**JASMIN directory workflow process**

This document explains the proposed storage and workflow of data files on the JASMIN C2S311a Lot 2 land workspace. It is expected that the marine data directories may have similar directory structure, but the processes differ, so this document just focuses on the land data workspace.

**Incoming data**

Data can be uploaded to group workspace via the arrival’s server, direct upload by an approved user for large volumes of data or upload by a team member from a local drive. The incoming directory is a temporary space for data uploaded via the arrival’s server, once the data is checked and verified it will be moved to Level0. All other data will be uploaded directly into Level0. Once the service is operational, a system for timely updates will need to be developed. All original data in source format will be stored in Level0 directory which will be inventoried and backed up to tape.

**Code**

There is a directory on the JASMIN workspace where all processing code should be saved [link]. The code should also be made available through the GitHub repository [https://github.com/glamod].

**Level0 parent directory**

The Level0 parent directory has an inventory sub-directory where the current Level0 data inventories are stored (see Figure 1). The Land/ sub-directory contains three other sub-directories one for each timestep (sub\_daily\_data, daily\_data and monthly\_data).

The data in Level0 in each timestep sub-directory should be named with a unique source ID (if already inventoried) and source name e.g. 1000084\_dwd.zip. If not inventoried then just name the data set by source name e.g. dwd.zip. The data which has been uploaded via the arrival’s server via the incoming data directory or directly uploaded by user or team member will be stored to Level0. All data in Level0 should be inventoried or in the process of being inventoried. The data in Level0/Land/ for each timescale should be assessed and based on a list of criteria (TBD) prioritised for merging.

If the data source has been prioritised for merging, the data files should be converted into the standard intermediate file format (IFF) single variable files. All the processing, e.g unzipping data files and conversion to IFF etc, can be conducted in the Level0/Land/processing sub-directories for each timescale.

