Fisheries data management at GFCM

Review and recommendations
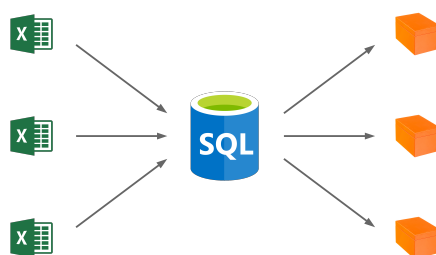
Arni Magnusson

Rome, 2021

Contents

1 Introduction

The General Fisheries Commission for the Mediterranean (GFCM) data center oversees a variety of fisheries data. The activities of the GFCM data center include data collection, quality control, database management, data processing, and dissemination of data (Goodman et al. 2014).

Data management takes place in a continuously changing environment. With each year, new data streams are added and technologies move forward. At the same time, user expectations change with time, as reflected in scientific and professional standards. Some of the core objectives in current data management are laid out in the FAIR data principles, that data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).

The scientific demand for open access to data (Molloy 2011) has increased rapidly in fisheries as in other domains of science, along with open access to reproducible analytical workflows, typically R scripts. Together, the data and the analytical workflows form the basis of scientific advice. In fisheries, these modern standards are encapsulated in the question: is the management of this stock based on open and transparent science? And if not, which criteria are still missing?

The main objectives of this review are to describe the management of GFCM fisheries data, identify possible improvements, and indicate how these improvements could be implemented. In addition to this review, this project also involved the creation of a new data framework called Stock Assessment Results (STAR) that organizes the GFCM stock assessment results and the formulation of the scientific advice.

The first part of this review has a wide scope, aiming to encompass fisheries data management at GFCM in general, while the second part focuses on STAR, describing the current functionality and proposing further development.

2 Overview of GFCM data management

2.1 Databases

The following GFCM fisheries databases can be browsed and queried on the GFCM RStudio Server, using the RODBC package, along with the MSSQL package (Magnusson 2021b):

- **DCRF** — data collection reference framework
- **StockAssessment** — stock assessment results
- **STATLANT37A** — catches in FAO area 37
- **BlackSeaSciDB** — Black Sea fisheries research
- **AIS** — vessel location
- **Bycatch** – bycatch
- **SBHS** – sensitive benthic habitats and species

This review covers the **DCRF** and **StockAssessment** databases in some detail, but other databases listed above are only described briefly here.

The **STATLANT37A** database contains data on catches in FAO area 37. It originates from the FAO FishStat database, but is now maintained by GFCM separately from FishStat.

The **BlackSeaSciDB** is the product of the BlackSea4Fish project, organizing a thorough overview of past and current fisheries research in the Black Sea. The central focus is on publications, the authors of those publications, and the data and results presented in those publications.

The **AIS** database contains vessel location data at the minute resolution, covering the entire GFCM area of application. Evaluation trials have focused on the impact of the COVID-19 pandemic on fisheries in the region. The amount of available data is very large, so the approach has been to import selected spatial datasets to address specific fisheries-related questions.

The **Bycatch** database contains data on incidental catch of vulnerable species. The data are composed by four modules: on-board observations, self-sampling, stranding data, and questionnaires. The participating countries are, so far, Cyprus, Morocco, Tunisia, and Turkey.

The **SBHS** database contains data on sensitive benthic habitats and species, submitted on a voluntary basis by a number of participating organizations. This database was formerly called VME, but was renamed to avoid an overlap with other VME data collections of FAO.

2.2 DCRF

The GFCM Data Collection Reference Framework (DCRF, GFCM 2018) organizes the collection and transmission of fisheries-related data in the Mediterranean and Black Sea. Data are submitted by member countries via Excel templates and then imported into the DCRF database. For further information, see:

http://www.fao.org/gfcm/data/dcrf/en/

2.2.1 Human resources

This framework is fully functional and provides a reliable pipeline to receive fisheries data from GFCM member countries. However, the procedures are labor-intensive, both for the member countries and GFCM staff. At times of high work-loads, it has been a challenge for GFCM to allocate enough resources in terms of staff and consultants to carry out the DCRF work.

This reviewer analyzed DCRF to identify whether the framework could be further developed to reduce the amount of annual work required, but concluded that the system is well designed and the workflows are efficient.

Recommendation 2.2.1:

⇒ It is recommended that GFCM continues to ensure the required resources in terms of staff and consultants to oversee the DCRF processes.

2.2.2 Quality control

An R package qc4dcrf is used for DCRF quality control. This package is developed and maintained in a private GitHub repository on https://github.com/gfcm. It currently includes sensitive data as example datasets, so the repository is private, providing read access only to the GFCM data center and DCRF development team.

Using this package, the quality control process produces a dashboard indicating in detail the number of data submissions that do not meet specific quality criteria. The dashboard lists overall percentages and also allows following up with country representatives to address and improve certain data submissions.

Recommendation 2.2.2:

⇒ It is recommended that all sensitive data be removed from the qc4dcrf package, so the GitHub repository can be made public. This will conveniently separate the general tool from the specific data. The repository will showcase to the world the high quality of GFCM data management, allow publications in reports and journals describing the system, and facilitate reuse of quality control methods and approaches for GFCM data streams other than DCRF.

2.2.3 DCRF database

The DCRF database is organized in tables that correspond to 'Tasks' that stem from GFCM Recommendations.

The Tasks and corresponding Recommendations are listed in the DCRF manual, available at http://www.fao.org/gfcm/data/dcrf/en/. For example, Task I is 'Global Figures of National Fisheries' (landings, number of vessels, total capacity, total engine power by country) and Task VII.2 is 'Length Data' (size distribution of landings, by species and area).

Table 1. 'Rec' tables in the DCRF database as of July 2021. Number of rows, number of columns, and storage size in megabytes.

Table	Rows	Cols	MB
Rec_I	122	28	0.2
Rec_II1	0	25	-
Rec_II2	0	26	-
Rec_III	0	32	-
Rec_IV1	0	58	-
Rec_IV2	0	63	-
Rec_IV3	0	29	-
Rec_V1	0	26	-
Rec_V2	0	21	-
Rec_V3	0	27	-
Rec_VI1	0	26	-
Rec_VI2	0	33	-
Rec_VI3	0	26	-
Rec_VI4	0	28	-
Rec_VII2	0	29	-
Rec_VII31	0	23	-
Rec_VII32	0	31	-
Rec_VII41	0	23	-
Rec_VII42	0	30	-
Rec_VII5	0	28	-
Rec_VII6	0	28	-

The final tables in the DCRF database are fully designed with the appropriate number of columns and data types, but are effectively empty, containing zero rows, except Rec_I which has 122 rows (Table 1).

While the DCRF quality control process is still underway, the DCRF data currently reside in tables whose names start with Tmp (Table 2). Until the quality control process is complete, the years of data have accumulated in the Tmp tables. For example, Tmp_VII2 currently has data from the reference years 2015, 2016, 2017, 2018, and 2019.

Table 2. 'Tmp' tables in the DCRF database as of July 2021. Number of rows, number of columns, and storage size in megabytes.

Table	Rows	Cols	MB
Tmp_II1	1643	23	0.3
Tmp_II2	46008	25	9.6
Tmp_III	3446	33	1.3
Tmp_IV1	341149	59	178.3
Tmp_IV2	37812	64	22.1
Tmp_IV3	166	30	0.1
Tmp_V1	1502	27	0.4
Tmp_V2	1571	22	0.3
Tmp_V3	26713	28	7.2
Tmp_VI1	1471	27	0.4
Tmp_VI2	1007	34	0.4
Tmp_VI3	32215	27	6.9
Tmp_VI4	713	29	0.3
Tmp_VII2	100235	30	27.4
Tmp_VII31	1622	24	0.4
Tmp_VII32	115318	32	37.9
Tmp_VII41	57	24	0.1
Tmp_VII42	45	31	0.1
Tmp_VII5	270	29	0.2
Tmp_VII6	212	29	0.1

The Data Center has plans to move data from the `Tmp` tables to the `Rec` tables after the quality control process is complete, so `Tmp_VII2` becomes `Rec_VII2`, etc.

Recommendation 2.2.3:

⇒ It is recommended that the completion of the DCRF quality control process is prioritized, so data can be moved from the `Tmp` tables to fully vetted `Rec` tables, for the years that have been quality controlled. This will make it clear which parts of the DCRF database are ready to be queried and used, marking an important milestone in the data managament of the DCRF framework.

2.2.4 DCRF read access

The DCRF data include landings, length frequencies, and maturity data that underly the stock assessment input data. In practice, though, there is not a direct link between the DCRF data and the stock assessment input data.

Currently, only the GFCM data center and DCRF development team can query data from the DCRF database. It would greatly increase the potential use of the DCRF database if other users, including stock assessors, are granted read access to the database, e.g. via the GFCM RStudio Server.

Recommendation 2.2.4:

⇒ It is recommended that stock assessors and other users have read access to the DCRF database. This will allow them to write R scripts that document how the stock assessment input data are prepared, specifically the catch at age, weight at age, and maturity at age. Such scripts would greatly strengthen the connection between DCRF and stock assessment input data.

2.3 Surveys

2.3.1 Databases, data management, and data access

The most extensive fisheries surveys in the GFCM areas are the international MEDITS (bottom trawl) and MEDIAS (acoustic) surveys. A number of other fisheries surveys are conducted regularly in the Mediterranean and Black Sea.

Currently, the GFCM data center does not host any fisheries survey data. Some of the survey data are already well managed by entities other than GFCM, with high-performing databases, quality control processes, and providing access to fisheries scientists to query the database directly. However, there are other surveys whose data management may not be at that level.

Recommendation 2.3.1:

⇒ It is recommended that GFCM considers hosting data from selected fisheries surveys in well maintained databases and providing data services, where this can be beneficial for fisheries research and management of the fish stocks within the GFCM area.

2.4 Stock assessments

2.4.1 Input

Stock assessment input data and scripts are organized in the SharePoint area for each working group, in a folder called `Input_data_and_scripts`. These files are in a free format, specific to the software used in each stock assessment, often bundled together in a zip file.

