

mStatGraph Matlab Package

Plotting and Statistical Analysis for
Oceanographers, Meteorologists and
Ecologists

USER MANUAL

By

Humberto L. Varona

Version 1.7

May, 2023

Software to analyze, compare and validate analysis and reanalysis datasets with an observed dataset (DSCompare).

Overview

Software to carry out statistical analysis, plot figures and maps of spatial distribution of physical and biogeochemical parameters, calculation of the parameters of the carbonate system of rivers and of MLD, ILD and BLT.

Version

1.7

Release date

April, 15th 2023

License

MIT

DOI

[10.5281/zenodo.8152683](https://doi.org/10.5281/zenodo.8152683)

Download URL

<https://zenodo.org/deposit/8152683>

Cite as

Varona, Humberto L., Noriega, C., Araujo, J., Lira, S. M. A., Araujo, M., & Hernandez F. (2023). Plotting and Statistical Analysis for Oceanographers, Meteorologists and Ecologists (mStatGraph). Version 1.7. Zenodo. <https://doi.org/10.5281/zenodo.8152683>

How to install

Matlab 2021b compatible software

- Open Matlab.
- Go to APP tab.

- Click on the "Install App" button.
- Select the mStatGraph.mlappinstall file.
- In the Install dialog click on the "Install" button.

Create a directory and copy into it the following databases: etopo_2.mat, gshhs_f_coast.mat, gshhs_f_rivers.mat, and wdb_f_borders.mat.

How to run

Type in the Matlab command window:

```
>> mStatGraph<Enter>
```

or find mStatGraph in the APP tab of Matlab.

Statistical tests

- *The minimum is the smallest value observed in a dataset. It represents the lowest point or the lowest measurement recorded.*
- *The maximum is the largest value observed in a dataset. It represents the highest point or the highest measurement recorded.*
- *Amplitude refers to the range between the minimum and maximum values in a dataset. It provides a measure of the spread or variability of the data.*
- *The mean, also known as the average, is the sum of all values in a dataset divided by the total number of observations. It is a measure of central tendency that represents the typical or average value of the data.*
- *The trimmed median is a robust measure of central tendency that provides a compromise between the median and mean. It is calculated by taking the average of the median and the midrange (the average of the minimum and maximum values).*
- *The standard deviation is a measure of the dispersion or variability of the data points around the mean. It quantifies how much the individual data points deviate from the mean. A higher standard deviation indicates greater variability, while a lower standard deviation indicates less variability.*
- *The standard error measures the variability of the sample mean. It represents the standard deviation of the sample means that would be obtained if multiple samples were taken from the same population. The standard error provides an estimate of the precision or reliability of the sample mean as an estimate of the population mean.*
- *Variance is a measure of the dispersion of data points around the mean. It is calculated as the average of the squared differences between each data point and the mean. A*

higher variance indicates greater variability in the data, while a lower variance indicates less variability.

- *The coefficient of variation (CV) is a relative measure of variability that expresses the standard deviation as a percentage of the mean. It is calculated by dividing the standard deviation by the mean and multiplying by 100. The CV allows for the comparison of the variability of different datasets with varying means.*
- *The median is the middle value in a dataset when the data is arranged in ascending or descending order. It represents the value that separates the higher half from the lower half of the data. The median is a measure of central tendency that is not affected by extreme values.*
- *The mode is the value or values that occur most frequently in a dataset. It represents the peak or the most common value(s) in the distribution.*
- *The mode frequency is the number of times the mode value(s) appear in a dataset. It indicates the count or frequency of the most frequently occurring value(s).*
- *Skewness measures the asymmetry of a distribution. It quantifies the extent to which the data is skewed to the left or right. Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side.*
- *Kurtosis measures the degree of peakedness or flatness of a distribution. It quantifies the tails and outliers in the data. Positive kurtosis indicates a relatively peaked distribution with heavier tails, while negative kurtosis indicates a flatter distribution with lighter tails.*
- *Quartiles divide a dataset into four equal parts, each containing 25% of the data. The first quartile (Q1) represents the 25th percentile, the second quartile (Q2) represents the median, and the third quartile (Q3) represents the 75th percentile.*
- *The interquartile range (IQR) is a measure of statistical dispersion. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The IQR represents the range of the middle 50% of the data and provides a measure of the spread that is less influenced by extreme values.*
- *Percentiles are values that divide a dataset into hundred equal parts. The Pth percentile represents the value below which a given percentage of the data falls. For example, the 25th percentile represents the value below which 25% of the data falls.*
- *Mann-Kendall test is a non-parametric test used to assess the presence of trends in a time series data. This test is widely employed in various fields, including oceanography, to analyze changes or variations in variables such as water temperature, salinity, or sea level. The test compares the rank differences between successive data pairs and calculates the test statistic known as Kendall's Tau. A positive value of Tau indicates an increasing trend, while a negative value indicates a decreasing trend.*
- *Mann-Whitney test, also known as the Mann-Whitney U-test, is a non-parametric test used to compare two independent samples and determine if there are significant*



differences between them. In oceanography, this test could be used to compare two groups of data, such as the concentrations of a specific chemical compound in two different areas or time periods. The test is based on comparing the rankings of the data in both groups and calculates a test statistic called U . If the U value is sufficiently small or large, it can be concluded that there are significant differences between the samples.

