DSCompare Matlab Package

Software to analyze, compare and validate analysis and reanalysis datasets with an observed dataset

USER MANUAL

By

Humberto L. Varona

And

Tonia A. Capuano

Version 2.1

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

Overview

Computational tool that analyzes, compares and validates analysis and reanalysis datasets with an observed dataset using statistical tests such as Mann Whitney (U-test), t-test, F-test, Root Mean Square Error (RMSE), correlation coefficient, BIAS, normalized BIAS, trend, scatter index and maximum anomalies.

Version

2.1

Release date

May, 5th 2023

License

MIT

Download URL

https://doi.org/10.5281/zenodo.8152618

Cite as

Varona, Humberto L., Capuano, Tonia A., Noriega, Carlos, Araujo, Julia, Araujo, Moacyr, & Hernandez, Fabrice. (2023). Software to analyze, compare and validate analysis and reanalysis datasets with an observed dataset (DSCompare). (2.1). Zenodo. https://doi.org/10.5281/zenodo.8152618

Comparison tests

- *Mean value of oceanographic parameters: This refers to the average value of various oceanographic parameters, such as temperature, salinity, or dissolved oxygen, measured at different locations or time intervals.*
- *Mann-Whitney test: It is a non-parametric test used to compare the distributions of two independent samples. It determines whether there is a significant difference between the medians of the two datsets.*
- *Two-sample t-test: This parametric test is used to compare the means of two independent samples. It assesses whether the difference between the samples means is statistically significant.*

- *Two-sample F-test: This test is used to compare the variances of two independent samples. It statistically determines if the variability between the two datasets is significantly different.*
- *Bias/Normalized bias (nBias): Bias refers to the systematic deviation between the measured values and the true values. Normalized bias is expressed as a percentage or a ratio, providing a standardized measure of the discrepancy, (Equations 1 and 2). Bias is a measure that reflects the error or constant deviation between simulated and observed values. It provides insight into whether the simulated values tend to systematically overestimate or underestimate the observed value, with positive values indicating overestimation and negative values indicating underestimation.*

$$Bias = \frac{\sum_{i=1}^{n}(O_i - S_i)}{n} \qquad \text{Equation 1}$$

Where O_i are the observed or reference values and S_i the simulated or theoretically estimated values, while n is the number of observations.

$$nBias = \frac{\sum_{i=1}^{n}(O_i - S_i)}{n}\frac{1}{\bar{S}} \qquad \text{Equation 2}$$

Where $\bar{S}$ is the mean value of S_i.

- *Standard deviation: It measures the dispersion or variability of a set of values around their mean. A higher standard deviation indicates greater variability within the dataset.*
- *Coefficient of determination (Equation 4): It is a measure of how well a regression model fits the data. R-squared indicates the proportion of the variance in the dependent variable whichis predictable from the independent variable(s).*

$$R = \left(\frac{\sum_{i=1}^{n}(S_i - \bar{S})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2 \sum_{i=1}^{n}(O_i - \bar{O})^2}}\right)^2 \qquad \text{Equation 4}$$

- *RMSE (Root Mean Square Error): This is a measure of the average prediction error of a regression model. It represents the square root of the mean of the squared differences between simulated and observed values (Equation 5). RMSE condenses the overall simulation error into a single value: a lower RMSE value indicates better model performance and higher accuracy, meaning that the simulated values are closer to the observed values on average. It serves as a useful measure for evaluating the quality and precision of simulated data.*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - S_i)^2}{n}} \qquad \text{Equation 5}$$

- *Maximum anomalies: Anomalies are deviations from the expected or average values. Maximum anomalies refer to the most extreme or significant deviations from the norm within a dataset.*
- *Scatter index: It quantifies the degree of scatter or dispersion of data points around a regression line. A higher scatter index suggests greater variability and less precision in the relationship between variables (Equation 6). A lower Scatter Index value signifies a higher level of agreement or consistency between the two sets of*

measurements. This index quantifies the dispersion or spread of the differences relative to the averaged reference value. In practice, it provides an indication of how closely the measurements align with the reference values, with a smaller Scatter Index indicating a tighter agreement between the two datasets.

$$SI = \frac{RMSE}{\bar{S}} \qquad \textit{Equation 6}$$

- *Covariance: Itmeasures the linear relationship between two variables, indicating the degree to which changes in one variable correspond to changes in another (Equation 7).*

$$Cov(S_i, O_i) = \frac{\sum_{i=1}^{n}(S_i - \bar{S})(O_i - \bar{O})}{n} \qquad \textit{Equation 7}$$

- *Trend analysis: It examines the long-term patterns and directionality of a dataset over time and helps identify if there is a significant upward or downward trend, indicating systematic changes in the data over the specified period.*

How to install

Matlab 2021b compatible software

- Open Matlab.
- Go to APP tab.
- Click on the "Install App" button.
- Select the DSCompare.mlappinstall file.
- In the Install dialog click on the "Install" button.

