Computers, Environment and Urban Systems 61 (2017) 120–128

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Utilizing Cloud Computing to address big geospatial data challenges

Chaowei Yang *, Manzhu Yu, Fei Hu, Yongyao Jiang, Yun Li

NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form 20 October 2016
Accepted 20 October 2016
Available online 2 November 2016

Keywords:
Big Data
Cloud Computing
Spatiotemporal data
Geospatial science
Smart cities

ABSTRACT

Big Data has emerged with new opportunities for research, development, innovation and business. It is characterized by the so-called four Vs: volume, velocity, veracity and variety and may bring significant value through the processing of Big Data. The transformation of Big Data's 4 Vs into the 5th (value) is a grand challenge for processing capacity. Cloud Computing has emerged as a new paradigm to provide computing as a utility service for addressing different processing needs with a) on demand services, b) pooled resources, c) elasticity, d) broad band access and e) measured services. The utility of delivering computing capability fosters a potential solution for the transformation of Big Data's 4 Vs into the 5th (value). This paper investigates how Cloud Computing can be utilized to address Big Data challenges to enable such transformation. We introduce and review four geospatial scientific examples, including climate studies, geospatial knowledge mining, land cover simulation, and dust storm modelling. The method is presented in a tabular framework as a guidance to leverage Cloud Computing for Big Data solutions. It is demostrated throught the four examples that the framework method supports the life cycle of Big Data processing, including management, access, mining analytics, simulation and forecasting. This tabular framework can also be referred as a guidance to develop potential solutions for other big geospatial data challenges and initiatives, such as smart cities.

1. Introduction

Earth observation and model simulation produce tera- to peta- bytes of data daily (Yang, Raskin, Goodchild, and Gahegan, 2010). Non-traditional, geospatial data acquisition methods, such as social media (Romero, Galuba, Asur, and Huberman, 2011), phone conversations (Frias-Martinez, Virseda, Rubio, and Frias-Martinez, 2010) and unmanned aerial vehicles (Einav and Levin, 2013), produce geospatial data at even faster speeds. In addition to the large Volume (Marr, 2015; Hsu, Slagter, and Chung, 2015), geospatial data exist in a Variety of forms and formats for different applications, their accuracy and uncertainty span across a wide range as defined by Veracity, and data are produced in a fast Velocity through real time sensors (Fig. 1). With unprecedented information and knowledge embedded, these big geospatial data can be processed for adding Value to better scientific research, engineering development and business decisions (Lee and Kang, 2015). They are envisioned to provide innovation and advancements to improve our lives and understanding of the Earth systems (Mayer-Schönberger and Cukier, 2013) when transformed from the first four Vs to the last V (value) through advancements in a variety of geospatial domains (Fig. 1).

Such transformations pose grand challenges to data management and access, analytics, mining, system architecture and simulations (Yang, Huang, Li, Liu, and Hu, 2016). For example, the first challenge is how to deal with the Variety and Veracity of Big Data to produce a fused dataset that can be utilized in a single decision support system (Kim, Trimi, and Chung, 2014). Another issue is how to deal with the velocity of Big Data to have scalable and extensible processing power based on the fluctuation of the data feed (Ammn and Irfanuddin, 2013). Supporting on-demand or timely data analytical functionalities also pose significant challenges for creating the Value (Fan and Liu, 2013; Chen and Zhang, 2014; Jagadish et al., 2014).

Cloud Computing has emerged as a new paradigm to provide computing as a utility service with five advantageous characteristics (Fig. 1 bottom two layers): a) rapid and elastic provisioning computing power; b) pooled computing power to better utilize and share resources; c) broadband access for fast communication; d) on demand access for computing as utility services; and e) pay-as-you-go for the parts used without a significant upfront cost like that of traditional computing resources (Yang, Xu, and Nebert, 2013). Service-oriented architecture is adopted in Cloud Computing and enables "everything as a service", including Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) (Mell and Grance, 2011). While redefining the possibilities of geospatial science and Digital Earth (Yang et al., 2013), Cloud Computing engaging Big Data enlightens potential solutions for big geospatial data problems in various geosciences and relevant domains.

However, utilizing Cloud Computing to address Big Data issues is still in its infancy, and it is a daunting task on how the five advantageous

* Corresponding author.
E-mail address: cyang3@gmu.edu (C. Yang).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010

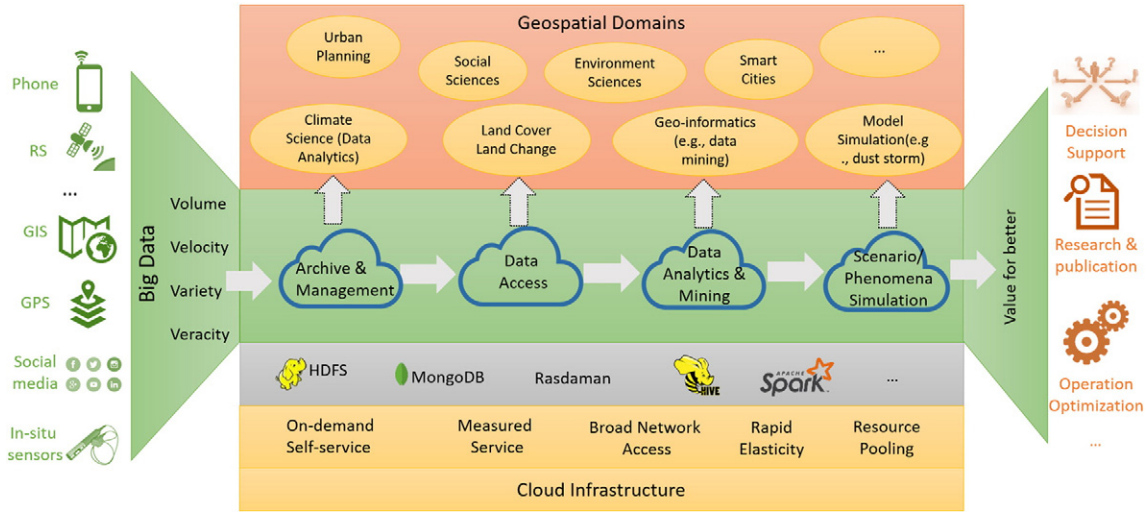


Fig. 1. Cloud Computing provides critical supports to the processing of Big Data to address the 4Vs to obtain Value for better decision support, research, and operations for various geospatial domains.

characteristics can address the first four Vs of Big Data to reach the 5th V (Fig. 1). This paper illustrates how Cloud Computing supports the transformation with four scientific examples including climate studies, knowledge mining, land-use and land cover change analysis, and dust storm simulation. These four examples are highly representative and can be easily adopted to other environmental and urban research fields, such as smart cities (Batty, 2013; Mitton, Papavassiliou, Puliafito, and Trivedi, 2012; Odendaal, 2003). The big geospatial data life cycle (data management and access, analyses/mining, phenomena/scenario simulation) are examined through the four examples and detailed in each example section (Table 1). For example, 2.1 is filled in the intersection cell of on-demand self-service and volume of Table 1. This means that 2.1 details how the volume (of big climate data) are addressed with the on-demand self-service of Cloud Computing.

