Response to comments on "A Data Model to Manage Data for Water Resources Systems Modeling"

By

Adel Abdallah and David Rosenberg

Correspondence to Adel M. Abdallah (amabdallah@aggiemail.usu.edu)

Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, Utah, United States

December 13, 2018

General response

We thanks the two reviewers and the editor for the constructive feedback on our manuscript to more clearly describe several aspects of the methods, results, and limitations of the work. Please refer to the "TrackChanges" attached document for all changes made in comparison to the first submission. We made the following main changes:

- Modified the graphical abstract figure and Figure 1 to better present the workflow and generality of software tools
- Discussed how WaMDaM supports units conversion
- Added a new column to the Methods table to support data quality and uncertainty
- Moved many URLs to citations and updated the relevant GitHub repository to point to the five iterative designs
- Rewrote Section 6 to include subsections on i) how users of recently published studied could use WaMDaM and its tools, ii) current limitations, iii) future work, and iv) invitation for feedback
- Removed redundant text and checked for grammar
- Replicated the results after the updated minor changes to the data model and the Wizard.

Below, we list reviewer comments in black and provide our responses and changes in blue (this color).

Comments from the editor

Two reviewers are in agreement that the paper has merit but requires revision and improvement to be suitable for publication in EMS. Please give careful consideration to the comments provided and revise the paper in a timely manner.

Thank you. We have revised the paper to address these comments as described below.

I note Reviewer 1's comment that there is no solid evaluation of the utility of the data model. Please think about how to demonstrate this either through real case studies which I'm sure you have got in hand, or through a virtual case studies.

The utility of the data model has been demonstrated in the Bear River Watershed, Utah in five use cases. We modified the sentence in line156 to say: "We demonstrate the utility of the data model in five use cases". We provide more specifics to Reviewer 1 comments below.

Reviewer 2 notes that it is difficult to deduce the primary contribution of the paper from the number of tools and scripts provided. Please think about how to make the core contribution more clear.

- We added the following sentence in lines 141-144 to the Introduction section to better communicate contribution related to software and tools:
  "Here, we contribute a generalizable data model called the Water Management Data Model (WaMDaM) to help organize, join, compare, and analyze multiple water resources datasets and models." We also introduce software tools that demonstrate key functionalities of the design.

- We also added Section 6.1 "How can modelers use WaMDaM database and its software?" where we illustrate the benefits of WaMDaM and its six software tools by presenting how the research groups of five recently published systems modeling studies can use the WaMDaM software tools.

Finally, from my perspective, I would like to see your introduction/literature review section show how your data model could have benefited prior modeling efforts reported in EMS. In other words, I'd like to "attract" readers to your article by having you identify several recent studies presented in EMS wherein your model - had it been available - would have improved or otherwise benefited their work. Then add a few paragraphs citing those studies and describing how the contribution would help such studies in the future. This type of citation will trigger a notice to those authors that their studies were cited and will draw them to look at your paper and to potentially use it in their next/on-going modelling work - ultimately using your tools and citing your paper. I hope you see the value in this additional effort/addition to the paper.

- This is a great suggestion. We have added section 6.1: "How can modelers use WaMDaM database and its software?". The section introduces 5 recent systems modeling studies and describe how they can benefit from use of WaMDaM.

-Reviewer 1

R1-Comment 1

This paper has merits, and reasonably valuable contributions in the modeling of water resources systems. Specifically, this submission introduces the WaMDaM that is surely helpful for researchers to combine the water source systems and analyze the combined data. With sufficient revisions, I believe that it can be suitable for submission, but also it would be a more viable submission by revision.

Thank you. We revised the paper and clarified the issues raised. Please see our further responses below.

R1-Comment 2

My first concern over this current submission is that it falls short of justification that allows the authors to make a concrete background about the use of a relational database. Currently, the use of a non-relational database has largely increased, and I think a non-relational database would be surely helpful for this type of data integrations. At least, a non-relational database provides more extensible components and controlled vocabularies, and also has many open sources. Although I am not suggesting the authors to use a non-relational database, having a short discussion about this would make a better manuscript.

- We added this paragraph at the end of Section 3.3 that summarizes the benefits of non-relational databases.
  "We note that non-relational databases or are increasingly being used and have the advantage that they scale and adapt without being limited to a schema (Hoberman, 2014). The core contribution of WaMDaM is the specification of the design requirements to store and organize water resources systems data and presenting logical and physical models to implement several water resources systems modeling use cases. Our implementation of the physical model as a relational database is just one way to solve the problem to organize and query heterogeneous data and serve that data to a model. Implementation in a non-relational database could likely satisfy the same use cases. The ability for WaMDaM to scale and adapt to manage much bigger datasets and models will be addressed in future work as we work with larger datasets."

In Section 4, we also now highlight the advantages of using the relational method to:

"i) support direct access to all data (Requirement #7), ii) be platform independent and implement as open-source on different operating systems for different relational database systems (Requirement #8), iii) support a standardized and stable Structured Query Language (SQL), and iv) follow common use and familiarity with the RDBMS within the water resources community"

R1-Comment 3

The authors need to do a better work on the usefulness of this development. Actually, there is no virtual evaluation of the proposed software. Many researchers have proposed software and tools like this submission, but we do not know if they will be useful.

Section 5 demonstrates and evaluate the usefulness of the data model and its supporting software in five use cases using 13 national and regional different datasets in the Bear River Watershed, USA. Results provide key water system information such as i) find input data within a model study area, ii) identify flow directions and connections among natural and engineered system components, iii) identify and compare water supply, demand, and reservoir data across multiple datasets and models, iv) show data similarities and differences among modeling scenarios, and v) select data, serve the data to a model, and run multiple model scenarios.

- We have added the following sentence to the end of first paragraph of Section 5 to emphasize that the five use cases support common operations that water resources systems analysts and modelers perform to develop and use models.
  "Together, the five use cases support common operations that water resources systems analysts and modelers perform to develop and use models."

We agree that even well-designed software and tools might not be adopted by users. The paper does not claim that the tool will be taken up.

