Design, construction, and application of a regional ocean database: A case study in Jiaozhou Bay, China

Yuan Yuan, Isabel Jalón-Rojas, Xiao Hua Wang, Dehai Song

LIMNOLOGY
and
OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 2019

doi: 10.1002/lom3.10304

Design, construction, and application of a regional ocean database: A case study in Jiaozhou Bay, China

Yuan Yuan,[1,2]* **Isabel Jalón-Rojas**,[1,2] **Xiao Hua Wang**,[1,2] **Dehai Song**[3]

[1]School of Physical, Environmental and Mathematical Sciences, The University of New South Wales, Canberra, Australian Capital Territory, Australia

[2]The Sino-Australian Research Centre for Coastal Management, The University of New South Wales, Canberra, Australian Capital Territory, Australia

[3]Key Laboratory of Physical Oceanography, Ministry of Education, at Ocean University of China, and Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

Abstract

Access to geophysical data poses a challenge in Oceanography and Marine Sciences. Ocean databases are an efficient solution to storing, standardizing, validating, and sharing marine and environmental data. However, ocean databases of regional seas and coastal systems are still rare. In this article, we describe a regional ocean database for Jiaozhou Bay, a coastal system affected by increasing anthropogenic activities, such as land reclamation, a cross-bay bridge construction, and land-based pollutant discharges. Understanding the impact of human activities on Jiaozhou Bay requires a comprehensive range of research on biogeochemical and physical changes, and on the current environmental state of the bay. The Jiaozhou Bay database described here aims to inform this key research and also to serve as a reference for the future development of regional ocean databases. It includes three kinds of variables: on-site, satellite, and modeling data. The database structure is unified and incorporates file names, header information (including metadata), measured values and data-quality flags, which facilitate data exchange and use. We propose a system of data-quality flags, based on 13 checks that encompass quality control of the different kinds of variables (e.g., satellite 2D data and on-site profile data). Three examples, one for each type of data, illustrate the use and value of the Jiaozhou Bay database in better understanding the oceanographic characteristics of this bay and the impact of intensive anthropogenic activities on its marine environment.

Ocean databases: an overview

Ocean databases store marine and environmental data from different disciplines and sources. They are valuable data management solutions, inasmuch as accessing, combining, verifying and analyzing these multi-disciplinary datasets present a significant challenge for scientists and managers (Jian et al. 2018). The first ocean database, Crubase, contains cruise data of up to 320 measured variables from 3000 to 4000 oceanographic stations (Vladimirov 1995; Vladimirov and Miroshnichenko 1997). It also includes temporal and geolocation information for every set of variables, but data-quality information is not given. Further marine database management systems have been developed for the Black Sea (Vladimirov and Miroshnichenko 1997; Vladimirov et al. 1999) and the Aral Sea (Lyubartsev et al. 2004). These systems were designed as interactive software to store and analyze multi-disciplinary oceanographic variables such as temperature, salinity, dissolved oxygen, chlorophyll *a* (Chl *a*), and primary production. Each oceanographic station was regarded as a basic data organization unit. Header information, measured values, and data provenance were stored. Numbers from 0 to 9 were used to flag the data quality, representing different levels of quality control.

The increasing interest in ocean research has motivated a growth in oceanographic data from field measurements, satellite observations, and other sources. Therefore, more-complex and diverse ocean databases have been created, such as the Australian Ocean Data Network (AODN) and the Ocean Data Portal (ODP). AODN is an interoperable online network of marine and climate data resources collected from ocean-going ships, autonomous vehicles, moorings, satellites, and other platforms (portal.aodn.org.au/). ODP is a project supported and implemented by the International Oceanographic Data and Information Exchange (IODE) program, and provides multi-disciplinary access to global ocean observations in real-time and delayed modes (www.oceandataportal.org/).

Because of its collection of heterogenous and multi-variable ocean data, the World Ocean Database (WOD) is a significant

*Correspondence: yuan.yuan1@student.adfa.edu.au

reference for new databases. WOD is the largest, most comprehensive collection of scientific information about global oceans (Levitus et al. 2013). It provides historical ocean-profile data archived from the National Oceanographic Data Centre (NODC) Ocean Climate Laboratory (OCL). These data were collected from buoys, ships, gliders, and other instruments. The first database version released by NODC/OCL was the World Ocean Atlas 1994, which included vertical profiles of six variables (temperature, salinity, oxygen, phosphate, nitrate, and silicate). The latest version, the World Ocean Database 2013 (WOD13), consists of 27 variables from more than 200,000 oceanographic cruises. WOD13 has multiple units for data organization such as profile, cast, station, and cruise; data can be organized based on depth, time, geographical location, or research project (Boyer et al. 2013). WOD13 also incorporates header information such as time, location, meteorological data, instrument, project, and originator's methods. Quality flags are given, based on statistical analysis of individual values and profiles; the original flags provided by data contributors are also included. WOD13 is a professional tool for collecting and distributing worldwide ocean data. However, WOD13 does not focus on regional seas; there are only sparse, or even no, WOD13 data in coastal seas due to the limited data collection.

Despite the contemporary development of ocean databases and the increasing concern about coastal issues, comprehensive databases for regional seas and coastal systems are not very common. Aside from the Black Sea and Aral Sea databases, which use relatively simple methods to organize data and flag data quality, most data compilations are related to a specific research interest, type of data or research project. For example, the EBSSED database contains all available data on superficial sediment textures in the eastern Bering Sea shelf study area (Smith and McConnaughey 1999). The GIS database provides baseline information on the different factors influencing coastal erosion processes and the value of assets at risk (www.eurosion.org/database/). Dynalit is a long-term hydromorphological database of high-quality meteorological data and observed coastal parameters (coastline, bathymetry, sedimentary fluxes, turbidity, sea level, and sea states) at 28 French coastal sites (www.insu.cnrs.fr/node/3957). Therefore, there is a lack of comprehensive databases for small marine systems that include a wider range of cross-disciplinary variables from different sources.

Jiaozhou Bay

Jiaozhou Bay (JZB) is a shallow semi-enclosed bay, located on the southern coast of the Shandong Peninsula (east China, Fig. 1a). It has a length of 33 km, a width of 28 km, and an average water depth of 7 m (Liu 1992). The greatest depth is 64 m, in the middle channel (Fig. 1b).

The bay is surrounded by Qingdao, a city with increasing economic development. As a consequence, JZB has been significantly affected by continuously increasing anthropogenic activities. Over the last century, land reclamation has significantly modified the coastline of JZB (Fig. 1b), decreasing the bay area by 40% (Chen and Chen 2012). From 2007 to 2011, a 27-km-long bridge (called the JZB Bridge here) was built across the bay. In addition, JZB also receives most of its pollutants from Qingdao. Waste-water discharge into the bay increased from 2.1 to 5.6 tonnes per annum per person between 1990 and 2012 (Liang et al. 2015). The annual discharge of terrigenous dissolved inorganic nitrogen, the dominant pollutant, increased by 10 times from the 1980s to the 2000s (Wang et al. 2006; Qian et al. 2009; Dong et al. 2010). This nutrient-rich water can cause algal blooms, which potentially lead to oxygen depletion.

