Development of climate data storage and processing model

Okladnikov I.G. (igor.okladnikov@gmail.com), Gordov E.P. (gordov@scert.ru), Titov A.G. (titov@scert.ru)
Institute of Monitoring of Climatic and Ecological Systems, Siberian Branch of Russian Academy of Sciences, Tomsk, Russia (http://www.imces.ru)
Institute of Computational Technologies, Siberian Branch of Russian Academy of Sciences, Tomsk branch, Tomsk, Russia (http://ict.nsc.ru)
Tomsk State University, Tomsk, Russia (http://www.tsu.ru)

Storage and processing model for climate datasets developed in the framework of a distributed research center for monitoring and predicting regional climate and environmental changes is presented. This research center is focused on processing and analysis of large volumes of climate data and able to help researchers (who are not experts in the field of information technologies) in the environmental and climate change study.

This work is supported by the Ministry of education and science of the Russian Federation, Agreement #14.613.21.0037.

Introduction

The considered model is based on a «shared nothing» distributed computing architecture [Valduriez P. Shared-Nothing Architecture / Encyclopedia of Database Systems. Eds.: Ling Liu, M. Tamer Özsu. –2009. – P. 2638-2639], where each computing node is independent and self-sufficient. On each compute node a dedicated software for processing and visualization of geospatial data - processing core - is installed; it provides programming interfaces to communicate with other nodes. The nodes are interconnected by local network or Internet and exchange data and control instructions via a SSH connection and WPS. Geospatial data is stored in sets of NetCDF binary files organized in a special directory hierarchy in the framework of a file system used on each compute node. To speed up data reading and processing three approaches are proposed: preliminary calculation of intermediate products, distribution of data across multiple storage systems (with or without redundancy), and caching and reuse of previously obtained products.

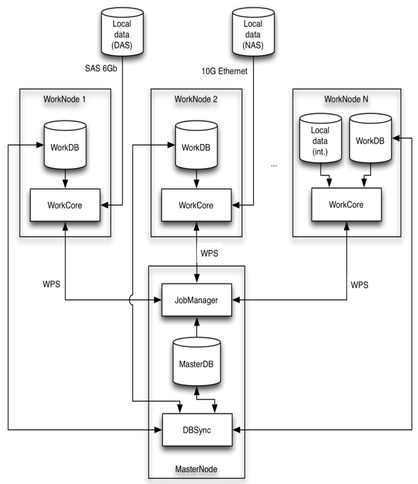

Fig.1. Architecture of a computing network

Spatial data archive

MRI/JMA APHRODITE's Water Resources Project data, ECMWF ERA-40 and ERA Interim Reanalysis, DWD GPCC data, JMA/CRIEPI JRA-25 and JRA-55 Reanalysis, MACC Project data, GMAO MERRA reanalysis, NCEP/DOE AMIP II Reanalysis, NOAA-CIRES Twentieth Century Global Reanalysis Version II, NCEP Climate Forecast System Reanalysis (CFSR), PlaSim model output, meteorological observational data for the territory of the former USSR for the 20th century.

Architecture

Georeferenced datasets are currently actively used for modeling, interpretation and forecasting of climatic and ecosystem changes on different spatial and temporal scales. Due to inherent heterogeneity of environmental datasets as well as their huge size (up to tens terabytes for a single dataset) a special software supporting studies in the climate and environmental change areas is required. An architecture of information-computational system (ICS) based on spatial data services is developed (Fig.2).

Major components:

1.Structured georeferenced data sets stored on high performance storage systems.
2.Modular computational core for data processing and visualization.
3.Metadata catalog based on ISO 19115 standards and CF conventions, published according to OGC CSW (Catalog Service for the Web) service specifications.
4.Computational and mapping web services for spatial data processing based on OGC WMS, WFS, WCS.
5.Geoportal as a joint access point to spatial geophysical data and metadata, web mapping and geo processing services, as well as ready products in the framework of SDI paradigm.
6.Web client represents client software (including desktop GIS, such as uDIG or QuantumGIS) relying on spatial data services.