2. Utilizing Cloud Computing to support climate analytics

The interrelated climate changes, such as greater incidence of heavy downpours and increased riverine flooding, are increasingly compromising urban infrastructure (Rosenzweig, Solecki, Hammer, and Mehrotra, 2011). Meanwhile human activities (e.g. the burning of fossil fuels) heavily impacted the global environment in the past 50 years (Bulkeley and Betsill, 2005). In order to understand climate change and its impacts to environmental and urban issues, the big climate data observed in the past and simulated for the future should be well managed and analyzed. However, both observation and simulation produce Big Data. For example, the next IPCC report will be based on 100 + petabytes of data, and NASA will produce 300 + petabytes of climate data by 2030 (Skytland, 2012). These data differ in format, spatiotemporal resolution, and study objective (Schnase et al., 2014). Big Data can help advance the understanding of climate phenomena and help identify how impacts of climate change on society and ecosystems can be remedied, such as detecting global temperature anomalies and investigating spatiotemporal distribution of extreme weather events, especially over highly populated regions (such as urban areas, Das and Parthasarathy, 2009; Debbage and Shepherd, 2015).

There are several challenges in the use of Big Data: a) the volume and velocity of big climate data have far exceeded the stand-alone computer's storage and computing ability; b) the variety of climate data in format and spatiotemporal resolution make it difficult to find an easy-to-use tool to analyze climate data; c) the veracity in model simulation is a concern for climate scientists of the uncertainties and mixed model qualities (Murphy et al., 2004). The combined complexities of volume, velocity, variety, and veracity can be addressed with cloud-based, advanced data management strategies and a service-oriented data analytical architecture to help process, analyze and mine climate data.

Table 1
The Big Data challenges as illustrated in the four examples are addressed by relevant cloud advantages to reach the Big Data Value and achieve the research, engineering and application objectives.

	On-demand Self-service	Broad network access	Resource pooling	Rapid elasticity	Measured service
Volume	2.1	4.1	2.1	2.1, 3.1, 3.2, 4.1, 4.2, 4.3, 5.1	4.1
Veracity	2.1	3.1, 5.3			
Velocity			2.1	4.1	4.3
Variety		3.1, 5.2	2.1		
Value	2.1, 3.2			2.1	3.2

2.1. Advanced spatiotemporal index for big climate data management

The hundreds of petabytes of climate data can only be managed in a distributed and scalable environment. Cloud Computing could help the management as follows: a) provisioning on-demand flexible virtual machines (VM) according to the volume of climate data; and b) automatically deploying HDFS, Hadoop Distributed File System, on the VMs to build a distributed filesystem. Data can be maintained in native format instead of sequenced text for saving storage space. A logical data architecture is also built to facilitate fast identification, access, and analyses (Li, Hu et al., 2016; Li, Yang et al., 2016). The core architecture is a spatiotemporal index (Li, Hu et al., 2016; Li, Yang et al., 2016) for the multi-dimensional climate data stored on HDFS. The index maps data content onto the byte, file and node levels within the HDFS. Nine components are used for the index and include: space, time and shape information describe the data grid's logical information which correlates to data query, byte offset, byte length, compression code, node list and file path identify specific location on the HDFS. This index enables users to directly locate and access data with exact spatiotemporal and content description.

In details, the space and time attributes in the spatiotemporal index will identify the grids overlapped with a spatiotemporal bounding box. The node list attribute is leveraged to deliver the computing programs to the node where the grids are stored. Then the computing programs can read the data as a data stream with high data locality, according to the byte offset, byte length, and compression code attributes. The

shape and data type attributes can be used to reshape the data stream into a multiple-dimension array.

The monthly MERRA data for 26 years, MAIMNXINT[1] (about 90 Gb), are used to evaluate the cloud based and spatiotemporal indexed Big Data management efficiency. This experiment analyzes the monthly mean value of specific climate variables (by changing their numbers) in a specified spatiotemporal range from 36 VM-based HDFS cluster connected with 1 Gigabit (Gbps). Each node is configured with eight CPU cores (2.60 GHz), 16 GB RAM and CentOS 6.5. Results with and without using the index (Fig. 2) show when the number of processed variables increased the run time without the index increased by a factor of ~9.1, whereas the run time with the index only increased by a factor of 1.8. Based on the time constraints, a flexible number of VMs can be provisioned on demand to finish the tasks within a specific time frame (Li, Hu et al., 2016; Li, Yang et al., 2016; Yang et al., 2015). Therefore, on-demand service and elasticity in combination with a high level management effectively accommodate the big climate data management and analytical demands.

2.2. Anything as a service to ease the climate modelling experiments

Climate simulation poses challenges on obtaining enough computing resources for scientific experiments when analyzing big simulation data or running a large number of model simulations according to different model inputs. Cloud Computing addresses this experiment as follows: a) the climate models can be published as a service (MaaS; Li et al., 2014) and enough VMs can be provisioned with specific model configurations for each ensemble modelling run on demand; b) the application is deployed as a service (Lushbough, Gnimpieba, and Dooley, 2015) with a popular web portal to support model operation and monitoring; and c) the workflow involving different analytics is operated as a service (WaaS; Krämer and Senner, 2015) with intuitive GUIs. The big climate data analytics are supported by Cloud Computing at the computing infrastructure level.

The architecture of the cloud-based service-oriented workflow system for climate model study includes (Fig. 3): a) the model service is responsible for compiling and running models on VMs, which are provisioned based on the snapshot of the system containing the modelling software environment to run a model; b) the VM monitor service provides the cloud platform with VM status information for resource scheduling; c) the data analysis service feeds the model output as the input for analytics, while analyzing data in parallel to address data intensive issues. Data publishing service enables users to access the analysis results in real time via the Internet. All of these services are controllable through a GUI, which enables users to drag and connect services together to build a complex workflow so the system can automatically transition to the applications specified by the workflow and run on the cloud with automatically provisioned VMs. As an example, Li (2015) built ModelE as a service to study the sensitivity of ModelE, and the experiment showed that this cloud-based method reduced time consumption by 10 times over the traditional method.

The challenges in climate research addressed by Cloud Computing are summarized in Table 1. First, the large volume of climate data from observation and simulation are stored in the distributed and scalable environment provisioned by the cloud platform (2.1). Second, the variety challenge in climate data is addressed using the spatiotemporal index to unify them from the aspects of space and time (2.1). Third, the variety challenge in climate models is relieved by building the service-oriented system to simplify the model setup, running and output analysis (2.2). These methods can be extended to other geospatial domains which involve high dimensional data and complex models, such as remote sensing, image processing and agent-based modelling of environmental and urban events.

[1] http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings/merra/merra_products_nonjs.shtml.