This arrangement provides a definite reference of the exact model input and model settings used in a given stock assessment. Furthermore, the input data and scripts are a significant step towards supporting reproducible analysis for GFCM stock assessments. A level of transparency where data and scripts are made available and can be rerun on anyone's computer are considered important and high-quality criteria of modern scientific work.

Recommendation 2.4.1:

> ⇒ It is recommended that GFCM encourages and supports reproducible analytical workflows, where stock assessors provide and archive the final set of input data and scripts, which can be rerun on anyone's computer to produce the same numerical result. This would not be a hard requirement or necessarily follow a fixed format, but should be an encouraged practice, especially in the case when stock assessors already possess the technical expertise.

2.4.2 SAF

Stock assessments are documented in Stock Assessment Forms (SAFs). These are written as a Word document following a standard structure. The main sections are:

1. Basic identification data (species, areas, countries)
2. Stock identification and biological information (growth, maturity)
3. Fisheries information (fleet, trends, management, reference points)
4. Fisheries independent information (surveys)
5. Ecological information (impact on non-target species, environmental indices)
6. Stock assessment (data, methods, results, diagnostics)
7. Stock predictions (forecast)
8. Draft scientific advice (stock status relative reference points)

The SAFs are an essential and well-functioning part of the GFCM stock assessment procedures, a cohesive document describing the underlying data and methods, modelling decisions, ecological information, forecast, and other details related to the stock assessment.

With the stock assessment input data files and scripts (Section 2.4.1) and STAR metadata and time series (Section 2.4.3 and Section 4) acting as technical and machine-readable annexes, the SAFs are the definitive reference material to read about each stock assessment.

Recommendation 2.4.2:

> ⇒ It is recommended that the three stock assessment components (Input, SAF, and STAR) are somehow linked to each other, so that one can navigate between all three when examining a stock assessment.

2.4.3 STAR

Stock assessment results are organized in a new framework called STAR, launched in January 2021, as a part of this project. STAR is described in detail in Section 4 of this report, along with recommendations for further development.

2.5   Scientific advice

The GFCM Secretariat organizes the scientific advice for fisheries management in a Table of Advice, containing key information for every stock where GFCM provides advice. It is crucial that the Table of Advice contains correct information and calculated quantities describing the status of each stock must be accurately derived from the stock assessment results.

The Table of Advice has one row for each stock and the following columns:

1. Geographical subareas
2. Species
3. Stock assessment method
4. Current levels of fishing mortality and biomass
5. Reference points
6. Quantitative status, relative to reference points
7. Stock status text
8. Advice text
9. Working Group comments

Generally, the information for the Table of Advice is completed near the end of a stock assessment working group meeting, in a process coordinated by the chairs and the secretariat.

A primary objective of the STAR framework is to streamline and support the process of producing the Table of Advice. Specific recommendations related to this are listed in Section 4, along with other recommendations on STAR development.

2.6   GitHub and DevOps

2.6.1   Open projects and data dissemination

The GFCM GitHub site (https://github.com/gfcm) is used to organize and showcase the development of GFCM scientific and data management projects. As of July 2021, the site hosts the following projects:

- gfcmBeamer — LaTeX package to create slides in GFCM Secretariat format
- gfcmSTAR — R package to work with STAR
- MSSQL — R package to work with SQL Server databases via RODBC
- qc4dcrf (private) — R package to perform quality checks for DCRF
- star — STAR homepage and Excel template
- wgm — Code repository for the Working Group on Methodology

In addition to organizing projects, such as the ones listed above, GitHub can also be used as a convenient and efficent platform to disseminate scientific data. GitHub is a widely used environment by fisheries scientists and is seen as a standard for open access and scientific collaboration.

Datasets on GitHub can be browsed and downloaded from a normal web browser and also accessed directly from R scripts and other software for data analysis.

Recommendation 2.6.1:

⇒ It is recommended that GFCM considers using the GitHub site to disseminate selected data products and make available for download. The GitHub site can be used to augment and extend the GFCM data portal (`http://www.fao.org/gfcm/data/en/`) using cross-linked references.

The GitHub site complements the existing DevOps site that GFCM uses. In general, GitHub is used for public repositories to showcase, share, and collaborate widely, while DevOps is used for private repositories to organize sensitive data and files. Both GitHub and DevOps are at a more technical and analytical level than SharePoint, which is well suited for Word, Excel, PowerPoint, and PDF files.

2.6.2 DevOps private repositories

The GFCM Visual Studio Server, `https://gfcm.visualstudio.com`, requires password-protected login to access the following DevOps repositories:

- AIS-Analysis — Analysis of spatial fishing density

- Data-representation-fisheries — Benchmark scripts and input files

- FLR_gfcm — FLR course material

- GADGET — Gadget model runs

- GFCM-DCRF — DCRF quality control dashboard

- MSE_Adriatic — MSE analyses for Adriatic anchovies and sardines

- SAF_InputData — Stock assessment input data files

- SGSABS — Black Sea stock assessments

- Socioeconomics — Socioeconomic analyses

- SoMFi — Analyses for the production of the SoMFi report

- StockAssessment_Basic — Stock assessment course material

- StockAssessmentResults — STAR development and uploaded templates

- WGSAD — Demersal stock assessments

- WGSASP — Pelagic stock assessments

2.7 Questionnaire survey

2.7.1 Improving access to data and awareness of existing tools

The GFCM data center provides numerous and diverse data products to users in the GFCM community, the secretariat, scientists, and other users. An initiative was made by the secretariat in spring 2021 to conduct a questionnaire survey to identify data needs, possible bottlenecks, and get user feedback.

Key questions on such a questionnaire could include questions such as:

- What are the main GFCM data you need to access in your work?

- How do you currently access these data?

- Can you identify areas in GFCM data management that could be improved, to make certain data more convenient to work with?

- Any other comments or suggestions related to GFCM data management?

The questionnaire could be circulated within the secretariat and other key users in the GFCM community. The first three questions are specific and should be easy to answer, while the fourth question is presented as optional and more open-ended.

By keeping the questionnaire simple, all recipients should be able to complete the survey with minimal time commitment, while allowing in the fourth question an opportunity to raise discussion points. Once the answers have been submitted, the results would be analyzed by key personnel in the secretariat and a follow-up round of discussion could be organized online.

Recommendation 2.7.1:

⇒ It is recommended that GFCM conducts a questionnaire survey along the lines of what is described above. The main objectives would be to identify data needs, possible bottlenecks, and result in improving access to data and raising awareness of existing tools and approaches to work with GFCM data products.

3 Recommendations for GFCM data management

3.1 List of general recommendations

The following list is a condensed summary of all recommendations from Section 2. See the cross-linked section numbers for context and further details.

DCRF

(2.2.1) Ensure the required resources in terms of DCRF staff and consultants.

(2.2.2) Make the `qc4dcrf` GitHub repository public.

(2.2.3) Move DCRF data from the `Tmp` tables to fully vetted `Rec` tables.

(2.2.4) Give stock assessors and other users read access to the DCRF database.

Surveys

(2.3.1) Host data from selected fisheries surveys in GFCM databases.

Stock assessments

(2.4.1) Encourage and support reproducible analytical workflows.

(2.4.2) Cross-link the stock assessment Input, SAF, and STAR components.

GitHub

(2.6.1) Use GitHub to make selected data products available for download.

4 STAR

4.1 Executive summary

Stock Assessment Results (STAR) is a new framework developed by the secretariat to organize GFCM stock assessment results. It streamlines and automates the information flow from stock assessments to scientific advice, strengthening quality assurance and data dissemination. The STAR homepage is:

https://github.com/gfcm/star

The central part of the framework is the STAR database, scheduled to launch in 2021, which will serve as a foundation to provide a variety of quality-controlled data products related to the management of fisheries in the Mediterranean and Black Sea.

The STAR data, describing the quantitative status of each stock, will be made available on a dedicated website when the assessment and scientific advice of that stock has been validated by the SAC. The published STAR datasets serve as a modern technical annex to the GFCM stock assessment and advisory process, as well as an open data portal to disseminate scientific results.

4.1.1 Benefits

- Reliably find stock assessment results, reference points, and advice in one place, a quality-controlled database.

- Reduce the workload on GFCM fisheries officers of looking up metadata in various places and following up with stock assessors to clarify what the correct values are.

- Trace exactly how the scientific advice for each stock was formulated, with formulas directly linked to stock assessment results and reference points.

- Streamline the process of producing the Table of Advice and increase the reliability of the process.

- Use database queries to generate summaries, e.g. the status of all stocks in the Mediterranean and Black Sea, in a given year or across many years.

- Use database as back-end to present stock assessment results online, both tables and plots.

- Use time series of biomass and fishing mortalities over time to show current and past stock status, recent and long-term trends, and as a basis for the choice of reference points.

- Share time series of biomass and fishing mortalities with scientific community, FAO State of World Fisheries and Aquaculture (SOFIA), and inclusion in other analyses.

4.2 Background and design

4.2.1 Background

The STAR project evolved as a part of the broader project of reviewing fisheries data management at GFCM. About one quarter into the review project, it was identified and mutually agreed that the highest priority recommendation would be to create a framework to organize the GFCM stock assessment results and the formulation of the scientific advice.

The stock assessment and advisory process is of critical importance to GFCM as a regional fisheries management organization. Before STAR, a substantial part of the secretariat work was to search for key metadata (current stock status, reference points, etc.) in various places. Stock assessors did not always provide the metadata in the agreed manner, so values needed to be harvested from a large number of lengthy documents. When stock assessors provided inconsistent metadata values in different places, the secretariat would contact stock assessors to determine the correct values.

4.2.2 Primary objectives

The first objective of STAR was to make the above stock assessment and advisory process more streamlined and reliable, by having stock assessors submit the assessment results and reference points in a dedicated input template. This template would be the only way to submit values to a central STAR database, which would then serve as the definite source of information about stock assessment results.

The second objective of STAR was to not only store the *current* stock status, but the entire *time series* of recruitment, biomass, catches, and fishing mortalities. These time series data are valuable for fisheries management, showing current and past stock status, recent and long-term trends, and as a basis for the choice of reference points. Moreover, GFCM had received requests by data analysts and influential research groups asking for such time series, but had to respond that these were not available.