JZB is therefore a potential natural laboratory in which to study the impact of intensive anthropogenic activities on the marine environment of embayments. The biogeochemical and physical changes in this system and its current environmental state are key issues that require diverse data. Many marine variables, such as temperature, salinity, water level and Chl *a* concentration, are accessible from different sources, including satellite, on-site, and modeling datasets. However, these data can be difficult to obtain and validate.

Aims

The above review indicates that (1) there is a lack of regional ocean databases (ROD) that efficiently combine and validate multi-disciplinary datasets, and (2) addressing research and management issues in coastal systems such as Jiaozhou Bay requires data from different sources, which can pose a challenge. The overall objective of this study is to design and create a ROD specific to JZB, based on the collection of quality-controlled marine data. The specific aims of the JZB database are twofold. First, to help assess the current state of the JZB marine environment from a cross-disciplinary and cross-variable view. Second, to serve as a reference for the future development of RODs, providing a methodological framework for and an example of data collection, database organization and data-quality flag systems.

The JZB database involves variables from three kinds of sources: on-site, satellite, and model. We propose a data storage and organization adapted to these different types of variable for easy use and quick access, as well as a data-quality flag system that encompass quality control of all the different variables. Three applications then illustrate the practicability and relevance of the JZB database.

Data and methods

Data sources

As in other marine environments, the available data for JZB can be classified into three categories: on-site (observed), satellite, and modeling data.

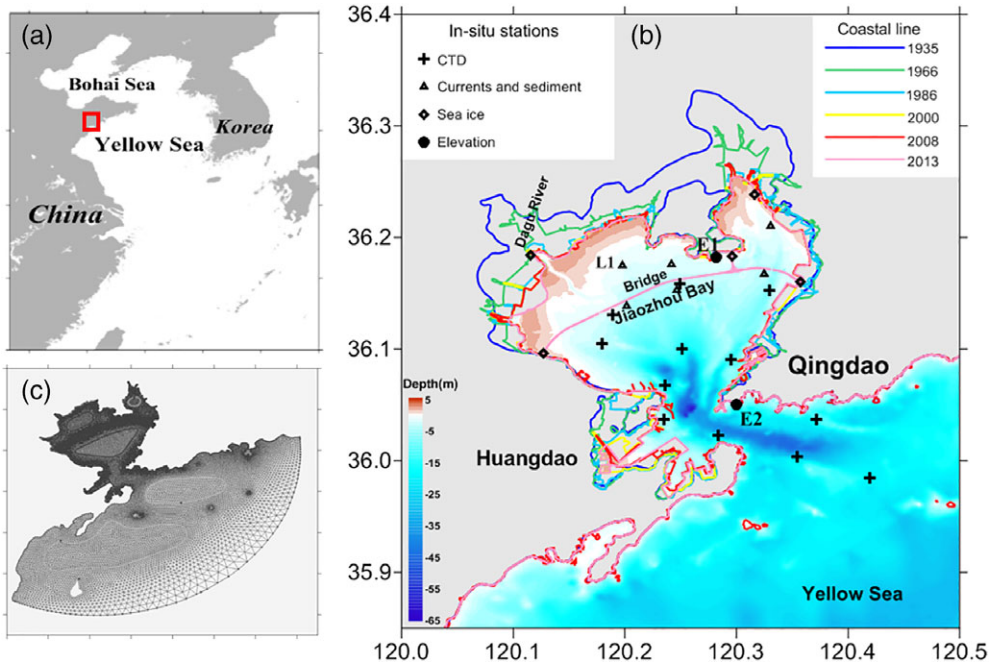


Fig. 1. (**a**) Location of JZB on the Yellow Sea coast. (**b**) Bathymetry map of JZB showing the coastline evolution from 1935 to 2013 (colored lines) and observation stations for: conductivity-temperature-depth (CTD, +), currents and sediment (Δ), sea ice (◊), and elevation (●), modified from Yuan et al. (2016). (**c**) JZB model mesh and grid distribution.

On-site data

At present, the JZB database includes three on-site datasets but it will be updated with further measurements. The first dataset is ship-based measurements to investigate seasonal variability in the marine environment in JZB, conducted in January, April, July, and October 2013. A ship-mounted conductivity-temperature-depth (CTD) sensor measured temperature, salinity, Chl *a* concentration, turbidity, pH, and dissolved-oxygen concentration at multiple depths at 12 stations (Fig. 1b).

The second dataset consists of current velocities and suspended-sediment concentrations measured by a ship-mounted rotor current meter (RCM-9) at six stations (Fig. 1b) during spring and neap tides in November and December of 2013.

The JZB database also includes sea-ice observations. Trained observers recorded the maximum ice thickness, ice type, floe size, and ice concentration at six coastal stations (Fig. 1b) from 2012 to the present.

Satellite data

Satellites provide ocean data covering whole study sites, with a relatively high temporal resolution. A total of six satellite sensors covers the JZB region: the Korean Geostationary Ocean Color Imager (GOCI), the Multi-Functional Transport Satellite (MTSAT), the Moderate Resolution Imaging Spectroradiometer (MODIS), the Advanced Very High Resolution Radiometer (AVHRR), and the Visible Infrared Imaging Radiometer Suite (VIIRS), Himawari-8.

MODIS is aboard two polar-orbit satellites (Terra and Aqua), observing the entire Earth's surface every 1–2 d. MODIS measures in 36 spectral bands, with wavelengths ranging from 0.4 μm to 14.4 μm. As a consequence, information on the various environments, such as the atmosphere, ocean, land and glacier, can be acquired simultaneously. VIIRS is one of a new generation of moderate-resolution imaging, designed to extend and improve upon a series of measurements initiated by MODIS. Both MODIS and VIIRS provide ocean color products and also the sea surface temperature (SST) due to their multiple spectral bands (36 for MODIS and 22 for VIIRS).

GOCI is the world's first geostationary ocean color satellite to cover the northeast Asian region (Ryu et al. 2012). GOCI-derived high-level products can be generated from the GOCI data-processing system (GDPS), based on given algorithms.

The Group for High Resolution Sea Surface Temperature (GHRSST) is a collection of global, multi-sensor, high-resolution SST products. It involves different sensors such as MODIS, VIIRS, AVHRR, and MTSAT. AVHRR is a radiation-detection imager aboard low-earth-orbit satellites, managed by the National Oceanic and Atmospheric Administration (NOAA). It detects the global cloud cover and surface temperature. MTSAT (MTSAT-1R and MTSAT-2) is a series of geostationary weather satellites operated by the Japan Meteorological Agency (JMA). It provides imagery over the Asia-Pacific region.

Table 1 compares the characteristics of these six satellite sensors and lists the products available for the JZB region: Chl *a* concentration, suspended-sediment concentration (SSC),

Table 1. Comparison of satellite data for JZB.