Create a directory and copy into it the following databases:

Filename	Description
gshhs_f_coast.mat	Database with the global coastline
gshhs_f_rivers.mat	database with the location of the major global rivers
wdb_f_borders.mat	database with the administrative political division of countries and states

these 3 files can be downloaded from https://zenodo.org/record/8152618

How to run

Type in the Matlab command window:

>> DSCompare <Enter>

or find DSCompare in the APP tab of Matlab.

Operation mode

Figure 1 shows the main screen of the DSCompare Matlab package, which compares an analysis/reanalysis dataset with a reference dataset (observed data). Only datasets stored in standard NetCDF format and complying with the CF-1.6 convention can be used. Datasets produced by hydro-thermodynamic and regional circulation models, such as ROMS, CROCO and NEMO models have a non-standard NetCDF format, since the reference variables of these are different as longitude and latitude are two-dimensional variables, and depth and time may have different names, e.g., **time_counter**, **level**, **depthu**, **depthv**, **depthw**, **depth**, etc.

All these datasets have to be standardized, that is, they have to be fully compatible with the nco, CDO and ncdump tools as the latter will be used in the preparation of the datasets for use in DSCompare. In the case of ROMS and CROCO models, the ROMSTOOLS (Penven et al., 2003) and CROCOTOOLS packages can be easily adapted for this objective, and for the output of the NEMO model there is a converting tool called fcNEMOtoStd v1.3 (Varona, 2023a) .

