Eddy Detection System Manual

Version: 1.0

Date: 2023-12-22

Author: Roderick Bakker

Correspondence: roderick.bakker@bios.asu.edu

Contents

General description and instructions

This manual accompanies the R scripts for the historical eddy detection system. Descriptions of the data processing and analysis are provided for each script.

Note: denotes a part of the analysis which is essential to proper data selection but won't hinder the running of the script.

Important: denotes a part of the analysis which is essential for operation of the scripts, failure to follow the instructions in these statements can lead to errors in the script, or downstream scripts which rely on the output of the current script.

[] denotes a variable inside an R script.

{} denotes a function inside an R script.

"" denotes a file, either for import or export.

Individual scripts should be opened from the RStudio IDE after opening the "historical_eddy_detection_system.R" project file.

1) Geo-Temporal Filter

a) Water Profiles: "1_a_water_profiles.R"

Data availability: https://www.bco-dmo.org/dataset/3918

BATS CTD data can be downloaded from the above link. Downstream files provided with the source code are from the v005 version downloaded on 2023-09-12.

When using data from other hydrographic records care should be taken that the format is the same as the BATS data used here, being in a long matrix/table format.

i. The BATS CTD data is processed in a way that allows the output files to be used for downstream calculations as well as for the shiny application.
ii. Potential densities are calculated at each available depth after which the dates are formatted into a more human readable format.
iii. Latitude and longitude filters are applied to extract the highest amount of casts in the smallest possible region. A QC plot of station coordinates during occupation is provided to help determine the best filter settings. A depth filter is also applied to extract the depths of interest and reduce the amount of data for downstream calculations.

Note: The depth filter should ideally span the entire local thermocline.

iv. The 2db binned data is binned every 10m and averaged to reduce the noise and amount of data. Binning is done on depth and not pressure because the downstream eddy analysis is performed in depth space.

note: The gc() command can be run after this procedure to help free up working memory but is not essential to the script.

v. To reduce the amount of irrelevant data, casts with a maximum depth of less than 500m are dropped from the data.

Note: This depth needs to be adjusted depending on the position of the local thermocline.

vi. Climatic averages along depth are calculated from the filtered data for each month. Note: These are used as a crude reference in the shiny application by which to judge shifts in water masses.

vii. QC plots are provided for the climatic monthly averages and comparisons of these averages to the individual casts from that month (e.g., comparison of climatic mean from September to casts taken in September).

viii. The data files and QC plots are exported to their appropriate folders using parallel mapping. A total of 4 files are exported. The 1st file "1_a_data_profiles.csv" contains the formatted, but unfiltered data prior to application of the binning procedure. The 2nd file "1_a_data_profiles_filt.csv" has been filtered for latitude, longitude and depth and binned to 10m intervals. The 3rd file "1_a_data_profiles_filt_select.csv" is the same as the second file, but the shallow casts (<500m) have been removed. The 4th file "1_a_data_profiles_means.csv" contains the climatic monthly averages.

ix. The file "1a_data_profiles_filt_select.csv" should be moved to the input folder of 1_geo_temporal_filter. This will be the file that supplies the coordinates for station occupation to be overlayed with the eddy radii.

x. The files "1a_data_profiles_filt_select.csv" and "1a_data_profiles_means.csv" should be moved to the shiny data folder.

b) Global Eddy Record: "1_b_global_eddy_record.R"

Data availability: https://www.aviso.altimetry.fr/en/my-aviso-plus.html

The .nc files containing the cyclonic and anticyclonic eddy tracks can be downloaded from the AVISO website. An account is required (registration is free, but the authentication process can take several days). Downstream files provided with the source code are derived from the META3.2 DT (issue: 1.0, publication date: 2022-02-15), downloaded on 2022-05-25.

xi. The .nc files are opened in R using the ncdf4 package. A variety of variables is available in the .nc files. In this script only the variables required for the downstream analysis are extracted from the files [eddy_variables_select], but know that many more are available [eddy_variables_options].

xii. Latitude and longitude filters are set for broad scale filtering to reduce the amount of data. These filters are applied during the data extraction procedure.

xiii. To aid in downstream analysis of the eddy behaviours their origins (first detection by satellite) are extracted from the .nc data files and appended to the broad scale filtered data.

xiv. The eddy tracks data is then formatted in the same style as the water profiles data. **Important:** Longitudes are set to 180°E-W, instead of the original 360°W, to resemble the coordinates in the water profiles. If the water profiles are in the 360°W format this transformation should be omitted from the script.

xv. QC plots are provided for the amounts and origins of the eddies.