R1-Comment 4

In the manuscript, I saw the authors have revised this software and collect feedbacks from collaborators. Thus, if the authors conduct a few open-ended questions with a focus group that could actually use this tools or collaborators, it would be useful and legitimate the publication of this work.

Lines 543-546: I could not find related version updates and feedbacks over five years in GitHub pages. If you briefly provide the feedback from collaborators it can be more viable submission.

- We updated the GitHub repository ReadME file to clearly direct visitors to the specific URL of the posted ER diagrams of the five versions. We also made it clear in the manuscript to better describe the scope and nature of the feedback under a new subsection 3.4. The section now reads:

  We iteratively revised this data model design in five key versions over the course of five years to satisfy the design requirements and use cases. The changes were in response to feedback from collaborators at the University of Manchester, University of California, Davis, and University of Massachusetts, Amherst on WaMDaM design and tools, and we acknowledge the need for larger and more diverse community testing and feedback to serve a wider audience of users. We also incorporated feedback on an earlier design and its description (Abdallah and Rosenberg, 2014). The five key designs are available on GitHub at https://github.com/WamdamProject/WaMDaM_Information_Model/tree/master/Earlier_ER_diagrams

- We added sections 6.2 and 6.3 that emphasize:
  "In response to earlier feedback, we are collaborating on a software ecosystem to make WaMDaM interoperable with Hydra Platform web-services (Knox et al., 2014), and OpenAgua (Rheinheimer, 2018). WaMDaM users will be able to import data stored in Hydra Platform as a new source of data. Users will also be able to export WaMDaM data into Hydra Platform and visualize networks and their data in OpenAgua. We are also integrating WaMDaM as a new HydroShare resource type to publish populated WaMDaM SQLite files and extract their metadata which enables their search and discovery (Horsburgh et al., 2015)."

R1-Comment 5

Finally, there is a lot of redundancy in the manuscript, especially, between discussion and conclusion. I think this manuscript can be shortened although I ask the authors additional explanations.

- We rewrote or removed redundant parts of the manuscript and revised Section 6 to focus on the key points under four headlines
  - 6.1 How can modelers use WaMDaM database and its software?
  - 6.2 Current limitations
  - 6.3 Future work
  - 6.4 Invitation to community involvement and feedback

R1-Comment 6

Other comments:
Line 27: Avoid to use an abbreviation in abstract (USA > United States).

- We use the United Sates now

R1-Comment 7

Line 251: Provide citations or more explanation about "Several recent efforts."

- We added four citations to support the claim. The sentence in lines 135-138 now reads:
  Several recent efforts to increase data consistency and transparency, such as the Open Water Data Initiative (Blodgett et al., 2016), Observations Data Model 2 (Horsburgh et al., 2016), the Open

and Transparent Water Data Act (Cantor et al., 218; Dodd, 2016) and have recommended data standards to integrate fragmented water information data into consistent and interoperable data systems.

R1-Comment 8

Line 305: I could not make a relation between the WaMDaM and Gray's rule that you cited. Please give more explanations.

- We expanded the sentence to further describe the importance of using Gray's rule in designing WaMDaM.

  "The use cases helped guide the WaMDaM design by answering key water management data questions. These use case questions sidestep less important aspects that may overcomplicate the design.

R1-Comment 9

Lines 313-316: It might be better to have the same question form for the fifth use case like other previous four cases.

Thanks.

- We rewrote the fifth use case into this question

  5. How do the input data developed in earlier use cases affect model outputs?

R1-Comment 10

Lines 352-356: I do not understand how a relational database can share the same attribute values across the systems to improve storage efficiency. In other words, it is difficult to find physical implementation of this part.

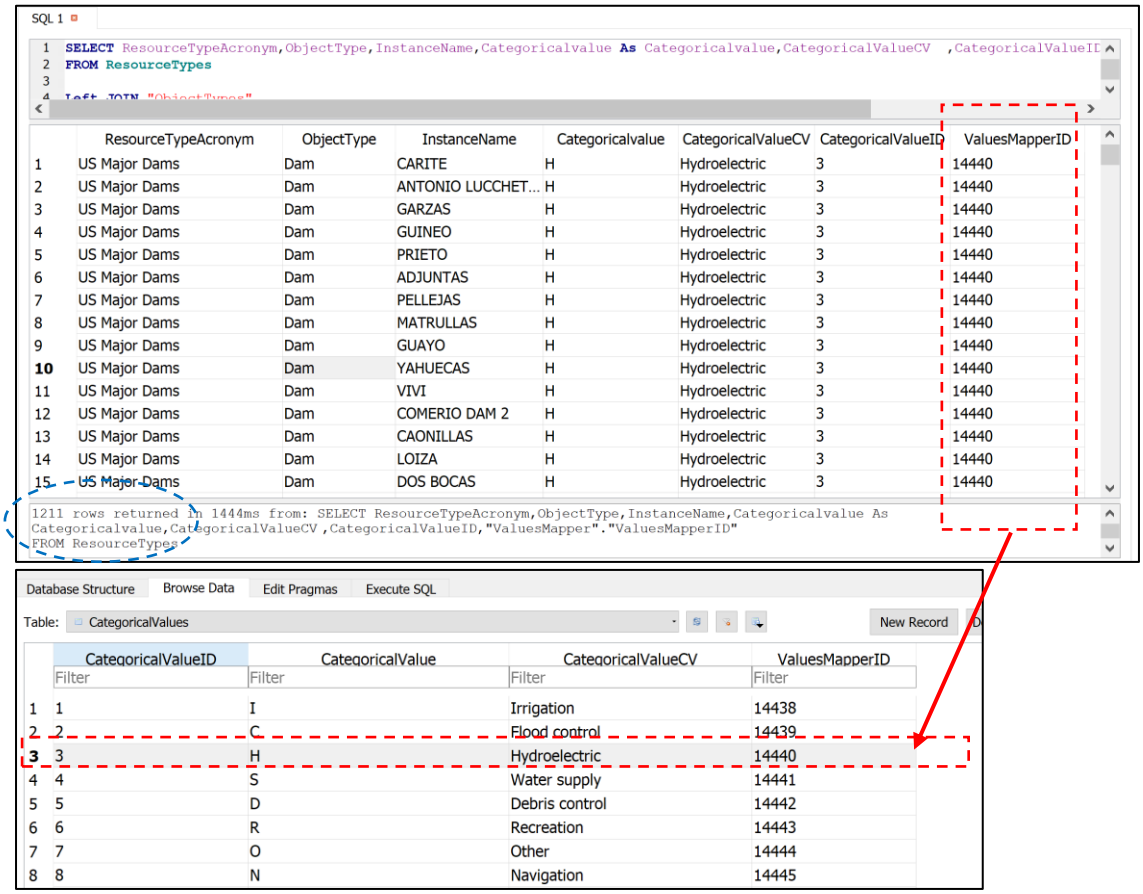
We address this comment in three points. First, we recognize that the previous use of "many systems" in the second part of the sentence "……..share the same value of an attribute across many systems and components" was likely confusing. The word "system" could refer to a water resources system model or to a database system.