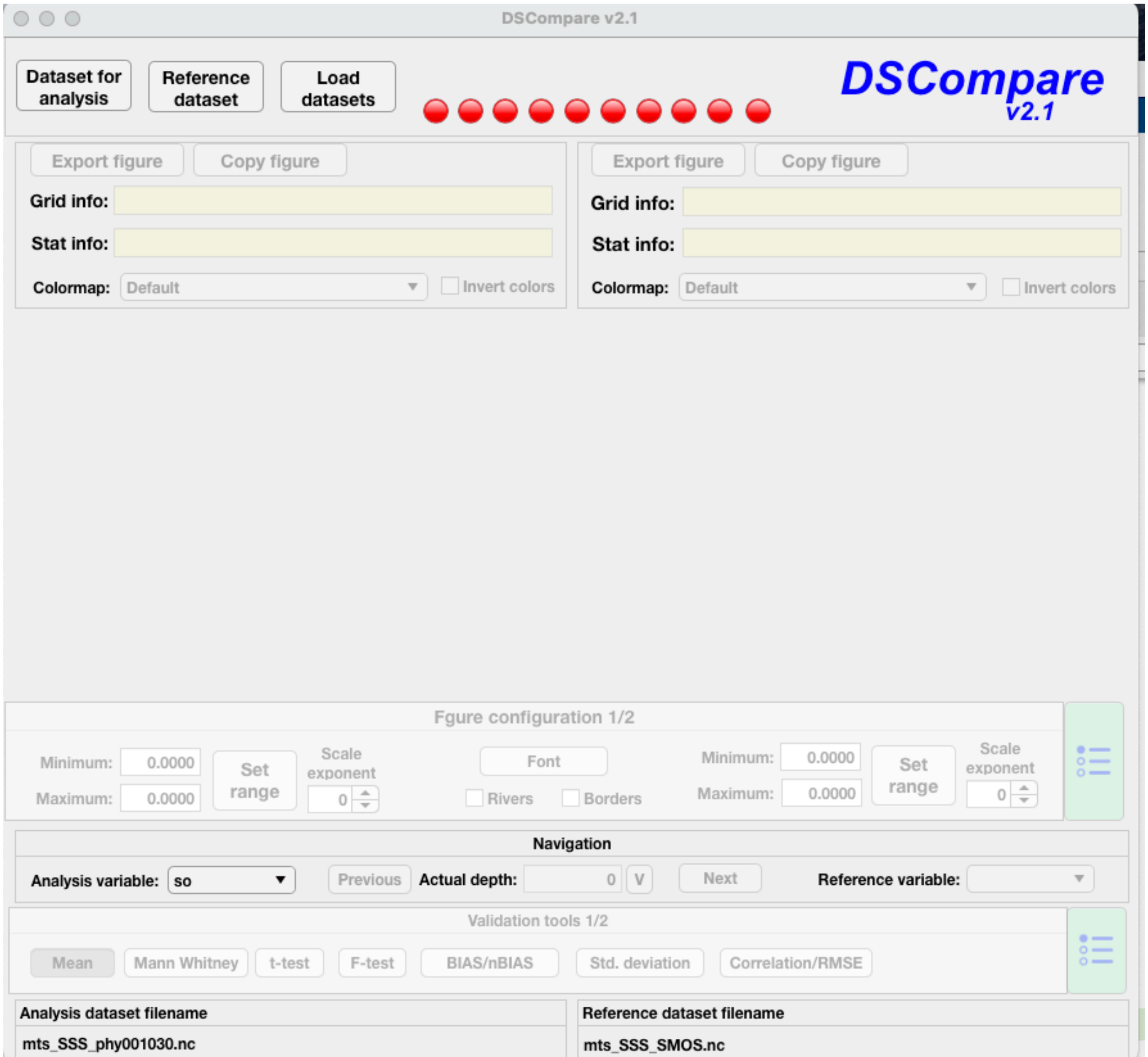


Figure 1. Main screen of DSCompare Matlab package.

DCompare workflow

1. Select the dataset to be validated by clicking on the "Dataset for analysis" button.
2. Select the variable to be validated using the "Analysis variable" drop-down menu.
3. Select the reference dataset by clicking on the "Reference dataset" button.
4. Select the reference variable using the "Reference variable" drop-down menu.
5. Load both variables into memory using the "Load datasets" button (figure 2).
6. The selected variables will be loaded and all comparison parameters for the surface will be calculated.
