DATA SCIENCE JOURNAL

Kawasaki, A, et al. 2018. Data Integration and Analysis System (DIAS) as a platform for Data and Model Integration: Cases in the Field of Water Resources Management and Disaster Risk Reduction. *Data Science Journal,* 17: 29, pp.1–14. DOI: https://doi.org/10.5334/dsj-2018-029

RESEARCH PAPER

Data Integration and Analysis System (DIAS) as a Platform for Data and Model Integration: Cases in the Field of Water Resources Management and Disaster Risk Reduction

Akiyuki Kawasaki[1], Petra Koudelova[2], Katsunori Tamakawa[3], Asanobu Kitamoto[4], Eiji Ikoma[1], Koji Ikeuchi[1], Ryosuke Shibasaki[1], Masaru Kitsuregawa[1,4] and Tosho Koike[3]

[1] Earth Observation Data Integration and Fusion Research Initiative (EDITORIA), The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, JP

[2] Faculty of Civil Engineering, The Czech Technical University in Prague, Thakurova 7, 166 29, Prague 6, CZ

[3] International Centre for Water Hazard and Risk Management (ICHARM), Public Works Research Institute (PWRI), 1-6, Minamihara, Tsukuba, 305-8516, JP

[4] National Institute of Informatics (NII), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JP

Corresponding author: Akiyuki Kawasaki (kawasaki@hydra.t.u-tokyo.ac.jp)

The development of data and model integration platforms has furthered scientific inquiry and helped to solve pressing social and environmental problems. While several e-infrastructure platforms have been developed, the concept of data and model integration remains obscure, and these platforms have produced few firm results. This article investigates data and model integration on the Data Integration and Analysis System (DIAS) platform, using three case projects from water-related fields. We provide concrete examples of data and model integration by analyzing the data transfer and analysis process, and demonstrate what platform functions are needed to promote the advantages of data and model integration. In addition, we introduce the Digital Object Identifier (DOI), a valuable tool for promoting data and model integration and open science. Our investigation reveals that DIAS advances data and model integration in five main ways: it is a "sophisticated and robust integration platform"; has "rich APIs, including a metadata management system, for high-quality data archive and utilization"; functions as a "core hydrological model"; and promotes a "collaborative R&D community" and "open science and data repositories". This article will appeal especially to researchers interested in new methods of analysis, and information technology experts responsible for developing e-infrastructure systems to support environmental and scientific research.

Keywords: data and model integration; platform; dam; hydroelectric power; flood control

1. Introduction

1.1. Background

If global environmental challenges are to be addressed effectively, a coordinated international approach that plans, implements and manages data, analytics and e-infrastructures is required (Belmont Forum 2015; Yarime 2017a). Countries around the world are developing e-infrastructures that will facilitate the combination of digitally-based technology (hardware and software), resources (data, services, digital libraries), and communications (protocols, access rights and networks). This, in turn, will support cutting-edge, collaborative research in a wide range of scholarly fields (RCUK 2010).

The integration of natural scientific data and socio-economic data is especially important for solving social problems and supporting decision-making. In recent years, we have gained a better understanding of the earth's environment by integrating and analyzing different data, including satellite, in-situ, model output, socio-economic and geospatial data. Combining socio-economic data with numerical models of

various scales allows us to visualize current situations and project future scenarios with regard to global environmental changes.

Integration of various data and models is leading to the development of methods by which global environmental issues can be addressed. The field of water resources management and water-related disaster risk reduction is a case in point. The development of a platform for integrating data and models has furthered scientific inquiry and helped to solve social and environmental problems.

Several e-infrastructure technologies and platforms for data and model integration have been developed. For example, the National Science Foundation (NSF) in the United States developed EarthCube as cyberinfrastructure to support the geoscience research community and society. EarthCube offers interoperability standardization, and technologies for better integration, visualization, and data analysis (Tarboton et al. 2011; Yang et al. 2017). Australian Geoscience is developing the Data Cube to store and utilize Earth observation data, with sophisticated data processing functions to analyze continental time-series of these data. At the moment, Data Cube focuses mainly on remote sensing data (Lewis, 2017). INSPIRE is European Spatial Data Infrastructure established and operated by member states of the European Union. INSPIRE enables European governments to share environmental spatial data, which, in turn, promotes policy-making across boundaries (INSPIRE Network Services Drafting Team, 2011).

Nonetheless, the concept of data and model integration remains obscure. The technologies and platforms mentioned above have generated many individual results, but few firm results in the context of data and model integration.

1.2. Objectives

The Data Integration and Analysis System (DIAS) is a platform (established 2006) for solving social and environmental problems, and promoting inter- and trans-disciplinary research and development (Kawasaki et al. 2017). In this article, we draw on results from the DIAS platform to investigate cases of data and model integration in the field of water resources management and water-related disaster risk reduction. By investigating three case projects taken from DIAS, we provide concrete examples of data and model integration by analyzing the data transfer and analysis process, and demonstrate what platform functions are needed to promote data and model integration. In addition, we introduce the Digital Object Identifier (DOI), a valuable identifier system for promoting data and model integration and open science.

This article will appeal to a wide audience, including researchers interested in new methods of analysis, and information technology experts responsible for developing e-infrastructure systems to support environmental and scientific research.

While the many applications and tools available through the DIAS platform support research in a variety of fields (Kawasaki et al. 2017; Moiz et al. 2018; Pandey et al. 2018), this article focuses on one field: water resources management and water-related disaster risk reduction. We have selected three cases for study:

- Case I: Real-time flood predictions and flood-control dam operation optimization system in Japan.
- Case II: Hydroelectric power (HEP) dam operation schemes in Japan, which aim to reduce the risk of floods while improving power generation efficiency.
- Case III: A system that assesses flood-risk reduction investment as a means for development in Pakistan.

2. Outline of DIAS

DIAS collects and files data from satellites, ground observation stations, numerical weather prediction models and climate change projection models. Once this data is integrated with geographic and socio-economic information, DIAS generates results for managing global environmental issues and natural disasters.

The underlying framework of DIAS's common base application platform is shown in **Figure 1**. Real-time datasets are provided by various organizations on a real-time basis. The DIAS archive includes satellite data and ground observation data, such as XRAIN (X-Band MP Radar Network), up-to-the-minute local composite rainfall data collected by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) in Japan; and Himawari-8 images, taken by the geostationary weather satellite Himawari-8.