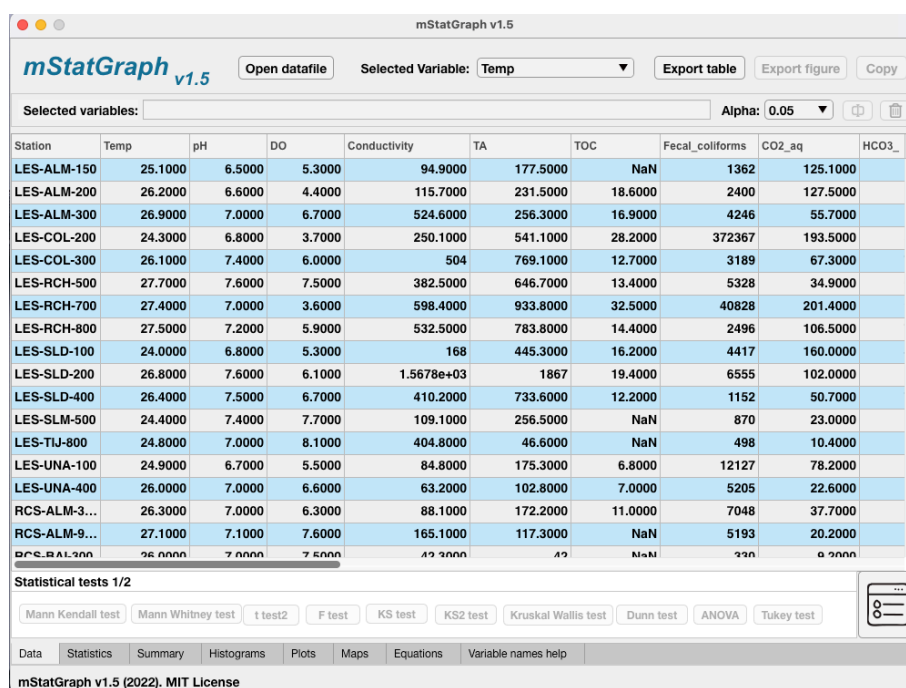
- Two-sample t -test is a parametric test used to compare the means of two independent groups. In oceanography, this test could be applied to compare the mean of a physical or chemical variable at two different locations or time points. The t -test compares the differences between the means of the groups, taking into account the variability within the groups. Upon conducting the test, a t -test statistic and a p -value are obtained. If the p -value is less than a predefined threshold (usually 0.05), it is considered that there are significant differences between the means of the groups.
- The Kolmogorov-Smirnov one-sample test is a non-parametric test used to evaluate if a sample of data follows a specific probability distribution. In oceanography, this test could be used to assess whether data from a variable, such as water temperature, fits a normal distribution or another theoretical distribution. The test compares the empirical cumulative distribution function of the data with the theoretical cumulative distribution function and calculates a test statistic based on the maximum absolute difference between the two functions.
- The Kolmogorov-Smirnov two-sample test is a non-parametric test used to compare the distributions of two independent samples. In oceanography, this test could be employed to determine if the distributions of two physical or chemical variables are different, for example, comparing the distributions of oxygen concentration in two different zones or time periods. The test compares the cumulative distribution functions of the two samples and calculates a test statistic based on the maximum absolute difference.
- The Kruskal-Wallis test is a non-parametric test used to compare the medians of three or more independent groups. In oceanography, this test could be utilized to analyze if there are significant differences in a biological or chemical variable among different locations or time periods. The test is based on the ranks of the data and calculates a test statistic known as H . If the H value is sufficiently large and the associated p -value is less than a predetermined threshold, it is concluded that there are significant differences between the medians of the groups.
- The Dunn test is a multiple comparisons procedure used as a post-hoc test after finding significant differences in the Kruskal-Wallis test. It allows identifying which specific groups have significant differences from each other. It provides a correction for multiple tests and calculates adjusted test statistics.
- ANOVA is a parametric test used to compare the means of three or more independent groups. In oceanography, it can be applied to analyze if there are significant differences in a physical or chemical variable among different locations or time periods. The test decomposes the total variability in the data into components due to differences between groups and within groups. It calculates an F -test statistic and an associated p -value. If the p -value is less than a predetermined threshold, it is concluded that there are significant differences between the means of the groups.

- *The Tukey test, also known as the Tukey's Honestly Significant Difference (HSD) test, is used as a post-hoc test after finding significant differences in ANOVA. It allows for the comparison of all possible pairs of means. This test helps identify which specific groups have significantly different means. It provides a correction for multiple comparisons and calculates adjusted test statistics.*

The aforementioned statistical tests play crucial roles in different aspects of data analysis and interpretation within the field of oceanography. They provide researchers with powerful tools to investigate trends, compare groups or distributions, examine relationships, and reduce complex datasets, ultimately contributing to a deeper understanding of the intricate dynamics of the ocean environment.