7. Through the buttons "Next" and "Previous" it will be possible to change the depth if both datasets have 4 dimensions (lon, lat, depth, time). With the "V" button a specific depth can be selected.

8. Finally, the validations will be performed through the buttons found in "Validation tools"; by default, 7 tools are shown (Mean, Mann Whitney, t-test, F-Test, BIAS/nBIAS, Std. Deviation and Correlation/RMSE). By clicking on the button , you can obtain 3 more tools (Maximum anomalies, Scatter index/Covariance and Trend).

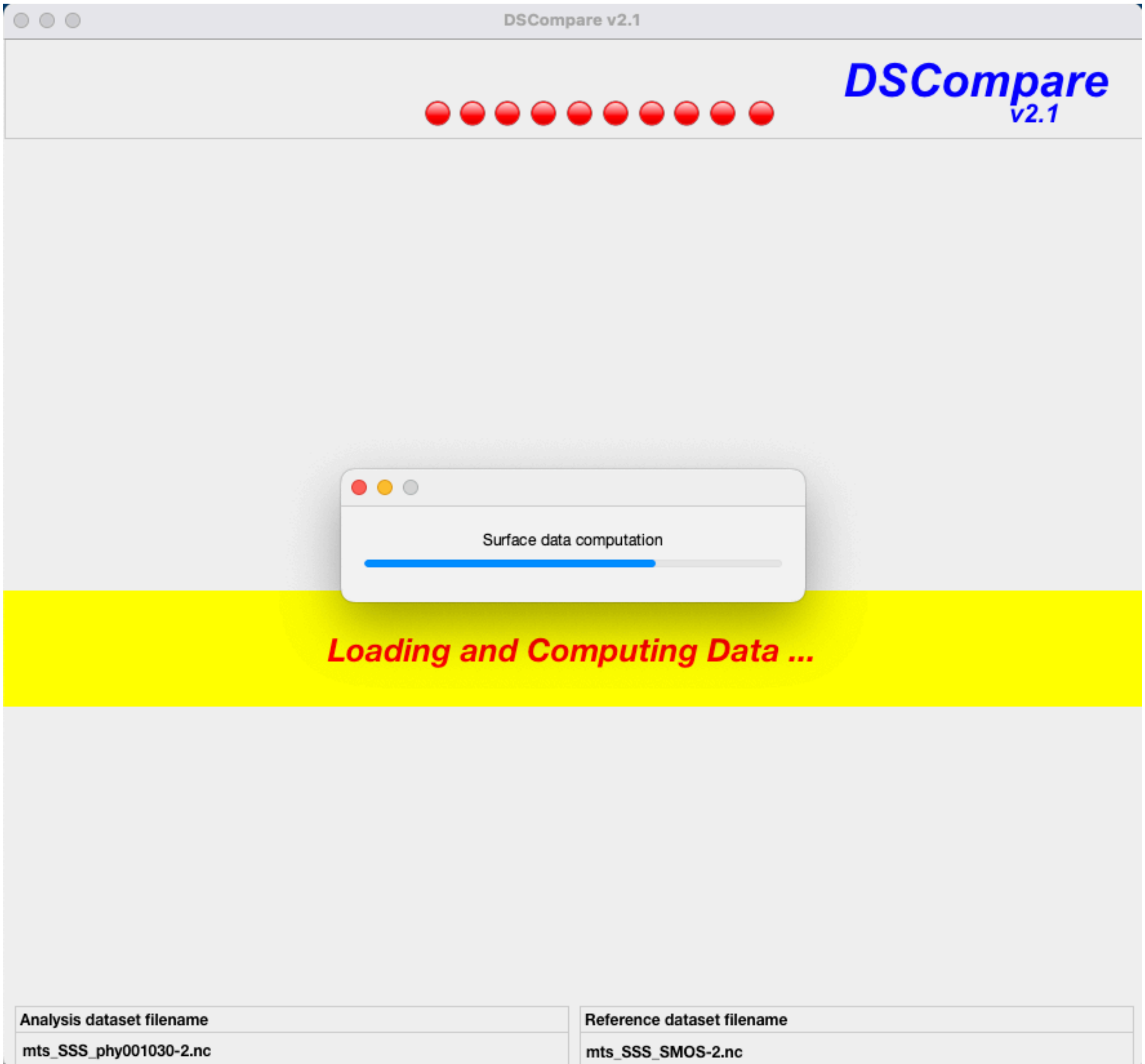
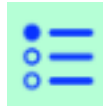
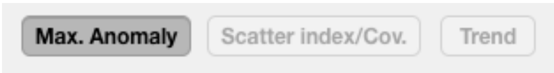


