

Standardization of data formats, metadata, access protocols and statistical visualization of SS3 stock assessment outputs

Anne-Elise Nieblas*, Sylvain Bonhommeau†, Taha Imzilen‡,
Dan Fu§, Fabio Fiorellato§, Julien Barde¶

November 27, 2017

SUMMARY

*Stock assessment model outputs are rarely available for users who want to explore these datasets. Some systems like RAM Legacy database can make parameterization of the model available but these parameters do not come along with data nor metadata. Traceability is an increasingly desired end product of any work flow, including a stock assessment work flow. To be able to verify the history, location, or application of a process or data by means of documented recorded identification allows a work flow to be fluidly and rapidly exchanged between users and environments and ensures continuity and replicability. This can be achieved by applying standard protocols to each step of the work flow. As part of this work, we attempt to define some standards in terms of stock assessment output in order to enhance traceability and replicability of stock assessment work flows. As an aspect of this, we have developed code that will transform the outputs of stock synthesis 3 into NetCDF so that metadata can be embedded, and that these data can be hosted by open source online access points. These data can then be visualized by a standard statistical visualization tool. Here, we provide an example of a standardized work flow, showcasing standardized data formats, metadata, access protocols, and statistical visualization as applied to the 2017 IOTC skipjack (*Katsuwonus pelamis*) stock assessment.*

*IRD, UMR MARBEC (IRD/Ifremer/Univ.Montpellier/CNRS), IRD Réunion, 97744 St Denis, La Réunion, France; anneelise.nieblas@ird.fr

†IFREMER- DOI, rue Jean Bertho, 97822 LE PORT CEDEX, La Réunion, France

‡IRD - UMR MARBEC 248, Av. Jean Monnet, 34200 Sète, France

§IOTC, Le Chantier Mall, Victoria Mahé, Seychelles

¶IRD, UMR MARBEC (IRD/Ifremer/Univ.Montpellier/CNRS), IOC, Rue de l'Institut, Ebène, Maurice

KEYWORDS: Indian Ocean, large pelagic fish, skipjack tuna, *Katsuwonus pelamis*, scientific cloud, stock assessment, grid computing, online processing, traceability, data standards, metadata, access protocols, Shiny

1. Introduction

Motivation is mounting to provide collaborative tools to aid the work flow and collaboration within and between scientific groups. At the IOTC, a new virtual research environment (VRE; [Candela et al., 2013]) has been created using the BlueBridge infrastructure ([Coro et al., 2017]; European Union grant agreement No 675680) to facilitate the parametrization, parallelization, and execution of various steps of the stock assessment work flow and the visualization of the results for the Stock Synthesis version 3 (SS3) model ([Methot, 2013, 2017]) to users with varying levels of expertise (IOTC_SS3; [Nieblas et al., 2017a,b]). Within these sorts of tools, which have users of diverse expertise that are using diverse computing environments, it is necessary to have a standard manner to manage the data inputs and outputs. Furthermore, a description of the data and the processes required to develop them is necessary to replicate the work flow and keep track of the changes that have occurred within the process.

At the WPDCS13, we will present an example of this standardized work flow showcasing the 2017 SKJ stock assessment model grid, how and where these runs are stored and how they can be used by a statistical visualization application in order to compare model runs.

2. Standards for the management of stock assessment outputs

Efforts have been made to find relevant standards for stock assessments to make this work flow as generic as possible and therefore transferable to other cases (e.g., to different runs, species, models). Here, we describe how we have complied with global standards to standardized the data formatting, metadata, access protocols and statistical visualization of the stock assessment model outputs.

There is currently no agreed-upon standard by which to standardize the outputs of stock assessment models. Here, we follow the [climate and forecast conventions](#) (CF conventions) that are widely-used within the marine science community (i.e., NOAA's [Global Ocean Observing Systems](#)) to provide recommendations for which variables to output, naming conventions for output files, variable names and units, metadata formats, data format (i.e., NetCDF), and access protocols that can be applied regardless of the stock assessment model used.

2.1 Tools to complement other initiatives

By increasing models' uncertainty calculations, we increase the number of model runs, and thus increase the amount of outputs to be described. Systems like the RAM legacy provide an excellent database for describing the outputs of

reference runs, but their database is reliant on manual collection of stock assessments, sourced from public websites, the literature, the stock assessment officers themselves ([Ricard et al., 2012]). Our aim is to complement the RAM legacy database (and other initiatives like it, e.g., ICES), such that we describe at least the final run for each stock assessment, along with the multiple other model comparisons, using standard data and metadata formats that could directly inform these other databases. In the case of the IOTC, there can be an entire grid of model runs (between 100 - 200 different parameterizations), and we suggest that the information to run this grid be stored for future years. Our tools would complement the RAM legacy database such that model data as well as code would be stored such that past stock assessment could be easily replicated.