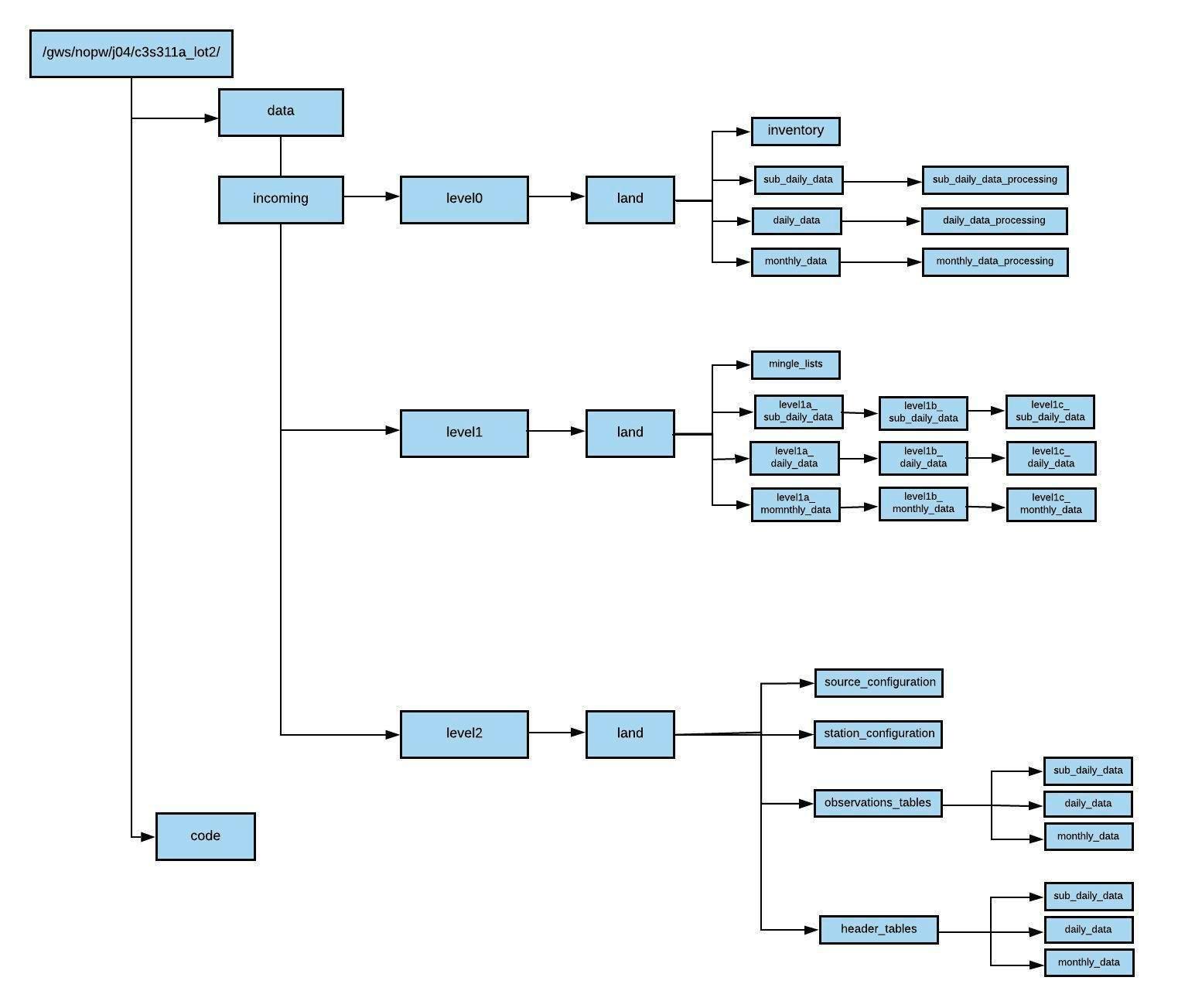


Figure 1 Schematic of proposed JASMIN group workspace for land data

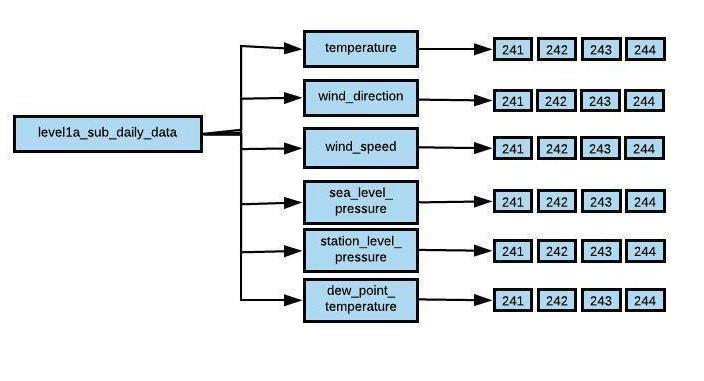


Figure 2 Schematic example of the Level1a sub-daily data sub-directory and the separate sub-directories by each variable.

You can store temporary data files in the Level0/Land/processing sub-directories. However, this is not a permanent long-term storage area and all temporary files should be deleted after a period of time, but all processing code should be saved in the code directory for replication in the future.

**Level1 parent directory**

Once the prioritised data source files have been converted to the IFF they are moved from the processing directories to the Level1/Land/Level1a sub-directory. The single variable station IFF files should be saved in the relevant timestep directory and in sub-directories for each variable (See Figure 2). Within these sub-directories the IFF station files should be separately saved in sub-directories by source (see figure 2). Name the IFF using the station ID - record number (for multiple station configurations) combined with the variable name (exactly as defined below) combined with source ID using \_ separators e.g IE9999-1\_sea\_level\_pressure\_245.psv

temperature

dew\_point\_temperature

wind\_direction

wind\_speed

sea\_level\_pressure

station\_level\_pressure

The mingle list is produced on the data in each of the Level1a sub-directories. The mingle list for each timestep should be stored in the mingle list sub-directory. Once a mingle list has been produced, the IFF station files are merged into multi variable Merged File Format (MFF). The MFF are saved to the relevant timestep Level1b sub-directories. Save as a pipe [|] separated file ext = [.psv] Multiple variables per file. Name the MFF using the merged station ID.

