces statistical downscaling data with 1 km grid spacing.

The data volume of the d4PDF is estimated to be a few petabytes, which is significantly larger than the volumes of past datasets. In DIAS, if analysis servers are available in the same local network as a data server storing the d4PDF, users can analyze the d4PDF on the analysis servers without downloading the data. However, currently, such analysis servers are not available for the d4PDF, and therefore, users are required to download the d4PDF to their local computers. In this case, the large data volume of the d4PDF would cause the following concerns: a lack of disk space for users who want to download the d4PDF, a long period required for downloading from the data server in DIAS to the users' computers, and a high load on the data server. Considering that the data volume is extremely large (a few petabytes) to download to local computers, a user-friendly system is required for retrieving and downloading data in a manner that satisfies user requests.

Conventional web-based retrieval systems for climate simulations, shown in Table 1, are typically used for retrievals and/or visualizations in the field of climate research. All the systems are designed to retrieve data by using metadata associated with the data files themselves rather than using the physical values stored in data files. If the data volumes to be downloaded to the local computers of users were reasonable, these conventional web-based retrieval systems would be quite useful. In the present study, we developed the "System for Efficient content-based retrieval to Analyze Large volume climate data" (SEAL) under SI-CAT to provide users services to find necessary data. SEAL was developed by combining conventional technologies in the field of information science. SEAL allows users to find data files according to the metadata associated with the contents, such as physical values, of the corresponding data files. SEAL reduces the data volumes of the files that the users downloaded by users from the data server to their local computers.

Table 1 Examples of conventional web-based retrieval systems for climate simulations

System name	Features
Data integration and analysis system program[a)]	Retrieval
Regional and global climate[b)]	Retrieval, visualization
Downscaled CMIP3 and CMIP5 climate and hydrology projections[c)]	Retrieval
National climate change viewer[d)]	Visualization
Regional climate change viewer[e)]	Visualization
Multivariate adaptive constructed analogs datasets[f)]	Retrieval
North American regional climate change assessment program[g)]	Retrieval
World data center for climate[h)]	Retrieval

[a)] http://www.diasjp.net
[b)] http://regclim.coas.oregonstate.edu
[c)] https://gdo-dcp.ucllnl.org/downscaled_cmip_projections
[d)] https://www2.usgs.gov/landresources/lcs/nccv/viewer.asp
[e)] http://regclim.coas.oregonstate.edu/visualization/rccv/index.html
[f)] https://climate.northwestknowledge.net/MACA
[g)] https://www.narccap.ucar.edu
[h)] https://cera-www.dkrz.de/WDCC/ui/cerasearch

Methods/Experimental

Data

To design SEAL, we used the d4PDF comprising global and regional simulation data. The global simulation data were produced by a global atmospheric model with a horizontal grid spacing of 60 km developed by the Meteorological Research Institute (MRI) (hereafter, MRI-AGCM; Mizuta et al. 2012. The regional simulation data cover all of Japan and were produced by a non-hydrostatic regional climate model with a 20 km grid spacing, developed by MRI (hereafter, MRI-NHRCM; Sasaki et al. 2011; Murata et al. 2013). Among them, we used the surface atmospheric data stored in the regional simulation data, which consists of three datasets. The first dataset comprises data on historical climate simulations from September 1950 to August 2011 (Mizuta et al. 2017). The other datasets include data on future climate simulations, where global average surface air temperature increased by 2°C or 4°C after the industrial revolution (hereafter +2K near-future climate simulations or +4K future climate simulations, respectively) (Mizuta et al. 2017; Fujita et al. 2019). In all the datasets, the horizontal grid sizes in the x and y directions are 191 and 155, respectively. Thus, the physical values for 29,605 grid points are stored in the datasets. The geographical longitude and latitude for each grid point are defined as $\lambda(x, y)$ and $\phi(x, y)$ ($1 \leq x \leq 191$ and $1 \leq y \leq 155$), respectively. These surface simulation data are available in the GRIdded Binary (GRIB) format.