The third objective of STAR was to streamline and support the process of producing the Table of Advice (Section 2.5). STAR would not only store stock assessment results, but also draft the scientific advice using dynamic calculations and automatically generated text within the Excel template. In this way, STAR would ensure a direct and traceable connection between the stock assessment results and the scientific advice.

4.2.3 Initial development

The STAR design and development was initiated in December 2020, by a development team consisting of Arni Magnusson, Roberto Emma, Miguel Bernal, Betulla Morello, and Federico De Rossi. Collaboration on the design and development moved rapidly, in time to launch the STAR Excel template at the joint plenary session of the assessment working groups meeting in January 2021. The timing was of importance, as these meetings were to produce the majority of GFCM stock assessments that year.

4.2.4 Metadata

In GFCM stock assessments, the term 'metadata' refers to a variety of information related to the assessment, essentially everything except the time series. For a given stock assessment, the metadata include:

- Species name, Geographical subareas, Countries
- Assessment type, Reference year, Reporting year
- Validation status, Year benchmarked, Assessment method
- Expert group, Contact person
- Reference point values
- Current biomass and fishing mortality relative to reference points

Before STAR, these metadata were submitted by stock assessors by assigning SharePoint properties to the Word documents containing the Stock Assessment Forms. However, this process was not very reliable, since stock assessors often omitted supplying these properties on SharePoint.

4.2.5 Design

In the context of STAR, a framework consists of a system and a workflow. The system is a bundle of computer software modules, while the workflow is in the human domain: an agreed way to work, rules and recommendations, coordination and capacity building for the various users.

The STAR diagram (Figure 1) captures the design of the framework, starting with submitted results coming from many stock assessments that are imported into a central relational database. The import procedure involves quality control, catching and handling possible errors that stock assessors may have made when filling in the template.

This means that products that are based on data stored in the database are quality-conrolled. Examples of such products could include automated construction of the Table of Advice, dissemination of scientific time series data, online browser of stock assessments, summary tables and plots, and so on.
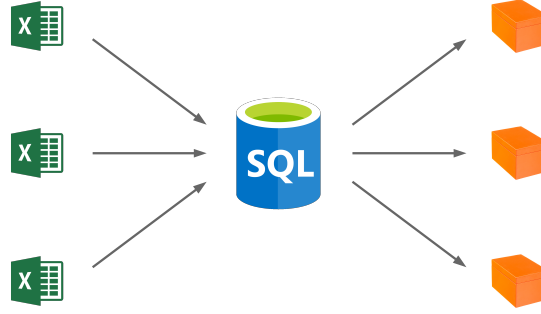
Figure 1. STAR design diagram, showing submitted results, central database, and quality-controlled products.

The main STAR system components are:

- **Excel template**, for organizing and submitting stock assessment results, as well as formulating the scientific advice

- **R package**, providing tools for various STAR-related tasks, e.g. importing, quality checking, and reformatting data from the submitted Excel templates

- **Database**, a central relational database to store the data and allow the creation of STAR-related data products

Overall, STAR strengthens and upgrades the processes that existed to organize the stock assessment results and formulation of scientific advice (Table 3).

Table 3. Data management of stock assessment results before and after STAR.

Stock assessment results	Before	After
Metadata values	GFCM secretariat need to search in SAFs	Stock assessor submits in STAR template
Data submission	SharePoint properties	Excel template
Database design	SharePoint properties	SQL database
Stock assessment & Advice	Not directly connected	Directly connected
Time series	Not collected	Collected

There is a degree of similarity between the DCRF and STAR frameworks, which are both centered on Excel templates, a database, and an R package for quality control. This is not a complete coincidence, since a design aim in STAR is to use platforms and solutions that people are already familiar with, both users and administrators.

There is also a degree of similarity between the STAR framework and the ICES Stock Assessment Graphs (SAG, ICES 2021). The initial design process of the STAR project included a study of SAG, evaluating the similarities and differences between the needs of GFCM and ICES when organizing stock assessment results. The main conclusion was to design STAR to not only store stock assessment results, but to also draft the scientific advice using dynamic calculations and automatically generated text within the Excel template.

4.3 System components

4.3.1 Excel template

The STAR Excel template was the first system component to be launched. Its design and implementation determines to a great extent the requirements and potential of the other system components.

The homepage for STAR and the Excel template is:

> https://github.com/gfcm/star

Excel was chosen as the software platform to implement the data submission template, mainly because most users can be expected to be familiar with entering and pasting values into Excel.

Version history

A condensed version history of the Excel template is provided in Table 4, and the full detailed version is available at https://github.com/gfcm/star/releases. The current state of the Excel template is relatively stable and mature. Upcoming development could include adding a QC sheet for quality control.

Table 4. STAR Excel template version history.

Date	Version	Main changes
30 Apr 2021	2.1.2	Current F and B are not calculated if VPA-Type Model is blank
23 Apr 2021	2.1.1	Added 3-Alpha Species Code
22 Apr 2021	2.1.0	Moved age/length dimensions to Time Series sheet
11 Apr 2021	2.0.0	Added metadata fields: countries, assessment type, reporting year, Automatic calculation of B0.33 and B0.66
18 Jan 2021	1.0.0	First release

The version numbering of the STAR Excel template and the R package **gfcmSTAR** are independent of each other. See details at https://github.com/gfcm/star#user-documentation and https://github.com/gfcm/gfcmSTAR#development.

The development of the Star Excel template has been organized in the DevOps repository **StockAssessmentResults** (Section 2.6.2) inside the 'develop' folder. It contains 37 numbered Excel files, where `star_37.xlsx` is the newest, corresponding to the official release 2.1.2 (see version history below).

It has proven useful to organize the development in small incremental development versions (numbered 1–37) with fewer official releases (1.0.0 to 2.1.2). In this way, the DevOps repository keeps a more detailed track of development versions, along with a `ChangeLog.txt` listing changes between development versions, a complete list of `Formulas.txt` and `TableObjects.txt`.

Metadata sheet

The design of the template is rather plain (Figure 2), with no macros or hidden features. Sheets are used to separate different parts, formulas are used to calculate stock status and the advice, and table objects are used to make it easy to read the data into other software.

Figure 2. STAR Excel template version 2.1.2, showing the Metadata sheet where stock assessors enter the stock assessment results. The gray areas are not edited by the user, but automatically calculate key quantities for the scientific advice.

Colored cells are used to help the stock assessor distinguish between cells that must be filled in (red), cells that must be filled in if defined (green), and cells that are calculated automatically and should not be edited (gray). The table object in the Time Series sheet expands automatically to contain the entire length of the longest time series.

The logical flow behind the advice begins with the time series of fishing mortalities and biomass, which are used to calculate current F and B. The quantitative stock status is then calculated by comparing the current F and B to defined reference point levels. The quantitative stock status is finally used as direct input for the scientific advice.

Advice sheet

As the STAR template is completed with values, the Advice sheet (Figure 3) automatically constructs the scientific advice entries to be used in the official Table of Advice. These entries fit exactly into the general format of the scientific advice (Section 2.5), forming a direct connection between the stock assessment results and the scientific advice.
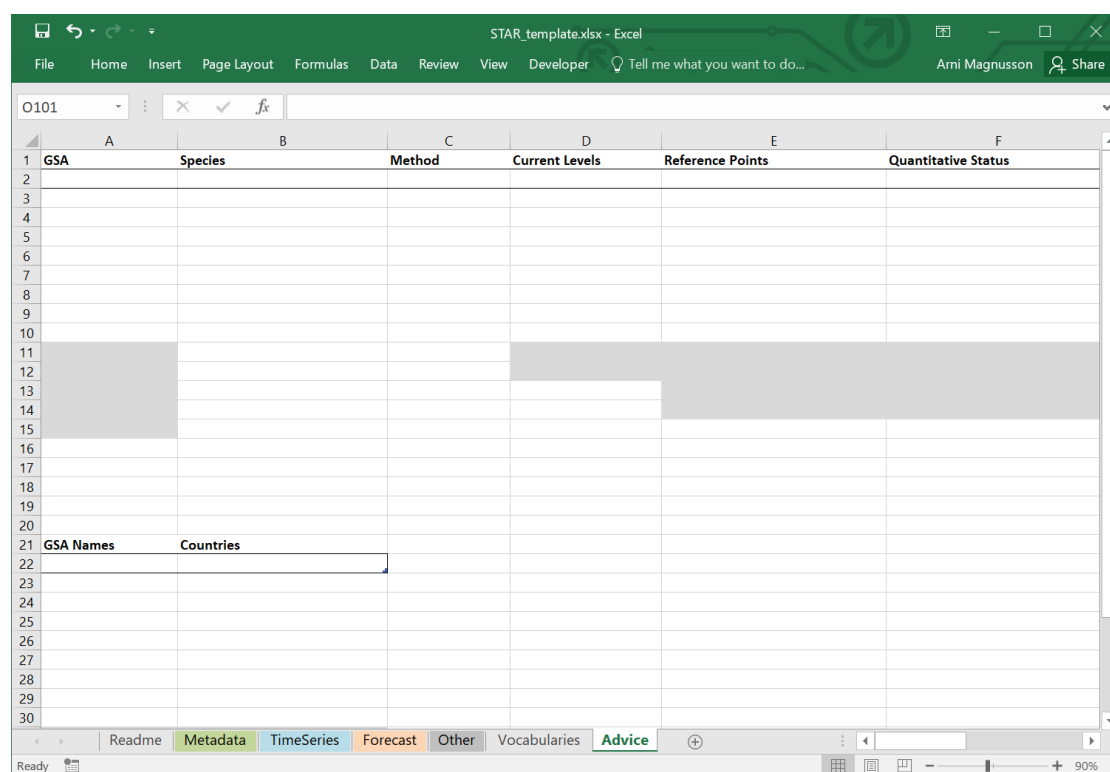

Figure 3. STAR Excel template version 2.1.2, showing the Advice Sheet where the scientific advice is produced automatically and fully formatted. The user does not enter anything in the Advice Sheet, but gray cells show intermediate calculations and white cells inside black borders are ready to be exported directly to the Table of Advice.

