ISSN:1738-6764, eISSN:2093-7504

Implementation of ESGF Data Node for International Distribution of CORDEX-East Asia Regional Climate Data

Jeongmin Han [1,*] and Jaewon Choi [2]

[1] APEC Climate Center; Research Fellow; goal@apcc21.org
[2] APEC Climate Center; Researcher; jwchoi@apcc21.org
* Correspondence

https://doi.org/10.5392/IJoC.2021.17.1.061

Manuscript Received 4 March 2021; Received 24 March 2021; Accepted 24 March 2021

Abstract: *As the resolution of climate change scenario data applied with regional models increased, Earth System Grid Federation (ESGF) was established around major climate-related organizations to jointly operated and manage large-scale climate data. ESGF developed standard software to provide model output, observation data management, dissemination, and analysis using Peer to Peer (P2P) computing technology. Roles of each institution were divided into index and data nodes. Therefore, ESGF data node was established at APEC Climate Center in Korea on behalf of Asia to share data on climate change scenarios of CORDEX-East Asia (CORDEX-EA) to study climate changes in Eastern Asia. Climate researchers are expected to play a large role in researching causes of global warming and responding to climate change by providing CORDEX-EA regional model data to the world through ESGF data node.*

Keywords: CORDEX-EA; ESGF; APEC Climate Center; P2P

1. Introduction

As international standard experiments to produce scenario data that are used to respond to global climate change are increasing rapidly, the volume of data is also continuously increasing. in order to respond to CMIP6/AR6, a study on CMIP6 for the IPCC 6th Assessment Report (AR6, scheduled to be published in 2021) is actively underway internationally. It is preparing to produce climate change information using the CMIP6 participation global model data updated in 2020, and the socio-economic scenario (SSP) [1].

In particular, several universities (UNIST, POSTECH, PNU, KNU), and the National Institute of meteorological science in Korea participated in the CORDEX project, and in the CORDEX Phase 1 study based on CMIP5, it was produced by the Statistical Downscaling method. the second-stage research produces more detailed (25km) regional climate change information by raising the model resolution compared to the horizontal resolution (50km) of the existing first-stage, and uses data from the second-stage regional climate change scenario newly produced in the Earth System Grid federation (ESGF) [2, 3].

ESGF has formed an alliance of several organizations for joint management and operation of climate data, and for the distribution of CORDEX East Asia data produced in Korea, the APEC Climate Center complies with the international standard system for effective management, sharing and distribution of large-capacity data. Through this, we intend to play an international leading role in information production and provision through participation in ESGF and CORDEX [4].

2. Materials and Methods

2.1 Materials

CORDEX (Coordinated Regional Climate Down-scaling Experiment) is an international project supervised by the WCRP (World Climate Research Program), and produces detailed climate change scenarios standardized for each region using the epidemiological and statistical downscaling technique. [5-7]. The vision of CORDEX is to advance and coordinate the science and application of regional climate downscaling through global partnerships [8].

There is a total of 14 areas of CORDEX Core, and Korea belongs to CORDEX-East Asia [9]. In addition, it is responsible for producing and operating CORDEX-East Asia data of the 14 regional model data, South Asia is distributed in India, but since Korea is the only East Asia region, it is essential to establish an ESGF data node for joint research to prepare for climate change [7, 9, 10]. CORDEX-EA domain is shown in Figure1.

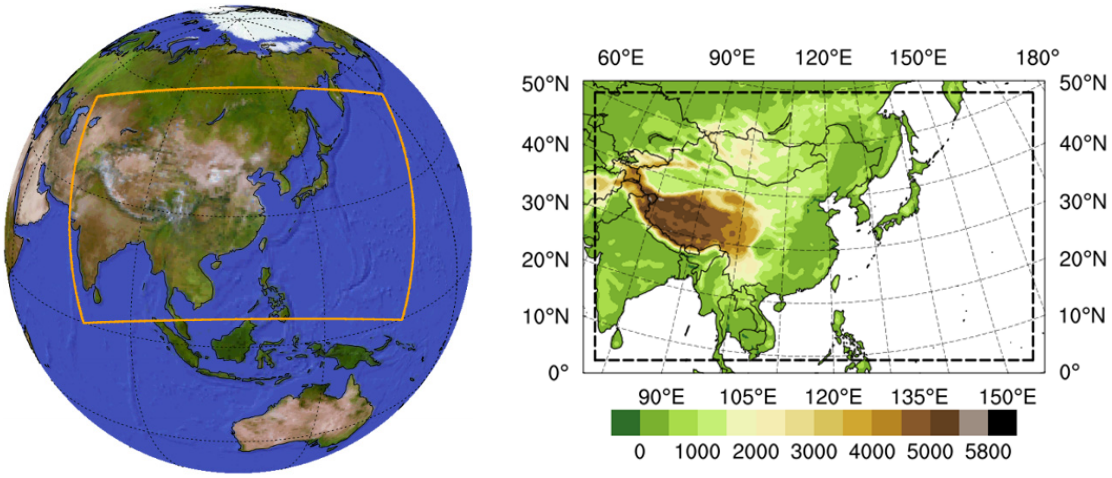

Figure 1. The domain of CORDEX-EA (Coordinated Regional Climate Down-scaling Experiment-East Asia) is contained within the coordinates of 0 to 50 of latitude and 60 to 180 longitude

