Batch Processing in PAMGuard

Douglas Gillespie[1]

[1]Sea Mammal Research Unit, University of St Andrews

2025-02-20

Tutorial Version 2.0

Learning Outcomes

In this tutorial you wil learn to:

1. Learn what we mean by batch processing
2. Install the Batch Processor in PAMGurd
3. Set up a PAMGuard configuration to run batch processes
4. Run a process of a multiple sets of raw data files
5. Run additional offline tasks on the generated PAMGuard output

Table of contents

List of Figures

1 Introduction

The PAMGuard batch processing module allows you to run the same PAMGuard configuration on multiple sets of data (Figure 1). This is particularly useful if you want to process data from deployments of multiple autonomous recorders in exactly the same way. It can also be used to reprocess data from multiple old cruises with a new detector, for instance with one of the new Deep Learning detectors / classifiers that have recently been added to PAMGuard.

> PAMGuard configuration files
>
> PAMGuard configurations are held in psfx files. They contain all of the information about the modules you've selected and how you've configured those modules.
> They can be easily copied or emailed between users.

PAMGuard users will know that setting up the same configurations on multiple datasets is tiresome. For each dataset, you need to copy the configuration (psfx) file, then change the input folder for your sound files, the output folder for the binary data and the output database. If you get this wrong, then you might overwrite some data. Then, when you decide that those weren't exactly the detector settings you wanted, you have to do it all again for all your data sets.

The PAMGuard batch processing module addresses this problem by running the same configuration on as many data sets as you want. You set up a series of "jobs", and the batch processor will work through them, using the same configuration, until they are complete. Generally, it will run multiple jobs concurrently on a single machine.

As well as processing raw audio data, it can also run 'offline tasks' on already processed data. For instance if you wanted to change the click classifier settings, and reprocess all of the click binary files in multiple datasets, the batch processor can help do this in an efficient manner.

> My personal data processing record
>
> My personal record is reprocessing over 80 Terrabytes of multibeam active sonar data, spread over 14 external USB hard drives, in less than 8 days. Do achieve this, I set up 6 jobs processing data on the first 6 hard drives, with the desktop PC set to run three jobs at a time. When the first three jobs were complete I replaced those hard drives and set up the next three jobs, and so on until processing was complete. In all, it only required half an hour or so of my time every 2 or 3 days.
> Data from this analysis are in the recent paper Montabaranom et al. Seals exhibit localised avoidance of operational tidal turbines.

1.1 Soundtraps

SoundTraps are slightly more sophisticated than many recorders. Depending on how they are configured, they may not just be recording raw audio data, but also the output of an onboard detector. The ones we're using were configured to sample raw data at a sample rate of 576kHz. The soundtrap then ran an automatic transient (click) detector on the incoming data and each detection was saved as a short clip just over a millisecond long. The incoming high frequency data were then decimated to 96kHz and all of the 96kHz data stored. Both the detections and the recordings are compressed using a lossless compression algorithm and stored together in a SUD file as shown in Figure 2.

The soundtrap system does not attempt to run automatic detectors for lower frequency sounds. This would be too complicated for the low power processor that soundtraps use. The combination of high frequency click detection and lower frequency recording allows us to detect the very high frequency clicks of species such as the harbour porpoise, and at the same time, make recordings which can be