xvi. Two data files are exported. The 1st contains the eddy tracks after the broad scale filtering. The 2nd file contains summary statistics of the eddies, after broad scale filtering.

xvii. The 1st file "1_b_eddy_tracks_full" should be moved to both the input folder of 1_geo_temporal_filter and the shiny data folder.

c) Geo-Temporal Filter: "1_geo_temporal_filter.R"

xviii. The geo-temporal filter uses "1_a_data_profiles_filt_select.csv" and "1_b_eddy_tracks_full.csv" files as input.

xix. To reduce the amount of irrelevant eddies (lifespan < 5 days) the total lifespans (in days) of the eddies is calculated. These will be filtered out later in the script.

xx. The mean coordinates of all casts in "1_a_data_profiles_filt_select.csv" are used as a reference point around which to filter for eddies potentially interacting with the station. In this script eddies with centers at 1° distance from the mean station coordinates are retained.

xxi. QC plots of the eddies are provided after the geographical filter.

xxii. The amount of days spent by the eddies within 1° of the station are calculated. Eddies which spent less than 5 days within 1° of the station are dropped from the data, as well as eddies with a lifespan of less than 5 days, as calculated in step 1cii.

xxiii. The geo-temporal filter is applied by calculating the distance between the station coordinates and the eddy centers on the days where an eddy center was within 1° of the station during occupation. Eddies were retained if the radius of the eddy was greater than the distance of the station coordinates to the eddy center ([effective_radius] > [distance]).

xxiv. QC plots of the eddies are provided after the geo-temporal filter.

xxv. Files are exported with the data after application of the broad scale filter, artefact reduction and geo-temporal filter. For each of these steps 2 files are exported, 1 file containing the eddy tracks and 1 file containing summary statistics, for a total of 6 files.

xxvi. The file "1_data_eddy_initial_consensus.csv" should be moved to the 3_k_means input folder and the "1_data_eddy_filt.csv" the shiny data folder.

2) Bootstrapping: "2_bootstrapping.R"

i. The bootstrapping procedure uses the "1_a_data_profiles_filt_select.csv" file as input.

ii. In the first step desired variables for bootstrapping are selected.

Note: Here oxygen values were included as they were available for the entire dataset, but they are not essential to any downstream analysis.

iii. The desired depths are selected at discrete intervals, here subsurface depths (100-300m) and 1000m depth were included, but they are not essential to the detection system. Note: For increased processing speed the amount of depths selected should be limited to the minimum amount required to run the downstream analysis (e.g., 3-7 depths along the thermocline) as the bootstrapping procedure is computationally intensive.

iv. Average values are calculated of all casts for every month in a single year at each discrete depth (e.g., average of all casts from September 2020 at 100, 200 & 300m depth). Note: Should you want to run the bootstrapping procedure on individual casts this step can be omitted, but depending on the amount of data this could significantly increase the CPU time required for the calculations.

v. An index column is added for the sliding window (running mean) method to operate on, and the data is nested into a list for the sliding window to iterate through.

vi. The {anomaly_calculation} function defines how the bootstrapping will operate. A random sample (monthly average of profiles) is selected from the data, this random sample will then be compared to the mean of 7 different, randomly chosen samples.

vii. For every list item in [profiles_nested] the function is replicated n times (default = 150), after which the sliding window is moved 5 years and the process is repeated.

viii. Summaries of minimum, maximum, median and mean bootstrap outcomes are provided per unique combination of year, month, variable and depth.

ix. A QC-plot for all summary values is provided after bootstrapping.

x. A second QC-plot is generated to check for proper saturation of the bootstrap distributions. Here 3 years (2007, 2008 & 2009) from a single month (June) are extracted for comparison of their distributions. 2007 corresponds to a period with anticyclonic activity, 2008 cyclonic and 2009 a non-eddy period.

xi. If the distributions show saturation (no singletons on outer edges of distributions) and a relatively normal distribution, the plots and data files can be exported.
Note: The distributions don't have to be perfectly normal as the median bootstrap outcomes are used in downstream analysis, but if the curves are excessively skewed further investigation may be warranted.

xii. Two files are exported, the first file "2_data_bootstrap_outcomes.csv" contains the raw bootstrap outcomes. The second file "2_data_bootstrap_outcomes_summary.csv" contains the summarized values, this file should be moved to the 3_k_means, 4_model_validation and 5_model_application input folder, as well as the shiny data folder.

