Research Article

Web-Based Data Integration and Interoperability for a Massive Spatial-Temporal Dataset of the Heihe River Basin EScience Framework

Qingchun Guo,^{1,2,3} Yaonan Zhang,⁴ Zhenfang He,^{3,4} and Yufang Min⁴

¹State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, Shaanxi 710061, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³Liaocheng University, Liaocheng, Shandong 252059, China

⁴Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou 730000, China

Correspondence should be addressed to Yaonan Zhang; yaonan@lzb.ac.cn

Received 7 October 2014; Revised 25 January 2015; Accepted 3 February 2015

Academic Editor: Yongqiang Zhang

Copyright © 2015 Qingchun Guo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To solve the messy problem of data types and form a unified data-processing solution, data in the Heihe River Basin were first classified into five types to integrate them and achieve unified management of data and metadata, preventing the loss of metadata, in the data model of eScience framework. Considering many of the challenges that exist in the construction of the online spatial-temporal data integration and interoperability eScience platform, we used the open data interfaces and standards such as the Common DataModel (CDM) interface, common scientific data modelling (i.e., NetCDF, GRIB, and HDF), and Open Geospatial Consortium (OGC) standards. Through the eScience platform, we also provided online data processing tools by collecting free tools (e.g., NetCDF tool, quality control tool). This eScience platform enables researchers to make full use of scientific research information and results and facilitates collaboration, especially between the GIS community and other members of the earth science community, with the purpose of establishing an online platform of uniform spatial data from the Heihe River Basin via common scientific data modelling.

1. Introduction

Modern geoscience often requires massive datasets and a huge amount of computation for numerical simulation and data handling, and data needs increase daily [1–4]. In earth and environmental science, data management is becoming more challenging [5–7]. Volumes of geographic data have been collected with modern acquisition techniques such as global positioning systems (GPS), remote sensing, wireless sensor, and surveys [8–12]. The increase in data volume has led to more distributed archiving, and it is consequently more difficult to analyse data at a single location and store it locally [13, 14]. “Big Data” has become a ubiquitous new term for researchers. It concerns not only the amount of data but also timeliness (velocity), variety, and veracity. The eScience environment makes full use of people, data,

and computing resources. Software enables convenient data applications and saves manpower and resources [15]. In addition to researchers, governments and private industries also have enormous interest in both collecting spatial data and using massive dataset resources for various application needs, especially Web-based needs [16–19]. The eScience environment supports the combination environment mentioned above and provides online workflow including integration, access, analysis, visualization, and quality control for various data sources and formats [18, 20, 21]. Compared to traditional research methods, eScience applications enable researchers to make full use of scientific research and facilitate international collaboration [22, 23]. Chen et al. proposed a geoprocessing eScience workflow model to integrate logical and physical processes into a composite process chain for sensor observations [24]. A case study on e-Cadastrals

was proposed to analyse the results of a survey to several European Cadastral Agencies and estimate the benefits of spatial data infrastructure [25]. The challenge of constructing a suitable data model for basin-scale research is real and requires improved data interoperability, developing better algorithms and good case studies.

Studies in the Heihe River Basin have examined the weather, ecology, hydrology, and water resources of cold and arid regions. With the development of remote sensing and wireless sensor network technologies, the amount of data that needs acquisition, storage, processing, and transmission has multiplied [26, 27]. Data have been and continue to be accumulated for the study of the Heihe River Basin, becoming the basis for forecasting and decision analysis [28]. As the data from the Heihe River Basin have become more diverse in time period and format, more complex, dynamic, and high-dimensional, and more metadata than ever before, it is more difficult to analyse and effectively visualize [29, 30]. Therefore, constructing an eScience context has been recognised as an urgent need for the unified management of data and metadata and unifying data formats from different disciplines. The eScience context provides the gamut of spatial data through Web services with both efficient data processing algorithms and effective visualization approaches for data applications via the THREDDS (Thematic Real-time Environmental Distributed Data Services) data server (TSD), OGC (Web Coverage Servers) WCS/WMS (Web Map Service), NcML-GML (NetCDF Markup Language-Geography Markup Language), and object-oriented components technology [9, 10, 12, 31, 32]. The internet-based platform provides a convenient way to achieve spatial-temporal data mining, integration, analysis, and distribution and can help researchers to make full use of existing data resources. Maximizing existing resources can reduce the duplication of work and charges for data acquisition and collection, which enhances collaboration, especially between the GIS community and the rest of the earth science community.

The research issues that need to be solved for online-integrated data management of the Heihe River Basin are as follows. First, when constructing an online data integration and interoperability platform, the data contributed by different users have different formats, unprecedentedly large size, unexpedited metadata, high dimensionality, and heterogeneous data sources. These data include remote sensing data with the formats such as HDF5, JPGE, and TIFF, raster data, radar data that are collected and processed by special software, climate observations in GRIB, NetCDF, and ASCII formats, shapefiles, and free-text files (e.g., TXT, word, and CSV). Multiple service interfaces and algorithms for data integration and interoperability are needed by data in different formats, which increases the difficulty of data sharing and interoperability. The data's computational algorithms are directed at various time scales and research objectives, for example, using daily versus monthly averages for forecasts of short-term weather and large-scale climate changes. Second, seamless management is difficult because the data has been separated from the metadata. In addition, it is difficult to download data on a specified period from a dataset, to transmit data or share various fields and software systems.