Basic design

The basic design of SEAL is based on three important concepts. The first concept is practical utility to satisfy user needs. These users mainly comprise researchers who estimate climate change impacts on nature and agriculture, as well as office staff of local governments, who need to make decisions on adaptation measures for climate change. Next, we explored the needs of users associated with SI-CAT. First, we explored physical variables used as retrieval criteria and found that precipitation and temperature are typically used for research on meteorology and climatology as well as impact estimations. Next, we

explored the retrieval criteria and processed index values, using physical variables such as daily precipitation. Table 2 summarizes the index values. Among the requests of the users with regard to index values, we excluded index values for the wet bulb globe temperature, which is not stored in the d4PDF.

The second concept relates to preservation of physical values stored in raw data. We did not apply any user-dependent processing, such as bias corrections, to the raw data. The third concept involves the general application of technologies developed for SEAL to various datasets. Data volumes of climate simulations will grow with improvements in computer performances in the near future. Such large-volume climate simulation data will encounter the same situations as those experienced with the erstwhile data in the d4PDF.

Table 2 Summary of the index values for each physical variable

Name of physical variable	Name of index value
Precipitation	01. Cases where hourly precipitation exceeds a given criterion.
	02. Number of occurrences where hourly precipitation exceeds a given criterion.
	03. Cases where accumulated precipitation in each given hour exceeds a given criterion.
	04. Cases where daily precipitation exceeds a given criterion.
	05. Number of occurrences where daily precipitation exceeds a given criterion.
	06. Annual maximum daily precipitation.
	07. Three days accumulated precipitation.
	08. One-month precipitation.
	09. Accumulated and averaged precipitation in each given month.
	10. Number of dry days.
	11. Number of continuous dry days.
	12. Number of days of continuous precipitation.
Temperature	13. Cases where daily maximum temperatures exceed a given criterion.
	14. Cases where daily maximum temperatures fall below a given criterion.
	15. Daily averaged temperatures.
	16. Monthly averaged temperatures.
	17. Averaged temperatures in each given month.
	18. Daily minimum and maximum temperatures.
	19. Number of days classified as tropical day, extremely hot day, extremely tropical night, frost day, and ice day.
	20. Number of days classified as tropical day, extremely hot day, extremely tropical night, frost day, and ice day in each given month.
	21. Cases where daily maximum temperatures exceed a given criterion and daily minimum temperatures fall below a given criterion.
Precipitation and temperature	22. Cases where daily maximum temperatures fall below a given criterion and daily precipitation exceeds a given criterion.

Figure 1 shows a conceptual diagram of SEAL, which comprises a relational database, data providing function, and user interface. Among the three main features, the relational database using PostgreSQL plays a key role and is designed to register temporally and spatially compressed data. The relational database is a collection of data items with pre-defined mutual relationships. The data items are defined as sets of tables with columns and rows. The relational database allows us to treat relationships of complex data by preliminarily defining relationships among the tables. In general, Structured Query Language (SQL) is used as an interface for communications with the relational database. The relational database is managed by relational database management systems such as PostgreSQL, MySQL, MariaDB, Oracle Database, and Microsoft SQL Server. To achieve a semi-permanent operation of SEAL after the end of the SI-CAT project, a relational database management system used for SEAL should be distributed without charge and should have proven performance of a stable operation performance. In addition, a wealth of technical information about the relational database management system should be provided. As a result, we narrowed the relational database management systems down to PostgreSQL and MySQL. Among them, we decided to use PostgreSQL because it allows us to easily install PostGIS (e.g., Marquez 2015), which supports geographic objects allowing location queries in SQL statements. We designed the relational database for precipitation and temperature according to the needs of the SI-CAT's members. In addition, we designed the relational database such that various other physical variables could be applied. The data-providing function allows the users to download temporally and spatially extracted data based on retrieval results obtained through the relational database. In addition, the web-based user interface allows the users to easily use the relational database without knowledge about PostgreSQL.

Detailed design

Relational database

The users require physical values for regions such as administrative districts or basins rather than grid cells because they are mostly interested in certain regions. Using the geographical longitudes and latitudes for the grid points, the geographical longitudes