3) K-means Clustering: "3_k_means.R"

i. The K-means clustering uses the "1_data_eddy_initial_consensus.csv" and "2_data_bootstrap_outcomes_summary.csv" files as input.

ii. The analysis is performed separately for each month to account for seasonality and reduce the amount of observations to be reviewed at once.

iii. The median bootstrap outcomes and selected depths (here, 500-800m) are filtered from the summarized bootstrap outcomes. After which they are joined to the initial eddy consensus. **Important:** The input depths should be adjusted here according to user preference.

iv. The month on which the K-means clustering will be applied is set as [month_filter]. This value will have to be adjusted to run the script for each individual month (1-12). Note: All output files will be appended with the value of [month_filter] to keep track of which month the output files originated from.

v. The data for the K-means clustering is extracted from the formatted data as [k_input]. The years to which the data corresponds is extracted to a separate variable as [k_years]. **Important:** The [k_input] and [k_years] variables need to be arranged by year and formatted in exactly the same way. This is imperative to ensure that [k_years] will line up correctly with the cluster outcomes.

vi. The centers argument of the {kmeans} function is set to 7, but can be varied for better data resolution.

Note: For our data 5-9 clusters worked best, but optimal settings will depend on data spread.

vii. The K-means cluster outcomes are plotted in temperature-salinity (TS) plots facetted by depth [p_1_ts_median] and in a principal component analysis (PCA) plot [p_2_PCA]. These plots can be reviewed individually before they are combined for export as [p_3_ts_PCA].

viii. The clusters are exported to the output folder and the combined TS-PCA plot to the QC_plots folder.

ix. The monthly clusters can be combined using the "3_k_means_cluster_concatentation.R" script. This will combine the monthly .csv cluster files to "3_data_eddies_clusters_concatenated.csv". Through manual revision this file can be used to generate an intermediate consensus.

x. Here a copy was saved as "3_data_eddies_clusters_intermediary.csv" and values in the [type] column were changed manually to match whether an eddy was present or not. If not already available, the [track] column was manually supplanted with the relevant eddy id where possible.

xi. Once revised, the "3_data_eddies_clusters_intermediary.csv" should be moved to the 4_model_validation input folder.

4) Model Validation: "4_model_validation.R"

I. The model validation uses the "2_data_bootstrap_outcomes.csv" and "3_data_eddies_clusters_intermediary.csv" files.

Note: This script can also be used for model validation of the final eddy census.

II. Relevant data is extracted from the bootstrap outcomes and joined to the intermediary census. Here we used the median, minimum and maximum values from 400-900m depth. The minimum and maximum values act as confidence intervals to help the pattern recognition discern between eddy and non-eddy states.

III. The analysis is performed using a random forest pattern recognition model with repeated cross-fold validation. The cross-fold validation was done with 10 folds (data chunk size) and 10 repeats (iterations).

Note: More repeats may lead to better model accuracy, but we found the increase in accuracy started to saturate after 8 repeats.

IV. The settings for the pattern recognition model are defined in the {eddy_identifier_training} function. The only arguments for this function are the input data and the model predictors.

Note: For the input data each variable, depth and bootstrap outcome type (minimum, maximum & median) should be placed in their own column. This should already be the case after the {pivot_wider} function has been applied, but it is worth checking before moving along with the analysis. The predictors are taken from the input data columns.

V. In the validation stage the data is separated in training and validation data (90% : 10%). The model is trained with 90% of the data and applied to the remaining 10% of data. This process is repeated n times (here n = 200).

VI. Statistics for each iteration are provided. These statistics are also summarized to average values of all iterations.

   i. [data_model_stats] provides average statistics on the model accuracy, false-negatives, false-positives, and probability scores.

   ii. [data_model_errors] provides an overview of each false-negative and false-positive observation. The summarized version provides a count (n) of how often this particular error was encountered.

   Note: the summary of model errors is a good file to use for revision of the intermediary to the final census.

VII. The statistics and errors for each iteration and their summary are exported to the output folder. Either to the stats_intermediary or stats_final folder, depending on which version of the eddy census was used.

VIII. After revision of the intermediary census the final census (here “4_data_eddies_final.csv”) should be moved to the 5_model_application input folder.

5) Model application: “5_model_application.R”

IX. The model validation uses “2_data_bootstrap_outcomes.csv” and “4_data_eddies_final.csv” files.

X. The model training works in the same way as in the previous script, except that 100% of the data is used to train to the data.

   Note: Here data from Jan-1993 to Dec-2020 was used, but this can be extended past 2020 if data is available.

XI. After model training eddies are predicted from data before Jan-1993.

XII. QC-plots are provided of temperature anomalies over time for the predicted eddies and counts of months with eddy activity for both predicted (<1993) and satellite observed eddies.

XIII. Two files are exported, one with a summary of the predicted eddies and one with the predicted eddies and their anomalies along depth.