IJDC | *General Paper*

Using a Computational Study of Hydrodynamics in the Wax Lake Delta to Examine Data Sharing Principles

Qian Zhang
University of Illinois at Urbana-Champaign

Heidi Imker
University of Illinois at Urbana-Champaign

Chunyan Li
Louisiana State University

Bertram Ludäscher
University of Illinois at Urbana-Champaign

Megan Senseney
University of Illinois at Urbana-Champaign

Abstract

In this paper we describe a complex dataset used to study the circulation and wind-driven flows in the Wax Lake Delta, Louisiana, USA under winter storm conditions. The whole package bundles a large dataset (approximately 74 GB), which includes the numerical model, software and scripts for data analysis and visualization, as well as detailed documentation. The raw data came from multiple external sources, including government agencies, community repositories, and deployed field instruments and surveys. Each raw dataset goes through the processes of data QA/QC, data analysis, visualization, and interpretation. After integrating multiple datasets, new data products are obtained which are then used with the numerical model. The numerical model undergoes model verification, testing, calibration, and optimization. With a complex algorithm of computation, the model generates a structured output dataset, which is, after post-data analysis, presented as informative scientific figures and tables that allow interpretations and conclusions contributing to the science of coastal physical oceanography.

Performing this study required a tremendous amount of effort. While the work resulted in traditional dissemination via a thesis, journal articles and conference proceedings, more can be gained. The data can be reused to study reproducibility or as preliminary investigation to explore a new topic. With thorough documentation and well-organized data, both the input and output dataset should be ready for sharing in a domain or institutional repository. Furthermore, the data organization and documentation also serves as a guideline for future research data management and the development of workflow protocols. Here we will describe the dataset created by this study, how sharing the dataset publicly could enable validation of the current study and extension by new studies, and the challenges that arise prior to sharing the dataset.

Received 20 October 2015 ~ *Revision received* 9 January 2017 ~ *Accepted* 23 January 2017

Correspondence should be addressed to Qian Zhang, School of Information Sciences (iSchool), UIUC, 112 LIS, 501E. Daniel Street, Champaign IL 61820-6211. Email: zqian1@illinois.edu

An earlier version of this paper was presented at 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Computational research now relies on high performance computing (HPC) and high throughput computing (HTC). A very large-scale numerical simulation that used to run weeks to months can now, with the power of HPC, be completed within hours with multiple cases executed on clusters in parallel. This power creates datasets ranging in scale from GB to TB. As an example, if a simulation generates 10 GB of data and 30 simulations are run per day, terabytes of data will be generated per week. What kind of systems (locally or remotely) can handle storage and preservation of such big data generation and analysis day by day? Even if some simulation results are ignored when found not to be sound and thus are deleted, what about the fate of sound simulation data that underpin major findings presented in the literature? Is there a 'home' for this data so that others may reproduce the study? What other uses could this data have beyond the initial work? In this paper, we will apply these questions to a complex, computational coastal oceanographic study of the Wax Lake Delta, an unusual sub-delta of the Mississippi River Delta Complex.

The computational study applies a three-dimensional numerical model ECOM-si (Blumberg, 1994) to simulate the circulation in the Wax Lake Delta under winter cold front conditions. This model uses real-time topography and bathymetry of the area to reproduce tides and the circulation between December 2012 and January 2013, encompassing a total of seven cold front events. The whole simulation package resulted in an over 70 GB of data in total.

Background

History of the Wax Lake Delta

The Wax Lake Delta (WLD) originated from the artificial Wax Lake Outlet (WLO), which was a man-made 30-mile long channel dredged by U.S. Army Corps of Engineers (USACE) in 1941 to divert ~30% of the water from the Mississippi River to the Gulf of Mexico. The intent was also to help reduce flooding in Morgan City, LA which is located northeast of the WLO. Since then, sediments rapidly deposited outside of the mouth of the outlet and formed a bayhead delta that represents the embryonic stage of a new major Mississippi River Delta lobe (delta growth is >6m/yr since 1950s; Roberts et al., 1980; Roberts, et al., 2003). This new lobe is the river-dominated sand-rich WLD, which is located downstream of the WLO and receives 34 million tons of sediment per year. From 1941 to 2011, sediment completely filled the delta and the delta growth rate has changed from 1km^2/yr (Allen, Couvillion, and Berras, 2011) to 2.98km^2/yr (Carle, 2013). Development of a network of channels separated by vegetated islands/bars was associated with this growth, and the area continues to evolve as the delta advances seaward (e.g., Roberts et al., 2003; Wellner et al., 2005; Edmonds and Slingerland, 2007; Day et al., 2011; Roberts et al., 2015).

 doi:10.2218/ijdc.v11i2.433

Importance of the Wax Lake Delta

Although the WLD was initially constructed artificially by USACE, it has maintained itself ever since and undergone natural changes through sediment deposition with little anthropogenic influence. The delta is important because it represents the next major lobe of the Mississippi River Delta Complex (e.g., Roberts and Sneider, 2003). Most of the modern Mississippi River Delta developed naturally over the last 600-800 years but nearly a third of its total area is now lost due to rapid sea level rise induced delta-plain subsidence and deficiency of terrigenous wetland sediment. Through exposure to catastrophic tropical storms (such as Hurricane Katrina in 2005) and winter storms (such as cold front passages) associated with storm surge-induced flooding, the rest of the Mississippi River Delta system degrades. However, the WLD is the only active delta that recovers quickly after damage and continues to regrow. Therefore, the discussion on whether it is feasible to build new land in the Mississippi River Delta has become a topic of great interest, and because of its nascent and rapid evolution, the WLD is consequently often used as the primary natural model by geoscientists, engineers, and ecologists in the study of both sedimentary processes and deltaic formation in the region. Moreover, the healthy and self-maintaining delta has served as a model not only for protection against storm surges but also for post-storm restoration of wildlife habitat. Therefore, study of the WLD provides insights into this delta's ability to serve as an unusual example of environmental and ecosystem sustainability not only in the Mississippi River Delta Complex but also for deltaic coasts around the world.