2.2 Previous Research

The APEC Climate Center is the only one in Asia to build an ESGF data node to provide CORDEX East Asia data. The ESGF data node of APEC Climate Center serves to service CORDEX East Asia data to users through real-time communication with the Swedish supercomputer center. Representative index nodes include DOE/LLNL/NASA (USA), IPSL (France), NCI (Australia) DKRZ (Germany), SMHI-NSC (Sweden), and BSDC (UK) Have [11, 12].

In particular, data nodes are being operated in Korea, India, China, and Japan, but the characteristics of the data are different because it provides data on global climate change scenarios in Asia.

India's ESGF data node produces and provides CORDEX regional model data, but does not provide data for East Asia including Korea, China, and Japan because it is in charge of South Asia among the 14 areas set by CORDEX. India has built its services only in South Asia area.

China and one part are operating ESGF data nodes, but they provide only CMIP5 global climate change scenario data, which is far from the CORDEX regional climate model data that this study intends to provide.

In summary, ESGF nodes in Europe and America serve as data services for each region, so the ESGF data node of the APEC Climate Center is the only data for East Asia region used in this study.

2.3 Data Node Requirements

ESGF consists of index node, data node, computation node, and IDP node [5, 10]. The data node of APEC Climate Center belongs to the ESGF Tier2 node, and in order to be linked to the ESGF index node, it must comply with the requirements of the ESGF alliance. If the requirements are not met, there is a policy to exclude it from the role as a node, so a stable system configuration is required to continue the role of the ESGF data node [3, 13, 18]. The major differences between Index node and data node sites with respect to ESGF infrastructure are the complexity of provided services and service level agreements. The list of requirements includes

- Have an uptime of >90%.
- Provide the data node part of ESGF software stack.

- Install the most recent version of ESGF software within two weeks.
- Prompt upgrade in case of detected security breaches (<7 days).
- Be responsible for the node maintenance and operation.
- Exclusion of data nodes: There is an ongoing proposition to enable ESGF to exclude a data node that does not satisfy all the ESGF node operation requirements or a data node that will degrade the federation usability.
- Are involved in primary data publication.
- Provide sufficient storage and network bandwidth as required by their supported data projects.

3. System Design

3.1 CORDEX-East Asia Data

The CORDEX East Asia Phase 2 data is produced in Korea, CORDEX East Asia Phase 1 data is provided through the CORDEX East Asia Data Center website, and Phase 2 data is newly created from 2019. Phase I data with a resolution of 50 km were produced in the first phase, and phase II data with a resolution of 25 km were in production. The first stage data consisted of one HadGEM2-AO GCM, but the second stage data consisted of two types of CMIP5 from HadGEM-AO, MPI-ESM-LR, and GFDL, and one type of CMIP6 from UK-ESM [14]. There are 5 types of RCM data produced through this Table1.

Table 1. As of November 2020, there are a total of 665 simulations (including evaluation, historical and scenario runs) in the CORDEX-EA

variable	long name	1h	3h	day	mon
tas	Near-Surface Air Temperature		●	●	●
tasmax	Daily Maximum Near-Surface Air Temperature			●	●
tasmin	Daily Minimum Near-Surface Air Temperature			●	●
pr	Precipitation	●	●	●	●
ps	Surface Air Pressure		●	●	●
hurs	Near-Surface Relative Humidity			●	●
sfcWind	Near-Surface Wind Speed		●	●	●
rsds	Surface Downwelling Shortwave Radiation		●	●	●
sfcWindmax	Daily Maximum Near-Surface Wind Speed			●	
rlds	Surface Downwelling Longwave Radiation		●	●	●
hfls	Surface Upward Latent Heat Flux		●	●	●
hfss	Surface Upward Sensible Heat Flux		●	●	●
rsus	Surface Upwelling Shortwave Radiation		●	●	●
rlus	Surface Upwelling Longwave Radiation		●	●	●
evspsbl	Evaporation		●	●	●
mrro	Total Runoff			●	
mrso	Total Soil Moisture Content			●	
snw	Surface Snow Amount			●	●
prc	Convective Precipitation		●	●	●
prw	Water Vapor Path			●	●
ua200	Eastward Wind		●	●	●
va200	Northward Wind		●	●	●
ta200	Air Temperature		●	●	●
zg200	Geopotential Height		●	●	●
ua500	Eastward Wind		●	●	●
va500	Northward Wind		●	●	●
ta500	Air Temperature		●	●	●
zg500	Geopotential Height		●	●	●
ua700	Eastward Wind			●	●