DIAS users also have access to climate change simulation data from the Coupled Model Intercomparison Project (CMIP; Meehl et al. 2000; Kawasaki et al. 2017), and the Nonhydrostatic ICosahedral Atmospheric Model (NICAM; Satoh et al. 2014). These datasets are stored in an extra large volume disk array system with

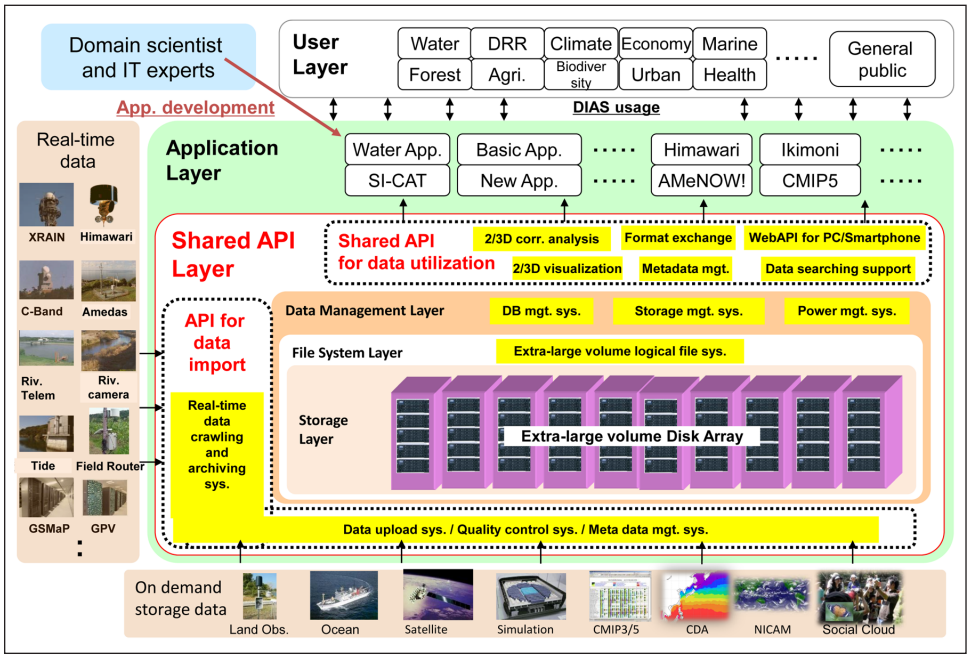

Figure 1: DIAS's common base application platform.

API (Application Programming Interface) software used to import the data. APIs work to convert or re-format data into manageable forms and are useful in creating the DIAS storage archive. API software consists of several tools, including the "Real-time data crawling and archiving system", "Data upload system", "Quality control system", and "Metadata management system".

Once a dataset is archived in the DIAS system, "Shared APIs for data utilization" are implemented. These APIs include functions – "2/3D visualization tools", "Format exchange tools," "2/3D correlation analysis tools", "Data searching support tools" – that work to support data utilization.

The DIAS Application Layer allows domain scientists working in environmental fields, including water resources management, hydro-metrology, agriculture, biodiversity, urban planning, public health, economics, and disaster risk reduction, to develop specific tools and applications by working together with Information Technology experts. These applications are connected directly to the User Layer, where they can be put into service.

DIAS is linking scientists around the world, allowing them to tackle global problems in a coordinated manner. DIAS is part of the Global Earth Observation System of Systems (GEOSS), and is linked to GEOSS data centers internationally through the exchange of meta-data via GCI (GEOSS Common Interface). An overview of DIAS and its infrastructure system, application development, and Research and Development community, is given in Kawasaki et al. (2017).

3. Case I: Real-time flood prediction and flood-control dam operation optimization system

3.1. System overview and uniqueness

The purpose of this system is to provide early flood forecasts and subsequent real-time decision support for flood-control dam operators when a flood is predicted. The development and implementation of the system for the Upper Tone River basin in Japan is described in Shibuo et al. (2016). Shibuo recognized two major problems concerning flood forecasting capabilities, both of which are addressed by DIAS.

The first problem lies in the accuracy of flood predictions, which are based on initial conditions in the basin and precipitation forecasts. This problem is addressed by (i) employing an advanced distributed hydrological model, which simulates continuously the physical state of the basin; (ii) employing an ensemble precipitation forecast model that uses errors in past forecasts to calculate uncertainty; and (iii) combining these two models to generate an ensemble flood forecast.

The second problem concerns communication: how best to disseminate flood information to authorities quickly and effectively. This problem is addressed by (i) integrating the modeling system with the data archive and a real-time data processing model on the DIAS system to ensure smooth and timely flow of data, and (ii) developing an interactive virtual reservoir simulator that shows how the operation of flood-control

dam gates will affect predicted stream flow. Integrating and combining models and data in real-time requires a sophisticated and robust data and model integration platform, which DIAS provides.

The unique features of this system include the integration of advanced hydrological models; a real-time data processing scheme; a large data archive; real-time data retrieval from various providers (in-situ weather, water level and streamflow stations, radar, weather forecasts); real-time and continuous simulations; and a virtual interactive reservoir simulator that allows dam operators to forecast scenarios and establish the optimal operation of dam gates under different conditions.

3.2. Data

The system integrates various types of data from multiple sources and at different temporal scales (**Figure 2**). The static data used to set up the basin for the hydrological model are described in Wang et al. (2009b). Forcing data that are needed to run the hydrological model are acquired continuously from in-situ observation stations and weather radars. Non-observed variables are taken from numerical weather prediction outputs. Telemeter stations provide reservoir inflow, outflow and water level data. All these data are received and processed in real-time. A new rainfall forecast is acquired every three hours, immediately after it is issued by the Japan Meteorological Agency (JMA).

3.3. Model system structure and methodology

The integrated model system consists of five components embedded into DIAS (**Figure 2**): a real-time data processing model (M1); a hydrological model, WEB-DHM (Water and Energy Budget-based Distributed Hydrological Model) (baseline mode) (M2); an ensemble precipitation forecast model (M3); an ensemble streamflow prediction (ESP) model, WEB-DHM (pred. mode) (M4); and an interactive virtual reservoir simulator model, dam operation decision support (M5). **Figure 2** shows the connections and data flow between these components.

