# The need to manage the looming "big flood data" problem.

**C Druery[1]**, D McConnell[1], C Smythe[2]
[1]Advisian, Sydney, NSW
[2]Sunshine Coast Regional Council, Nambour, QLD

**ABSTRACT**

The Queensland Floods Commission of Inquiry (2012) recognised the need for Councils to maintain up-to-date flood information (Recommendation 2.7). Implicit within this recommendation is the understanding that flood information is not static, flood models and flood maps change over time.

There are many reasons why changes occur, these include:

- Availability of new/enhanced detailed base information (eg new LiDAR survey)
- Changes to the built environment
- Implemented flood mitigation measures
- Framework and policy changes
- New modelling approaches/techniques
- Changes to catchment characteristics (eg climate change or river mouth bathymetry)

In addition to maintaining good base flood information for land use planning and disaster management purposes, significant amounts of flood information can be produced for the assessment of proposed infrastructure.

Furthermore, each flood model produces flood information for a range of flood probabilities and sometimes a range of flood durations. For each of these there are different flood characteristics, such as depth, velocity, water surface level and hazard.

Advances in flood modelling methods, such as from 1D to 2D, and enhanced computer power allowing for the analysis of increasingly higher resolution flood model domains, means that the size of the output data from flood models is ever increasing.

The combination of all flood model and flood mapping datasets presents a "big flood data" problem for the Sunshine Coast Regional Council, and likely for all Councils across Australia.

This paper explores the approach that Sunshine Coast Regional Council has taken to develop a best practice framework for record keeping, data discovery and continuing management of flooding information.

The framework addresses issues common to all agencies in managing evolving flood data, including:

- Governance and quality
- Dataset receipt and tracking
- Rollback, decision and "point in time" data history (legal challenges)
- Archiving
- Data searches
- Management of study overlap and consistency

- Data updates vs dataset replacement
- Structured information distribution (internal and external)
- Usability and drill-down detail for differing end user needs

## INTRODUCTION

The Sunshine Coast in South-Eastern Queensland contains some 12,000km of waterways across 5 major catchments, covering an area of more than 2,200km$^2$.

The very nature and effort associated with flood modelling has logically led to the Sunshine Coast Regional Council (SCRC) gradually expanding its flood model coverage across the LGA over multiple decades. Over this time, the volume of flooding information has grown to more than 10TB!

Recognising this substantial investment in, and ongoing commitment to, flood risk management, SCRC partnered with Advisian to develop a “best practice” system to assist in managing this vast and growing dataset, as well as provide access to flooding (and related) information in a consistent format by a wide range of end-users.

The widely-used waterRIDE™ platform provided the visual and analytical basis for the system, and was augmented with the development of a “Data Manager” module to integrate and streamline the data management process. Herein, this is referred to as “the system”.

## SOURCES OF “BIG FLOOD DATA”

By its very nature, flooding data is large. Flood models are generally developed as part of formal flood studies, but may also be created as part of flood impact assessments or scenario analysis. Any given flood model represents catchment conditions at a single “point in time” and, therefore, is subject to change. However, as models are updated the superseded models remain valuable and should not simply be replaced.

Whilst these flood models may change for a number or reasons, the outcome is the creation of ever increasing volumes of flooding data.

### Flood Modelling Approaches

Flood modelling datasets are large, especially those from modern 2D models. Typical dynamic model runs occupy up to 10GB per run. Adding to the volume of information, models are generally created for a range of flood probabilities spanning from rare to frequent (eg PMF, 1000yr, 500yr, 200yr, 100yr, 50yr, 20yr, 10yr, 5yr, and 2yr).

More recent 2D modelling approaches often include 4 or 5 separate runs for each probability, being different storm durations for the same probability. This means the 10GB single quickly becomes a 500GB overall dataset. Where an LGA convers multiple catchments, the data volumes increase rapidly.

As the availability of detailed topographic information from LiDAR survey increased in recent years, so has the development of more detailed flood models. As detail increases, data volumes grow exponentially (eg data storage requirements for a 5m gridded model are 4 times larger than for a 10m model, and 25 times larger for a 2m model).