va700	Northward Wind		●	●
ta700	Air Temperature		●	●
hus700	Specific Humidity		●	●
ua850	Eastward Wind	●	●	●
va850	Northward Wind	●	●	●
ta850	Air Temperature	●	●	●
hus850	Specific Humidity	●	●	●
ua925	Eastward Wind	●	●	●
va925	Northward Wind	●	●	●
ta925	Air Temperature	●	●	●
hus925	Specific Humidity	●	●	●
ua100m	Eastward 100m Wind		●	●
va100m	Northward 100m Wind		●	●

As for the historical data, the first stage data was from 1979 to 2005, for 27 years, but the second stage data was from 1979 to 2005, CMIP5 data, 27 years and 1970 to 2014, CMIP6 45 years. In the scenario, 95 years (2006-2100) data were produced for RCP4.5 and RCP8.5 in the first stage, but in the second stage, 95 years (2006-2100) for CMIP5 and 86 for CMIP6 for RCP2.6 and RCP8.5. It is the year (2015~2100). The ERA-I forcing is 27 years (1979-2005) for the first stage, but 37 years (1979-2015) for the second stage. The experimental data were RCM of 25 km, and HeadGEM2-AO, MPI-ESM-LR, GFDL-ESM2M, UK-ESM models were used [6, 9, 10].

The institute and models that produce CORDEX data in East Asia are as shown in the Table 2 below [1].

Table 2. The institute and models that produce CORDEX-EA

RCP	HadGeM2-AO	MPI-ESM-LR	GFDL2M
RCP8.5 (2006~2100)	CCLM(POSTECH)	MM5(UNIST)	WRF(PNU) RegCM4(KNU)
RCP2.6 (2006~2100)	RegCM4(KNU) MM5(UNIST) CCLM(POSTECH)	WRF(PNU) CCLM(POSTECH) MM5(UNIST)	WRF(PNU) RegCM4(KNU)
HISTORICAL (1979~2005)	CCLM(POSTECH)	MM5(UNIST)	WRF(PNU) RegCM4(KNU)

3.2 Configuration of ESGF Data Node System

To build an ESGF data node, we prepared 2 servers and 1 storage set. The server is a model produced by Dell and Fujitsu, and the operating system is prepared based on CentOS 6, and the latest patches are in progress. The storage consists of Raid6 and has 2 parities and 1 spare. The system specifications and configuration are shown in Figure2.

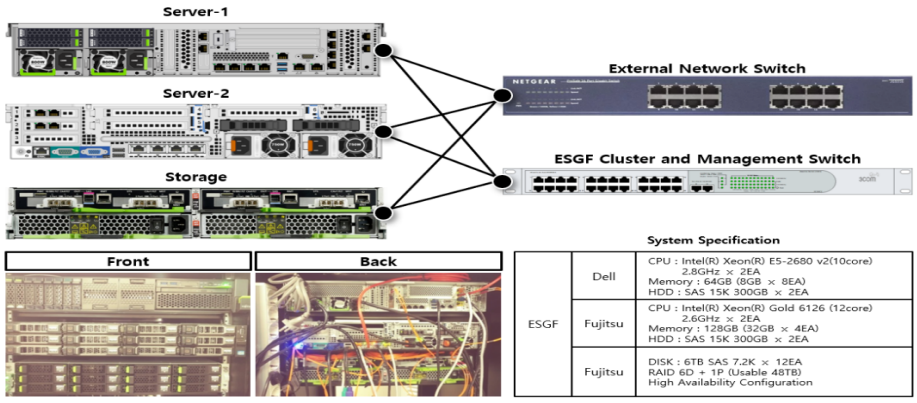

Figure 2. For stable service, ESGF Data Node was built using two servers and one storage set. ESGF Data node has System specification like that

ESGF data node was built using two servers with different specifications and a set of storage devices. Servers are made by Dell and Fujitsu, respectively, because the server specifications are different, so it is not possible to configure the mirroring method, and there has been a problem that the work content of one server is not recognized by the other server in real time. Therefore, server clustering was designed to share the service area using the storage area using Raid6, a more stable storage configuration format [15].

When the running server becomes unstable, the other server in standby blocks the unstable server and operates instead, and the hardware is configured in the form of Active-Standby. In addition, a software configuration and a range of errors were set, and a program for the operating processor was added so that the replacement service can operate smoothly due to the occurrence of a failure.