Metadata must be effectively archived. There is an urgent need for the design of a user-friendly interface to facilitate interoperability, geocomputation, and geovisualisation. Third, unification of the heterogeneity of existing protocols and standards in service-oriented architecture is also needed, such as those applied to establish GIS standards. Finally, experts from different disciplines are familiar with their specific data format and processing software. They encounter problems in converting one format to another or losing information in converting the data formats.

To address these issues, spatial and temporal data from the basin in an eScience context was classified to five categories according to the data source. The data model is a key factor of data sharing. Selecting the data model to integrate spatial-temporal data is very important for the efficient and seamless management in basin eScience. The model determines whether we can efficiently describe the state or evolution of a geographical entity and effectively solve various problems related to time. Considering the compatibility of data, it is hard to achieve a single, completely uniform data model, so the goal may be a few data models. The Heihe River Basin eScience platform selected the Network Common Data Form (NetCDF) (<http://www.unidata.ucar.edu/software/netcdf/>), Hierarchical Data Format (HDF) (<http://www.hdfgroup.org/projects/>), and GRIB in Binary (GRIB) (<http://www.wmo.int/pages/prog/www/>) to integrate and interoperate the data according to requirements analysis. The platform integrates multisource and differently formatted data into a few data models. It can solve the contradiction between the high performance computing and the low efficiency of the processing software, creating management standards for massive spatial-temporal datasets and metadata. The new organization mode and collaborative environment can unite disciplines, regions, and time scales and achieve the complete value chain of the data integration, acquisition, transport, storage, processing, application, and service on this eScience platform via Web services [33]. The eScience environment framework is convenient for cooperation between experts in various fields and simplifies data postprocessing analysis and data retrieval. It can be accessed openly and freely by everyone. The primary focus is to improve data and application interoperability via data models, interfaces, standards, and Web services.

The emphasis of this paper is on the process and methods of unifying data from the Heihe River Basin in NetCDF and on introducing the modules of online data integration and interoperability.

2. Study Areas

The Heihe River is the second largest inland river in China. It is 821 km long and originates from Qilianshan Mountain. The Heihe River Basin is a typical large inland river basin that covers an area of approximately 130,000 km² in the arid zone of northwestern China. It is located in the middle section of the Hexi Corridor of Gansu Province, which is composed of upper, middle, and lower reaches. Its upper reaches originate in the Qilian Mountains, where the headwater streams form

FIGURE 1: The study area map of Heihe River Basin.

strong drainages, and its lower reaches end in the desert Inner Mongolia. The middle reaches are primarily oases surrounded by the Gobi Desert, and the landscape includes heterogeneously distributed farmland, forest, and residential areas [34]. The study area is shown in Figure 1.

The Heihe River Basin, as a typical study area for earth science, has been the object of recent research on weather, climate and remote sensing, ecology, and hydrology, which frequently require analysing hundreds of thousands of variables. The data processing produces large numbers of results, explanations, and other information in various formats and files. In the Heihe River Basin, long-term monitoring, testing, and research are the main sources of data and the important

basis of earth system science research. Managing and processing the long-term monitoring data are some of the important tasks for basin research. Therefore, in the eScience context of the basin, issues such as the distribution, heterogeneity, and volume of data need to be addressed during the design and implementation of new data-oriented infrastructures, services, standards, and systems.

3. Methods

3.1. Common Data Model and Spatial-Temporal Data Model for EScience Context. The chosen spatial-temporal data model has to address platform compliance with model

FIGURE 2: The data integration and interoperability interfaces of spatial-temporal data.

interfaces and several protocols. It must also fit the service-oriented architecture through open and free interfaces because the eScience context is a service-oriented distributed environment that allows scientist to share distributed data resources and data processing components. Unidata's Common Data Model (CDM) has a unified interface to access NetCDF, HDF, and GRIB to build a bridge of interoperability for different data models. The CDM interface is applied to form the workflow of data integration, visualization, distribution, and analysis in the eScience framework. The data integration and interoperability interfaces of spatial-temporal data are shown in Figure 2.

3.2. Data Classification and Archive in the Heihe River Basin EScience Context. In the Heihe River Basin, to generate the NetCDF format from heterogeneous multisource data (e.g., geoTiff, ASCII, free-text, shapefile, and grid), it is important to classify and archive the Heihe River Basin data. According to the characteristics of the NetCDF and the data formats, these formats were classified into five types including station point, point, grid, image, and radial to address messy problems in data formats (Table 1). Station types describe the time series observation data that remain fixed in space and have an exact set of specified spatial coordinates (e.g., hydrologic station data and weather station data). Unlike station points, points can change their location and have no relationship with each other. They can record data in text formats. Grid includes structured and unstructured data. A number of traditional data files can be integrated and archived into a single NetCDF file (e.g., free-text data) via these types, as shown in Table 1. The NetCDF structure provides a powerful mechanism for dealing with complicated scientific workflows and resolves "messy" issues, such as traditional multiple files and heterogeneous data.