2.2 Core Variables for Stock Assessment Models

The list of variables that are simulated differs between stock assessment models but there is a core of common variables between models which can be standardized by conventions.

In this [spreadsheet](#) we establish a mapping between SS3 and virtual population analysis (VPA) stock assessment models to identify common variables (SS3 variables summarized by Table 1). By doing so, we aim to start a set of recommendations to adopt common conventions for these core variables as has been done for environmental variables by the CF conventions with the [CF Standard Name Table](#).

Table 1: **Core variables for stock assessment models**, using selected outputs and variable names from the SS3 model, to be mapped to similar core variables for the virtual population analysis (VPA) model.

Variable name	Definition
BBmsy	B/Bmsy
FFmsy	F/Fmsy
Vuln_bio_cpue	Vulnerable biomass
Obs_cpue	CPUE observations
Exp_cpue	Expected CPUE
SE_cpue	standard error of cpue
Dev_cpue	cpue deviations
Like_cpue	cpue likelihood
Likelogs_cpue	log likelihood of cpue

natage	numbers at age
meanbodywt	mean body weight
Bio_all_spr	Fished total biomass at R0
Bio_Smry_spr	fished summary biomass at R0
SPBzero	unfished spawning biomass at R0 and current year's life history
SPBfished	fished spawning biomass at R0 and current year's life history
SPBfished _R	fished spawning biomass per re- cruit
SPR	spawning potential ratio equal to SPBfished/SPBzero
SPR _{std}	standard deviation of the spawn- ing potential ratio
YbyR	yield per recruit
GenTime	unfished generation time equal to mean age weighted by fecundity at age
F_std	standard deviation of F
SPB	spawning biomass
Recruits	recruits
Tot_Exploit	total exploitation
aveF	average F across ages beginning at the summary age
maxF	maximum F among ages for each gender
sum_Apical_F	sum of the Apical F
FeqZminusM	F equals Z minus M ($F=Z-M$)
spr	spawning potential ratio
matage	m at age
zatage	z at age
catage	catch at age
growthseries	growth series
movement	movement rate between areas in a multi-area model
ageselex	age selection
sizeselex	size selection
Bio_all.ts	Time series of the biomass of all ages

Bio_smry_ts	Time series of the biomass for ages at or above the summary age
SpawnBio	Time series of the female spawning output
Recruit_0	Time series of recruitment of age-0 fish in this year
obs_cat_ts	time series of observed catch
F_ts	times series of F
Obs_len	length observations
Exp_len	expected length
RecDev_val	recruitment deviation
RecDev_lci	lci of recruitment deviation
RecDev_uci	uci of recruitment deviation

2.3 Standardization of metadata for stock assessment

Using the same recommendations as the global community, we adopted the set of [metadata elements recommended by NOAA](#) that provide a mapping with the International Organization for Standardization (ISO) [19115 standard](#) (widely used for geographic metadata). Compliant metadata are written for the execution of each run with a dedicated R function that describes the inputs, parameters, and processes that were necessary to generate a given run output. These metadata elements can be directly embedded within the NetCDF data file (see [Section 2.4](#)) by using the `ncdf4` R package. Once in the NetCDF files, we use another R standard function to extract the metadata from NetCDF files (using OPeNDAP access protocol) and transform them into the Open Geospatial Consortium (OGC) / ISO 19115 standard, which can then be harvested and displayed by multiple metadata catalogs (e.g., Geonetwork: see [example](#); [Figure 1](#)). By doing so, searching these metadata allow data and process discovery and the ability to replicate an assessment.

2.4 Standardization of data formatting and access protocols

To properly standardize a work flow, data outputs should be independent from the programming languages that have been used to create them (e.g., C++, R in the case of SS3 model runs) so that they could be stored, shared, and reused by a broad community of users ([\[Imzilen et al., 2017\]](#)). Though the IOTC_SS3