Motivation and Objectives for the Wax Lake Delta Study

In order to understand the evolution of the WLD, understanding the hydrodynamic processes of delta progradation is a prerequisite and thus of great interest of this study. Although some observational work (for example, Walker and Hammack, 1999; Kemp et al., 1980) demonstrated from both satellite and in situ collected data that winter cold front passages are important modifiers of deltaic deposition, very little work has actually numerically modeled the dynamic process of cold front passages in the region. Additionally, there is a need for high-resolution numerical models because of the WLD's complexity. To address this gap, authors Zhang and Li designed and implemented a numerical hydrodynamic simulation of the WLD region to model water level variation and associated bay-shelf water exchange during winter cold front passages (Zhang, 2015), which is essential for fully understanding the land-sea interaction. Furthermore, a better understanding of these two processes is critical to enable better coastal management of the dynamic ecosystem and its environmental, economic, and residential aspects.

On the other hand, there has been lack of understanding of hydrodynamics in areas that are constantly changing, partially because of the lack of data and partially because of the lack of a high-resolution numerical model that can resolve complex and inundated wetland areas. Thus the challenge is that detailed topography and bathymetry is needed for such a model, but integration of such data is difficult due to very limited and sparse real-time data availability. Because of these challenges, we seek to understand how to best manage and share the data from the WLD study so that others can take advantage of the effort already put into addressing these challenges.

Dataset Description

The WLD study consists of not only a very large dataset (approximately 74 GB) but also a numerical model, software, and scripts for data analysis and visualization, as well as detailed documentation. Figure 1 shows the data flow of this specific case study. The numerical model (ECOM-si) is the central component, which is the 'bridge' between input data and output data. The input transmitted into the numerical model is set up by three driving forces – tides, wind, and river discharge, as well as associated mesh grids and bathymetry configurations. The model output consists of user-defined variable data files in both time series and field distribution. The model simulations were completed using ECOM-si for the time period of December 15th, 2012 to January 20th, 2013, with the support of in situ measurements from 14 sites (Table 1).

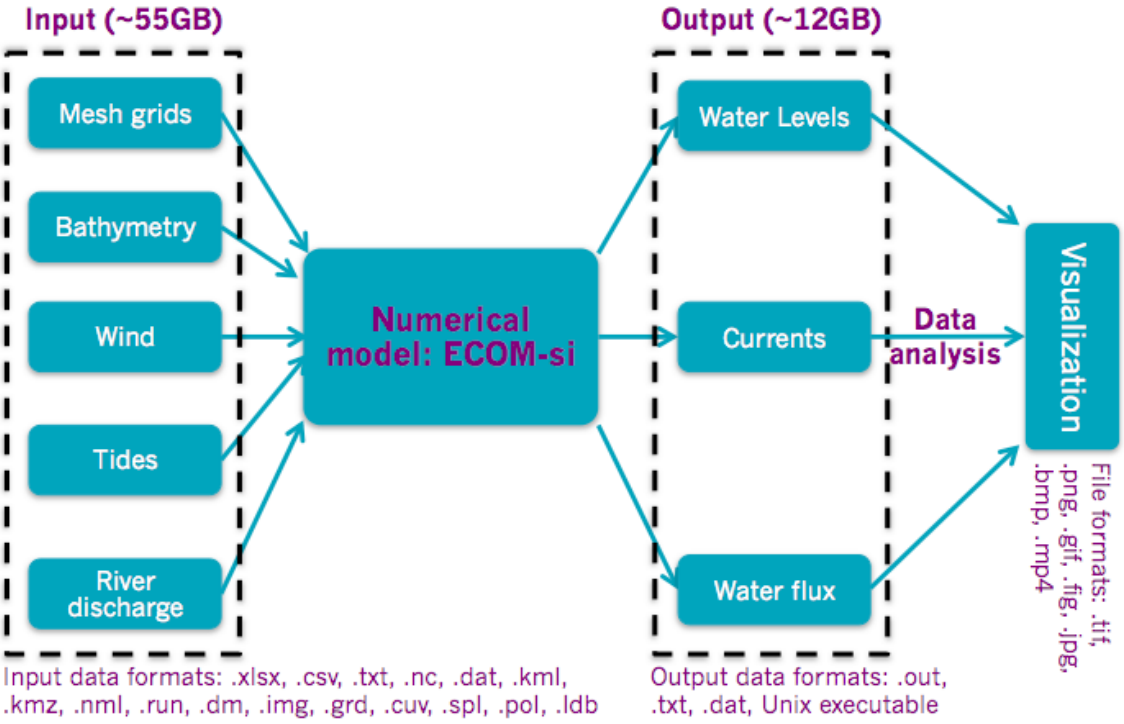

Figure 1. Data flow of the WLD hydrodynamics study: multiple input data are configured for numerical model setup, which initiate the simulation and output user-customized variables with time evolving. Both input and output datasets are highly structured, and must be visualized for further interpretation and analysis.

The data at each station/site are a combination of meteorological (winds, sea level air pressure (barometric pressure), air temperature, sea surface temperature, and humidity), oceanographic (water levels and currents), and surface-water (river discharge) data. Both the meteorological and oceanographic data along the Louisiana coast can be obtained from observational stations at NOAA's National Data Buoy Center[1] and the Center for Operational Oceanographic Products and Services[2]. A third important data source comes from the LSU Wave-Current-Surge Information System[3]

1 National Data Buoy Center (NDBC): http://www.ndbc.noaa.gov/to_station.shtml
2 The Center for Operational Oceanographic Products and Services (CO-OPS): http://tidesandcurrents.noaa.gov/map/
3 Wave-Current-Surge Information System (WAVCIS): http://www.wavcis.lsu.edu/

 doi:10.2218/ijdc.v11i2.433

lab, which provides current profiles in the continental shelf region off Louisiana using deployed instrument Acoustic Doppler Current Profilers (ADCP) on a near-real time basis. For surface-water data, the daily mean river discharge on site were obtained from USGS station 07381590 (the WLO at Calumet, LA) and USGS station 07381600 (the Lower Atchafalaya River at Morgan City, LA), respectively. Additionally, several surveys inside the WLD were conducted for collecting water level, current data, and most importantly for this study, topography and bathymetry data.