TABLE 1: Classification and archive of various data types.

Data type	Definition
Station point	Station observations/time series
Point	Unconnected point data
Grid	Structured grid and unstructured grid
Radial	Radar observation data
Image	Satellite observation data

3.3. The Design Flow and Function of Data Integration and Interoperability. Spatial-temporal data integration and interoperability platform were constructed based on the Web service of B/S architecture to enhance sharing and interoperability. The design flows of conversion and interoperability for data in the framework are shown in Figure 3. The CDM interface was used to access different scientific data including NetCDF, HDF, and GRIB. In addition to the CDM, we also used two other technologies (XML schemas and object-oriented components) to realize the data integration and interoperability framework. TSD data server, NCML-GML, and OGC WCS/WMS were achieved mainly through the XML schemas. The XML schemas resolve the problems in remote access to data and facilitate the interoperability of GIS and other data via Web services. Object-oriented component technology mainly addressed the issues that different domains develop different data processing algorithms in various computer languages (e.g., C, MATLAB, and Fortran). We needed to provide object-oriented components technology to construct components library via collecting data processing program. In addition, we can access these three data formats through the CDM interfaces via OPeNDAP or HTTP protocols. NcML-GML and the WCS/WMS achieved the Web service of GIS data encoded in NetCDF.

FIGURE 3: The design flow of data format conversion and interoperability in the framework.

FIGURE 4: The primary functions of spatial-temporal data integration and interoperability in the eScience framework.

The framework is convenient for the standard management of large-scale spatial-temporal data and facilitates cooperative research across disciplines.

Figure 4 shows the main functions of the spatial-temporal data in the eScience platform. The platform provides the services including NetCDF metadata extraction, NetCDF

dataset operation, data format conversion, data visualization, and data access. If the CDM interface cannot achieve special data processing, appropriate components in component libraries were selected according to the requirements of the researchers. If a component does not exist, a new component can be designed and added to the component libraries.

FIGURE 5: The key technology of the eScience interoperability platform of GIS community and other communities.

The metadata extraction services in NetCDF extract dataset attribute information including department, author, and coordinate system and attribute names. They also can be extracted via NcML. Dataset operation services include the basic operations such as appending spatial-temporal data, renaming, modifying and deleting attributes, variables, and dimensions. Data format conversion services convert the formats of point data, remote sensing data, radar data, and the grid data to NetCDF and convert NetCDF to raster and vector data format through third party software or an online tools library such as GIS software to promote the sharing and interoperability of data. The visualization services of NetCDF offer dynamic visualization of long series spatial-temporal data to achieve convenient comparisons and selection of the data in a study area via WCS/WMS or online tools. NetCDF access services acquire data online through THREDDS Data Server and existing protocols (e.g., OPeNDAP and ADDE). When the users are not interested in all the data, they can extract sections of data for certain variables at certain times or in certain regions from these datasets via the Web. Analysis of NetCDF realises arithmetic operations through the browser on the NetCDF datasets such as computing averages.

3.4. NetCDF Data Interoperability with TSD, OGC WCS/WFS, and NcML-GML Technology in EScience Framework. Web technology provides support for eScience development through innovative technologies and protocols, the message format and algorithms, and creative services such as Wikis, TSD, and WCS [35, 36]. The eScience framework is a service-oriented interoperability platform for large spatial-temporal datasets. The key technologies, THREDDS Data Server, OGC WCS/WFS, and NcML-GML, facilitate the interoperability of the scientists in different disciplinary areas, as shown in Figure 5. The THREDDS Data Server (TDS) is the Web server for scientific data and lists the datasets in a THREDDS catalogue, which is simply an XML file offering available datasets and services. Through the TDS, users can obtain the name and location of datasets from different institutions and then access the datasets through OPeNDAP, ADDE, or NetCDF/HTTP protocols [37]. TDS can serve any dataset that the NetCDF-Java library can read (e.g., NetCDF-3, NetCDF-4, HDF-4, HDF-5, HDF-EOS, GRIB-1, and GRIB-2). It can also provide data access (subset) services (e.g., OGC

WMS and WCS), data collection services (e.g., aggregation), and metadata services (e.g., NcML). Researchers can obtain select parts of these datasets via Web browser (e.g., certain variables at certain times or regions).

An NcML document is an XML document describing the content and structure of the data stored in a NetCDF file and represents a generic NetCDF dataset (<http://www.unidata.ucar.edu/software/netcdf/ncml/>) [38]. In our eScience context, it can be used as a “public interface” for spatial-temporal online data, conforming to the NetCDF data model. NcML describes the metadata of the NetCDF data and does not encode the data. The purpose of NcML is to define and redefine NetCDF file. The NcML has the function as follows:

- (i) Metadata to be added, deleted, and changed.
- (ii) Variables to be renamed, added, deleted, and restructured.
- (iii) Aggregated data from multiple CDM files (e.g., Union, JoinNew, and JoinExisting).