As shown in the Figure3 below, the CORDEX East Asia data center or ESGF index node performs a role of searching data and providing the searched data to users. The actual data is provided on the ESGF data server, so if there is a large data request or an error occurs, the clustered console determines whether there is a failure through monitoring and immediately switches the service to another server on standby. The running server was built to restore the software that was judged to be a failure by itself to a normal state and to switch to the standby state for the next service.

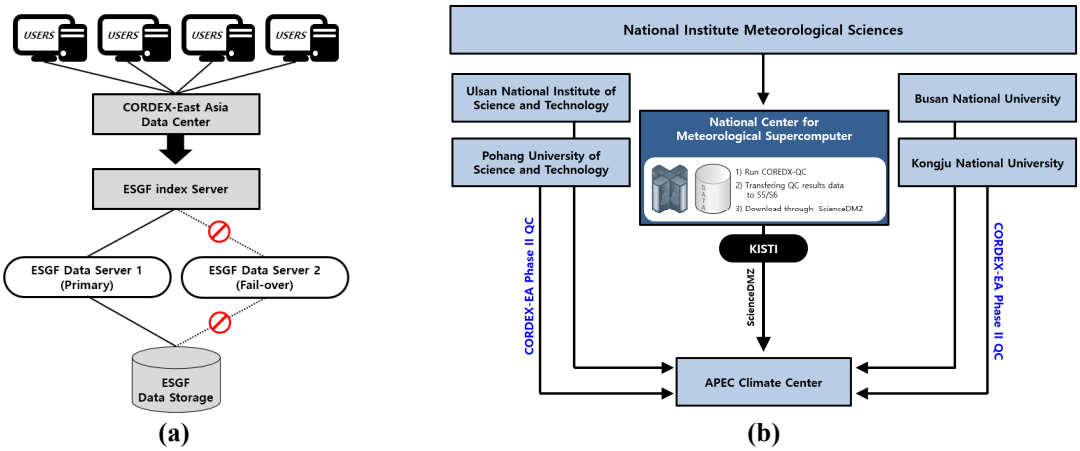

Figure 3. If there is a large data request or an error occurs, the clustered console determines whether there is a failure through monitoring and immediately switches the service to another server on standby(a) and Due to the network separation, data produced by each university are transmitted to APEC Climate Center through the meteorological supercomputer center. (b)

The production of CORDEX data is in charge of the National Institute of meteorological Science and each university, and the distribution of the data is in charge of the APEC Climate Center, so all data produced must be collected by the ESGF data node in the APEC Climate Center. However, since each institution's network is diverse, the transmission form is also inevitably varied. In order to solve this problem, the data transmitted from neighboring countries were reviewed by the Meteorological Academy and then transmitted to the APEC Climate Center through the Science DMZ network through the Korean Meteorological Supercomputer. Each university used the Internet network to collect it in the storage of the external communication network of the APEC Climate Center.

4. Implementation of ESGF Data Node

4.1 Meta-data production and distribution

The ESGF data node built in the APEC Climate Center is configured in the form of parsing and processing XML documents based on java. Basically, http web communication is used and the data structure in PostgreSQL is configured as a database. It also uses apache Solar for search and Globus FTP for data distribution. The basic web server is composed of apache tomcat and apache https, as described above, the data collected by each university is the data obtained by downscaling the regional data from the global data by applying the unique model of each university as the regional model data [13, 16].

In order to distribute the collected data as shown in the Figure4, the first thing to do is check whether the produced data complies with the rules set by CORDEX [10, 17]. For the data that passed the inspection,

metadata is extracted through the attribute information that the data should have, and the extracted metadata is stored in the database, and at the same time, it is created as an xml document. Based on the created xml document, a web-based “thredds” document is created by copying the actual file, and meta information is transmitted by accessing the index node system. The transmitted meta information is stored in the index node database and used as metadata for search. When the user searches for the desired file through the search and executes the download, the user connects to the server where the actual file is located and receives the data. In addition, although the data node is linked to one index node, the index node is linked to each other, so that the meta information is updated in real time, so that the same result can be obtained even if a search is performed on another index node [14].