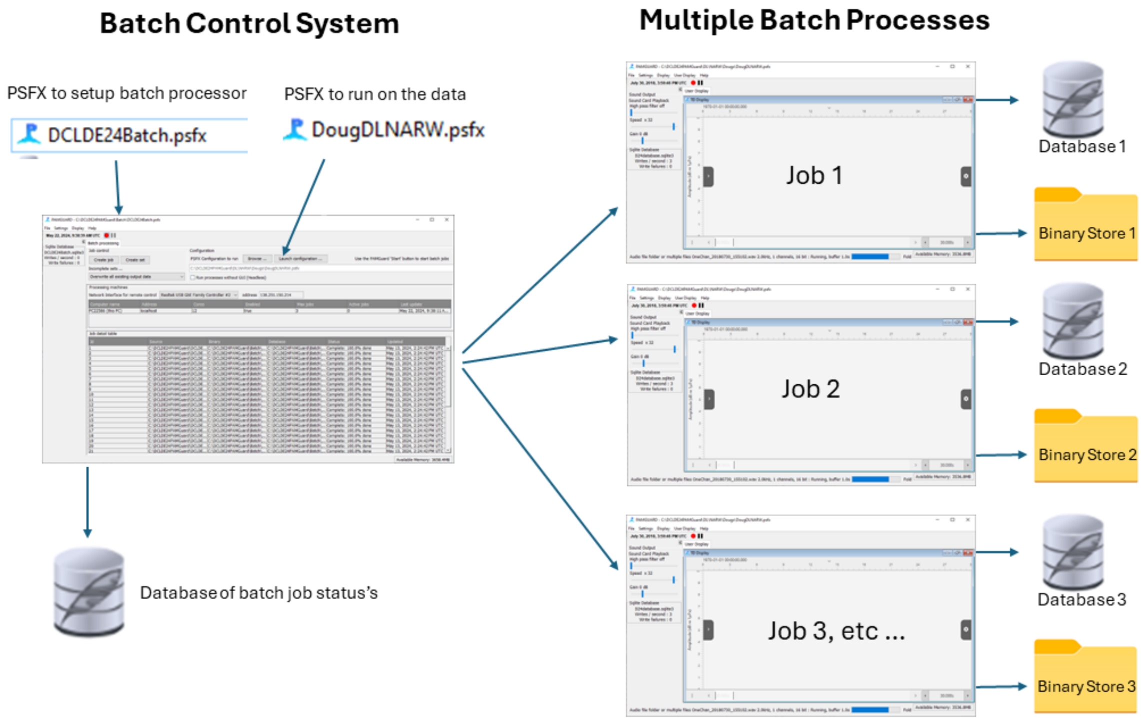


Figure 1: Diagram of the batch processor configuration

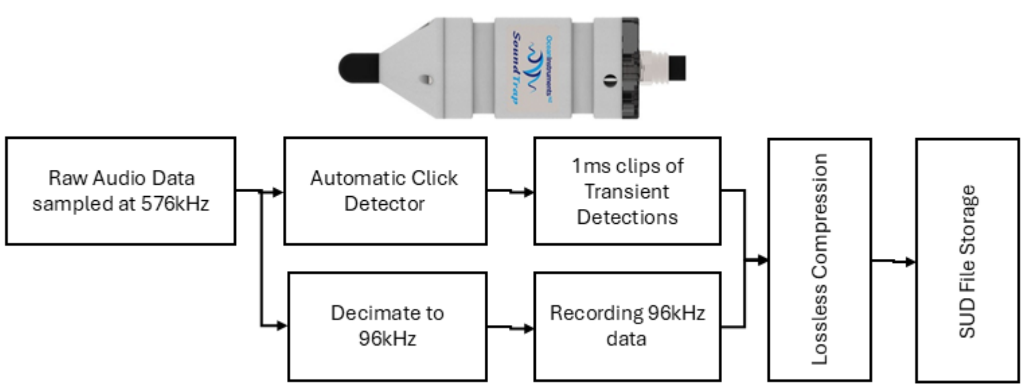


Figure 2: Schematic diagram of the data flow through a soundtrap

processed offline for a wide variety of other species. This system allows for much longer deployments than would be possible if all of the high frequency data were recorded. For more information about soundtraps, visit the Ocean Instruments web pages.

This tutorial follows on directly from another PAMGuard Tutorial Introduction to Static Monitoring which you might want to complete before proceeding with this tutorial, which will focus on batch processing with a pre-built configuration.

2 Installation

2.1 Software

This tutorial will work with PAMGuard Version 2.02.16 or later and the Batch Processor Plugin version 2.0 or later, both of which are available on the PAMGuard website. Once you've installed PAMGuard, copy the downloaded Batch module (BatchProcessing_2_0.jar) into the plugins folder, which you'll find in your PAMGuard installation folder (probably C:\Program Files\Pamguard\plugins). If you've any older versions of the Batch Processor plugin in that folder, make sure you remove them. Once this is done, the Batch Processor module will appear in the Add Modules / Utilities menu just like any other PAMGuard module, and it's help pages will be available in the PAMGuard online help.

2.2 Sample Data

The data you'll be using for this tutorial comes from a deployment of five SoundTrap 300 recorders off the West coast of Scotland, which form part of the Compass Project. We've taken a single days data for each recorder since the full dataset would be too large for a tutorial exercise and might take many days to process.

The data are available on Zenodo at (Link to the Zenodo site). The Zenodo dataset contains three files: SoundData.zip, configuration.zip and pamguardoutput.zip. For this tutorial you only need SoundData.zip and configuration.zip. You'll be recreating the files in pamguardoutput.zip, so you don't need them, but take a look if you want to.

Unzip SoundData.zip and configuration.zip to a convenient location on your hard drive.

3 Configure PAMGuard

Two PAMGuard configurations are required to run the batch processor. A configuration that contains the PAMGuard modules that you want to run on your data, and a second configuration that is going to control the Batch Processor. We'll call these the **Run configuration** and the **Batch Configuration** respectively.

3.1 The Run configuration

The run configuration you'll be using is the configuration you should have ended up with at the end of the Introduction to Static Monitoring tutorial. If you completed the static monitoring tutorial, you can use the configuration you had at the end, or you can use compass_settings_static_logger.psfx which you should have downloaded in configuration.zip for this tutorial.