Satellite		GOCI	MTSAT	MODIS	AVHRR	VIIRS	Himawari-8
Spatial resolution		0.5 km	1.25 or 4 km*	0.25, 0.5, or 1 km*	~1 km	750 m	0.5, 1, or 2 km*
Temporal resolution		1 h (8 times a day)	25 min	1–2 d	Twice a day	2 d	10 min
Operational period		13 Jul 2010 to now	28 Jun 2005 to 06 Jul 2015	**Terra:** 24 Feb 2000 to now **Aqua:** 04 Jul 2002 to now	Jun 1979 to now (Jun 2006 for the JZB area)	02 Jan 2012 to now	18 Dec 2014 to now
Available high-level products	Chl *a*	√		√		√	√
	SSC	√					
	WQL	√					
	PP	√		√			
	WCV	√					
	SST		√	√	√	√	√
	SSS			√			

*The resolution depends on the band.

water-quality level (WQL), primary production (PP), water current vector (WCV), SST, and sea surface salinity (SSS).

Modeling data

We applied the Finite Volume Community Ocean Model (FVCOM, Chen et al. 2003) to JZB in order to load the JZB database with hydrodynamic datasets. FVCOM is an unstructured-grid, free-surface, 3D primitive-equation coastal ocean circulation model. The unstructured grid has flexible resolution that readily adapts to complicated topographies at small scales, such as in JZB. The 864 piers of the JZB Bridge (Fig. 1b) were included in model grid cells, as these may have a significant effect on the currents in JZB. A high-resolution grid, especially intensified around the piers, was used (Fig. 1c). Each pier occupied a whole cell, identified as a no-water area. The grid was formed by 85,271 model nodes and 165,617 grid cells, with the finest grid cells of 7 m around the piers and the coarsest grid cells of 2000 m in the open sea. A secondary grid of the same resolution but without the piers was also used to simulate scenarios prior to the bridge construction. Currents were divided into external and internal modes in the mode-split approach used for the solution of the circulation model. The external time step was 0.6 s, the internal time step 1.2 s. Seven sigma layers were used in the vertical direction. The coastline and the bathymetry of 2013 was retrieved, at a spatial resolution of 200 m, from nautical charts and remote-sensing images from Google Maps. Currently the model is a barotropic model forced only by tides, which dominate the hydrodynamics in JZB (Gao et al. 2014). The tide is generally expressed in terms of tidal constituents. The forcing of eight major tidal constituents (M_2, S_2, K_1, O_1, N_2, P_1, K_2, and Q_2) and three shallow-water constituents (M_4, MS_4, and MN_4) were applied at the open boundary. These 11 tidal constituents were provided by the East China Seas Model (Song et al. 2013), and represent most of tidal energy and tidal elevation in the coastal area. The model was calibrated and validated with observations of: (1) tidal elevations during 2016 at two stations (E1 and E2, Fig. 1b); and (2) current velocities during 2013 at six stations (Fig. 1b). Current velocities were recorded as 25-h time series at different depths during both spring and neap tides. The performance of the model was assessed using the correlation coefficient R^2 and the absolute error between model and observation values. This error was calculated for the amplitudes and phases of the four major constituents (M_2, S_2, O_1, and K_1; Fig. 2c).

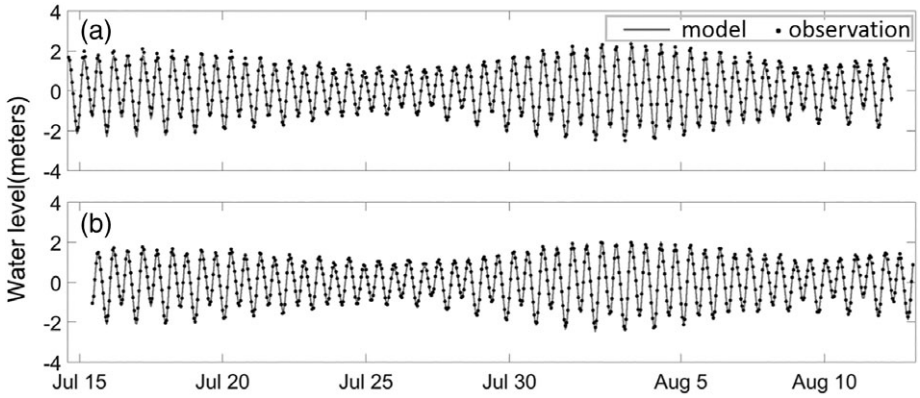
The results showed good agreement between the simulated and observed tidal elevation (R^2 = 0.99 for both E1 and E2; Fig. 2). The differences between the observed and modeled amplitudes and phases of the tidal harmonics were less than 0.02 m and 1°, respectively.

Modeled and observed currents were compared for 33 time series of 25 h each, recorded at six current stations at different depths. These results also showed reasonable agreement between simulation and observations. This is illustrated in Fig. 3 for two tidal cycles at L1 (Fig. 1b) during spring and neap tides (R^2 = 0.79, 0.88, 0.71, 0.69 for Fig. 3a–d, respectively). Overall, 85% of the R^2 values for these 66 comparisons (33 time series each for direction and velocity) were higher than 0.5.

In addition, Gao et al. (2014, 2018,b) developed hydrodynamics, wave, and sediment models for JZB based on the coastline in 2008, which will also provide further modeling data for the JZB database.

Database organization

The JZB database is organized into files. Each file stores the data from the same platform or instrument, at one station for the on-site observations or for the whole domain for the satellite and modeling data. Datasets are first classified into two groups, reference data (R) and normal data. Reference data, such as the coastline location, are used to evaluate the quality



(c)

Constituent		Station					
		E1			E2		
		Model	Observation	Error	Model	Observation	Error
M_2	Amplitude (m)	1.40	1.41	0.01	1.29	1.31	0.02
	Phase (°)	133.68	134.01	0.33	129.89	129.92	0.03
S_2	Amplitude (m)	0.41	0.39	0.02	0.38	0.37	0.01
	Phase (°)	187.13	186.93	0.20	181.50	181.02	0.48
O_1	Amplitude (m)	0.21	0.22	0.01	0.21	0.20	0.01
	Phase (°)	296.63	296.80	0.17	293.42	294.37	0.95
K_1	Amplitude (m)	0.35	0.35	0.00	0.35	0.35	0.00
	Phase (°)	9.38	9.42	0.04	7.41	6.60	0.81

Fig. 2. Comparison of the simulated (lines) and observed (dots) tidal elevations at (**a**) E1 and (**b**) E2 from 15 July to 15 August 2015. (**c**) Comparison of the simulated and observed amplitudes and phases of the four major constituents, with absolute errors.

of normal data. Normal data include on-site (I), satellite (S), and model (M) data. When a new dataset is added to the JZB database, data are converted into the internal format: time is transformed to local time, the values of variables are in the International System of Units, and the number of significant digits is uniform.

Files are labeled so as to facilitate access to the data. The name of each file contains the data source type (R, I, S, or M), the start date, the end date, the instrument, the name of the variable, and the version number. The version number is for cases in which there are later corrections to a file. If more than one variable is stored in a file, the word "Multi" appears in the file name. In general, data files are formatted as NetCDF (.nc). For example, the file name *I20150715_20150814_CTD_Multi_Ver00.nc* indicates that data in this file are on-site measurements collected by CTD sensors from 15 July to 14 August 2015, and the file contains more than one variable; the version of this file is the original (Ver00). The file name *S20150101_20151231_GOCI_Chl_Ver01.nc* indicates that the data stored in this file are Chl *a* concentrations retrieved from GOCI from 01 January to 31 December 2015; the file has been corrected once.