We take average monthly temperature NetCDF data of the Heihe River Basin as an example. The data are in a CF-complaint NetCDF format, and the visualisation is shown via online tools in Figure 6. The NcML of the data is seen in Appendix.

The aggregation function of the NcML is useful for time series data combinations. Multiple time series NetCDF data can be aggregated into a single, logical dataset with several types of aggregation including Union, JoinExisting, and JoinNew. To facilitate interdisciplinary work between earth sciences and the GIS communities, NcML-GML is developed to use NetCDF datasets in GIS software, providing them with all the necessary metadata in the form of GML (Geography Markup Language) extensions to NcML. GML is written in XML schema for the storage of geographic information with the GIS community semantics. NcML-GML supports referencing information of spatial-temporal data and realizing the function of the platform that describes the coverage data derived from NetCDF data file. NcML-GML and WCS/WMS can map the NetCDF model into the model of GIS and facilitate the interoperability between these two models and different scientists. Through the technology above, users can obtain metadata and the slices of data they require from remote NetCDF files on a Web server accessible directory.

FIGURE 6: The key integrated NetCDF data technology based on Web via object-oriented method with MATLAB and Java mixed solution.

FIGURE 7: A simplified sequence diagram for the NetCDF integration.

3.5. *The Key Object-Oriented Component Technologies of Spatial-Temporal Online Data Integration and Interoperation.* To improve the calculation speed and convenient visualization of the data, we selected the mixed solution of MATLAB and Java to complete data integration based on the Web as one example of the object-oriented method to build components. Quality control components are also built by the object-oriented method. Figure 6 shows the technical framework.

The MATLAB and Java mixed solution is to complete the custom framework and interface via Java and Web technology for special data processing and computation through MATLAB with powerful matrix and numerical analysis capabilities. The mixed solution can solve the problems posed by MATLAB's poor interactivity and the fact that MATLAB programs cannot run outside the environment. In addition, the characteristics of Java language such as crossing platforms and exception handling, multithreading, and stable and fast operation could also be utilized.

Figure 6 shows the workflow of mixed solutions technology; first, the MATLAB code completes core algorithms of NetCDF integration and generates the m files. Second, the m files are then transformed into a component which will interact with server-side through the Java language without the MATLAB environment through MATLAB compiler and MATLAB builder JA. Finally, an encapsulation function

would be called to achieve the core calculation of MATLAB and online computing on the Web through the Java program with the MATLAB dynamic library.

Figure 7 depicts an example of the processes within a scientific workflow. The NetCDF integration process contains an integrated chain invoking first the data processing component and then the integration process. After creating new NetCDF, the add data component continues to increase the variable to NetCDF, extending time dimension or adding other variables. CDM is available as free software to process NetCDF and is actually called several times as part of different scientific workflows.

3.6. *Online Spatial-Temporal Data Quality Control Methods on ESience Platform.* Spatial data quality has been recognised as an important issue in GIS. However, online spatial-temporal data quality control has received little attention from data processing. Irregularities cause unreliable results because any initial spatial data error can be propagated through the spatial data processing. Based on glaciers, permafrost, deserts, and atmospheric, ecological, environmental, hydrological, and other elements, monitoring systems established in Heihe River Basin realize automatic data transmission and connect with the basin eScience context. Before data integration and analysis, we achieve real-time

data detection and calibration to ensure data quality control on the eScience platform.

In this study, we mainly focus on the online data quality control of outliers in the spatial-temporal data before conversion to NetCDF. This is very important for data quality control, especially for data transmission in wireless sensor networks before data formats are converted to NetCDF files. We provided quality control components for Web services in the components library. In addition, we will continue to enrich our components library to facilitate data processing and data quality control.

According to the data request, the basin eScience platform provides online outlier detection methods, including extreme test method, 3δ test method, Dixon's test method, and Grubbs' test method. The platform will provide convenient detection of abnormal data points, which will help users to understand the data change rules over time and the intrinsic relationships among the data. Based on the physical characteristics and statistical experience of the various elements, the extreme test method gives the maxima and minima values of the real-time data. For the 3δ test method, according to the theory of error, the random error δ obeys a normal distribution. As the standard differential is generally unknown, S counted with a Bezier formula is typically used instead of δ . In formula (1), $\bar{x}$ is the true value, and x_i is observation data. Consider

$$\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i, \quad i = 1, 2, 3, \dots, n, \quad (1)$$

$$S = \sqrt{\frac{\sum_{i=1}^n v_i^2}{N-1}}, \quad i = 1, 2, 3, \dots, n. \quad (2)$$