The configuration (Figure 3) uses an FFT module to compute a spectrogram of the 96kHz data which feeds a Whistle detector and a long team spectral average (LTSA) generator. The 96kHz data also input to a Noise band Monitor and are also decimated to 10kHz fo feed a second copy of the Whislte and Moan detector to search for lower frequency tonal sounds. You'll also see the "ST Click Detector" which is not connected to the data from the sound acquisition. This is because it's

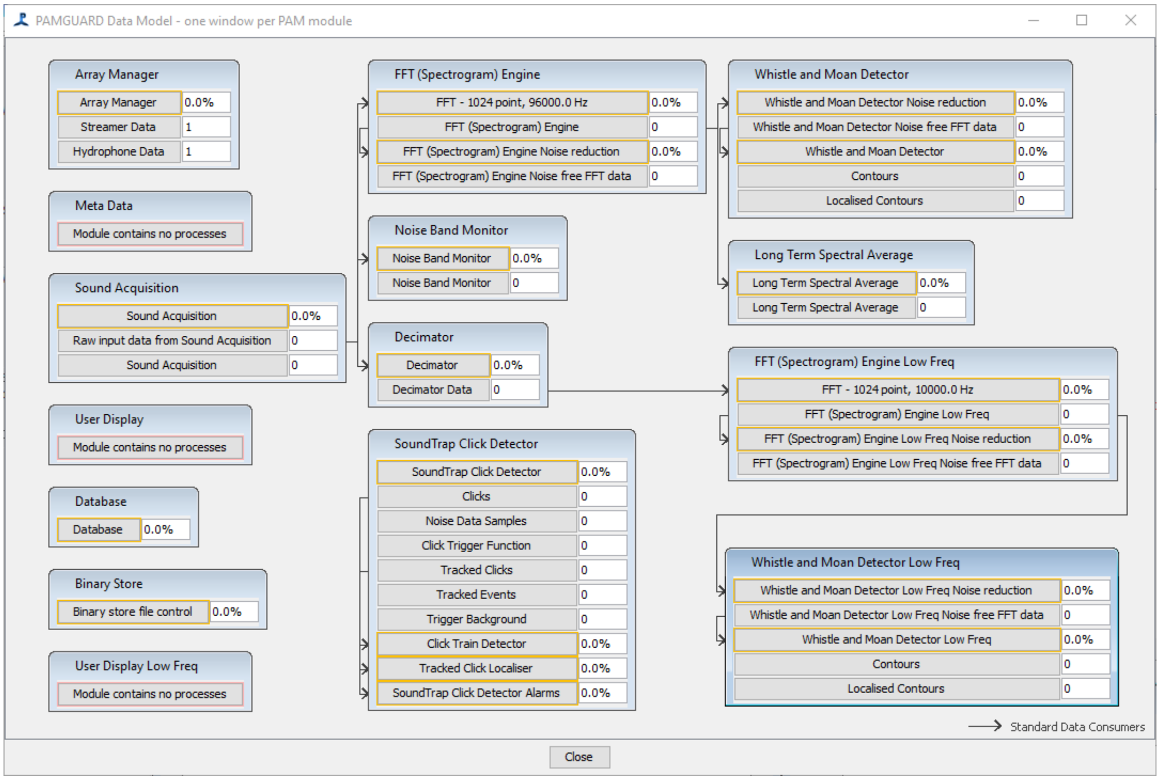


Figure 3: PAMGuard Data Model Diagram for SoundTrapClicksNWsls.psfx

a special version of the click detector, modified to receive the detection clips from the SUD files, which you'll remember are at the higher sample rate of 576kHz.

As well as the above sound processing and detection modules, there is the Array Manager and Meta Data modules (both always present in every PAMGuard configuration), a Binary Store and Database for the output data we're about to generate, and a User Display which will show a spectrogram of the 96kHz data overlaid with whistle detections.

You've probably opened the compass_settings_static_logger.psfx file yourself by now. Close it again before you start to create the Batch Configuration.

3.2 The Batch Configuration

You'll be making this configuration from scratch so that you learn to use the batch processor.

Start PAMGuard and create a new configuration: launch PAMGuard from the Windows Start menu, and when it asks you to "load PAMGuard configuration from ..." press **Browse / Create New...** . In the dialog that opens, navigate to where you want to work, and enter the file name CompassBatch (Figure 4), then hit the Select button or press Enter.

From PAMGuards *File / add modules / utilities* menu add a Database and a Batch processing module to the configuration. Then from the *File / Database / Database Selection* menu dialog, press **Browse / Create ...**, and enter the name of the database (e.g. CompassBatch) which is going to hold information about the jobs you're running and how they are progressing. Your PAMGuard configuration should now look like Figure 5 .