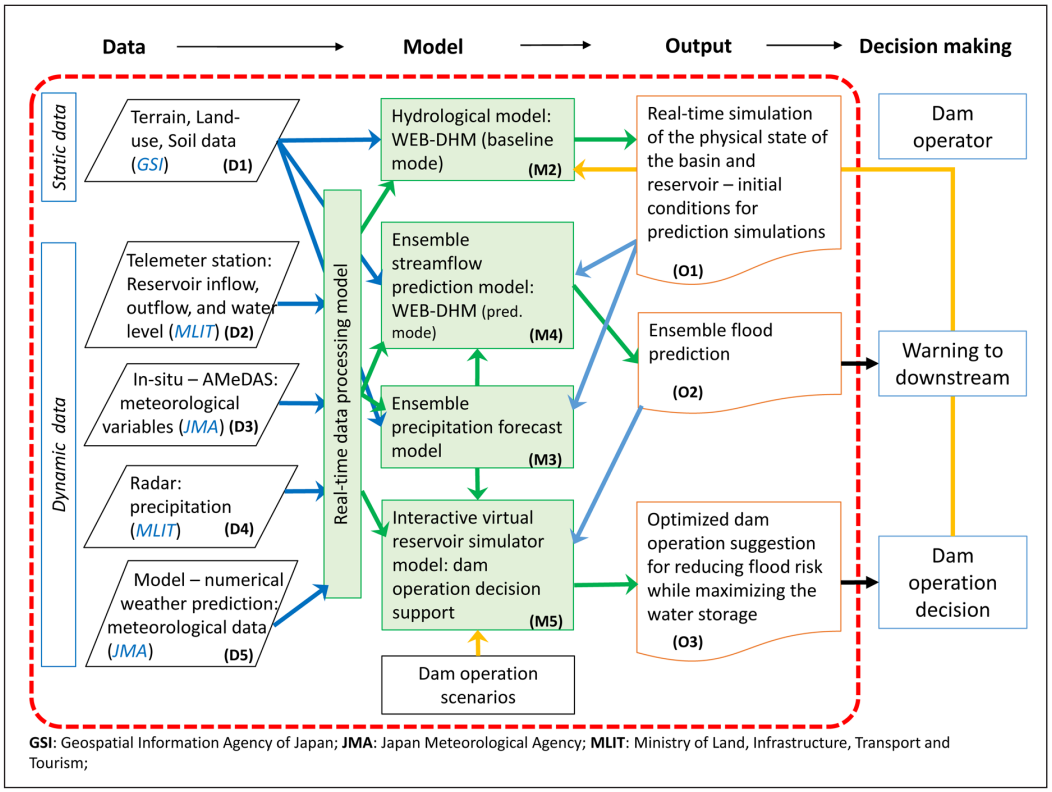

Figure 2: The schematic diagram of the DIAS flood prediction and flood-control dam operation optimization system. The red dashed line encompasses the system components. The column at left shows various types of data used by the system (D1–D5), the middle column shows individual model components (M1–M5), and the column at right depicts the system outputs (O1–O3). The arrows indicate the flow of data and information from the data source to the system (blue arrows), and between the system components (green arrows). The black arrows leading from the red box show the system output applications. The yellow arrows pointing back to the system components indicate interactive input from dam operators. The data flow and exchange between the components are assured by the DIAS system.

The real-time data processing model links the various data and models and assures automatization of input/output processes (M1). Through a built-in Geographic Information System (GIS) function, the model converts the data, supplied by multiple providers in different formats and at different times, into forms that suit the other components of the system. This ensures a seamless flow of data between individual model components, and between the interactive virtual reservoir simulator model and the end user, as well as synchronization of the models. All processes, from data retrieval to result visualization, run continuously and at real time.

The hydrological model (M2) simulates continuously (with an hourly time step) the real physical (hydrological) state of the river basin. The WEB-DHM (Wang et al. 2009a and 2009b) is employed and run in baseline mode, i.e. it proceeds by one time step whenever new input is received from the real-time data processing model (M1). The WEB-DHM model reflects a series of physical relationships that describe hydrological processes, i.e. movement of water and energy through the system (basin). It also includes a flood-control dam operation module that allows for the inclusion of data taken from dam gate operations. The hydrological model outputs include a set of key hydrological state variables for each computational grid in a basin at each time step, i.e. data on initial conditions are updated continuously (O1). Because this data is current and regular, the streamflow prediction model produces more accurate flood forecasts.

The ensemble precipitation forecast model (M3) uses a method developed by Saavedra et al. (2010). Quantitative Precipitation Forecasts (QPF), provided by the JMA in a gridded format, are evaluated against observation. Errors in both intensity and spatial displacement are considered, incorporated into a forecast error value, and converted to perturbation weights (the larger the error, the higher the perturbation weight). These empirically derived perturbation weights are used to produce 51 precipitation forecast members (M3), following the method of Saavedra et al. (2010). When the JMA issues a new 15-hour weather forecast (this is done every three hours), a new ensemble of precipitation forecasts is produced.

The ensemble streamflow prediction model, WEB-DHM (pred. mode) (M4), predicts streamflow based on the 51 ensemble precipitation forecast members (M3). It employs the WEB-DHM model, configured for the given basin, and uses the ensemble precipitation forecast as precipitation forcing data. Other meteorological variables, such as air temperature, wind speed, solar radiation, humidity and air pressure are taken from JMA as Meso-Scale Model (MSM) Gird Point Value (GPV). Information on initial conditions comes from a baseline WEB-DHM simulation (O1). The ESP model (M4) generates 51 different time-series on possible streamflows covering the period that corresponds to the weather forecast, i.e. the next 15 hours. Accordingly, the model provides a range of possible flood flows through the point of concern (O2). In the ESP model, reservoir outflow is set to equal reservoir inflow, i.e. no reservoir regulation function is considered.

The interactive virtual reservoir simulator model, dam operation decision support (M5), is a variation of the ESP model, but it includes a flood-control dam operation function. It is an interactive model, which allows the user to introduce various flood-control dam-release scenarios for the period of the weather forecast. These scenarios are modeled by multiplying reservoir inflow with factors ranging from 0 to 2, where 0 represents no water release from the dam, 1 means release and inflow are equal, and 2 means the release is twice the inflow. In the case of heavy rain upstream, the amount of water released from the dam might need to exceed inflow, if the volume of inflow threatens reservoir capacity. To avoid such situations, a priori releases from the dam should be considered when heavy rain is predicted. Water volumes, both in reservoirs and downstream, can then be recalculated according to different dam-release scenarios.

By comparing results from various scenarios and studying predicted streamflows, dam operators can make informed, evidence-based decisions about flood-control operations. Whatever those decisions may be, information on the operation of dam gates is fed into the hydrological model and ESP model simulations. This ensures that the baseline simulation of the hydrological state of the basin is as accurate as possible.

4. Case II: Hydroelectric power (HEP): dam operation support for flood control, and improving power generation efficiency

4.1. System overview and uniqueness

Increasingly intense, torrential rainfall has caused vast damage in Japan and elsewhere around the world and overwhelmed flood control plans. HEP dams, which operate without flood-control obligations, are considered a promising option for limiting the effects of excessive precipitation and reducing flood risk (Ando et al. 2017). A HEP dam operation support system, being developed on DIAS, helps HEP dam operators to decide when to release water. This system is the result of a R&D project involving The University of Tokyo, the International Centre for Water Hazard and Risk Management (ICHARM), Nippon Koei Co., Ltd. (a construction consulting company), and two electric power companies in Japan.

4.2. Model system structure and methodology

The system is based on the one developed for real-time flood prediction and flood-control, described in the previous chapter. However, two additional models – a Cloud data assimilation model, CALDAS-WRF (M3b), and a downstream flood damage assessment model, RRI (Rainfall-Runoff-Inundation) (M6) – are included in this system (**Figure 3**) to address the specific needs of HEP dams.

The Ensemble Kalman Filter was applied to the ensemble precipitation forecast model (M3a). The selection of ensemble size is a trade-off between accuracy and computational cost. Miyoshi et al. (2012) experimented with different ensemble sizes (20, 27, 34, and 41 members). They found that the biggest improvement came when ensemble size increased from 20 to 27; an increase from 34 to 41 produced only a small difference. Therefore, 33 ensemble precipitation forecasts were generated in this case, allowing for more accurate predictions of precipitation distribution upstream from dam reservoirs. A Global Spectral Model (GSM), provided by the JMA in a gridded format, was evaluated against observations taken from zones of different sizes over and around the area of interest. By applying normally distributed random numbers with zero mean and unity standard deviation to the perturbation weights to generate GSM perturbation factors, the GSM value was spread into 33 members.