Figure 2. Loading of datasets.

Clicking on the buttons below will display the results of all the tests. Figure 3 shows an example of the Mann-Whitney test.

More tests can be activated by clicking on the button 



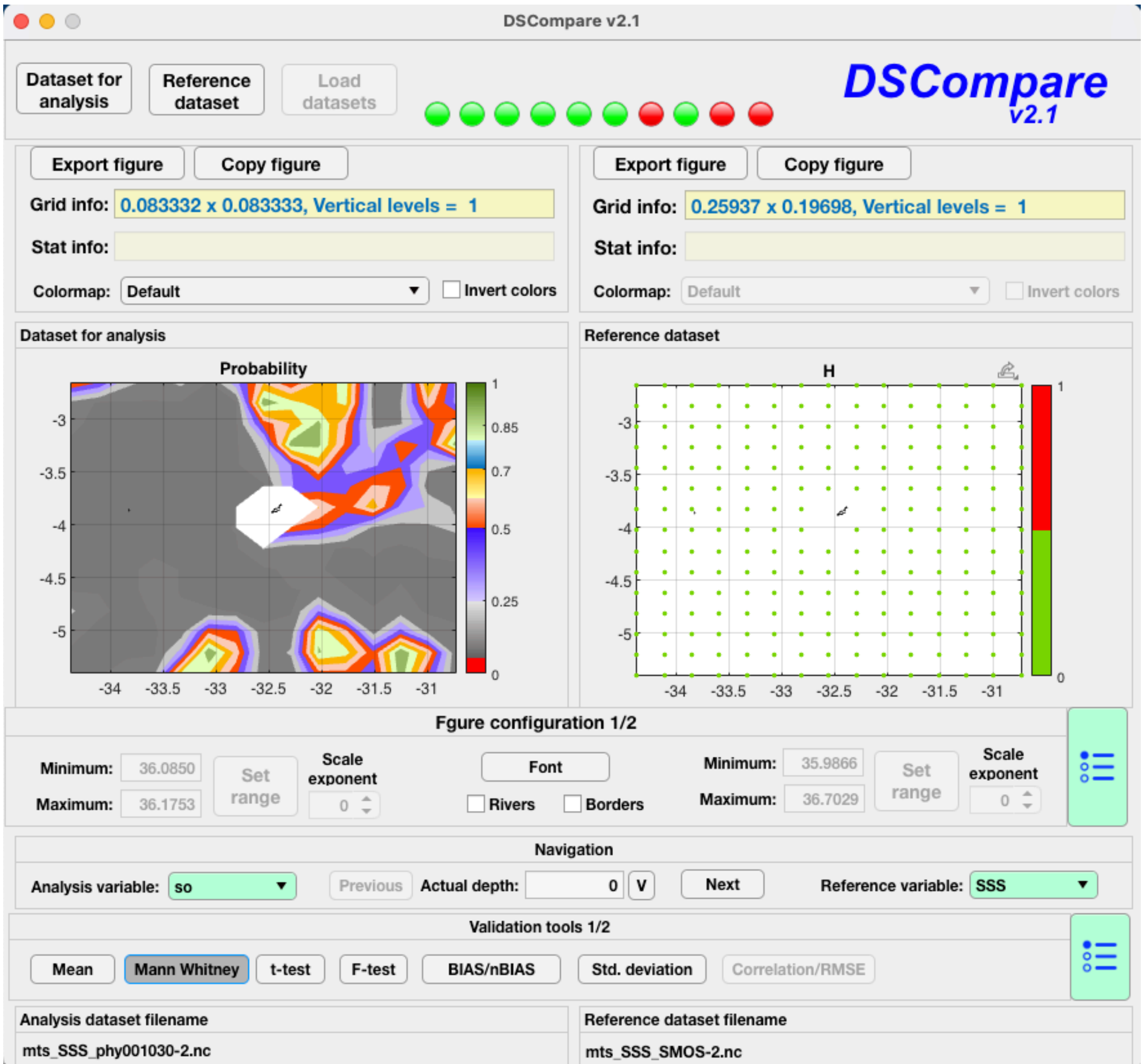


Figure 3. Comparison using the Mann - Whitney test.

Preprocessing of the datasets

The datasets to be analyzed and the reference dataset must have very similar geographical limits; this information can be retrieved using CDO (Schulzweida et al., 2006) from the command line in the terminal:

```
cdo sinfo dataset_name.nc
```

Output:

```
File format : NetCDF
   -1 : Institut Source   T Steptype Levels Num    Points Num Dtype : Parameter ID
    1 : unknown  unknown  v instant      40   1      6560   1  F64  : -1
    2 : unknown  unknown  v instant      40   1      6560   1  F32  : -2
  Grid coordinates :
    1 : lonlat                  : points=6560 (80x82)
                           lon : -59.5 to 19.5 by 1 degrees_east
                           lat : -30.211 to -3.238 by 0.333 degrees_north
  Vertical coordinates :
    1 : pressure                : levels=40
                         depth : 5 to 4478 millibar
  Time coordinate :
                          time : 489 steps
    RefTime =  0001-01-01 00:00:00  Units = days  Calendar = standard
  YYYY-MM-DD hh:mm:ss  YYYY-MM-DD hh:mm:ss  YYYY-MM-DD hh:mm:ss  YYYY-
MM-DD hh:mm:ss
  1980-01-01 00:00:00  1980-02-01 00:00:00  1980-03-01 00:00:00  1980-04-01 00:00:00
  1980-05-01 00:00:00  1980-06-01 00:00:00  1980-07-01 00:00:00  1980-08-01 00:00:00
.................................................................................
   ............................
2020-05-01 00:00:00  2020-06-01 00:00:00  2020-07-01 00:00:00  2020-08-01 00:00:00
  2020-09-01 00:00:00
```

The "sinfo" operator displays the limits of all dimensions of the NetCDF file.

The "sellonlatbox" operator is used to select a geographic region from a dataset:

```
cdo sellonlatbox,-59.5,19.5,-30.211,-3.238 reference_dataset.nc output.nc
```

The two datasets to be compared have to cover the same temporal period. To select a time interval, type:

```
cdo seldate, 1980-01-01, 2020-09-01 reference_dataset.nc output.nc
```

They must also match in spatial resolution, both horizontally (this is resolved in DSCompare) and vertically; the latter can be done by means of the "intlevel" operator.