For an observation data point x_i , if its residuals v_i meet $v_i = |x_i - \bar{x}| > 3S$, $i = 1, 2, 3, \dots, n$, x_i is marked as outlying data. For Dixon's test method, suppose the overall observation data are normally distributed. In the sample $x_1, x_2, x_3, \dots, x_n$, n is the number of the samples, and the observation data are arranged in order of size $x'_1, x'_2, x'_3, \dots, x'_n$. Depending on the number of samples, we select a different formula, such as formula (3). We marked $r_{10}, r'_{10}, r_{11}, r'_{11}, r_{21}, r'_{21}, r_{22}$, and r'_{22} as r_{ij} and r'_{ij} . To determine the significance level α , look up the threshold $D(\alpha, n)$ in the threshold table of the Dixon test. If $r_{ij} > r'_{ij}$, $r_{ij} > D(\alpha, n)$, then x'_n is judged as an abnormal value. If $r_{ij} < r'_{ij}$, $r_{ij} > D(\alpha, n)$, then x'_1 is judged as abnormal value. Otherwise, there are no abnormal values. Dixon's test method is suitable for real-time data quality control. Consider

$$r_{10} = \frac{x'_n - x'_{n-1}}{x'_n - x'_1}, \quad r'_{10} = \frac{x'_2 - x'_1}{x'_n - x'_1} \quad (3 \leq n \leq 7),$$

$$r_{11} = \frac{x'_n - x'_{n-1}}{x'_n - x'_2}, \quad r'_{11} = \frac{x'_2 - x'_1}{x'_{n-1} - x'_1} \quad (8 \leq n \leq 10),$$

$$r_{21} = \frac{x'_n - x'_{n-2}}{x'_n - x'_2}, \quad r'_{21} = \frac{x'_3 - x'_1}{x'_{n-1} - x'_1} \quad (11 \leq n \leq 13),$$

$$r_{22} = \frac{x'_n - x'_{n-2}}{x'_n - x'_3}, \quad r'_{22} = \frac{x'_3 - x'_1}{x'_{n-2} - x'_1} \quad (14 \leq n \leq 30). \quad (3)$$

For Grubbs' test method, we assumed normal independently measured samples $x_1, x_2, x_3, \dots, x_n$, where n is the number of the samples, the residual absolute value of the data is x_i , and $\bar{x}$ is the maximum. $\bar{x}$ is the average of the samples. Then, we constructed the statistic $(x_i - \bar{x})/S$, with the formula for S being given by formula (2). At the selected significance level α , we obtain the threshold $G(\alpha, n)$ by formula (4). α is usually a value of 0.05 or 0.01. Consider

$$P\left(\frac{x_i - \bar{x}}{S} \geq G(\alpha, n)\right) = \alpha. \quad (4)$$

If $|x_i - \bar{x}| \geq G(\alpha, n)S$, then x_i is abnormal value and $G(\alpha, n)$ can be given by the lookup table.

In the eScience platform, we also collect a range of open and free data processing tools and provide them online, such as visualization tools to facilitate collaboration like the NetCDF tools (<http://www.unidata.ucar.edu/downloads>).

4. Results

4.1. The Case Study of Spatial-Temporal Data Integration and Interoperability on EScience Platform. In this paper, observation data from the Mafengou subbasin wireless sensor transmission site in the Heihe River Basin is used as a case study for abnormal data quality control. The temperature data are transmitted every 30 minutes with a total of 73 records. Figures 8(a), 8(b), 8(c), and 8(d) compare the dataset before and after outlier quality control with four methods, the extreme test, 3δ test, Dixon's test, and Grubbs' test, respectively. Figure 8(a) shows that three outliers were found by the extreme test method. Figure 8(b) shows that the 3δ test method found the obvious abnormal data. Figure 8(c) shows that five outliers were found by Dixon's test method. Grubbs' test method is the best, finding seven outliers, as shown in Figure 8(d).

Figure 9 shows the NetCDF tools display of the grid map of the temperature for the Heihe River Basin. The NetCDF tools can also browse remote data model datasets (e.g., NetCDF, GRIB, and HDF) via the TSD data server. An online tools library facilitates data processing and the interoperation of the eScience context using tools with which researchers are familiar.

4.2. Raster Data Integration. To demonstrate the data integration, we took average monthly temperature raster data of the Heihe River Basin as an example and integration components of the component library via Web services as shown in Figure 10. The tool mainly achieved integration and aggregation of the data. First, it converted grid data to ASCII and then integrated the data online as NetCDF to complete the long series data integration. In this example, the grid size is 500 meters, the line number is 899, the column number is 1041, the coordinate x of the left bottom corner is 666083.7 meters, and coordinate y is 4008999.5 meters. These parameters and coordinate system were needed on the webpage. When generating the m function files, we choose grid size, rank number of the grid, and the left bottom corner coordinates as the function's parameters for NetCDF

FIGURE 8: (a), (b), (c), and (d) compare a dataset before and after the outliers quality control with the methods extreme, 3δ , Dixon's, and Grubbs' tests, respectively.

data integration, time as an unlimited dimensional variable parameter, and the coordinate system as the metadata according to the CF. The components can also add variables to the NetCDF via aggregation. Figure 10 shows a visualization map of one-month data from the NetCDF datasets of the Heihe river upstream temperature data in November, 2005.