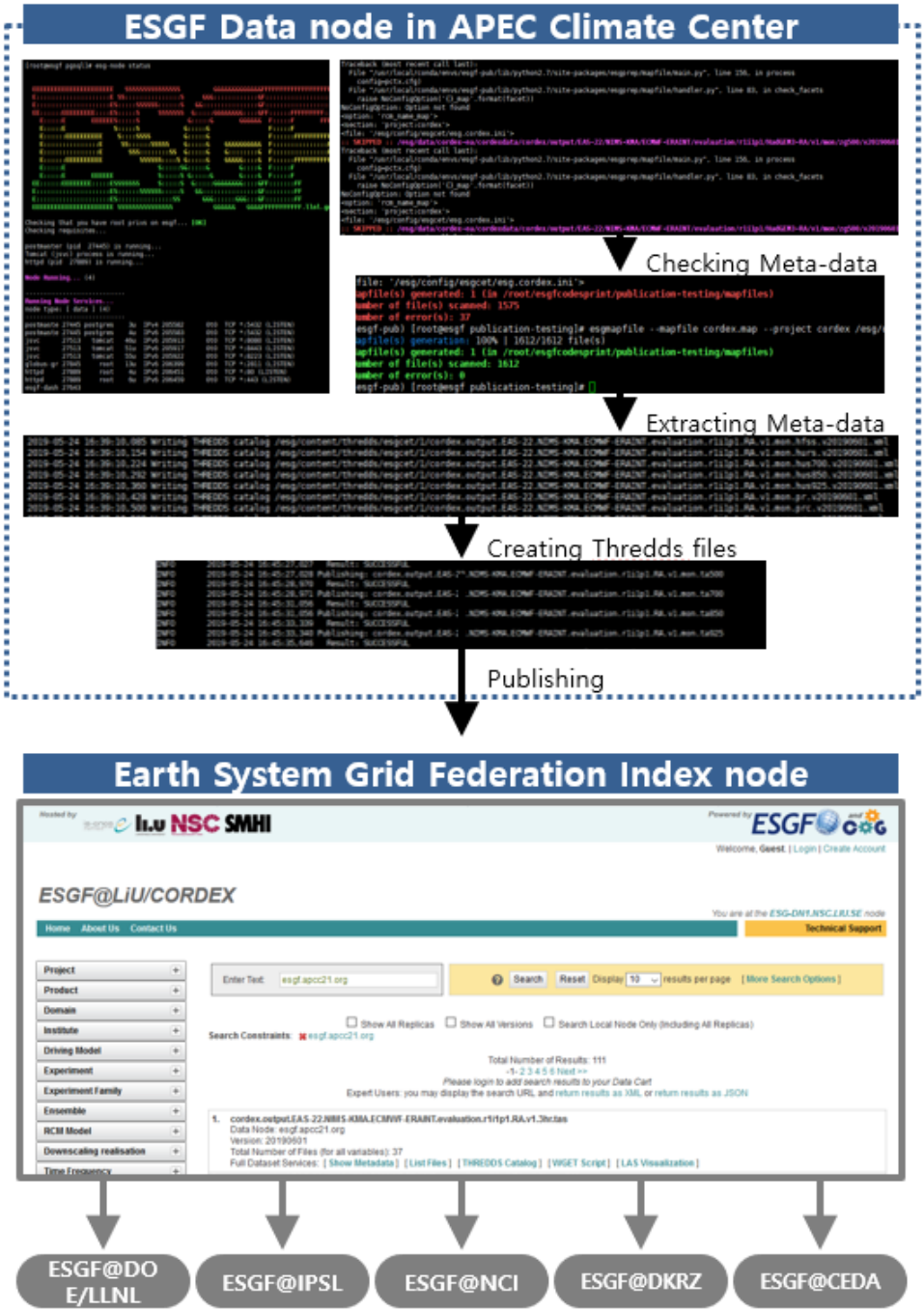

Figure 4. The thredds catalog file is extracting meta-data from the data in the Korean ESGF Data Node. When the generated meta-data is transmitted to the Swedish ESGF Index Node, the meta is shared with the ESGF Index Node in each country

4.2 Data retrieval and delivery

The metadata generated by the data node is published to the index node. Based on the provided metadata, the index node provides various services to users as shown in the following screen. The main services provided by the ESGF node are as follows.

First, the solar search engine provided by Apache was used to search for ESGF data nodes. This search engine is an engine that stores and searches as an index, a structure that stores words and addresses in key and value based on morpheme analysis. In order to search data distributed in each node, metadata is stored in the XML Repository method, and XML documents are parsed and stored as indexing. The method of searching through the ESGF data node can utilize the basic search function and the extended search function. As shown in the Figure5 below, if you use the extended search function, you can search by variable name, production date or data set period. You can also use wildcards like '*' in the search name.

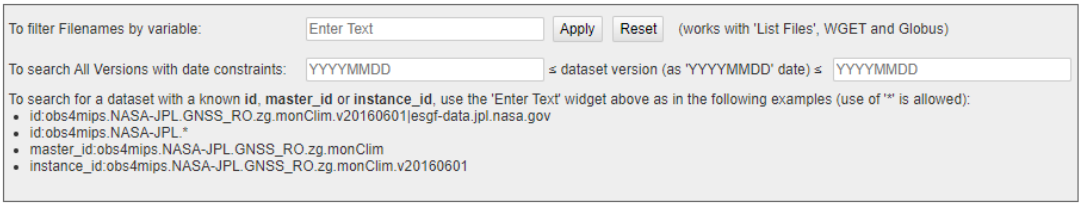

Figure 5. ESGF provides simple search and extended search for convenient data search, you can use the wildcards on extended search

Second, the results retrieved from the ESGF data node are grouped into a data set and displayed as a list. The search list is basically displayed in the form of a web list, or users can view the search results in XML or JSON format. This is to support a new type of extended service by using the search result of the ESGF data node. As shown in the Figure6 below, this type of provision is to provide web crawling that extracts necessary information from other service, so user can get the information of the file directly without downloading the file.

