# ICCAT GBYP ELECTRONIC TAGS DATABASE GOES SHINY

S. Tensek[1]

*SUMMARY*

*The relational database was created in PostgreSQL and it incorporates structured data provided by ICCAT GBYP electronic tags. The script for automatic populating of the database has been developed in R, for the data provided in the format used by Wildlife Computers and CLS. All tags sensor observations on depth and temperature have been joined to their corresponding maximum probable geolocations. In addition, the interactive application was created using Shiny by RStudio which allows visualisation of the track, temperature and depth series, time at temperature, time at depth, minimum and maximum depth in relation to the sea bottom and the Argos post-release track.*

*RÉSUMÉ*

*La base de données relationnelle a été créée dans PostgreSQL et intègre des données structurées fournies par les marques électroniques de l'ICCAT-GBYP. Le script pour le remplissage automatique de la base de données a été développé en R, pour les données fournies dans le format utilisé par Wildlife Computers et CLS. Toutes les observations des capteurs des marques sur la profondeur et la température ont été reliées à leurs géolocalisations probables maximales correspondantes. En outre, l'application interactive a été créée à l'aide de Shiny by RStudio qui permet la visualisation du tracé, la température et la série de profondeurs, le temps à la température, le temps passé en profondeur, la profondeur minimale et maximale par rapport au fond marin et le tracé Argos après la remise à l'eau.*

*RESUMEN*

*La base de datos relacional se creó en PostgreSQL e incorpora datos estructurados facilitados por las marcas electrónicas del ICCAT GBYP. El script para alimentar de forma automática la base de datos se ha desarrollado en R, para los datos facilitados en el formato utilizado por Wildlife Computers y CLS. Todas las observaciones de los sensores de las marcas relacionadas con profundidad y temperatura se han vinculado con sus geolocalizaciones más probables correspondientes. Además, se creó la aplicación interactiva utilizando Shiny by RStudio que permite visualizar las series de rastreo de ruta, temperatura y profundidad, tiempo en temperatura, tiempo en profundidad, profundidad máxima y mínima en relación con el fondo del mar y rastreo de la ruta tras la liberación mediante Argos.*

*KEYWORDS*

*Tagging, bluefin tuna, electronic tags, data collections*

[1] ICCAT GBYP, Calle Corazón de María, 8, 6th floor, 28002 Spain, stasa.tensek@iccat.int

## 1. Introduction

The Atlantic-Wide Research Programme for Bluefin Tuna (GBYP) is a special research programme of ICCAT whose main goals are to reduce uncertainty in stock assessment and to provide robust management advice, which requires improved knowledge of key biological processes and parameters. Since currently almost all the data used in the stock assessments are obtained from the fisheries-dependent data, it is important to obtain data from alternative sources, such as tagging studies, in order to verify the assumptions made when conducting the assessments.

One of the major research tasks under the ICCAT GBYP is to carry out the large, wide and intensive scientific tagging program to address several important biological and ecological topics regarding Atlantic bluefin tuna (*Thunnus thynnus*). According to the general programme, in Phase 1 (2010-2011) the Tagging Design and the Tagging Manual were adopted and from Phase 2 (2011-2012) onwards, it was proceeded with the tagging activities. At the beginning, the priority was given to the deployment of the conventional tags, but due to low recovery rates, in Phase 4 the Steering Committee decided to focus on the deployment of electronic tags instead, keeping the conventional tagging only as a complementary activity. Tagging activities of the ICCAT GBYP have been reported by Di Natale *et al.* (2015), Di Natale *et al.* (2016), Di Natale *et al.* (2017) and Tensek *et al.* (in press).

GBYP started with electronic tag activities in 2011 when the first few pop-up satellite tags were deployed. Over the course of 7 years, up to 2017, within the framework of this Program, or in joint actions with other institutions, more than 325 electronic pop-up tags were deployed. The exact number is unknown because some of the cooperating institutions haven't provide any data to the ICCAT GBYP, despite repeated requests. Moreover, the number of recovered and useful datasets is even smaller because some tags have either never emitted, or the transmitted data were corrupted, or the tags prematurely detached after only a few days.