4.3. The Integration of Wireless Sensor Network Station Data. Station point data from a wireless sensor transmission site in the Heihe River Basin was used as a case study for integrating point data. The data were transmitted every 15 minutes and examined via the quality control components mentioned in Section 4.4 before conversion to NetCDF. The soil humidity data of the observation data were defined in NetCDF. The integration of NetCDF is divided into two parts: the first describes the information of station number, latitude, longitude, and altitude, and the other describes the measurements such as meteorological and hydrological elements. The visualization map of the NetCDF dataset for soil humidity of the Mafengou subbasin wireless sensor station in October is shown in Figure 11. The lines named soil humidity1 and soil humidity2 were the data of NetCDF from different time, and the soil humidity10 lines were aggregations

of two NetCDF datasets with one of soil humidity1 NetCDF and soil humidity2 NetCDF.

For the time series data of the observation station, variables aggregated can integrate different time series NetCDF data into one NetCDF database via NcML files in order to add time series data. The following codes are a program example aggregating different time series soil humidity data, shown in Figure 8.

```
<NetCDF xmlns="http://www.unidata.ucar.edu/
namespaces/NetCDF/ncml-2.2">

  <aggregation    dimName="time"    type=
"joinExisting">
    <NetCDF location="humidity1.nc"/>
    <NetCDF location="humidity2.nc"/>
  </aggregation>

</NetCDF>
```

4.4. The Integration Examples of Image and Radial Data. In integrating the array structure of image data into the NetCDF data model, we consider the following variables: line

FIGURE 9: Example using freely available software (NetCDF (4.2) Tools) from the online library, which can process and visualize NetCDF and NcML files and remotely access NetCDF, GRIB, and HDF files via TSD.

(the number of satellites scan lines), elem (the element point per scan line number), and the band (the band number of observations). The geographical location is described by latitude and longitude, and the observation values of each band are defined as the main data variable of NetCDF.

In integrating the radial data (e.g., radar data), we mainly consider the radial data to be located by azimuth, elevation, angle, and orientation. A scan record is made up of a number of adjacent radial data records. The main variables of the NetCDF data model in the program include gate (the number of the pulse of a radial data record), radial (the number of radial data of recording a scan), scan (scan number), distance (the distance of pulse), time (the time of the data record), eleva (elevation angular), and azim (azimuth). In this paper, NetCDF is used as a case study of technology to implement the framework.

5. Conclusion

The eScience platform provides effective interfacing and interacting strategies for data processing, sharing scientific research and decision support with the general public. It is an important method to solve the common problem of information islands by offering public Web data access. Online integration of heterogeneous data sources provided a uniform interface for users to access, analyse, and seamlessly manage the data and give a standard format for data processing programs. The problem of messy data formats was resolved by the eScience platform. It improved the ability of the users to investigate complex phenomena such as climate change, hydrological change, and soil dynamics. Finally, the eScience environment will be gradually used to support decision-making in the Heihe River Basin.

FIGURE 10: The raster data were integrated to NetCDF through the components library based on Web service.

In further research, we will examine HDF and GRIB data processing methods and gradually establish a single online spatial data process in the eScience context for the Heihe River Basin, developing a suite of efficient parallel algorithms and constructing a geoscience data-supporting library suitable for high performance parallel computation.

This study constructed the Heihe River Basin data integration and interoperability eScience context, which integrated the spatial-temporal data and different formats into NetCDF data models. The framework was constructed based on HDF, NetCDF, and GRIB for uniform management of the spatial-temporal data and metadata, which were long-term, massive, and multidimensional. In addition, we can access and analyse these data formats (e.g., HDF, NetCDF, and GRIB) through the CDM interface, which provided a convenient method for data mining, integration, and the analysis of spatial-temporal data. The framework can establish the eScience cooperative work environment and support the efficient application of the data via Web services. It is especially beneficial to the GIS and the earth science communities for cooperative communication via eScience platform.

The data integration and interoperability eScience platform of the combination of technological solutions can achieve the following goals: (i) the integration of real-time and historical data; (ii) solving the data application

FIGURE 11: Aggregated different time series soil humidity data with one NetCDF.

problems cross fields, areas, and disciplines; (iii) conveniently accessing and analysing the data resources from different institutions; and (iv) addressing issues about heterogeneous existing standards and existing protocols of Web data access. The combination of solutions chosen could be interesting for achieving the goals, but one kind of technology cannot achieve them.

Through the platform, to generate the NetCDF format from heterogeneous multisource data (e.g., geoTiff, ASCII,

free-text, shapefile, and grid), it is different from other data share platforms and it is important to manage and share scientific data. And Heihe River Basin eScience platform is superior to other data share platforms in sophisticated analysis algorithm workflows, access to powerful computational resources, analysis, and interactive visualization interface. Our continuing work will provide scientists access to a wide range of datasets, algorithm applications, access to computational resources, services, and support.