As mentioned earlier, the challenge of the WLD is due to its constantly changing topography and bathymetry, which can be influenced by inundation due to local or remote winds, high river discharge, and other physical processes. Compiling topography and bathymetry data of this complex area is the foundation to investigating this area. In this study, five data sources are used to incorporate the bathymetry in the WLD: the first one is the public model data from National Geophysical Data Center's (NGDC) three arc-second (~90 meters) U.S. Coastal Relief Model[4], which itself is a synthesis of NGDC's NOS hydrographic surveys, multibeam bathymetry, track line bathymetry, USGS, and other federal government agencies and academic institutions.

In addition to being extracted from numerical ocean models, the second source of bathymetry data are obtained from Light Detection and Ranging (LIDAR) measurements. In this study, a topographic LIDAR survey was conducted over the WLD. The other three data sources are all based on field trips conducted by the Li research group into the WLD for bathymetry survey in the winters of 2011, 2012, and 2013. Those in situ measurements provided more accurate and more complete bathymetric information and are thus very valuable.

All the raw data in Table 1 originated from multiple external sources, including government agencies (NOAA's NDBC, CO-OPS, USGS, NGDC), community repositories (WAVCIS), and deployed field instrument surveys (three field trips, LIDAR measurements), which cumulated in a very large dataset (~55 GB) with complex data structures and file formats (for example, .xlsx, .csv, .txt, .nc, .dat, .kml, .kmz, .nml, .run, .2dm, .img, .grd, .cuv, .spl, .pol, .ldb).

Each raw dataset must go through the processes of data visualization, data analysis, and data cleaning in order to generate reasonable information needed to support the research. Take NOAA's NDBC wind data at station FRWL1[5] in 2012 for instance, with the time series plotting of wind, it can be identified that there exists an abnormal spike on 12/18/2012 at 08:00 am, where the wind speed is 99.0m/s associated with the direction of 999o. These measurements are in clear contrast to the wind magnitude of less than 3m/s one hour before and after, not to mention that a wind direction of 999° contradicts the standard nomenclature that dictates measurements are expressed between 0° and 360°. This analysis indicates that the wind information at this timestamp is 'bad' or 'missing,' and a placeholder was used instead.

The simplest way to deal with this may be to just delete the problematic data point. Alternatively, we could 'approximate' the data point in question via, for example, an interpolation method by using the valid data points before and after this time instance. Of course, we could also look for records from other sources (such as WAVCIS) to check the data availability and accuracy, etc., and then 'borrow' their information to replace the poor-quality data.

4 Coastal Relief Model (CRM): http://www.ngdc.noaa.gov/mgg/coastal/crm.html

5 FRWL1 – Station ID of Fresh Water Canal Lock, LA: http://www.ndbc.noaa.gov/view_text_file.php?filename=frwl1h2012.txt.gz&dir=data/historical/stdmet/

Table 1. Summary of data stations obtained from the observational stations across the Gulf coast (asterisks indicate publicly available data from the corresponding station).

Location			Data					
Station	Latitude	Longitude	Wind	Barometric Pressure	Air Pressure	Water Level	Current	River Discharge
Delta #1	29.506357°	-91.472°			*	*	*	
Big Hogs Bayou	29.518045°	-91.355°			*	*	*	
CSI03	29.411°	-92.061°	*	*	*	*	*	
CSI06	28.867°	-90.483°	*	*	*	*	*	
CSI09	29.1015°	-89.978°	*	*	*	*		
LAWMA	29.4483°	-91.337°	*	*	*	*		
TESL1	29.6667°	-91.237°	*	*	*	*		
FRWL1	29.555°	-92.305°	*	*	*	*		
GISL1	29.263°	-89.957°	*	*	*	*		
PSTL1	29.178°	-89.258°	*	*	*	*		
PILL1	28.932°	-89.407°	*	*	*	*		
PORT FOURCHON	29.113°	-91.198°				*		
WLOC	29.698°	-91.372°				*		*
LARMC	29.693°	-91.212°				*		*

This process is more complicated because it usually involves even more visualizing, examining, and cleaning, as well as merging and integrating multiple datasets in various formats, considering different resolution in both time and space. However, the benefits are an increase in both data size and quality reliability, if done contentiously. Regardless, the result is a new set of wind data, which is compatible with the numerical model. Similarly, other data variables (e.g., bathymetry data, river discharge, and tidal information) all need to go through similar quality control steps which involve judicious, meticulous, and time-consuming data cleaning and fusion.

However, the fitness for use of the input dataset alone is not sufficient for a successful simulation. Before applying any real use case, the model has to be verified and validated. This is essential to ensure that the model is solving the problem in question in an accurate way. In this context, this is represented as a set of governing equations that couple the conservation of continuity, momentum, temperature, salinity and density in differential forms. To achieve that, the governing equations must be simplified by ignoring as many nonlinear (high-order) terms as possible until the most exact solution can be obtained.

 doi:10.2218/ijdc.v11i2.433

When the numerical solver agrees with the exact result, we can say that the model is verified. Because the simulation model is designed to emulate the real world, it needs calibration, customized configuration, and potentially optimization in order to generate results as close to the physical world as possible. Of course, the model cannot perfectly represent every aspect of reality, so a series of validation trials are needed to tune the model into a more accurate one that aligns with our research interests. It is known that the simulation model consists of complex computing steps and highly sophisticated numerical algorithms, so by comparing the numerical results and real-time observations and measurements, the most suitable model setups are configured, which includes value selection of certain parameters, re-formulating or ignoring some terms in the governing equations, etc. Finally, the model is ready to use for further investigation.

In this study, a total number of 37 successful simulation cases were executed and analyzed. Each simulation generated a series of highly structured datasets (~12 GB, data type example: .out, .txt, .dat, Unix executable). The output in Figure 1 only shows the model products that directly contributed to this study and the corresponding results, but the actual output data are more extensive. Major scientific conclusions of the study included (1) water intrusion and flow partition within the delta distributary network; (2) a relationship between currents and winds, as well as impact on sediment transport-induced geomorphology change in the delta; (3) cold-front-induced flushing event analysis and consequences; and (4) energy distribution breakdown and dominant forcing exploration (Zhang, 2015).