Each file is divided into two components: the header (discussed below) and the measurements or simulation data (discussed above). The header has four parts.

1. **Background header:** this contains basic information about the set of observations: date, time, location, meteorological information (if available), and sea-floor depth (if available).
2. **Originator header:** this contains the information provided by originator of the data: instrument, ship (platform), institute, project, originator's units, scales, original quality flags (if available), and method.
3. **Variables:** the names of the variables are listed in this header information, and can include: longitude, latitude, time, temperature, salinity, oxygen, pH, chlorophyll, suspended-sediment concentration, etc.
4. **Data-quality flags:** data-quality flags specific to the JZB database are created (see Section *Data quality control and flags*). A flag value is assigned to each observation.

The information in the header for the file *S20150101_20151231_GOCI_Chl_Ver01.nc* (mentioned above) is shown in Fig. 4.

Data quality control and flags

Data quality control is a fundamental requirement in the development of a database in order to identify incorrect or inconsistent data. Multiple tests are essential to detect

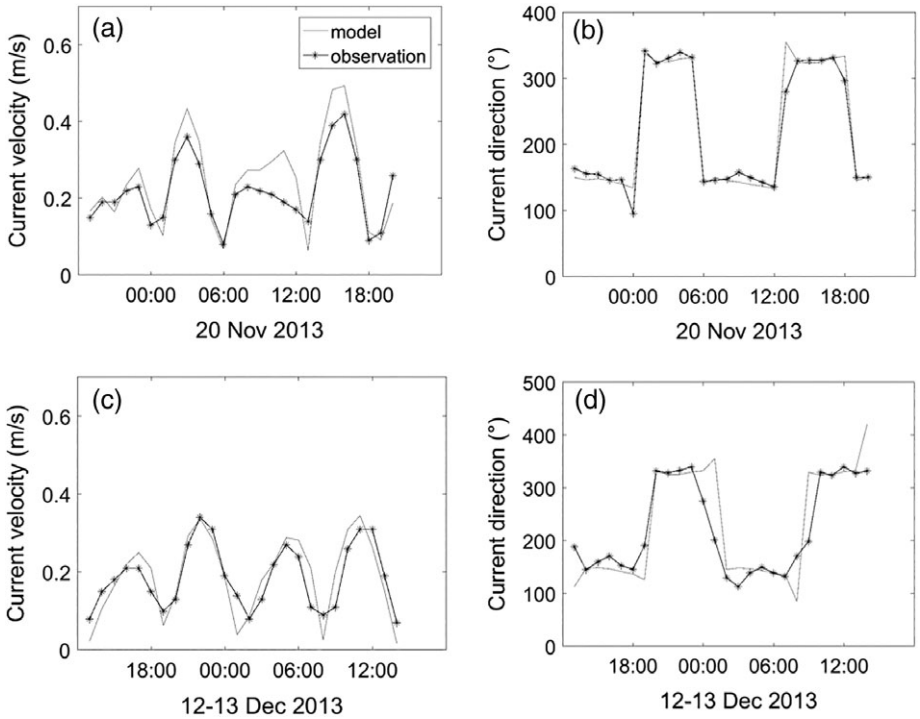


Fig. 3. Comparison of simulated (lines) and observed (asterisks) surface current velocity (**a**, **c**) and direction (**b**, **d**) at station L1 (Fig. 1b) during spring (**a**, **b**) and neap (**c**, **d**) tides in November and December 2013.

potential errors. Data failing the quality-control tests are not removed but marked with standardized quality-control flags. The originator's flags are also provided in the originator header when possible (Section *Database organization*).

Most quality-flag systems in existing ocean databases are too simple (Vladimirov and Miroshnichenko 1997; Vladimirov et al. 1999; Lyubartsev et al. 2004), partly due to the difficulty entailed in designing a comprehensive and efficient data quality control system. Ocean databases may contain different kinds of variables from different sources, as in the JZB database; not all checks are relevant for all data, which may result in inconsistent flags. For instance, satellite data are usually distributed over the sea surface, forming a two-dimensional dataset, whereas the CTD data usually comprise measurements at different depths, forming profile datasets. Depth-dependent checks are not relevant for sea-surface observations, and checks for time series should not be applied to profile datasets. A uniform quality-flag system for the whole database, avoiding these problems, was therefore necessary.

A quaternary-flag system and 13 quality-control checks are proposed to encompass the quality control of all the different variables. These quality checks are listed in Table 2 and explained below.

Location check

The location check verifies if the reported position of an observation is located in the JZB area. If an observation is not located in the JZB area, subsequent checks (e.g., sea/land check, standard-deviation check) will be not applied.

Sea/land check

Reference data involving coastline evolution in recent decades are used to determine whether an observation is located in the water (check passed) or on land (check failed). This check is required for observations very near the coast, and is useful, for example, when ships provide observations while located in tidal flats. Satellite data at the JZB Bridge location are marked as land in order to warn users about a possible contamination of the data (see discussion in Section *satellite-data application*).

Range check

The range check detects potential outliers by comparing the values of each variable with set extreme values. These extreme values are set large enough to encompass variations at different time scales, including seasonal and interannual variability, using literature values.

Standard-deviation check

Statistical checks identify statistically questionable values at different time scales: monthly, seasonal, and annual. The monthly standard-deviation check compiles all values for a given variable for every month (irrespective of the year or instrument) into an integrated dataset, and calculates simple statistics (mean, standard deviation and number of observations). A value more than five standard deviations from the

```
filename                          = 'S20150101_20160101_GOCI_Chl_Ver02.nc'
Background header:
        start_time                = '20150101021640'
        end_time                  = '20151231041647'
        northernmost_latitude     = '36.399'
        southernmost_latitude     = '35.801'
        easternmost_longitude     = '120.499'
        westernmost_longitude     = '120.001'
        spatial_resolution        = '500 m'
Originator header:
        platform                  = 'COMS'
        sensor                    = 'GOCI'
        algorithm                 = 'OC2'
        originator_chlorophyll_units = 'mg/m^3'
        originator_quality_flags
                size :            1095 × 11873
                dimensions :      time × location
Variables:
        time
                size :            1095 × 1
                dimensions :      time
                attributes :
                                  long_name   = 'time'
                                  units       = 'seconds'
        lon
                size :            1 × 11873
                dimensions :      location
                attributes :
                                  long_name   = 'longitude'
                                  units       = 'degrees'
        lat
                size :            1 × 11873
                dimensions :      location
                attributes :
                                  long_name   = 'latitude'
                                  units       = 'degrees'
        chl
                size :            1095 × 11873
                dimensions :      time × location
                attributes :
                                  long_name   = 'chlorophyll a concentration'
                                  units       = 'mg/m^3'
                                  _fillvalue  = '-999'
Data quality flags:
        chl
                size :            1095 × 11873
                dimensions :      time × location
```

Fig. 4. Information for in the header of the file *S20150101_20151231_GOCI_Chl_Ver01.nc.*

mean is considered as a potentially incorrect value. The check is therefore flagged as "fail." Similarly, seasonal and annual standard-deviation checks are implemented.