4.3.2   R package

The initial need for an R package was to convert STAR templates from version 1.0 to version 2.1. The stock assessors in Jan–Feb 2021 had submitted their results in templates version 1.0, differing substantially from version 2.1 (Table 4) that was designated as the basis of the STAR database.

It was decided that the best way to convert STAR templates from version 1.0 to version 2.1 was to leave the submitted Excel templates unchanged, but to import the stock assessment results into R, convert them into appropriate R structures compatible with version 2.1, and export to standardized comma-separated-values (CSV) files. These CSV files are then imported into the STAR database.

The homepage for the **gfcmSTAR** package (Magnusson 2021a) is:

> https://github.com/gfcm/gfcmSTAR

R was chosen as the software platform to implement the STAR import/export and quality control, mainly because the R language handles these tasks well and also based on the success of the **qc4dcrf** package used for quality control of the DCRF framework.

Version history

A condensed version history of the R package is provided in Table 5, and the full detailed version is available at https://github.com/gfcm/gfcmSTAR/releases. The current state of the R package is relatively stable and mature. Upcoming development could include adding more QC functions.

Table 5. R package version history.

Date	Version	Main changes
8 Nov 2021	2.3.0	Added `write.report()` to write a report to a text file
31 Oct 2021	2.2.1	Improved `report()` and `qc()` to give more detailed information
30 Oct 2021	2.2.0	Added `diff.stars()` and `report()` to manage bulk import
26 Jul 2021	2.1.0	Added support to read STAR templates version 2.1
9 Jun 2021	2.0.1	Improved `qc()` to handle a directory of STAR templates
8 May 2021	2.0.0	Added `export.many.csv()` to export many STAR objects
5 May 2021	1.1.0	Added `import.many.csv()` to import many CSV files
4 May 2021	1.0.1	Improved `read.template()` to include SharePoint properties
30 Apr 2021	1.0.0	First release

The version numbering of the STAR Excel template and the R package **gfcmSTAR** are independent of each other. See details at https://github.com/gfcm/star#user-documentation and https://github.com/gfcm/gfcmSTAR#development.

Dependencies

To import STAR templates, the **gfcmSTAR** package requires the ability to read Excel files. This is achieved by relying on the **XLConnect** package, which in return requires the **rJava** package and the underlying Java platform.

The choice of relying on **XLConnect** and Java was made by the STAR development team after evaluating the strengths and weaknesses of all R packages that can read Excel files. An important criterion was the ability to read Excel table objects that contain the STAR template data.

Table 6. R packages that can read Excel files.

Package	Dependencies	Table objects	Widely used	OS support
openxlsx	3 pkgs	No	Yes	Win, Linux, Mac
readxl	20 pkgs	No	Yes	Win, Linux, Mac
XLConnect	1 pkg + Java	Yes	No	Win, Linux, Mac
xlsx	2 pkgs + Java	No	Yes	Win, Linux, Mac

There are mainly four R packages that provide the ability to read Excel files (Table 6). Of those, XLConnect is the only one that fully supports Excel table objects. The STAR development team acknowledged that the Java dependency can be a hassle, but decided that support for Excel table objects was a top priority.

Functions

The package help page that comes with **gfcmSTAR** lists the following functions in version 2.3.0, categorized by functionality.

Import:

`import.many.templates`	import many templates
`import.many.csv`	import many CSV files
`read.star.csv`	read a pair of CSV files
`read.template`	read STAR template
`read.template.v10`	read STAR template version 1.0
`read.template.v21`	read STAR template version 2.1
`read.properties`	read SharePoint properties
`set.classes`	convert data types
`template.version`	detect STAR template version

Export:

`export.many.csv`	export many STAR objects to CSV files
`write.star.csv`	write STAR object to CSV files

Manipulate strings:

`append.id`	modify existing `Assessment_ID`
`combo`	combine reference year, species, and GSA
`gsa.names`	convert GSA codes to full GSA names

Lookup tables:

`lookup.gsa`	geographical subareas (GSA)
`lookup.species`	species

Quality control:

`diff.stars`	show differences between STAR objects
`identical.stars`	check if STAR objects are identical
`peek`	examine a metadata field
`qc`	run all quality checks
`qc.exists`	file exists
`qc.xlsx`	file extension is 'xlsx'
`qc.star`	file is a STAR template
`qc.vpa`	`VPA_Model` is `Yes` or `No`
`qc.ts.names`	check if time series columns are intact
`qc.ts.numbers`	check if time series are numbers and not strings
`report`	report files successfully imported
`write.report`	write a report to a text file

The package comes with a full help page for each of the above functions.

4.3.3 Database

As of August 2021, the STAR database has been designed but not yet created. It was designed in conjunction with the STAR Excel template, to ensure that all stock assessment results from the input templates will be efficiently organized in a dedicated STAR database (Table 7).

Table 7. History of the STAR database design.

Date	Version	Notes
11 Apr 2021	2	Compatible with STAR Excel template version 2.x.
18 Jan 2021	1	Compatible with STAR Excel template version 1.x.

The STAR database is designed as a standard relational database supporting SQL queries. It consists of two core tables, called *Metadata* and *TimeSeries* (Figure 4). These tables correspond to data found in the input template sheets called Metadata and TimeSeries.
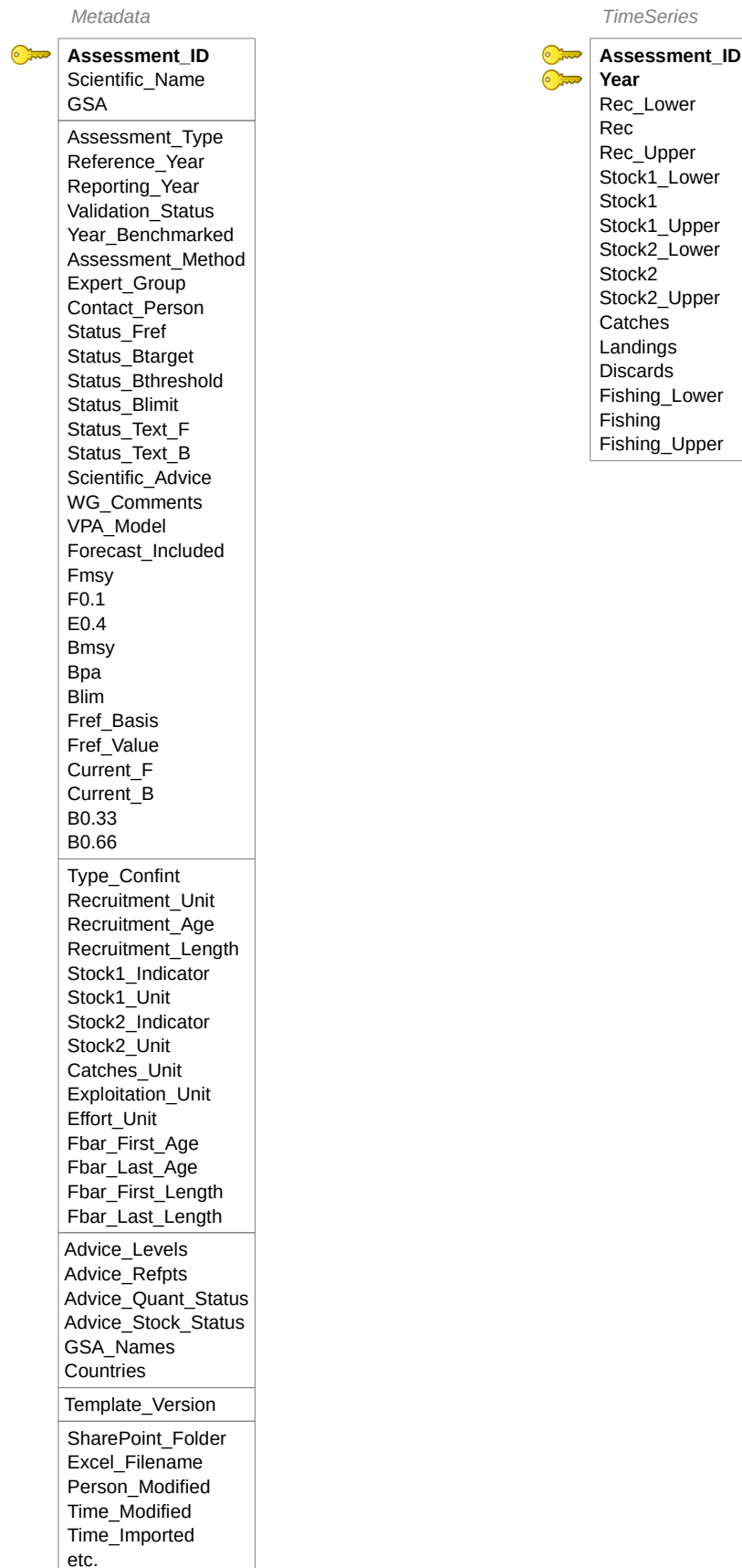
Metadata

🔑 **Assessment_ID**
Scientific_Name
GSA

Assessment_Type
Reference_Year
Reporting_Year
Validation_Status
Year_Benchmarked
Assessment_Method
Expert_Group
Contact_Person
Status_Fref
Status_Btarget
Status_Bthreshold
Status_Blimit
Status_Text_F
Status_Text_B
Scientific_Advice
WG_Comments
VPA_Model
Forecast_Included
Fmsy
F0.1
E0.4
Bmsy
Bpa
Blim
Fref_Basis
Fref_Value
Current_F
Current_B
B0.33
B0.66

Type_Confint
Recruitment_Unit
Recruitment_Age
Recruitment_Length
Stock1_Indicator
Stock1_Unit
Stock2_Indicator
Stock2_Unit
Catches_Unit
Exploitation_Unit
Effort_Unit
Fbar_First_Age
Fbar_Last_Age
Fbar_First_Length
Fbar_Last_Length

Advice_Levels
Advice_Refpts
Advice_Quant_Status
Advice_Stock_Status
GSA_Names
Countries

Template_Version

SharePoint_Folder
Excel_Filename
Person_Modified
Time_Modified
Time_Imported
etc.