- We rewrote the sentence with more specific words and an example:
  "To improve storage efficiency and enable consistent reuse of data, the data system must be able to reuse the same attribute value across multiple object instances (e.g., use same dam purpose of "Hydroelectric" across multiple reservoirs in the same dataset)"

Second, this section introduces the design requirement that is implemented later in Section 3.2 in the fifth and sixth paragraphs. Please refer to the examples in that section.

Third, we provide a simple example here that shows how the design enables sharing the same categorical value of "Hydroelectric" across 1,211 many dam instances (see Figures below). In contrast, the US Dams shapefile dataset stores the categorical value "Hydroelectric" 1,211 times. This WaMDaM approach of storing values once and sharing them is more efficient and allows the option to register the term one time with a controlled vocabulary.

Below, we provide an excerpt of query results to show how the ValuesMapperID enables sharing the same categorical value “Hydroelectric” across 1,211 dams. Note that ValuesMapperID points to one Hydroelectric categorical value that is shared for all the Dams in the US Dams data source.



SQL 1

```
SELECT ResourceTypeAcronym,ObjectType,InstanceName,Categoricalvalue As Categoricalvalue,CategoricalValueCV  ,CategoricalValueID
FROM ResourceTypes

Left JOIN "ObjectTypes"
```

	ResourceTypeAcronym	ObjectType	InstanceName	Categoricalvalue	CategoricalValueCV	CategoricalValueID	ValuesMapperID
1	US Major Dams	Dam	CARITE	H	Hydroelectric	3	14440
2	US Major Dams	Dam	ANTONIO LUCCHET...	H	Hydroelectric	3	14440
3	US Major Dams	Dam	GARZAS	H	Hydroelectric	3	14440
4	US Major Dams	Dam	GUINEO	H	Hydroelectric	3	14440
5	US Major Dams	Dam	PRIETO	H	Hydroelectric	3	14440
6	US Major Dams	Dam	ADJUNTAS	H	Hydroelectric	3	14440
7	US Major Dams	Dam	PELLEJAS	H	Hydroelectric	3	14440
8	US Major Dams	Dam	MATRULLAS	H	Hydroelectric	3	14440
9	US Major Dams	Dam	GUAYO	H	Hydroelectric	3	14440
10	US Major Dams	Dam	YAHUECAS	H	Hydroelectric	3	14440
11	US Major Dams	Dam	VIVI	H	Hydroelectric	3	14440
12	US Major Dams	Dam	COMERIO DAM 2	H	Hydroelectric	3	14440
13	US Major Dams	Dam	CAONILLAS	H	Hydroelectric	3	14440
14	US Major Dams	Dam	LOIZA	H	Hydroelectric	3	14440
15	US Major Dams	Dam	DOS BOCAS	H	Hydroelectric	3	14440

```
1211 rows returned in 1444ms from: SELECT ResourceTypeAcronym,ObjectType,InstanceName,Categoricalvalue As
Categoricalvalue,CategoricalValueCV ,CategoricalValueID,"ValuesMapper"."ValuesMapperID"
FROM ResourceTypes
```

Database Structure | Browse Data | Edit Pragmas | Execute SQL

Table: CategoricalValues — New Record

	CategoricalValueID	CategoricalValue	CategoricalValueCV	ValuesMapperID
	Filter	Filter	Filter	Filter
1	1	I	Irrigation	14438
2	2	C	Flood control	14439
3	3	H	Hydroelectric	14440
4	4	S	Water supply	14441
5	5	D	Debris control	14442
6	6	R	Recreation	14443
7	7	O	Other	14444
8	8	N	Navigation	14445

R1-Comment 11

Line 488: A limited documentation of source codes can be one of the disadvantages of open source.

We agree.

- We added the sentence at lines 269-270
  At the same time, we recognize that open-source software require documentation to be reusable

R1-Comment 12

Lines 538-542: The further descriptions can be moved to a footnote.

- We removed these lines and embedded the link to the schema within Figure 4’s caption.

R1-Comment 13

Lines 594-598: It would be better to revise Figure in order to show metadata and its related components. The current figure seems that only attributes and scenarios have metadata.

Lines 681 and 700: Can I relate the ValuesMappers and Mappings in Figure 2? If not, how about to illustrate these objects?

We moved Figure 4A from the Appendix (now Figure 4 in the paper) to better connect the described ideas with the implementation and to show how the metadata elements are mapped to Instances and data values (for each attribute and instance). Figure 4 also emphasizes an important contribution of the work which is the design and implementation of this logical model.

Moving Figure 4 to the manuscript also answers the reviewer's question and shows that the ValuesMapper and Mappings are related. Here the ValuesMapper table allows values for each of the seven data types to be associated with many *Scenarios* and many *Instances*. At the same time, the *Mapping* bridge entity allows one attribute to have one or many values (e.g., a dam purpose can be "irrigation" and "hydropower" values. We point out that the fifth point in section 3.2 describes how and justifies the need for how the ValuesMappers and Mappings are connected.

The seventh point in Section 3.2 further describes the implementation of metadata tables.