All available electronic tag datasets have already been provided to Matt Lauretta (in charge of collecting and analysing all satellite tags data sets on behalf of the SCRS BFT Species Group) and Tom Carruthers (the modelling expert in charge of developing the technical aspects of both MSE and OM on behalf of the ICCAT GBYP Core Modelling MSE Group) and are currently been used for the bluefin tuna stock-assessment purposes. The basic analysis of the electronic tag data up to Phase 6 has already been performed by Tensek *et al.* (2017). The complex analysis of the data has not been performed yet.

The data obtained by electronic tags comprise a set of unstructured and semi-structured data of different type which are received in different formats and follow a different folder hierarchy. These data can be kept in the same repository and can be named by the individual tag ID, but without performing the additional effort, they cannot be further organized. In this way the data are not easily accessed or searched through and it is difficult to do any type of comparison between two or more tags. The first step towards performing any more complex analysis comprehends preparing the data first. Data preparation is a tedious and a time consuming process and, according to the some data science experts, usually more than the half of the work in data analysis is preparing the data for analysis. Nevertheless, once data are cleaned and assembled, they can be easily introduced in a relational database or a spreadsheet. Structured data has the advantage of being easily entered, stored, queried and analysed.

In the GBYP Phase 7 the first attempt was made towards creating the relational database which would gather all relevant data provided by electronic tags and in this way facilitate their extraction for any further analysis. The database was created in PostgreSQL and can be accessed directly or through R. A code was developed in R which serves for automatic cleaning, formatting and writing the data into the database, but also for its querying. In addition, the interactive application was developed using Shiny which permits quick visual insight into the data of each tag.

## 2. Materials and methods

### *2.1. Type of data*

All the electronic tags deployed so far by the ICCAT GBYP have been made by the same manufacturer, Wildlife Computers, and therefore the datasets produced by these tags follow more or less the same pattern and have the similar format. The format of files has slightly changed over the years, but the changes were limited to introducing some new columns and changing the column order. This fact makes the datasets produced by different tags comparable and easier to compile in a unique table.

ICCAT GBYP receives the data on electronic tags from various sources: Wildlife Computers Portal, Wildlife Computers Factory and CLS Company. The scheme of the different files that are received, including their source and format is provided in **Table 1**. Tag sensor data on light, depth and temperature, including the Argos post-release track can be freely downloaded from the Wildlife Computers Portal as several CSV files for each tag. These data don't include the processed locations (maximum probable geolocations estimated by a state-space model), but they can also be calculated through the Portal and freely downloaded in a single CSV file. GBYP also receives the data from CLS. This contracted service provides the delivery of the tag sensor data including the processed locations (calculated by, what is considered, the most reliable model), several graphical representations and the Argos post-release track. In case that the tag has been physically recovered, it is sent to the Wildlife Computers Factory where the sensor data are downloaded directly from the tag and then sent to GBYP. The data obtained this way are complete, in contrast to the data obtained by the satellite transmission, and serve for providing more accurate processed locations.

Tag sensor data (raw observations) include readings of the tag sensors on depth and temperature which are saved in the following spreadsheets:

- SERIES – contains series of temperature and depth readings in a given sample interval
- SERIES RANGE – consolidates the minimum and maximum values of depth and temperature in a given period. These values don't necessarily correspond to series values
- SST – contains set of readings qualified as possible sea surface temperature (from 0-5 m depth) in a given sampling period
- HISTOS – accumulates histogram data i.e. percent of time spent in each bin during the predefined period, including Time At Depth (TAD), Time At Temperature (TAT), definition of TAD limits (bins) and TAT limits (bins)
- MINMAXDEPTH – contains minimal and maximal depths experienced by the tag in a given period of time
- MIXLAYER – information about the mixed layer including the percent of time spent in and out of the layer during the predefined period.

The spreadsheet ARGOS contains a series of locations of the tag after the release (pop-off). General and summary information of the tag deployment and release, including the information on the success of the transmission is given in the file SUMMARY. In addition to these files, the raw data produced by the tag include the file LIGHTLOC which compiles information used by the Wildlife Computer software GPE3 for calculating the geolocation estimates and also some other files.