### Physical Changes to the Floodplain

Adding to the volume of flooding data at SCRC is that for any given flood study, there may be multiple versions of the model that have iteratively evolved reflecting ongoing changes to the built environment within the floodplain. As changes are proposed, investigated, approved and constructed, a new version of the model will likely be created reflecting the changed flooding behaviour.

In some cases, unexpected natural changes to the floodplain may also lead to the need for re-modelling. A typical example is significant changes to a river mouth as a result of natural accretion of sand or conversely the scour of sand following a flood event.

### Regulatory Change

Regulatory change has the potential to lead to significant changes in the volume of flood data. For example, regulatory changes associated the consideration of climate change or with the Queensland Floods Commission of Inquiry. These regulatory drivers have led to an expanded suite of analyses for the various catchments across the SCRC LGA. Further, the Floods Commission of Inquiry has formally recognised that Councils must maintain, and keep t flood information current.

### Types of Flooding

In conjunction with expanding flood modelling capacity, the recognition of "additional" mechanisms of flooding has increased data volumes. Traditional riverine and coastal inundation flooding investigation has been augmented with surface water runoff inundation (overland flow), as a legitimate source of flood risk that must be modelled and ultimately managed.

### End User Detail Requirements

A 5m flood model may be considered sufficiently detailed to provide a good representation of hydraulic behaviour across the floodplain. However, when using the information to determine flood depths across a land parcel or evacuation route, the 5m base framework may be too coarse.

Draping the model results onto a finer scale Digital Elevation Model (DEM) may yield a more appropriate level of detail for those tasks (albeit not changing the underlying hydraulics). In the case of a 5m grid model that is draped onto a 1m DEM, the volume of flooding data will increase by 25 times!

Advances in computing hardware have enabled faster model run times, however practitioners have recognised the benefits of choosing improved refinement of hydraulic models over improved runtimes. Similarly, evolution of hydraulic software (such as rain on grid modelling) are also yielding result files with a greater level of spatial detail. In both instances the improved spatial detail means larger output files.

## THE NEED TO MANAGE "BIG FLOOD DATA"

Given the very large datasets, limited corporate system resources, and increasing demands being placed on the flooding team, SCRC found itself providing access to flooding information as simple, peak grids of key hydraulic parameters through ArcGIS on a “per catchment basis”. Overlap between datasets and any inconsistencies were managed by highlighting affected areas with notes recommending confirmation from the flooding team.

Data storage limitations imposed corporately made it difficult to store flood data centrally, with information often distributed amongst available corporate storage locations and valuable time-varying data was left on external hard drives. The knowledge of where data was located was often limited to a single individual, presenting a significant corporate risk. With large amounts of flood information not readily available to other users within Council, conceptually simple flood queries had to be processed by the flooding team, when they could have been answered directly had the information been accessible in an easily usable and interpretable format to the enquirer.

In addition, with critical resources occupied with servicing such basic requests, these resources were then unavailable for strategically more important tasks, such as managing and improving the datasets.

Simply, the volume of data overwhelmed Councils capacity to manage it.

**Data History**

Flood data changes over time, with the “current conditions” flooding data only current as at a specific point in time. As new datasets become available, the prior “current conditions” are superseded. However, as the history of Councils “current conditions” flood data drives official planning decisions it is imperative that an auditable record be kept.

Such an information history becomes critical when attempting to understand the driving factors behind decisions that were made at a specific time.

**Quality Control**

As flood data volumes grew, it became an increasingly labour-intensive process for SCRC to maintain quality controlled, auditable access to flooding information. Ideally, all end users would be able to access a common set of flooding datasets and extract the same information when repeatedly querying a specific location.

However, the reality was that SCRC’s experienced flood engineers were often required to manually intervene where there was overlap between the disparate flooding datasets or to provide judgement on “what flood level should be used”, raising Councils risk profile in both providing, and its capacity to provide, flooding information, as well as significantly reducing the efficiency of the process.

**Usability and Clarity**

As datasets become large, speed of access and general usability become important considerations. Traditional GIS tools only provided access to a subset of the flooding

information with overall data access speed becoming somewhat limiting (as well as the corporately selected GIS solution being unstable for flood data management purposes).

The sheer volume of data became increasingly difficult to store on SCRC's corporate network, with data considered "non-critical" residing on USB hard drives held with the flooding team. As a result, access to such information became a somewhat inefficient process.

