General requirements for data

Input data is specific to the component models that you use, but typically consists of climate, topography, land use, rainfall, and management practices. Examples of these are provided in Tables 1 and 2.

Table 1. Model calibration and validation (required data sets)

Base layers and data	Common data formats	Description and use
Observed flow data preferably at the same time-step as model	Time series	Typically, you can use daily time-step gauging station data (ML/day or m3/s) in the hydrological calibration /validation process, or as a substitute for modelled runoff as an observed flow time series at relevant nodes. You need to assess the length of time for which records are available, data quality and data gaps to determine how useful the data sets are for calibration/validation. It is preferable to have data sets that are at least 10 years in duration and that cover both wet and dry periods. When extracting gauged flow time series data from databases such as HYDSTRA (commercial software used across Australia for archiving gauge data), it is important to align the time bases for flow with other climate data sets. For example, SILO data is collected each day at 09:00am, so each data point is the total rainfall for the previous day. Therefore, when extracting flow data from HYDSTRA, ensure that the "end of day" option is selected, so that the flow data will align with the SILO rainfall data. It is important to understand the conventions used by your organisation as well as those that send you data - they may not be the same!
Observed water quality data	Time series	Data used in the water quality calibration/validation process. You need to assess the length of time for which records are available, data quality and data gaps to determine how useful the data sets are for calibration /validation. It is preferable to have data sets that cover both storm event and ambient conditions that include both wet and dry periods.
Existing reports	Report or spreadsheet	Existing reports for the region may assist in hydrology and water quality calibration process, eg. load estimation.

Table 2. Optional data sets

Optional data sets	Common data formats	Description and use
Visualisation layers (eg. roads, towns, soils, streams)	Polygons, polylines or points	You can add (drag and drop) additional layers into the Layer Manager and turn these layers on or off as required. These layers are not used in model development and are more a visualisation tool. The layers need to have the same projection and resolution as the DEM or sub-catchment map (but can have different extents).
Aerial and satellite imagery	Image	You can use aerial or satellite imagery to ensure the node-link network generated (either drawn manually or generated by Source) is correct. You can also use imagery to check the accuracy of the DEM-generated sub-catchment map. The layers need to have the same projection and resolution as the DEM or sub-catchment map (but can have different extents).
Local or relevant data on best management practices	Reports or spreadsheets	This includes information on locally-relevant best management practices that could be used when creating scenarios.

Existing hydrology and water quality reports and data	Relevant format	Existing reports for the region may help when you parameterise water quality models (eg. EMC/DWC derivation. For example, an existing IQQM model of a region that uses the Sacramento rainfall runoff model can be used to parameterise a Sacramento model in Source. Climate data sets may be used to speed up calibration of rainfall runoff models.

Note that all spatial data must use the same supported projections:

- Albers Equal Area Comical;
- Lambert Conic Conformal; or
- Universal Transverse Mercator (UTM);

The exception is SILO gridded climate data, which is formatted in a geographic coordinate system.

Table 3 summarises the minimum necessary and optional input data needed to create a catchment model using Source.

Table 3. Building models (required data sets)

Base layers and data	Common data formats	Description and use
Digital Elevation Model (DEM)	Grid	A pit-filled DEM is used to compute the sub-catchment boundaries and node-link networks. Source can automatically generate sub-catchment boundaries according to a user-specified minimum drainage area (stream threshold) and flow gauging station positions. Selecting a small minimum sub-catchment area value will generate a large number of sub-catchments. This will increase the size of the project and run-time.
Sub-catchment map	Grid	A sub-catchment map can be used in place of a DEM. This defines the sub-catchment boundaries within Source. You then need to draw the node-link network for the catchment.
Functional Unit (FU) map	Grid	A functional unit (FU) map divides the sub-catchment into areas of similar hydrological behaviour or response (eg. land use). Source uses FU maps to assign functional unit areas. The FU map needs to have the same projection and resolution as the DEM or sub-catchment map (but it can have different extents, provided the FU map at least covers the extent of the sub-catchment defined by the modeller.
Gauging station nodes	Point	Optionally, a shape file or ASCII text file that lists the gauging station coordinates, and an identifier such as gauge name or number that is used to define gauging station nodes. The coordinates of the gauges need to be in the same proejction as the DEM or sub-catchment map. Incorporating the location of gauging stations as nodes can be particularly useful when calibrating the rainfall runoff model at a gauge.
Rainfall and PET data	Grid or time series	Rainfall and potential evapotranspiration (PET) time series are used as inputs to the rainfall-runoff models. The most commonly-used files are SILO daily rainfall-runoff and PET ASCII grids. Using a daily ASCII grid format allows you to update the rainfall data at a later stage and re-run the model. If local data is available, you can also attach your own rainfall data files to rainfall runoff models for each FU within a sub-catchment.
Point source data (if storages are to be modelled)	Time series	Outflow and/or constituent data. The time series needs to have the same time-step, an should be run for at least the same duration, as the climate or flow inputs to the model.
Storage details (if storages are to be modelled)	Time series	Includes coordinates, maximum storage volume, depth/surface area volume relationship, observed inflow and outflow data, losses, extractions, release rules, dam specifications, gates, valves, etc.
Stream network layer (optional)	Polyline	A stream network layer ensures the node-link network you build over the sub-catchment map is correct. You can then check the accuracy of the DEM generated sub-catchment map.