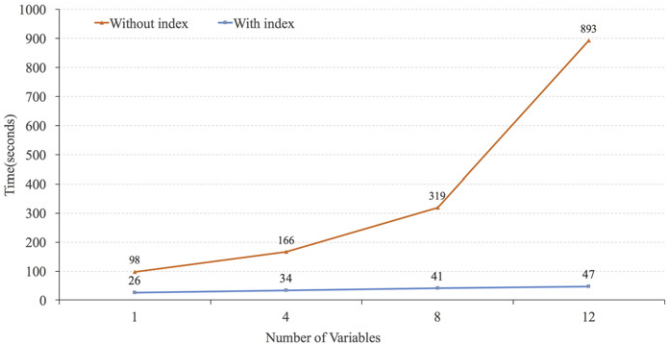


Fig. 2. Run time of the daily global mean calculation for different numbers of variables.

3. Supporting knowledge mining from big geospatial data

We have collected big geospatial data with different spatiotemporal stamps and resolutions for environment and urban studies using various methods, e.g., Global Positioning System (GPS), remote sensing, and Internet-based volunteer (Jiang and Thill, 2015; Yang et al., 2011). The increment in volume, velocity, and variety of the spatiotemporal data poses a grand challenge for researchers to discover and access the right data for research and decision support (Yang et al., 2011). One method of addressing this Big Data discovery challenge is to mine knowledge from the big geospatial data and their usages (Vatsavai et al., 2012) for query expansion, recommendation and ranking. The mined knowledge includes but is not limited to domain hot topics, research trends, metadata linkage and geospatial vocabularies similarity. This process is challenged with Big Data volume, velocity and variety. Such a mining process poses two challenges: a) how to divide Big Data into parallelizable chunks for processing with scalable computing resources; and b) how to utilize an adaptable number of computing resources for processing the divided Big Data. Take the MUDROD project for NASA Physical Oceanography Distributed Active Archive Center (PO. DAAC) as an example, the 2014 web log (contains geospatial data usage knowledge) was over 150 million records and the mining task takes >5 h to complete using a single server (6 cores, 12G memory and Win 7 OS). For high traffic websites with a large number of users sending requests concurrently, logs are produced at a much higher velocity, exceeding a single server's data-processing capability. In addition, logs are semi-structured or unstructured data stored in various formats (e.g. Apache HTTP, FTP, NGINX, IIS log format or user-defined format). Each format requires a specific processing protocol complicating the integration of different formats for further processing. The uncertainty affects the quality of mined knowledge with common noise (e.g., from web crawlers) requiring computational intensive crawler detection algorithms to preprocess original logs (Jiang et al., 2016).

3.1. Accelerating user log mining through data parallelism

The first step to processing big log files is to proceed in parallel by conducting the same operations on a dynamic number of VMs based on data volume and time constraints (Gordon, Thies, and Amarasinghe, 2006). To divide the original logs into the same number of VMs of a cluster, two data parallelism methods are applied to efficiently split logs, including time-based and IP-based log partition. In the time-based log partition (Fig. 4a), logs of consecutive dates are grouped into the same file. Once the original logs are split into k files (i.e., k = number of VMs in the cluster), the difference of the sum of logs in each file is minimized. This partitioning is solved as a linear partition problem (Skiena, 1998). In IP-based log partition (Fig. 4b), logs of the same IP are grouped into the same file using the greedy algorithm (Korf, 2011). Different from the time-based partition, the alteration of arrangement of IP is allowed.

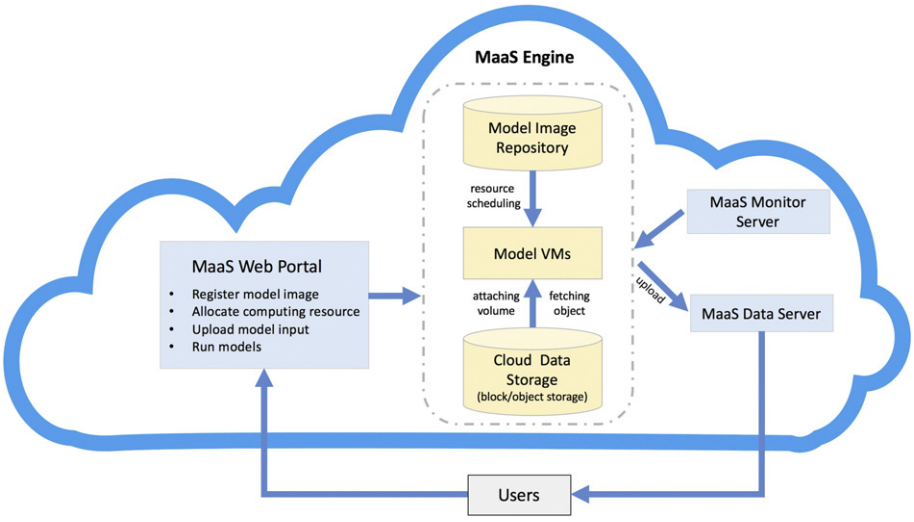


Fig. 3. The cloud-based service-oriented workflow system for climate model study.

Once the original logs are distributed in a virtual cluster, log pieces are processed in parallel in all VMs so log mining efficiency is notably increased.

3.2. Provisioning on-demand computation resources within a virtual cluster

In contrast to setting up a cluster manually, Cloud Computing facilitates the provisioning of a virtual cluster automatically with a dynamic number of VMs (Krämer and Senner, 2015). More computing resources can be deployed to process big historic data while a dynamic number of VMs can be provisioned to handle real-time data streams. On-demand computing resources are necessary to meet the requirement of dynamical log data volumes. For example, in the January 2014 PO. DAAC log mining task with more VMs in the cluster, less processing time was spent on finishing the task (Fig. 5a). Both the time-based partition and the IP-based partition dramatically accelerated the mining processes. However, the time-based partition changed sessions generated by log processer (Fig. 5b).

For the entire 2014 PO. DAAC logs, the total processing time was reduced 70%, from 190 to 49 min, as the VMs increased from 1 to 4 (Fig. 6).

Like geospatial data usage log, geospatial data can also be efficiently processed by a cluster leveraging data parallelism paradigm. Geospatial data can be partitioned into smaller parts based on different aspects, such as latitude, longitude, time or file size, and then distributed among VMs for parallel processing.

As summarized in Table 1, the broad network access and rapid elasticity enables data parallelism methods to efficiently segment Big Data and preparing the data for processing in parallel (3.1). The on-demand self-service, measured service and rapid elasticity add or remove computing nodes in a short time to meet the dynamic computing requirement (3.2).