Metadata is connected to 1) node and link instances within a scenario (the source and method that created a node or link), and 2) values of attributes of instances within a scenario (e.g., the source and method of a Hyrum Reservoir time series storage existing conditions scenario)

Here as described in the seventh point in Section 3.2, "each source or method is associated with a person (author) who set up the source or created the method. Each person belongs to an organization."

R1-Comment 14

Line 648: I lost here. What are the six key features?

Sorry. We now refer to "requirements" to connect with the previous descriptions. We use the word "requirements" throughout instead of "features" to avoid any confusion. We modified the sentence to the following:

Here we describe the first six design requirements introduced in Section 2.1 that are needed to interconnect schema components and specify the fourteen required elements.

R1-Comment 15

Lines 877-879: it might be better to move to a footnote with URL.

We kept this URL in text. Footnotes sometimes can be distracting.

R1-Comment 16

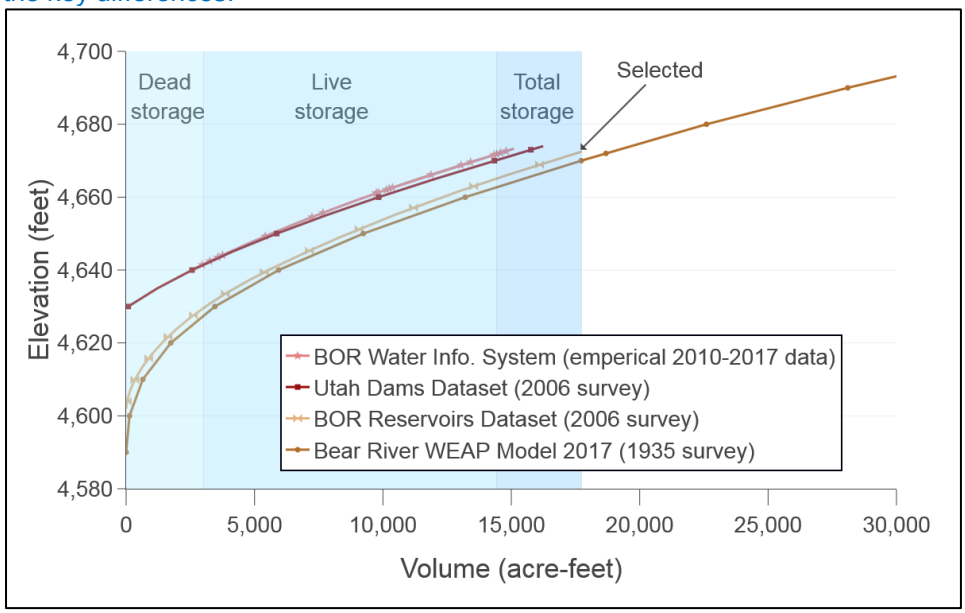
Use Case 3.2: I could not find Figure 7A and B.

We apologize for this typo. We removed the reference to A and B subplots and the text now points directly to Figure 7.

R1-Comment 17

Line 1355: I do not know which curve shows the bathymetry survey in 1935.

We added the bathymetry survey year into the legend for each curve. We removed the old bottom dashed brown curve that is identical to the current bottom line to simplify comparisons and focus on the key differences.



R1-Comment 18

Lines 1411-1415: It would be good to add another table for BRSDM, in order to help audiences.

We added this table to the Appendix as Table A4. We elected not to have it within the paper for two reasons 1) the text describe its content (i.e., identical network, and 80% share values), 2) to focus on the key unique and interesting results with a limited number of tables in the manuscript.

R1-Comment 19

Lines 1593-1597: footnote?

We deleted this URL.

-**Reviewer 2**

R2-Comment 1

- This paper presents a data model, manifested as a relational database model, for water resources planning models (i.e. simulation and/or optimization models of water systems organized around nodes and links and associated attributes). This is sufficiently novel to warrant publication, with minor revisions, and is relevant to EMS.

Thank you. We address the comments and revised the manuscript.

R2-Comment 2

I have attached my comments for specific areas/lines in the manuscript in an attached Word file (with my PDF comments extracted by an online comment extractor tool); page numbers refer to PDF page numbers. I have three general comments.

Thank you for extracting the comments. We respond to each of them below.

R2-Comment 3 (we compiled all the relevant comments here)

First, it is unclear which tools should be considered as presenting a novel contribution of this work. Several tools and scripts are mentioned, but they seem to be either vague in function or not particularly important. On the other hand, the Excel-based data loader is presented as seemingly important (e.g., as in the Graphical Abstract), but not really fully described. A more explicit enumeration of related tools/scripts should be included.

Second, and related to the first comment, given that it seems the most important contribution of this work is really the data model, more emphasis should be placed on representing the data model more clearly in the figures, while at the same time separating/generalizing tools/scripts from the data model. Though the Excel data loader seems useful, is it really a novel contribution?

If so, shouldn't there also be a data extractor, to connect the data model with actual water system models? As is, the entire data workflow does not appear to be fully represented; loading data is shown, but retrieving data is not. I could be missing something, but in any case the entire typical workflow (disparate data through to water system model & output/decisions)--and where the key contributions of this work fit in to that workflow--should be made quite clear; presently it is not.

Figure 1: I suggest two modifications. First (trivial), stack vertically to more efficiently compare current vs WaMDaM and give more space for larger text.

Second, consistent with my other comments, I suggest a box (or sets of boxes) on the right side of the "A generalized data model" in panel B representing scripts/APIs/SQL/etc., which are needed to move data from the data model (implemented as a database) to water system-specific functional needs (data, network connectivity, data, etc.).

I also suggest revising the Excel Workbook component to a more generic "data loader", as the Excel workbook seems somewhat arbitrary; presumably any script/API could be written to load data, and the Excel Workbook is just an example developed for demonstration, and not necessarily a core theoretical contribution of this work.

Graphical Abstract: See my comment for Fig. 1-B and update this figure as needed.

Great suggestions to better describe the manuscript contributions! Please see the new Figure 1. The new figure now clearly shows the generic workflow of Importer and Exporter software tools that depend on scripts, SQL, and API.

