International Environmental Modelling and Software Society (iEMSs)
7th International Congress on Environmental Modelling and Software
San Diego, California, USA, D.P. Ames, N. Quinn (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Usage of the Deltares Open Archive

Gijsbers[1], Peter, Andre Grijze[1], Marc Van Dijk[1], Onno Van Den Akker[1]
[1] *Deltares, P.O.Box 177, 2600 MH Delft, The Netherlands*
(peter.gijsbers-at-deltares.nl)

Abstract: Water agencies conducting daily operational forecasting need an archive solution that enables them to review (and legally defend) forecasts and decisions made, conduct post-event performance analysis, create canned datasets for training and create datasets for model calibration and hindcasting. Each use case has different storage needs, both in terms of data types stored, moment of archiving, backup and end-of-life strategies. In addition, each use case has different discovery and retrieval needs, sometimes followed by additional data creation (e.g. a post-event evaluation report) to be stored in relation to the original data. Data discovery strategies for scientific purposes typically start with an area and period of interest, followed by more query details such as locations, variables and quality/status indications. Many of the operational use cases have an additional item in common: they can be related to an 'event', e.g. a storm, flood, drought or spill event. An event can be labelled by a meaningful name, covers an area and a period of interest (e.g. Hurricane Sandy hit the NJ-NY coast around 28 October – 31 October 2013). Events can thus be used to discover relevant data in the archive. Events can also be related to use cases, e.g. training events or calibration events. Since use cases differ in data needs, as well as end-of-life and backup strategies, these can be related to events as well. Using the above mentioned concepts, this paper discusses the intended usage of the Deltares Open Archive in daily operations and science.

***Keywords**:* open; hydrologic archive; forecasting; events; use case

1 INTRODUCTION

Many operational water agencies recognize the need for an archive with water related data. When being asked what portion of the data collected needs to be archived, the typical answer is 'everything'. This answer is driven by a lack of deeper consideration what data actually is needed for what purpose, how long a dataset is actually relevant for that purpose and how users actually expect to find the data when the archive has become many terabytes in size. Storing 'everything' will also become very expensive, in hardware, in labour to keep the data available on appropriate devices and in the ability to find the data as needed.
To design an appropriate data archive solution, a more detailed use case analysis is needed which identifies:

- what specific uses exist for the archived datasets
- what portion of the operational dataset needs to be preserved for a specific use
- how long the dataset needs to be preserved for this use
- how people intend to search for the dataset they need
- what the data accessibility requirements are (e.g. in terms of standardized or proprietary data formats)
- what the performance requirements are both in terms of discovery and retrieval speed

The end uses determine what datasets and metadata needs to be stored, thus having a major impact on the solution chosen.

This paper discusses the intended usage of an archive for water agencies. These use cases are to a large extend inspired by discussions with various operational water agencies, most notably the Bureau of Meteorology (Australia). The analysis made clear that a hydrological archive for operational water agencies should go beyond the ability to store time series data and rating curves. Also text products, reports and communication records as wel l as model run governance data should be stored in relation to the activity conducted. Existing hydrological archive systems such as WISKI (Kisters, 2014), Aquarius (Aquatic informatics, 2014) and CUAHSI HIS (CUAHSI, 2014) do not meet those needs. Hence a new archive system is being developed by Deltares. Technical details of this archive are discussed in more detail by Grijze et al. (2014). The Australian Bureau of Meteorology is

the 'launching customer' where the Deltares archive is implemented as part of the implementation of HyFS, the new operational Hydrological Forecasting System of the Bureau. The implementation of HyFS includes migration of the existing Australian HyModel archive to the new Deltares archive. Deltares is also co-operating with the Dutch Rijkswaterstaat to prepare migration of the current Matroos forecast archive to the new Deltares Open Archive. More implementations, often in combination with Delft-FEWS based forecasting and water information systems, are in the pipeline.

2 THE NORMAL AND THE ABNORMAL

2.1 The 'normal' situation

Normal day-to-day water management practice deals with water conditions which may be considered routine to the water managers. Observational data is gathered to operate the system. Usually this data is also archived to enable water systems analysis task as such as modelling, trend analysis, statistics and reporting. While numerical weather predictions are used as part of the operation, the water agency seldom sees a need to archive this dataset, partly because they expect the meteorological service to archive such datasets. Data quality often is a big challenge for hydrological archives, especially when the day-to-day routine involves little quality control. An archive which has poor quality data in store reduces its usefulness as an additional effort is needed at usage time to generate a dataset of suitable quality for water systems analysis and modelling.