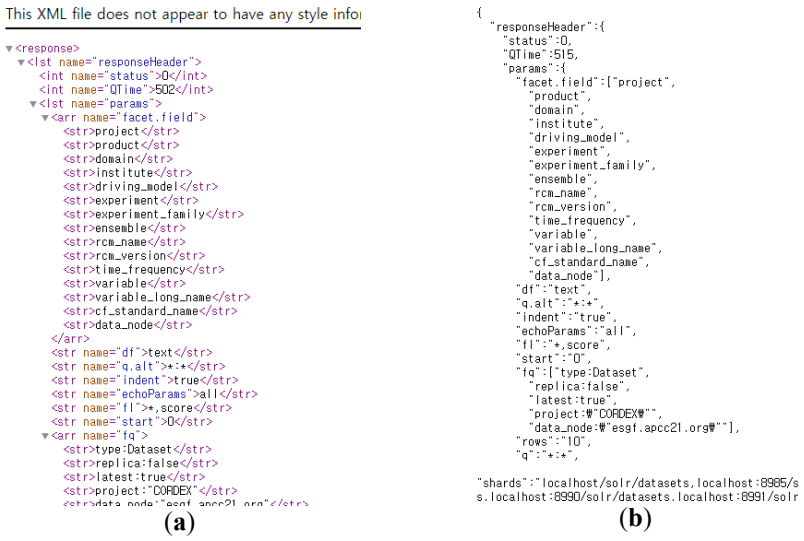

Figure 6. ESGF provides XML(a) and JSON file(b) formats

Third, the results searched through the ESGF node are provided as a search list, and only the version based on the data production date and the total number of files are briefly displayed. Therefore, when users want to obtain more information of a file, a meta information view menu is added to avoid the hassle of downloading the file and checking the property information of the file, so that the user can check whether the user wants the data through the meta information of the data set. As shown in the Figure7 below, meta information includes file ID and version, time stamp, data node name, date time start, domain name, etc. if you select the file list, the file list of the data set is displayed, and the user can see the meta information of each file. You can also download files directly from here.

(a) **(b)**

Figure 7. Meta information includes file ID and version, time stamp, data node name, date time start, domain name, etc. **(a)** and File list views of the data set is displayed, and the user can see the meta information of each file **(b)**

Fourth, for the searched data, THREDDS, HTTP, OPeNDAP, and "wget" Script are provided so that data can be downloaded in various ways. The method of downloading with OPeNDAP is useful when extracting only values from a file by extracting data based on the condition of properties, variables, and coordinates based on the browser. On the other hand, if you select the "wget" script menu, a script is automatically created for the list searched in the system and a file is provided. As shown in the Figure8 below, this is provided so that users can conveniently use it when downloading files in a Linux or Mac environment, and users can download many files at once by running a script on Linux or Mac. The user selected the desired method and provided the data to be used.