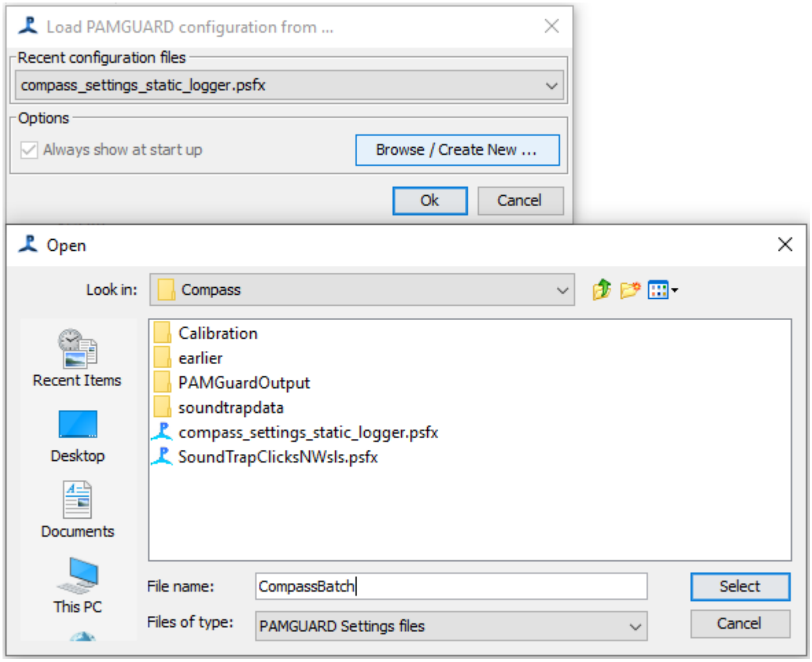


Figure 4: PAMGuard dialog for creating a new configuration file

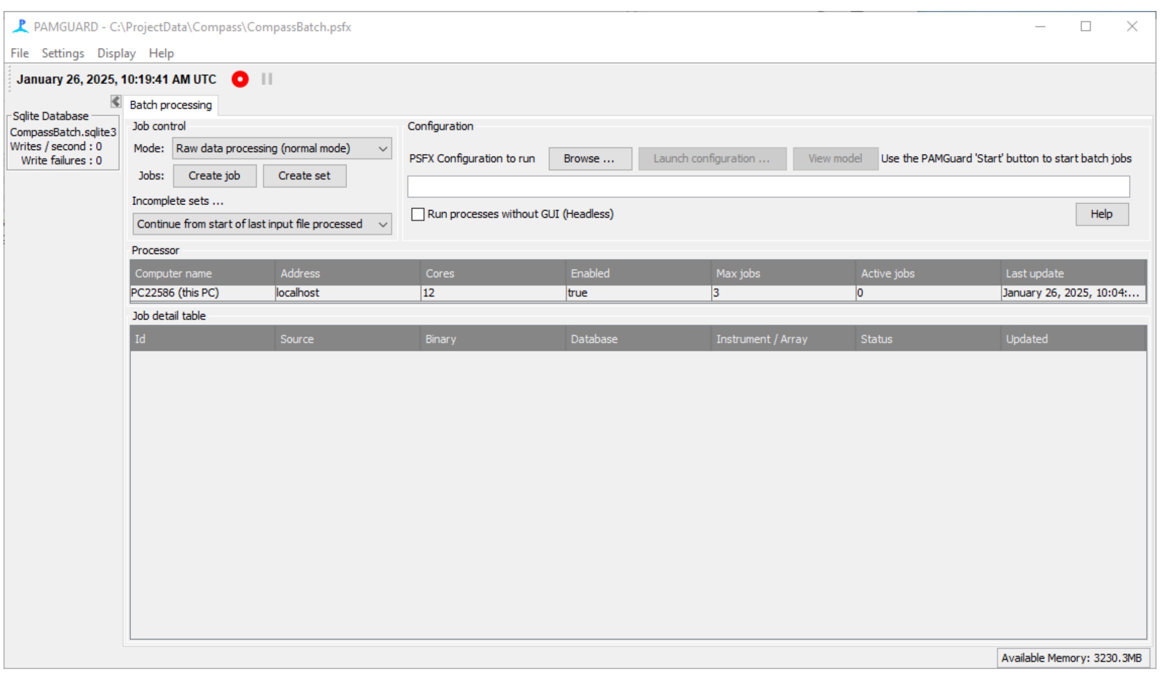


Figure 5: PAMGuard configuration with database and batch processor modules

3.2.1 Select the configuration file

At the top of the display, in the configuration panel, press the **Browse ...** button and select the Run Configuration compass_settings_static_logger.psfx.