Operation mode

mStatGraph allows you to load files with ".dat" extension with comma-separated columns and the first row containing the name of the variables. The variables will be displayed in the table located in the "Data" tab. In this tab there are buttons for all the statistical tests that are implemented, as well as options for the computation of parameters of physical oceanography and the carbonate system in rivers. For each statistical test the significance level can be set (0.01, 0.05, and 0.10). The variables (columns) can be deleted with button  and renamed with button .

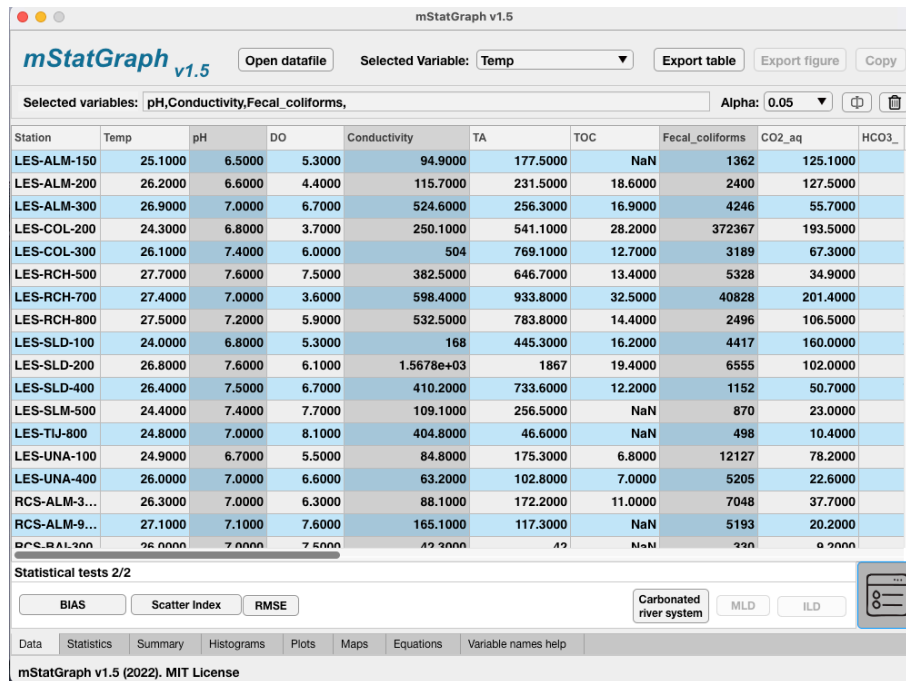


Station	Temp	pH	DO	Conductivity	TA	TOC	Fecal_coliforms	CO2_aq	HCO3_
LES-ALM-150	25.1000	6.5000	5.3000	94.9000	177.5000	NaN	1362	125.1000	
LES-ALM-200	26.2000	6.6000	4.4000	115.7000	231.5000	18.6000	2400	127.5000	
LES-ALM-300	26.9000	7.0000	6.7000	524.6000	256.3000	16.9000	4246	55.7000	
LES-COL-200	24.3000	6.8000	3.7000	250.1000	541.1000	28.2000	372367	193.5000	
LES-COL-300	26.1000	7.4000	6.0000	504	769.1000	12.7000	3189	67.3000	
LES-RCH-500	27.7000	7.6000	7.5000	382.5000	646.7000	13.4000	5328	34.9000	
LES-RCH-700	27.4000	7.0000	3.6000	598.4000	933.8000	32.5000	40828	201.4000	
LES-RCH-800	27.5000	7.2000	5.9000	532.5000	783.8000	14.4000	2496	106.5000	
LES-SLD-100	24.0000	6.8000	5.3000	168	445.3000	16.2000	4417	160.0000	
LES-SLD-200	26.8000	7.6000	6.1000	1.5678e+03	1867	19.4000	6555	102.0000	
LES-SLD-400	26.4000	7.5000	6.7000	410.2000	733.6000	12.2000	1152	50.7000	
LES-SLM-500	24.4000	7.4000	7.7000	109.1000	256.5000	NaN	870	23.0000	
LES-TIJ-800	24.8000	7.0000	8.1000	404.8000	46.6000	NaN	498	10.4000	
LES-UNA-100	24.9000	6.7000	5.5000	84.8000	175.3000	6.8000	12127	78.2000	
LES-UNA-400	26.0000	7.0000	6.6000	63.2000	102.8000	7.0000	5205	22.6000	
RCS-ALM-300	26.3000	7.0000	6.3000	88.1000	172.2000	11.0000	7048	37.7000	
RCS-ALM-900	27.1000	7.1000	7.6000	165.1000	117.3000	NaN	5193	20.2000	
RCS-ALM-200	26.0000	7.0000	7.5000	17.3000	17.3000	NaN	236	0.2000	

The variables can be selected by clicking on each column, for more than one column it will be necessary to combine the click with the Win (command) key for Linux and Windows (MacOS) operating systems. In the editbox "Selected variables" all selected variables will appear separated by comma. The variables Longitude, Latitude, Depth, Time, and Station are not selectable.



