ONLINE COLLABORATIVE ENVIRONMENT TO RUN THE EASTERN BLUEFIN TUNA STOCK ASSESSMENT WORKFLOW

Taha Imzilen, Sylvain Bonhommeau, Tristan Rouyer, Laurence T. Kell, Julien Barde[1]

SUMMARY

In this note, we describe an approach which makes possible to execute online the set of codes used for eastern bluefin tuna (BFT-E) stock assess-ment at ICCAT. 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 new approach has a lot of scientific benefits, but very few participants would be able to reproduce it without specific skills. Since november 2015, the H2020 BlueBridge project is funding an engineer to set up a WebSite (or VRE for Virtual Research Environment) where users will be able to 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: a workspace to share documents or data, Web pages or RStudio server to process data online and an automated report service to dynamically generate documents packaging these results. Such a collaborative environment enables 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 is expected to give more transparency and collaboration within working groups.

RÉSUMÉ

Dans cette note, nous décrivons une approche qui permet d'exécuter en ligne le jeu de codes utilisés pour l'évaluation du stock de thon rouge de l'Est (BFT-E) à l'ICCAT. En 2014, le groupe d'espèces sur le thon rouge de l'ICCAT a pu exécuter des milliers de scénarios de modèles en utilisant la parallélisation des codes R et Fortran sur un superordinateur. Cette nouvelle approche présente beaucoup d'avantages scientifiques, mais très peu de participants seraient en mesure de la reproduire sans compétences spécifiques. Depuis novembre 2015, le projet BlueBridge project H2020 finance un ingénieur pour créer un site Web (ou VRE pour « Virtual Research Environment »), où les utilisateurs pourront paramétrer et exécuter diverses étapes du flux de travail d'évaluation du stock de thon rouge de l'Est. En reconditionnant les codes fournis par le groupe d'espèces sur le thon rouge de l'ICCAT, les mêmes codes sont maintenant exécutables en ligne. Ils peuvent être paramétrés, exécutés et édités par n'importe qui à partir d'une simple page Web et les sorties des données sont livrées dans un format de données standard. À ce stade, ce VRE est accompagné de divers services Web collaboratifs : un espace de travail pour partager des documents ou des données, des pages Web ou du serveur RStudio pour traiter des données en ligne et un service de rapport automatisé pour générer dynamiquement des documents qui présentent ces résultats. Un tel environnement collaboratif permet de stocker et d'accéder à l'ensemble des données et des codes sources pour reproduire les résultats passés ou pour essayer de nouveaux paramétrages du modèle avec des outils habituels ou de simples formulaires Web. Une telle approche devrait donner plus de transparence et de collaboration au sein des groupes de travail.

[1] IRD - UMR MARBEC 248, Av. Jean Monnet, 34200 Sète, France; taha.imzilen@ird.fr, julien.barde@ird.fr; Phone: +33 499 57 32 32 Fax: +33 499 57 32 15.
[y]IFREMER - Le Port, La Réunion, France; sylvain.bonhommeau@ifremer.fr; Phone: +33 499 57 32 66 Fax: +33 499 57 32 95.
[z]IFREMER - UMR MARBEC 248, Av. Jean Monnet, 34200 Sète, France; syl-vain.bonhommeau@ifremer.fr; Phone: +33 499 57 32 66 Fax: +33 499 57 32 95.
[x]ICCAT Secretariat, C/ Corazón de María 8, 28002 Madrid, Spain; Laurie.Kell@iccat.int; Phone: +34 914 165 600 Fax: +34 914 152 612

RESUMEN

Se describe un enfoque que hace posible ejecutar online el conjunto de códigos utilizados para la evaluación del stock de atún rojo del este (BFTE) de ICCAT. En 2014, el Grupo de especies de atún rojo de ICCAT ha podido ejecutar miles de ensayos de modelos utilizando la paralelización de códigos R y Fortran en un superordenador. Este nuevo enfoque tiene muchos beneficios científicos, pero muy pocos participantes podrían reproducirlo sin unas capacidades específicas. Desde noviembre de 2015, el H2020 BlueBridge project está financiando un ingeniero para crear un sitio web (o VRE de Entorno de investigación virtual) en el que los usuarios puedan parametrizar y ejecutar varios pasos del trabajo de la evaluación de atún rojo del este. Reempaquetando los códigos facilitados por el grupo de especies de atún rojo de ICCAT, los mismos códigos son ahora ejecutables online. Pueden ser parametrizados, ejecutados y editados por cualquiera a partir de una simple página web y los resultados se presentan en un formato de datos estándar. En esta etapa, este VRE viene con varios servicios web de colaboración: un espacio de trabajo para compartir documentos o datos, páginas web o servidor RStudio para procesar los datos online y un servicio de comunicación automatizado para generar documentos de forma dinámica empaquetando estos resultados. Dicho entorno de colaboración permite almacenar y acceder a todo el conjunto de datos y códigos fuente para replicar resultados pasados o intentar nuevas parametrizaciones del modelo con herramientas usuales o formularios web simples. Se prevé que dicho enfoque proporcione más transparencia y colaboración entre los grupos de trabajo.