### Data Value

The creation of high quality flooding information requires considerable effort. In order to leverage the investment in such datasets, SCRC recognised that only providing access to parts of the datasets was stifling use of the information by various stakeholders across Council.

Through effective data management, SCRC envisioned simpler and improved access to the wealth of flooding information available – and to a greater number of users.

## SYSTEM OPERATION CONCEPT

The key operating rationale behind the system was to provide a "seamless, current point of truth" for flooding information at any location, whilst still providing ready access to quality-controlled source datasets where suitable.

This was ultimately achieved by ensuring that any decisions regarding data usage, overlap, quality, refinement etc were "burnt into" the seamless datasets being made available to SCRC end-users. This eliminated any data-specific interpretation requirements of the base datasets, allowing users to focus on their problem-specific analysis.

### Master Grids

This "burnt in" framework was implemented using a concept of *master grids,* created from supporting *source data.* The master grids are a continually evolving, seamless mosaic of *all* of Councils flooding data for a set of commonly used design Annual Recurrence Intervals (ARI's). Specifically, SCRC prepared master grids for the 10yr, 100yr ARI design floods along with the 100yr plus Climate Change scenario.

The master grids store *peak* water level, depth, velocity, velocity X depth and hazard, as well as a reference to the source dataset/study (and in turn its metadata) at each cell.

The master grids provide the end user with a single point of access to peak flood information, and a link to the underlying dynamic source information. They also provide a consistent, detailed framework on which to interrogate and analyse the flooding datasets, irrespective of the framework on which the source information was originally created (such as differing resolution model grids or flexible mesh model frameworks).

Any decisions regarding the management of overlap or different quality datasets were made at the time of creating (or updating) the master grids, thereby ensuring absolute consistency in the provision of information to end users.

**Underlying Source Data**

The source data used to build the master grids comes from the flood models used in the relevant flood studies and flood impact assessments. Access to the underlying datasets was designed to be controlled, only being made available to relevant end users.

Similarly to updating master grids, any adjustment or processing of the source datasets (eg removal of modelling artefacts) was carried out at the time of registration into the system, ensuring that all datasets registered and marked as “live” in the system were “ready to use”. Versioning (eg draft, pending, adopted) of source datasets provides a further layer of access control.

**Metadata**

Metadata forms an essential part of both the storage and usage parts of the system. Each flood study registered into the system receives a unique ID which facilitates links between all datasets in the system and their source information.

Metadata is entered against each flood study and includes:

- Study Name
- Study Author
- Study Date
- Coverage spatial extents
- Model Type
- Model Detail
- Model Quality – a measure of the quality of the modelling (resolution, model type etc)
- Study Quality – a measure of the quality of type of study (formal flood study, flood impact assessment, rapid hazard assessment, etc)
- Overall Quality – a combination of Model Quality and Study Quality
- Flooding Type (Riverine, Overland and Storm Tide)
- Design Events modelled
- Scenarios modelled (calibration, climate change, blockage)
- Catchment ID
- Locales covered by study
- Status (draft, adopted, retired, rollback etc)
- Information type (peak or time series)
- Reference GIS layers
- Study reports and references

Thorough and consistent metadata allowed SCRC to develop business rules to help end users make the most appropriate use of available information, particularly in relation to quality.

**Information Flow and Access**

Figure 1 illustrates the flow (and store) of information through the system.

Broadly, datasets reside in three locations:

1) Other Datasets: reference GIS and other spatial/Council datasets

2) Flood Data – LIVE: “current” master grids and “live” source datasets

3) Flood Data – Archived: archived master grids and source datasets

Differing levels of access are then provided to various users, according to their likely use of information and access level permissions:

- General Users: read-only access LIVE data and other datasets
- Intermediate Users: read-only access to LIVE, Archived and other datasets
- Administrators: read/write access to LIVE and Archived data stores