TimeSeries

🔑 **Assessment_ID**
🔑 **Year**
Rec_Lower
Rec
Rec_Upper
Stock1_Lower
Stock1
Stock1_Upper
Stock2_Lower
Stock2
Stock2_Upper
Catches
Landings
Discards
Fishing_Lower
Fishing
Fishing_Upper

Figure 4. STAR database design, version 2.

Recommendation 4.3.3:

⇒ It is recommended that the STAR database be created, as a relational database supporting SQL queries, similar to the design laid out in Figure 4.

4.3.4 Existing metadata

The STAR development team has identified that instead of creating a new database from scratch, it is practical to add the two tables (Metadata and TimeSeries) to the existing StockAssessment database (mentioned in Section 2.1). This will be helpful to establish and maintain continuity between existing metadata from pre-STAR stock assessments and upcoming STAR stock assessment results. Currently, the *SAF* table of the StockAssessment database contains metadata from assessments in Reference Years 2003 to 2019.

Recommendation 4.3.4:

⇒ It is recommended that the GFCM data center copies all existing metadata from the old *SAF* table to the new *Metadata* table. This way, all GFCM stock assessments will be merged and can be queried together.

4.3.5 Submitted templates

Stock assessment results that stock assessors have submitted as STAR Excel templates are organized in the DevOps repository StockAssessmentResults (Section 2.6.2) inside the 'uploads' folder. See Section 4.5.2 for a detailed list of directories and number of Excel templates.

Properties

The STAR database design (Figure 4) includes fields describing the SharePoint folder, Person modified, and Time modified. There are several ways to harvest these from the SharePoint system and one way is to download an Excel file containing these properties. This is done by selecting the STAR folder on the Teams channel for each working group meeting and then 'Open in SharePoint' and 'Export to Excel'.

The Excel files containing SharePoint properties are organized in a folder called `uploads/2021/properties` within the StockAssessmentResults DevOps repository.

4.3.6 Documentation

For stock assessors

Stock assessors mainly need to learn how to enter the results of their assessment in the STAR Excel template.

As mentioned in https://github.com/gfcm/star#user-documentation, guidelines for users are provided in four places:

1. Overview of the STAR framework and wider context in the launch event [presentation](#)

2. Usage comments within the [Excel STAR template sheets](#)

3. Sheet highlighting the [logic](#) behind the Excel formulas

4. Changes between versions are listed in the [NEWS.md](#) file

For STAR administrators

The STAR administrators of the GFCM data center will benefit from using the R package **gfcmSTAR** effectively to import stock assessment results from Excel templates, export to CSV, and perform quality control procedures.

In an R session, the following commands bring up the package help page:

```
library(gfcmSTAR)
?gfcmSTAR
```

The package help page contains a list of functions (see 'Functions' in Section 4.3.2) with links to a full help page for each function.

The package also comes with an example script that demonstrates bulk import of Excel templates, export of CSV files, and quality control. This script can be found at:

https://github.com/gfcm/gfcmSTAR/blob/master/inst/scripts/import.R

The script can be adapted to a specific year and working group folder containing STAR templates to semi-automate the import, export, and quality control.

This report

Section 4 of this report adds a significant amount of documentation on the STAR framework, its objectives, design, functionality, and how the system components work together.

4.4 Quality assurance and quality control

4.4.1 QA in the Excel template

Quality assurance (QA) is about STAR design features that prevent user mistakes or make them less likely to occur, especially when stock assessors enter data into the Excel templates.

Dropdown menus

Many data fields in the STAR Excel template are entered by the user by selecting from a list of pre-specified values, using dropdown menus. This prevents users from misspelling species or country names and avoids a variety of possible typos

or misunderstandings when entering data in those fields. It also results in the STAR database having more consistent values that can be queried and filtered.

Dropdown menus in the Metadata sheet are: species name, geographical subareas, countries, assessment type, reference year, reporting year, validation status, year benchmarked, VPA-type model, forecast included, and F or E target reference point.

Dropdown menus in the TimeSeries sheet are: type of confidence intervals, recruitment unit, stock size indicator, stock size unit, catch unit, exploitation unit, and effort unit.

Automatic calculations

There are also many data fields in the STAR Excel template that are automatically calculated. This not only streamlines the work for stock assessors, working group chairs, and the secretariat, but vastly reduces the likelihood of errors in those fields.

Automatic calculations in the Metadata sheet are: current F, current B, B0.33, B0.66, F/Fref, B/Btarget, B/Bthreshold, B/Blimit, stock status text (exploitation), and stock status text (biomass).

Automatic calculations in the Advice sheet are: comma-separated GSAs, comma-separated GSA names, comma-separated countries, text describing current levels with appropriately rounded numbers (Fc = x, Bc = y), text describing reference points, text describing quantitative status with appropriately rounded numbers, text describing stock status (exploitation and biomass).

4.4.2 QC in the Excel template

Quality control (QC) is about design features that enable users and administrators to identify mistakes in STAR template data fields. In some cases, QC helps the stock assessor to identify and correct mistakes before submitting the STAR template. In other cases, QC helps the secretariat to identify mistakes and contact stock assessors to correct the mistakes.

As of August 2021, the Excel template does not have QC features. A new sheet called QC could be added to the STAR template, e.g. between Vocabularies and Advice. This sheet would contain a relatively simple list of checks that have the general form:

```
IF(logical_test, TRUE, FALSE)
```

Each test checks if the user has made a certain type of error in the STAR template. The highest priority would be to check for errors that are commonly made in practice, based on the STAR experience from 2021, see Section 4.5.

The outcome of a QC test can be conditionally formatted, so a `FALSE` value can be shown bold and red. Stock assessors would be asked to check their QC sheet before submitting their STAR templates, making everyone's work easier.

Recommendation 4.4.2:

⇒ It is recommended that the GFCM data center develops a new sheet called QC in the STAR Excel template, checking for common errors in the template.

4.4.3 QC in the R package

The R package `gfcmSTAR` comes with QC functions that check if a file exists, if it looks like a STAR template, etc. There is also a wrapper function `qc()` to run all QC functions on a given file, see 'Functions' in Section 4.3.2.

The QC functions are useful when importing submitted STAR Excel templates, to catch errors that stock assessors may have have made when entering the results of their assessments and to identify the precise and location of an error. In the R environment it is easy to apply the QC functions to a large collection of STAR templates and quickly identify Excel files that have errors.

When a STAR Excel template is imported into R using the `gfcmSTAR` package, every data field is converted to the appropriate type. In R, those are `character`, `integer`, `numeric`, `logical`, and `POSIXct` (date and time). This conversion augments the QC process by raising warnings when data fields contain values that do not conform to the expected type.

Recommendation 4.4.3:

⇒ It is recommended that the GFCM data center considers adding more QC functions to the R package. The development of additional QC functions should address specific errors that are commonly made in practice.

4.5 Experience from 2021

4.5.1 Countries metadata

A couple of months after the January assessment working group meetings, the STAR development team conducted a comparison between the newly submitted metadata and the metadata collected from assessments in previous years. This led to the discovery that an important metadata field 'Countries' was missing from the STAR Excel template version 1.0.0.

As a result, the Countries metadata needs to be filled in afterwards for the initial batch of STAR templates. This addition of Countries metadata will need to be filled in manually for STAR templates submitted before April 2021. Depending on what the GFCM data center recommends, this can be done at the Excel level, R level, CSV file level, or using a SharePoint interface to the STAR database.

- At the Excel level, one could start with an empty STAR template version 2.1.2 and manually complete all data fields by copying from the originally submitted template version 1.0.0.

- At the R level, one can pass `countries=c(CountryA,CountryB,CountryC)` to the `read.template` function when reading in a STAR template.

- At the CSV file level, one can edit the CSV file manually and add Countries using a text editor.

- A SharePoint interface can be developed for this specific task, allowing write access to the STAR database.

The GFCM data center may also develop and recommend a route that is slightly different from the above options.

Recommendation 4.5.1:

⇒ It is recommended that Countries metadata are added manually to the first batch of assessment results that were submitted with STAR template version 1.0.0.

4.5.2 STAR templates submitted Jan–Apr 2021

The STAR framework was introduced to the stock assessors at the joint plenary session of the assessment working groups on 18 January 2021, and the first stock assessment results were submitted via STAR templates on 22 January.

The stock assessors uploaded the STAR templates to a folder within the Teams channel for their working group meeting. The secretariat later copied the templates to a centralized location in the DevOps repository **StockAssessmentResults** (Section 2.6.2) inside the 'uploads' folder. In the period from January to April 2021, a total of 79 templates were submitted (Table 8).

Table 8. STAR templates submitted Jan–Apr 2021. The Period column indicates the first and last submission dates from each stock assessment meeting. The Folder column is the subdirectory inside the **StockAssessmentResults** 'uploads' folder. Templates is the number of STAR templates within each folder.

