Computing core of a software package for “cloud” analysis of climate change and the environment

I G Okladnikov
Institute of Monitoring of Climatic and Ecological Systems SB RAS, Tomsk, Russia
Institute of Computational Technologies SB RAS, Tomsk Branch, Tomsk, Russia

Abstract. The computing core of an information and computing software package based on a dedicated software framework for carrying out scientific research related to statistical processing and analysis of spatial geophysical data archives obtained both from observations and modelling is presented. In its development, accumulated experience in the development of information-computational web GISs for processing of large amounts of spatial data has been used. The basic components of the computing core are represented by modules for searching, selecting, and processing spatial data arrays and by a core manager governing the processing and data flows. Its modular structure provides a possibility to operatively expand the functionality of the software complex using various procedures for mathematical and statistical analysis, processing and graphical representation of the results in the form of graphs, diagrams, and fields on a map of the territory under consideration.

1. Introduction
Climatic changes stimulated the development of environment observation and modeling systems, which has already led to the emergence of archives of spatial data reaching petabyte volumes. Analysis of such data becomes impossible without an appropriate computational and informational support. Development of effective tools for computational and informational support of analysis of large data sets, and organization of their effective use for obtaining and practical application of new knowledge is one of the areas of modern climate science. In fact, to solve this problem it is necessary to create a software environment based on modern information and telecommunication technologies [1, 2] that meets the requirements of a spatial data infrastructure (SDI, [3, 4]), which implies the use of modern technologies for processing of geophysical data that allow integrating various software solutions for performing cloud computing using remote high-performance computing resources. The development of thematic information and computational complexes that form the necessary infrastructure should be based on Web-GIS technologies [5-8]. Their application is a promising way to improve the efficiency of multidisciplinary regional and global research in the field of Earth sciences, including the analysis of climate change and its impact on the spatiotemporal behavior of plant ecosystems.

Nowadays, there are several web-oriented information and computing systems dedicated to the processing of spatially referenced geophysical data. Some of them are: a dedicated system for analyzing meteorological data in real time [9]; a much more functional GES-DISC Interactive Online Visualization ANd aNalysis Infrastructure (GIOVANNI) system developed at NASA for processing and visualizing satellite observations (http://daac.gsfc.nasa.gov/techlab/giovanni/); the RIMS, one of
the most advanced systems for integrated online processing of heterogeneous data on climate, hydrology, remote sensing, etc. (http://RIMS.unh.edu/), and others.

However, despite of a number of attempts made in the field of informatization of Earth sciences, there is still no high-performance tool that provides unified user and software interfaces and combines ample opportunities for processing, analyzing, and visualizing datasets obtained from various sources for the integrated study of global and regional climate change.

This paper presents the computing core of a software package for analysis of the climate change and the environment, its key components, and capabilities.

2. Results
One of the key elements of the software package is a modular computing core. It is a set of software components interacting through a unified application program interface (API) and performing search, selection, processing, and visualization of spatial data. The software package is focused on "cloud" distributed processing of large sets of spatial climatic data and can be deployed on several computational nodes interconnected by data transmission channels. A separate computing core is installed on each node, linked with its own spatial data archive and metadata database [9]. The selection of a required computing core, its launch and execution control are carried out by a geoportal using a principle of “data locality”. The geoportal also implements the logic of web applications, communication with mapping web services, and provides work with the metadata database. The archives of spatial data associated with the computing core can be modeling data (reanalysis, global and regional climatic and meteorological models) and/or observational data (meteorological stations, satellite observation data, etc.) of various spatial resolutions, set at different spatiotemporal domains and presented in the NetCDF or HDF format [10].

The computing core was developed using the GNU Data Language (GDL, http://gnudatalanguage.sourceforge.net/) and Python (http://python.org) programming languages, which provide proven procedures for mathematical processing and visualization of spatial data, as well as programming interfaces for reading/writing files in NetCDF and ESRI Shapefile formats, and access to PostgreSQL DBMSs. To manage the core input-output, form a computation pipeline, and control the execution of modules the computing core manager is used (Figure 1). It provides management of the core modules and transfer of intermediate results between them. Data processing is performed within the framework of a computation pipeline, which is represented by a sequence of calls to computing core modules with transfer of data from outputs of one module to inputs of another. The computation pipeline is formed using a special task file in the XML format, prepared by the web portal based on the results of user actions in the graphical interface. The task file contains key characteristics of the spatial data sets to be processed, a description of the computation pipeline in the form of a sequence of calls to computational modules, and descriptions of the intermediate data arrays transferred between them, as well as parameters for writing processing results to a graphic file. A DTD scheme of this XML task file is presented below.

```xml
<!ELEMENT task (service, metadb, computing, (data|destination)+, processing+)>
<!ATTLIST task
  uid ID #REQUIRED
  owner CDATA #IMPLIED
  description CDATA #IMPLIED
>
<!ELEMENT service (timePeriod)>
<!ATTLIST service
  processingClass CDATA #REQUIRED
  calculationParameter CDATA #REQUIRED
  trend CDATA #REQUIRED
  isBaseLayer CDATA #REQUIRED
>
<!ELEMENT timePeriod (dateStart, dateEnd)>
<!ATTLIST timePeriod
  type CDATA #REQUIRED
>
<!ELEMENT dateStart EMPTY>
<!ATTLIST dateStart
  year CDATA #REQUIRED>
```
According to this scheme, XML task files are created by the geoportal. For example, this is a task file describing the calculation of the average air temperature for 1961 at a height of 2 m using the ECMWF ERA-40 reanalysis data in each grid cell with a resolution of 2.5x2.5 degrees for the entire Earth.