This button appears on many of the tabs and allows you to activate more options for statistical tests, computation of parameters or to change the labels on the figures.



When the data is loaded, mStatGraph sorts the variables by name. In the following table we show some of the basic variables.

Variables	Description
Station	Collection station name
Lon Long Longitude	Geographic longitude
Lat Latitude	Geographic latitude
Depth Level	Depth or atmospheric level
Time Date Datetime	Time/Date of sampling or analysis or reanalysis data
Temp Tmp Pottemp Temperaure SSS	Ocean or air temperature
Salt Salinity SSS	Ocean salinity
uCurr vCurr	Velocity components of the current

uWind vWind	Velocity components of the wind
TA	Total alkalinity of ocean or river water
pH PH	pH of ocean or river water

For example, if the file contains geographical longitude and latitude, the "Maps" tab will be activated, allowing the plotting of maps with the sampling stations or the spatial distribution of each parameter; if it contains temperature, pH, TA, the option to compute $\text{CO}_2(\text{aq})$, HCO_3^- , CO_3^{2-} , and DIC will be activated. If the file contains data of current velocity or wind velocity components, current and wind rose figures can be plotted.

mStatGraph v1.5

Open datafile Selected Variable: Temp Export table Export figure Copy

Selected variables: Alpha: 0.05

	Lat	Temp	pH	TA	CO2aq	HCO3	CO32	pCO2	DIC
0033	-12.1485	26.3000	6.8000	101.0000	35.0386	100.9393	0.0303	1.0640e+03	136.0082
5254	-14.2655	25.5000	7.0000	102.6000	22.7127	102.5039	0.0480	674.1406	125.2646
0637	-15.2974	26.0000	7.0000	102.8000	22.5913	102.7027	0.0486	680.1771	125.3427
3419	-10.5981	25.2000	6.8000	104.3000	36.7691	104.2387	0.0306	1.0820e+03	141.0385
6214	-17.2479	26.2000	7.1000	117.0000	20.3599	116.8601	0.0700	616.4868	137.2899
1119	-13.5994	27.1000	7.1000	117.3000	20.1537	117.1571	0.0714	625.9578	137.3822
5413	-16.6168	26.0000	6.9000	123.7000	34.2297	123.6070	0.0465	1.0306e+03	157.8832
9124	-13.2931	26.7000	6.9000	126.1000	34.5463	126.0038	0.0481	1.0610e+03	160.5982
1284	-16.2598	25.3000	6.6000	132.4000	73.8816	132.3508	0.0246	2.1804e+03	206.2571
9214	-16.0091	28.5000	6.9000	134.6000	35.9790	134.4936	0.0532	1.1620e+03	170.5258
5408	-13.1495	26.8000	6.8000	139.0000	47.8804	138.9156	0.0422	1.4746e+03	186.8382
3366	-13.3985	25.7000	6.7000	147.3000	64.9011	147.2305	0.0347	1.9374e+03	212.1663
4430	-14.0145	27.5000	7.2000	148.0000	20.0806	147.7713	0.1144	630.7250	167.9663
2860	-16.3928	26.8000	6.7000	149.4000	64.7959	149.3279	0.0360	1.9956e+03	214.1599
4646	-13.0891	27.2000	6.9000	153.0000	41.6235	152.8821	0.0589	1.2964e+03	194.5646
5873	-16.4134	26.7000	7.0000	154.3000	33.5712	154.1519	0.0741	1.0310e+03	187.7971
5590	-15.9470	26.7000	7.0000	155.8000	33.8975	155.6504	0.0748	1.0410e+03	189.6228
4782	-12.6867	26.2000	7.0000	172.2000	27.6706	172.0261	0.0820	1.1442e+03	200.7075