Regarding the GEOLOCATION estimates, they can be provided in different formats, depending on the source. CLS provides geolocation in HTML format (also in XML, but HTML was used for the purpose of populating the database) and the geolocations processed by Wildlife Computers GPE3 software are in CSV format. In theory, there is an infinite number of possible estimates, depending on the mathematical model used for its calculation and the parameters used in the model (maximum speed of animal). Currently all the tags that had transmitted more than few days were processed by CLS using the Grid Filter, but the older ones had also been processed by CLS using the Kalman Filter. In addition, some tags were processed by GPE3 model, and some of them with multiple runs, each one with different value of the parameters. All these geolocation sets are kept in the data depository. One of the key differences of the geolocations provided by CLS and GPE3, apart from the format, is that CLS provides one geolocation per day, while GPE3 provides several.

### *2.2 Data cleaning and formatting*

The first process in creating the database was assembling the data, which consisted in combining the datasets of the same type produced by different tags. Since each record has a tag ID, the records of all tags can be piled together inside the appropriate table. Each type of file was put in its own table. The variables (columns) of different datasets didn't always follow the same order, but they were matched by the column name. In cases where datasets didn't have the same number of variables, the larger one was preserved and the corresponding values in smaller dataset were left empty. This method allowed for creating tables for all type of files except for the geolocations.

During the binding process, the data was formatted in order to be comparable and the type of data was changed where necessary. One of the greatest challenges encountered was formatting the dates and times, because they came in several formats. Given that some CSV files were received several years ago, some of them have been opened in Excel and saved, which provoked that some dates got automatically formatted according to the local settings. For example, here are some of the different formats of the dates and times that were found in the datasets:

15:59 3-ene-2013; 01/23/2013 15:59:00; 15:59:00 23-jan-2013; 23-01-23 15:59:00; 01/23/2013 15:59; 23-Jan-2013 15:59:00; 23-ene-2013 15:59:00. All dates and times were changed to the same format.

The files HISTOS required special treatment before the merging. Since that file contained data of 4 types (time at depth, time at depth limits, time at temperature, time at temperature limits), each type was treated separately and 4 tables were produced.

Other challenge was unifying the geolocation datasets, which came in several formats, even from the same source and consisted of multiple sets for the same tag. Regarding the datasets produces by GPE3, they have been merged and a single geolocation per day was calculated as a daily mean value. As regards the datasets delivered by CLS, they came in three different formats and neither included the tag ID as a variable in the table. These datasets were also assembled and new variables were added for the tag ID, the processing method and the date of processing and populated with the corresponding values. Later, all datasets produced by GPE3 and CLS were merged according to the common fields. This table therefore includes all the sets of daily geolocations that have ever been calculated by any model or by any run. At the end, the additional table was created, which was comprised of only one set of daily geolocations for each tag, which was considered as the best one. The criteria for selecting the best one was the method used for calculating the geolocations and the date of the processing. The priority was first given to the processing method (CLS Grid Filter is considered better than CLS Kalman Filter and both are considered better than GPE3) and then to the processing date (more recent date is considered better).

### *2.3 Creation of the database*

All processes on reading, cleaning, formatting and assembling the data was done by R. In this way, a series of tables were produced, which were afterward written to PostgreSQL database. A special attention was made to renaming the columns in order to make them readable by PostgreSQL. Almost all interaction with the database was done through R, using the package "RPostgreSQL", including reading, writing and joining the tables, and other queries.

The database incorporates various types of tables: the original ones that comprehend all the data from the datasets and are not formatted and the simplified ones that contain only the relevant variables and are modified and formatted adequately in order to facilitate later work. In most of them it was necessary to add a column for date only (without time) to be able to combine it later with geolocations which are in the same format. Additionally, the tables with information on depth and temperature were joined to the geolocation table using the common fields for tag ID and date. In the resulting combined tables, each value of temperature or depth was associated with the corresponding geolocation. The combined tables were also written in the database. In the simplified and combined tables, the duplicates were removed and they were ordered by tag ID and date.