This study was part of a state project titled 'Delta development and coastal marine accretion during cold front passages and floods: Relevance to river diversions (2013-2015)', funded by the State of Louisiana Coastal Protection and Restoration Authority (CPRA) from 1st September 2013 – 31st August 2016, NOAA through the NGOMEX09 project, and NOAA through GCOOS and WAVCIS. The project collaborators included PI Dr. Harry Roberts, co-PIs Dr. DeWitt Braud, Dr. Ron Delaune, Dr. Chunyan Li, and Dr. John White from Coastal Studies Institute (CSI) at Louisiana State University (LSU). The numerical model used in this study was shared by Dr. Jun Lin from College of Marine Sciences at Shanghai Ocean University, China. Some fieldwork that deployed and retrieved the instruments was assisted by Mr. Eddie Weeks and the Field Support Group of CSI. The study period was from December 2013 to May 2015. Furthermore, this work was supported in part by the US National Science Foundation awards DBI-1356751 (KURATOR) and SMA-1439603 (SKOPE).

Data Reuse and Re-Discovery

Research Lifecycle and Data Lifecycle

We have touched two lifecycles in this computational oceanography case study. The first one is the research lifecycle (e.g., research lifecycles proposed by the University of Central Florida Libraries[6] and by the University of Michigan Library[7]), generally consisting of proposal planning and writing, project startup, working with data, and completion of project via a final report/publication. Almost all the research work

6 Research lifecycle at University of Central Florida Libraries: http://library.ucf.edu/about/departments/scholarly-communication/research-lifecycle/

7 Research lifecycle proposed by University of Michigan Library: http://guides.lib.umich.edu/DiscoveryPoE

follows those processes from the beginning to the end, and this computational case study is not an exception. The second cycle is the data lifecycle (e.g., University of Virginia Library[8], DCC Curation Lifecycle Model[9], DataONE data lifecycle[10], California Digital Library Research and Scholarship Lifecycle[11]), which is mainly comprised of data discovery, data collection, data analysis, and data sharing, the processes of which are not necessarily orderly and linear.

It has been observed in this study that some activities in the data lifecycle, such as collecting, integrating, and analyzing data, play key roles in the research process of the research lifecycle, but the research encompasses more than just the data-centric steps (Carlson, 2014). Another observation is that the data lifecycle occurs over the entire research process but is only highlighted during particular phase of research. Thus the data lifecycle can be considered as a subset of the research lifecycle. Sometimes data is a product of research (such as the output data); but when some unexpected results or data errors occur (such as during the data integration process), the research stage may need to move backward into the data lifecycle to conduct additional processing or even start over again.

This WLD case study is trying to answer the question: While a Ph.D. dissertation was completed and a journal paper is currently under co-authors' review for submission, what else can be gained based on the efforts into both the research project as a whole and the data? The end of the research lifecycle is not complete, and taking the data lifecycle into account will extend the value of the research products. By sharing related research data, other impacts such as big data aggregation, research reproducibility, and data reuse can be realized (Borgman, 2012; Fisher and Zigmond, 2010; Pejša, 2012; Piowwar and Vision, 2013; Pejša et al., 2014). In the next section, we will discuss the specific benefits of WLD data sharing in more detail.

Data Sharing Motivations

In recent years, the trend for increasing access to the data that supports research discovery has received more attention from many mainstream publishers (e.g., PLOS One[12], Science[13], Nature[14], and Proceedings of the National Academy of Sciences[15]). Since 2013, data sharing has become an explicit mandate in the OSTP memo requiring funding agencies to develop a plan ensuring "digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze"(Holdren, 2013). In addition, "policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job

8 Data lifecycle at the University of Virginia Library: http://data.library.virginia.edu/data-management/lifecycle/
9 DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model
10 DataONE data lifecycle: https://www.dataone.org/data-life-cycle
11 California Digital Library Research and Scholarship Lifecycle: http://www.cdlib.org/about/approach.html
12 Data sharing policies for all of PLOS journals: http://journals.plos.org/plosone/s/data-availability
13 Editorial policies for Science: http://www.sciencemag.org/authors/science-editorial-policies#dataavail
14 Data policies for Nature: http://www.nature.com/authors/policies/availability.html
15 Data policies for Proceedings of the National Academy of Sciences: http://www.pnas.org/site/authors/journal.xhtml

creation" (Holdren, 2013). But what are the specific benefits of sharing data in this computational oceanography study?

The President's Council of Advisors on Science and Technology report (PCAST, 2011) addressed the challenge in biodiversity ecosystems as "temporal, spatial, and methodological heterogeneity", which is exemplified in the Wax Lake Delta. Although it is much smaller in size compared to the Mississippi River Complex system, the success of land building on this delta is so unique that it is poised write a new chapter of deltaic research to help understand and address coastal problems on a global scale.

Additionally, considering the continually changing and dynamic (wet and dry or inundation) locality, it's financially and logistically impossible to set up multiple monitoring systems on one hand and unrealistic to reproduce the in situ data condition on the other. Therefore, sharing such a rare and unique dataset will accelerate the realization that such data is a vital regional and national asset to be both protected and exploited. Furthermore, reuse of those freely available and accessible hydrodynamics data will enable effective mobilization, integration, and incorporation into other 'grand challenge' of scientific endeavors, such as shoreline protection and marsh creation.

Data sharing has different scopes: on the one hand, data can be shared either directly through individual agreements or indirectly through presentations, publications, and other research products. On the other, data can also be shared by depositing into a reliable data repository or archival system to make it discoverable and accessible to a broad audience. The latter extends the value beyond either the life of the research project cycle or the initial set of researchers involved.

Such sharing mechanisms enhance research transparency and thus the understanding of research outcome, and promote data (re-)discovery and reuse, which in turn inspire new collaboration and innovative ideas for other research topics, especially for "bridging the converging interests" (Keralis, 2012) across domain boundaries. The data sharing discussed in this paper is the latter scope; that is, deposition of data in a public repository so that everyone can have access.