Statistical tests 2/2

BIAS Scatter index RMSE

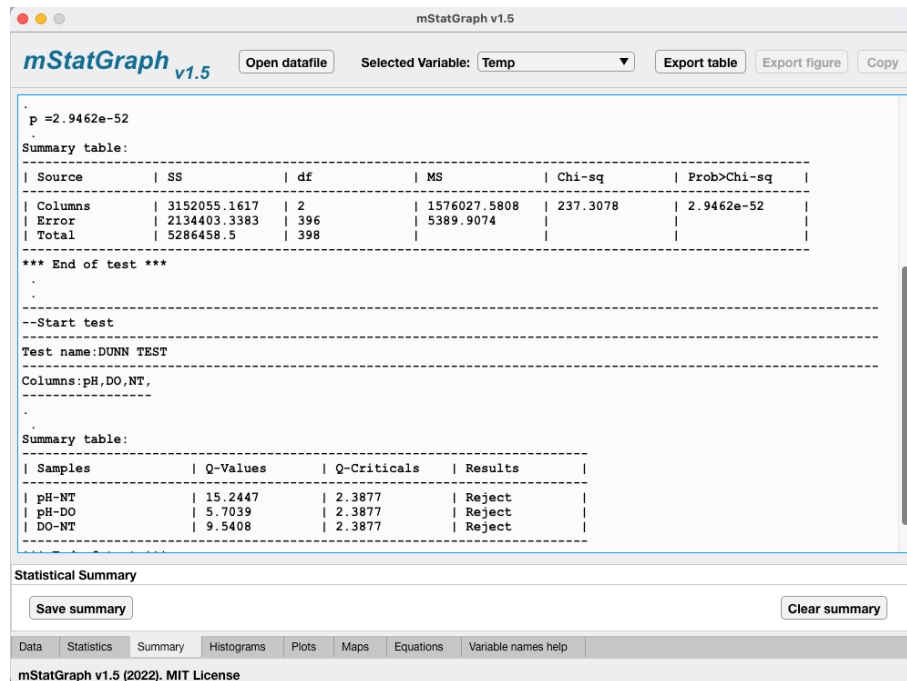
Carbonated river system MLD ILD

Data Statistics Summary Histograms Plots Maps Equations Variable names help

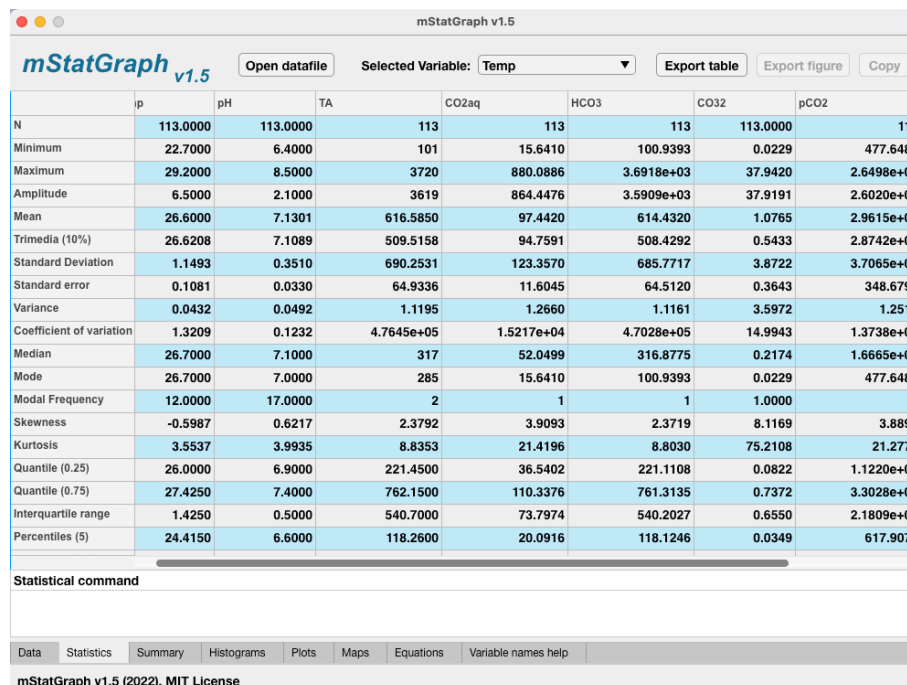
mStatGraph v1.5 (2022). MIT License

To compare two variables, at least two variables (columns) must be selected, thus automatically activating: Mann Whitney test, two-samples t-test, two-samples F-test, RMSE, Scatter index, and bias. If more than two are selected in addition to the above mentioned tests, the Kruskal Wallis test (the Dunn test will only be activated after the Kruskal-Wallis test) and ANOVA (the Tukey test will only be activated after the ANOVA test) will be activated.

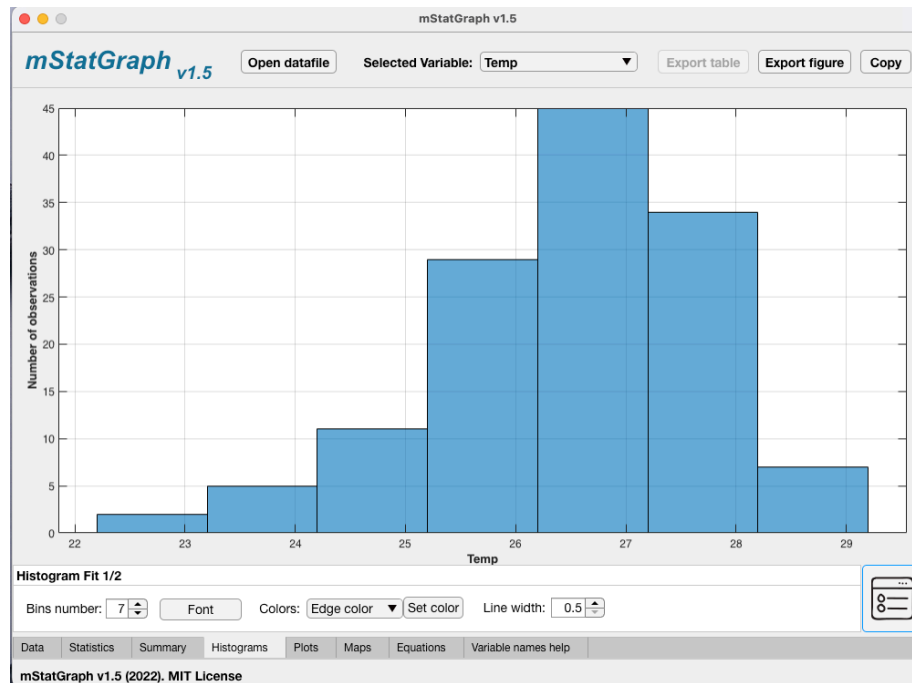
All statistical tests and computation of ocean parameters will be included in a summary found in the "Summary" tab. Summaries can be saved in a ".txt" file. The summary can be cleared using the "Clear summary" button.



