IOTC-2016/WPM07/XXX

Online collaborative environment to run Stock Assessment workflow: an option for IOTC?

Taha Imzilen[*], Sylvain Bonhommeau[†], Tristan Rouyer[‡],
Laurence T. Kell[§], Emmanuel Chassot[*],
Julien Barde[*]

SUMMARY

In this note, we present an approach which has been recently used to execute online the set of codes used for eastern bluefin tuna (BFT-E) stock assessment at ICCAT. We aim at discussing the possible interest of the approach to the IOTC Community for focusing on Indian ocean tuna and tuna-like stocks and running other assesment models such as Stock Synthesis 3 (SS3). In 2014, ICCAT BFT-E working group has been able to execute thousands of model runs by using parallelization of R and Fortran codes on a supercomputer. This approach has a lot of scientific benefits, in particular for better including the multiple sources of uncertainty in assessment results and associated scientific advice. However, very few participants would be able to reproduce it without specific technical skills. Since November 2015, a WebSite or VRE for Virtual Research Environment has been set up within the H2020 BlueBridge project to enable users to easily parametrize and execute various steps of the BFT-E stock assessment workflow. By repackaging codes provided by ICCAT BFT-E working group, the same codes are now executable online. They can be parametrized, executed and edited by anybody from a simple web page and data outputs are delivered in standard data format. At this stage, this VRE comes with various collaborative Web Services: (i) a workspace to share documents or data, (ii) Web pages or RStudio server to process data online, and (iii) an automated report service to dynamically generate documents packaging these results. Such a collaborative environment enables

[*]IRD - UMR MARBEC 248, Av. Jean Monnet, 34200 Sète, France; taha.imzilen@ird.fr; Phone: +33 499 57 32 32 Fax: +33 499 57 32 15.

[†]IFREMER - Le Port, La Runion, France; sylvain.bonhommeau@ifremer.fr; Phone: +33 499 57 32 66 Fax: +33 499 57 32 95.

[‡]IFREMER - UMR MARBEC 248, Av. Jean Monnet, 34200 Sète, France; sylvain.bonhommeau@ifremer.fr; Phone: +33 499 57 32 66 Fax: +33 499 57 32 95.

[§]ICCAT Secretariat, C/Corazón de María, 8. 28002 Madrid, Spain; Laurie.Kell@iccat.int; Phone: +34 914 165 600 Fax: +34 914 152 612.

to store and access the whole set of data and source codes to replicate past results or to try new parametrizations of the model with usual tools or simple web forms. Such an approach might bring more transparency and collaboration within working groups.

KEYWORDS: bluefin tuna, scientific cloud, stock assessment, grid computing, online processing

1. Introduction

The goal of the present work is to execute online the whole workflow for BFT-E Stock Assessment as done in 2014 on Ifremer supercomputer. We argue here that this kind of approach is generic enough and then worth to be tried with other RFMOs focusing on other models and species. With our first attempt, we wanted to showcase for next Stock Assessment working groups that such an approach can be achieved online without having to deal with complex command lines. BlueBridge H2020 project provides Web collaborative environments (VRE for Virtual Research Environments; [Candela et al., 2014, 2013]) to enable any kind of users to check this work with dedicated data access and processing services related to ICCAT BFT-E (current Web Site). In this note, our main goal is to showcase how this VRE has been be used by ICCAT to process data available in 2016 and how it could be of potential interest to IOTC. We will first describe how the ICCAT workflow has been split in different steps to (i) process data, (ii) visualize and choose the most relevant output, (iii) run projections, (iv) generate and package plots within automated reports. These different steps can be executed "as usual" from R by using RStudio online integrated within the VRE, or from Web Forms where users do not have to deal with programming languages but focus on parametrization. It has to be noted that this work also demonstrates the need for standard data format to store and expose Stock Assessment data wherever they come from. In the same way, the Web Processing Services used comply with best practices to execute processes on remote server. By doing so, such a work can be re-used in different contexts by porting codes to different IT infrastructures. At this stage, the feedback from the community of users is needed to improve this work and provide a more robust application in the coming years.