A cloud data assimilation model, CALDAS-WRF (M3b), was developed to analyze rainfall distribution in a specific basin with a high degree of accuracy. Cloud area is specified using a coupled atmosphere and a land data assimilation system, part of the Weather Research and Forecasting Model. This model simultaneously assimilates data on soil moisture, heat and moisture within clouds, and vertically integrated cloud water content over land. Because cloud data assimilation uses up-to-date information about the location of rain systems, it has the potential to improve predictions about where rain will fall, and how heavily. This method promises significant improvement in predicting heavy rainfall and the flooding that may follow.

The hydrological model, WEB-DHM (ensemble streamflow prediction mode) (M4), calculates ensemble streamflow according to the ensemble forecast precipitation (O4). It employs the WEB-DHM model, configured for a given basin, but is run using input forcing data from the ensemble precipitation forecast model

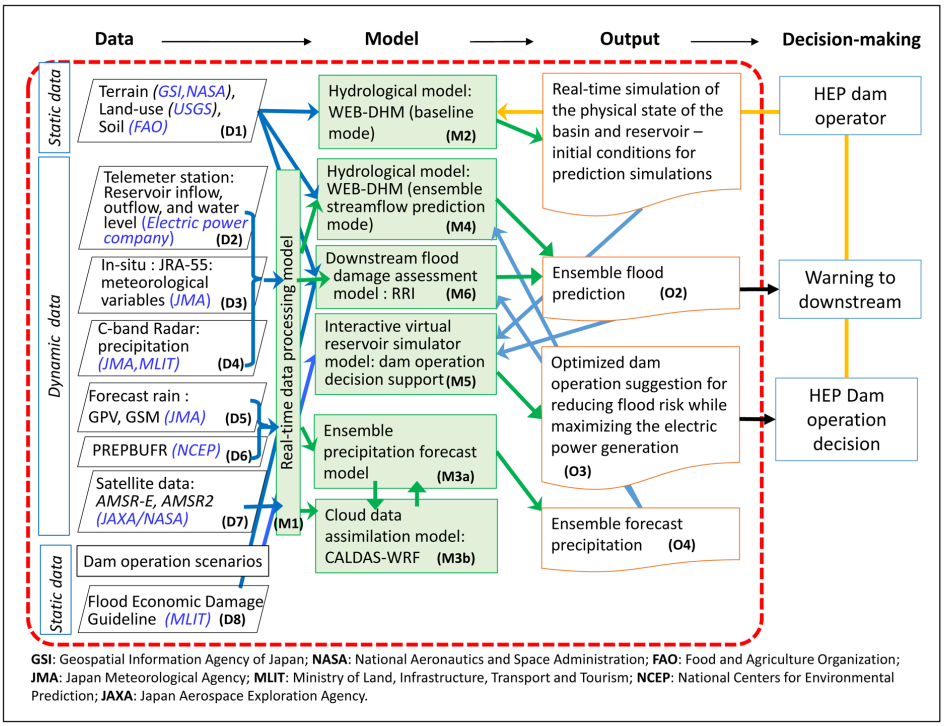

Figure 3: Hydroelectric power (HEP) dam operation schemes in Japan, aiming to reduce the risk of floods while improving power generation efficiency. The red dashed line encompasses the system components. The column at left shows various types of data used by the system (D1–D7), the middle column shows individual model components (M1–M6), and the column at right depicts the system outputs (O1–O3). The arrows indicate the flow of data and information from the data source to the system (blue arrows), and between the system components (green arrows). The black arrows leading from the red box show the system output applications. The yellow arrows pointing back to the system components indicate interactive input from dam operators. The data flow and exchange between the components are assured by the DIAS system.

(M3a), precipitation area predictions from cloud data assimilation model, CALDAS-WRF (M3b), and meteorological variables taken from the JMA's Reanalysis 55 (JRA-55, 2013) (D3).

The interactive virtual reservoir simulator model, dam operation decision support (M5), is a variation of the hydrological model, WEB-DHM (ensemble streamflow prediction mode) (M4), but this M5 model includes a HEP dam operation function. The controllers of HEP dams are under no obligation to work to control floods. In order to promote a priori releases of water from HEP dam reservoirs, the benefits of doing so need to be demonstrated quantitatively. This is possible when the effect of releasing water to limit flood damage is compared with the economic losses sustained by electric power companies when precipitation predictions are missed and reservoirs are not refilled by floods. Ando et al. (2017) illustrate how a priori releases of water can reduce flood damage. Their study, which considered 4 HEP dams in the Oi River basin, calculated that timely a priori releases of water could reduce economic losses by 22 billion JPY in the event of a large scale flood. (The calculation, which draws on information from the Japanese Ministry of Land, Infrastructure, Transport and Tourism (MLIT 2016), is based on worst-case scenarios).

The downstream flood damage assessment model, RRI (M6), was developed to assess the economic damage caused by floods. It draws on the Guideline of Flood Economic Damage (MLIT 2005). This two-dimensional model simulates rainfall runoff and flood inundation concurrently (Sayama et al. 2012). The M6 model can help HEP dam operators to decide when to release water from reservoirs, and when to issue warnings to residents downstream.

5. Case III: Investment in flood-risk reduction as a means for development in Pakistan

5.1. System overview and uniqueness

The purpose of this system is to assess quantitatively how economic growth in developing countries can be boosted by preventing or reducing flood disasters. The system shows the benefits that derive from data and model integration enabled by the DIAS platform in the field of development investment. The development of the system and its pilot application in Pakistan are described in Ota (2014) and Yokomatsu et al. (2014). Outputs are in the form of economic indices.

The system integrates a flood simulation model and a dynamic macroeconomic model that simulates long-term economic growth (**Figure 4**). The flood model combines an advanced land surface and hydrological model with a river routing and inundation model. This ensures that basin runoff is calculated accurately, and an estimate of inundation is obtained. The inundation calculation considers potential flood protection structures along the river channel. Various flood protection scenarios (instances of flood-risk reduction investment, such as levees on a river channel) can be introduced and their effect on economic growth examined. The flood-related variables – extent of inundation, depth, duration – are converted into indices that measure mortality and the effects of disasters on human capital and financial and physical assets. These indices are calculated for floods of different water levels and for people in different income groups. The flood inundation distribution is combined with population density and income group distribution (study area) to estimate damage rates for each income group/flood level. The effect of flood-risk reduction investment is introduced by reducing flood-related variables such as extent of inundation and depth. The damage rates are inputs for the economic model (Yokomatsu et al. 2014), which forecasts long-term economic growth with and without disaster protection investment. The outputs are GDP growth and a Gini coefficient.

The unique features of this system include: the integration of a flood simulation model with an economic model to convert physical variables into economic parameters; the integration of a regional-scale flood inundation model with advanced hydrological models to improve discharge simulation; the incorporation of a macroeconomic model that can quantify the long-term economic effects of disaster risk reduction investment (rather than simply measuring loss after a disaster); and a function that allows for specific consideration of developing economies.