In addition, there was a help table created, which defines the latest relevant dates for each tag. It was done in order to remove all data which correspond to period before the pop-off when tag was not attached to the fish any more (or the fish was dead), but it was still keeping the data. As already pointed out by Tensek *et al.* (2017), it is very important to remove this type of data before performing the analysis, in order not to create bios. The latest relevant dates were determined manually, by carefully checking each dataset and finding the approximate moment when the tag detached or the fish died. In the tables that combine temperature and depth values with geolocations all these non-relevant data were removed.

Data in the database can be updated with new data either for the new tag, or with more detailed datasets for the tag that already exists in the database, or with new geolocation estimates. The procedure for entering new data was developed in a way that the new data is appended to the existing tables, but then the tables are again checked for the duplicates and they are removed. The table with best geolocation sets is calculated again and the new table overwrites the old one. All the combined table which are based on best geolocation also need to be updated and they are produced again and the new ones overwrite the old ones.

### *2.4 Creation of the Shiny Application*

In order to develop an interactive and practical tool for visualisation of the tags tracks and temperature and depth plots, the simple Shiny application was developed in R. The application automatically connects to PostgreSQL database and, depending on the query, reads the respective data subset. Currently the database is hosted in the local server and the Shiny application only runs locally on the same engine.

The Shiny application user interface incorporates a drop down menu from which a user can select the ID number of tag. The interactive content is presented in four tabs, one for each type of information (**Figure 1.)**.

The "Track" tab (**Figure 2**.) incorporates an interactive zoomable map developed by R package "leaflet" which shows daily estimated geolocations in different colours, depending on the month. The point of deployment is marked and each geolocation point displays the date when the mouse is hovered over it.

The "Temp&Depth" tab (**Figure 3.**) accommodates 4 plots, for temperature and depth values from time series and histogram data. Temperature and depth series are represented by linear plots which also comprise a smooth fitted line. Histogram data are represented by tile plots and a colour of the tile corresponds to the percentage of time spent within the bin in a given period.

The "MinMaxDepth" tap (**Figure 4**.) contains a linear plot for minimum and maximum daily depth. It also shows the depth of the bottom of the sea in a corresponding geolocation. The data on the depths of the bottom are downloaded from NOAA using the "marmap" package.

Finally, the "Argos" tab (**Figure 5**.) incorporates a leaflet map which shows Argos post-release locations. The pop-off point is market and each point displays the date when hovered over with a mouse.

## 3. Results

The code that was developed in R serves for compiling the datasets produced by Wildlife Computers and those provided by CLS. Data are automatically cleaned, formatted and written to the appropriate table in the PostgreSQL database. All records on temperature and depth are associated with the corresponding geolocation. This code functions for miniPATs and MK10 type of tags produced by Wildlife Computers since, at least, 2011 and is compatible with 3 formats of outputs produced by CLS from 2011 onward. The code for entering new data in the database and updating the existing one was also developed. Although the code was developed for PostgreSQL, it can be easily modified to connect to other type of SQL database.

The database that is created in PosgresSQL server can be accessed either directly, or through R. Querying the database and any modification or update can be done through R as well.

Finally, the Shiny application was developed for the visualisation of the tag track, Argos track, temperature and depth series, histograms of time at temperature and time at depth and minimum and maximum depth in relation to the sea bottom. The application is interactive and it allows selecting a tag ID from the predefined list of IDs.

## 4. Discussion

Currently all the datasets of the electronic tags deployed by the ICCAT GBYP are stored in a single repository on the ICCAT Secretariat server. The access to the files is allowed to some Secretariat staff only and the files are provided to the external experts on request. The process of selecting the files to be transferred to the experts has to be done manually. The access to the dataset is also not easy because the depository has difficult and not consistent folder structure with various subfolders and any analysis is impossible before previously investing a lot of time in cleaning and preparing the data. The lack of structure makes compilation a time and energy consuming task. Having all the electronic tagging data already formatted and clean in a relational database permits easy extraction of any data subset according to the criteria set in the query. Since no effort needs to be involved in previous preparation of data, any further analysis becomes more accessible as it is a straightforward process.