The proposed knowledge discovery method of mining web log can be integrated with domain data portals to help environmental or urban scientists quickly discover useful information and knowledge. Though it should be pointed out that the user specific profiling (knowledge) data may also be of privacy and security concerns. In urban studies, spatial data mining and geographic knowledge discovery has emerged as an active research field in recent years. GPS data, high-resolution remote sensing data and internet-based volunteered geographic information are collected to extract unknown or unexpected information (Mennis and Guo, 2009; Jiang and Thill, 2015). These data sets are of unprecedentedly large size and the data parallelism paradigm can be leveraged to utilize Cloud Computing for efficient processing, e.g., for analyzing jobs-housing correlations (Long and Thill, 2015) in a smart cities context.

4. Supporting land-use and land-cover change analysis

Land-Use and Land-Cover Change (LULCC) has emerged as a fundamental component of environmental change and sustainability research. Landsat alone has produced 6 petabytes of data (Turner et al., 2003; Hansen and Loveland, 2012). The Land Change Monitoring, Assessment and Projection (LCMAP) pressed the need to generate science-quality land change products from current and near-real time Earth observations (Dwyer, 2014). However, several Big Data challenges exist as follows: a) storing, accessing and sharing big land use data; b) rapidly modelling LULCC with large-scale training set and complex algorithms; and c) rapidly changing analyses and predictions with LULCC data.

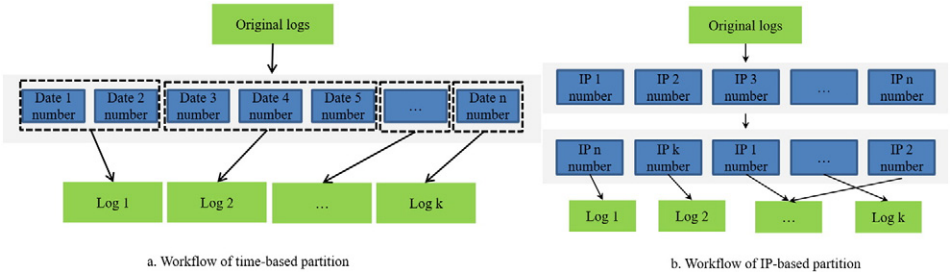


Fig. 4. The workflow of time-based partition (a) and IP-based partition (b).

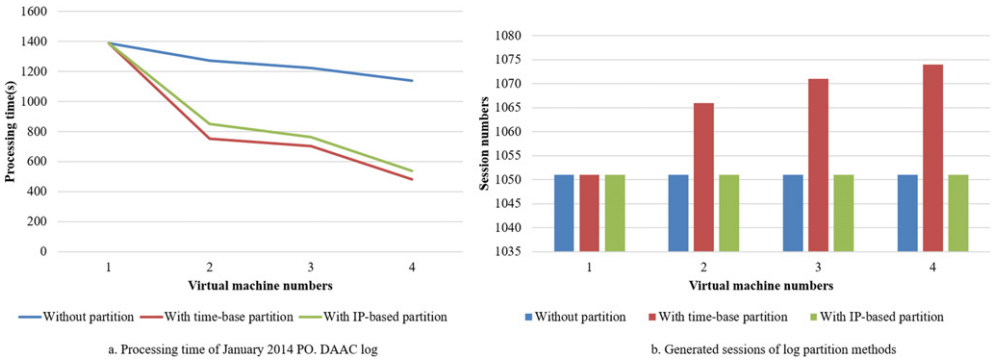


Fig. 5. Processing time of January 2014 PO. DAAC log (a) and generated sessions of log partition methods (b).

4.1. Storing, sharing and analyzing big land cover data on the cloud

Petabytes of historical LULCC and terabytes of streaming LULCC data require expensive and on-premise hardware that is hard to maintain and administrate, while cloud storage is outsourced to third-party cloud providers performing updates and maintenance (USGS, 2016). Additionally, cloud storage supports immediate access and exposes them through one simple web service interface, high reliability through redundancy and distribution of data and pay-as-you-go pricing (Calder et al., 2011). As the longest, continuous record of Earth's land surface observed from space, Landsat data is available via Amazon S3 since 2015. Most Landsat scenes from 2015 are available along with a selection of cloud-free scenes from 2013 and 2014. All new Landsat 8 scenes are available daily and often within hours of production (AWS, 2015). In addition to the improvement of data access, land cover imagery stored on the cloud is combined with land cover modelling published on the cloud to ease the LULCC research workflow, result sharing, and reproducing. For example, ArcGIS Online allows quick visualization and analysis of Landsat data on AWS. Mapbox uses Landsat on AWS to power Landsat-live, a browser-based map that is constantly refreshed with the latest imagery from the Landsat 8 satellite (AWS, 2015).

4.2. Rapid modelling with big training set and complex algorithms

Among the three types of LULCC models on image classification, land use suitability, and environmental impact of land cover change (Eastman, 2012), algorithms are complex and usually involve large training sets to build a robustic model. However, most can be converted to generic data mining problems. For example, the land change modeler of IDRISI, a popular GIS land change modelling tool, is based on logistic regression and neural networks (Eastman, 2003). The parallelization of these data mining algorithms is well studied in Cloud Computing communities and is supported by open source, large-scale processing frameworks, such as Spark MLlib (Meng et al., 2016). To leverage these technologies, a middleware was developed to convert the training set of the land cover images into the format that existing technologies can digest and convert the results back into ones that the LULCC requires (Fig. 7).

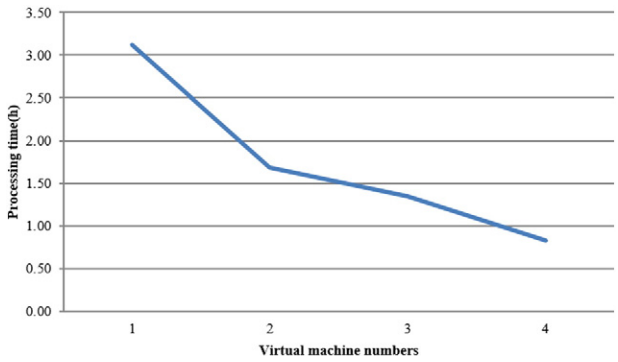


Fig. 6. Processing time of 2014 PO. DAAC log.

4.3. Rapid change analysis and prediction with big LULCC data

A classification or prediction model can be generated either through traditional approach or the one proposed in section 4.2. It would still be computationally intensive if each image and pixel is processed sequentially in LULCC models. This is handled with a virtual cluster to accelerate the processing through the following steps: a) parallelization of the study area into sub-areas; b) distribution of the LULCC data to the VMs where analyses run simultaneously; and c) aggregation of the results into a result dataset. As an example, a series of high-resolution global forest cover change maps from Google Earth Engine (Hansen et al., 2013) through its intrinsically-parallel computational access to Google cloud (Moore, 2015) demonstrates the possibility of utilizing a Cloud Computing platform to accelerate large land cover image classification (Fig. 8).