Data sharing can advance the 'big data' world by minimizing duplication of research effort, which besides being more efficient, ensures the uniqueness of the dataset and ownership of the data (Fischer and Zigmond, 2010). Meanwhile, open data enables recombination of data from heterogeneous sources spanning multiple times and places to ask new questions (Whitlock, 2011).

In retrospect, most results from our WLD study are drawn based on either multiple data sources or multiple data variables. Without those data being shared and publicly discoverable, there would have been no way to complete this research investigation with a limited number of field surveys that are inherently non-comprehensive and costly. Thanks to data sharing, the bathymetry data used here are a composite of five sources, with staggered resolution on the local and global scale, respectively.

With all this information, data integration can be carried out by, for example, both interpolation and extrapolation approaches. From this point, access to open data is less about inspecting one's findings from an individual dataset but more about the capability of acquiring and synthesizing data. In fact, this might be the greatest advantage of data sharing – achieving and combining data from multiple sources and then comparing or blending in innovative ways (Butler, 2006).

In spite of diverse sources and data types, integrated datasets follow a common community convention such that most people can easily understand and interpret them. The more systemically data are collected, organized, and processed, the more reliably the community can integrate datasets from multiple sources. Furthermore, the more likely multiple datasets can be mixed, the more potential for stronger correlations. By

assembling data from numerous sources and identifying new relationships, scholars investigate new questions in not only abundance but also depth and breadth (Borgman, 2012).

Data sharing could also enable asking new questions and expanding the scope of research exploration and collaboration (Borgman, 2012; Fisher and Zigmond, 2010; Piowwar and Vision, 2013; Pejša et al., 2014). Take the current WLD case study as an example. With thorough documentation and well-shaped data organization, both raw and model-generated datasets are ready for sharing in a domain repository or an institutional data repository. Recall the fact that the input data covers the domains of meteorology and oceanography and that the output data compute more variables (although unlisted in Figure 1) such as salinity, temperature, sediment transport, turbulence, and particle tracking as byproducts. Those data can be (re-)discovered and reused in fields beyond computational oceanography to study topics of, for instance, the survival of catastrophic storms in atmosphere and climatology, fishery and coastal wildlife habitats in ecosystem sustainability, sedimentary architecture and delta/water body evolution in geomorphology, nitrogen removal and carbon sequestration in biochemistry, as well as coastal hazard prediction and mitigation from flooding and land loss, etc. (CPRA, 2012).

A LSU professor (K. Xu, personal communication, March 16, 2016) in Geological Oceanography mentioned that one of his ongoing projects is working on the geomorphology evolution of this area by studying the interaction among currents, plants, and sediment transport in one of the small islands in WLD. In order to achieve that goal, the field support group has set up tripods to closely monitor changes in vegetation growth, sediment transport, topography, currents, salinity, and waves. All of the WLD data from the project described in this paper (both observation and simulation data) can readily contribute to the needs of Xu's study except for vegetation. Even though our datasets may not fall in the same time period as theirs, the WLD data could at least prove useful for a preliminary analysis. Therefore our whole dataset can be used as prerequisite data for exploring sediment diversion process that will potentially enhance coastal preservation and restoration in both Louisiana and the world coastal systems.

Furthermore, the three field trips described earlier did not only survey bathymetry and topography information, but also collected water samples at different water depths. According to another biochemistry oceanographer (S. Ates, personal communication, March 18, 2016) at LSU, the water sample data, together with the oceanographic data in this WLD study, can be used to investigate the nutrient elements in the delta water body, water residence time, and the relationship between water flow-induced turbidity, algal blooms, and phytoplankton primary productivity/biomass from the perspective of wetland biogeochemistry. Additionally, Dr. Ates is excited because the WLD flow partition conclusion in this study may provide a good indication on the biodiversity distribution in the delta, too.

All of these data can be either raw and unstructured, or cleaned, integrated, well-structured, and ready for use. In any case, one may stand 'on the shoulders of giants' by reusing existing – and shared – data to efficiently use time, effort, and research funding. Once the data is released, it is up to the data re-user to determine new lines of appropriate research and not the original data creator. When data are shared more quickly and openly, analysis and results based on each other's data are drawn more readily.

Publishing data can work in the similar way as publishing a journal article, where the incentives are given to those who create the data. For example by minting a unique

persistent identifier such as DOI (e.g., Figshare[16], zenodo[17]), the dataset is easily discovered, identified, and cited once published. As further encouragement, whenever the shared data is reused, it is ideal if the usage is tracked and associated metrics are made available, such as the number of views, downloads, and citations (Pejša et al., 2014).

Another benefit of data sharing is to improve the transparency of research, thus allowing others to verify the research more easily and provide helpful feedback (Pejša et al., 2014; Stodden, 2010). Scientific discoveries happen as a result of sequences of seemly smaller but indispensable steps. If the WLD data from this study were made publicly available, an attempt to reproduce the study independently could be made thereby strengthening (or amending) the validity of the research findings. Although the results could be negative, sharing data conveys to other researchers that those publishing the initial research findings have sufficient confidence in their work to welcome any steps taken to validate, correct, or extend the study.

That said, on a larger scale, taking the same model input and model setup will generate the model output in a similar sense as the current results, and on a smaller scale, data analysis scripts with the same data input can be verified with the same output. However, not everything can be reproducible. For example the mesh grid generation is irreproducible because it makes use of the open software Delft3D-RGFGRID (Delft3D-RGFGRID, 2013) that works like a semi-automatically drawing tool for manual, user-defined delineation of the computational mesh grids. The exact same grid area cannot be reproduced because the user determines the shape manually and no two users would draw the exact same shape; indeed, the same user will draw slightly different shapes.

Regardless of some complications, once data sharing-enabled verification occurs, the researchers, scientific findings, and dataset themselves all obtain more recognition and credibility. The subsequent reuse can thus strengthen the record of scholarship, which can be reflected as citation impact and reputation, etc., directly associated with individual's productivity and competitiveness.

The motivation to release data depends to a large degree on the amount of time, labor, and resources required, which vary by both the purposes for which data were collected and the approaches to handling data such as documentation, cleaning, and converting into reusable formats (Borgman, 2012; Stodden, 2010; Tenopir et al., 2015). For the WLD data, the drive for sharing is first based on the belief that data and publication support each other, where the release of data and publications are coupled as a whole and complementary to each other (Borgman, 2012).