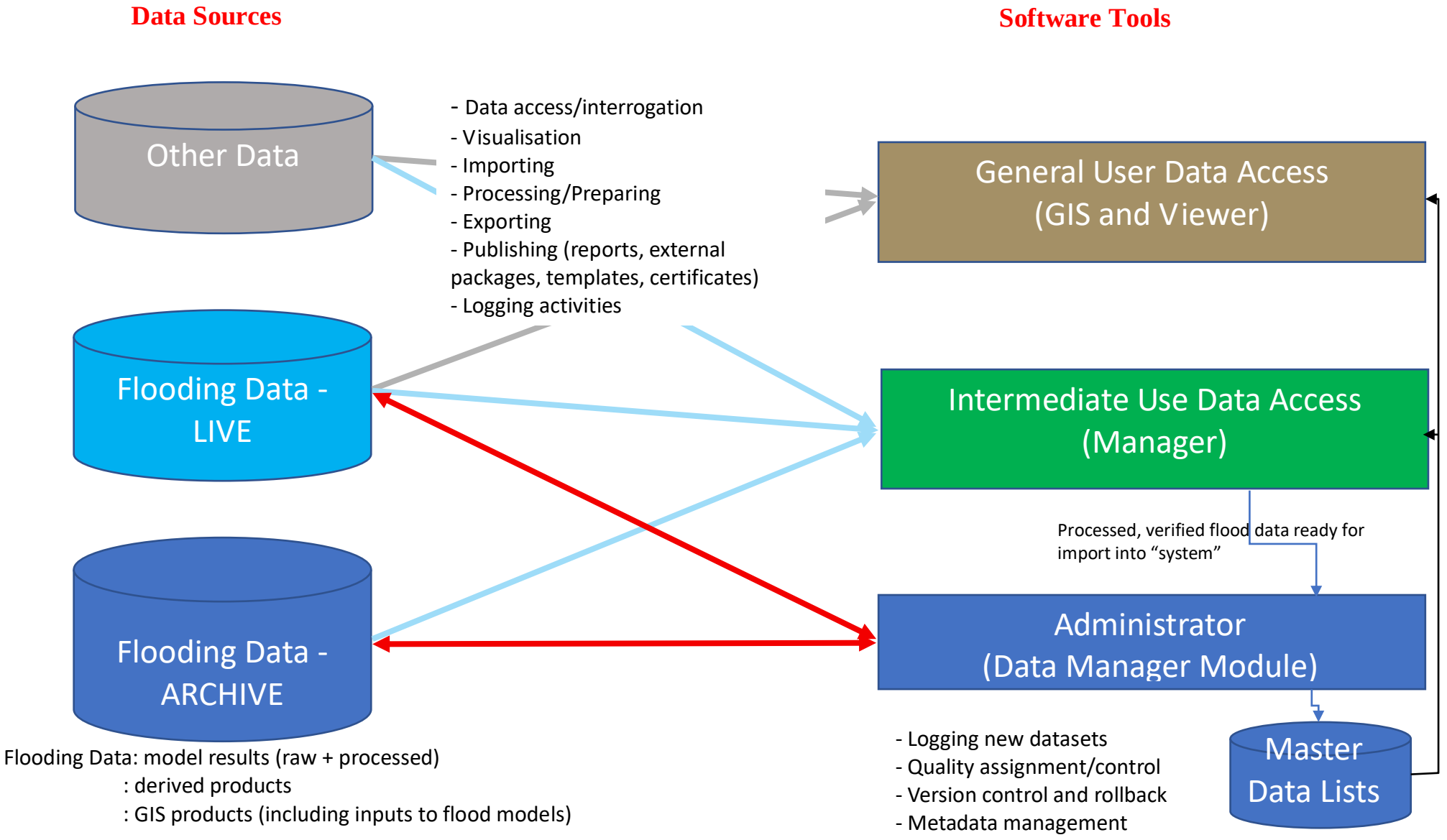


**Figure 1 – Conceptual system framework for data flow and data access**

## INITIAL DATA PREPARATION AND IMPORT

Most flood modelling outputs require further processing to remove any artefacts of the modelling process that could cause confusion to “non-modeller” end users. When preparing source data for inclusion into the system, it is important that any key decisions as to how datasets should be interpreted are made prior to the registration and commissioning of the datasets. In this manner, all datasets incorporated into the system are considered to be “ready to use”, albeit they can be at quite different overall quality levels.

Considerations that Council addressed when importing studies included consistency, study overlap, and data cleaning requirements.