Currently the GBYP electronic tag database is kept on the personal computer and it cannot be remotely connected. The Shiny application is also only locally run. The possibility of the integration of the electronic tag database within the ICCAT Secretariat has already been discussed, but it has to be further evaluated. Currently the database is hosted locally on the PostgreSQL server, but it can easily be migrated to other SQL server and the R code for database populating and updating can be easily modified to support it.

The possibility of putting the database and the Shiny application online and that way available to the external users has not been discussed yet. In any case, it can be relatively easily done, if decided so, and the access can be granted only to trusted users.

**Bibliography**

Di Natale A., Idrissi M., 2015. Review of the ICCAT GBYP tagging activities 2010-2014. Collect. Vol. Sci. Pap. ICCAT, 71(3): 1125-1143.

Di Natale, A., Tensek, S. and Pagá García, A., 2016. Preliminary information about the ICCAT GBYP tagging activities in Phase 5. Collect. Vol. Sci. Pap. ICCAT, 72(6): 1589-1613.

Di Natale, A., Tensek, S. and Pagá García, A., 2017. ICCAT Atlantic-Wide Research Programme For Bluefin Tuna (GBYP) Activity Report for the Last Part of Phase 5 and the First Part of Phase 6 (2015-2016). Collect. Vol. Sci. Pap. ICCAT, 73(7): 2424-2503.

Tensek, S., Di Natale, A., Pagá García, A., 2017. ICCAT GBYP PSAT Tagging: The First Five Years. Collect. Vol. Sci. Pap. ICCAT, 73(6): 2058-2073.

Tensek, S., Pagá García, A., Di Natale, A. (in press). ICCAT GBYP Tagging Activities in Phase 6. Document SCRS/2017/042: 12 p.

**Table 1.** Types of electronic tags datasets received by ICCAT GBYP including their description, format and possible source (only the files which were used for the creation of the database are listed here)

| TYPE | DESCRIPTION | FORMAT | POSSIBLE SOURCE | | |
|---|---|---|---|---|---|
| | | | Downloaded from WC Portal | Provided by CLS | Provided by WC Factory (recovered tag) |
| Sensor data (observations) | Readings of the tag sensors (temperature, depth) while the tag was deployed. Locally saved in the tag and after popping off, transmitted to the satellite. | Histos.csv | data with gaps | | complete data |
| | | MinMaxDepth.csv | data with gaps | | complete data |
| | | MixLayer.csv | data with gaps | | complete data |
| | | Series.csv | data with gaps | | complete data |
| | | SeriesRange.csv | data with gaps | | complete data |
| | | SST.csv | data with gaps | | complete data |
| Geolocation estimates = track | Calculated using some state-space model | GPE3.csv | X | | - |
| | | .html (in 3 different formats) | | X | - |
| Argos locations | Post-release locations transmitted through satellite | Argos.csv | X | | - |
| Other | Data used by GPE3 software for calculating geolocations | LigthLoc.csv | data with gaps | | complete data |
| | General info about deployment | Summary.csv | data with gaps | | complete data |