2. Materials and Methods

The work presented in this note has been driven by an IT engineer in charge of describing, repackaging codes (provided by ICCAT) usually executed on local machines (and on a supercomputer since 2014) to deploy and execute them online. It is important to note that this engineer is not an expert in stock assessment but that he has been acting as a mediator between ICCAT working group and the team administrating the IT infrastructure of the Bluebridge project.

Codes and data have been provided by ICCAT, the underlying IT infrastructure [Candela et al., 2015] has been provided by BlueBridge H2020 project:

- Codes:

  - VPA Fortran code,
  - R codes to feed the Fortran VPA model and process outputs. In particular following packages:
  - knitr (LaTeX+ R codes) codes to generate automated reports,
  - inputs: ICCAT datasets,
  - processes: R codes to feed Fortran VPA model and process outputs as well as to generate report,
  - outputs: R data object (from FLR package) have been transformed into netCDF data formats, .html (maps, plots), pdf (reports),

- Grid Softawre and Hardware:
  - online RStudio,
  - Web Processing Server to execute above codes as Web Services (OGC WPS and OpenCPU) from remote "clients" (various porgramming languages, Web Browsers, GIS, etc.),
  - A grid of servers to store, access and process data with above codes

To set up a similar work with IOTC, we should work directly with IOTC scientists running the stock assessment model in order to compile the code online. We are currently investigating the opportunity to use SS3 online but other assessment models such as surplus production models used for some billfish species as well as ASAP (Age Structured Assessment Program) used for bigeye could be envisaged.

3. Description of the Workflow

By itself the workflow used in the present work is generic and might be worth to be reused for different use cases (other species and stock assessment working groups). It has been deliberately split in different steps (through R functions). Some steps are the same from a group to another. Parallelization is made possible by the infrastructure when needed [Coro et al., 2015]. In the case of BFT-E the model uses VPA Fortran but other models could be chosen in the future and it is thus important to separate steps that might be re-used by other models:

Four main steps:

- STEP 1 (heavy): Analysis (retros) of the BFT-E ICCAT datasets,

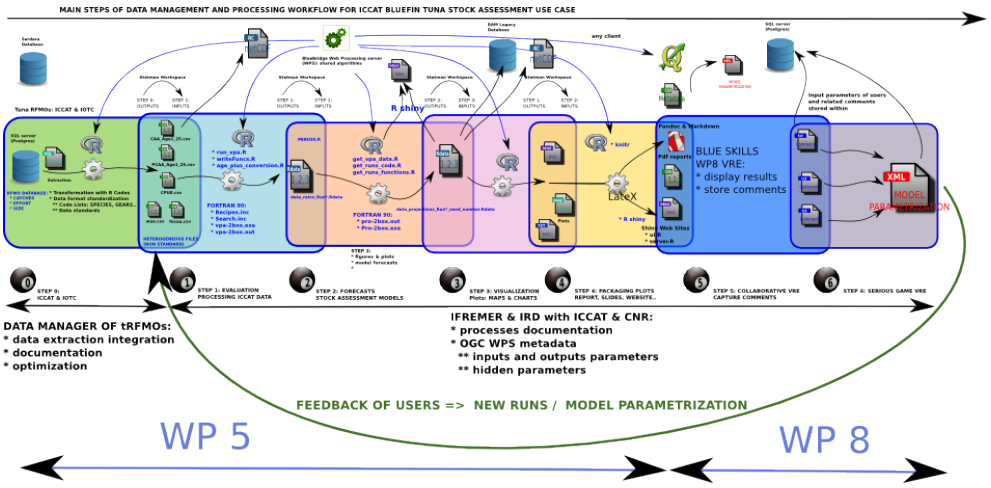

Figure 1: The example of ICCAT Workflow for Eastern Bluefin Tuna

- STEP 2 (light and generic): Visualization of data analysis (retros),
- STEP 3 (heavy): Projections which is the most demanding step in terms of machine resources,
- STEP 4 (light and generic): Writing the main structure and plots of the executive summary by using an automated report (see results with OPeNDAP)

Steps 2 and 4 have been improved to use netCDF files as inputs to generate plots. By adopting common data formats, steps 2 and 4 might be directly reused by any other working group (see section 4.1). Step 1 and Step 3 underlying codes depend, of course, on stock assessment model but the main benefit is related to machine resources and possible parallelization.