2.2 Events: the abnormal situation

The whole situation changes when daily practice is disturbed by more abnormal or even extreme weather events. Deviations from the normal situation require heightened attention or even action from the organisation. Good quality data becomes more important during operations. In addition, accountability of the organisation for its actions gives them an interest to archive more data which is relevant to the event. Events can vary in duration and geographic extent, just like the data related to monitor and forecast the occurrence and onset. Which data is relevant to archive thus partly depends on the nature of the event.

A thunderstorm on a warm summer day, typically originating from a convective cell, is of short duration (minutes-hours) and may be very local, causing local drainage problems or flash floods. Relevant data sets include water level observations, rainfall radar images (observed and now-cast) and where feasible supported by high resolution numerical weather predictions.

Wind driven events, such as extreme storm surges may last one or two tidal cycles covering an extensive coast line. Relevant datasets include tide tables, field and buoy observations and forecasts using hydrodynamic and wave models, driven by forecast wind fields.

Storms may also carry heavy rains, causing drainage and flooding problems, either local or downstream. The duration of the event may vary from days to weeks, depending mostly on the response of the catchment. Catchment response is strongly affected by the initial condition before the weather event arrives. Hydrologic forecasts thus depend on observations reflecting the current state (wetness) of the system, in combination with weather forecasts.

River floods may also be caused by remote excessive snow melt, where the initial condition (and extent) of the snow pack, in combination with the long-term temperature forecast are most important to predict the melt process and the resulting runoff.

Drought events can last for years. Dependent on the catchment characteristics and topography, groundwater conditions or the snow pack in the mountain regions may play a major role in the occurrence and extent of a drought. Observations reflecting catchment conditions, in combination hydrological models using seasonal and climatic input can provide a basis for decision making. As can be seen the kind and extend of data to archive thus partly depends on the nature of the event.

3 ARCHIVE USE CASES FOR WATER MANAGEMENT AGENCIES

3.1 Reviews and inquiries on event handling by the water agency

As indicated, events require heightened attention or even action from the organisation. Given the accountability of organisations for their actions, they may be faced with post-event reviews or even formal inquiries by external parties. Such inquiry will extensively focus on the decisions and actions taken (e.g. warnings issued, dam releases, flood defence actions or evacuation) with emphasize on the timeline of information availability that guided the decision making. To support a post-event inquiry or review, an archive thus should hold information that allows reconstruction of such timeline. This includes both scientific data as well as non-scientific data such as forecast products and records of communications. This is different from a need for exact reproducibility of results or even storage of all data including intermediate calculation results.

The scientific data record should provide insight in the observations and numerical weather predictions that were available at the moment the hydrological forecast was produced. This, together with the model configuration used for the calculation, the initial model states, the manual settings used by the forecasters and the final calculation results should enable forecast reconstruction without the need to store all the intermediate model results. At least as important for an inquiry, if not more important is the non-scientific data record. This should assist in the reconstruction how the forecasters came to their final issued forecasts, what communication was conducted when and to whom and how and when the decisions were made and communicated.

To conclude, an archive which is supposed to support a post-event inquiry or review should hold from start to end all scientific (observations, available NWP forecasts, hydrological forecast calculations and issued forecasts) and non-scientific data (forecast products, records of communications) to allow reconstruction of a timeline for decision making. Post-event addition of the review findings to the archived dataset would make it more valuable for future use.

Organisations frequently have a legal requirement to keep such dataset for a legally prescribed period of time. A review or inquiry is a process which may take a few days or weeks. Completeness, readability and discoverability are the most important requirement, while it is acceptable when the retrieval process takes some time. Devices may be chosen which fit those needs.

3.2 Training material

Operational water agencies need to train their staff in the forecasting and warning process. Preferably this is done with realistic training material. Records of interesting events may provide suitable material, especially when the dataset has the extent as required for the review/inquiry case. Training may be a reason to keep the dataset for a longer period than the legally required term for review and inquiry. Given that training preparation normally does not require urgent data access, similar requirements can be defined for completeness, readability and discoverability and the duration of the retrieval process.

3.3 Post-event performance analysis

Hydrological forecasts performance assessment is another activity which operational organisations may want to conduct to assess their current skill and understand where they can improve their forecast capabilities. Such performance assessment only needs data for the variable of interest (e.g. water levels or precipitation). To enable such task, the archive needs observations, the final calculated forecasts and the issued forecasts as these can be used to calculate performance indicators addressing peak accuracy, lead time or timing of threshold crossing. Once the assessment is conducted, and the associated report is stored in the archive, some of the datasets may become less relevant. Post-event analysis does not require urgent access to these datasets.

