Harmonisation Output File Format Definition

Sam Hunt & Ralf Quast

National Physical Laboratory, Fast Opt

3/31/2017

FIDUCEO has received funding from the European Union's Horizon 2020 Programme for Research and Innovation, under Grant Agreement no. 638822

Harmonisation Output File Format Definition

Contents

1 Introduction

To allow for ease of comparison between different approaches to the harmonisation problem within the FIDUCEO project a common output data format is proposed, to be read by a shared diagnostic tool which can perform the set of tests and generate the set of plots agreed upon.

The output is split between two types of file. A first output file contains the output of the optimisation and simple statistics describing the success of the process. A second set of files then contains the residuals of the fit per match-up series for diagnostic purposes. The purpose of this document is to define the form of these files.

2 File Format Description

The data will be stored in NetCDF files, the structure of which will be described in this section.

2.1 Naming Conventions

Firstly, with three methods currently in development and host of test datasets in use which are refined over time a clear naming structure should be defined in order to avoid confusion. This naming structure should be reflected in the harmonisation output filenames and metadata.

2.1.1 Software Implementation Naming

Abbreviations for the method type are shown in

Table 1. To allow a more precise traceability of the software implement used within the data files this abbreviation should be appended with versioning information as follows:

MM_V.V_ TTTTTTT_ CC

where:

MM – harmonisation method abbreviation as defined in

- Table 1.
- *V.V* – is the software implementation version number
- *TTTTTTT* – is the tag from the version control system
- *CC* – Software configuration ID (i.e. number 00-99 which map to defined configurations)

For example version 1.0 of the Fast Opt software with a VCS tag of 3d22cde and in the defined 01 configuration would be named as:

FO_1.0_3d22cde_01

Table 1 - Harmonisation Approach Abbreviated Names

Method	Long Name	Short Name
Fast Opt Approach	Fast Opt	FO
NPL Full Errors in Variables Approach	NPL EIV	EV
NPL ODR Pack with Monte Carlo Approach	NPL ODRMC	OM

2.1.2 Dataset Naming Convention

The input dataset should be referred to with the following naming structure:

INSTR_TYPE_M_CORREL_MCN_YYYYMMDD_YYYYMMDD

where:

- INSTR – Instrument abbreviated name and can take the value:
  - *AVHRR* - For the AVHRR
  - *MW___* - For the microwave instrument
  - *HIRS__* - For HIRS
- *TYPE* – Dataset type abbreviation. This indicates the type of simulation. For example it could indicate a simulation of realistic match-up data or simulations of match-up data from around the globe. Currently can take the value:
  - RSIM – Real simulation
- *M* – Dataset model variant, indicating number of harmonised parameters, for example for the AVHRR case:
  - *3* – for the time independent model
  - *4* – for the time dependent model
- *CORREL* – Dataset correlation structures (as defined by Emma). The description is 4 characters long with each character representing a different form as:
  - *R___* - random correlation present in data
  - *_S__* - systematic correlation present in data
  - *__A_* - averaging correlation present in data
  - *___R* - rectangular correlation present in data

  So in the case of the AVHRR data with its full correlation structure we would have RSA_. More forms can be added to this list as necessary.

- *MCN* – Monte Carlo trial number, from *000* to *999*. If not a Monte Carlo dataset fill with ___.
- *YYYYMMDD* – Three dates are in order: match-up start date and match-up end date

As an example a realistic AVHRR match-up simulated dataset for match-ups between 1st January 1980 and 1st January 2017 created on the 24th March 2017 with the full correlation structure (no Monte Carlo) would have the name:

AVHRR_RSIM_4_RSA_____19800101_20170101

2.2 Harmonisation Output File

This file is intended to include all the information a user requires from the harmonisation process. This is includes the values of the harmonised parameters and associated covariance matrix and some summary statistics to assess the overall success of the optimisation.

Filename

harm_MMMMMMMMMMMMMMMMM_DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD.nc

- *MMMMMMMMMMMMMMMMM* – is the harmonisation software name defined in 2.1.1.
- *DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD* – dataset abbreviated name as defined 2.1.2.