Time-inversion check

A time inversion occurs when an observation has a time record ahead of the observation directly preceding it. The second observation involved in a time inversion fails the check.

Spike check

A single observation quite different to adjacent ones in a time series is regarded as a spike. A spike is characterized by a sharp change in both the magnitude and the gradient (of the lines joining the successive data points). Checking for spikes in data is a research field in its own right, with many publications addressing the problem. The spike check below uses a relatively simple criterion, and clearly cannot perform as well as more sophisticated methods, but it should be sufficient for flagging spikes in most oceanographic datasets. This will be verified with actual data and more guidance given in its use in a future publication.

The spike check here checks for a spike in every three sequential data points, call them V_1, V_2 and V_3, with the potential spike at V_2. It assumes the data are equally spaced,

Table 2. Definition of quality flags for the JZB database. *Position* is the position of the check digit in the 13-digit quaternary flag, which should be read from right to left.

Data type	Position	Quality test
General data	0	Location check
	1	Sea/land check
	2	Range check
	3	Monthly standard-deviation check
	4	Seasonal standard-deviation check
	5	Annual standard-deviation check
Time-series data	6	Time-inversion check
	7	Spike check
Profile data	8	Location-consistency check
	9	Depth-positive check
	10	Depth-inversion check
	11	Depth-duplication check
	12	Gradient check

with time step δt. Essentially, it checks that the gradients of the lines joining V_2 to each of the other two points are of opposite sign and whether their geometric mean is greater than some set threshold value M_{TH}. If both conditions are true, V_2 is flagged as a spike. The two conditions are combined in the one equation: if

$$(V_2 - V_1)(V_3 - V_2) < -M_{\mathrm{TH}}\delta t^2 \tag{1}$$

the data point V_2 is flagged as a spike.

The (positive) threshold mean slope M_{TH} is set by the user after consideration of the relevant JZB parameters. If M_{TH} is set too small, many normal peaks in the data may be flagged as spikes; if M_{TH} is set too large, actual spikes may not be flagged. If a spike is large enough, the value may also be flagged in the range check.

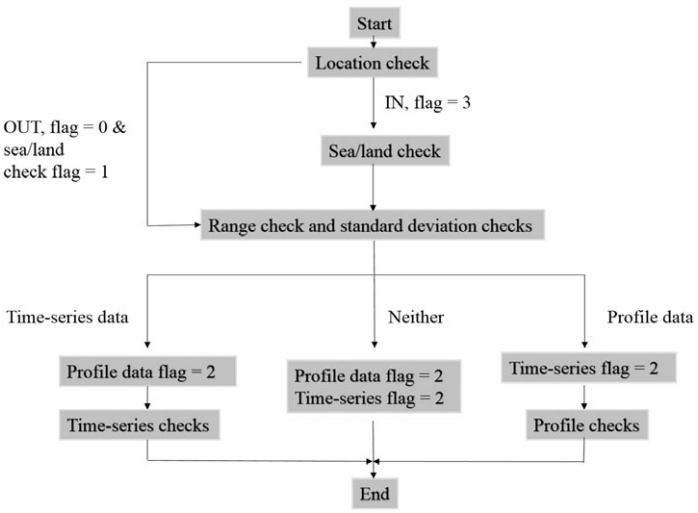


Fig. 5. Flow chart of the data quality control process.

Location-consistency check

The location-consistency check verifies whether the reported position of an observation is consistent with prior positions from the same platform, such as a ship-mounted CTD.

Depth-positive check

The depth-positive check is performed to identify negative depth values, which represent measurements recorded out-of-water during the instrument placement or recovery.

Depth-inversion and depth-duplication check

A depth inversion occurs when a record has a shallower (deeper) depth than the preceding value during a descending (ascending) profile. A depth duplication occurs when an observation has the same depth as the previous one in the profile. The second observation involved in a depth inversion fails this check.

Gradient check

The gradient check detects excessive decreases or increases in a variable over a depth range. The gradient is defined as:

$$\text{gradient} = \frac{V_2 - V_1}{z_2 - z_1} \tag{2}$$

where V_1 is the previous value, V_2 is the checked record, and z_1 and z_2 are the respective depths of V_1 and V_2. A negative (positive) gradient describes a decrease (increase) in the value over the depth. A value will fail the gradient check if its gradient, Eq. 2, exceeds set gradient extreme values, defined after consideration of the JZB.

All the checks operate automatically in order, as illustrated in Fig. 5. The result of each quality check is indicated by a quaternary digit: the quality flag of each value has 13 quaternary digits (one digit per check). Each quality flag is read from right to left (i.e., the digit on the far right corresponds to the location check, *Position* 0 in Table 2). The digits used in the flags and their meaning are listed in Table 3.

However, some users may only be concerned about whether a value passes or fails a quality control. In that case, the quaternary-flag system is too complex. Therefore, a simplified binary version of the quality flags has been added to the quality-flag system (Table 3).

Temperature data from a ship-based survey in JZB during 2016 are used here to illustrate the result of the quality control of profile data (Fig. 6). In particular, the full flag of the fourth entry for a CTD measurement is taken as an example: 3033122333333 (boxed in Fig. 6). The CTD data are usually recorded as profiles, so the profile checks are relevant (flag value = 0, 1 or 3), whereas the time-series checks are not relevant (flag value = 2). Most relevant checks were passed (flag value = 3), except the depth-duplication check that failed (flag value = 0); the location-consistency check was not done because of a missing value (flag value = 1). The corresponding simplified binary flag is 1,011,011,111,111, showing that most check results were acceptable except the depth-duplication check and the location-consistency check.

Table 3. Possible check-digit values and their interpretation for the quaternary and binary flags in the JZB database.

Quaternary	Binary	Interpretation	Result
0	0	Check relevant and done but failed	Failure
1	0	Check relevant but not done	
2	1	Check not relevant	Acceptable
3	1	Check relevant, done and passed	

The storage required for both the quaternary and binary flags is large, as both have 13 digits. In order to reduce the size of each file, both flags are converted to decimal values. The maximum space occupied by decimal flags is eight digits for quaternary flags and four digits for binary flags. Both the decimal-quaternary and decimal-binary flags are placed in a data file for each observation in the JZB database. For the example illustrated in Fig. 6, the decimal flag of the quaternary (binary) value 3,033,122,333,333 (1011011111111) is 54,374,399 (5887). Users can therefore extract either flag according to their requirements.

Applications

Three examples of the use of the JZB database are given below to highlight the convenience and usefulness of taking validated data from a ROD. The data were extracted from target data files in the JZB database. These target files were found by searching the file names for the time and data type that we were interested in. Data to be analyzed are selected depending on their quality flags. The data that failed the location check, the sea/land check, the range check or the monthly standard-deviation check were eliminated for these three applications. Furthermore, the on-site data must also have passed the profile checks. The numerical quality flags are easily processed when database users are choosing which checks should be passed. Figures can be plotted from the selected data using graphing software such as MATLAB or Surfer.