4. Promote standards to store, access and process stock assessment data

Standardization is a key issue when trying to replicate some work or to adapt it outside of its original context. This is worth for data formats, data access and processing services. To reach a certain level of standardization data structures and pieces of codes have to be reorganized to avoid replication of efforts.

4.1 A standard data format to store Stock Assessment data

Our main goal in this part of the work consists in defining a standard data structure implemented with a well knowk data format (netCDF) which enables

to store and expose data coming from both outputs of retros and projections of stock assessment models. Such an effort could be made as well to transform declarative data from member states and related corrected data from scientists of RFMOs. Such a data format could be used for example to facilitate the storage of data in the RAM Legacy database.

4.11 Benefits of standard data formats

In the first version of the work, data generated by the workflow where saved as *R data* object making them hardly reusable outside the R environment. However, data outputs should be independent from underlying programming languages (e.g. R and Fortran) so that they could be stored shared and reused by the community of users. This was identified in 2014 and, since, we have been working on converting the native R data format into widely used data formats.

The netCDF data format is a good candidate as it is widely used for model outputs and ocean observatories. In such a data format, it becomes possible to store multiple run outputs and to expose them through a single access point by using open source servers like Thredds which offers multiple way to remotely access these data. In particular OPeNDAP protocol access enables an access from most existing programming languages. Data can now be shared online, remotely accessed to visualize and compare outputs. Here is an example of projection output made publicly available with Thredds. By using Thredds, datasets can be harvested by metadata catalogs like Geonetwork which brings a valuable service for Stock Assessment Data Discovery. This require a specific work on metadata enrichment.

Step 2 and Step 4 of the workflow are consuming data directly from a server. In this way, it becomes possible to store and replay past results, plots, and reports with mulipe programming languages and applications.

4.12 Current data structures for retros and projections

As illustrated by Figures 2 and 3, dedicated R functions have been created to turn R Data Objects generated by the workflow (using FLR package) into more generic data formats.

As FLR is widely used by different stock assessment working groups, these functions might be reused and, when necessary, enriched to store additional outputs from other models and species. In particular, by doing so, steps 2 and 4 to plot time series and generate reports might be reused directly whatever underlying species and models.

The data structure of files coming out of the retrospective analysis uses the following naming convention: **data_retro_Run**.nc**. Each file is made of 8

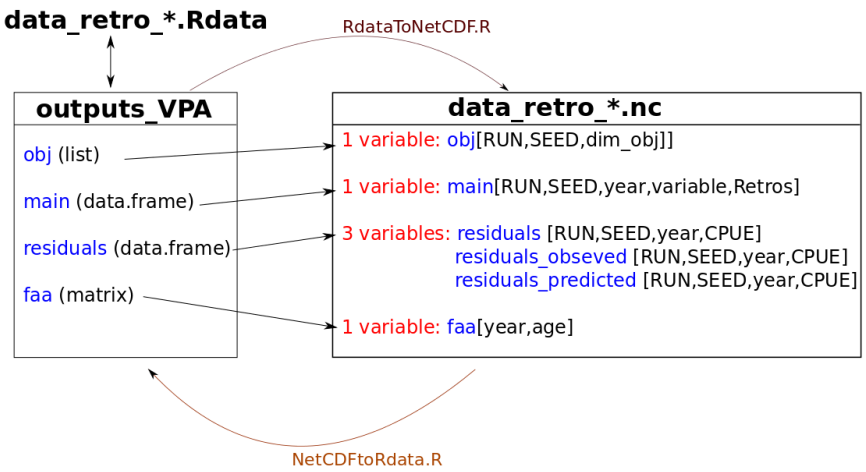

Figure 2: R function to turn retros R data output in netCDF files

dimensions and 6 variables (which depend on all or part of previous dimensions):

- 8 dimensions :
  - RUN (integer): number of Run,
  - SEED (double): number of seed,
  - CPUE: ESPMarTrap,JLL EastMed,Nor PS,JP LL NEA,SP BB1,SP BB2, SP BB3.
  - Retros: number of retros,
  - age : Age 1,Age 2,Age 3,Age 4,Age 5,Age 6,Age 7,Age 8,Age 9,Age 10plus,
  - variable: SSB (spawning stock biomass),Recruits, F2.5, Fplusgroup.
  - year (double): year time.
  - dim_obj:
    - obj_func: Value of the objective function (no unit),
    - obj_func_with_cte,
    - nb_param: Number of parameters estimated by the model (no unit),
    - nb_data: Number of data used by the model (no unit),
    - AIC:Akaike Information Criteria (no unit),