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. By doing so, 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. A web site was thus needed to access this work and related services. BlueBridge H2020 project provides collaborative environments (VRE for Virtual Research Environments Candela *et al.* [2014]) to enable any kind of users to access these services and run processes of ICCAT BFT-E (current Web Site, so called VRE in such infrastructures Candela *et al.* [2013]). In this note, our main goal is to showcase how this VRE can be used to process 2016 data. We will first describe how the ICCAT work ow 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 don't 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 servers. 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) and deploying 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,

- knitr (+ 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 in 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 programming languages, Web Browsers, GIS),
  - A grid of servers to store, access and process data with above codes

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:

4 main Steps:

- STEP 1 (heavy): Analysis (retros) of the BFT-E ICCAT datasets,
- STEP 2 (light) : Visualization of data analysis (retros),
- STEP 3 (heavy) : Projections which is the most demanding step in terms of machine resources,
- STEP 4 (light) : 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.

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

4.1 A standard data format to store Stock Assessment data

In the first version of the work, data generated by the workflow were saved as R data object making them hardly reusable outside the R environment. However, data outputs should be independent from underlying programming languages (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 mots of existing programming languages.

Data can now be shared online, remotely accessed to visualize and compare outputs. 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.

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..) and clients (Web Browsers, Qgis).

This is made possible by implementing the Web Processing Service standard as recommended by 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 using 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.

5. Online collaborative environment for ICCAT BFT-E

The BlueBridge project and underlying IF infrastructure provide a set of useful services to access and parametrize such a workflow. Each collaborative environment comes with a list of members who can share documents and messages within this public or private environment. Data and codes can be accessed and shared by all members.

In addition we tried to add some relevant services for the community.

5.1 RStudio online to run BFT-E Stock Assessment workflow

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

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

5.2 Run the BFT-E 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 used a Web Processing Service which makes possible to run these codes from a Web Page directly in the VRE as shown in **Figure 3**.

A similar Web form can generated in any Web page (out of the VRE) as shown in **Figure 4** (check online demonstration).

Users can focus on parametrization and don't 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.

6. Results and Discussion

This report showcase 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++). 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 dis-cuss 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 ma-chine resources.

At some point a similar approach could be replicated with additional models executed on other infrastructures / clouds (Amazon, google..) but the main question regards the specifications for users graphical or programmatic inter-faces: 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 re-search and innovation programme under the BlueBRIDGE project (Grant agreement No. 675680).

References

Candela, L., Castelli, D., and Pagano, P. Virtual research environments: An overview and a research agenda. Data Science Journal, 12:GRDI75{GRDI81, 2013. doi: 10.2481/dsj.GRDI-013.

Candela, L., Castelli, D., Manzi, A., and Pagano, P. 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.

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.

Coro, G., Candela, L., Pagano, P., Italiano, A., and Liccardo, L. 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.