3.3 Create jobs

You should have five sets of sud files, each in a different folder. You can create all five jobs at once from the **Create Set** button in the Job Control Panel. Press the Create Set button, then in the dialog, press the Select button in the top right corner. Navigate to the folder containing the five folders of soundtrap data and select just that root folder (in the example in Figure 6 it's C:\ProjectData\Compass\soundtrapdata).

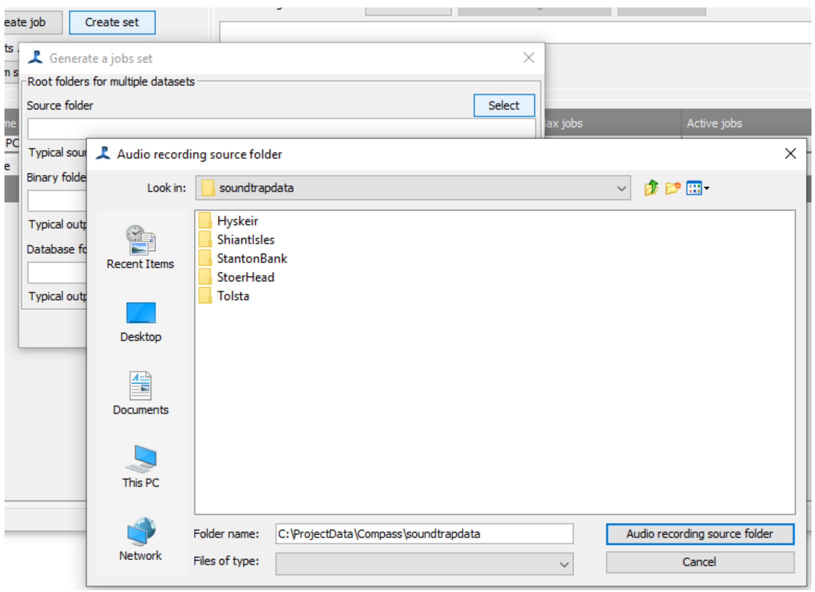


Figure 6: Selecting the source folder for multiple batch jobs

The Batch processor will only look one level of folders down from the folder you select, so make sure you've selected the right folder. If you've got it right, you'll be told that its found five sub folders containing data.

Next select folders for the binary output and for the databases that will be automatically generated. I've chosen the folder C:\ProjectData\Compass\PAMGuardOutput for both.

Press OK, and the set of five batch jobs should show in the main PAMGuard batch display (Figure 7). If you look at your folders and files with Windows Explorer, you'll not see any of the output binary folders and databases at this stage. Don't worry, they will be created automatically as each batch job starts.

3.4 Set individual job calibration and location data.

By default, the batch processor will take hydrophone data from the Run Configuration and use it with each set of data processes. If you're making noise measurements, you will need to set the correct hydrophone calibration values for each job, or the noise measurements will not be accurate. If you're processing data from static hydrophones, you may also want to set the correct location of each deployment.

Callibration and location values for each dataset are provided in the files CompassMetaData.xlsx and CompassMetaData.csv (both are the same data).