Appendix

The NcML of the data is as follows:

```
<?Xml version="1.0" encoding="UTF-8"?>

<NetCDFxmlns=http://www.unidata.ucar.edu/
namespaces/NetCDF/ncml-2.2
location="E:/200511.nc">
<dimension name="y" length="1041"/>
<dimension name="x" length="899"/>
<attribute name="Conventions" value="CF-
1.0"/>
<attribute name="Source_Software" value=
"ESRI ArcGIS"/>
  <attribute name="History" value=
  "Translated to CF-1.0 Conventions by
  NetCDF-Java CDM (NetCDFCFWriter)"/>
<variable name="r200511" shape="y x" type=
"float">
  <attribute name="long_name" value=
  "r200511"/>
  <attribute name="esri_pe_string" value=
  "PROJCS"/>
  <attribute name="coordinates" value=
  "y x"/>
  <attribute name="grid_mapping" value=
  "albers_conical_equal_area"/>
  <attribute name="units" value="K"/>
  <attribute name="missing_value" type=
  "float" value="-3.4028235E38"/>
</variable>
<variable name="y" shape="y" type="double">
  <attribute name="units" value="km"/>
  <attribute name="long_name" value="y
coordinate of projection"/>
  <attribute name="standard_name" value=
  "projection_y_coordinate"/>
</variable>
<variable name="x" shape="x" type="double">
  <attribute name="units" value="km"/>
  <attribute name="long_name" value="x
coordinate of projection"/>
  <attribute name="standard_name" value=
  "projection_x_coordinate"/>
</variable>
<variable name="albers_conical_equal_area"
shape="" type="int">
  <attribute name="grid_mapping_name"
value="albers_conical_equal_area"/>
  <attribute name="longitude_of_central_
meridian" type="double" value="105.0"/>
  <attribute
name="latitude_of_projection_origin"
type="double" value="0.0"/>
```

```
<attribute name="false_easting" type=
"double" value="0.0"/>
<attribute name="false_northing" type=
"double" value="0.0"/>
<attribute name="standard_parallel" type=
"double" value="25.0 47.0"/>
<attribute
name="_CoordinateTransformType"
value="Projection"/>
<attribute name="_CoordinateAxisTypes"
value="GeoX GeoY"/>

</variable>
```

```
</NetCDF>
```

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (91125005/D011004), the National Natural Science Foundation of China (41290255), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB03020601), the CAS/SAFEA International Partnership Program for Creative Research Teams (KZZD-EW-TZ-03), and China Data Sharing Infrastructure of Earth System Science.

References

- [1] K. V. Knyazkov, S. V. Kovalchuk, T. N. Tchurov, S. V. Maryin, and A. V. Boukhanovsky, "CLAVIRE: e-Science infrastructure for data-driven computing," *Journal of Computational Science*, vol. 3, no. 6, pp. 504–510, 2012.
- [2] M. Lutz, J. Sprado, E. Klien, C. Schubert, and I. Christ, "Overcoming semantic heterogeneity in spatial data infrastructures," *Computers & Geosciences*, vol. 35, no. 4, pp. 739–752, 2009.
- [3] NRC, *Grand Challenges in Environmental Sciences*, National Academy of Sciences, Washington, DC, USA, 2001.
- [4] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM SIGMOD Record*, vol. 34, no. 3, pp. 31–36, 2005.
- [5] J. L. Goodall, J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment, and I. Zaslavsky, "A first approach to web services for the National Water Information System," *Environmental Modelling & Software*, vol. 23, no. 4, pp. 404–411, 2008.
- [6] J. S. Horsburgh, D. G. Tarboton, M. Piasecki et al., "An integrated system for publishing environmental observations data," *Environmental Modelling & Software*, vol. 24, no. 8, pp. 879–888, 2009.
- [7] S. Fiore and G. Aloisio, "Data management for eScience," *Future Generation Computer Systems*, vol. 27, no. 3, pp. 290–291, 2011.
- [8] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.
- [9] W. Li, C. Yang, D. Nebert et al., "Semantic-based web service discovery and chaining for building an Arctic spatial data