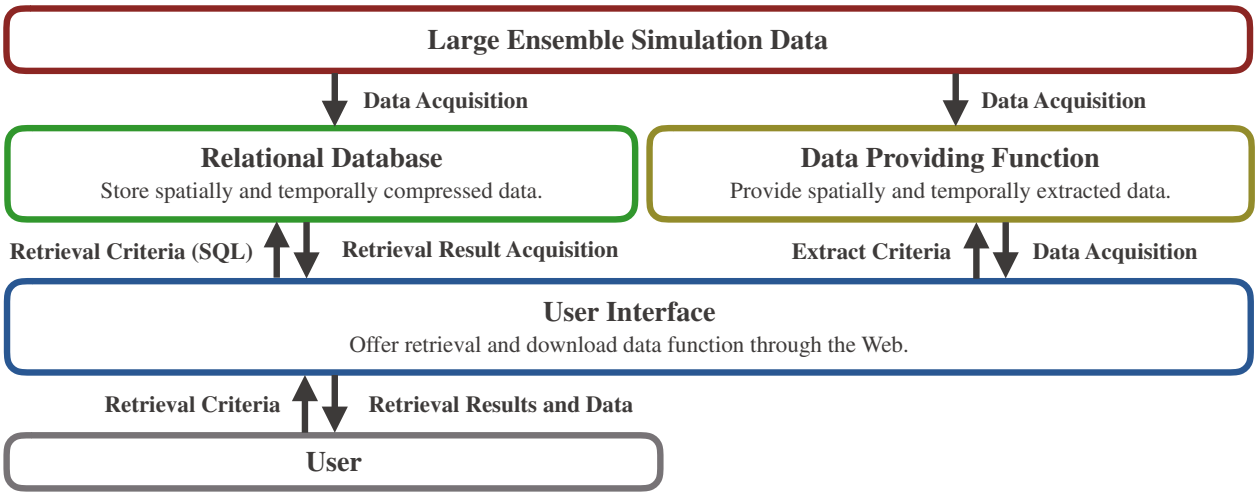

Fig. 1 Conceptual diagram of SEAL

and latitudes for four corners of a certain grid cell are defined as

$$C_{\mathrm{nw}}(x,y) = \left\{ \frac{\lambda(x,y)+\lambda(x-1,y)}{2}, \frac{\phi(x,y)+\phi(x,y+1)}{2} \right\} \text{(northwest)}, \tag{1}$$

$$C_{\mathrm{ne}}(x,y) = \left\{ \frac{\lambda(x,y)+\lambda(x+1,y)}{2}, \frac{\phi(x,y)+\phi(x,y+1)}{2} \right\} \text{(northeast)}, \tag{2}$$

$$C_{\mathrm{se}}(x,y) = \left\{ \frac{\lambda(x,y)+\lambda(x+1,y)}{2}, \frac{\phi(x,y)+\phi(x,y-1)}{2} \right\} \text{(southeast)} \tag{3}$$

and

$$C_{\mathrm{sw}}(x,y) = \left\{ \frac{\lambda(x,y)+\lambda(x-1,y)}{2}, \frac{\phi(x,y)+\phi(x,y-1)}{2} \right\} \text{(southwest)}, \tag{4}$$

where $2 \leq x \leq 190$ and $2 \leq y \leq 154$. The regions are explained by combinations of grid cells. In addition, for some variables, such as temperature, users require physical values on a daily, monthly, and yearly basis rather than on an hourly basis.

Spatial and temporal compressions were applied to physical values of the SI-CAT data. Then, spatially and temporally compressed physical values were registered on the relational database. These compressions reduce the disk size required for the relational database. The compressions also reduce the retrieval time because they decrease the number of records stored in the relational database. In principle, the number of grid cells for each region does not affect retrieval times because records for a certain region requested by the users are uniquely identified using an index of the relational database. On the other hand, retrieval times are different for records on an hourly, a daily, a monthly, or a yearly basis as a smaller temporal resolution increases the number of targeted records. In fact, the retrieval times for records on the smallest hourly basis are much shorter than corresponding working times using the conventional web-based retrieval systems. Therefore, an increase in the retrieval time due to smaller temporal resolutions is not viewed by most users as an issue. As a result, SEAL works well even for smaller spatial and temporal resolutions.