Period	Folder	Templates
22 Jan–16 Feb	`2021/WGSASP_General`	21
22 Jan–24 Feb	`2021/WGSAD_Western_Mediterranean`	28
26 Jan–15 Feb	`2021/WGSAD_Central_Eastern_Med_and_Adriatic_Sea`	29
20 Apr	`2021/WGSAD_Benchmark_SOL_GSA_17`	1
	Total	79

4.5.3 WGSASP

The WGSASP assessment working group submitted 21 templates by placing them in the Teams folder called STAR. The Excel templates were of version 1.0.0. The administrators copied the files for safe permanent storage in the **StockAssessmentResults** DevOps repository inside a folder called:

> uploads/2021/WGSASP_General

When the `import_WGSASP_General.R` script is run, a report summarizes how many files were successfully imported and lists any import errors and QC issues:

```
*** Count

$Import
 21 files
 17 imported
  2 failed   (5, 13)
  2 removed  (9, 17)

$QC
 3 issues (5, 9, 13)

*** Filenames

$Failed
 [5] star_ANE_GSA01 ref2019.xlsx => Error in if (VPA_Model)
[13] star_PIL_GSA04 ref2019.xlsx => Error in if (VPA_Model)

$Removed
 [9] STAR_PIL_17_18.xlsx
[17] star_template.xlsx

$QC
 [5] star_ANE_GSA01 ref2019.xlsx => vpa
 [9] STAR_PIL_17_18.xlsx         => ts.numbers
[13] star_PIL_GSA04 ref2019.xlsx => vpa
```

In addition to the errors and QC issues listed above, the import script shows the following warnings:

```
[6] star_ANE_GSA06 ref2019_model1.xlsx
  Warning in class(Metadata$Year_Benchmarked) <- "integer" :
  NAs introduced by coercion


[7] star_ANE_GSA06 ref2019_model2.xlsx
  Warning in class(Metadata$Year_Benchmarked) <- "integer" :
  NAs introduced by coercion
```

For reference, the complete list of STAR templates submitted by WGSASP:

```
 [1] STAR_2019_ANE_7.xlsx               STAR_2019_PIL_7.xlsx
 [3] star_ANE_20_ref2019.xlsx           star_ANE_22_ref2019.xlsx
 [5] star_ANE_GSA01 ref2019.xlsx        star_ANE_GSA06 ref2019_model1.xlsx
 [7] star_ANE_GSA06 ref2019_model2.xlsx star_ANE_GSA16 2019.xlsx
 [9] STAR_PIL_17_18.xlsx                star_PIL_20_ref2019.xlsx
[11] star_PIL_22_ref2019.xlsx           star_PIL_GSA01 ref2019.xlsx
[13] star_PIL_GSA04 ref2019.xlsx        STAR_PIL_GSA04W_VIT_Ref  year_2019.xlsx
[15] star_PIL_GSA06 ref2019.xlsx        star_PIL_GSA16 2019.xlsx
[17] star_template.xlsx                 STAR-2019-ANE-1718.xlsx
[19] STAR-2019-ANE-9.xlsx               STAR-2019-PIL-3.xlsx
[21] STAR-2019-PIL-9.xlsx
```

Commentary

In files 5 and 13, the VPA-Type Model cell in the template had been left blank. This prevents the calculation of current stock status and should be addressed and fixed before importing to the STAR database. Likewise, file 9 had Time Series entered as strings and not numbers, also preventing the calculation of current stock status. These problems are already evident to the stock assessor filling in the STAR template, where stock status entries are shown as blank.

After inspection, file 17 turned out to be a partial draft version of `PIL_1718` which already had a full version submitted. File 17 was therefore manually removed as an invalid submission.

Files 6 and 7 had string entries for the Year Benchmarked, namely "2021 (no finished)", which raises a warning during import into R. This does not prevent the import of these templates, but after conversion to integers this leaves Year Benchmarked as `NA`.

Files 6 and 7 were two submissions for the same reference year, species, and GSA (2019, ANE, GSA 6) using different assessment methods. To prepare both entries for STAR database import, they were given the unique identifiers `STAR_2019_ANE_6_a4a_spict` and `STAR_2019_ANE_6_a4a`.

Summary

After removing the invalid submission, a total of 20 assessment results were submitted from WGSASP. Out of those, 17 were successfully imported, or 85%. The remaining 3 need only minor improvements to make importing them possible.

4.5.4 WGSAD Western Mediterranean

The WGSAD Western Mediterranean assessment working group submitted 28 templates by placing them in the Teams folder called STAR. The Excel templates were of version 1.0.0. The administrators copied the files for safe permanent storage in the **StockAssessmentResults** DevOps repository inside a folder called:

    uploads/2021/WGSAD_Western_Mediterranean

When the `import_WGSAD_Western_Mediterranean.R` script is run, a report summarizes how many files were successfully imported and lists any import errors and QC issues:

```
*** Count

$Import
 28 files
 25 imported
  3 failed    (7, 26, 28)
  0 removed

$QC
 3 issues (7, 26, 28)

*** Filenames

$Failed
 [7] STAR_2019_DPS_5.xlsx  => Error in if (VPA_Model)
[26] STAR_2019-SBR-13 .xls => Error : IllegalArgument
[28] star_template.xlsx    => Error in if (VPA_Model)

$Removed
 None

$QC
 [7] STAR_2019_DPS_5.xlsx  => vpa
[26] STAR_2019-SBR-13 .xls => xlsx
[28] star_template.xlsx    => vpa
```

For reference, the complete list of STAR templates submitted by WGSAD Western Mediterranean:

```
 [1] STAR_2019_ DPS_6.xlsx        STAR_2019_ARA_5.xlsx
 [3] STAR_2019_ARA-1_IEO.xlsx     STAR_2019_ARA-2.xlsx
 [5] STAR_2019_ARA-6.xlsx         STAR_2019_ARS_91011.xlsx
 [7] STAR_2019_DPS_5.xlsx         STAR_2019_DPS-03.xlsx
 [9] STAR_2019_DPS-1.xlsx         STAR_2019_HKE_4.xlsx
[11] STAR_2019_HKE_5.xlsx         STAR_2019_HKE-1.xlsx
[13] STAR_2019_HKE-6.xlsx         STAR_2019_MUR_5.xlsx
[15] STAR_2019_MUT_01.xlsx        STAR_2019_MUT-6.xlsx
[17] STAR_2019_MUT10.xlsx         STAR_2019_NEP-6.xlsx
[19] STAR_2019-ARA-91011.xlsx     STAR_2019-DPS-91011.xlsx
[21] STAR_2019-HKE-1567.xlsx      STAR_2019-HKE-891011.xlsx
[23] STAR_2019-MUT-7.xlsx         STAR_2019-MUT-9.xlsx
[25] STAR_2019-NEP-9.xlsx         STAR_2019-SBR-13 .xls
[27] Star_Hake 01&03.xlsx         star_template.xlsx
```

Commentary

In file 7, the VPA-Type Model cell in the template had been left blank. This prevents the calculation of current stock status and should be addressed and fixed before importing to the STAR database. The problem is already evident to the stock assessor filling in the STAR template, where stock status entries are shown as blank.

After inspection, file 28 turned out to be an empty template and therefore not a valid submission.

File 26 had been saved in `xls` format, but STAR templates are provided as `xlsx` and must be saved in that format. This conversion to a very old format (before MS Office 2007) is harmful, as the old file format removes all table objects that are used as data containers in the STAR template.

Summary

After removing the invalid submission, a total of 27 assessment results were submitted from WGSAD Western Mediterranean. Out of those, 25 were successfully imported, or 93%. The remaining 2 need only minor improvements to make importing them possible.

4.5.5　WGSAD Central, Eastern Med and Adriatic Sea

The WGSAD Central, Eastern Med and Adriatic Sea assessment working group submitted 29 templates by placing them in the Teams folder called STAR. The Excel templates were of version 1.0.0. The administrators copied the files for safe permanent storage in the **StockAssessmentResults** DevOps repository inside a folder called:

> `uploads/2021/WGSAD_Central_Eastern_Med_and_Adriatic_Sea`

When the `import_WGSAD_Central_Eastern_Med_and_Adriatic_Sea.R` script is run, a report summarizes how many files were successfully imported and lists any import errors and QC issues:

```
*** Count

$Import
 29 files
 21 imported
  7 failed   (5, 9, 13, 14, 15, 20, 25)
  1 removed  (21)

$QC
 7 issues (5, 9, 13, 14, 15, 20, 25)

*** Filenames

$Failed
 [5] star_2019_MUT-15.xlsx                 => Error in $<-.data.frame
 [9] star_DPS 17_19.xls                    => Error : IllegalArgument
[13] star_HKE_19 bnchmk settings_unstable model.xls => Error:Illegal
[14] star_HKE_19 STECF accepted _stable model.xls   => Error:Illegal
[15] star_template.xlsx                    => Error in if (VPA_Model)
[20] star-2019-MTS-17.xlsx                 => Error in $<-.data.frame
[25] star-2019-MUT-22.xlsx                 => Error in if (VPA_Model)

$Removed
[21] star-2019-MUR-26 (1).xlsx

$QC
 [5] star_2019_MUT-15.xlsx                    => ts.names
 [9] star_DPS 17_19.xls                       => xlsx
[13] star_HKE_19 bnchmk settings_unstable model.xls => xlsx
[14] star_HKE_19 STECF accepted _stable model.xls   => xlsx
[15] star_template.xlsx                       => vpa
[20] star-2019-MTS-17.xlsx                    => ts.names
[25] star-2019-MUT-22.xlsx                    => vpa
```

In addition to the errors and QC issues listed above, the import script shows the following warnings:

```
[10] STAR_DPS_GSAs12-16.xlsx
  Warning in class(Metadata$Fbar_First_Age) <- "integer" :
  NAs introduced by coercion
  Warning in class(Metadata$Fbar_Last_Age) <- "integer" :
  NAs introduced by coercion
```

For reference, the complete list of STAR templates submitted by WGSAD Central, Eastern Med and Adriatic Sea:

```
 [1] star_2019_ARS_GSAs18-19.xlsx      STAR_2019_MUT_12-14.xlsx
 [3] star_2019_MUT_GSA19.xlsx          star_2019_MUT_GSAs17-18.xlsx
 [5] star_2019_MUT-15.xlsx             star_2019-ANK-17.xlsx
 [7] star_2019-CTC-17.xlsx             star_BOY_GSA17.xlsx
 [9] star_DPS 17_19.xls                STAR_DPS_GSAs12-16.xlsx
[11] star_EOI_GSA_18.xlsx              star_HKE_12_16.xlsx
[13] star_HKE_19 bnchmk settings_unstable model.xls
     star_HKE_19 STECF accepted _stable model.xls
[15] star_template.xlsx                star-2019-HKE-1718.xlsx
[17] star-2019-HKE-20.xlsx             star-2019-HKE-22.xlsx
[19] STAR-2019-LithMor-27.xlsx         star-2019-MTS-17.xlsx
[21] star-2019-MUR-26 (1).xlsx         star-2019-MUR-26.xlsx
[23] star-2019-MUT-16.xlsx             star-2019-MUT-20.xlsx
[25] star-2019-MUT-22.xlsx             star-2019-MUT-24.xlsx
[27] STAR-2019-PagEry-27.xlsx          star-2019-SJA-17.xlsx
[29] star-2020-MUT-25.xlsx
```

Commentary

Seven files have errors that prevent import into R. The QC detects issues in the same seven files, and indicates the exact nature of these errors. In files 15 and 25, the VPA-Type Model cell in the template had been left blank. This prevents the calculation of current stock status and should be addressed and fixed before importing to the STAR database. The problem is already evident to the stock assessor filling in the STAR template, where stock status entries are shown as blank.