3.4 Real-time diagnostic verification against historic events

Historic events provide diagnostic value to the forecaster to compare the current situation and simulated forecast with observed situations from the past (Demargne, et al., 2010). Quick and easy

access in a matter of minutes, preferably by the forecasting system, is needed to allow forecasters to use historic information during a calamity situation.

3.5 Model Calibration

Model development and calibration is a common activity conducted for water systems analysis as well as forecasting. It requires a complete, quality controlled record of observations, both in terms of forces (e.g. precipitation and temperature) and water conditions against which to compare the model results (e.g. water levels, flows, wave heights, water quality indicators). Bad quality data in the archive requires additional effort to make the dataset suitable for model calibration. Preferably such quality control effort is conducted before the data is stored in the archive. Alternatively, the archive should allow update of the dataset after quality control. Model calibration does not require urgent access to these datasets. Open access to data, preferably using standard services and/or data format, is important to enable usage by a wide range of calibration tools.

3.6 Model Verification

Generally, a calibrated model is verified against a relevant portion of the dataset which has not been used for model calibration. Model verification of a forecasting model is best done by hindcasting and comparison against forecast forcing. Since these datasets can get voluminous, data administrators may choose to store forecast forcing only for interesting events. Again, this activity does not require urgent access to these datasets, while the forecasting system is the most obvious tool using the data.

3.7 Other use cases

A variety of other water systems analysis use cases could be imagined, ranging from statistical time series analysis for trends or extreme events to model development for water systems analysis. Typically these use cases require open access to long and complete records of quality controlled observations.

3.8 Time line of data production and usage

Figure 1 illustrates what the general timeline data production and usage is in relation to the different use cases. Table 1 summarize the different data needs by the different uses cases. As can be seen Training and event review require access to all data while other use cases only need a portion. In general, a continuous record of observations is needed, while most other data types only are needed for specific periods of interest. Data relevant to real time diagnostics for forecast verification is the only dataset which requires high availability and fast access (i.e. within seconds). All other use cases can cope with slower response times for discovery and retrieval as proper planning of data retrieval from the archive can prevent delay of the work process.