USLE (Universal Soil Loss Equation) and/or gully density layers (optional)	Grid	Can be used to spatially vary EMC/DWC values in the constituent generation process (use the Scale EMCs and DWCs with the Hazard Map constituent generation method available through the Spatial data pre-processor plugin). The layers need to have the same projection and resolution as the DEM or sub-catchmetn map (but can have different extents).

Data formats

For gridded spatial data files, formats should be in ESRI text format (.ASC) or ESRI binary interchange (.FLT). Vector data should be in shape files. Gridded rainfall data can be ordered from either:

- http://www.longpaddock.qld.gov.au/silo/; or
- http://www.bom.gov.au/jsp/awap/rain/index.jsp

It is recommended that overlaying Digital Elevation Models (DEM), functional units or sub-catchments have the same projection and resolution (but they can have different extents).

Zero-padded data

Certain file formats require data to be zero-padded. In Table 4, the first column represents months, and is not zero-padded. Some applications will sort this data as is shown in the second column. The third column is zero-padded and sorts correctly.

Table 4. Zero-padded data (sorting example)

Non-zero padded data	Default sorting order	Zero-padding (always sorts correctly)
1	1	001
2	10	002
10	100	010
20	2	020
100	20	100
120	200	120

Time and dates in data files

The TIME framework (used by Source) uses a subset of the ISO-8601 standard. The central part of this subset is the use of the format string:

yyyy-MM-ddTHH:mm:ss

Dates should comply with the ISO 8601 standard where possible but more compact formats will be read if unambiguous. For example:

- the dates 24/01/2000 (Australian) and 01/24/2000 (USA) are unambiguous; but
- the date 2/01/2000 is ambiguous and depends on the local culture settings of the host machine.

The TIME framework will always write dates in the following format and it is recommended that you follow the same format and use zero padding within dates. For example, "2000-01-02" is preferred over "2000-1-1" to avoid ambiguity:

yyyy-MM-dd

Annual data can often be entered by omitting a day number and using month number "01" (eg 01/1995, 01/1996, 01/1997).

Where a date-time specifier only contains a date, the reading is assumed to have occurred at time 00:00:00.0 on that date.

The smallest time-step that Source can currently handle is one second. When reading a data file, Source examines the first few lines to detect the date-time format and time-step of the time series:

- If the format is ISO 8601 compliant, this format will be used to read all subsequent dates;
- Failing that, an attempt is made to detect the dates and time step with English-Australia ("en-AU") settings, for backward-compatibility reasons; and
- Last, the computer configuration is used for regional and language settings.

Possible problems with time-steps

Incorrectly-formatted date and/or time entries will result in errors if Source is unable to interpret your data file (eg. *LoadDataFileIOException*). You may also need to check your data if you use an ambiguous date format.

There are two known problems where a time step may be incorrectly detected:

- When reading a file on a computer with US settings, because of the mm/dd/yyyy date format. This may happen if the whole of a daily time series covers less than 13 days, or less than 12 months for a monthly time series; or
- When reading a file which has years in two digit format (eg. 30/01/99) instead of four digit format (eg. 30/01/1999). An error will occur if your time series has the two-digit years 29 and 30. In this case 29 is read in as 2029, and 30 is read in as 1930. Note that data exported from Excel in *.csv format will be saved with the displayed date.

Both problems can be avoided by using the recommended ISO 8601 format to prevent ambiguity.

Predicted or calculated data

The predictions produced by an integrated model developed with Source depend on the selected component models. Example outputs include flow and constituent loads as a time series.

Missing entries

Missing entries are usually specified as -9999. Empty strings or white space are usually also read as missing values. Occasionally, other sentinel values are used, such as "-1?" in IQQM files.

Decimal points

Always use a period (".", ASCII 0x2E) as a decimal separator for numerical values, irrespective of the local culture/language/locale settings for Windows.