Files 9, 13, and 14 had been saved in xls format, but STAR templates are provided as xlsx and must be saved in that format. This conversion to a very old format (before MS Office 2007) is harmful, as the old file format removes all table objects that are used as data containers in the STAR template.

Files 5 and 20 had altered column names in the Time Series sheet. For example, the Fishing column had been renamed to Fishing (1-3), Rec_lower to Colonna2, and Stock1_upper to StDev3. In STAR templates, the Time Series column names must remain intact. The content of these columns is standardized and they map directly to the corresponding fields in the STAR database.

After inspection, file 15 turned out to be an empty template. File 21 turned out to be a duplicate of file 22 and therefore not a valid submission.

File 10 had string entries for Fbar first age and Fbar last age, namely "0.17 (age 0)" and "1.15 (age 3)", which raises a warning during import into R. This does not prevent the import of this template, but after conversion to integers this leaves Fbar first age and Fbar last age as NA.

Summary

After removing the invalid submission, a total of 27 assessment results were submitted from WGSAD Central, Eastern Med and Adriatic Sea. Out of those, 21 were successfully imported, or 78%. The remaining 6 need only minor improvements to make importing them possible.

4.5.6 WGSAD Benchmark SOL GSA 17

The WGSAD Benchmark SOL GSA 17 expert group submitted 1 template by placing it in the Teams folder called STAR. This Excel template was of version 2.1.0. The administrators copied the file for safe permanent storage in the StockAssessmentResults DevOps repository inside a folder called:

```
uploads/2021/WGSAD_Benchmark_SOL_GSA_17
```

The `import_WGSAD_Benchmark_SOL_GSA_17.R` script consists of essentially three lines of R code, given that it only has to handle a single STAR template:

```
sol17 <- read.template("STAR_SOL_GSA17.xlsx")
qc("STAR_SOL_GSA17.xlsx")
write.star.csv(sol17)
```

The first command imports the Excel template into R, the second command runs a suite of QC tests, and the third command exports the STAR object from R into a set of two standardized CSV files ready for database import.

The import of this file was 100% successful, passing all QC tests:

```
* checking 'STAR_SOL_GSA17.xlsx' with qc.exists ... OK
* checking 'STAR_SOL_GSA17.xlsx' with qc.xlsx ... OK
* checking 'STAR_SOL_GSA17.xlsx' with qc.star ... OK
* checking 'STAR_SOL_GSA17.xlsx' with qc.vpa ... OK
* checking 'STAR_SOL_GSA17.xlsx' with qc.ts.names ... OK
* checking 'STAR_SOL_GSA17.xlsx' with qc.ts.numbers ... OK
```

4.5.7   Outstanding task: Import CSV files into database

A total of 64 STAR templates have now been imported into R, quality-controlled, and exported to CSV files (Table 9). These files have been placed in subfolders on a dedicated mounted drive `/mnt/star-templates` on the RStudio Server. The next step (Figure 5) is to import the CSV files into the STAR database.

Table 9. Number of quality-controlled STAR templates that were successfully exported to CSV format, ready for import into the STAR database.

Folder	Templates
`WGSAD_Benchmark_SOL_GSA_17`	1
`WGSAD_Central_Eastern_Med_and_Adriatic_Sea`	21
`WGSAD_Western_Mediterranean`	25
`WGSASP_General`	17
Total	64

Recommendation 4.5.7:

⇒ It is recommended that the CSV files that passed quality control and have been exported to `/mnt/star-templates` be imported into the STAR database.

4.5.8   Outstanding task: Fix errors in Excel templates

A total of 11 STAR templates have errors that need to be fixed to make import into R possible. In most cases, the required improvements are minor and require only a small amount of work.

Table 10. List of STAR templates that have errors preventing import into R.

Group	STAR template	Error
WGSASP	`star_ANE_GSA01 ref2019.xlsx`	vpa
WGSASP	`star_PIL_GSA04 ref2019.xlsx`	ts.numbers
WGSASP	`STAR_PIL_17_18.xlsx`	vpa
WGSAD-WM	`STAR_2019_DPS_5.xlsx`	vpa
WGSAD-WM	`STAR_2019-SBR-13 .xls`	xlsx
WGSAD-CEMA	`star_2019_MUT-15.xlsx`	ts.names
WGSAD-CEMA	`star_DPS 17_19.xls`	xlsx
WGSAD-CEMA	`star_HKE_19 bnchmk settings_unstable model.xls`	xlsx
WGSAD-CEMA	`star_HKE_19 STECF accepted _stable model.xls`	xlsx
WGSAD-CEMA	`star-2019-MTS-17.xlsx`	ts.names
WGSAD-CEMA	`star-2019-MUT-22.xlsx`	vpa

Recommendation 4.5.8:

⇒ It is recommended that STAR templates with errors preventing import into R be fixed, either by the stock assessors who submitted them or by the secretariat.

4.6 New assessment and advisory workflow

4.6.1 Before the assessment working group meeting

To prevent any confusion about what is the current version of the STAR Excel template, it might be best to instruct stock assessors to download it from the STAR repository:

https://github.com/gfcm/star

An introductory STAR slide, or slides, can be a part of a standard presentation at the beginning of a meeting.

A helpful email could be sent to inidivual stock assessors privately, explaining errors that were found in the STAR template they submitted last year.

4.6.2 During the assessment working group meeting

When a stock assessor has the final model run ready, they complete a STAR template with the assessment results. The two main sheets are Metadata and TimeSeries, while stock assessors are also invited to submit additional information in the sheets called Forecast and Other.

In each working group, there are many stock assessors who are expert R users. They may find the R package gfcmSTAR useful to check that a completed STAR template can be imported into R and passes the quality checks, using two simple commands:

```
read.template("STAR_2019_HKE_5.xlsx")
qc("STAR_2019_HKE_5.xlsx")
```

Stock assessors upload completed STAR templates to a dedicated site. In 2021, this was a STAR folder on the Teams channel for each working group meeting, directly connected to SharePoint.

It is recommended to use descriptive filenames, such as STAR_2019_HKE_5.xlsx or STAR_2019_ARA_91011.xlsx. The STAR system, however, is robust to any kind of filenames. At the time of import into R, a standardized Assessment_ID is constructed from the reference year, species, GSA, and an optional suffix to make it unique. This unique ID is then used in the STAR database.

4.6.3   After the assessment working group meeting

Copy to DevOps

When the templates have been submitted to the dedicated upload site (Teams in 2021), the STAR administrators copy the files for safe permanent storage in the **StockAssessmentResults** DevOps repository inside a folder that has a standardized name, for example:

```
uploads/2021/WGSASP_General
```

The meeting (or "event") has the same name as the Teams channel, and the year indicates when the event took place.

R, QC, CSV

The next step is to import the STAR templates into R, run QC tests, and export to CSV format. This is a crucial step in the STAR framework, a pipeline of data flow from the Excel templates to the SQL database, while guaranteeing a level of data quality (Figure 5).

Figure 5. STAR import procedure, showing the flow of data from the Excel template, import into R, export to CSV format, and import into the database. The R import step runs a suite of QC tests to check for common errors.

The procedure is semi-automated using the R package **gfcmSTAR**, along with the `import.R` script that comes with it. The import script is adapted to work with a given set of Excel templates by specifying the year and event, directing it to the DevOps folder containing the Excel templates. Furthermore, the script contains a section to handle files that are not valid templates, or cases where `Assessment_ID` needs to be set manually.

Sections 4.5.3, 4.5.4, and 4.5.5 of this report are annotated examples detailing the procedure of bulk import of Excel templates into R, QC, and exporting to standardized CSV files, ready to be imported into the database. The full import scripts are in the `uploads/2021` DevOps folder.

Import into database

After the R import QC procedure, the CSV files are on the RStudio Server on the dedicated drive `/mnt/star-templates`. Since the CSV files are designed to exactly match the database table design, the CSV-to-database import should be relatively straightforward, once the database import procedure has been scripted.

4.6.4 Analysis of submitted STAR templates

Reading from database

The recommended way to analyze STAR data is to access the data from the STAR database, or from products derived from the database. All data corrections in the STAR framework focus on making the STAR database the authoritative quality-controlled source of STAR data for analysis.

The database is designed for convenient queries, containing all the STAR data in two tables called Metadata and TimeSeries (Figure 4). Take, for example, an analysis of F/F_{ref}, the current fishing mortality relative to the target reference point, where a minimal script could be:

```
library(RODBC)
con <- odbcConnect("STOCKASSESSMENTDSN-Pub", "user", "pass")
x <- sqlQuery(con, "SELECT Status_Fref
                    FROM Metadata
                    WHERE Reference_Year = 2019")
boxplot(x$Status_Fref, horizontal=TRUE, xlab="F / Fref")
abline(v=1, lwd=5, col="green3")
```

Reading from CSV

In the interim period until the database is operational, the STAR data can only be accessed from a large number of Excel templates or a large number of CSV files. Luckily, this is made relatively easy by the `read.many.csv()` function in the R package **gfcmSTAR**. Likewise, the `peek()` function makes it easy to access values from a cluster, a list of STAR objects:

```
library(gfcmSTAR)
setwd("/mnt/star-templates/2021")

wgsasp <- import.many.csv("WGSASP_General")
wgsad_wm <- import.many.csv("WGSAD_Western_Mediterranean")
wgsad_cema <- import.many.csv(
   "WGSAD_Central_Eastern_Med_and_Adriatic_Sea")
wgsad_sol17 <- import.many.csv("WGSAD_Benchmark_SOL_GSA_17")

cluster <- c(wgsasp, wgsad_wm, wgsad_cema, wgsad_sol17)
Status_Fref <- peek(cluster, "Status_Fref")
boxplot(Status_Fref, horizontal=TRUE, xlab="F / Fref")
abline(v=1, lwd=5, col="green3")
```

This script can now be run successfully on the RStudio Server and the resulting plot is shown in Figure 6.
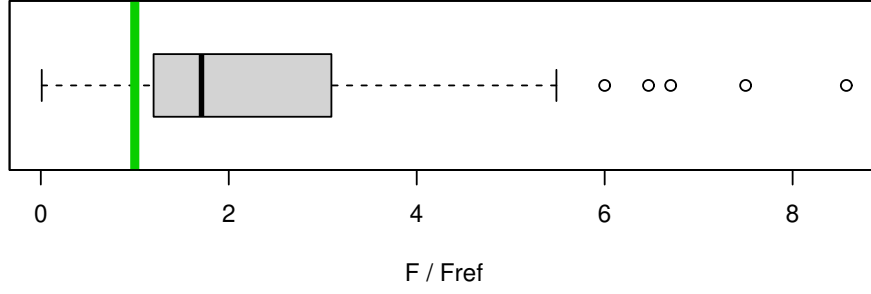
Figure 6. Current fishing mortality relative to the target reference point (F/F_{ref}) from 50 fisheries in the Mediterranean. The data are based on 64 valid STAR templates that were submitted in 2021, of which 50 included a numeric value for F/F_{ref}. The green vertical line shows $F = F_{\text{ref}}$.