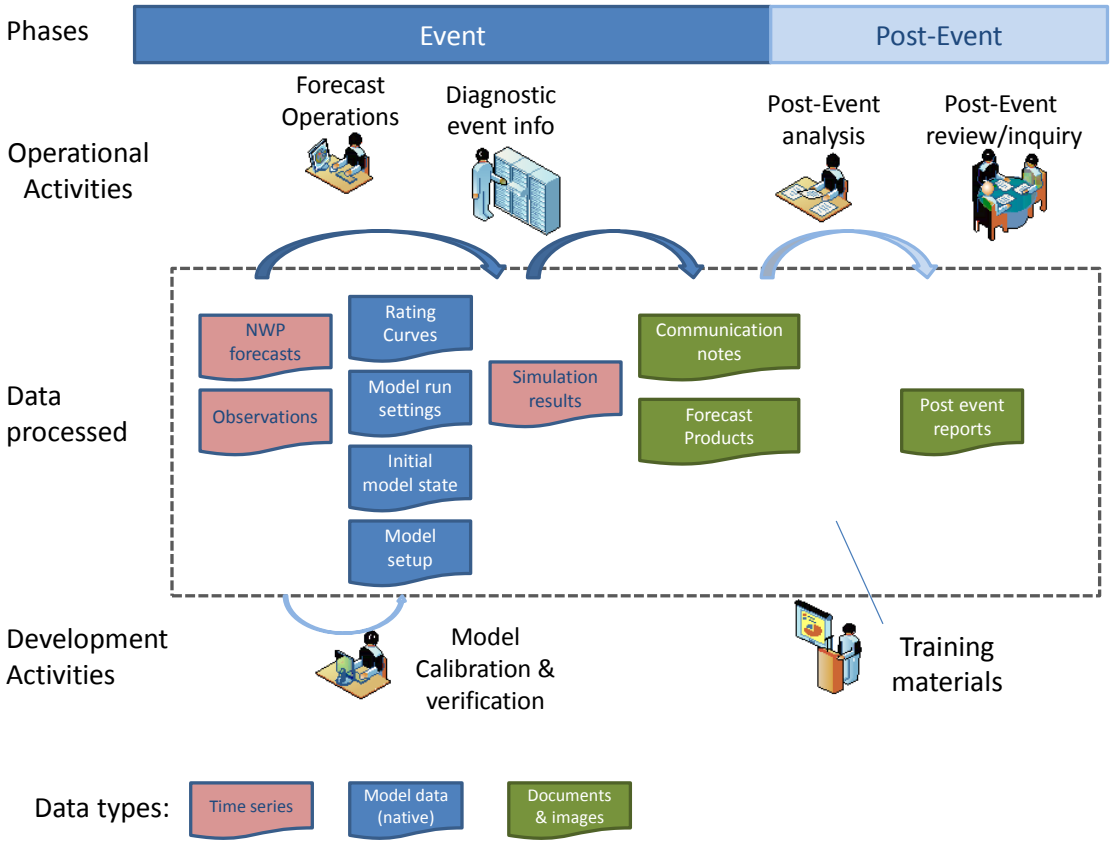

Figure 1 Data production and uses in relation to different use cases

Table 1 Relevance of data types for different use cases

Use case / Data type	Event review / inquiry	Training materials	Event analysis (Skill)	Informative event/ Diagnostic	Model calibration	Model verification
Observation	x	x	x	x	x	x
Simulation results (water forecasts)	x	x	x			
External forces (NWP forecasts)	x	x				x
Initial model state	x	x				
Model setup	x	x				
Modelrun settings	x	x				
Rating curves	x	x			x	x
Forecast products	x	x				
Communication notes	x	x				
Post event analysis reports	x	x	x			

4 ARCHIVE RELATED PROCESSES

4.1 Metadata for discovery

When the event is known by the person searching for data, keys used to discover the data are the event name (if any), the start and end date (time) of the event and the area where the event happened. When a general search for events will be conducted, the area of interest will be known, while other search criteria may be needed such as threshold crossings or value crossings. Searching

by threshold crossing makes searching easier as less local knowledge is required at the moment of searching. To enable searching by threshold crossings a metadata tagging mechanism is required to highlight occurrence of certain conditions (precipitation, flow, water level or prevailing wind conditions in direction and speed). This tagging could be done during data storage or afterwards by a local expert, or by a tool that automatically can analyse data and add the relevant metadata.

For those use cases that have no relation to events, search criteria are more focussed on obtaining long time series for relevant locations and quantities. In some situations, searches may be desired by related locations (e.g. upstream off) but this requires additional topological knowledge. Table 2 identifies the search criteria to support a use case, where M=mandatory, D= desirable, O=optional and n.a. = not applicable.

Table 2 Search criteria to support a use case

Use case **Search criteria**	Event review / inquiry	Training	Event analysis (Skill)	Informative event / Diagnostic	Model calibration	Model verification
By Date (start-end)	M	M	M	O	M	M
By Area	M	M	M	D	D	D
By Event label	M	M	M	M	n.a.	D
By Location	D	D	M	M	M	M
By Location relation	n.a.	n.a.	n.a.	n.a.	O	O
By Variable	D	D	M	M	M	M
By Threshold crossing	D	D	M	D	O	D
By Value crossing	O	O	O	O	O	O

4.2 Meta data generation

Preferably metadata is generated during storage. Some metadata can only be generated at a later moment. Tagging an event with a name, start and end date, is generally a manual activity that has to be conducted as a post-event activity. As indicated, data mining tools may be applied to enrich the dataset with more metadata at a later stage, e.g. marking threshold crossings.

4.3 Reliable production of datasets

To support inquiries, comprehensive datasets need to be stored. The most reliable solution will continuously store all relevant scientific and non-scientific data and its meta data for the most extensive use case (inquiries/training) and remove unnecessary data as soon as is known that is not needed (i.e. no event has happened). Such continuous storage process should be reliable, having fall-back options that prevent data gaps when temporarily the data store cannot be reached.

4.4 Open access

Model calibration and water systems analysis are activities that may be conducted by a variety of software tools. To increase the range of applications, data should be accessible through open industry standards (e.g. OGC based services and/or data formats) as proprietary data formats reduce the ability of researchers and developers. The archive should allow storage of data from multiple sources. If proprietary formats are used, data conversion should be applied, either at storage time to keep a consistent format in the archive, or at retrieval time to provide the data in a standard format. Any time metadata requirements should be met to enable data discovery.