5.2. Data

The static data used in the hydrological models are taken from publicly available global datasets: Digital Elevation Mode (DEM) data from the Shuttle Radar Topography Mission (SRTM); soil type data from the Food and Agriculture Organization of the United Nations (FAO); land use data from the United States Geological Survey (USGS); glacier coverage data from the International Centre for the Integrated Mountain Development (ICIMOD); and dynamic vegetation and snow cover data from MODIS Terra (**Figure 4**). Weather and hydrological forcing data come from in-situ observation stations of the Pakistan Water and Power Development Authority (WAPDA) and the Pakistan Meteorological Department (PMD) (D2), and reanalysis data from

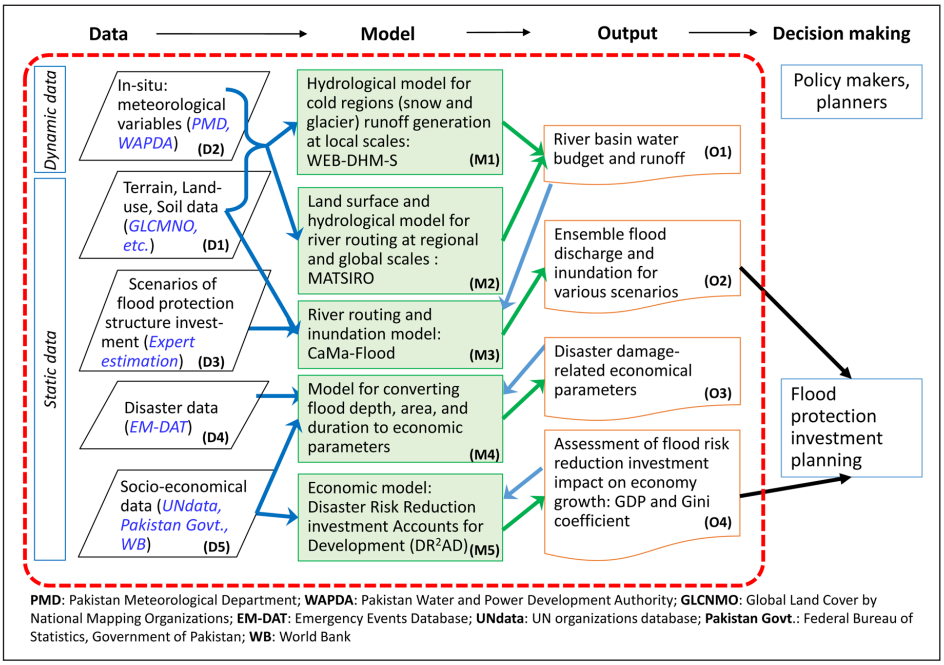

Figure 4: The schematic diagram of the flood-risk reduction investment for development system. The red dashed line encompasses the system components. The column at left shows various types of data used by the system (D1–D5), the middle column shows individual model components (M1–M5), and the column at right depicts the system outputs (O1–O4). The arrows indicate the flow of data and information from the data source to the system (blue arrows), and between the system components (green arrows). The black arrows leading from the red box show the system output applications. The data flow and exchange between the components are assured by the DIAS system.

Global Land Data Assimilation System (GLDAS). Disaster data is derived from the Emergency Event Database (EM-DAT) (D4); population distribution and socio-economic data from United Nations databases and the Pakistani Federal Bureau of Statistics; and wealth-related data from the World Bank (D5). Land cover data for the DRR investment Accounts for Development (DR²AD) model were taken from the Global Land Cover by National Mapping Organizations (GLCNMO) Global Map V.1 (D1).

5.3. Model system structure and methodology

The integrated model system consists of five model components (**Figure 4**): a hydrological model for cold regions (snow and glaciers) runoff generation at local scales, WEB-DHM-S (WEB-DHM with advanced snow and glacier physics (Shrestha et al. 2015)) (M1); a land surface and hydrological model for river routing at regional and global scales, MATSIRO (Minimal Advanced Treatments of Surface Interaction and Runoff model; Takata et al. 2003) (M2); a river routing and inundation model, CaMa-Flood (Yamazaki et al. 2011; 2012) (M3); a model for converting flood depth, area, and duration to economic parameters (M4); and an economic model, Disaster Risk Reduction investment Accounts for Development (DR²AD) (Yokomatsu et al. 2013) (M5). The cold region model simulates land surface and hydrological processes in the area of a river basin. The WEB-DHM-S is used to simulate surface and subsurface runoff from snow- and glacier-affected areas of the basin. The energy balance-based snow and glaciers component means the model can be applied to cold regions such as the Upper Indus basin.

The regional and global scale land surface and hydrological model, MATSIRO, simulates land surface and hydrological processes in a more simplified manner than does WEB-DHM-S. As the model can be applied at both regional and global levels, there is a consistency of scale between its application and that of the CaMa-Flood inundation model. The MATSIRO is used to simulate surface and subsurface runoff from the basin area, where snow and glacial processes are not involved.

The river routing and inundation model simulates river flow and inundation depth and duration. The distributed global river routing model CaMa-Flood (M3) is used to chart the runoff generated by the land surface and hydrological models (O1). For each grid point, the model determines water storage and depth, river discharge, and the extent of inundated areas. Because the CaMa-Flood allows for the inclusion of hypothetical flood protection structures, such as high levees, consideration can be given to different flood scenarios.

The flood-damage conversion model (M4) translates the flood model output variables (flood discharge, inundation area, depth, duration) into six disaster-related parameters (flood return period, physical damage rate, human damage rate, land damage rate, financial damage rate, and disaster mortality rate). These parameters inform calculations for different scenarios. The calculations are entered into the economic model, which simulates the long-term effects that investment in disaster risk reduction (flood protection measures, in this case) has on economic growth. The dynamic stochastic macroeconomic model DR^2AD (M5) calculates GDP and a Gini coefficient that considers specifically the economies of developing countries. GDP and Gini coefficient multiple time series show a positive link between flood protection measures and future economic growth.

6. How Digital Object Identifiers (DOI) promote data and model integration

6.1. Introducing a DOI to create a long-term and stable data platform

DIAS collects and collates valuable scientific data and models. The gathering and distribution of data are complex issues. Adoption of best practice measures, including the FAIR (Findable, Accessible, Interoperable, Re-usable) Data Principles (FORCE11 2017), is important if DIAS is to be a trustworthy, stable and secure data repository.

Long-term and stable access to data also depends on the management practices of researchers. Researchers must be educated about the danger of losing valuable data. It is common practice to keep data on private hard disks, but without strict archiving and access methods data may be lost or stolen. Moreover, some data, such as satellite data and climate change prediction model outputs, require large volume data storage. Small laboratories and research offices may not have the capacity to manage such a platform.

These potential problems are why there is a need for platforms such as DIAS that offer long-term and stable access to huge volumes of data. We are striving to improve data management on the DIAS platform to ensure its reliability and security. Below we write of an achievement in this regard, and challenges that remain to be solved.