    - AICc: Akaike Information Criteria (no unit),
    - BIC:Bayesian Information Criteria (no unit),
    - chi_square,

- 6 variables (with related dimensions in brackets) :
  - main(Retros, variable, year),
  - residuals(CPUE, year): residuals in catch (no unit),
  - residuals_obseved(CPUE, year): observed residuals in catch (no unit),
  - residuals_predicted(CPUE, year): predicted residuals in catch (no unit),
  - faa(age, year, SEED, RUN): fishing mortality by age (in ŷ-1),
  - obj(dim_obj, SEED, RUN).

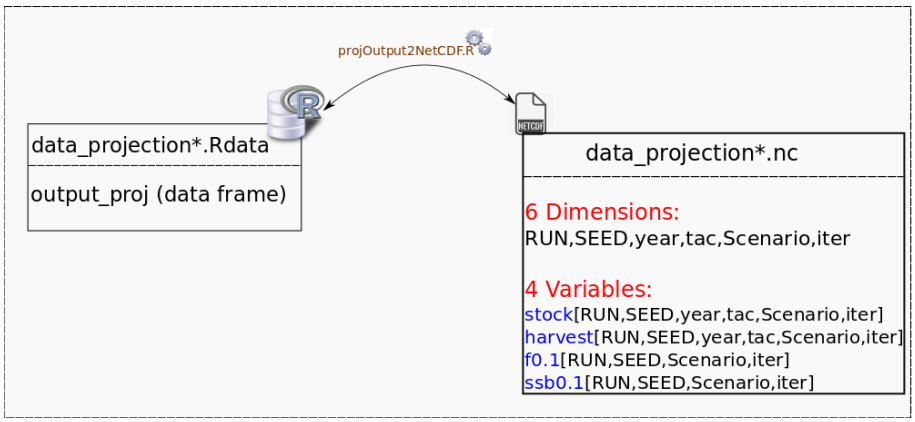

Figure 3: R function to turn projections R data output in netCDF files

Regarding the data structure of files coming out of the projections, we use the following naming convention: **data_Proj_Run**.nc**. Each file is made of 6 dimensions and 4 variables (which depend on all or part of previous dimensions):

- 6 dimensions :
  - RUN (integer): number of Run,
  - SEED (double): number of seed,
  - Scenario (integer): Recruitment scenario (1= "low recruitment": average recruitment over 1970-1980; 2= "medium recruitment": average recruitment over 1955-2006; 3= "high recruitment": average recruitment over 1990-2000),

  - iter (integer): number of iteration,
  - tac (double) : Total Allowable Catch (in kg),
  - year (double): year time.

- 4 variables (with related dimensions in brackets) :
  - stock(iter, Scenario, tac, year, SEED, RUN)(float): abundance by age and year (in numbers),
  - harvest(iter, Scenario, tac, year, SEED, RUN)(float): number of individuals caught by age and year (in numbers),
  - f01(iter, Scenario, SEED, RUN)(float): Fishing mortality as a procy for fishing mortality at MSY (in y^-1),
  - ssb01(iter, Scenario, SEED, RUN)(float): Spawning Stock Biomass (in kg).

4.2 Standards for Web Processing Services (WPS)

Processes like data can be accessed and executed remotely from various programming languages (R, Java, Python, Javascripts, etc.) and clients (Web Browsers, Qgis, etc.). To get such services, most of the work consists in complying with good practices which mainly consist in restructuring the codes to isolate generic functions where inputs and outputs are well defined. Once there, as illustrated by Figure 4 one or multiple functions will be used to set up a **process** and related **parameter(s)**. A process is just a part of a **workflow**. Each process can be parametrized through a set of **input parameters** and can produce a set of **output parameters**. By doing so, by storing different types of parametrization, it becomes possible to execute the whole workflow and thus to replicate pas workflows (meaning the work done by stock assessment working groups along the years).