4.5 Archive data management

System administrators have different devices with different capabilities available to create an archive infrastructure. Archive data management requires understanding the data volumes involved and making decisions which datasets should be stored for what period of time on which devices. A data

storage strategy is crucial to keep the archive maintainable with good performance. Event tagging for different use cases can be used to define such strategy. Preferably, such strategy can be encapsulated in instructions for a data management tool that can assist in data transfer (or removal) to the different devices.

4.6 Preserving integrity of the archive data store and catalogue

Data stored in an archive should not be lost. Preserving integrity of the data and the catalogue to find the data are essential. Solutions should be considered that accommodate restoring datasets and re-building of the associated catalogue.

4.7 Migration of existing archives

Water agencies often have existing hydrological archives. The data within these archives is valuable and needs to be pre-served when implementing a new archive system. Various strategies could be imagined, ranging from a mechanism to keep supporting the legacy system up till transfer of datasets and metadata to the new archive. Transfer has an advantage that an organisation can leave legacy technology behind.

5 ICT ORGANISATIONAL CONSIDERATIONS

Service level requirements for infrastructure availability vary per use case. Archived data used during real time forecasting operations requires nearly continuous 24/7 accessibility of the dataset with good performance. Most other use cases can cope with an archive which has slower response time and is not guaranteed 24/7. In combination with continuous growth of the data volume, the difference in use cases provides flexibility to the ICT department to fit the archive into hardware investment and replacement programs. Discovery and retrieval of important data should remain at acceptable performance while the data volume growths. Methods to manage archive growth would be beneficial to meet these needs.

Many data archives hold proprietary data which should be protected against misuse. Security mechanism thus should be incorporated in the solution to control access to the data as well as tagging of metadata (events).

A data archive has to be a sustainable solution which can last for many years. Maintainability is crucial as the archive may need to survive personal or organisational change, both in terms of developers as system administrators. Complex solutions typically involve more components with more interactions and thus more potential points of failure. Simple designs based on industry standard technologies are easier to develop, maintain, administer and repair, thus increasing the sustainability of the archive.

Backend systems, such as a data archive, are often managed by specialized ICT departments within a water agency. Typically, these departments have made organisation wide decisions for the technology stack that they are willing to support. This includes preference for hardware suppliers, operating systems and back end software such as relational database management systems and application servers. Ability to acquire maintenance services are an important consideration. Open source technology only is considered an advantage by ICT departments if support can be acquired. OS platform independency is a valuable asset at any time.

6 CONCLUSIONS

Water management agencies are often struggling with the question what hydrological data needs to be archived. Frequently they end up with archiving 'everything for ever', as they are afraid that an essential piece will be missed. The in-depth analysis of this paper demonstrates that it is not needed to store 'everything' and it is not needed to keep equal lifetimes for each piece of archived data. Normally, it is sufficient to store (quality controlled) observations with some other basic data. Preferably this data is stored and accessible 'for ever' to accommodate long term trend analysis. Only

during interesting situations (i.e. events) there is a need to store more data. Reviews and legal inquiries require reconstruction of decision timelines. This requires storage of both scientific (e.g. observations, forecasts, simulations) and non-scientific (e.g. communication records) data. Once the reviews and legal inquiries have been completed, most data may be released from the archive. Only if the material is used for training it may need longer preservation. Other purposes generally pose less extensive requirements on the data archive needs. Events may also be of interest for real time diagnostic verification of the current forecast. In this case, specific requirements are posed to fast and easy access of associated observation records.

REFERENCES

Aquatic Informatics 2014. Water Data Management Solutions http://aquaticinformatics.com/products (last accessed 12.03.2014)

CUAHSI 2014. CUAHSI HIS sharing hydrologic data http://cuahsi.org/his.aspx (last accessed 12.03.2014)

Julie Demargne, James Brown, Yuqiong Liu, Dong-Jun Seo, Limin Wu, Zoltan Toth, Yuejian Zhu, 2010. Diagnostic verification of hydrometeorological and hydrologic ensembles. Atmospheric Science Letters. 11(2), 114–122, DOI: 10.1002/asl.261

Grijze A., Gijsbers, P.,Van den Akker, O.,Van Dijk, M., De Rooij, E., Pelgrim E. 2014. Technology behind the Deltares Open Archive. In: Ames, D.P., Quinn, N. (Eds.) Proceedings of the 2014 International Congress on Environmental Modelling and Software, San Diego, California, USA, Session A2: Sharing Scientific Environmental Data and Models.

Kisters 2014. WISKI Your Hydrological workbench http://www.kisters.eu/water/software.html (last accessed 12.03.14)