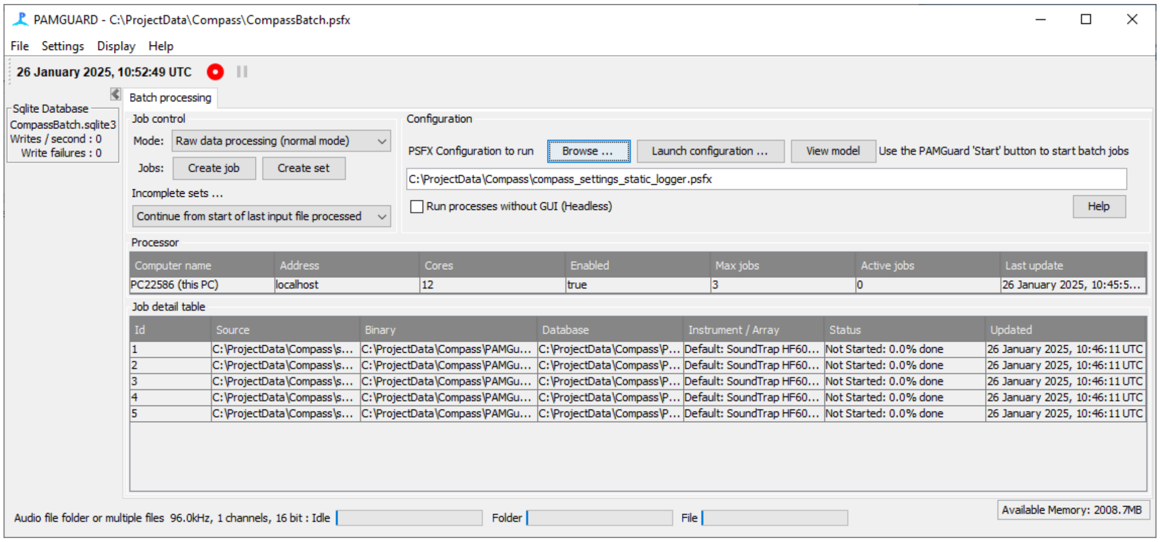


Figure 7: PAMGuard configuration with a set of five batch jobs

For each job in the table, right click on the row and select *Add or Edit calibration / Array Data* from the dropdown menu. This will open the Array Manager dialog (Figure 8a), using the current default values from the Run Configuration. Enter the SoundTrap serial number in the Instrument Id field. Next double click on the hydrophone element (there is only one for this SoundTrap). Enter the hydrophone sensitivity as the negative of the High Gain full scale value from the spreadsheet (Figure 8b). Then open the 'streamer', again by double clicking or by pressing the *Edit* button. Right click on the menu button to the right of the display position and select *Edit* from the dropdown menu. When the Hydrophone array reference position dialog opens, click on "Decimal" in the top right hand corner and copy/paste the appropriate values from the spreadsheet (Figure 8c).

Work your way through all five jobs and set the data for all of them.

! Feedback

We agree that this can be a bit of a pain for a deployment of many recorders and are trying to think of an easier way of doing this, such as importing data from a database or spreadsheet.

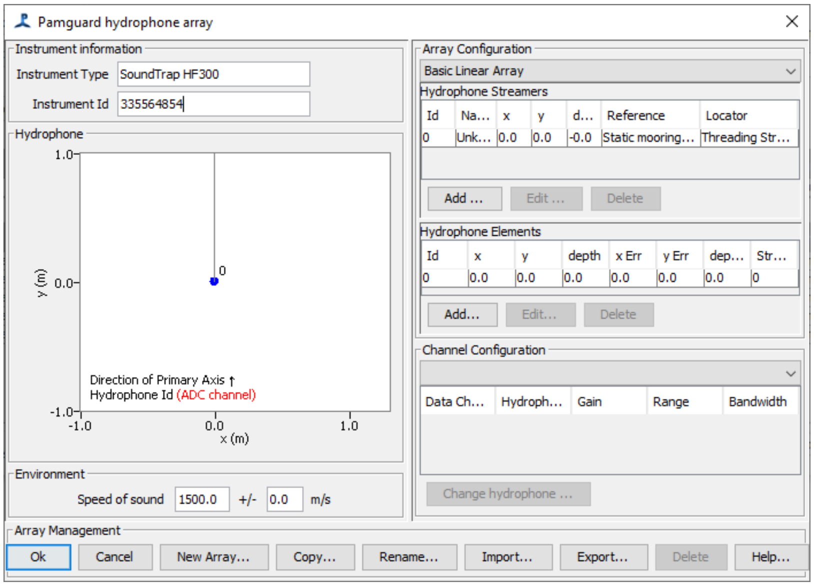
4 Run the batch jobs

4.1 Number of concurrent jobs

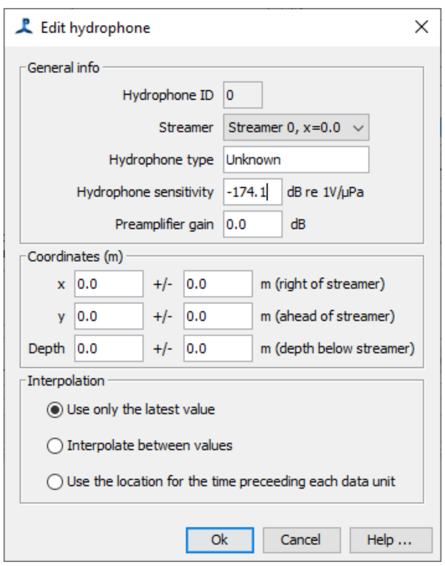
The final thing to decide before starting processing is how many concurrent jobs your computer can manage efficiently. How many jobs can run concurrently depends on the complexity of the configuration and on how powerful the computer is. On a good Intel I7 desktop with 32 GBytes of RAM and 20 processor cores, I'd probably run three jobs at a time. On my laptop, which only has 12 cores and 16 GByte RAM, I'll stick to 2. To set the number of jobs right click on the Processor table (which will have a single line in it for your current machine) and either increase or decrease the maximum number of jobs.