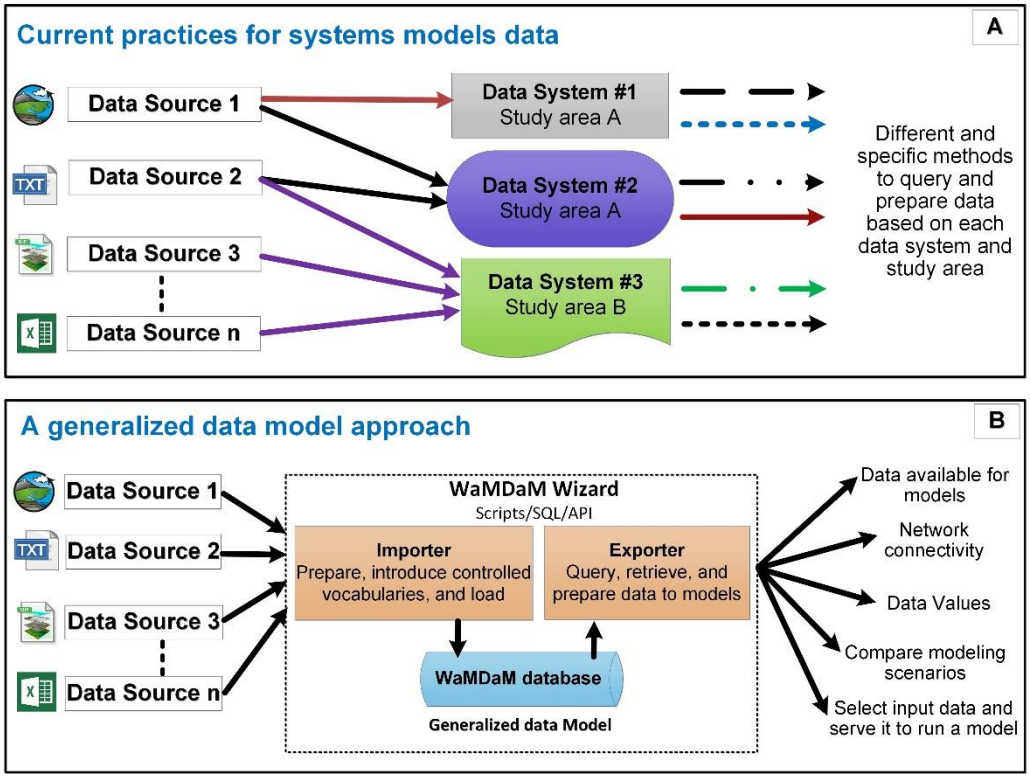


Figure 1: (A) Current data practices use different systems and data manipulation methods for each data source and study area while (B) a generalized data model integrates across the structure and syntax of data sources and allows modelers to undertake multiple efforts such as identify data for models, compare networks, data values, and scenarios, and serve data to models.

The reviewer is also correct, the core contribution of the presented work is the design, description, implementation, and demonstration of uses case of the Water Management Data Model (WaMDaM). The presented software tools support this contribution and enable users to interact with the data model by loading data from difference sources then querying it. We only provide minimal description of these tools because they are essential to the usefulness of WaMDaM design. The mentioned tools are just one possible implementation and serve as an example to other potential tools.

We made the following changes and additions to better communicate how the software tools support the contribution of WaMDaM.

- We added this first sentence into Section 2
  "We focus on addressing the essential first step in modeling which data analysis and synthesis with the ultimate goal to prepare and serve data into models"

- We rewrote the following sentence in Section "4. WaMDaM Related Software" to focus on the purpose of the mentioned tools.
  "We created software tools to demonstrate WaMDaM's functionality and allow users to more easily interact with its database."

- We added this sentence
  "The Wizard provides key functionalities of the design and it is just one of many of possible ways to import or export data of the database"

- We rewrote the third paragraph in section 4.1 "WaMDaM Wizard" to enumerate the software tools. The paragraph now reads:
  The Wizard has tools to i) prepare and pivot a shapefile, time series, or seasonal data into the data structure of the workbook template, ii) import time series stream flow data from WaterOneFlow CUAHSI web-services, iii) import time-series WaterML files for reservoir inflow,

release, storage, elevation from the U.S. Bureau of Reclamation (USBOR) Water Information System web service (https://water.usbr.gov/), iv) import network data stored in WEAP using its Application Programming Interface (API) into the workbook template, v) use the provided controlled vocabularies in the workbook to register and relate native terms across sources as discussed in Section 4.2 , vi) adapt and use the example Jupyter Notebooks to run data query and analysis across data sources and serve data into the model, and vii) compare and verify differences in topology or input data values across modeling scenarios.

R2-Comment 4
Finally, it is somewhat ambiguous how the data model can be used to solve the data management problems it presents in a way that doesn't result in continued distribution of data in different formats / vocabularies, etc. More specifically: how is the data model intended to be implemented/deployed? If it is implemented on a single user's desktop, then data residing on that desktop's WaMDaM instance must be exported for use elsewhere, in which case it would be exported as, say, CSV files, Excel files, etc., in which case we end up where we started in terms of disparate data formats, etc. On the other hand, if the data model could be deployed as a central database on a server--likely accessed via a web API--then the database could serve as the "final" destination of the incoming data. However, in this case one needs much more than simply a water network data model; a user management system with permissions, etc., is also needed. Though this veers from the theoretical, the deployment and access intent that supports the data management challenge addressed should be noted, along with potential future data model extension needs to make that deployment possible.

We rewrote Section 6 and introduced section 6.1 with a header that addresses this comment. "How can modelers use WaMDaM database and its software?"

As mentioned in Section 3.3 and 3.4, SQLite was selected as a simple demonstration of the data model.

We added a citation to the GitHub repository DOI that directs the readers to the three other implementations of the WaMDaM blank databases (Abdallah, 2018b).