In this WLD study, the main methods and findings have been addressed in the scholarly publications (i.e. one dissertation and one journal paper), so data sharing as a next step will provide support for reproducible research and add value to the publications and vice versa (Borgman, 2007; Bourne, 2005; Pepe, Mayernik et al., 2010).

Equally important, the release of data will serve as a third copy backup of the whole data package. The third reason is to make the rare but important dataset publicly available to enable derivative work. The data have high value and are irretrievable once lost. Sharing the WLD bathymetry data will not only fill in the necessary data for that area but contribute to our knowledge of the Mississippi River Complex by enabling prediction of the next phase of the change and growth. At the same time, it might encourage researchers interested in this area to release their own data, which ultimately

16 Figshare: https://figshare.com/
17 zenodo: https://zenodo.org/

could result in opening up a brand new research topic in the WLD. Likewise, disclosing the data in this region may encourage researchers to disclose their data in other regions of the world, making a more complete and accurate profile of the ocean. Finally, since all the research details are well documented together with data cleaning, analysis and management, there is actually no more extra effort or diligence needed on our part as the data creators to share the whole data package.

Data Sharing Complications

In spite of the benefits of data sharing, there is a gap between expectations and reality that is important to address. What would prohibit or limit our ability to make the WLD data available and how could these issues be overcome?

Unpredictable questions from data users, whose professional background primarily determines how they will interpret and understand the data pieces, is one data sharing concern (Stodden, 2010). For example, potential users may even ask about IT issues, e.g., file opening and associated software installation. In that case, the data producers fear drowning in endless support, which may degrade their efficiency.

In addition, it is hard to imagine who will attempt to reuse the data and if that reuse is scientifically sound, which sometimes requires a case-by-case consideration (Mayernik, 2011; Borgman, 2012). Misinterpretation of the data could unexpectedly discredit the data, the data creators, and even the community (Hilgartner, 1997; Hilgartner and Brandt-Rauf, 1994). Because of this, the farther distance the reuser is from the origin of the data creation, the greater risk of misinterpretation, and the more detailed the documentation that is required.

In fact, some researchers may even consider data sharing a 'double-edged sword' since the more shared, the higher the risk that mistakes and errors get recognized. While in and of itself, exposure of mistakes is a beneficial for science as a whole, some disciplines are highly competitive and the concern is that such discoveries may cause disproportionally harsh criticism and even jeopardize one's reputation (Cope, 2015). Therefore, self-protection is a common psychological factor.

There are many reasons that researchers don't want to share data prior to publication. For example, they may need to complete internal quality control to ensure that the data is valid to begin with. Other reasons may be partially due to the possibility that extra work might be needed in response to peer-review publication, which can result in the changes in dataset; and partially due to the worries over difficulties in intellectual property as well as lack of incentives and rewards for doing so (Stodden, 2010).

Furthermore, making data publicly available may result in loss of control, and data may become disconnected from the creators and/or the metadata (Borgman, 2012). Without consistent standards and metrics for data citation, data sharing concerns may dissuade researchers from sharing their data in such a competitive academic world. The WLD data is not immune from these concerns since the current project status that at least one journal paper is not published yet. A compromise may be to publish the dataset under an embargo until the official release of the paper.

What can be shared?

A realistic issue about data sharing is that data may come from multiple sources and have different policies and licenses for use (Stodden, 2010). Plus, some datasets evolve through a 'chain of custody,' i.e., multiple hands and a series of cleaning steps, which is very common for academic research teams. Even if some of the data producers wish

share the data, they must obey individual sharing agreements from different data sources, which is why data sharing is complicated and usually taken case-by-case.

Obviously, a relatively optimistic scenario is that the data can be distributed but under some restriction on remixing, changes, and commercial usage, such as Creative Commons[18] copyright licenses. However, since unbounded data sharing is not yet common-place, in some cases researchers consider datasets for 'internal use only,' and sharing is expected to only occur among group members or with collaborators.

In the WLD study, the LIDAR data regarding to the topography information falls into such category. The large high-resolution WLD topography data was essential for the results of our study, which underscores its value, but a separate research group collected this data. The original creator's sharing preference, which is delayed public disclosure, must be respected.

This situation highlights how retrospective attempts to discuss data sharing once a study is well underway can be awkward, complicated, and even create tension in professional relationships. Ignoring these dynamics is not realistic. However, while we anticipate researchers will change their practices towards more open sharing in the future, it raises an important question for us in the short term: if some of the input data cannot be shared, does that mean the output data generated from such input shouldn't be shared either? After discussion and mutual agreement with the data owners, one potential solution is to set an embargoed period for the data.

Furthermore, these conundrums are not exclusive to the sharing of data, specifically. As research technology progresses, there is interest in openly sharing not only research data, but tools and computational models as well since software and hardware implementation are crucial for resultant data production but rarely mentioned (Stodden, 2010).

Making code open will accelerate the pace of discovery by influencing and expanding the way in which the community generates and consumes data (Pejša et al., 2014). As noted above and in Figure 1, the WLD data package doesn't contain just data alone, but also the associated data analysis methods, scripts, as well as the analysis results. In this specific study, all the data, code, and documentation (e.g. log files and readme files) need to be shared to truly be transparent and enable verification and extension of our study. Without the numerical model, the key processes that bridge the input and output components would remain a mystery.

Likewise, without sharing the output, there is no way to reproduce the numerical results for verification purpose, and each step is opaque without documentation. However, if all of the data, code, and other scholarly objects are not offered under the same licenses, such interoperable data package becomes even more challenging.

Where to share?

Another realistic issue is the fact that the whole data package in this WLD study is very large (over 70 GB with single data file sizes of over 5 GB), which exceeds the size limit for most data repositories.

Furthermore, since data loss is an enormous risk (Berman, 2014), there is difficulty in choosing a trustworthy dissemination platform that can provide a sustainable and secure infrastructure. On the one hand, much of the data that needs to be preserved is not currently protected or even shared so it seems any repository would be better than none. On the other hand, data repositories cannot 'preserve' without a better sense of what contributes to the value of these datasets. For example, data without enough metadata for discovery or without complete documentation might cause difficulty in

[18] Creative Commons copyright licences: https://creativecommons.org/

understanding the dataset and thus are probably less frequently to be used than those that have more metadata and documentation.