The physical value of grid-cell number g at time t for region number k is defined as $p_k(g,t)$. The spatially compressed physical values are defined as the sum of physical values of each grid cell at time $t = t'$ for region number k such that

$$U_k(t') = \{p_k(g_0,t'), p_k(g_1,t'), \ldots, p_k(g_{n_k},t')\}, \tag{5}$$

where n_k is the number of grid cells in region k. Then, the sum of the physical values is defined as

$$u_k(t') = \sum_{i=0}^{n} p_k(g_i,t')(i = 0,1,\ldots,n_k). \tag{6}$$

The temporally compressed physical values are defined as the sum of the physical values of m continuous time bins at grid cell $g = g'$ for region number k, where m is the number of time bins.

$$V_k(g') = \{p_k(g',t_0), p_k(g',t_1), \ldots, p_k(g',t_m)\}. \tag{7}$$

Then, the sum of the aforementioned physical values is defined as

$$v_k(g') = \sum_{j=0}^{m} p_k(g', t_j)(j = 0, 1, \ldots, m). \quad (8)$$

As a result, the sum of the physical values in continuous m time bins for region number k is defined as

$$q_k = \sum_{i=0}^{n_k} \sum_{j=0}^{m} p_k(g_i, t_j), \quad (9)$$

which is registered on the relational database. The values in Eq. (9) are not normalized by factors such as number of grid cells n_k and number of time bins m. The values in Eq. (9) are converted to normalized physical values in SQL scripts at the time of retrieval by the users. Normalization factor n_k is always applied to the retrieval results. In addition, applications of factor m are determined depending on the physical values. Only normalization factor n_k is applied to Eq. (9) for determining the physical values that emphasize a temporal summation (such as precipitation). Then, Eq. (9) is rewritten as follows:

$$\overline{q}_k = \frac{1}{n_k} \sum_{i=0}^{n_k} \sum_{j=0}^{m} p_k(g_i, t_j). \quad (10)$$

For determining the physical values that emphasize a temporal average (such as temperature), both normalization factors n_k and m are applied to Eq. (9), which is rewritten as follows:

$$\overline{q}_k = \frac{1}{n_k} \frac{1}{m} \sum_{i=0}^{n_k} \sum_{j=0}^{m} p_k(g_i, t_j). \quad (11)$$

Data-providing function

The data-providing function allows the users to download temporally and spatially extracted raw data based on the results of data retrieval obtained from the relational database. DIAS provides a basic function, using which the users can download binary data (big-endian and 4-byte floating-point without headers) appropriate for the Grid Analysis and Display System (GrADS; e.g., Doty et al. 1995). Most researchers attempting estimation of climate impact require raw data with human-readable formats such as the text or csv formats. To increase the conveniences of such users, we developed a function that converts the GrADS binary format into the text or csv formats.

User interface

The SI-CAT members prefer to proceed with data retrieval without knowledge about the GRIB format and PostgreSQL. Such users also prefer to convert raw data to the text or csv formats without knowledge of command line interfaces. Then, we developed a web-based user interface for using the relational database and the data-download function without users requiring knowledge about PostgreSQL and command line interfaces.

Implementation

In this study, we used surface data of MRI-NHRCM in the d4PDF, where the temporal resolution is 1 h. As shown in Table 2, temporal resolutions for precipitation have various requirements. The precipitation is stored as two time resolutions in the hourly scale (i.e., $m = 1$ in Eq. 10) and daily scale (i.e., $m = 24$ in Eq. 10) to satisfy the requirements of low retrieval time and high practical utility. Using the daily values, the retrieval time of accumulated precipitation with time scales of more than 1 day was reduced to 24 h compared with the hour-based retrieval time. The daily mean temperature was stored (i.e., $m = 24$ in Eq. 11) as the daily maximum and minimum temperatures. In most cases, the users require the physical values for their administrative district. Therefore, we decided to calculate the physical values for 47 prefectures, considering the 20 km grid spacing. We would like to emphasize that using the shapefile developed by Environmental Systems Research Institute, Inc., we can calculate the physical values for any combination of grid cells according to user requests. The shapefile represents geospatial vector data. For example, in the future, we shall calculate the physical values for basins in Japan at the request of users interested in river engineering. The physical values for the basins will be used for a relevant web interface with SEAL. Below, we present the calculations of the physical values for 47 prefectures. We summed the physical values of the grid cells that overlap a region of each prefecture. Among the 47 prefectures, Tokyo metropolis and Okinawa prefecture have isolated islands (i.e., these islands are distant from their respective main regions). Thus, the physical values for Tokyo metropolis were summed over the grid cells of the main island. In addition, the physical values for the Okinawa prefecture were summed over the grid cells of the Okinawa main island. The number of grid cells differs depending on the prefectures.