The broad network process, rapid elasticity and measured service improves the storage, access and analytics of big LULCC data for the data volume and velocity challenges (Tables 1, 4.1). The rapid elasticity allows large-scale data processing framework middleware to support the rapid modelling of large training sets and complex algorithms (4.2). The rapid elasticity and measured service also make it possible for the proposed parallel computing framework to provide near real-time classification, land cover change and prediction maps (4.3). The solutions proposed in LULCC can also be adopted in other scientific issues such as climate change, ecosystem service and habitat and biodiversity modelling.

5. Supporting dust storm forecasting

Dust storms are serious hazards to health, property, and the environment worldwide, especially urban areas (Knippertz and Stuut, 2014; WMO, 2011). During and after a dust storm, traffic accidents increase because of the rapidly decreasing visibility; air quality and human health are compromised when dust particles remain suspended in the atmosphere; efficiency of renewable energy sources is reduced when dust interferes with the energy capture mechanics (Wilkening, Barrie, and Engle, 2000). Therefore, it is crucial to predict an upcoming dust event with high spatiotemporal resolution to mitigate the environmental, health, and other asset impacts of dust storms (Benedetti et al., 2014). A standard requirement for such prediction requirement is to simulate one day phenomena within a two-hour computational time (Xie, Yang, Zhou, and Huang, 2010). This is easy to achieve with a coarse-resolution (1/3 degree) dust model forecast for the U.S. Southwest using a single CPU that takes ~4.5 h to complete processing. For

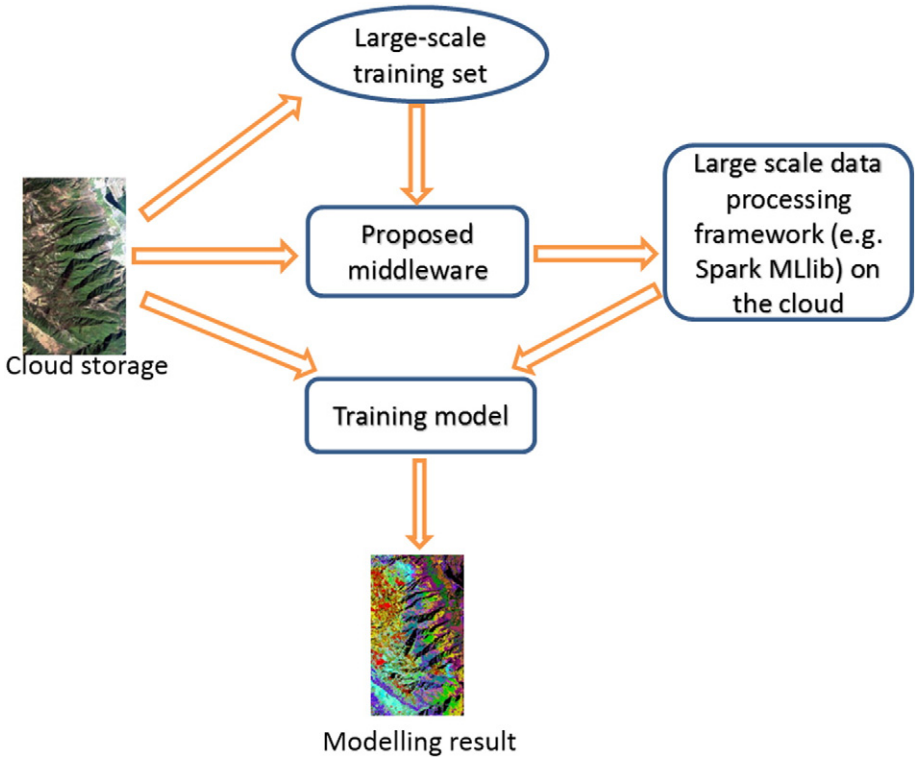


Fig. 7. The role the proposed middleware plays in the model building process.

high-resolution simulation (e.g. 3 km by 3 km), the volume of the model output data increases from 100 Gb to 10 Tb. The computational time increases by a factor of 4 in each of three dimensions (latitude, longitude and time steps). This results in an overall increase of a factor of 64 ($4 \times 4 \times 4 = 64$) or 12 days to complete the processing. This challenge of reducing from 12 days to 2 h is a Big Data problem in how to deal with the large volume of data processing/computing, how to ingest the variety of content input from geographic, atmospheric and ecosystem data and how to improve the veracity of model forecast data by ingesting high quality model input data.

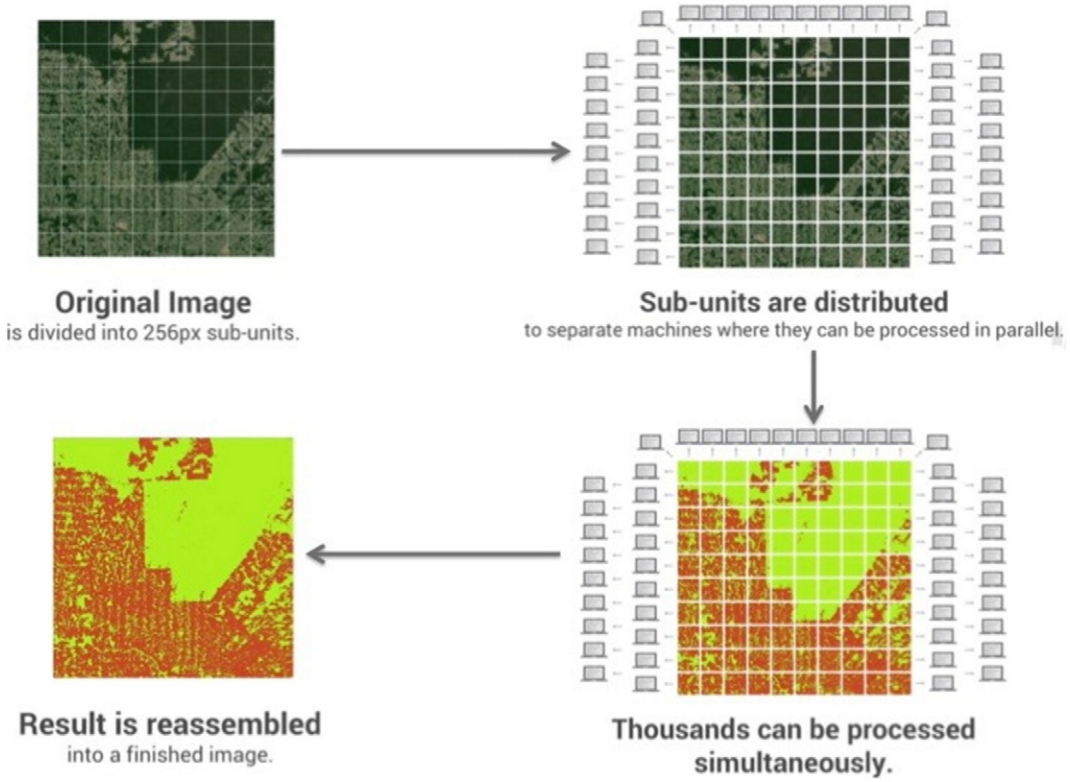


Fig. 8. Google Earth Engine divides Big Data to process in parallel using multiple computers (Moore, 2015).