A DOI is an identifier system designed to facilitate long-term access to resources such as research data (Takeda et al. 2015). It is an alphanumeric string assigned to an object with arbitrary granularity. For a DOI to take effect, we first submit metadata describing an object and its URL. A DOI string can then be resolved into a URL, offering permanent access to data.

While DOIs were originally designed for academic journals and articles, their success has led to wider use, including for research data. DIAS has developed a system whereby requests for the assignment of DOIs to DataCite are communicated directly to the Japanese registration agency (Japan Link Center).

6.2. Advantages of DOIs

DOIs are useful in allowing long-term access to data. DOIs have an advantage over URLs because the ID for an object and the ID for a location are separated and integrated through a resolver. If a location is changed, it is a simple task to change the setting of the resolver and redirect access to the new location. DOIs are useful when assigned to represent objects, as this allows for consistency across academic resources.

DOIs can be assigned to data on enabled platforms, but not on platforms such as personal websites. The role of DIAS in collecting and sharing data, and promoting open science, is expected to grow.

DIAS assigned its first DOI to the "GAME-Tibet POP/IOP Dataset" (DOI: https://doi.org/10.20783/DIAS.496) (Kitamoto 2017b). This data was archived under the GAME (Global Energy and Water cycle Exchanges (GEWEX), Asia Monsoon Experiment) framework (GEWEX 2001). The field experiment, implemented in the Tibetan Plateau at both the plateau and meso scales, clarified land surface-atmosphere interactions in the context of the Asian monsoon system. Pre-phase observations were conducted in 1997 and precise, intensive observations in 1998. Careful analysis, based on the large data archive, led to new findings about the water cycle mechanism in monsoonal regions of Asia (JMSJ GAME special issue).

DOIs are expected to increase data utilization in new research fields related to water and energy cycle changes. Once a DOI is assigned, its unique identity ensures it can be used universally for data and model integration. This, in turn, allows for the development of software that uses the DOI to access and analyze the same data. The DOI can also be used to identify which dataset informed the findings in a certain article. In the future, it will be easy to collect and summarize the conclusions drawn by different scientists from the same dataset. Furthermore, citing DOIs in articles ensures credit is given to the creators of data, something that does not happen as often as it should.

7. Discussion

In previous chapters, we discussed the mechanics of data and model integration. In this chapter, we consider the advantages of integration.

7.1. A sophisticated and robust integration platform

It is difficult to integrate datasets (**Figures 2, 3, 4**) using a cloud-based data system. Data is static, dynamic and multi-layered, making transfers difficult. The DIAS platform makes these transfers easier because it uses an extra high-speed network, a specialized data processing and quality management technique/system, and a high performance analysis server for processing huge amounts of observation and satellite data in real-time. In addition, DIAS has enormous data storage capacity (26 PB). Real-time data and archived data are stored in a single analysis environment for high-speed processing.

7.2. Incorporating rich APIs, including metadata management systems, to ensure high-quality data archives and efficient data utilization

DIAS has prepared various APIs for processing and archiving huge amounts of data quickly and efficiently. APIs are vital in developing advanced applications for integrating and utilizing data. The inclusion of APIs in the DIAS platform distinguishes it from many other large-scale data storage systems.

The same is true of the use of metadata management systems in DIAS. These systems play an important role in maintaining high data quality for storage and utilization. DIAS includes various data-quality management tools, which contribute to advanced data and model integration.

7.3. Core hydrological model for data and model integration

DIAS uses WEB-DHM as a core hydrological model. The WEB-DHM model estimates soil moisture with improved streamflow prediction capabilities for hazard assessment and water resources management (Wang et al. 2009a and 2009b). Static and dynamic data are processed efficiently and effectively by separating the WEB-DHM model into 4 components.

The WEB-DHM model allows the coupling of various physical models related to the water cycle, including a snow and glacier model (Shrestha et al. 2015), a frozen soil hydrology model (Bao et al. 2016), a dynamic vegetation model (Sawada et al. 2014; Sawada et al. 2015), and a dam operation model (Saavedra et al. 2010; Shibuo et al. 2016). The model is applicable to multiple types of basin hydrology and climate as it establishes initial values and parameters using global datasets, as well as estimates air forcing using data assimilation techniques. Furthermore, autonomous parameter adjustment allows for more accurate assessments and predictions of the effects of floods, drought and other consequences of climate change (Sawada and Koike 2016). Through this process of integration, we are better placed to develop applications that support decision-making on water resource management issues.

7.4. Collaborative R&D community

The DIAS R&D meeting function has been uniting scholars for 17 years. Scholars working in environmental fields and in Information Technology hold regular, fruitful discussions (Kawasaki et al. 2017). These collaborative meetings have led to the development of sophisticated applications for DIAS users. We believe DIAS is unique in providing a platform for scientists and other scholars to pursue data and model integration.

7.5. Open science and data repositories

The "Open Science" concept, a global agenda for realizing better science (Yarime 2017b), is important to the future of DIAS. In Japan, "Open Science" was selected as one of the strategic keywords/phrases for the 5th Science and Technology Basic Plan (CSTI 2015). In Europe, the European Open Science Cloud aims to build infrastructure to expand e-science (European Commission 2017). In the United States, The Center for Open Science (2017) is developing a platform for sharing research data and information within a user community. In keeping with this global trend, in April 2016 DIAS established an "Open Science Special Interest Group (Open Science SIG)" to discuss issues such as DOI assignment, data citation, and how the journal submission process can be made easier (Kitamoto 2017a). We are working to make DIAS open science ready.

Reproducibility and transparency of research are issues in open science. Article authors are required to deposit data to a repository where this information can be accessed in the long term. Repositories must satisfy requirements imposed by publishers to be deemed trustworthy. These basic criteria are listed in the Earth System Science Data (ESSD) Journal (2017):

1. Persistent identifier: data sets must have a digital object identifier (DOI).
2. Open access: data sets must be freely available (once a standard registration and login process is complete).
3. Liberal copyright: anyone is free to copy and disseminate data sets as long as the original authors are given appropriate credit (equivalent to the Creative Commons Attribution License).
4. Long-term availability: repositories must ensure data sets are permanently available.

As a bare minimum, data repositories must use DOIs if they wish to be deemed trustworthy.

While data repositories are indispensable to the research process, trustworthy ones are rare. We expect more and more researchers to use DIAS on the basis that it is committed both to open science and to nurturing a culture in which data providers are credited for their work.

8. Conclusions

8.1. Summary

The importance of data and model integration in the field of water resources management and disaster risk reduction has long been discussed, yet there are few concrete examples of such integration. In this article, we cited examples of data and model integration from the DIAS environment. The processing diagram shows the datasets and models used in each example, with reference to output. By investigating three case projects taken from DIAS, we provided concrete examples of data and model integration, and demonstrated what platform functions are needed to promote the advantages of data and model integration.