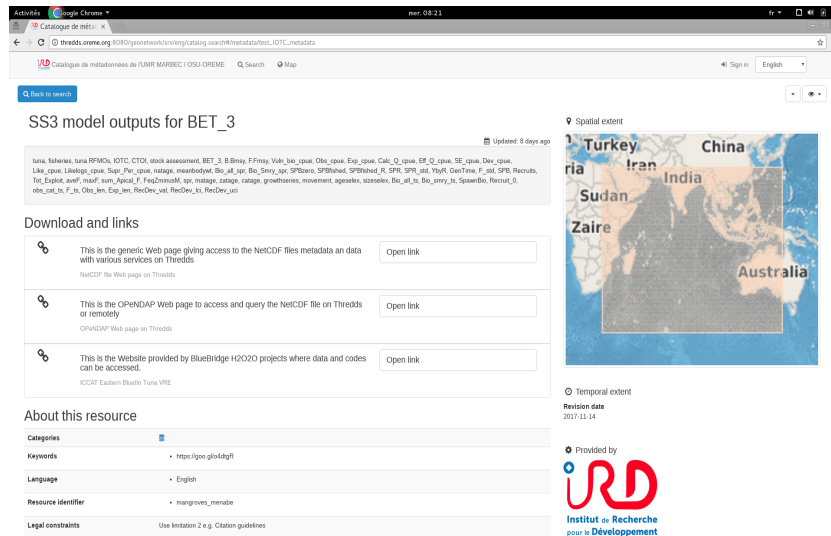


Figure 1: Example of stock assessment metadata displayed by Geonetwork once transformed in OGC metadata

VRE work flow relies primarily on R, we transform the outputs of the model into NetCDF data format.

Indeed, NetCDF is a widely-used data format which can be used by most scientific programming languages and is becoming a global standard of ocean models, ocean data, and satellite data. In 2012, NetCDF became an official [Open Geospatial Consortium standard](#). This format enables us to store model outputs on open source data servers, such as [Thredds](#), such that data and metadata are available from a single online access point (Figure 2).

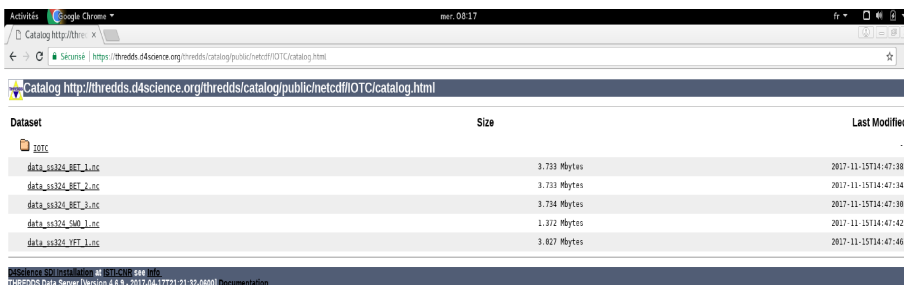


Figure 2: Thredds server to access stock assessment outputs once transformed into NetCDF files

We created an [OutputsSS2NetCDF R function](#) which turns SS3 model out-

puts, created using the `r4ss` R package (a common method for reading SS3 outputs; [Taylor et al., 2011]), into NetCDF files, embedding metadata and outputting select, core variables (see Table 1).

Indeed, by using NetCDF, data become natively available with the http-based OPeNDAP access protocol once they are deployed on data servers, such as open source Thredds servers (Figure 2 or see example). Such servers offer multiple remote-access protocols. It then becomes possible to interact remotely with the data on the server in order to extract metadata or to subset data in a programmatic way, as NetCDF can be read by most existing programming languages.

This function, implemented in R, is generalized for `r4ss` output, and can be applied to any species from any regional fisheries management organization; thus making any SS3 output available remotely. It then becomes possible to visualize and compare outputs of many different runs and thus enhance traceability of past stock assessments.

2.5 Naming conventions for NetCDF outputs of model runs

A standard naming convention allows users to easily identify between files of different stock assessments and model runs, immediately exhibiting key metadata. We are developing a standard naming convention for stock assessment model runs, which currently includes:

- generic information e.g., `stock_assessment_output`
- model e.g., `SS3`
- species e.g., `SKJ`
- management area e.g., `IOTC` or `Indian_Ocean`
- temporal extent of inputs (start date and end date) e.g., `1975_2015`
- key parameters e.g., `h70_MAt_t3_rtss_tm15` (steepness, maturity, x , tagging x , x)

So, for this example, the full name of the file would be:

```
stock_assessment_output_SS3_SKJ_IOTC_1975_2015_h70_MAt_t3_rtss_tm15.nc
```

2.6 Standardization of statistical visualization (Shiny)

The Shiny, [SS3 Diagnostic Plots](#), has been developed to provide a quick and easy manner to visualize, assess, and compare model runs. The Shiny links to the various model runs using OPeNDAP access protocol on the Thredds server, and then uses these data as inputs to inform interactive diagnostic plots. Currently,

the Shiny plots the catch-per-unit-effort (CPUE) fit to observations, standardized CPUE residuals, size frequency plots of aggregated length frequency, mean length distribution, and mean weight distribution, biomass plots of biomass, depletion (%), and spawning stock biomass (SSB) by area and summed over area, F/F_{msy}, recruitment deviation, recruitment by area, and total recruitment (summed over area), and finally, size selectivity. Including additional plots is possible, and suggestions for which plots to include are welcome.