To achieve this we have been implementing the Web Processing Service standard as recommended by the Open Geospatial Consortium (OGC). Such a standard facilitates porting a code or a process from an environment to another if needed. To enable a thinner integration within the websites of partners we have used OpenCPU protocol as well. It is possible to run the processes on such a server remotely by using most of scientific programming languages (R, Python, Java..) which are implementing these standards as well. At this stage data and processes used by the workflow can be accessed within a dedicated Website or remotely.

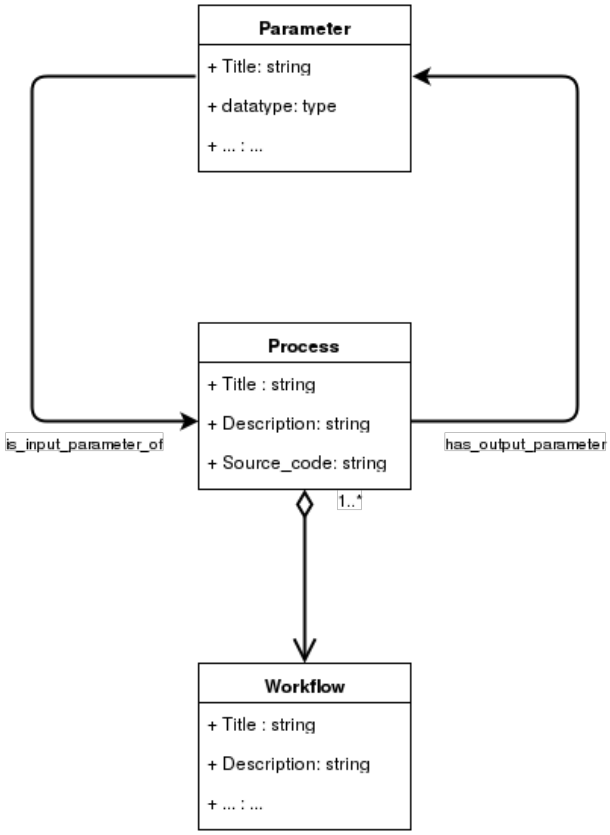

Figure 4: Structuring codes to manage workflows

5. Online collaborative environment for ICCAT BFT-E

The BlueBridge project and underlying infrastructure provide a set of useful services to access and parametrize such a workflow. Each collaborative environment (or VRE, Virtual Research Environment) comes with a list of members who can share documents and messages within this public or private environment. Data and codes can the be accessed and shared by all members. In addition to these generic services, users usually need for relevant services for their specific community (eg RStudio). The Figure 5 illustrates how services can now be managed in the cloud and accessible to any member of the group.

5.1 RStudio online to run a Stock Assessment workflow

Rstudio server works exactly like the desktop application. Each user can access its private workspace separately. However, sharing the same RStudio limits configuration issues and ensure that every member will be able to compile successfully the codes as all packages are already installed when a newcomer joins

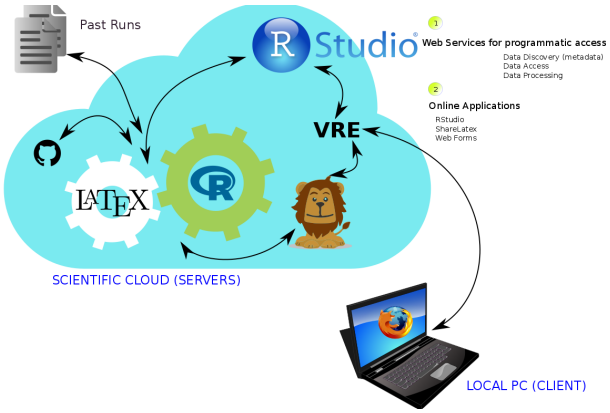

Figure 5: VRE: a collaborative Website to access services managed in a scientific cloud

the environment. When one provides new codes compiled with RStudio of the VRE, everybody will be able to execute them.