The basic function in DIAS provides the option of printing raw data in the binary format to a standard output. Thus, a Python script was developed to receive raw data in the binary format as the standard input and print them in the text or csv format as the standard output.

Figure 2 shows a screenshot of the web-based user interface, which is currently available only in Japanese. The web-based user interface consists of five parts. The first part comprises selection fields for common conditions of retrieval conditions (names of datasets, experiment types, physical variables, and retrieval types), as shown in Fig. 2a. The second part shows a selection and input fields for unique conditions of retrieval associated with the retrieval types, as shown in Fig. 2b. The third part comprises

Fig. 2 Screenshot of the web-based user interface of SEAL. **a** Selection fields for common conditions of retrieval (names of datasets, experiment types, physical variables, and retrieval types). **b** Selection and input fields for unique conditions of retrieval associated with the retrieval types. **c** Information fields for explanation of the retrieval types. **d** Information fields for supplements of the input fields, acknowledgements, and contact details. **e** Result fields for retrieval

information fields for explanations of the retrieval types, as shown in Fig. 2c. The fourth part comprises information fields for supplements of the input fields, acknowledgements, and a contact, as shown in Fig. 2d. The fifth part shows the result field of the retrieval, as shown in Fig. 2e. If the users press the download buttons placed in the retrieval results, the web-based interface calls the data-providing function and delivers the raw data in the binary, text, or csv formats through a Multipurpose Internet Mail Extension type called "application/octet-stream."

Case studies

Spatial and temporal compression ratios

The spatial and temporal compressions contribute to size reductions of data analyzed by the users to find the necessary data. To clarify the reductions quantitatively, we calculated the spatial and temporal compression ratios. The spatial compression ratios change depending on the area of each prefecture because the numbers of corresponding grid cells differ for each prefecture in Japan. The compression ratios are defined as the reciprocals of the numbers of the corresponding grid cells. Hokkaido prefecture has a maximum of 302 grid cells, while both Tokyo metropolis and Osaka prefecture have a minimum of 13. Thus, the spatial compression ratios are ∼0.3% for Hokkaido prefecture and ∼8% for both Tokyo metropolis and Osaka prefecture. In addition, the temporal compression ratio is ∼4% for the daily data. As a result, the data sizes of the daily data are reduced by ∼0.01% at a maximum and ∼0.3% at a minimum after applying both the spatial and temporal compressions. Such data size reductions may help reduce the amount of time required for exploring necessary data compared with methods using the conventional web-based retrieval systems.

Time for exploring and retrieving necessary data

To clarify the advantages of SEAL, the time required for retrieving necessary data was quantitatively examined using the conventional methods and SEAL on our local server. The local server was equipped with Intel Xeon E7-4820 (CPU: 40 cores) and 512 GB physical memory. One CPU core was used for all analyses, which are described as follows. Figure 3 shows a situation in which a user requires to download hourly data of precipitation and temperature stored in the + 4K future climate simulations for targeted days, on which the daily precipitation exceeds 100 mm in Hokkaido prefecture, Japan. The conventional methods require 3–90 h for downloading hourly data of precipitation around Hokkaido prefecture, ∼31 h for exploring the targeted days, and 0.02–0.5 h for downloading hourly temperature data for those targeted days. SEAL requires ∼0.002 h for finding the target days and 0.03–1 h for downloading the hourly data of precipitation and temperature for those target days. Data sizes of the hourly data are reduced to ∼0.5% compared with the original data. Hence, SEAL can reduce time required for retrieving necessary data to less than 1% of that required by the conventional methods.

Retrievals of heavy precipitation

As discussed earlier, SEAL contributes toward reducing the time required for exploring necessary data. This advantage may allow the users to find extreme events, such as heavy precipitation, quickly. To examine its practical utility, we performed retrieval using SEAL operating on the local server by assuming a situation in which users specializing in river engineering require data for their research. Using SEAL, we explored data for Tokyo metropolis, Japan, in the + 4K future climate simulations,