5.1. Accelerating large volume computation and processing

To deal with the challenge of reducing computing time from 12 days to 2 h, Huang, Yang, Benedict, Rezgui et al. (2013) and Huang, Yang, Benedict, Chen et al. (2013) proposed an adaptive, loosely-coupled model strategy, linking a high resolution/small scale dust model with a coarse resolution/large scale model. This strategy runs the low-resolution model and identifies subdomains, Areas of Interest (AOIs) with predicted high dust concentrations (Fig. 9a). A higher-resolution model for these AOIs is executed in parallel. With the support of Cloud Computing, clusters for high-resolution model runs for specific AOIs are established rapidly in parallel and are completed more efficiently than an execution of a high-resolution model over the entire domain. The execution time required for different AOIs when Cloud Computing handles all AOIs in parallel is <2.7 h (Fig. 9b).

5.2. Ingesting a big variety of dust model input

With the increase of spatiotemporal resolution of a dust forecast model, the challenge is to access dynamic data with different formats, content and uncertainties (Yang et al., 2011). The capability of broad network access of Cloud Computing can serve the access and preprocessing of a larger variety of the model input data with advanced network bandwidth and scalability. Huang, Yang, Benedict, Chen et al. (2013) and Huang, Yang, Benedict, Rezgui et al. (2013) showed that Amazon cloud instances can complete most of the forecasting tasks in less time than HPC clusters (Fig. 10), indicating that Cloud Computing has potential to resolve the concurrent intensity of the computing demanding applications.

5.3. Improving data veracity of dust forecasts

One of the most significant factors affecting the veracity of model output is the uncertainty of model initial condition (Lin, Zhu, and Wang, 2008). These uncertainties can be investigated and characterized through sensitivity tests using various model variables (Zhao et al., 2010; Liu et al., 2012). To reduce the uncertainty of the initial conditions, data assimilation techniques have been applied to dust models by assimilating the observations into the model to correct model initial conditions (Niu et al., 2008; Sekiyama, Tanaka, Shimizu, and Miyoshi, 2010; Liu et al., 2011). With the increasing variety of data sources, sensitivity tests and data assimilation can be conducted with minimum effort of preprocessing and integration into the model, thus enabling the efforts to improve model accuracy, and eventually reduce model uncertainty (Lin et al., 2008; Darmenova, Sokolik, Shao, Marticorena, and Bergametti, 2009). The entire complex process can be precisely preserved in a VM image that can be reused with minimum effort & reducing future manual errors.

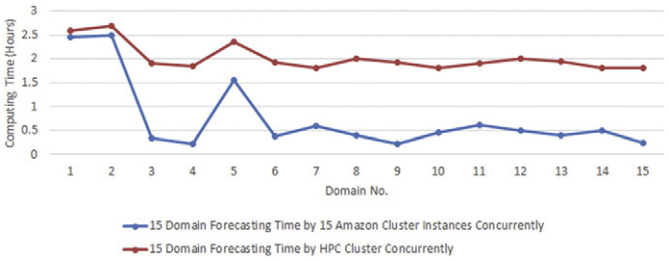


Fig. 10. NMM-dust execution time for 15 forecasting tasks on Amazon EC2 and HPC cluster (Huang, Yang, Benedict, Chen et al., 2013; Huang, Yang, Benedict, Rezgui et al., 2013).

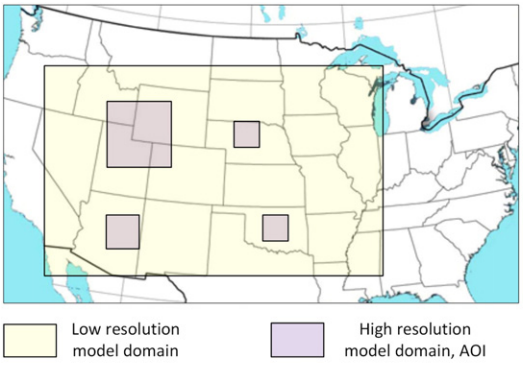
Therefore, the challenges for large-scale scientific prediction can be addressed by engaging the features of Cloud Computing with a net effect of accelerating a dust forecast task (Tables 1, 5.1). With broad network access, the ingestion of a larger variety of input data is achieved, and the ingestion is preprocessed on the cloud without consuming the computing resources designated for the core of model simulations (5.2). The selection of model input data is more sophisticated, improving the representation of the model's initial conditions and potentially data veracity of model's simulation output (5.3). These approaches are easily adaptable in other scientific computation or simulation models that require results within a short period of time, including the prediction of floods, hurricanes, and air pollution.

6. Conclusion

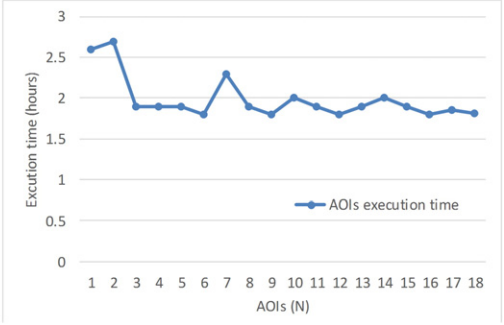
Big geospatial data pose grand challenges during the lifecycle of data storage, access, manage, analysis, mining, and modelling. The four examples illustrate the capability of Cloud Computing to address the 4 V challenges to reach Value with the five Cloud Computing advantages of on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service (Table 1). The boxes filled with section numbers indicate that these sections leverage the features of Cloud Computing to address the relevant challenges within big geospatial data. Table 1 is also of value as a guide to evaluate solutions for other big geospatial data challenges.

While research have been conducted to utilizing Cloud Computing to address big geospatial data challenges, many challenges remain to be addressed:

- Big geospatial data storage and management remains a high priority, including how to optimize different traditional (e.g., MySQL, PostgreSQL) and emerging database Management systems (e.g., NoSQL, HDFS, SPARK, HIVE) on the cloud environment for distributed storage, access and analytics (Agrawal, Das, and El Abbadi, 2011)



a. Low resolution and high resolution model domains



b. NMM-dust model execution time for each AOI in parallel

Fig. 9. Low-resolution model domain area and sub-regions (Area of Interests, AOIs) identified for high-resolution model execution (Huang, Yang, Benedict, Rezgui et al., 2013; Huang, Yang, Benedict, Chen et al., 2013).