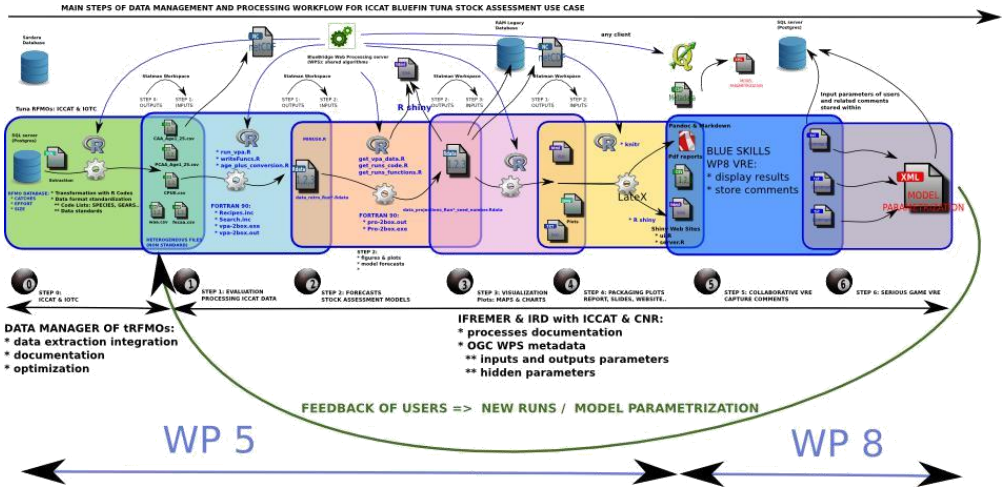

Figure 1. Workflow.

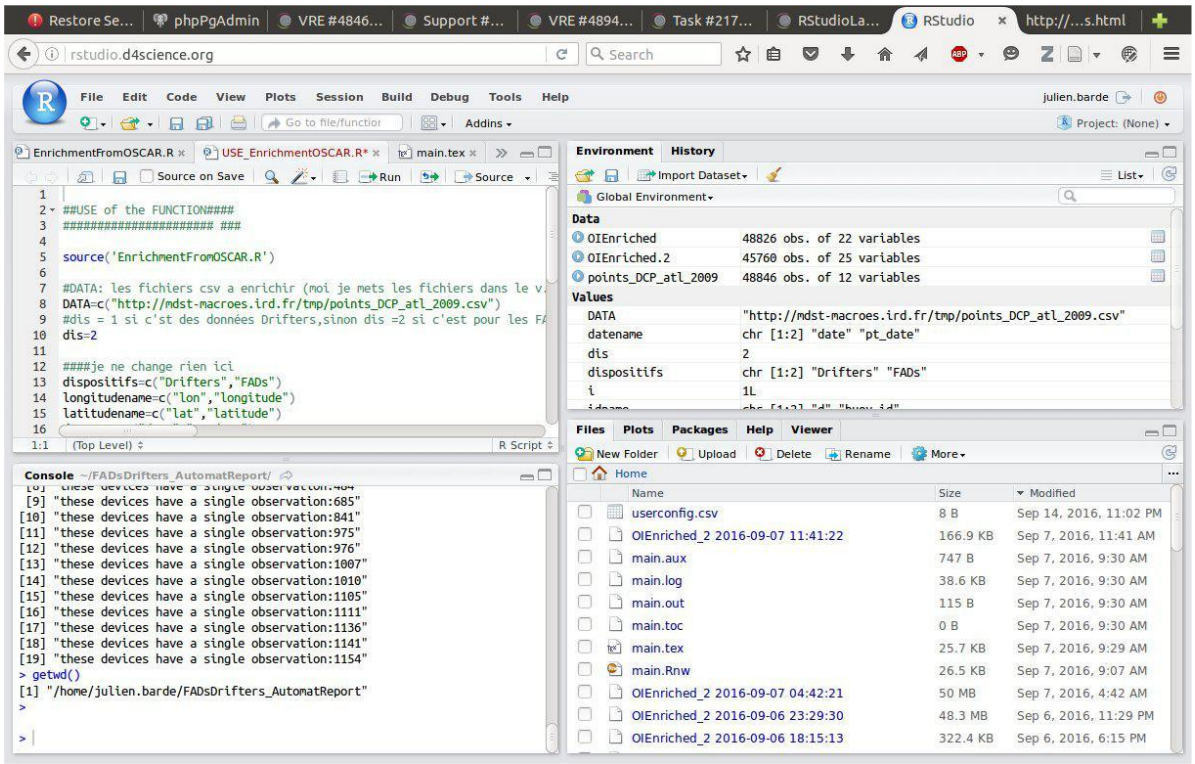

Figure 2. Users can directly edit and run the codes with Rstudio online.