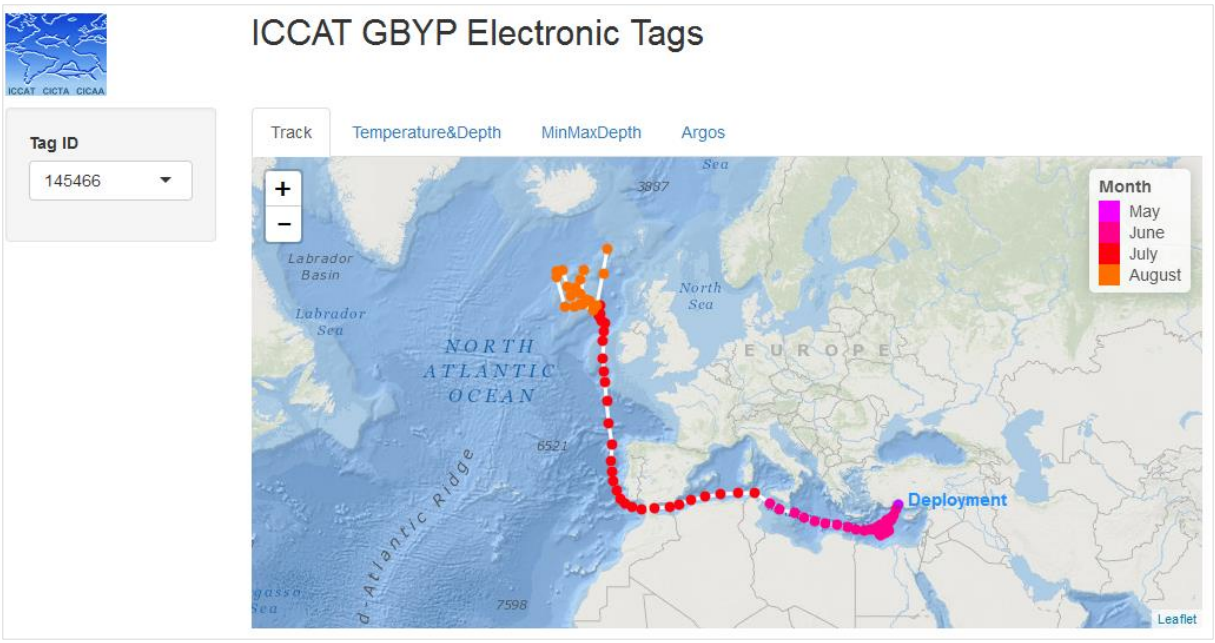


**Figure 1.** The interface of the ICCAT GBYP Electronic Tags Shiny Application

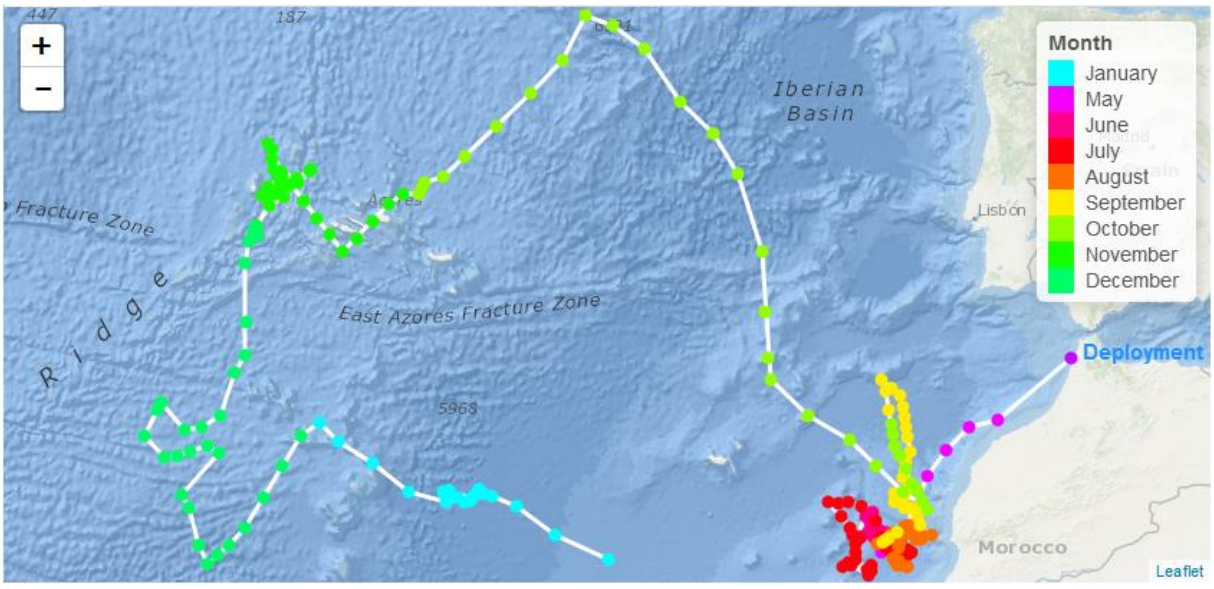


**Figure 2.** Track of the electronic tag (114009) plotted by the ICCAT GBYP