On-site data application

Temperature and salinity data measured from 05 January 2016 to 11 January 2016 by a ship-mounted CTD were incorporated into the JZB database and are analyzed here to describe the current thermohaline distribution in JZB during winter. Figure 7 shows the distribution of temperature and salinity at the surface and at the bottom of the bay. The temperature was vertically homogeneous, as shown by the similar patterns of surface and bottom temperatures (Fig. 7a,c). Temperature increased from the inner to the outer bay. Salinity was spatially homogeneous, with a variation less than 1 psu throughout the whole JZB (Fig. 7b,d). The contour lines show the distribution of saltier water flowing into the bay. This kind of data can be useful for researchers in various fields. On-site data can be used as strong evidence for managers evaluating the ecological state of the bay and formulating regulations. These data can be also used to validate

depth	temperature	full flag(quaternary)	full flag(decimal)	simple flag(binary)	simple flag(decimal)
-0.04	4.342	2220222333333	44216319	1110111111111	7679
0.05	4.337	3333322333333	67088383	1111111111111	8191
0.10	4.328	3333322333333	67088383	1111111111111	8191
0.10	4.334	3033122333333	54374399	1011011111111	5887

Data	Quality test	Position	Quaternary flags	Binary flags	Quality control results
General data	Location check	0	3	1	Pass
	Sea/land check	1	3	1	Pass
	Range check	2	3	1	Pass
	Monthly standard-deviation check	3	3	1	Pass
	Seasonal standard-deviation check	4	3	1	Pass
	Annual standard-deviation check	5	3	1	Pass
Time series data	Time-inversion check	6	2	1	Not relevant
	Spike check	7	2	1	Not relevant
Profile data	Location-consistency check	8	1	0	Relevant but not done
	Depth-positive check	9	3	1	Pass
	Depth-inversion check	10	3	1	Pass
	Depth-duplication check	11	0	0	Fail
	Gradient check	12	3	1	Pass

Fig. 6. Example of quality-control results for a temperature profile.

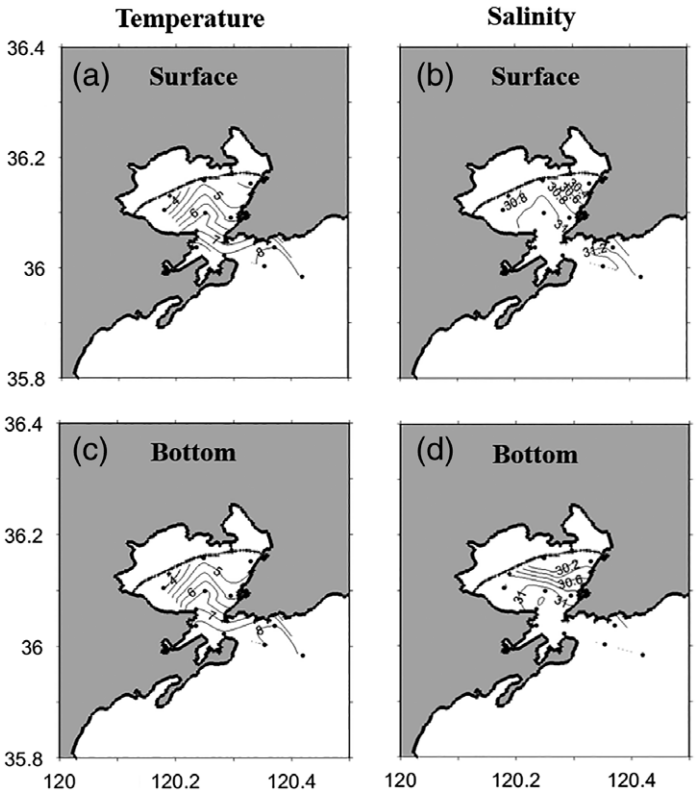


Fig. 7. Surface (**a**, **b**) and bottom (**c**, **d**) temperature (**a**, **c**; °C) and salinity (**b**, **d**; psu) distributions in January in JZB. Dots show the locations of observation stations.

satellite and modeling data, which may be less accurate but usually cover a larger area or longer periods of time.

Satellite-data application

The distribution of Chl *a* concentration based on GOCI data is used here as an example of a satellite-data application. These data are L2 products, downloaded from the GOCI website and incorporated into the JZB database after quality control. Figure 8b shows the distribution of Chl *a* concentration in the bay on 13 September 2017 (after the JZB Bridge construction). The Chl *a* concentration decreased from the top of the bay to the mouth, with high values along the coast.

It is noteworthy that the Chl *a* concentration at the JZB Bridge appears significantly high (Fig. 8b). Even if this were actually the case, there was probably contamination of the reflectance signal (which is related to Chl *a* concentration, see below) by the bridge surface, giving inaccurate values for the Chl *a* concentration. However, the GOCI platform provides no warning about this possible contamination. In order to test this hypothesis, we applied the normalized-difference water index (NDWI), an index used to distinguish open water from soil and terrestrial vegetation in remote-sensing images (McFeeters 1996). The NDWI is given by:

$$\mathrm{NDWI} = \frac{\mathrm{Green} - \mathrm{NIR}}{\mathrm{Green} + \mathrm{NIR}} \tag{3}$$

where Green and NIR are the reflectances in the green and near-infrared bands, respectively. Figure 8a gives an example of the NDWI in JZB, showing that the reflection from the bridge surface affects the reflectance data. Given that Chl *a* concentrations from GOCI are calculated using reflectance data from different bands (Moon et al. 2012), including the green band, we can conclude that Chl *a* data are affected by the bridge. To take this into account, Chl *a* data at the bridge position are flagged as *land* in the sea/land check during the data quality control. However, users make the final decision about the validity of these data. This highlights the usefulness of the JZB database in providing validated data.

Modeling-data application

The recent construction of the JZB Bridge has opened up a debate about its impact on the marine environment of the

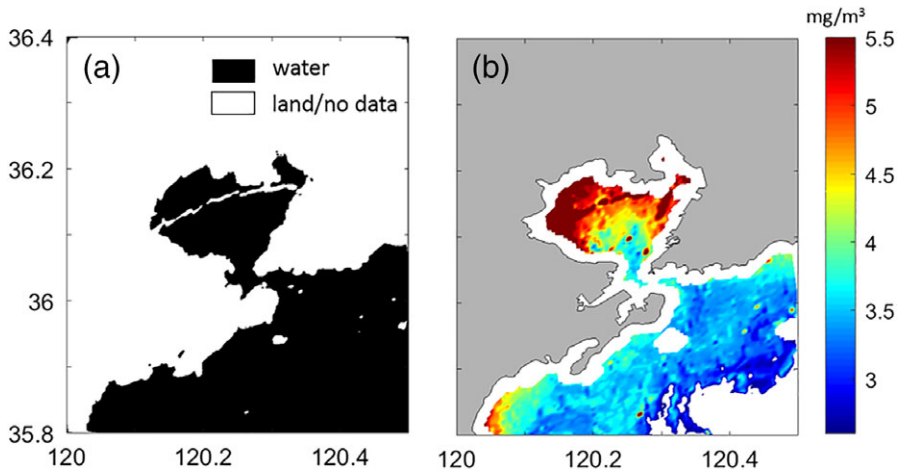


Fig. 8. (**a**) Normalized-difference water index (NDWI) from GOCI data. Black areas represent water, white areas land or zones with no data. (**b**) Distribution of Chl *a* concentration based on L2 products from GOCI on 13 September 2017.