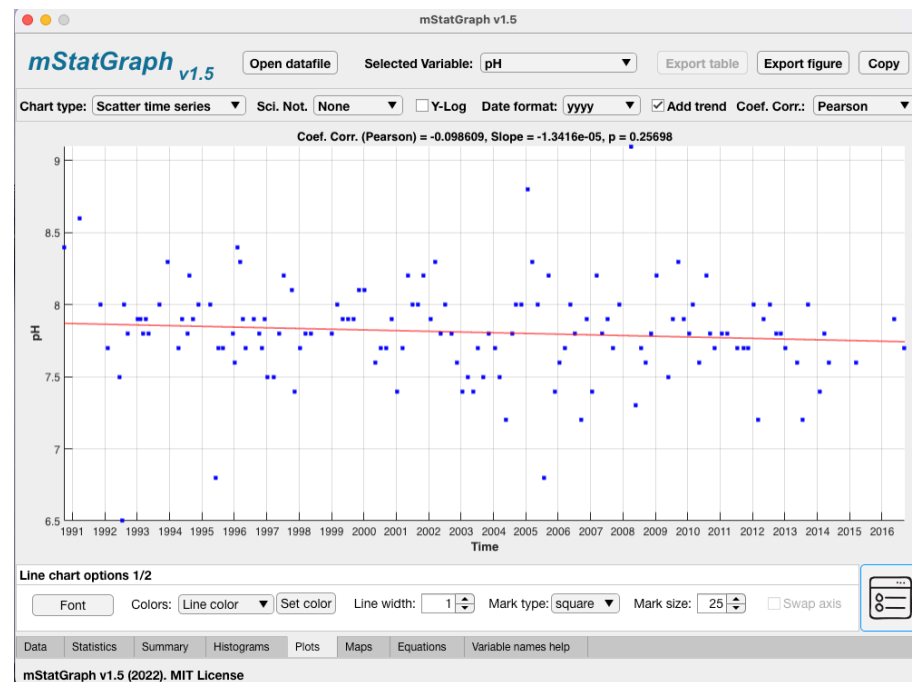
The basic statistics of all variables are computed automatically after the data file is loaded and displayed in the table under the "Statistics" tab. These statistics can be saved in a ".txt" file.



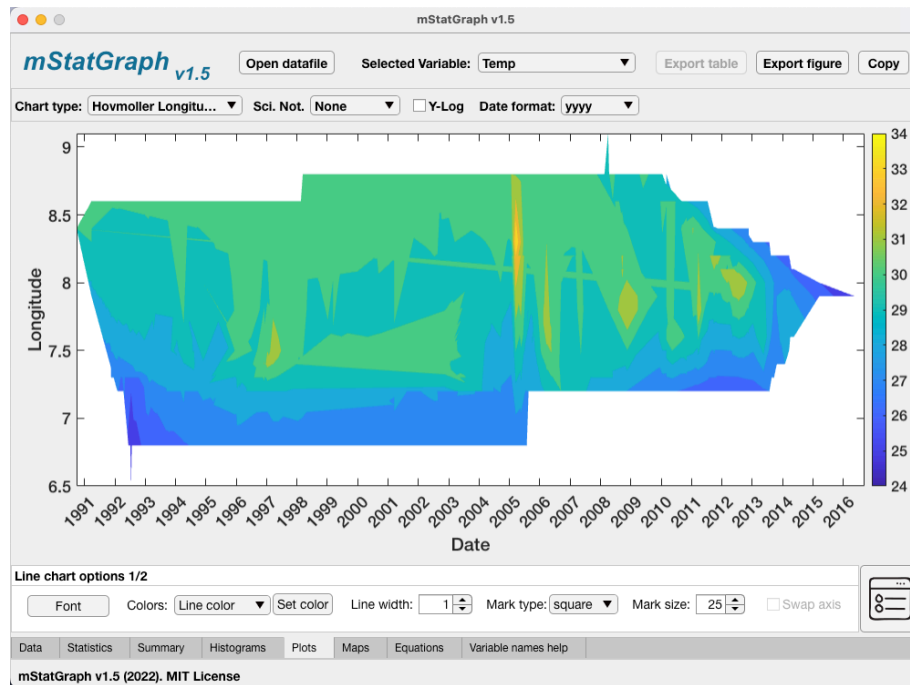
Example of a histogram (Variable selected: Temperature):