Solutions to these problems are needed for better quality control and proper reuse (Stodden and Miguez, 2013). A sustainable data ecosystem is needed to enable cooperation and efficiency which further requires active engagement with academic communities (Bourne et al., 2015).

Conclusions

This paper aimed to discuss how the Wax Lake Delta computational oceanography research could be made more valuable through sharing. Although publications usually focus on major scientific findings, we offered a unique perspective by focusing instead on the 'supporting evidence' – the data. The first half of the paper describes the research project and data generation that makes use of the computational ocean model ECOM-si to determine the circulation and wind driven flows in the delta during cold fronts, while the second half discusses open data in the context of this specific study.

Here our goal was to go beyond the concept of making data visible and accessible to emphasize the real and complex issues that data sharing evokes. By describing realistic considerations that complicate data sharing and reuse in practice, some issues are raised. For example, researcher concerns, the storage gap, infrastructure sustainability and reliability, licensing for a mixture of dataset and code, and sharing policies for indirectly generated datasets, etc.

Despite the challenges and the fact that effective data sharing is non-trivial, the sharing of the Wax Lake Delta data has the potential to enable validation and reproducibility of the current study and also enable more efficient future research through reuse of the data described here.

Our sharing solution is to divide the whole dataset into seven separate deposits, with no publication delay for six of the seven deposits and a last deposit with a one year embargo, which will make the metadata publicly available but restrict the data files from access until the peer-reviewed journal paper get published (Zhang, 2016; Zhang and Li, 2016a; Zhang and Li, 2016b; Zhang and Li, 2016c; Zhang and Li 2016d; Zhang, Li and Braud, 2016; Zhang and Li, 2017).

Importantly, the embargo option does ensure that the data will be available at the time of article publication. While we have successfully navigated this specific data sharing situation for the WLD study, we note that some scenarios will be even more complicated. For example, a new term, 'non-consumptive research[19]', has been coined in the digital humanities, which defines a situation when data is use-protected or copyrighted and thus cannot be made publicly available. Although the term arose in response to the challenges encountered during text mining studies, there may be similarities to other situations where data cannot be shared broadly due to ownership or sensitivity issues.

As European Commissioner for Digital Agenda, Neelie Kroes, said: "Data is the new gold" (Kroes, 2011). There has been growing awareness and progress towards contributing open research data and advancing discovery by using past data. Current challenges are just reminders that there is a great deal of work yet to be performed. Research data management, from planning to dissemination to reuse, is the collective

19 The Data Capsule for Non-Consumptive Research: Final Report: https://scholarworks.iu.edu/dspace/bitstream/handle/2022/19277/HTRCSloanReport_ScholarWorks.pdf?sequence=1&isAllowed=y

coordination responsibility among a broad community including researchers, institutions, libraries, journals, scientific societies, funding agencies, archivists, legal and policy environment. As a data producer, we want to contribute our data products and make them valuable to peers. As a data user, we hope that we have reliable and efficient access to other's data that could enable new research. As a data curator, we are devoted to providing the best data management and policy practices to make each scholar's work be of value over time.

References

Allen, Y.C., Couvillion, B.R., & Barras, J.A. (2012). Using multi-temporal remote sensing imagery and inundation measures to improve land change estimates in coastal wetlands. *Estuaries and Coasts, 35*(1), 190-200. Retrieved from http://link.springer.com/article/10.1007/s12237-011-9437-z

Berman, F. (2014). Despite growing data, infrastructure stands still: Why the gap puts research data at risk. IEEE The Institute. Retrieved from http://theinstitute.ieee.org/ieee-roundup/members/achievements/despite-growing-data-infrastructure-stands-still

Blumberg, A.F. (1994). A primer for ECOM-si. Technical Report. HydroQual Inc., Mahwah, N.J.

Borgman, C.L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet.* Cambridge, MA: MIT Press.

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/abstract

Borgman, C.L., Wallis, J.C., Mayernik, M.S., & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. *Proceedings of the 7th Annual International ACM/ IEEE Joint Conference on Digital libraries (JCDL '07)*. New York: ACM. Retrieved from http://dl.acm.org/citation.cfm?doid=1255175.1255228

Bourne, P. (2005). Will a biological database be different from a biological journal? *PLoS Comput Biolology, 1*(3), e34.

Bourne, P.E., Lorsch, J.R., & Green, E.D. (2015). Perspective: Sustaining the big-data ecosystem. *Nature, 527*(7576), S16-S17. Retrieved from http://www.nature.com/nature/journal/v527/n7576_supp/full/527S16a.html

Butler, D. (2006). Mashups mix data into global service: Is this the future for scientific analysis? *Nature, 439*(7072), 6–7. Retrieved from http://www.nature.com/nature/journal/v439/n7072/full/439006a.html

Carle, M.M. (2013). Spatial structure and dynamics of the plant communities in a prograding river delta. Wax Lake Delta, Atchafalaya Bay, Louisiana. Baton Rouge, Louisiana: Louisiana State University, Ph.D. dissertation, 139p.

Carlson, J. (2014). The use of life cycle models in developing and supporting data services. In Joyce M. Ray (Eds.), *Research Data Management: Practical Strategies for Information Professionals*, 63-86. West Lafayette, IN: Purdue University Press.

Cope, J. (2015). Why not to share your data: The "crackpot" argument. eRambler Blog. Retrieved from http://erambler.co.uk/blog/crackpot-argument-against-data-sharing/

CPRA (Coastal Protection and Restoration Authority). (2012). *Louisiana's comprehensive plan for a sustainable coast.* Baton Rouge, Louisiana: CPRA.

Day, J.W., Kemp, G.P., Reed, D.J., Cahoon, D.R., Boumans, R.M., Suhayda, J.M., & Gambrell, R. (2011). Vegetation death and rapid loss of surface elevation in two contrasting Mississippi delta salt marshes: The role of sedimentation, auto compaction, and sea-level rise. *Ecological Engineering, 37*, 229–240.