4.2 Run the jobs

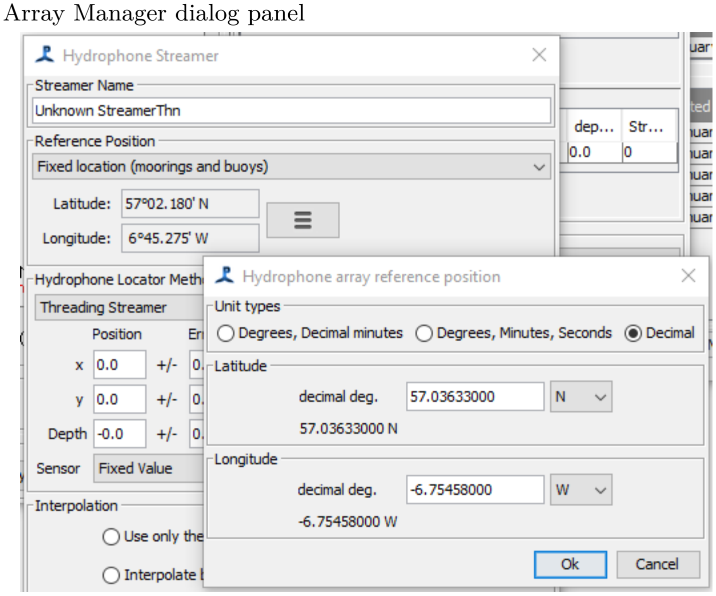
You're now ready to start the batch jobs. Press the PAMGuard start button and wait a few seconds for the first jobs to start. The jobs table will show the status of each job (Figure 9), which will update as each processed file completes. When a job ends, another will start until all jobs are complete.



(a) Array Manager dialog panel



(b) Hydrophone Sensitivity



(c) Location Data

Figure 8: Entering the hydrophone calibration and location data

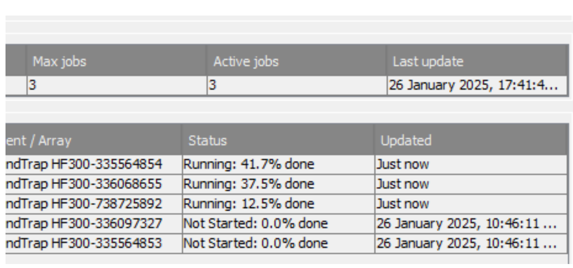


Figure 9: Batch Job Status

Look in the folder where you've told it to put the binary files and databases. You should see database files and folders of binary output start to appear as each job starts.

You can change the number of concurrent jobs while they are running. If you reduce the number, then no jobs will be stopped, but when one ends, another will not start until the number of active jobs drops below the set maximum. If you increase the number of jobs, it's likely that another job will start immediately.

Once jobs are complete, there will be a menu option available to open that dataset with the PAMGuard Viewer. For further information, see the online help.

> Impatient ?
>
> Unless you have a powerful computer, running these sample jobs may take an hour or more. At this point you may want to stop the jobs and take a look at the preprocessed data we provided for you.

5 Running Offline Tasks

Several PAMGuard modules have options in Viewer mode to re-run certain tasks, such as click classification.

Indeed, in the exercises above, we deliberately 'forgot' to set up the SoundTrap click detector to classify porpoise clicks, so we should run that task now.

To run offline tasks go back to your batch configuration then in the top of the Job Control Panel, select Mode: Offline tasks (viewer mode) from the two options. (You could also start an entirely new batch configuration, and set up the jobs again, but that's not necessary for us here). The display will change to also show a table of possible offline tasks to run on the data (Figure 10). Note that you may have to resize the display and move the dividers between the different tables to see this properly. Also, if you're now using the pre-processed data, the paths to the data in your batch configuration might all be wrong, so check them!

5.1 Configure a click reclassification task

The batch processor has automatically extracted a list of available tasks from the modules in your run configuration. In this case, there are five tasks, all from the click detector. Recalculating delays and bearings is not relevant to single channel data, but we can run the click classification. Click on the **Configure Task** button for Reclassify Clicks, then in the dialog (Figure 11a), then press the **New** button to open the Click Parameters dialog (Figure 11b). In the bottom right from **Set Defaults** select *Porpoise* which will set default values for porpoise click classification.

5.2 Run the task on all five datasets

If you're using the batch processor configuration that you used to process the raw data, then it will still be saying that all jobs are complete. To reset them ready for running the offline tasks, right click anywhere in the jobs table and select *Reprocess all jobs* from the dropdown menu.

Make sure that you've selected the appropriate checkbox to run the Reclassification offline task.

Then, just as before, press the red start button. Each dataset will open in turn and the selected offline task will run on each dataset, following the same rules as earlier as to how many jobs will run concurrently.

Running the click classification is a lot quicker than processing the raw data, so this should only take a couple of minutes to run on most computers. Once jobs are complete, you can view the data for each job using the PAMGuard viewer.