We agree that a web-server with full web services would increase the value of this work by offering the stored data into a wider audience of modelers but that is out of the scope of this paper. We see that as an important next programming step that is now discussed in Section 6.3 of Future Work. Section 6 also discusses how a populated SQLite database can be shared publically on HydroShare with a permanent Digital Object Identifier (DOI) where it can be discovered and reused by others.

Excel, csv, and text files will continue to be standard format. A well-structured and described excel sheet like the WaMDaM workbook template or files output by WaMDaM can be an intermediate medium to input data or share query results. This use is similar to output obtained from downloading the time series data query from CUAHSI (provide link to Hydro Client @ https://data.cuahsi.org/)

R2-Comment 5
In addition to addressing these issues, the manuscript should be double checked for grammar/punctuation/etc. as always.
An external Professional Engineer and native English speaker double checked the manuscript for grammar and punctuation and we made changes where needed.

R2-Comment 6
Lines 141-144: While these points are all valid, there is also some benefit from struggling with data in the same way that the model building process itself is beneficial beyond model results. Building models results in a better understanding of the system, and can facilitate collaboration (e.g., in participatory modeling or shared vision modeling). Similarly, struggling with data management forces a continual renewal of individual or institutional knowledge about that data, despite the inherent data management / modeling inefficiencies involved. It's worth noting that data management inefficiency ("wrangling") isn't always undesirable.

We agree that data query and synthesis are essential to building models and that wrangling with the data can help in understanding the water system. The wrangling problem that the WaMDaM data model and its implementation address concerns the large number of methods used to work with different file formats, structures, and semantics which all can cause confusion, be prone to errors, and time consuming. Also, each new user must again wrangle with these problems. Users can benefit from a data system that helps to reduce this type of data wrangling.

Users will still wrangle with selecting the appropriate controlled vocabulary (use an existing term or a new term? If a new term, what term to use?). This type of wrangling is a good thing as it forces the user to figure out how to relate their data to existing data. Also after wrangling, there is a product: the data are related, a controlled vocabulary can be used to relate the data to other data, and other users can further build on these products.

We feel that data wrangling over file formats and structures is better invested in querying data in the WaMDaM database (or a Web-service in the future) and running a full spectrum of analysis and comparisons that are possible because of the data organization and access.

- We have updated the sentence to read:
  "A common database design to organize and manage water resources system data can help modelers and managers spend less time to wrangle with data formats and structures and more effort on analysis to learn about the system and model it."

- We also edited line# 919-920 to better reflect this idea.
  "Modelers then can spend more time on data analysis and synthesis than on time consuming and error-prone steps to manipulate data to set up and run a model."

R2-Comment 7
Line 182 (and elsewhere): I believe it is Hydra Platform, not HydraPlatform (with space).

Thank you for catching this typo. Now we use Hydra Platform throughout.

R2-Comment 8
Section 2: There are two additional data concerns worth highlighting. First is data quality/uncertainty: often we are selecting/comparing data with potentially questionable or unknown quality or uncertainty (e.g., poorly collected data; input from another uncertain/poor quality model/method, etc.). As we may still use data of less than ideal quality/certainty in our models, tracking data source quality/uncertainty in our data pipeline becomes that much more important for not only qualifying our analyses, but also quantifying the possible uncertainty in our own analyses. How could WaMDaM account for this?

We thank the reviewer for highlighting this important data management aspect, data quality. We added a new field into the Methods entity called "DataQuality" with a "text" physical data type where the user can choose to document the potential uncertainty in the data and indicate the quality of data within the method that generated it.

- We modified this sentence at lines 409-410-to include this concept.

  "The Methods entity describes how values were created, an instance is defined, data quality, or the resource type works (e.g., simulation or optimization method for a model program). Modelers may document uncertainty in the data and indicate the quality of data by the method that generated it"

We note that WaMDaM requires providing the sources and methods of the reported data where each source and method is provided or created by a person within an organization (refer to sections 3.2). Tracking and potentially reusing metadata (source and method) for each attribute value and for each node or link, is one of the novel contributions of the WaMDaM design beyond any of the systems models and datasets we reviewed (Table 1 and Appendix A1). Thus, a user can also use metadata to follow up and ask the originator of the data about data quality issues.

Quality control is left to the user who loads data (i.e., supplies it) into the database. Users who later query the database will be equipped with information that describe the data source, method, and data quality. The user will also see the person who created the data and their organization.

There are potentially other ways to document data quality but we decided this modified design and metadata provides a good balance between what is ultimately needed and what is realistic to provide (will not burden users who want to provide data). Section 3.2 under the seventh point (lines 407-422) points out to the idea of balancing the "principles and practicality" of metadata as recommended by Duval et al. (2002).

R2-Comment 9

Second is data time step (or units) conversion. Where would/could that happen? Should it be assumed that any such conversion would be up to the end data user? Or could there be scope for such conversions within the WaMDaM framework? Here it is clearly the former (up to the end user), but it should at least be noted here that often we do not go directly from the database, no matter how perfect it is, to our model; instead, typically there is some intermediary conversion routine, similar to the data loader wizard, but for in between the database and the end use. Perhaps this is implicit within, say "select input data", but this intermediary should nonetheless be acknowledged. Without such an intermediary, the data model is not "generalized", and we end up back where we started.

Line 1585: I suggest also highlighting the need for translators for converting between time steps and units, which vary significantly between different water system models. This is not trivial, and falls very much in line with the general need to be able to convert data from a generalized data model such as WaMDaM to specific water system models.

We agree that unit conversion is important to support, especially to make the data model implementation useful for many systems models that potentially operate on the same data but use different units.

- We added the sentences in lines 428-430 to clarify how WaMDaM already supports converting units.

  "Units can be converted using a constant or linear multipliers. For example, a one litter value has a 0.001 constant fraction in reference to a 1.0 cubic meter volume unit. We adopted the list of controlled units from Hydra Platform (Knox, 2018)."

Controlled units are publically available at (http://vocabulary.wamdam.org/units/). To view all the units, click at "Download Vocabulary (CVS) button in the above URL to download all of units into a csv file. Users can submit new needed units to the controlled vocabulary as described in Section 4.2.