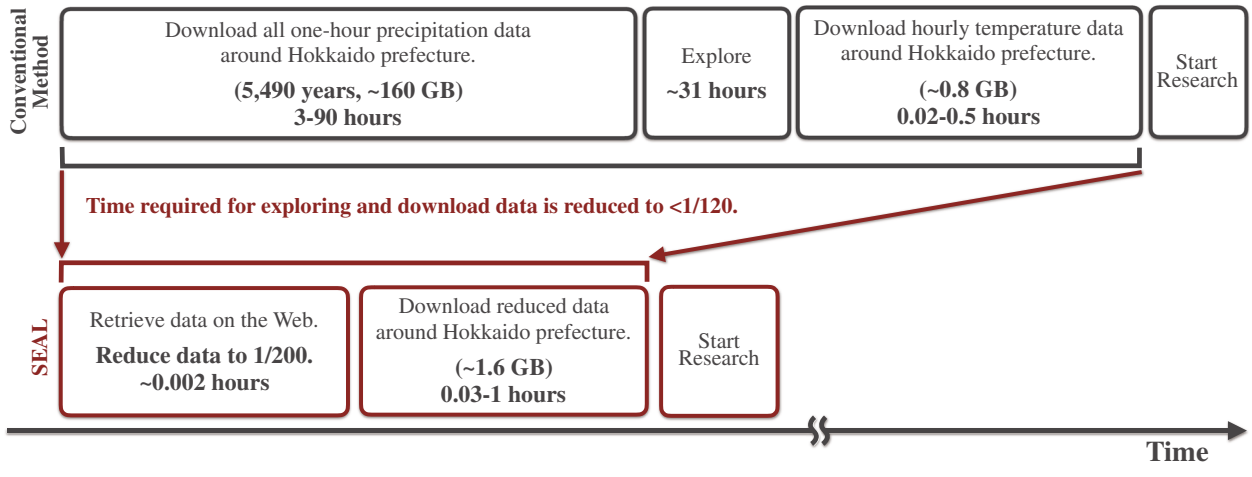

Fig. 3 Schematic of the expected advantages of using SEAL. The elapsed times shown in the diagram are actual measured values for the + 4K future climate simulations based on download speed from DIAS to JAMSTEC (4–120 Mbps)

where the number of days of continuous precipitation is greater than 20 days and accumulated precipitation is greater than 800 mm. Here, we define a precipitation day as a day for which the precipitation exceeds 0.1 mm. As a result, we found 6 events meeting this criterion, with a retrieval time of approximately ∼30 s. Among these events, the event with highest precipitation showed a value of ∼1109 mm over 24 days. Figure 4 shows the contour map of the accumulated precipitation for the event. The event is attributed to heavy precipitation centered around Shizuoka prefecture, Japan. When exploring the abovementioned events by using the conventional methods, 0.13–3.9 h are required for downloading the hourly data of precipitation for Tokyo metropolis, which corresponds to 13 grid cells. Additional time will be required for users to calculate the continuous precipitation. Consequently, SEAL is capable of reducing the time required for exploring the necessary data.

Results

Retrieving large volumes of data, such as in the case of the d4PDF, by using the conventional web-based retrieval systems entails disadvantages such as lack of user disk space, long download period, and high load on the data server. To provide services that can resolve such concerns, we developed SEAL, which allows users to efficiently and quickly explore necessary data under SI-CAT. Using SEAL, the users can find the necessary data without downloading them and/or requiring knowledge about the GRIB format and PostgreSQL because users can conduct all the tasks via the web-based user interface. Moreover, users can download the desired original data in the binary, text, or csv format based on the data retrieval results via the web-based user interface. The data sizes stored in the relational database of SEAL are reduced to ∼0.01% at a maximum and ∼0.3% at a minimum compared with the original data due to the use of the spatial and temporal compressions. In addition, SEAL can reduce the time required for retrieving necessary data to less than 1% of that in the case of conventional systems.

Discussion

Advantages of the relational database

Although the conventional framework, OPeNDAP (https://www.opendap.org), provides a function to retrieve values from a single time-series grid data (i.e.,