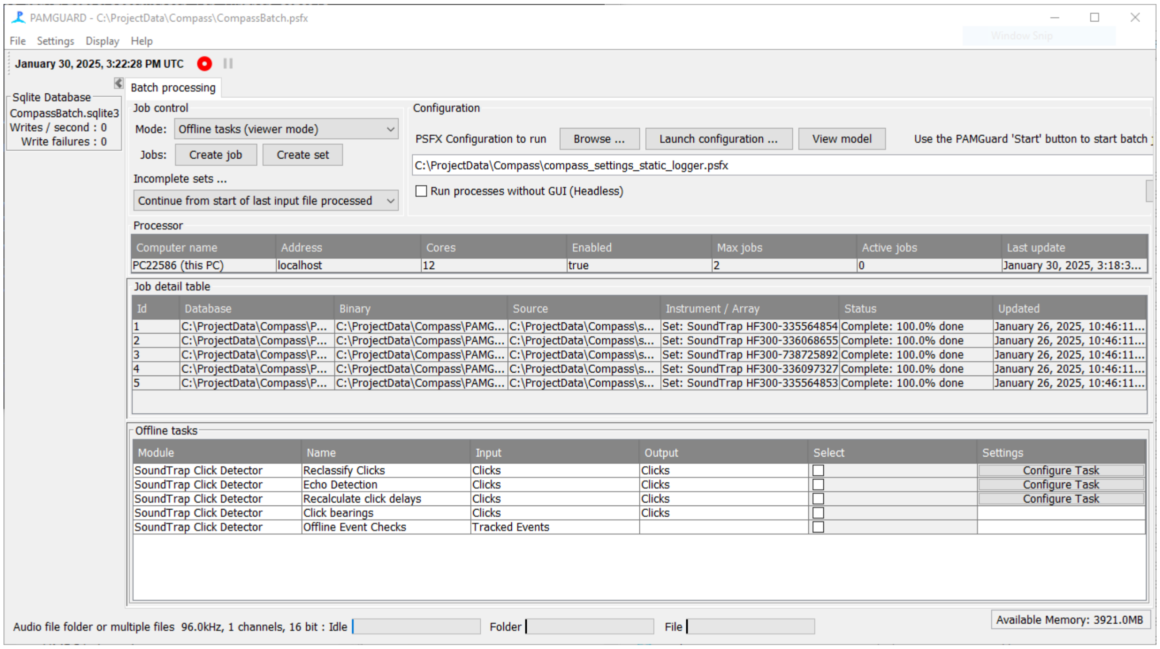
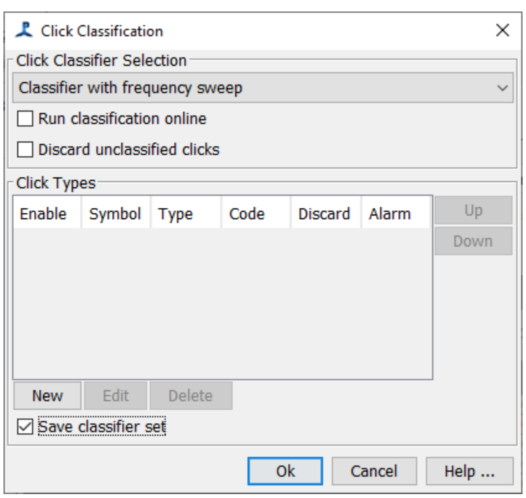
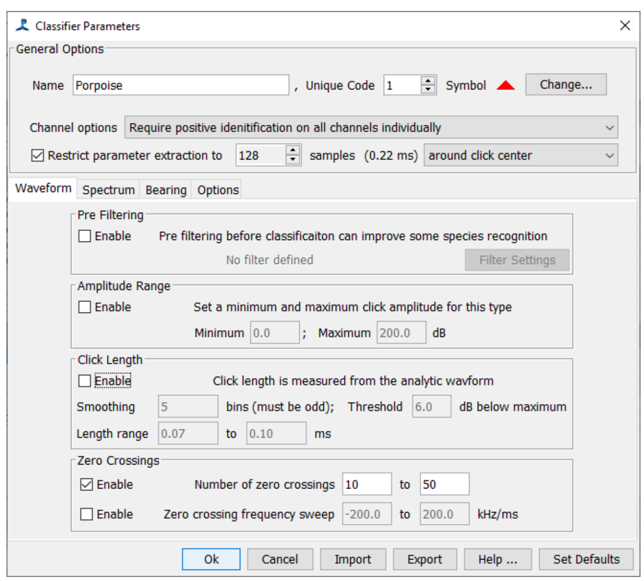


Figure 10: Configuration display for controlling offline tasks



(a) Click Reclassification configuration



(b) Setup for a Porpoise click classifier

Figure 11: Configuring click classifiers prior to running reclassification