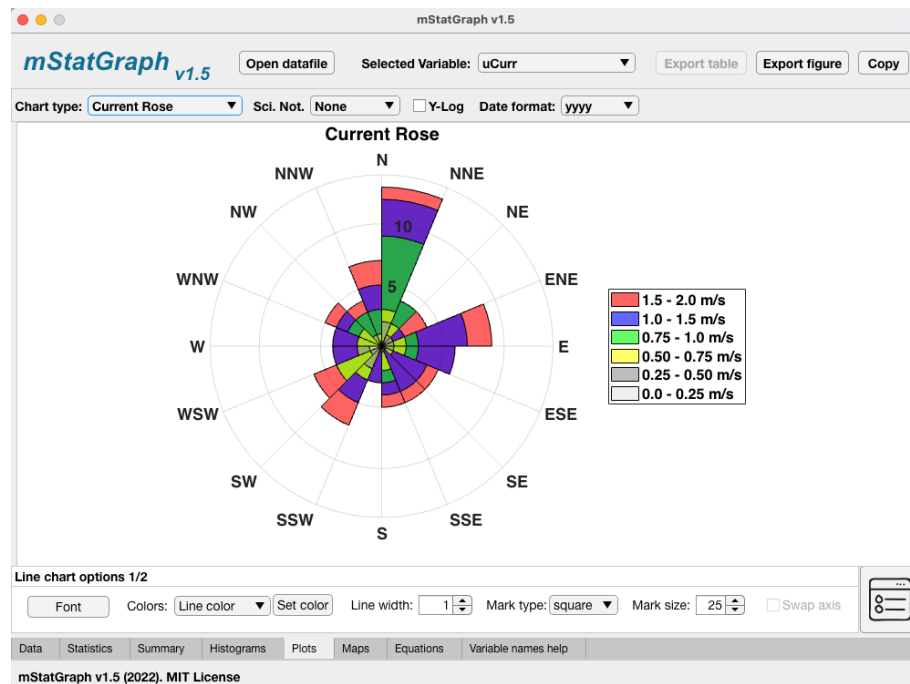
Example of a trend analysis (Variable selected: pH):



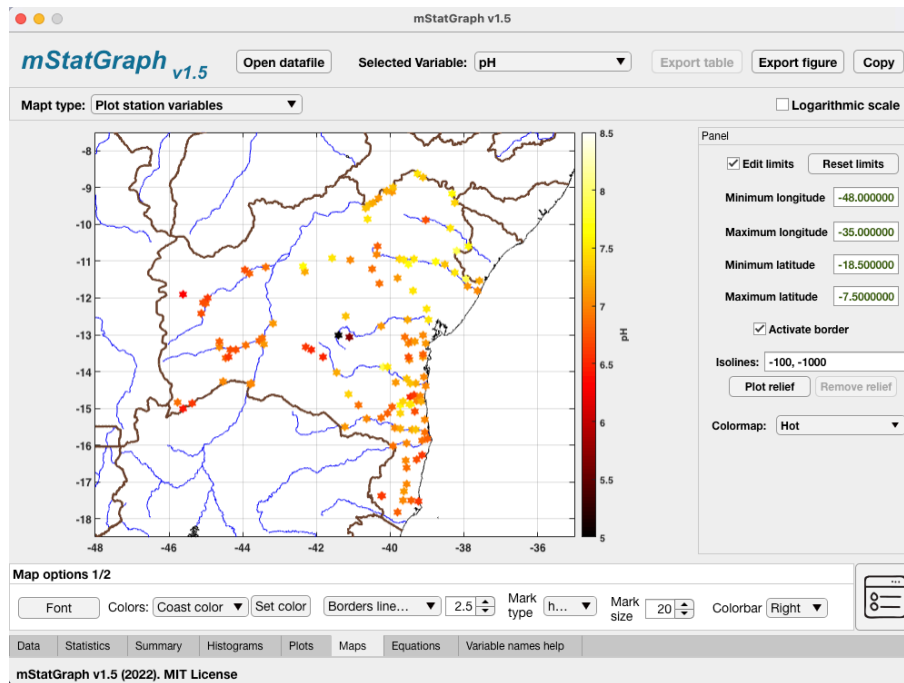
Hovmöller example (Variable selected: Temperature):



Example of marine currents rose:

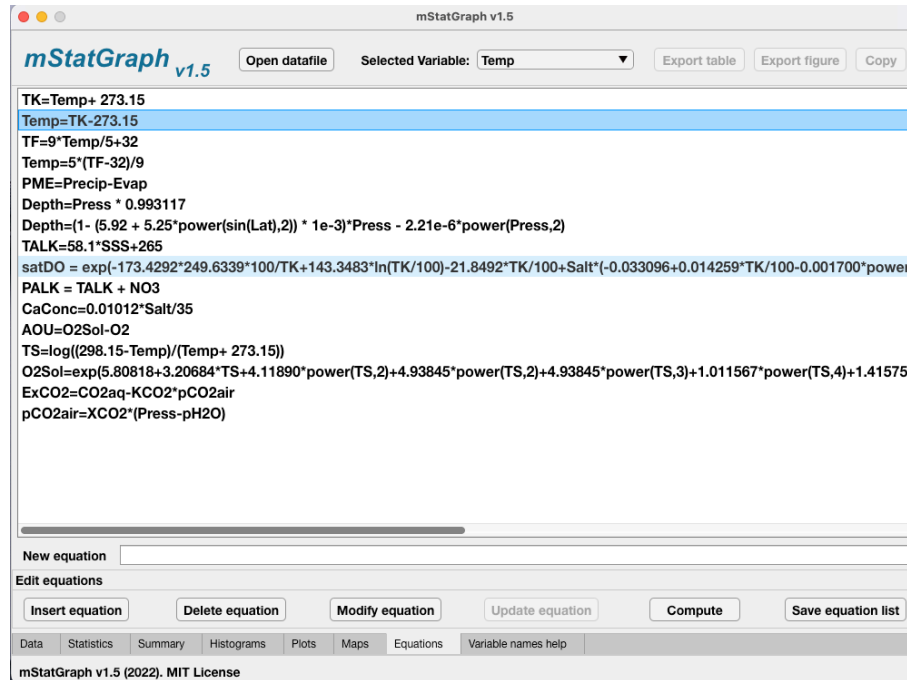


Example of a map with the spatial distribution of pH sampling in rivers:

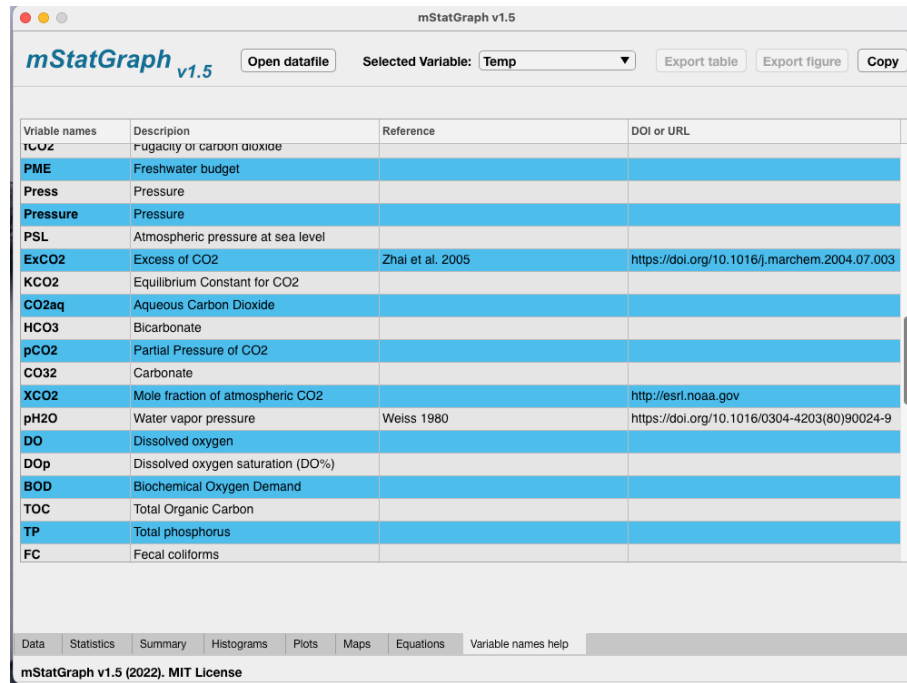


In the "Maps" tab you can customize the maps before export by changing the geographic coordinates of the boundaries, adding state or country borders, changing the shape of the sampling point, you can add bathymetry in the case of ocean data, and you can change the colormap.

mStatGraph allows you to enter equations to calculate new parameters or make unit conversions, these equations only allow as independent variables those found in the table of the "Data" tab and you can use all the predefined functions in MATLAB 2021b. The process is simple, write the equation in the "New equation" editbox and click on the "Insert equation" button, this way a new equation will be added to the list, this list can be saved through the "Save equation list" button. To calculate any equation just select it and click on the "Compute" button, this will add a new column to the table in the "Data" tab with the variable name to the left of the "=" sign.



The last tab ("Variable name help") lists the allowed variable names so that mStatGraph can classify them.



Variable names	Description	Reference	DOI or URL
TCO2	Trugacity of carbon dioxide		
PME	Freshwater budget		
Press	Pressure		
Pressure	Pressure		
PSL	Atmospheric pressure at sea level		
ExCO2	Excess of CO2	Zhai et al. 2005	https://doi.org/10.1016/j.marchem.2004.07.003
KCO2	Equilibrium Constant for CO2		
CO2aq	Aqueous Carbon Dioxide		
HCO3	Bicarbonate		
pCO2	Partial Pressure of CO2		
CO32	Carbonate		
XCO2	Mole fraction of atmospheric CO2		http://esrl.noaa.gov
pH2O	Water vapor pressure	Weiss 1980	https://doi.org/10.1016/0304-4203(80)90024-9
DO	Dissolved oxygen		
DOp	Dissolved oxygen saturation (DO%)		
BOD	Biochemical Oxygen Demand		
TOC	Total Organic Carbon		
TP	Total phosphorus		
FC	Fecal coliforms		