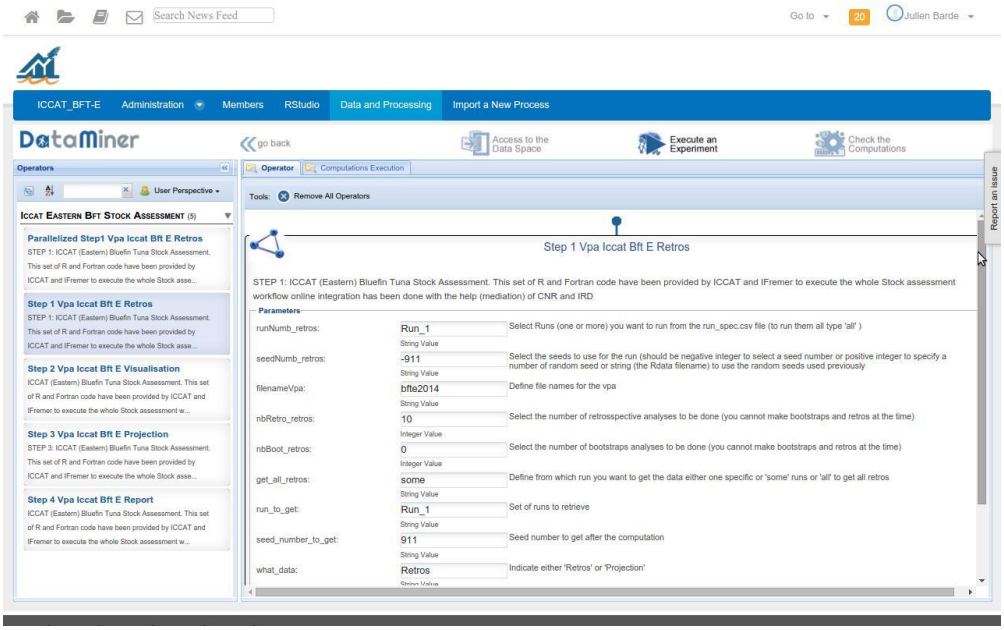

Figure 3. Step 1 can be executed from a Web page of the VRE.

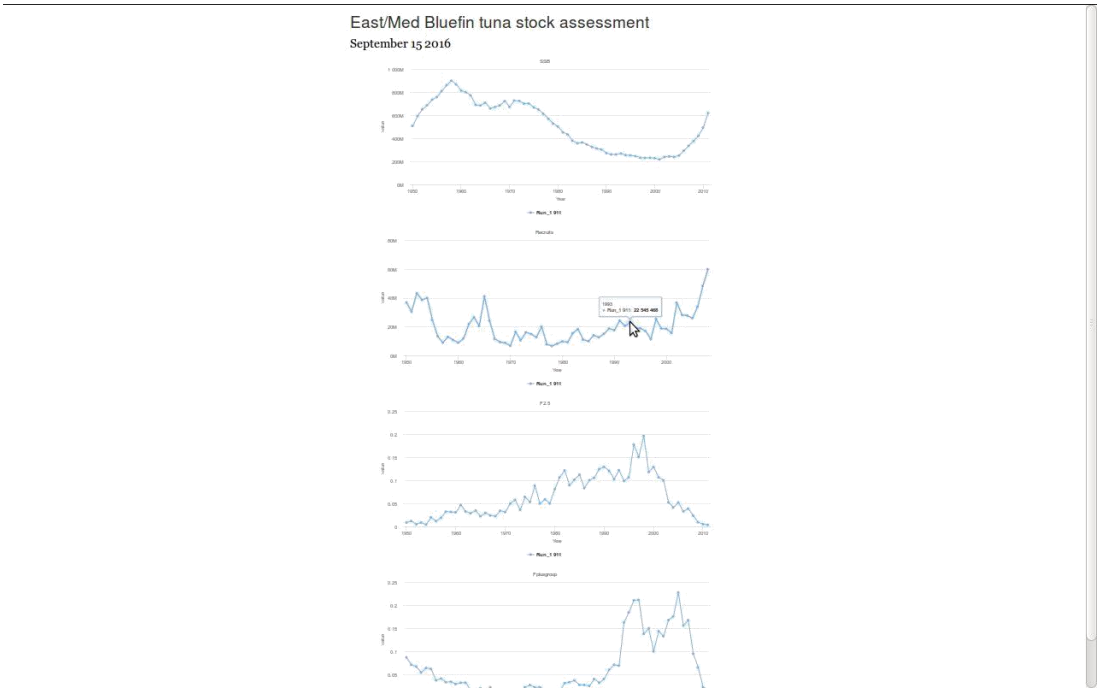

Figure 4. Visualization of retros at step 2 (results from step 1)