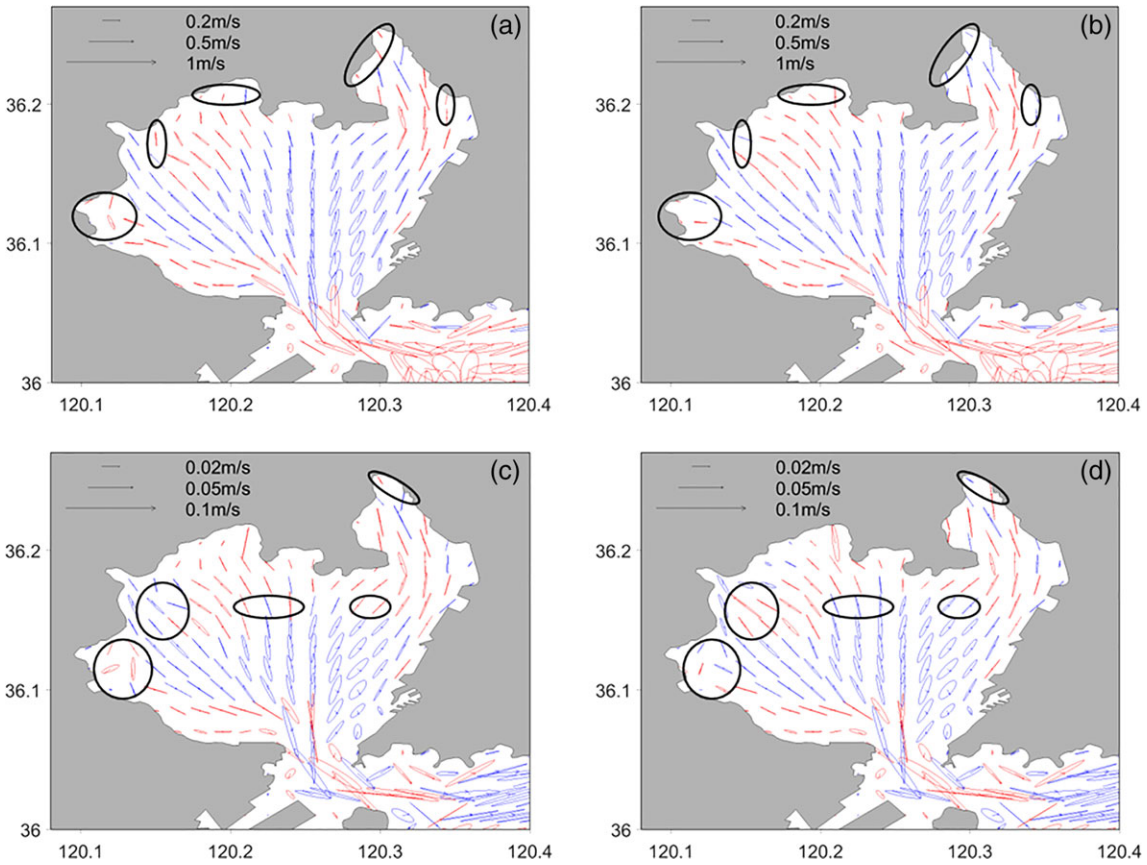


Fig. 9. Comparison of surface tidal ellipses of the M_2 (**a**, **b**) and K_1 (**c**, **d**), based on modeling data from the JZB database for two scenarios: with (**a**, **c**) and without (**b**, **d**) the JZB Bridge. Red and blue ellipses indicate clockwise and counterclockwise rotation, respectively. Black ellipses highlight the differences between the two scenarios.

bay, in particular the hydrodynamics. The hydrodynamic modeling data in the JZB database can be used to evaluate the effect of the bridge on tidal dynamics. We compared two modeling datasets, corresponding to periods before and after the construction of the bridge, with the same forcing; the only difference was that they were generated using different grids, one with and one without the bridge piers. This comparison showed similar distributions of the tidal amplitudes and phases of the four major constituents (M_2, S_2, O_1, and K_1) for the two scenarios (not shown). However, there were some minor differences in tidal currents, as illustrated by the surface tidal ellipses (Fig. 9). M_2 is the dominant constituent in JZB, with the largest tidal currents. Tidal ellipses in the inner bay were relatively narrow, showing that rotation was relatively unimportant. The differences between the scenarios with and without the bridge are highlighted by the black ellipses in Fig. 9, showing the reversed rotation of ellipses and changes in current magnitudes. These changes were mostly located near the coast in the inner bay. This is likely due to the accumulation of small changes in the propagation of tidal energy, especially after passing through the bridge. We conclude that, in general, the bridge construction had only limited effects on the tidal dynamics in JZB.

These results contrast with a previous modeling study performed in JZB. Li et al. (2014) implemented two different grids with the same forcing for the two scenarios: with and without the bridge. Their results suggested a large impact of the bridge on the hydrodynamics. However, the meshes in their two models were different; this may cause unexpected discrepancies in the numerical calculations and highlights the importance of using uniform conditions to obtain modeling data for databases. Implementing numerical modeling is a long and complex task. The incorporation of this kind of data into a ROD facilitates the use of these data by a wide range of users such as policy makers and coastal managers interested in evaluating the effects of planned marine engineering in ocean environments, and researchers not specialized in numerical modeling from multiple areas such as Sedimentology, Biology, Chemistry, Geology, and Engineering.

Conclusion

A regional ocean database has been designed and created for Jiaozhou Bay. This embayment is an excellent system for illustrating the usefulness of a ROD, as its multi-disciplinary

research and management issues require a wide range of validated environmental data. In particular, there is an increasing need for understanding the state of the environment in JZB and the impact of intensive anthropogenic activities on it. The JZB database has been designed to support future research in the bay, for example an assessment of the current state of the JZB marine environment, but also to be a reference for the future development of RODs for other coastal and estuarine regions.

The JZB database is organized into files with a uniform structure. Each file incorporates comprehensive header information, including metadata, the names and values of variables and data-quality flags. An efficient labeling system allows users to easily and quickly access and understand the data. The proposed quaternary flag system includes 13 quality-control checks and encompasses data quality control of all the different formats of data. Binary quality-control flags are also provided for users seeking simplified data-quality information. Each value forming the JZB database is accompanied by both kinds of quality flags in a decimal format, in order to save storage space.

Three applications of the JZB database clearly illustrate the usefulness of RODs in addressing multi-disciplinary research issues. In the particular case of JZB, the database has proven to be an excellent tool in understanding the oceanographic characteristics of the bay and the impact of intensive anthropogenic activities on its marine environment. Each application uses a different type of data, i.e., on-site, satellite and modeling data. The on-site data application showed the thermohaline distribution in JZB during winter. The satellite-data application described the distribution of Chl *a* concentration in JZB, showing a decrease from the inner bay to the mouth, and high values in coastal areas. We demonstrated, for the first time, that Chl *a* concentration products from GOCI at the bridge location were contaminated because the bridge surface has a different reflectance to water. This highlights the importance of the quality controls implemented in the database to provide valid data to users. The modeling data application evaluated the effects of the JZB Bridge on tidal dynamics, showing only a small impact.