DIAS offers important advantages for data and model integration. Its archives contain an enormous amount of observed and simulated model output data, as well as extensive real-time in-situ data. DIAS allows for advanced data and model integration through provision of a "sophisticated and robust integration platform", "rich APIs, including a metadata management system, to ensure high-quality data and efficient utilization", a "core hydrological model for data and model integration", and a "collaborative R&D community".

We demonstrated the importance of DOIs for promoting data and model integration. DOIs will be vital in expanding the capacities of the DIAS platform.

8.2. Data and model integration issues on the DIAS platform

Inter- and trans-disciplinary approaches that join science and society will be essential in solving complex global environmental problems. The Sustainable Development Goals, the Paris Agreement on Climate (Paris Agreement), and the Sendai Framework for Disaster Risk Reduction (Aitsi-Selmi et al. 2016), provide evidence for this statement.

DIAS offers an important opportunity. Using the DIAS platform, we can combine environmental science data and socio-economic information, then conduct advanced analysis better to understand global environmental change (Yarime 2017a). At the same time, DIAS can help lead the open science movement. For this to happen we need to improve the transparency of data management and computer codes, and incorporate APIs to strengthen integrated data analytics.

A vibrant R&D community will be vital in maintaining the technical advantages of the data platform and in keeping abreast of international trends. To this end, attempts should be made to attract additional financial support, including from the private sector. As Koudelova et al. (2017) note, education about the data platform and its benefits is important if its capabilities are to expand.

Acknowledgements

The authors thank the members of the EDITORIA science team for their support, the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) for funding the DIAS project, and all those who provided data for analysis. This investigation was partially supported by the Water Cycle Data Integrator, Academic-Industry Collaboration Program, The University of Tokyo.

Competing Interests

The authors have no competing interests to declare.

References

Aitsi-Selmi, A, Murray, V, Wannous, C, Dickinson, C, Johnston, D, Kawasaki, A, Stevance, AS, Yeung, T, et al. 2016. Reflections on a science and technology agenda for 21st century disaster risk reduction. Based on the scientific content of the 2016 UNISDR Science and Technology Conference on the Imple-

mentation of the Sendai Framework for Disaster Risk Reduction 2015–2030. *International Journal of Disaster Risk Science*, 7(1): 1–29. DOI: https://doi.org/10.1007/s13753-016-0081-x

Ando, T, Kawasaki, A and **Koike, T.** 2017. Evaluation methodology for flood damage reduction by preliminary water release from hydroelectric dams. *AGU Fall Meeting 209744*. New Orleans, USA, Dec. 2017.

Bao, H, Koike, T, Yang, K, Wang, L, Shrestha, M and **Lawford, P.** 2016. Development of an enthalpy-based frozen soil model and its validation in a cold region in China. *J. Geophys. Res. Atmos*, 121: 5259–5280. DOI: https://doi.org/10.1002/2015JD024451

Belmont Forum e-Infrastructures and Data Management Collaborative Research Action Steering Committee. 2015. A Place to Stand: e-Infrastructures and Data Management for Global Change Research: Belmont Forum e-Infrastructures & Data Management Community Strategy and Implementation Plan.

Center for Open Science. 2017. Retrieved from: https://cos.io/.

CSTI: Council for Science, Technology and Innovation, Cabinet Office, Government of Japan. 2015. Report on The 5th Science and Technology Basic Plan [Tentative Translation]. Retrieved from: http://www8.cao.go.jp/cstp/kihonkeikaku/5basicplan_en.pdf.

Earth System Science Data. 2017. Retrieved from: https://www.earth-system-science-data.net/.

European Commission. 2017. Research and Innovation, Open Science. Retrieved from: https://ec.europa.eu/research/openscience/index.cfm.

FORCE11. 2017. The Future of Research and Communications and e-Scholarship, The FAIR Data Principles. Retrieved from: https://www.force11.org/group/fairgroup/fairprinciples.

GEWEX: Global Energy and Water Exchanges. 2001. Special Issue: GEWEX Asian Monsoon Experiment. *Journal of the Meteorological Society of Japan*, Series II, 79(1B). Retrieved from: https://www.jstage.jst.go.jp/browse/jmsj/79/1B/_contents.

INSPIRE Network Services Drafting Team. 2011. Position Paper on the Implementing Rules for INSPIRE Services allowing Spatial Data Services to be invoked. https://inspire.ec.europa.eu/documents/.

JRA-55. 2013. Japanese 55-year Reanalysis. Global Environment and Marine Department, Japan Meteorological Agency. http://jra.kishou.go.jp/index.html.

Kawasaki, A, Yamamoto, A, Koudelova, P, Acierto, RA, Nemoto, T, Kitsuregawa, M and **Koike, T.** 2017. Data Integration and Analysis System (DIAS) Contributing to Climate Change Analysis and Disaster Risk Reduction. *Data Science Journal*, 16(41): 1–17. DOI: https://doi.org/10.5334/dsj-2017-041

Kitamoto, A. 2017a. DIAS and Open Science – Global Trends and DIAS's Directions Focusing on the Utilization of DOI. *Paper presented at 1st DIAS Open Science Seminar*, Tokyo, Japan. Retrieved from: http://www.diasjp.net/wp/wp-content/uploads/2017/07/dias-sympo2017-kitamoto.pdf.

Kitamoto, A. 2017b. Starting the Assignment of DOI in Data Analysis and Integration System (DIAS). Retrieved from: http://www.diasjp.net/infomation/topics-dias-first-doi-registration/.

Koudelova, P, Kawasaki, A, Koike, T, Shibuo, Y, Kamoto, M and **Tokunaga, Y.** 2017. Design and implementation of a training course on big data use in water management. *Data Science Journal*, 16(46): 1–18. DOI: https://doi.org/10.5334/dsj-2017-046

Lewis, A, et al. 2017. The Australian Geoscience Data Cube – Foundations and lessons learned. *Remote Sensing of Environment*, 202: 276–292. DOI: https://doi.org/10.1016/j.rse.2017.03.015

Meehl, GA, Boer, GJ, Covey, C, Latif, M and **Stouffer, RJ.** 2000. The Coupled Model Intercomparison Project (CMIP). *Bull. Amer. Meteor. Soc.* 81: 313–318. DOI: https://doi.org/10.1175/1520-0477(2000)081<0313:TCMIPC>2.3.CO;2

Miyoshi, T and **Kunii, M.** 2012. The local ensemble transform Kalman Filter with the Weather Research and Forecasting Model: Experiments with real observations. *Pure Appl. Geophys*, 169: 321. DOI: https://doi.org/10.1007/s00024-011-0373-4

MLIT: Ministry of Land, Infrastructure, Transport and Tourism. 2005. Guideline of Flood Economic Damage.

MLIT: Ministry of Land, Infrastructure, Transport and Tourism. 2016. Flood Hazard Maps in the Oi River Basin (Worst- case scenarios).

Moiz, A, Kawasaki, A, Koike, T and **Shrestha, M.** 2018. A systematic decision-support tool for robust hydropower site selection in poorly gauged basins. *Applied Energy*, 224(15): 309–321. DOI: https://doi.org/10.1016/j.apenergy.2018.04.070

Ota, A. 2014. Assessment of disaster prevention investment for decreasing economic damage by integrating dynamic equilibrium model and flood model – towards contributing to policy making for disaster prevention (Master's Thesis). Tokyo: The University of Tokyo.