We also added a sentence to Section 3.2 to explain that start and end times and time step are properties of a scenario. None of the Use Cases require scenario time step conversion but converting time steps for the scenario's data is an important and challenging data wrangling task that WaMDaM could help with in future work (using metadata describing scenarios and attributes). We added a sentence in Section 6.3 to describe this future work.

R2-Comment 10

Lines 345-350: Hydra Platform addresses this as well, so this should be noted, at least as a reference.

We added a citation to include Hydra Platform. We also note that Table 1 shows how Hydra Platform already supports this requirement as indicated by the "X" sign.

R2-Comment 11
Lines 373-376: This is not true with Hydra Platform, but is true with WEAP. It seems that this sentence was meant to apply to ArcHydro, in which case probably these two sentences need to be re-arranged somehow.

- We rewrote the sentences at lines 197-201 to better reflect what each data model or system is capable of. The updated sentences are:

  "Modular and extensible design is supported in most existing data systems and water management models such as Hydra Platform and the ODM (Harou et al., 2010; Knox et al., 2014). Other systems, such as ArcHydro and WEAP (Maidment, 2002; Yates et al., 2005) allow adding new data objects as in ArcHydro, but users are still forced to use core components and attributes that might not be needed for a case study ."

R2-Comment 12
Line 411 ("change network topology"): This is a good point. The ability to track changes in network configurations as "scenarios" is something I haven't seen before.

We thank the reviewer for pointing out this very important aspect of the WaMDaM design.

R2-Comment 13
Line 478: It is unclear here what is meant by "conditional data queries". Does this imply unfettered access to the database via, say, SQL or some SQL wrapper? If so, this can be extremely dangerous in a generalized database with multiple users who might do irreversible harm to the data, intentionally or otherwise. Free query of data might also be undesirable if data access should be controlled. Is there an assumption that all data should be accessible to all users? These issues should be clarified.

- We better explain this concept as "direct access to all data". We rewrote our explanation in lines 257-263
  Seventh, the data system must support direct access to subsets of data and metadata that enable their search and filtering based on a schema. In contrast, unstructured data storage known as the Binary Large OBject (BLOB) formats (Sears et al., 2006) do not allow direct access to subsets of stored values but rather to the block of data. Although storing BLOB data such as blocks of time series or arrays as in Hydra Platform and HEC-DSS (HEC, 2009) can be efficient and fast, users are only limited to such systems custom functionality to decode and access subsets of the content.

In the current implementation of WaMDaM data model in a local SQLite copy, users have full access to the database through direct SQL as in using the https://sqlitebrowser.org/ or using Python or other programing languages.

We note that contribution of this work is the design and demonstration of WaMDaM and yet there is great room for future work to develop web-services that are scalable and provide more controlled access. Such control would define user roles and maintain database integrity. We clarify this idea in Section 6.2.

R2-Comment 14
Lines 491-498: The purpose of these lines (starting with "Examples") is unclear.

- We rewrote the sentences and moved them to lines 550-555 (results).

  "The use cases apply one optimization and two priority-based simulation models for the Bear River study area: 1) the Watershed Area of Suitable Habitat (WASH) model that allocates water to maximize watershed habitat areas (Alafifi and Rosenberg, In review), 2) the Bear River Systems Dynamic Model (BRSDM) (Sehlke and Jacobson, 2005), and 3) WEAP model."

R2-Comment 15
Figure 2: I recommend double-checking the figure and color references to accommodate those with color blindness, perhaps with line types in addition to color. Should this figure be aligned with the schema diagram in the appendix? Also, black on dark colors is more difficult to read.

- We changed the font color for text inside the boxes to white. We tried our best to align the locations and colors of the conceptual diagram (Figure 2) with the logical diagram (Figure 4). The blue tables are for resource structure to the far left, green tables for networks to the top middle, orange for metadata to the bottom middle, and red for data values table at the far right.

The color scheme we used is designed for color-blind readers by the Department of Geography, University of Oregon (scheme type: Stepped-sequential scheme, 5 hues x 5 saturation/value levels) http://geog.uoregon.edu/datagraphics/color_scales.htm#Other%20Schemes

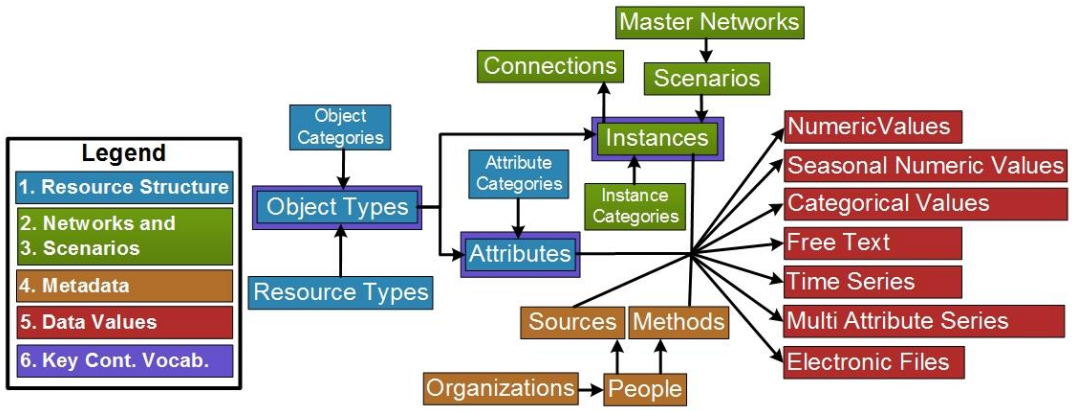


Figure 1: The conceptual diagram relating the first six design requirements for the water management data model. Key controlled vocabularies are introduced to the boxes outlined in purple.

R2-Comment 16
Lines 1091-1104: My interpretation of this is that Hyrum Reservoir is associated with two different networks. Is this interpretation correct?

Yes that is correct.

In this case, the result is a data model comprising a network of networks, which share some nodes, which is a very positive contribution. It might be worth mentioning here how data associated with different water system models would be identified. i.e., are two different datasets representing the same attribute associated with a particular network scenario?