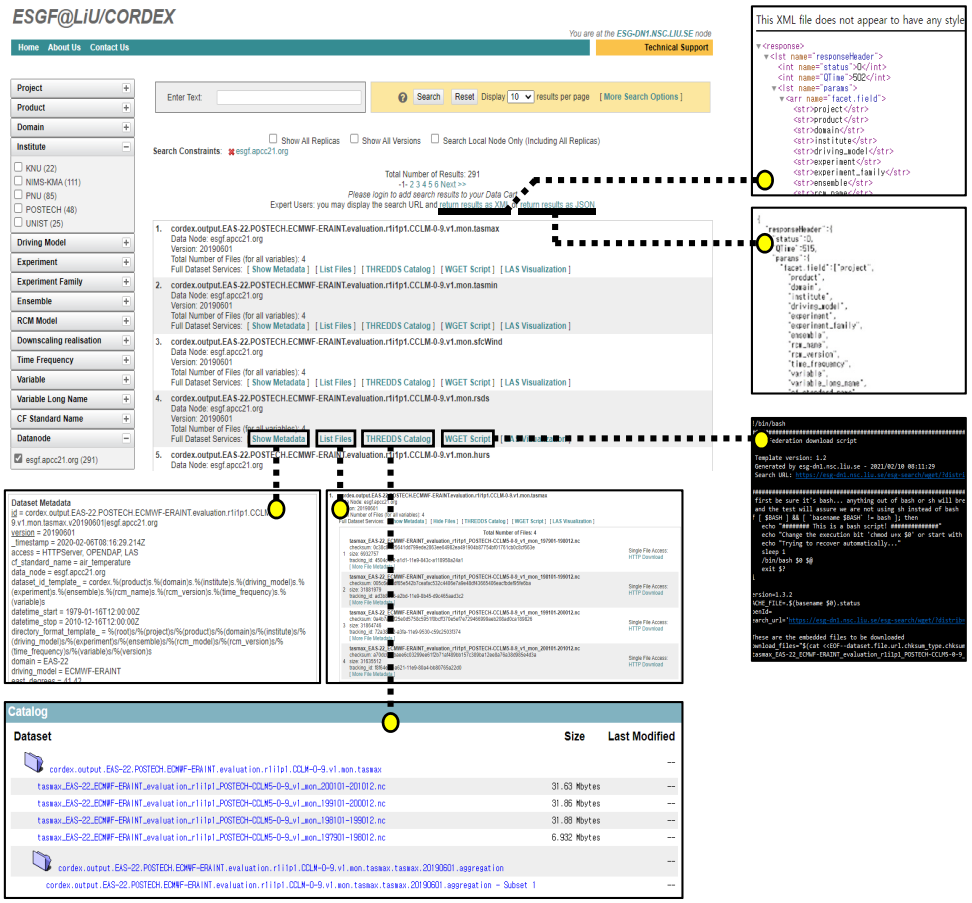

Figure 8. User can search for the desired data on the ESGF index node. You can get the data information of the searched list and download the file list in the format you want in XML or JSON format. In addition, not only can download files using "weget" or "OPeNDAP"

4.3 pefermance evaluation

Performance evaluation was conducted to check if ESGF's requirements were met for the ESGF Data Node built in APEC Climate Center. After staring the service, two servers were clustered in Active-standby format to comply with more than 90% uptime. If a failure occurs in one server during operation, another server in standby is configured to operate in place of the failed server. 329 days per year must be operated to meet certain requirements.

In addition, the essential software provided by ESGF was configured to be used as an application and I installed the latest version in 2.8RC. also, the intrusion detection system has connected to maintain the best state of system and is configured to play the role of distributing and publishing major data. To secure enough storage space, 160TB of storage was configured and the network bandwidth was expanded to transmit data of 2Gbit per second.

The construed system at APEC Climate Center received comprehensive evaluation from ESGF and obtained formal approval as ESGF's data node. As shown in the Figure9 below, after approval from ESGF, a total of 5,238(291 datasets) CORDEX-EA regional climate model data were distributed worldwide.

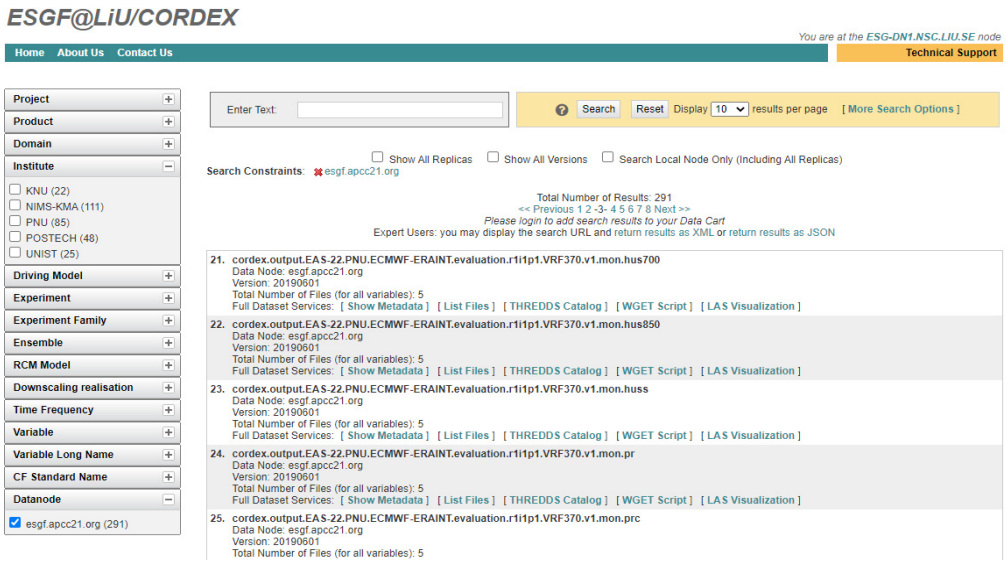

Figure 9. This is the result of searching Korea's data node from the Swedish index node. We can check a total of 291 data sets, and we can also check the name of the institution that produced the data.

5. Conclusions

In this study, APEC Climate Center completed the system configuration for building ESGF Data Node, and obtained the approval for index linkage from ESGF Index Node. In order to meet the requirements of ESGF, server redundancy was configured and the Archive required by CORDEX was followed. The APEC Climate Center uses a network separation policy, so the data received from the Institute of Meteorological Sciences uses Science-DMZ of the Korea Meteorological Supercomputer Center, and each university has tried to minimize the time constraints in collecting large amounts of data using the Internet network.