- Spatiotemporal Big Data mining requires real-time data processing, information extraction and automation to extract information and knowledge. More scalable spatiotemporal mining methods (Vatsavai et al., 2012) should be developed to take advantage of the elastic storage and computing resources of cloud platforms (Triguero, Peralta, Bacardit, García, and Herrera, 2015).
- Security is a challenge to assure protection for both sensitive data and the users' privacy. More research is needed to tracking and maintaining trust information to identify and prevent attacks on the cloud platform (Manuel, 2015).
- The usage behavior (e.g., when, where, and what VMs are used) on the cloud platform directly affects the energy efficiency and sustainability of the Cloud Computing resources. More tools are necessary to measure usage of resources, including computing resources and data for pricing purposes and to guide use of Cloud Computing services (Yang et al., 2016).
- Spatiotemporal thinking methodologies are critical, and more should be developed and formalized to optimize Cloud Computing for big geospatial data processing (Yang et al., 2015; Yang et al., 2016).
- Utilizing Cloud Computing and Big Data technologies in new initiatives, such as smart cities and smart communities (Batty, 2013; Mitton et al., 2012), should be investigated from the initiative context (Odendaal, 2003), application complexities (Long and Thill, 2015), relevant data selection, fusion, mining (Jiang and Thill, 2015), and knowledge presentation (Fox, 2015).

Acknowledgements

This research is supported by NSF Cyber Polar, Innovation Center, EarthCube and Computer Network System Programs (PLR-1349259, IIP-1338925, CNS-1117300, ICER-1343759) and NASA (NNG12PP37I) as well as Microsoft, Amazon, Northrop Grumman, and Harris. We thank the anonymous reviewers for their insightful comments and reviews. Dr. George Taylor edited an earlier version of the paper.

References

Agrawal, D., Das, S., & El Abbadi, A. (2011). Big Data and Cloud Computing: Current state and future opportunities. *Proceedings of the 14th International Conference on Extending Database Technology* (pp. 530–533) ACM.

Ammn, N., & Irfanuddin, M. (2013). Big Data challenges. *International Journal of Advanced Trends in Computer Science and Engineering*, *2*(1), 613–615.

AWS (2015). *Landsat on AWS.* https://aws.amazon.com/public-data-sets/landsat/.

Batty, M. (2013). Big Data, smart cities and city planning. *Dialogues in Human Geography*, *3*(3), 274–279.

Benedetti, A., Baldasano, J. M., Basart, S., Benincasa, F., Boucher, O., Brooks, M. E., et al. (2014). Operational dust prediction. In P. Knippertz, & W. J. -B. Stuut (Eds.), *Mineral dust: A key player in the Earth system* (pp. 223–265). Dordrecht: Springer Netherlands.

Bulkeley, H., & Betsill, M. M. (2005). *Cities and climate change: Urban sustainability and global environmental governance. 4.* (pp. 1–2). Florence: Psychology Press, 1–2.

Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolsvold, A., McKelvie, S., ... Haridas, J. (2011). Windows Azure Storage: A highly available cloud storage service with strong consistency. *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles* (pp. 143–157) ACM.

Chen, C. P., & Zhang, C. -Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347.

Darmenova, K., Sokolik, I. N., Shao, Y., Marticorena, B., & Bergametti, G. (2009). Development of a physically based dust emission module within the Weather Research and Forecasting (WRF) model: Assessment of dust emission parameterizations and input parameters for source regions in Central and East Asia. *Journal of Geophysical Research. Atmospheres*, *114*(D14).

Das, M., & Parthasarathy, S. (2009). Anomaly detection and spatio-temporal analysis of global climate system. *Proceedings of the third international workshop on knowledge discovery from sensor data* (pp. 142–150) ACM.

Debbage, N., & Shepherd, J. M. (2015). The urban heat island effect and city contiguity. *Computers, Environment and Urban Systems*, *54*, 181–194.

Dwyer, J. L. (2014). Development of Landsat information products to Support Land Change Monitoring, Assessment, and Projection (LCMAP). *AGU fall meeting abstracts. 1.* (pp. 3725).

Eastman, J. R. (2003). *IDRISI Kilimanjaro: Guide to GIS and image processing.* Worcester: Clark Labs, Clark University, 305.

Eastman, J. R. (2012). *IDRISI Selva manual.* Worcester, Massachusetts, USA: Clark University.

Einav, L., & Levin, J. D. (2013). *The data revolution and economic analysis (no. w19035).* National Bureau of Economic Research.

Fan, J., & Liu, H. (2013). Statistical analysis of Big Data on pharmacogenomics. *Advanced Drug Delivery Reviews*, *65*(7), 987–1000.

Fox, M. S. (2015). The role of ontologies in publishing and analyzing city indicators. *Computers, Environment and Urban Systems*, *54*, 266–279.

Frias-Martinez, V., Virseda, J., Rubio, A., & Frias-Martinez, E. (2010). Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development* (pp. 11) ACM.

Gordon, M. I., Thies, W., & Amarasinghe, S. (2006). Exploiting coarse-grained task, data, and pipeline parallelism in stream programs. *ACM SIGOPS Operating Systems Review*, *40*(5), 151–162.

Hansen, M. C., & Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment*, *122*, 66–74.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ... Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*(6160), 850–853.

Hsu, C. H., Slagter, K. D., & Chung, Y. C. (2015). Locality and loading aware virtual machine mapping techniques for optimizing communications in MapReduce applications. *Future Generation Computer Systems*, *53*, 43–54.

Huang, Q., Yang, C., Benedict, K., Chen, S., Rezgui, A., & Xie, J. (2013a). Utilize Cloud Computing to support dust storm forecasting. *International Journal of Digital Earth*, *6*(4), 338–355.

Huang, Q., Yang, C., Benedict, K., Rezgui, A., Xie, J., Xia, J., & Chen, S. (2013b). Using adaptively coupled models and high-performance computing for enabling the computability of dust storm forecasting. *International Journal of Geographical Information Science*, *27*(4), 765–784.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94.

Jiang, B., & Thill, J. C. (2015). Volunteered geographic information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems*, *53*, 1–3.

Jiang, Y., Li, Y., Yang, C., Armstrong, E. M., Huang, T., & Moroni, D. (2016). Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. *ISPRS International Journal of Geo-Information*, *5*(5), 54.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM*, *57*(3), 78–85.

Knippertz, P., & Stuut, J. B. W. (2014). *Mineral Dust.* Dordrecht, Netherlands: Springer.

Korf, R. E. (2011). A hybrid recursive multi-way number partitioning algorithm. *IJCAI proceedings-International Joint Conference on Artificial Intelligence*, *22*(1), 591.

Krämer, M., & Senner, I. (2015). A modular software architecture for processing of big geospatial data in the cloud. *Computers & Graphics*, *49*, 69–81.

Lee, J. G., & Kang, M. (2015). Geospatial Big Data: Challenges and opportunities. *Big Data Research*, *2*(2), 74–81.

Li, Z. (2015). *Optimizing geospatial cyberinfrastructure to improve the computing capability for climate studies.* (Ph.D. Dissertation, George Mason University. http://ebot.gmu.edu/handle/1920/9630).

Li, Z., Yang, C., Huang, Q., Liu, K., Sun, M., & Xia, J. (2014). Building model as a service to support geosciences. *Computers, Environment and Urban Systems*. http://dx.doi.org/10.1016/j.compenvurbsys.2014.06.004.