### Consistency

SCRC flood modelling datasets have been collected over many decades and have come from models with a wide range of detail and frameworks, including:

- Detailed 2D grid (eg 2m and 5m cell size)
- Coarse 2D grids (eg 10m and 20m cell size)
- Flexible mesh frameworks
- 1D flood models

The above diversity of data frameworks presents a challenge to end users unfamiliar with ways different models are developed. As a result, SCRC determined that a framework providing a consistent level of detail on which model results could be viewed, interrogated and analysed, was needed to address the requirements of a wide variety of end users.

The resulting approach was to leverage SCRC’s detailed 1m LiDAR based DEM as a common framework onto which all model results could be draped, forming the “single point of truth” Master Grids.

### Master Grid Creation

Master Grids were created by successively draping the various source model outputs for a given ARI onto the 1m DEM, within the system. During the draping process, experienced flood engineers could manage the creation of a highest quality master grid. Part of this process was managing the merging of overlapping studies, as well cleaning of modelling artefacts.

As SCRC’s studies were developed over a long period of time, and using a range of modelling techniques, there was considerable overlap between studies (for example, between mainstream studies with water levels that back-up tributaries overlapping with tributary-only studies that had an assumed mainstream tailwater level). Study overlap was managed in a number of ways:

- Direct enveloping: retaining the maximum hydraulic parameters where there was any overlap thereby ensuring a conservative outcome. Care was taken in applying this approach to ensure that any modelling artefacts such as artificially high tailwater levels did not influence the enveloping process.

- Truncation of datasets: deletion of parts of models considered hydraulically incorrect such as the tailwater region in a tributary flood study or inflow "ramps".

- Transition/blending across discontinuities: occasionally, there were discontinuities between water surfaces which were addressed using blending and transitioning tools across a designated interface area.

Depending on the type of model used, outputs may also require cleaning. For example, direct rainfall modelling results usually require extensive cleaning to remove artefacts of that modelling approach that result in a water surface existing at every cell.

The end result was a detailed, consistent 1m "master grid" of all hydraulic parameters, that also links to the source data (and metadata). One of the key benefits SCRC wished to obtain from the master grid creation process was an increase in detail of any parameter related to flood depth. This is clearly shown in the following figures when comparing the "as modelled" depth mapping (Figure 2) to the master grid depth map (Figure 3).