- infrastructure," *Computers & Geosciences*, vol. 37, no. 11, pp. 1752–1762, 2011.
- [10] L. Li, J. Guardiola, and Y. Liu, "Qualitative spatial representation and reasoning for data integration of ocean observing systems," *Computers, Environment and Urban Systems*, vol. 35, no. 6, pp. 474–484, 2011.
 - [11] J. Mennis and D. S. Guo, "Spatial data mining and geographic knowledge discovery—an introduction," *Computers, Environment and Urban Systems*, vol. 33, no. 6, pp. 403–408, 2009.
 - [12] G. W. Johnson, A. G. Gaylord, J. C. Franco et al., "Development of the Arctic Research Mapping Application (ARMAP): interoperability challenges and solutions," *Computers & Geosciences*, vol. 37, no. 11, pp. 1735–1742, 2011.
 - [13] T. Foerster, L. Lehto, T. Sarjakoski, L. T. Sarjakoski, and J. Stoter, "Map generalization and schema transformation of geospatial data combined in a Web Service context," *Computers, Environment and Urban Systems*, vol. 34, no. 1, pp. 79–88, 2010.
 - [14] L. S. R. Froude, "Storm tracking with remote data and distributed computing," *Computers & Geosciences*, vol. 34, no. 11, pp. 1621–1630, 2008.
 - [15] K. V. Knyazkov, S. V. Kovalchuk, T. N. Tchurov, S. V. Maryin, and A. V. Boukhanovsky, "CLAVIRE: e-Science infrastructure for data-driven computing," *Journal of Computational Science*, vol. 3, no. 6, pp. 504–510, 2012.
 - [16] T. Blanke, M. Hedges, and S. Dunn, "Arts and humanities e-science—current practices and future challenges," *Future Generation Computer Systems*, vol. 25, no. 4, pp. 474–480, 2009.
 - [17] L. Díaz, C. Granell, M. Gould, and J. Huerta, "Managing user-generated information in geospatial cyberinfrastructures," *Future Generation Computer Systems*, vol. 27, no. 3, pp. 304–314, 2011.
 - [18] G. Giuliani, N. Ray, and A. Lehmann, "Grid-enabled Spatial Data Infrastructure for environmental sciences: challenges and opportunities," *Future Generation Computer Systems*, vol. 27, no. 3, pp. 292–303, 2011.
 - [19] M. G. Tait, "Implementing geoportals: applications of distributed GIS," *Computers, Environment and Urban Systems*, vol. 29, no. 1, pp. 33–47, 2005.
 - [20] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-Science: an overview of workflow system features and capabilities," *Future Generation Computer Systems*, vol. 25, no. 5, pp. 528–540, 2009.
 - [21] G. Folino, A. Forestiero, G. Papuzzo, and G. Spezzano, "A grid portal for solving geoscience problems using distributed knowledge discovery services," *Future Generation Computer Systems*, vol. 26, no. 1, pp. 87–96, 2010.
 - [22] D. D. Pennington, "Collaborative, cross-disciplinary learning and co-emergent innovation in eScience teams," *Earth Science Informatics*, vol. 4, no. 2, pp. 55–68, 2011.
 - [23] J. Rowley, "E-Government stakeholders—who are they and what do they want?" *International Journal of Information Management*, vol. 31, no. 1, pp. 53–62, 2011.
 - [24] N. C. Chen, C. L. Hu, Y. Chen, C. Wang, and J. Y. Gong, "Using SensorML to construct a geoprocessing e-Science workflow model under a sensor web environment," *Computers and Geosciences*, vol. 47, pp. 119–129, 2012.
 - [25] M. T. Borzacchiello and M. Craglia, "Estimating benefits of spatial data infrastructures: a case study on e-Cadastrs," *Computers, Environment and Urban Systems*, vol. 41, pp. 276–288, 2013.
 - [26] J. Wu, "The effect of ecological management in the upper reaches of Heihe River," *Acta Ecologica Sinica*, vol. 31, no. 1, pp. 1–7, 2011.
 - [27] J. R. Millan-Almaraz, I. Torres-Pacheco, C. Duarte-Galvan et al., "FPGA-based wireless smart sensor for real-time photosynthesis monitoring," *Computers and Electronics in Agriculture*, vol. 95, pp. 58–69, 2013.
 - [28] Q. Dahe, "China's meteorological science data sharing process and prospect," *Scientific Chinese*, vol. 9, pp. 18–20, 2004.
 - [29] B. Plale, D. A. Gannon, J. Alameda et al., "Active management of scientific data," *IEEE Internet Computing*, vol. 9, no. 1, pp. 27–34, 2005.
 - [30] D. P. Lanter, "A lineage metadata approach to removing redundancy and propagating updates in a GIS database," *Cartography and Geographic Information Science*, vol. 21, no. 2, pp. 91–98, 1994.
 - [31] R. B. Jerard and O. Ryou, "NCML: a data exchange format for internet-based machining," *International Journal of Computer Applications in Technology*, vol. 26, no. 1–2, pp. 75–82, 2006.
 - [32] K. Stock, T. Stojanovic, F. Reitsma et al., "To ontologise or not to ontologise: an information model for a geospatial knowledge infrastructure," *Computers & Geosciences*, vol. 45, pp. 98–108, 2012.
 - [33] A. M. Castronova, J. L. Goodall, and M. M. Elag, "Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard," *Environmental Modelling & Software*, vol. 41, pp. 72–83, 2013.
 - [34] X. Li, L. Lua, W. Yangc, and G. Chenga, "Estimation of evapotranspiration in an arid region by remote sensing—a case study in the middle reaches of the Heihe River Basin," *International Journal of Applied Earth Observation and Geoinformation*, vol. 17, no. 1, pp. 85–93, 2012.
 - [35] G. Fox and M. Pierce, "Grids challenged by a web 2.0 and multicore sandwich," *Concurrency and Computation: Practice and Experience*, vol. 21, no. 3, pp. 265–280, 2009.
 - [36] L. D. Stein, "Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges," *Nature Reviews Genetics*, vol. 9, no. 9, pp. 678–688, 2008.
 - [37] R. P. Signell, S. Carniel, J. Chiggiato, I. Janekovic, J. Pullen, and C. R. Sherwood, "Collaboration tools and techniques for large model datasets," *Journal of Marine Systems*, vol. 69, no. 1–2, pp. 154–161, 2008.
 - [38] Y. B. Li, A. J. Brimicombe, and M. P. Ralphs, "Spatial data quality and sensitivity analysis in GIS and environmental modelling: the case of coastal oil spills," *Computers, Environment and Urban Systems*, vol. 24, no. 2, pp. 95–108, 2000.