For example, the Fast Opt method (software version 1.0, configuration 01, VCS tag 3d22cde) run on the date of this document's creation with the full AVHRR simulation for match-ups between 1st January 1980 and 1st January 2017 should have a filename:

harm_ FO_1.0_ 3d22cde_01_AVHRR_RSIM_4_RSA_____19800101_20170101.nc

Attributes

- software – *MM* from harmonisation software name defined in 2.1.1.
- software_version – *VV* from harmonisation software name defined in 2.1.1.
- software_tag – *TTTTTTT* from harmonisation software name defined in 2.1.1.
- job_id – *CC* from harmonisation software name defined in 2.1.1.
- matchup_dateset – dataset abbreviated name as defined in 2.1.2.
- cost – sum of squares of final cost/objective function.
- cost_dof – degrees of freedom of the cost function.
- cost_p_value – X^2 Probability (i.e. probability of a X^2 variable with given degrees of freedom taking a value greater than the objective function final value)

Dimensions

- Total number of harmonised parameters
    - Name – n
    - Length – combined number of parameters harmonised for all sensors in the whole dataset

Variables – Required for Diagnostic Plotting

- Harmonisation Parameters
    - Short name – parameter
    - Variable type – double
    - Description – Harmonisation parameters
    - Dimension – n
- Harmonisation Parameter Covariance Matrix
    - Short name – parameter_covariance_matrix
    - Variable type – double
    - Description – Harmonisation parameter covariance matrix
    - Dimension – n, n
- Harmonisation Parameter Sensor Name
  e.g. for the AVHRR dataset this would be an array as:
  [m02, m02, m02, m02, n19, n19, n19, n19…]
    - Short name – parameter_sensors
    - Variable type – char
    - Description – Sensors associated with harmonisation parameters
    - Dimension – n

Variables – Optional

Below are examples of potential additional variables that can be added if useful but are not required by the diagnostic plotting software (additional attributes can also be added if more appropriate):

- Harmonisation Parameter Correlation Matrix
    - Short name – parameter_correlation_matrix
    - Variable type – double
    - Description – Harmonisation parameter correlation matrix
    - Dimension – n, n
- Harmonisation Parameter Hessian Matrix
    - Short name – parameter_hessian_matrix
    - Variable type – double
    - Description – Harmonisation parameter Hessian matrix
    - Dimension – n, n
- Harmonisation Parameter Uncertainties
    - Short name – parameter_uncertainties
    - Variable type – double
    - Description – Harmonisation Parameter Uncertainties
    - Dimension – n, n

2.3 Harmonisation Residual Files

This set of files is intended to include the fit residuals for diagnostic purposes. The files are of a similar form to the input match-up data and as such are split into individual files for each match-up series. The variables H_res and k_res contain the residuals of the data and is of the same structure as H and K in the input match up data.

NB: For the NPL EIV approach H_res will be filled, however for the Fast Opt approach where the data are not treated as parameters this variable does not need to be included.

Filename

harm_MMMMMMMMMMMMMMMMM_DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD_res_Si_Sj.nc

- *MMMMMMMMMMMMMMMMM* – is the harmonisation software name defined in 2.1.1.
- *DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD* – dataset abbreviated name as defined 2.1.2.
- *Si* – abbreviated name of the first sensor of the match-up series pair.
- *Sj* – abbreviated name of the second sensor of the match-up series pair.

For example, a residual file for the example harmonisation file described in 2.2 for the Meteosat 2 - NOAA-19 pair would be:

harm_ FO_1.0_ 3d22cde_01_AVHRR_RSIM_4_RSA_____19800101_2017010_res_m02_n19.nc

Attributes

- software – *MM* from harmonisation software name defined in 2.1.1.
- software_version – *VV* from harmonisation software name defined in 2.1.1.
- software_tag – *TTTTTTT* from harmonisation software name defined in 2.1.1.
- job_id – *CC* from harmonisation software name defined in 2.1.1.
- matchup_dateset – dataset abbreviated name as defined in 2.1.2.
- sensor_i_name – name of the first sensor of the match-up series pair.
- sensor_j_name – name of the second sensor of the match-up series pair.

Dimensions

- Number of columns of H_res matrix
    - Name – n_col
    - Length – combined number of covariates for two sensors
- Total number of match-ups
    - Name – m
    - Length – match-ups in series

Variables – Required for Diagnostic Plotting

- Harmonisation Match-up Data Residuals
    - Short name – H_res
    - Description – Data residuals
    - Dimension – n_col, m
- Harmonisation Match-up Adjustment Factor Residuals
    - Short name – K_res

- Description – K residuals
- Dimension – m

Variables – Optional

As for the harmonisation output file additional variables (or attributes) can be added where useful, but are not required by the diagnostic plotting software.