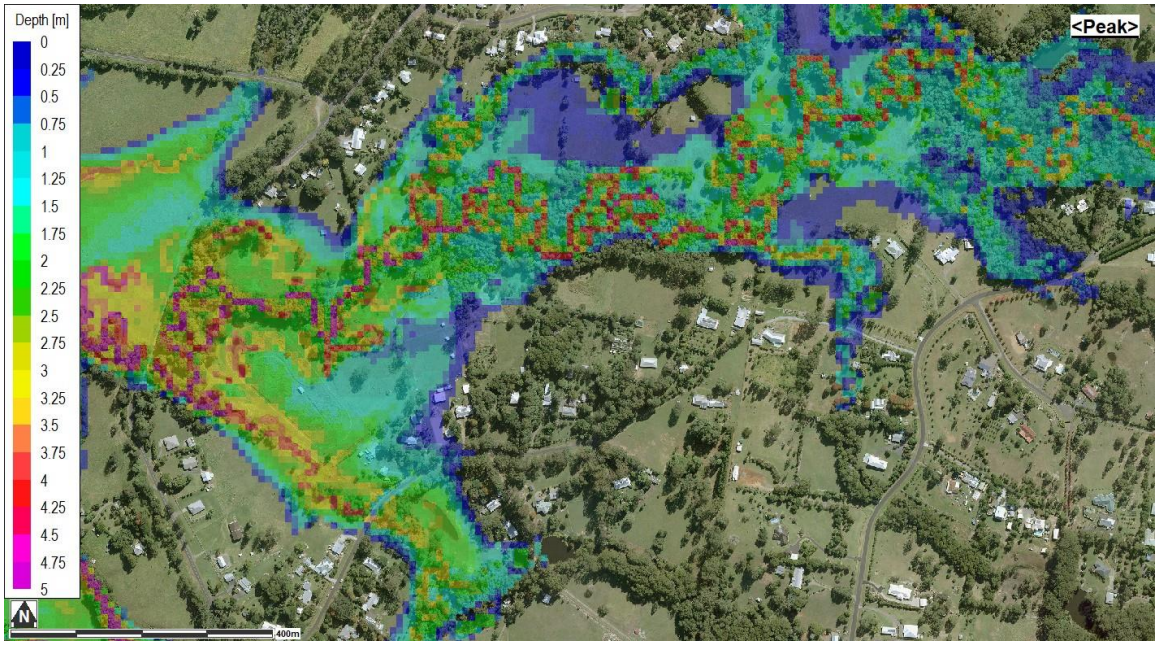


**Figure 2 – Peak Flood Depth - 10m Gridded Model**

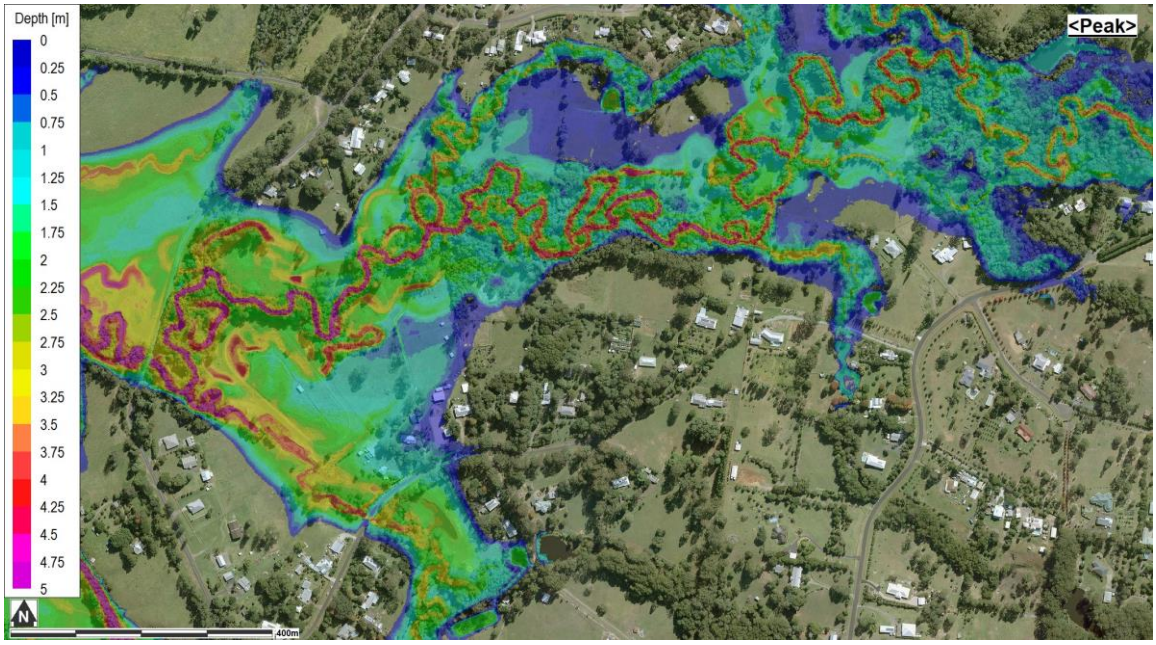


**Figure 3 – Peak Flood Depth - 1m Master Grid**

The Master Grids only contain the *peak* values of the various hydraulic parameters. As peaks are the most commonly used flooding information, this allows users to access

detailed information without the added storage requirements that draping the full time series onto the master grids would entail.

As a peaks-only dataset, each master grid covering the entire SCRC LGA is 130GB, comprising more than 3 billion cells.

**Data Import and Registration**

Prior to creation of the Master Grids, each flood study was imported, quality assured and registered in the system independently. Studies were imported as full time series results (where available) on their native "as modelled" spatial framework, ensuring absolute integrity of the original source information.

The registration process also involves assigning metadata for each study (individual model runs within a study receive the same metadata as the study).

## ONGOING DATA MANAGEMENT

The system has been designed to operate as a dynamic information store, facilitating receipt of new data and updating of the Master Grids, along with archiving and rollback functionality.

As the master grids represent the current flooding data, there are a number of circumstances in which they may require updating.

**Data Updates – Floodplain Development**

As part of SCRC planning policies, there are specific flood-related development controls. The primary requirement is that new development does not adversely impact surrounding properties/land, which is usually addressed with a small-scale flood impact assessment models. SCRC encourages developers to utilise Council's datasets and has traditionally provided "clipped" datasets that were manually extracted.