Delft3D-RGFGRID. (2013). Generation and manipulation of curvilinear grids for Delft3D-FLOW and Delft3D-WAVE, Hydro-Morphodynamics, Deltares. Delft3D-RGFGRID User Manual. Version: 4.00.30932.

Edmonds, D. & Slingerland, R. (2007). Mechanics of river mouth bar formation: Implications for the morphodynamics of delta distributary networks. *Journal of Geophysical Research, 112*, F02034. doi:10.1029/2006JF000574

Fischer, B.A. & Zigmond, M.J. (2010). The essential nature of sharing in science. *Science and Engineering Ethics, 16*(4), 783–799.

Hilgartner, S. (1997). Access to data and intellectual property: Scientific exchange in genome research. Intellectual Property Rights and the Dissemination of Research Tools in Molecular Biology: Summary of a Workshop Held at the National Academy of Sciences (pp. 28–39). Washington, DC: National Academies Press.

Hilgartner, S. & Brandt-Rauf, S.I. (1994). Data access, ownership and control: Toward empirical studies of access practices. *Knowledge, 15*, 355–372.

Holdren, J.P. (2013). Increasing access to the results of federally funded scientific research. Retrieved from White House, Office of Science and Technology Policy website: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Kemp, G.P., Wells, J.T. & van Heerden, I.LI. (1980). Frontal passages affect delta development in Louisiana. *Coastal Oceanographic and Climatological News, 3*, 4-5.

Keralis, S.D.C. (2012). The Denton Declaration: An open access manifesto. Retrieved from https://openaccess.unt.edu/denton-declaration

Kroes, N. (2011). *Data is the new gold.* Opening Remarks, Press Conference on Open Data Strategy, Brussels, Belgium.

Mayernik, M.S. (2011). Metadata realities for cyberinfrastructure: Data authors as metadata creators. Unpublished doctoral dissertation. UCLA, Los Angeles. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/14943/metadatarealities.pdf?sequence=2

Pejša S. (2012). NEES data repository [Data set]. Retrieved from https://nees.org/resources/4345

Pejša, S., Dyke, S.J., & Hacker, T.J. (2014). Building infrastructure for preservation and publication of earthquake engineering research data. *International Journal of Digital Curation, 9*(2), 83-97. doi:10.2218/ijdc.v9i2.335

Pepe, A., Mayernik, M.S., Borgman, C.L., & Van de Sompel, H. (2010). From artifacts to aggregations: Modeling scientific life cycles on the semantic web. *Journal of the American Society for Information Science and Technology, 61*(3), 567–582.

Piwowar, H., & Vision, T.J. (2013). Data reuse and the open data citation advantage. *Peer J. PrePrints, 1*, e1v1. doi:10.7287/peerj.preprints.1v1

President's Council of Advisors on Science and Technology (PCAST) Working Group on Biodiversity Preservation and Ecosystem Services. (2011). *Sustaining environmental capital: Protecting society and the economy.* Government Printing Office, Washington D.C. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast_sustaining_environmental_capital_report.pdf

Roberts, H.H., Adams, R.D., & Cunningham, R.H.W. (1980). Evolution of sand-dominant subaerial phase, Atchafalaya Delta, Louisiana. *American Association of Petroleum Geologists Bulletin, 64*, 264-279.

Roberts, H.H., Coleman, J.M., Bentley, S.J., & Walker, N. (2003). An embryonic major delta lobe: a new generation of delta studies in the Atchafalaya-Wax Lake delta systems. *GCAGS/GCSSEPM Transactions, 53,* 690-703.

Roberts, H.H., DeLaune, R.D., White, J.R., Li, C., Sasser, C.E., Braud, D., Weeks, E. & Khalil, S. (2015). Floods and cold front passages: Impacts on coastal marshes in a river diversion setting (Wax Lake Delta Area, Louisiana). *Journal of Coastal Research, 31*(5), 1057-1068.

Roberts, H.H. & Sneider J. (2003). Atchafalaya-Wax Lake Deltas the new regressive phase of the Mississippi River delta complex: Guidebook. In *A Field Seminar for the GCAGS Convention* (pp.71).

Stodden, V. (2010). *The scientific method in practice: Reproducibility in the computational sciences.* MIT Sloan Research Paper No. 4773-10. doi:10.2139/ssrn.1550193

Stodden, V. & Miguez, S. (2013). *Best practices for computational science: Software infrastructure and environments for reproducible and extensible research*. Available at SSRN 2322276.

Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS ONE, 10*(8), e0134826.

Walker, N.D., & Hammack, A.B. (1999). Impacts of river discharge and wind forcing on circulation, sediment distribution, sediment flux and salinity changes: Vermilion/Cote Blanche Bay System, Louisiana, Final Report. U.S. Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS.

Wellner, R., Beaubouef, R., Van Wagoner, J., Roberts, H.H., & Sun, T. (2005). Jet-plume depositional bodies: The primary building blocks of Wax Lake Delta. *Transactions of the Gulf Coast Association of Geological Societies, 55,* 867–909.

Whitlock, M.C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology and Evolution, 26*(2), 61–65.

Zhang, Q. (2015). Numerical simulation of cold front-related hydrodynamics of Wax Lake Delta. Dissertation. URN: etd-05132015-081026.

Zhang, Q. (2016). Public agency data of the Wax Lake delta. University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-4871125_V1

Zhang, Q. & Li, C. (2016a). Model dataset for the Wax Lake delta. University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-9511904_V1

Zhang, Q. & Li, C. (2016b). Current data of the Wax Lake delta. University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-1752285_V1

Zhang, Q. & Li, C. (2016c). Bathymetry data of the Wax Lake delta (2012-12-01). University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-4810873_V1

Zhang, Q. & Li, C. (2016d). Bathymetry data of the Wax Lake delta (late 2012). University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-1001307_V1

Zhang, Q., Li, C., & Braud, D. (2016). LIDAR data for the Wax Lake delta. University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-3764213_V1

Zhang, Q., & Li, C. (2017). Meterology and ocean data collected at LSU WAVCIS Lab. University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-2436375_V1