All these kinds of data and analysis can be extremely useful for a wide range of users, such as: model developers who need to validate model results by comparing with on-site or satellite data, researchers from different fields interested in the variability of different biogeochemical and physical variables at different time scales, and managers and governments interested in evaluating the current ecological state of the bay or the effects of planned marine engineering on the bay environment. The potentially wide range of users, together with the unified structure of the database and the efficient quality-control system, makes the JZB database a valuable resource for coastal research and management.

References

Boyer, T. P., and others. 2013. World Ocean Database 2013. NOAA Printing Office.

Chen, C., H. Liu, and R. C. Beardsley. 2003. An unstructured grid, finite-volume, three-dimensional, primitive equations ocean model: Application to coastal ocean and estuaries. J. Atmos. Oceanic Tech. **20**: 159–186. doi:10.1175/1520-0426(2003)020<0159:AUGFVT>2.0.CO;2

Chen, J. R., and X. Chen. 2012. Numerical simulation of the hydrodynamic evolution of the Jiaozhou Bay in the last 70 years. Acta Oceanol. Sin. **34**: 30–41.

Dong, Z., A. Lou and W. Cui. 2010. Assessment of eutrophication of Jiaozhou Bay. Mar. Sci. **34**: 35–39 (in Chinese with English abstract).

Gao, G. D., X. H. Wang, and X. W. Bao. 2014. Land reclamation and its impact on tidal dynamics in Jiaozhou Bay, Qingdao, China. Estuar. Coast. Shelf Sci. **151**: 285–294. doi:10.1016/j.ecss.2014.07.017

Gao, G. D., X. H. Wang, X. W. Bao, D. Song, X. P. Lin, and L. L. Qiao. 2018. The impacts of land reclamation on suspended-sediment dynamics in Jiaozhou Bay, Qingdao, China. Estuar. Coast. Shelf Sci. **206**: 61–75. doi:10.1016/j.ecss.2017.01.012

Gao, G. D., and others. 2018. Effects of wave-current interactions on suspended-sediment dynamics during strong wave events in Jiaozhou Bay, Qingdao, China, J. Phys. Oceanogr. **48**: 1053–1078. doi: 10.1175/JPO-D-17-0259.1

Jiang, Y., and others. 2018. A smart web-based geospatial data discovery system with oceanographic data as an example. ISPRS Int. J. Geo-Inf. **7**: 62, DOI: 10.3390/ijgi7020062

Levitus, S., and others. 2013. The world ocean database. Data Sci. J. **12**: WDS229-WDS234. doi: 10.2481/dsj.WDS-041

Li, P., G. Li, L. Qiao, X. Chen, J. Shi, F. Gao, N. Wang, and S. Yue. 2014. Modelling the tidal dynamic changes induced by the bridge in Jiaozhou Bay, Qingdao, China. Cont. Shelf Res. **84**: 43–53. doi:10.1016/j.csr.2014.05.006

Liang, S. K., S. Pearson, W. Wu, Y. J. Ma, L. L. Qiao, X. H. Wang, J. M. Li, and X. L. Wang. 2015. Research and integrated coastal zone management in rapidly developing estuarine harbours: A review to inform sustainment of functions in Jiaozhou Bay, China. Ocean Coast. Manag. **116**: 470–477. doi:10.1016/j.ocecoaman.2015.09.014

Liu, R. 1992. The characteristics of natural environment, p. 2–3. *In* R. Liu [ed.], Ecology and living resources of Jiaozhou Bay. Science Press, (in Chinese).

Lyubartsev, V. G., V. L. Vladymyrov, and V. V. Myroshnychenko. 2004. Multidisciplinary marine environmental database for the Aral Sea. J. Mar. Syst. **47**: 3–9. doi:10.1016/j.jmarsys.2003.12.004

McFeeters, S. K. 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. Int. J. Remote Sens. **17**: 1425–1432. doi:10.1080/01431169608948714

Moon, J. E., and others. 2012. Initial validation of GOCI water products against in situ data collected around Korean peninsula for 2010–2011. Ocean Sci. J. **47**: 261–277. doi: 10.1007/s12601-012-0027-1

Qian, G., H. Han, J. Liu, S. Liang, X. Shi, and X. Wang. 2009. Spatiotemporal changes of main chemical pollutants for the last thirty years in the Jiaozhou Bay. Period. Ocean Univ. China **39**: 781–788 (in Chinese with English abstract).

Ryu, J. H., H. J. Han, S. Cho, Y. J. Park, and Y. H. Ahn. 2012. Overview of geostationary ocean color imager (GOCI) and GOCI data processing system (GDPS). Ocean Sci. J. **47**: 223–233. doi:10.1007/s12601-012-0024-4

Smith, K.R., and R.A. McConnaughey. 1999. Surficial sediments of the eastern Bering Sea continental shelf: EBSSED database documentation. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-104.

Song, D., X. H. Wang, X. Zhu, and X. Bao. 2013. Modeling studies of the far-field effects of tidal flat reclamation on tidal dynamics in the East China Seas. Estuar. Coast. Shelf Sci. **133**: 147–160. doi:10.1016/j.ecss.2013.08.023

Vladimirov, V. L. 1995. An integrated data bank of an oceanographic cruise. Autom. Control Comput. Sci. **29**: 47–50.

Vladimirov, V. L., and V. V. Miroshnichenko. 1997. Multipurpose database management systems for marine environmental research, p. 355–364. *In* N. B. Harmancioglu, M. N. Alpaslan, S. D. Ozkul, and V. P. Singh [eds.], Integrated approach to environmental data management systems. Springer Netherlands.

Vladimirov, V. L., S. T. Besiktepe, and D. G. Aubrey. 1999. Database and database management system of the TU-Black Sea Project, p. 291–301. *In* S. T. Beşiktepe, Ü. Ünlüata, and A. S. Bologa [eds.], Environmental degradation of the Black Sea: Challenges and remedies. Springer Netherlands.

Wang, X. and K. Li. 2006. The seaward discharge of primary pollutants in Jiaozhou Bay, p. 3-22. In: Wang, X., K. Li, and X. Shi. [eds.], The marine environmental capacity of primary pollutants in Jiaozhou Bay. Science Press, Beijing (in Chinese).

Yuan, Y., D. Song, W. Wu, S. Liang, Y. Wang, and Z. Ren. 2016. The impact of anthropogenic activities on marine environment in Jiaozhou Bay, Qingdao, China: A review and a case study. Reg. Stud. Mar. Sci. **8**: 287–296. doi: 10.1016/j.rsma.2016.01.004

Acknowledgments

This paper benefited from editorial review by Dr Peter McIntyre from UNSW Canberra. Y. Yuan is supported by the China Scholarship Council and a UNSW Canberra Top-up Scholarship. Y. Yuan and D. Song are supported by the Ministry of Science and Technology of the People's Republic of China under contract (2015CB452905). The JZB database is located in the UNSW Data Archive (Research Data ID: D0235484). This is publication No. 56 of the Sino-Australian Research Centre for Coastal Management.

Conflict of Interest

None declared.

Submitted 09 August 2018
Revised 28 November 2018
Accepted 18 December 2018

Associate editor: Mike DeGrandpre