Many foreign institutions operate ESGF data nodes, but the provided data are different. In addition, while Asia is participating in ESGF in Japan, China and India, China and Japan are providing data different from those provided by the APEC Climate Center, and India provides the same CORDEX data, but East Asia data as South Asian data. National Institute of Meteorological Science and domestic universities are the only places that produce and provide them.

CORDEX-EA regional climate model data produced in Korea were distributed through the ESGF data node of the APEC Climate Center, and they play an important role in providing CORDEX data to the world.

Acknowledgments: This research was supported by the APEC Climate Center and KREONET(Korea Research Environment Open NETwork) which is managed by Korea Institute of Science and Technology Information

Conflicts of Interest: The authors declare no conflict of interest and the funders had no role in the design of the study.

References

[1] Gayoung Kim, Dong-Hyun Cha, Changyong Park, Chun-sil Jin, Evaluation and Projection of Regional climate over East Asia in CORDEX-East Asia Phase 1 Experiment, Asia-Pacific Journal of the Atmospheric Sciences, vol. 57, no. 5, Feb, 2020, doi: https://doi.org/10.1007/s13143-020-00180-8.

[2] Philip Kershaw, Ghaleb Abdulla, Sasha Ames, and Ben Evans, "ESGF Future Architecture Report," LLNL-TR-812915, pp. 1-14, July. 2020, doi: https://doi.org/10.5281/zenodo.3928223.

[3] J.P.Evans, "CORDEX-An international climate downscaling initiative, international Congress on Modeling and Simulation,"19th international Congress on Modelling and Simulation-Sustaining Our Future :Understanding and Living with Uncertainty, pp. 2705-2711, 2011, doi: https://doi.org/10.36334/modsim.2011.A1.bae.

[4] WCRP, 2020. Accessed: Feb. 12, 2021. [Online] Available: https://cordex.org/wp-content/uploads/2020/09/Domain-Criteria-Document-FINAL.pdf.

[5] O. B. Christensen, W.J. Gutowski, G. Nikulin, and S. Legutke, "CORDEX Archive Design, Version3," vol. 2, no. 5, pp. 1-23, 2020, doi: https://doi.org/10.1007/s10584-020-02835-x.

[6] Luca Cinquini, The state of the Earth System Grid Federation, ESGF F2F, 2018.

[7] ESGF, 2021. Accessed: Feb. 12, 2021. [Online] Available: https://esgf.llnl.gov/esgf-media/pdf/ESGF-Tier1and2-NodeSiteequirements-V5.pdf.

[8] CORDEX, 2021. [Online] Available: https://cordex.org/about/what-is-regional-downscaling.

[9] CORDEX, 2015. Accessed: Feb. 12, 2021. [Online] Available: CORDEX domains for model integrations, 9, https://cordex.org/domains.

[10] Matt Pryor and Phil Kershaw, Alan iwi, Highly Available ESGF Services for the copernicus Climate Data Store, ESGF F2F, 2018.

[11] Shengjin Wang, Hongru Yang, Quoc Bao Pham, Dao Nguyen Khoi, and Pham Thi Thao Nhi, "An Ensemble Frame work to Investigate Wind Energy Sustainability Considering Climate Change Impacts," Sustainability, vol. 12, no. 3, pp. 12-17, 2020, doi: https://doi.org/10.3390/su12030876.

[12] Joong-Bae Ahn, Yeon-Woo Choi, and Sera Jo, "Evaluation of Reproduced Precipitation by WRF in the Region of CORDEX-East Asia Phase 2," Atmosphere Korean Meteorological Society, vol. 28, pp. 85-97, 2018, doi: https://doi.org/10.14191/Atomos.2018.28.1.085.

[13] ESGF, 2021. Accessed: Jan. 12, 2021. [Online] Available: https://esgf.llnl.gov/esgf-media/pdf/ESGF_Governance_5_11_2017.pdf.

[14] CORDEX, "List of data aspects that have to be checked at data digestion time into an ESGF CORDEX archive," Digestive Diseases and Sciences, pp. 1-3, 2013.

[15] J. M. Han and H. L. Lee, "Developing a participatory climate service platform. APEC Climate Center," 2020.

[16] CORDEX, 2021. Accessed: Feb. 12, 2021. [Online] Available: https://cordex.org/domains/region-7-east-asia.

[17] CORDEX, 2014. Accessed: Feb. 12, 2021. [Online] Available: CORDEX Variables requirement table, 1-9.

[18] CORDEX, 2020. Accessed: Mar. 12, 2021. [Online] Available: https://cordex.org/wp-content/uploads/2020/12/summary_CORDEX_simulations_ Nov_ 2020.pdf.