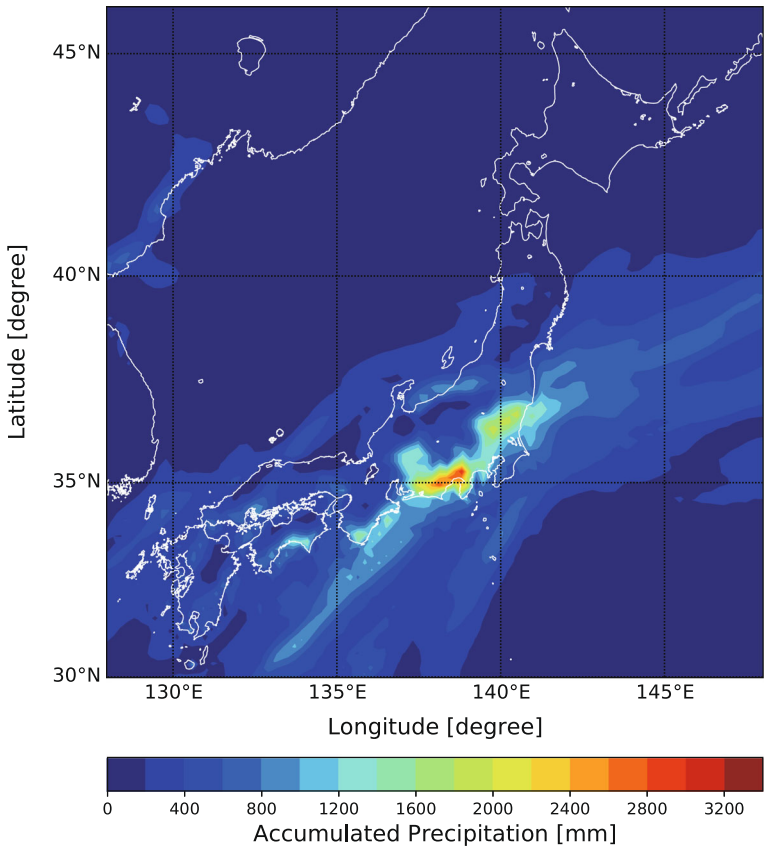

Fig. 4 Contour map illustrating heavy precipitation

a single data file), its retrieval speed is not fast because OPeNDAP scans all the values in a data file. In contrast, SEAL achieves high-speed retrieval of values by adopting the relational database, thus satisfying the criteria for multiple ensembles (i.e., all data files). This is because the relational database considerably improves the data retrieval speeds, which are highly affected by the database indices. Furthermore, the relational database has an advantage in that the multiple ensembles are scanned once.

Resolving limitations of conventional systems

As mentioned in the "Introduction" section, the conventional retrieval systems experience three disadvantages while retrieving necessary data from large data volumes. In the case described earlier, the size of the data users require to download to their local computers is reduced to ∼0.5% compared with that in the case of the conventional methods. Assume that a user needs to download hourly precipitation data around Hokkaido prefecture, which is stored in the + 4K future climate simulations of a regional climate model. The size of the necessary data is ∼0.8 GB which is much smaller than the data size of ∼160 GB using the conventional methods. Moreover, the time required to download the necessary data is 0.02–0.5 h, which is much lesser than that (3–90 h) required when using the conventional methods. Furthermore, the load on the data server is considerably lessened because the data sizes and required data retrieval time are considerably reduced, as mentioned earlier. Therefore, we conclude that all three issues related with retrieving large volumes of data are resolved by the proposed SEAL.

Conclusions

With increasing climate simulation data volumes, the conventional web-based data retrieval method suffers from three limitations while retrieving large data volumes, similar to those experienced when using the d4PDF. These include lack of user disk space, the long period required for data download, and high load on the data server. To resolve these concerns, we developed SEAL, which allows users to find data files by using metadata associated with the contents (such as physical values) of the data files under SI-CAT. SEAL allows the users to find the necessary data without downloading them, and they need not have knowledge about the GRIB format and/or PostgreSQL, because the users can proceed with all tasks via the web-based user interface. In addition, SEAL allows the users to download the desired original data in the binary, text, and csv formats based on the data retrieval results via the web-based user interface. The data sizes stored in SEAL's relational database are reduced to ∼0.01% at a maximum and ∼0.3% at a minimum of the original data due to the adoption of the spatial and temporal compressions. The relational database considerably improves the speed of retrieval, which is highly affected by the database indices, allowing for multiple ensembles to be scanned at once. In addition, SEAL can reduce data sizes and the total time required for retrieving necessary data to less than 0.5% and 1%, respectively. These reductions contribute to improvements in the load on the data server. As a result, SEAL works well as expected and provides solutions to all the concerns mentioned earlier. SEAL is currently being tested on a local server and will be released on DIAS during the Japanese 2019 fiscal year. The techniques developed for SEAL might be quite useful for simulation and observation of data when using grid spacing and/or time slicing in other research fields.