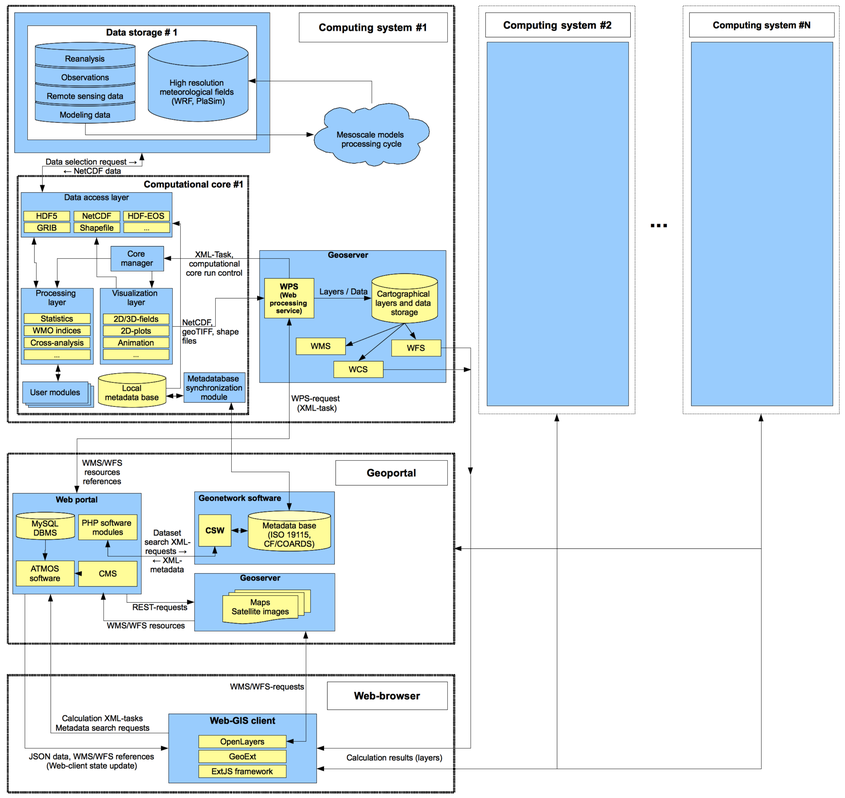

Fig. 2. ICS architecture

Metadata database

For a fast search and retrieval of a required data according to given parameters in the framework of the developed data storage and processing model, a metadata database, containing descriptions of space-time features of datasets available for processing, locations of data files, as well as descriptions and run parameters of analysis and visualization software components is developed. In general, the developed model supports functioning of the distributed research center for monitoring and predicting regional climate and environmental changes. It provides "data awareness" principle for spatial data location and processing.

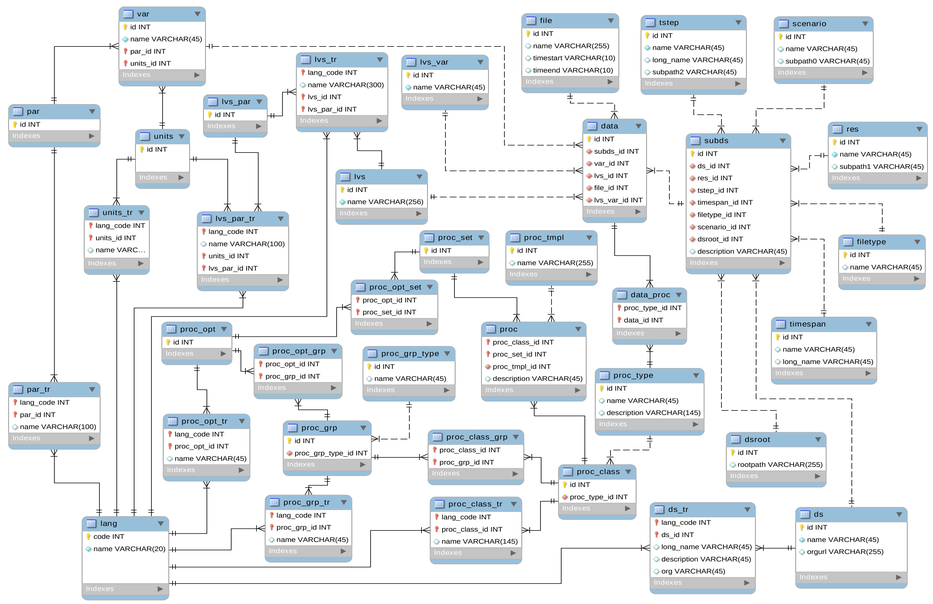

Fig. 3. Metadata database ER-diagram

Information-computational system prototype

According to the architecture presented a prototype of information-computational system for integrated analysis of global and regional climate change using heterogeneous georeferenced data is developed (Fig. 4).

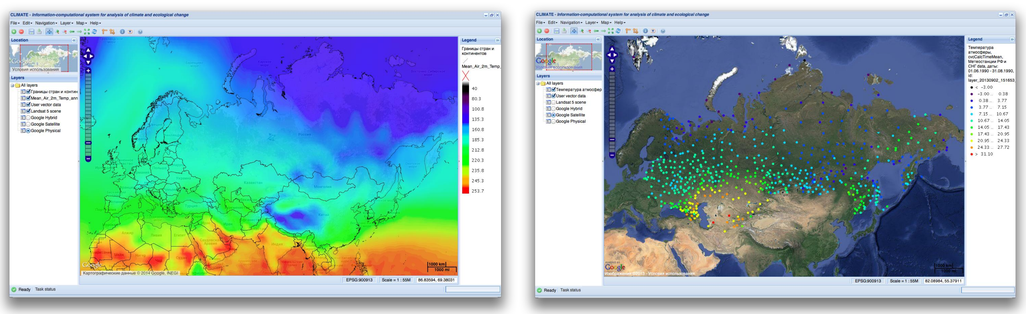

Fig. 4. Examples of spatial data processing by ICS in a web browser window