Li, Z., Yang, C., Liu, K., Hu, F., & Jin, B. (2016a). Automatic scaling Hadoop in the cloud for efficient process of big geospatial data. *ISPRS International Journal of Geo-Information*, *5*(10), 173.

Li, Z., Hu, F., Schnase, J. L., Duffy, D. Q., Lee, T., Bowen, M. K., & Yang, C. (2016b). A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. *International Journal of Geographical Information Science* doi:/10.1080/13658816.2015.1131830.

Lin, C., Zhu, J., & Wang, Z. (2008). Model bias correction for dust storm forecast using ensemble Kalman filter. *Journal of Geophysical Research. Atmospheres*, *113*(D14).

Liu, Z., Liu, Q., Lin, H. C., Schwartz, C. S., Lee, Y. H., & Wang, T. (2011). Three-dimensional variational assimilation of MODIS aerosol optical depth: Implementation and application to a dust storm over East Asia. *Journal of Geophysical Research. Atmospheres*, *116*(D23).

Liu, X., Shi, X., Zhang, K., Jensen, E. J., Gettelman, A., Barahona, D., ... Lawson, P. (2012). Sensitivity studies of dust ice nuclei effect on cirrus clouds with the Community Atmosphere Model CAM5. *Atmospheric Chemistry and Physics*, *12*(24), 12061–12079.

Long, Y., & Thill, J. C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, *53*, 19–35.

Lushbough, C. M., Gnimpieba, E. Z., & Dooley, R. (2015). Life science data analysis workflow development using the bioextract server leveraging the iPlant collaborative cyberinfrastructure. *Concurrency and Computation: Practice and Experience*, *27*(2), 408–419.

Manuel, P. (2015). A trust model of Cloud Computing based on quality of service. *Annals of Operations Research*, *233*(1), 281–292.

Marr, B. (2015). Big Data: Using SMART Big Data. *Analytics and metrics to make better decisions and improve performance*. Wiley 258pp.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt.

Mell, P., & Grance, T. (2011). *The NIST definition of Cloud Computing.*

Meng, X., Bradley, J., Yuvaz, B., Sparks, E., Venkataraman, S., Liu, D., ... Xin, D. (2016). Mllib: Machine learning in apache spark. *JMLR*, *17*(34), 1–7.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*, *33*(6), 403–408.

Mitton, N., Papavassiliou, S., Puliafito, A., & Trivedi, K. S. (2012). Combining cloud and sensors in a smart city environment. *EURASIP Journal on Wireless Communications and Networking*, *1*, 1.

Moore, R. (2015). *How a Google engineer, 66,000 computers, and a Brazilian tribe made a difference in how we view the Earth.* (http://earthzine.org/2015/01/27/how-a-google-engineer-66000-computers-and-a-brazilian-tribe-made-a-difference-in-how-we-view-the-earth/).

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*(7001), 768–772.

Niu, T., Gong, S. L., Zhu, G. F., Liu, H. L., Hu, X. Q., Zhou, C. H., & Wang, Y. Q. (2008). Data assimilation of dust aerosol observations for the CUACE/dust forecasting system. *Atmospheric Chemistry and Physics*, *8*(13), 3473–3482.

Odendaal, N. (2003). Information and communication technology and local governance: Understanding the difference between cities in developed and emerging economies. *Computers, Environment and Urban Systems*, *27*(6), 585–607.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 18–33). Berlin Heidelberg: Springer.

Rosenzweig, C., Solecki, W. D., Hammer, S. A., & Mehrotra, S. (Eds.). (2011). *Climate change and cities: First assessment report of the urban climate change research network* (pp. xvi). Cambridge: Cambridge University Press.

Schnase, J. L., Duffy, D. Q., Tamkin, G. S., Nadeau, D., Thompson, J. H., Grieg, C. M., ... Webster, W. P. (2014). *MERRA analytic services: Meeting the Big Data challenges of climate science through cloud-enabled climate analytics-as-a-service*. Environment and Urban Systems: Computers.

Sekiyama, T. T., Tanaka, T. Y., Shimizu, A., & Miyoshi, T. (2010). Data assimilation of CALIPSO aerosol observations. *Atmospheric Chemistry and Physics*, *10*(1), 39–49.

Skiena, S. S. (1998). *The algorithm design manual: Text. 1*.Springer Science & Business Media.

Skytland, N. (2012). *Big Data: What is NASA doing with Big Data today.* (Open. Gov open access article).

Triguero, I., Peralta, D., Bacardit, J., García, S., & Herrera, F. (2015). MRPR: A MapReduce solution for prototype reduction in Big Data classification. *Neurocomputing*, *150*, 331–345.

Turner, B. L., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., Eckley, N., ... Martello, M. L. (2003). Illustrating the coupled human–Environment system for vulnerability analysis: Three case studies. *Proceedings of the National Academy of Sciences*, *100*(14), 8080–8085.

USGS (2016). *Landsat 8 missions.* http://landsat.usgs.gov/landsat8.php.

Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012). Spatiotemporal data mining in the era of big spatial data: algorithms and applications. *Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for big geospatial data* (pp. 1–10) ACM.

Wilkening, K. E., Barrie, L. A., & Engle, M. (2000). Trans-Pacific air pollution. *Science*, *290*(5489), 65.

World Meteorological Organization (WMO) (2011). *WMO Sand and Dust Storm Warning Advisory and Assessment System (SDSWAS)—Science and implementation plan 2011–2015.* Geneva, Switzerland: WMO.

Xie, J., Yang, C., Zhou, B., & Huang, Q. (2010). High-performance computing for the simulation of dust storms. *Computers, Environment and Urban Systems*, *34*(4), 278–290.

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems*, *34*(4), 264–277.

Yang, C., Wu, H., Huang, Q., Li, Z., & Li, J. (2011). Using spatial principles to optimize distributed computing for enabling the physical science discoveries. *Proceedings of the National Academy of Sciences*, *108*(14), 5498–5503.

Yang, C., Xu, Y., & Nebert, D. (2013). Redefining the possibility of digital Earth and geosciences with spatial Cloud Computing. *International Journal of Digital Earth*, *6*(4), 297–312.

Yang, C., Sun, M., Liu, K., Huang, Q., Li, Z., Gui, Z., ... Lostritto, P. (2015). Contemporary computing technologies for processing big spatiotemporal data. *Space-time integration in geography and GIScience* (pp. 327–351). Netherlands.: Springer.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2016). Big Data and Cloud Computing: Innovation opportunities and challenges. *International Journal of Digital Earth*. http://dx.doi.org/10.1080/17538947.2016.1239771.

Zhao, C., Liu, X., Leung, L. R., Johnson, B., McFarlane, S. A., Gustafson, W. I., Jr., ... Easter, R. (2010). The spatial distribution of mineral dust and its shortwave radiative forcing over North Africa: Modeling sensitivities to dust emissions and aerosol size treatments. *Atmospheric Chemistry and Physics*, *10*(18), 8821–8838.