Electronic Tags Shiny Application

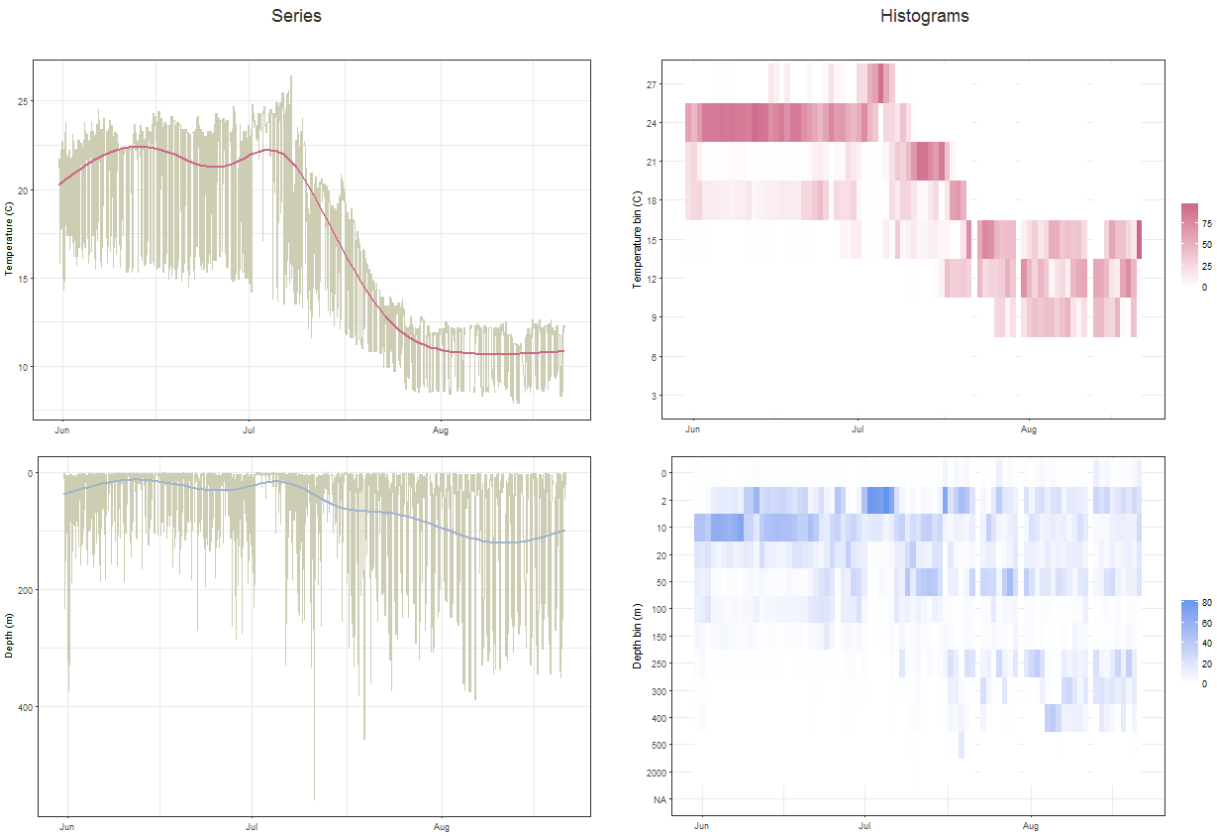


**Figure 3.** Temperature and depth series and time at temperature and time at depth of the electronic tag (145466) plotted by the ICCAT GBYP Electronic Tags Shiny Application