The data, quality control (QC) checks will be conducted on the Level1b MFF. Once QC has been conducted the data are saved as Quality Controlled File Format (QFF) in Level1/Land/Level1c sub-directories by timescale. Saved as a pipe [|] separated file ext = [.psv] Multiple variables per file. Name the QFF using the Merged station ID.

**The processed data stored in Level1 should be mirrored with NCEI, with all the NCEI merge processing code saved on the workspace code directory. This needs some more team discussion and planning for the first data release in September 2019 as most of the sub-daily data processing is being conducted by NCEI.**

Once all the merging and QC has been conducted the QFF data files will be converted into the Common Data Model (CDM) format and saved into Level2/Land directories separated by timescale. The workflow should run sequentially from merging, QC, conversion to CDM format and final files saved in Level2/Land directories. The data could be QC’d and converted into the CDM format by Giorgi regions (23) but needs more team discussion. The QFF will mean that the same code can be used for converting all the data files efficiently on an ongoing basis. The converting of the data files to the CDM format will need to be parallelised across multiple Lotus servers to reduce processing time.

Details of CDM file formats and requirements are outlined in the next section.

**Level2 parent directory**

Level2 contains all the QFF converted into the CDM format. In addition, at this point the following CDM data dependant files need to be produced in pipe [|] delimited text files named and stored in the following sub-directories in Level2:

Level2/land/source\_configuration/

containing psv files: **contact.txt, organisations.txt, source\_configuration.txt.**

Level2/land/station\_configuration/

containing psv files: **station\_configuration.txt, station\_configuration\_optional.txt.**

**It is important that the all these files match up with the QFF station files being converted into the CDM format.**

The level 2 data is in the final CDM format files consisting of header tables and observations tables (see Figure 1). There are two sub-directories within the Level2 land parent directory [header\_tables, observations\_tables] and subsequently three further sub-directories for each data timestep [sub\_daily\_data, daily\_data, monthly\_data]. The following section describes the content of the header\_tables and observations\_tables sub-directories and the file naming conventions.

**Header\_tables sub-directory**

The header\_tables sub-directory contains all the merged CDM formatted data header tables separated into three sub-directories by timescale (see Figure 1). One header table per station and saved as a pipe [|] delimited file (header\_table\_<specifier>\_<label>. psv). Note that there is a <specifier> component and the file names that ties together different files. It is acceptable for a data provider to generate separate Data Delivery Package (DDPs) that refer to the same <specifier>. For example, the source configuration and station configuration components might be delivered in an initial DDP and the actual data (header table and observation table) could be sent later. The <specifier> would be used by both the provider and recipient to identify the connection between the DDPs.

For example (header\_table\_BETA\_EG000062417.psv). The example shows that the specifier relates to the Beta release, this will change with each release, for example the first data release might be (header\_table\_FirstDataRelease\_EG000062417.psv).

**Observations\_tables sub-directory**

The observations\_tables sub-directory contains all the merged CDM formatted data observation tables separated into three sub-directories by timescale (see Figure 1). One observation table per station, saved as a pipe [|] delimited file (observations\_table\_<specifier>\_<label>. psv).

For example (observation\_table\_BETA\_EG000062417.psv). The example shows that the specifier relates to the Beta release, this will change with each release, for example the first data release might be (observation\_table\_FirstDataRelease\_EG000062417.psv).

* There must be the same number of header\_table and observations\_table files.
* The rows in the header\_table and observations\_table files must be paired by timestamps.
* Observations tables contain row for each observation value, observation id element column 1 relate to each observation value per report timestamp e.g, three variables and three entries per report date stamp.

The data files in Level2 directory should be ready for ingestion into the Geo-server and subsequently the Copernicus Climate Data Store (CDS).