Abbreviations

d4PDF: A database for Policy Decision making for Future climate change; DIAS: Data Integration and Analysis System Program; GrADS: Grid Analysis and Display System; GRIB: GRIdded Binary; SEAL: System for Efficient content-based retrieval to Analyze Large volume climate data; SEAL-F: SEAL-Finder; SI-CAT: Social Implementation Program on Climate Change Adaptation Technology; SQL: Structured Query Language

Acknowledgements

This study was performed as part of the Social Implementation Program on Climate Change Adaptation Technology (SI-CAT). This study utilized the database for Policy Decision making for Future climate change (d4PDF), which was produced under the SOUSEI and SI-CAT programs. We would like to thank Editage (www.editage.com) for English language editing.

Authors' contributions

YN, SK, FA, and DM proposed the topic and conceived and designed the study. YO, KK, and YI contributed to improving the design of the relational database and the user interface. MF, SS, YO, SK, SW, MI, RM, AM, and HK helped produce the original data used in the study. All the authors read and approved the final manuscript.

Funding

This work was supported by the Social Implementation Program on Climate Change Adaptation Technology (SI-CAT).

Availability of data and materials

Please contact the author for data requests.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Research Institute for Value-Added-Information Generation, Japan Agency for Marine-Earth Science and Technology, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan. [2] College of Humanities and Sciences, Nihon University, 3-25-40 Sakurajosui, Setagaya-Ku, Tokyo 156-8550, Japan. [3] Academic Center for Computing and Media Studies, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan. [4] Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan. [5] Laboratory of Meteorology, Hokkaido University, N10W8, Kita-ku, Sapporo, Hokkaido 060-0810, Japan. [6] Meteorological Research Institute, Japan Meteorological Agency, 1-1 Nagamine, Tsukuba, Ibaraki 305-0052, Japan.

Received: 7 August 2019 Accepted: 12 December 2019
Published online: 26 February 2020

References

E. Doty B, Kinter J, James L (1995) Geophysical data analysis and visualization using the grid analysis and display system. In: Szuszczewicz EP, Bredekamp

JH (eds). Visualization techniques in space and atmospheric sciences. NASA SP-519, National Aeronautics and Space Administration, Washington, D.C. pp 209–217
Fujita M, Mizuta R, Ishii M, Endo H, Sato T, Okada Y, Kawazoe S, Sugimoto S, Ishihara K, Watanabe S (2019) Precipitation changes in a climate with 2-k surface warming from large ensemble simulations using 60-km global and 20-km regional atmospheric models. Geophys Res Lett 46(1):435–442
Marquez A (2015) PostGIS Essentials. Packt Publishing, Birmingham
Mizuta R, Yoshimura H, Murakami H, Matsueda M, Endo H, Ose T, Kamiguchi K, Hosaka M, Sugi M, Yukimoto S, Kusunoki S, Kitoh A (2012) Climate simulations using mri-agcm3.2 with 20-km grid. J Meteorol Soc Japan Ser II 90A:233–258
Mizuta R, Murata A, Ishii M, Shiogama H, Hibino K, Mori N, Arakawa O, Imada Y, Yoshida K, Aoyagi T, Kawase H, Mori M, Okada Y, Shimura T, Nagatomo T, Ikeda M, Endo H, Nosaka M, Arai M, Takahashi C, Tanaka K, Takemi T, Tachikawa Y, Temur K, Kamae Y, Watanabe M, Sasaki H, Kitoh A, Takayabu I, Nakakita E, Kimoto M (2017) Over 5000 years of ensemble future climate simulations by 60-km global and 20-km regional atmospheric models. Bull Am Meteorol Soc 98(7):1383–1398
Murata A, Sasaki H, Hanafusa M, Kurihara K (2013) Estimation of urban heat island intensity using biases in surface air temperature simulated by a nonhydrostatic regional climate model. Theor Appl Climatol 112(1):351–361
Sasaki H, Murata A, Hanafusa M, Oh'izumi M, Kurihara K (2011) Reproducibility of present climate in a non-hydrostatic regional climate model nested within an atmosphere general circulation model. Sola 7:173–176

Publisher's Note