4.6.5 Open STAR data repository

The R code snippets in Section 4.6.4 demonstrate how the STAR framework can bring a large amount of important stock assessment results to the fingertips of an analyst, ready to be analyzed, tabulated, and plotted.

This has obvious value for GFCM to facilitate and strengthen the analysis behind the State of Mediterranean and Black Sea Fisheries (SoMFi, FAO 2020) report, as well as the general provision of advice to GFCM member states. Easy access to STAR data can also be of great help for many routine tasks done by the secretariat and expert group chairs.

Furthermore, the STAR framework can also become a landmark in Mediterranean and Black Sea fisheries science, by making all stock assessment results global and open access. This would bring GFCM to the forefront of regional fisheries management organizations that are currently working towards this goal, to be able to say that the management of their fish stocks is based on the modern principles of open and transparent science.

In the past, the GFCM stock assessment results have been made available as Stock Assessment Forms (SAF) and various other reports in PDF format. The STAR framework makes the data available in a structured and complete format that can be read directly into R, Excel or any other software.

Since all the CSV files have already been created as part of the process of importing STAR data from Excel templates to the database, it can take a relatively small amount of work for the GFCM data center to upload the CSV files to a public GitHub repository. The same hierarchical directory structure can be used as in the private DevOps repository.

GitHub comes with a built-in CSV viewer that allows users to browse the data online and download individual files or read them directly from GitHub into R and other software. GitHub facilities also allow users to download the entire collection of CSV files as a zip file. In other words, the basic services of a data repository are already built in.

Recommendation 4.6.5:

⇒ It is recommended that all STAR data are made available as CSV files in an open STAR data repository. The results from a stock assessment would be uploaded and made public once the assessment has passed all review stages and the advice has been released. A dedicated repository on the existing GFCM GitHub site would be a convenient platform, involving very little extra work for the data center.

4.6.6 Online visual browser

Although browsing and viewing CSV files online (see Section 4.6.5) would be a significant service for analysts with technical skills, there would be important benefits of also providing plots of the STAR data to browse and view online.

There are at least five kinds of plots to consider:

1. Plots from a single stock assessment. For example, time series of catches, biomass, fishing mortality, and recruitment, overlaying horizontal lines to show relevant reference points.

2. Plots from the assessments of the same stock carried out in different years. For example, empirical retrospective patterns in B and F time series, as a measure of the reliability of the assessment.

3. Plots summarizing all stock assessments that were carried out in a given year. For example, total catches by species group and average F/F_{ref} by species group.

4. Plots summarizing all stock assessments from all years. For esxample, how the proportion of stocks that are being overfished has changed over time.

5. Plots that combine STAR results with other data sources. For example, effort data, socio-economic data, environmental data, and spatial data.

Clearly, there could be a long list of possible plots and user interface features to consider, so the required cost could easily grow quickly, in terms of development and maintenance. Therefore, the goal of an online visual browser is perhaps best approached in small steps. For example, uploading to a webpage static versions of very few key plots based on STAR results that would be fundamentally useful for the work of the secretariat and/or advisory products.

If an online visual browser of STAR results becomes popular and valuable, it can grow organically with time, adding a dynamic user interface and other features.

Recommendation 4.6.6:

⇒ It is recommended that GFCM evaluates the usefulness of an online visual browser of plots based on STAR data. This development should prioritize plots that would be of fundamental value.

4.6.7 Automated Table of Advice

A primary objective of the STAR framework is to streamline and support the process of producing the Table of Advice. The STAR template version 1.0.0, which was used in the 2021 stock assessment working group meetings (WGSASP and WGSAD), demonstrated important benefits in making the construction of the Table of Advice a more streamlined and reliable process.

Based on feedback from the working group meetings, as well as further research and development, the current STAR template version 2.1.2 provides further automation and is more closely aligned with the desired format and content of the Table of Advice.

The automation could be taken a step further. Instead of manually pasting the computer-generated advice lines from a spreadsheet, an online service could construct the Table of Advice from submitted STAR templates. In this way, submitting STAR templates would become the only way an assessment working group meeting can add and modify entries in the Table of Advice.

Perhaps the experience from the upcoming 2022 assessment year will indicate whether the current level of automation, where the computer-generated advice line is pasted from a spreadsheet, is adequate and practical. Further automation, to create an online service that constructs the Table of Advice, would require substantial research and development from the GFCM data center, but the benefits may outweigh the cost.

The secretariat, working group chairs, and stock assessors share the task of STAR quality control (QC) and corrections of minor errors in submitted STAR templates. Therefore, a relevant question regarding further automation of the Table of Advice, is whether all QC tasks are completed during or after the working group meeting.

Recommendation 4.6.7:

⇒ It is recommended that GFCM considers whether further automation is useful for the construction of the Table of Advice. During the 2022 assessment year, the secretariat and working group chairs should carefully consider how the Table of Advice is constructed in the current STAR framework and evaluate the benefits of further automation, as well as QC tasks.

5 Recommendations for STAR development

5.1 List of STAR recommendations

The following list is a condensed summary of all recommendations from Section 4.
See the cross-linked section numbers for context and further details.

Database

(4.3.3) Create STAR database.

(4.3.4) Merge SAF metadata from previous years with new STAR Metadata table.

Excel template

(4.4.2) Develop a new sheet called QC in the STAR Excel template.

R package

(4.4.3) Add more QC functions to the R package.

Import assessment results into database

(4.5.1) Add Countries metadata manually to the 2021 STAR submissions.

(4.5.7) Import CSV files that passed quality control into database.

(4.5.8) Fix errors in 2021 STAR submissions preventing import into R.

Future goals

(4.6.5) Make CSV files available in an open STAR data repository.

(4.6.6) Evaluate the usefulness of an online visual browser of STAR plots.

(4.6.7) Evaluate whether the Table of Advice should be further automated.

6 Summary and conclusions

6.1 Current status of GFCM data management

Section 2 of this report gives an overview of some of the main GFCM data products and services.

Overall, the GFCM data center does an excellent job managing fisheries and other data. The main data products are organized in relational databases, stored on reliable out-of-house servers, continually backed up. Over the years, data systems have been successfully migrated to newer platforms, as technologies move forward.

The user interface for the secretariat and others to interact with the data is generally quite good. Communication and collaboration between the GFCM data center and other teams is outstanding, both internally and with external partners.

6.2 Recent improvements

6.2.1 Related to this project

Four additions to the GFCM data products and services can be highlighted, as valuable by-products that were created during this project of reviewing fisheries data management at GFCM:

- GFCM GitHub area – to organize development and disseminate products
- MSSQL package – to browse and examine GFCM database tables
- STAR template - to enter and submit stock assessment results
- gfcmSTAR package - to oversee STAR quality control and data import

6.2.2 Other advances in 2020–2021

The GFCM data center has in the last months launched three major data products that were not covered in this review:

- GFCM Database of Fisheries Restricted Areas
- Small-Scale Fisheries Mapping Tool
- GFCM Register of Ports

6.3 Future improvements

It is outside the scope of this review to predict future improvements to GFCM data management, but a number of recommendations are made in this review. A total of 18 recommendations have been divided into general recommendations (Section 3) and STAR recommendations (Section 5). Each recommendation has a link to the main text where the topic is covered in detail.

7   Acknowledgements

8   References

FAO (Food and Agriculture Organization of the United Nations). 2020. The State of Mediterranean and Black Sea Fisheries 2020. General Fisheries Commission for the Mediterranean. Rome. 139 pp.
https://doi.org/10.4060/cb2429en

GFCM (General Fisheries Commission for the Mediterranean). 2018. GFCM Data Collection Reference Framework (DCRF). Version: 21.2. 222 pp.
https://www.fao.org/gfcm/data/dcrf/en/

Goodman, A., et al. 2014. Ten simple rules for the care and feeding of scientific data. PLOS Computational Biology 10:e1003542. 5 pp.
https://doi.org/10.1371/journal.pcbi.1003542

ICES (International Council for the Exploration of the Sea). 2021. ICES Stock Assessment Graphs.
https://standardgraphs.ices.dk

Magnusson, A. 2021a. gfcmSTAR: Tools to work with GFCM Stock Assessment Results (STAR). R package version 2.3.0. 32 pp.
https://github.com/gfcm/gfcmSTAR/releases/download/2.3.0/gfcmSTAR.pdf

Magnusson, A. 2021b. MSSQL: Tools to work with Microsoft SQL Server databases via RODBC. R package version 1.0.0. 12 pp.
https://cran.r-project.org/web/packages/MSSQL/MSSQL.pdf

Molloy, J.C. 2011. The Open Knowledge Foundation: Open data means better science. PLOS Biology 9:e1001195. 4 pp.
https://doi.org/10.1371/journal.pbio.1001195

Wilkinson, M.D., et al. 2016. The FAIR Guiding Principles for scientific data manamagement and stewardship. Scientific Data 3:160018. 9 pp.
https://doi.org/10.1038/sdata.2016.18