At its discretion, Council will provide extractions of regional flood models to consultants wishing to undertake analysis in a given part of the regional catchment.

To facilitate this, the system allows SCRC to use a polygon to define an area for which datasets required for modelling are automatically clipped and issued as a "modellers data package". Datasets issued include: terrain, along with available flooding data such as water levels and depths, velocity, hazard, VxD, hydrographs, and time series outputs.

The extraction agreement requires that the DEM and results files are provided to Council. Upon receipt of these files, SCRC is able to import, quality assure, manipulate and register them as a new data source, then assign relevant metadata including status, and ultimately "drop in" to update relevant Master Grids, as required.

**Data Updates – New Flood Studies**

The system utilises a similar process when new flood studies are commissioned. Base datasets required for modelling are extracted within the area of interest as a “modellers data package”.

When the new study outputs are received, they are imported, quality assured, manipulated and then registered in the system and assigned a status (draft, adopted etc). Once the study has been adopted, the Master Grids can be updated with information from the new study, replacing that from any previous study(ies) in that area (assuming the newer study is assigned a better quality than the previous study).

The Master Grid defines which studies are considered “current” within the system, with direct linkage to metadata information available from the study ID assigned to each cell of the grid.

**Archiving and Rollback**

As changes are made to the Master Grids, previous versions are automatically archived. This allows users to review the “time history” of flooding information available at SCRC as well as facilitating the ability to roll back to a previous version/state should Council wish.

In this manner, there is a Quality Controlled time history of the flooding information that supported decisions made by Council. Should the need arise to investigate a decision, for example during a legal dispute, the system can be temporarily rolled back to the same information that was available at the time the decision was made.

**Storage Management**

As data volumes continue to grow, a key role of the system is to assist SCRC in managing its data storage requirements. As such, not all datasets are (nor do they need to be) “readily” available at all times. A layered storage architecture has been implemented to enable Council to optimise storage costs. In decreasing order of cost, the following structure is used to store the various data components of the system:

1. *Maximum Availability*: local network server storage at SCRC providing access to database backend, all metadata and Master Grids.

2. *Medium Availability*: local network storage (NAS or lower cost network storage) providing access to current source datasets and recently archived Master Grids.

3. *Low Availability*: low cost cloud-based storage providing a data store for archived source datasets and older archives of Master Grids.

The above layered structure ensures that regularly used datasets can be readily and quickly accessed, whilst less frequently used (or required) datasets are stored on inexpensive cloud-based servers.

**System Performance**

When working with such large datasets, it is important to consider data access and usage speeds. The system streams all datasets across the network utilising data pyramids, ensuring excellent performance when accessing and interrogating datasets.

All processing and manipulation effort associated with new dataset receipt is carried out on a local machine to ensure maximum performance, prior to being registered into the system.

## ACCESSING DATA BY END USERS

Whilst the core of the system is in managing the flooding datasets themselves, it integrates seamlessly with Councils existing standard-release waterRIDE™ applications. In this manner, SCRC has been able to manage its flooding data, without necessarily changing the way it is accessed.

A key aim of the system was to ensure that the rich array of flooding information could be reliably, consistently and quickly accessed by various types of end users within Council, without the need for all users to understand how the data is stored and organised.

As such, access to flooding datasets is facilitated using waterRIDE™ projects, including a single project for access to Master Grids and individual projects providing access to each flood study. Access to these projects is managed through permissions, providing a mechanism for controlling access to different types of end users.

The system also allows flood information searches to be carried out to identify available information and extract data histories at any location, including comparison between various Master Grids over time. Metadata associated with the source studies comprising any Master Grid can also be accessed by end users at any time.

## CONCLUDING REMARKS

The growing volume of a Councils flooding datasets presents a “big flood data” problem. The management of these large data sets in storage constrained environments requires effective process and systems to best leverage the value of these datasets. In addition, such processes and systems should enable the administrators of such datasets to prioritise their time to focus on managing and improving the information, rather than processing requests for other users.

The system, now implemented at SCRC, demonstrates that there are solutions to the looming “big flooding data problem” which promises to not only resolve the data management issues but to allow the opportunity for the data to be better leveraged through improved accessibility.