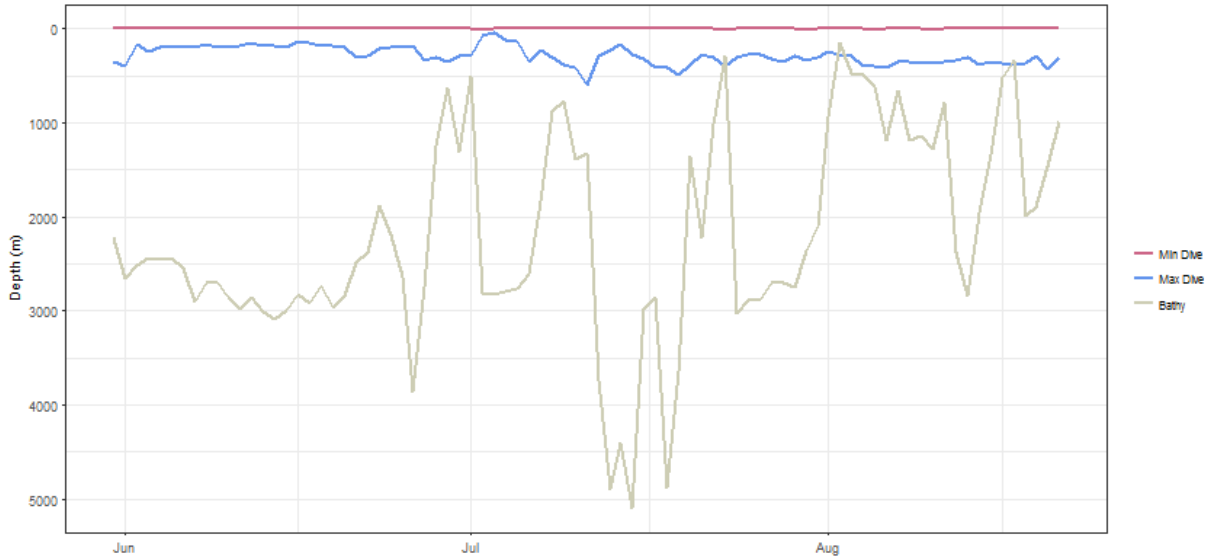


**Figure 4.** Minimum and maximum depth of the electronic tag (145466), including the depth of the sea bottom, plotted by the ICCAT GBYP Electronic Tags Shiny Application

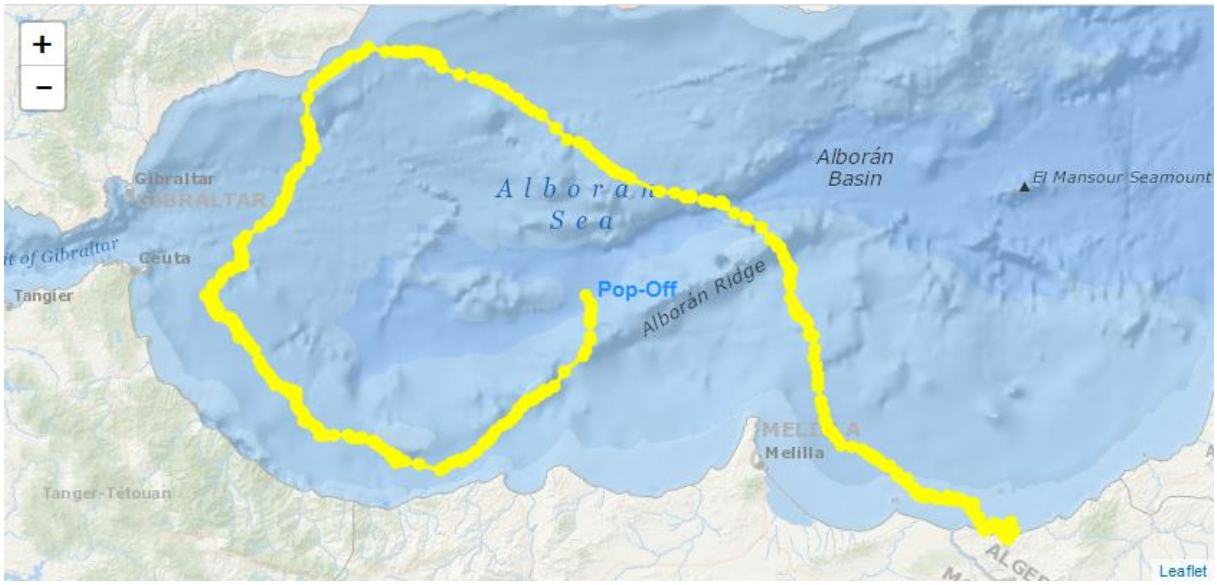


**Figure 5.** Argos post-release track of the electronic tag (150294) plotted by the ICCAT GBYP Electronic Tags Shiny Application