Pandey, BK, Khare, D, Kawasaki, A and **Mishra, PK.** 2018. Climate change impact assessment on blue and green water by coupling of representative CMIP5 climate models with physical based hydrological model. *Water Resources Management*. DOI: https://doi.org/10.1007/s11269-018-2093-3

RCUK (Research Councils UK). 2010. Delivering the UK's e-infrastructure for research and innovation: Report commissioned by the Department for Business Innovation and Skills. http://www.rcuk.ac.uk/documents/research/esci/e-infrastructurereviewreport-pdf/.

Saavedra, VO, Koike, T, Yang, K, Graf, T, Li, X, Wang, L and **Han, X.** 2010. Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts. *Water Resour. Res*, 46: 10544.

Satoh, M, et al. 2014. The Non-hydrostatic Icosahedral Atmospheric Model: description and development. *Progress in Earth and Planetary Science*, 1: 18. DOI: https://doi.org/10.1186/s40645-014-0018-1

Sawada, Y and **Koike, T.** 2016. Towards ecohydrological drought monitoring and prediction using a land data assimilation system: A case study on the Horn of Africa drought (2010–2011). *J. Geophys. Res. Atmos*, 121: 8229–8242. DOI: https://doi.org/10.1002/2015JD024705

Sawada, Y, Koike, T and **Jaranilla-Sanchez, PA.** 2014. Modeling hydrologic and ecologic responses using a new eco-hydrological model for identification of droughts. *Water Resour. Res*, 50: 6214–6235. DOI: https://doi.org/10.1002/2013WR014847

Sawada, Y, Koike, T and **Walker, JP.** 2015. A land data assimilation system for simultaneous simulation of soil moisture and vegetation dynamics. *J. Geophys. Res. Atmos*, 120: 5910–5930. DOI: https://doi.org/10.1002/2014JD022895

Sayama, T, Ozawa, G, Kawakami, T, Nabesaka, S and **Fukami, K.** 2012. Rainfall-Runoff-Inundation analysis of the 2010 Pakistan flood in the Kabul River basin. *Hydrological Science Journal*, 57(2): 298–312. DOI: https://doi.org/10.1080/02626667.2011.644245

Shibuo, Y, Ikoma, E, Saavedra, VO, Wang, L, Lawford, P, Kitsuregawa, M and **Koike, T.** 2016. Implementation of real-time flood prediction and its application to dam operations by data integration analysis system. *J. Disaster Res*, 11(6): 1052–1061. DOI: https://doi.org/10.20965/jdr.2016.p1052

Shrestha, M, Koike, T, Hirabayashi, Y, Xue, Y, Wang, L, Rasul, G and **Ahmad, B.** 2015. Integrated simulation of snow and glacier melt in water and energy balance-based, distributed hydrological modeling framework at Hunza River Basin of Pakistan Karakoram region. *J. Geophys. Res. Atmos*, 120: 4889–4919. DOI: https://doi.org/10.1002/2014JD022666

Takata, K, Emori, S and **Watanabe, T.** 2003. Development of the minimal advanced treatments of surface interaction and runoff. *Global and Planetary Change*, 38: 209–222. DOI: https://doi.org/10.1016/S0921-8181(03)00030-4

Takeda, H, Murayama, Y and **Nakajima, R.** 2015. Pilot project to register DOIs for research data. *Journal of Information Processing and Management*, 58(10): 763–770.

Tarboton, D, Maidment, D, Zaslavsky, I, et al. 2011. Advancing solutions for an EarthCube Design. What can be learned from the CUAHSI HIS experience? An EarthCube design approaches paper. EarthCube website. https://www.earthcube.org/document/2011/advancing-solutions-earthcube-design-cuahsi.

Wang, L, Koike, T, Yang, K, Jackson, TK, Bindish, R and **Yang, D.** 2009a. Development of a distributed biosphere hydrological model and its evaluation with the Southern Great Plains Experiments (SGP97 and SGP99). *J. Geophys. Res*, 114: D08107. DOI: https://doi.org/10.1029/2008JD010800

Wang, L, Koike, T, Yang, K and **Yeh, PJF.** 2009b. Assessment of a distributed biosphere hydrological model against streamflow and MODIS land surface temperature in the upper Tone River Basin. *J. Hydrol*, 377: 21–34. DOI: https://doi.org/10.1016/j.jhydrol.2009.08.005

Yamazaki, D, Kanae, S, Kim, H and **Oki, T.** 2011. A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resour. Res*, 47: W04501. DOI: https://doi.org/10.1029/2010WR009726

Yamazaki, D, Lee, H, Alsdorf, DE, Dutra, E, Kim, H, Kanae, S and **Oki, T.** 2012. Analysis of the water level dynamics simulated by a global river model: A case study in the Amazon River. *Water Resour. Res*, 48: W09508. DOI: https://doi.org/10.1029/2012WR011869

Yang, C, Huang, Q, Li, Z, Liu, K and **Hu, F.** 2017. Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, 10(1): 13–53. DOI: https://doi.org/10.1080/17538947.2016.1239771

Yarime, M. 2017a. Facilitating data-intensive approaches to innovation for sustainability: opportunities and challenges in building smart cities. *Sustain Sci*, 12: 881. DOI: https://doi.org/10.1007/s11625-017-0498-1

Yarime, M. 2017b. Learning and open data in sustainability transitions: evolutionary implications of the theory of probabilistic functionalism. *Environ Syst Decis*, 1–4. DOI: https://doi.org/10.1007/s10669-017-9668-z

Yokomatsu, M, Wada, H, Ishiwata, H, Kono, T and **Wakigawa, K.** 2014. An Economic Growth Model for Disaster Risk Reduction in Developing Countries. *Paper presented at 2014 IEEE International Conference on Systems*, Man, and Cybernetics, San Diego, CA, USA. DOI: https://doi.org/10.1109/SMC.2014.6974139

Yokomatsu, M, Wakigawa, K, Wada, H, Takeya, K, Okayasu, T, Sonoda, T, Takamatsu, H, Ishiwata, H, Amano, Y, Nagatomo, N and **Mimaki, J.** 2013. Disaster Risk Reduction Investments Accounts for Development: Model and Case Study of Pakistan. *Paper presented at 4th Conference of the International Society for Integrated Disaster Risk Management*. Northumbria University, Newcastle upon Tyne, UK.

How to cite this article: Kawasaki, A, Koudelova, P, Tamakawa, K, Kitamoto, A, Ikoma, E, Ikeuchi, K, Shibasaki, R, Kitsuregawa, M and Koike, T. 2018. Data Integration and Analysis System (DIAS) as a Platform for Data and Model Integration: Cases in the Field of Water Resources Management and Disaster Risk Reduction. *Data Science Journal,* 17: 29, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2018-029

Submitted: 18 March 2018 **Accepted:** 27 September 2018 **Published:** 23 October 2018