It might be worth mentioning here how data associated with different water system models would be identified “i.e., are two different datasets representing the same attribute associated with a particular network scenario?”

Figure 4: I suggest also clarifying whether the reservoir in each water system model is represented as the same physical data object in the data model.

- We also added Figure 3 and its caption along with these two sentences to clarify this concept. *“Each resource type (e.g., model) has its own native terms. Data of different models can be related using three controlled terms, object type (e.g., Reservoir), attribute name (e.g., Volume), and instance name (e.g., Hyrum) (Figure 3).*

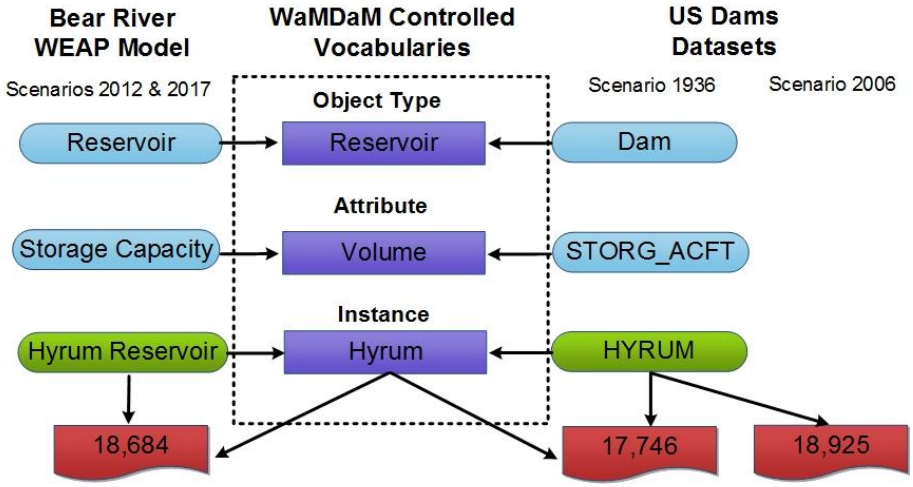


Figure 2: Relating native names with controlled vocabularies for object types, attributes, and instance names allows modelers to query and simultaneously access values across native terms. Identical storage are shared among scenarios of the Bear River WEAP Model while different storage values in the US Dams Datasets are stored separately.

We offer this query result that demonstrates how the search for object types that are registered with the controlled term "Reservoir" and instance name "Hyrum Reservoir" returns all the Hyrum reservoir instances in different datasets.

	ResourceTypeAcronym	MasterNetworkName	ScenarioName	ObjectTypeCV	ObjectType	InstanceNameCV	InstanceName
1	US Major Dams	US Major Dams	US Dams As-Is	Reservoir	Dam	Hyrum Reservoir	HYRUM
2	RWISE	RWIS Western States	RWIS data as-is	Reservoir	reservoir	Hyrum Reservoir	HYRUM RESERVOIR
3	WEAP	Bear River Network	Bear River WEAP Model 2017	Reservoir	Reservoir	Hyrum Reservoir	Hyrum Reservoir
4	WEAP	Bear River Network	Bear River WEAP Model 2010	Reservoir	Reservoir	Hyrum Reservoir	Hyrum Reservoir
5	WASH	Lower Bear River Network	base case scenario 2003	Reservoir	v	Hyrum Reservoir	j29
6	NHAAP	National coverage of online U.S. hydro...	Existing Hydropower Assets Database ...	Reservoir	Hydropower Plant	Hyrum Reservoir	Hyrum
7	BOR	Hyrum	Base case	Reservoir	Reservoir	Hyrum Reservoir	Hyrum Reservoir

Query:
https://github.com/WamdamProject/WaMDaM_UseCases/blob/master/4_Queries_SQL/Specific_Instance_Implementations.sql

R2-Comment 17
Line 1533: Regarding "the scripting features": it is unclear what this means

- We rewrote this sentence at line 909 to more accurately describe what is being done.
  "The WEAP API and SQL make it possible for users to use WaMDaM to set up scenarios, replicate, and extend the work."

R2-Comment 18
Line 1452 notes a "query method". Several scripts have been noted, including the loader wizard, a Jupyter Notebook script, etc., but it is unclear how these scripts actually connect to the database. Is it specifically using SQL (the "query method")?

The current mechanism for reading data should be clarified explicitly, and used as a basis for recommending a standardized API, as direct SQL queries is probably not an ideal way to access the data, unless the database is intended for few users to access.

- We replace the two words: "query method" in this line with "SQL queries"
  "The WaMDaM CVs, consistent data storage and SQL queries enabled selecting the…."

- We added this sentence to section 4.1 to explain how data is inputted into WaMDaM and queried. *"The WaMDaM Wizard uses SQLAlchemy to load data into the database and we use direct SQL script to query the database through a Python SQLite3 library."*

R2-Comment 19
This lack of clarity extends to the following paragraph (lines 1539-1541): "a user can access a consistent set of tools to store, organize, query, compare, select, visualize, and share water resources data". What are these tools? Does this refer to the Excel data loader? This should be clarified. See also my earlier comment related to use of SQL.

- We rewrote the third paragraph in section 4.1 "WaMDaM Wizard" to enumerate the software tools as *addressed* earlier in R2-Comment 2.

- We rewrote Section 6.1 to address these various tools. The text now describes 5 recently published systems analysis studies and explains specific WaMDaM tools each study could use to help better *manage* study data.

R2-Comment 21
Lines 1584-1597: Related to my previous comments related to SQL access, the intended use scope of WaMDaM should be clarified. Would the intent be for use by a single user? A small, controlled group of users? Or the world--enabled by a controlled web API with a user management system? If a larger (global) user base is intended, I think there's no way to avoid the need for an API wrapper around WaMDaM. This should be discussed.

We could not agree more. Currently, WaMDaM use is by a single user on a desktop. We have added Section 6.1 to clarify potential uses of WaMDaM while Section 6.3 discusses the need for an API design to distribute WaMDaM capabilities and services to a large audience and also protect its integrity as mentioned in earlier responses in R2-Comment 18