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE task SYSTEM "ctdl.dtd"
This task file corresponds to the following algorithm in a verbal form:
1) call the module `cvcCalcTiMean` for calculating the average for Data_1 data corresponding to the air temperature at a height of 2 m, with a spatial resolution of 2.5x2.5 degrees and a time step of 6 hours from the ECMWF ERA-40 reanalysis with an additional parameter timeMean = 'data'; save the intermediate result to an array named Result_1.
2) write the Result_1 array to the ‘output.tif’ graphic file in GeoTIFF format together with the ‘output.sld’ legend as a Styled Layer Descriptor - SLD in XML format by calling the cvcOutput module,
3) write the Result_1 array to the ‘output.nc’ file in NetCDF format by calling the cvcOutput module.

Functionally, this pipeline looks like this:
1) Result_1 = cvcCalcTiMean_1(Data_1, ModuleParameters_1);
2) cvcOutput(Result_1, 'output.tif', 'output.sld');
3) cvcOutput(Result_1, 'output.nc');

Based on the task file, the core manager forms a computation pipeline and provides sequential launches of the corresponding computing core modules passing data arrays with intermediate results from one to another (Figure 1).

The computing core processing modules (computational modules) are a set of independent classes written in the GNU Data Language programming language that provide numerical processing of geophysical data stored on data storage systems. A general scheme of operation of each module is as follows:
- module initialization,
- reading input data via a unified internal API (using data access modules),
- data processing by the internal logic of the module,
- return the result to the computing core manager via a unified internal API.

The input data for the module are spatial data arrays set on a given space-time domain. The output data of the computational module are also spatial data arrays which are results of processing of the input arrays. The computing core manager provides storage and transfer of intermediate results from one module to another, including passing results to data access modules for writing to output files.

Each software module for numerical data processing is responsible for one type of processing. By combining various modules in a sequence, it becomes possible to form workflows for spatial data processing of any complexity. The results of processing in raster form are saved in graphic files of the GeoTIFF format, and those in vector form in ESRI Shapefile files. In this case, arrays of data are written to files in the NetCDF format containing additional metadata. Subsequently, these files are passed to the geoportal for preparation and displaying of cartographic layers to the user in a GUI on an interactive map.

3. Conclusions

This work is aimed at developing a thematic software package for analysis of climatic and environmental changes by integrating the interdisciplinary (geographic, climatic, meteorological) archives of observational, modeling, and remote sensing datasets in a thematically distributed information and computing system with the GIS functionality and the ability of "cloud" processing of heterogeneous spatial data. This computer complex is the next step in the development of applied information and telecommunication systems. It provides the specialists in various fields of science with unique opportunities for reliable analysis of heterogeneous geophysical data. The use of proven
computational algorithms will ensure the reliability of the results obtained in specific subject areas. The availability of the system on the Internet and the ability to work with data without using special knowledge in programming will allow a wide range of researchers and decision-makers to concentrate on solving their specific problems.

![Diagram of the computing core manager](image)

**Figure 1.** General algorithm of the computing core manager.

The computing core, which is one of the key elements of the software package for the analysis of climatic and environmental changes, is an independent software unit that includes a set of modules for accessing and processing spatial data and a core manager that controls the execution of modules and data transfer between them. It can be easily adapted to various sets of geophysical data and tasks from various areas of the Earth sciences. The open source software used in its development makes it possible to deploy and use it non-commercially on various hardware and software platforms running the Linux operating system.

**Acknowledgments**
This work was performed in the framework of state budget theme no. AAAAA-A17-117013050037-0. The development of the core manager was performed in the framework of ICT SB RAS state assignment program no. 0316-2018-0002.
References

[1] Nativi S, Ramamurthy M and Ritschel B. EGU-ESSI Position Paper https://sites.google.com/a/imaa.cnr.it/egu-essi/Home/documents

[2] Gordov E P and Lykossov V N 2007 Development of information-computational infrastructure for integrated study of Siberia environment J. Computational technologies 12 (Special issue 2) 19-30

[3] Steiniger S and Hunter A J S 2012 Free and open source GIS software for building a spatial data infrastructure Geospatial Free and Open Source Software in the 21st Century ed E Bocher and M Neteler (Heidelberg: LNGC, Springer) pp 247-61

[4] Koshkarev A V, Ryakhovskii A V and Serebryakov V A 2010 Infrastructure of distributed environment of storage, search and transformation of geospatial data Open Education 5 61-73

[5] Frans J M van der Wel 2005 Spatial data infrastructure for meteorological and climatic data Meteorol. Appl. 12 7-8

[6] Gordov E P, Fazliev A Z, Lykosov V N, Okladnikov I G and Titov A G 2010 Development of a Web-Based Information-Computational Infrastructure for the Siberia Integrated Regional Study Environmental Change in Siberia. Advances in Global Change Research ed H Balzter (Dordrecht: Springer) pp 233-52

[7] Ataeva O M, Kuznetsov K A, Serebryakov V A and Filippov V I 2010 Portal for integration of spatial data “GeoMeta” (Moscow: CC RAS)

[8] Gordov E P, Okladnikov I G, Titov A G 2011 Application of web mapping technologies for development of information-computational systems for georeferenced data analysis Vestnik NSU (Series: Information technologies) 9(4) 94-102

[9] Okladnikov I G, Gordov E P, Ryazanova A A and Titov A G 2019 Elements of software package for “cloud” analysis of the environment and climate changes IOP Conf. Series: Earth and Environmental Science 386 012037 DOI: 10.1088/1755-1315/386/1/012037

[10] Okladnikov I G, Gordov E P and Titov A G 2016 Development of climate data storage and processing model IOP Conf. Series: Earth and Environmental Science 48 012030 DOI: 10.1088/1755-1315/48/1/012030