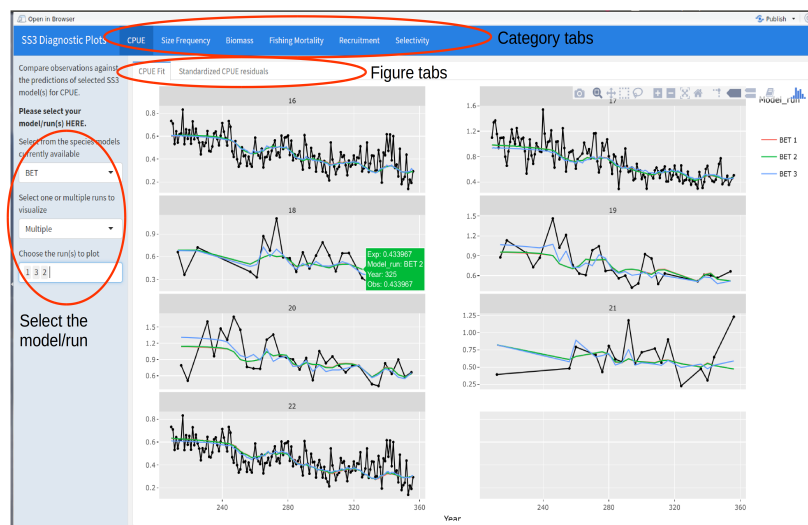


Figure 3: **The R Shiny allows users to visualize the results of current or past model runs**, allowing runs to be overlaid on interactive plots to compare different model parameterizations.

3. Benefits of standardization

Standardization of the stock assessment work flow means that parameters, processes, and data can be backed up for future use. This ensures traceability and replicability of past stock assessments. Furthermore, model outputs can be easily browsed and visualized with graphical interfaces (e.g., Shiny), enabling data and process discovery. Defining and applying agreed-upon standards means that a diverse array of models can be evaluated, with no dependency on RFMO, stock area, species, or even (eventually) model.

4. Conclusions

The BlueBridge project enables an online collaborative environment by providing the infrastructure necessary to parametrize, visualize, and access a workflow on a VRE. A collaborative environment such as the VRE uses simple interfaces to facilitate the storage and access of data on an open source Thredds server, whose data can be accessed using standard OPeDAP protocols. Stable data and metadata storage such as this ensure traceability and replicability of the stock assessment workflow. Finally, the VRE provides a standard visualization of the results of one or many model runs, allowing quick, easy, and direct comparisons between runs. This environment is available to members who can share documents, messages, data and codes in both a public and private space.

The end goal of the IOTC_SS3 VRE is to allow participants of stock assessments to use the VRE interactively at working groups in order to explore input parameters and results, to store and replicate past results, to give more transparency to the decision-making process, and to enhance collaboration within working groups. Improving the ease of use by providing descriptions of the complex SS3 model will bring more transparency and collaboration within working groups. Technical performance, document production, and harmonization of content are expected to be enhanced due to this increased participation. We hope to show the potential of this environment to foster collaboration and incorporation of scientific advice within working groups. We particularly encourage feedback on these tools and their application from the community of users to improve their utility in the future.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under the BlueBRIDGE project (Grant agreement No 675680).

References

- L. Candela, D. Castelli, and P. Pagano. Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal, 12:GRDI75–GRDI81, 2013. doi: <http://doi.org/10.2481/dsj.GRDI-013>.
- G. Coro, G. Panichi, P. Scarponi, and P. Pagano. Cloud computing in a distributed einfrastructure using the web processing service standard. Concurrency and Computation: Practice and Experience, 29, 2017.
- T. Imzilen, S. Bonhommeau, T. Rouyer, L. Kell, and J. Barde. Online collaborative environment to run the eastern bluefin tuna stock assessment workflow. Collect. Vol. Sci. Pap. ICCAT, 73:2528–2534, 2017.
- R. Methot. User Manual for Stock Synthesis: Model Version 3.24s., 2013.
- R. Methot. User Manual for Stock Synthesis: Model Version 3.30., 2017.
- A. Nieblas, S. Bonhommeau, T. Imzilen, D. Fu, F. Fiorellato, and J. Barde. An online tool to easily run stock assessment models, using SS3 and SWO as an example. IOTC Working Party on Methods, 8, 2017a.
- A. Nieblas, S. Bonhommeau, T. Imzilen, D. Fu, F. Fiorellato, and J. Barde. An online tool to easily run stock assessment models, using SS3 and YFT as an example. IOTC Working Party on Tropical Tunas, 8, 2017b.
- D. Ricard, C. Minto, O. Jensen, and J. Baum. Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish and Fisheries, 13:380–398, 2012.
- I. Taylor, I. Stewart, A. Hicks, T. Garrison, A. Punt, J. Wallace, and C. Wetzel. r4ss: R code for Stock Synthesis. R package version 1.16., 2011.