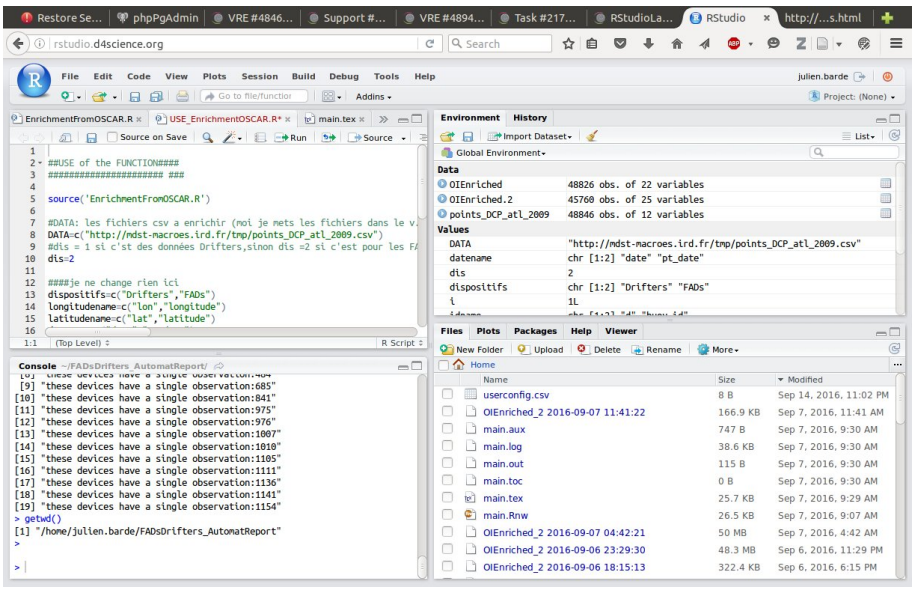

Figure 6: Users can directly edit and run the codes with Rstudio online

The first step with IOTC would consist in checking that R codes can be compiled with RStudio and add, if needed, extra OS or R packages.

5.2 *Run the stock assessment from external Web pages*

As not all users get skills in R or Fortran, we made an effort to package these codes on the server in a way which enables their execution from many "clients" like Web Forms or GIS for example. We complied with the guidelines of OGC Web Processing Service (see 4.2) which makes possible to run these codes from a Web Page directly in the VRE as shown in Figure 7.

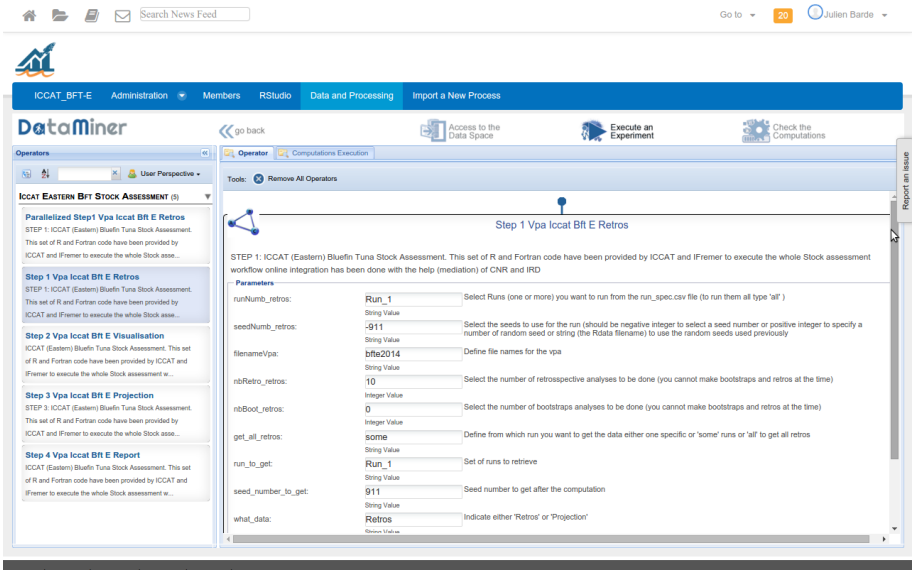

Figure 7: Step 1 can be executed from a Web page of the VRE

A similar Web form can be generated in any Web page (out of the VRE) as shown in Figure 8 (check online demonstration).

Users can focus on parametrization and do not deal anymore with programming. If the code is of interest out of the VRE, they can integrate a Web Form in any Web page to run it.

5.3 *Collaborative edition of reports*

Since last ICCAT meetings, we asked the BlueBrige project to investigate the feasibility of providing collaborative environment to edit automated reports with R and Latex. ShareLaTeX might be used soon to facilitate the collaborative edition of executive summary or other reports where most of the plots can be automated as done in Step 4 of ICCAT BFT-E workflow.

Having ShareLaTeX enabled within the infrastructure of BlueBridge would enable to compile both LaTeXand R codes in the VRE by using the same R server (already used by RStudio) which ensures that reporting activities and

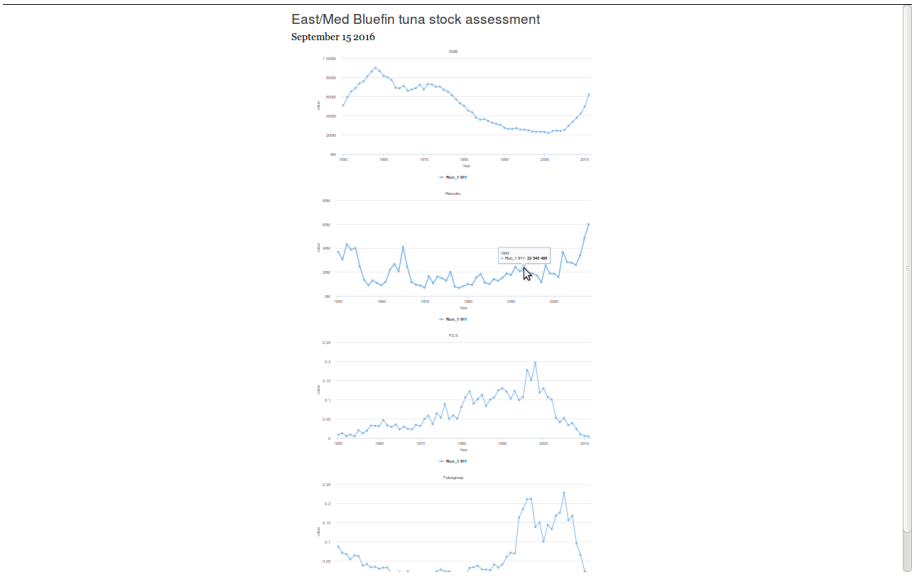

Figure 8: Visualization of retros at step 2 (results from step 1)

codes compilation can be done in the same environment.

6. Results and Discussion

This report showcases what can now be achieved with online collaborative environments. The use case of BFT-E relies on a set of R and VPA Fortran codes. However such an approach is generic enough to execute other types of models using other programming languages (Java, Python, R, C++, etc.). In the same way the scientific cloud offered by BlueBridge project might change but technical aspects used so far can run on other infrastructures.

More than the use case or the tools, the main benefit is the collaborative environment users can now expect to improve the way they produce and discuss results. This is as well relevant to ensure the reproducibility and then transparency of the workflow over the years. Tools and models will evolve but technology can facilitate the adoption of models requiring more skills and machine resources.

At some point, a similar approach could be replicated with additional models executed on other infrastructures / clouds (Amazon, Google, etc.) but the main question regards the specifications for users graphical or programmatic interfaces: in particular the integration of additional collaborative tools or the

management of different levels or parametrization according to users.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme 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: 10.2481/dsj.GRDI-013.

L. Candela, D. Castelli, A. Manzi, and P. Pagano. Pos (isgc2014) 022 realising virtual research environments by hybrid data infrastructures: the d4science experience. In International Symposium on Grids and Clouds (ISGC), volume 23, 2014.

L. Candela, D. Castelli, G. Coro, L. Lelii, F. Mangiacrapa, V. Marioli, and P. Pagano. An infrastructure-oriented approach for supporting biodiversity research. Ecological Informatics, 26, Part 2:162 – 172, 2015. ISSN 1574-9541. doi: http://dx.doi.org/10.1016/j.ecoinf.2014.07.006. URL http://www.sciencedirect.com/science/article/pii/S1574954114001022. Information and Decision Support Systems for Agriculture and Environment.

G. Coro, L. Candela, P. Pagano, A. Italiano, and L. Liccardo. Parallelizing the execution of native data mining algorithms for computational biology. Concurrency and Computation: Practice and Experience, 27(17):4630–4644, 2015. ISSN 1532-0634. doi: 10.1002/cpe.3435. URL http://dx.doi.org/10.1002/cpe.3435.