The depths can be displayed as:

```
ncdump -v depth  reference_dataset.nc
```

…

```
…
…
data:

depth = 5, 15, 25, 35, 45, 55, 65, 75, 85, 95, 105, 115, 125, 135, 145, 155,
    165, 175, 185, 195, 205, 215, 225, 238, 262, 303, 366, 459, 584, 747,
    949, 1193, 1479, 1807, 2174, 2579, 3016, 3483, 3972, 4478 ;
}
```

```
ncdump -v depth analysis_dataset.nc
```

```
…
…
…
data:

 depth = 5.5, 7.1, 10, 25, 50, 70, 95, 105 ;
}
```

We can then interpolate the dataset to be validated at the depths at which the data is found in the reference dataset with the following operator:

```
cdo intlevel,5.5,15,25,35,45,55,65,75,8595,105 analysis_dataset.nc ouput.nc
```

Note: The shallowest depth of the reference dataset is 5 m and that of the analysis dataset is 5.5 m, so it cannot be interpolated for the depth of 5 m, and the same happens with the deepest depth. In addition, it cannot be interpolated for depths greater than 105 m, which is the maximum depth of the analysis dataset. Vertical interpolation can also be performed through the Matlab package: NCVerticalInterp (Varona, 2023b).

Once the dataset to be analyzed and the reference dataset have the same spatial geometry and the same data frequency, they are ready to be used in DSCompare.

Note: The size of the datasets that can be loaded into DSCompare will depend on the size of the RAM available on the computer where the Matlab software is run.

Datasets can be reduced by extracting each variable separately through the CDO "selname" operator:

```
cdo selname,salt analysis_dataset.nc salt_output.nc
```

The CDO user manual can be downloaded from the following URL:

https://code.mpimet.mpg.de/projects/cdo/embedded/cdo.pdf

When loading datasets for analysis and reference data, the following error may appear because the NetCDF ls file is missing the dimension and the variable **depth**, **level** or **lev**.

This error occurs when only 1 level is compared.

```
ncdump -h sample.nc
```

```
netcdf sample {
dimensions:
	time = UNLIMITED ; // (132 currently)
	lon = 15 ;
	lat = 15 ;
variables:
	float time(time) ;
		time:standard_name = "time" ;
		time:long_name = "time" ;
		time:bounds = "time_bnds" ;
		time:units = "days since 1950-01-01 00:00:00.0" ;
		time:calendar = "gregorian" ;
		time:axis = "T" ;
	float lon(lon) ;
		lon:standard_name = "longitude" ;
		lon:long_name = "longitude" ;
		lon:units = "degrees_east" ;
		lon:axis = "X" ;
	float lat(lat) ;
		lat:standard_name = "latitude" ;
		lat:long_name = "latitude" ;
		lat:units = "degrees_north" ;
		lat:axis = "Y" ;
	float SSS(time, lat, lon) ;
		SSS:standard_name = "sea_surface_salinity" ;
		SSS:long_name = "Unbiased Sea Surface Salinity" ;
		SSS:units = "pss" ;
		SSS:_FillValue = NaNf ;
		SSS:missing_value = NaNf ;
		SSS:cell_methods = "time: mean" ;

// global attributes:
…
```

1- **We add the dimension and the variable "depth" with only one value**

```
ncap2 -A -s 'defdim("depth",1);depth[$depth]=0.0;depth@long_name="Depth";' sample.nc
ncdump -h sample.nc
```

```
netcdf sample {
dimensions:
	time = UNLIMITED ; // (132 currently)
	lon = 15 ;
	lat = 15 ;
	depth = 1 ;
variables:
	float time(time) ;
		time:standard_name = "time" ;
		time:long_name = "time" ;
		time:bounds = "time_bnds" ;
		time:units = "days since 1950-01-01 00:00:00.0" ;
		time:calendar = "gregorian" ;
		time:axis = "T" ;
	float lon(lon) ;
		lon:standard_name = "longitude" ;
		lon:long_name = "longitude" ;
		lon:units = "degrees_east" ;
		lon:axis = "X" ;
	float lat(lat) ;
		lat:standard_name = "latitude" ;
		lat:long_name = "latitude" ;
		lat:units = "degrees_north" ;
		lat:axis = "Y" ;
	float SSS(time, lat, lon) ;
		SSS:standard_name = "sea_surface_salinity" ;
		SSS:long_name = "Unbiased Sea Surface Salinity" ;
		SSS:units = "pss" ;
		SSS:missing_value = NaNf ;
		SSS:cell_methods = "time: mean" ;
		SSS:_FillValue = NaNf ;
	double depth(depth) ;
		depth:long_name = "Depth" ;

// global attributes:
…
```

2- **A time variable is added with all dimensions, including level ("delpth", "level", "lev").**

```
ncap2 -h -A -s "temporal_SSS[time, depth, lat, lon]=SSS" sample.nc
ncdump -h sample.nc
```

```
netcdf sample {
dimensions:
	time = UNLIMITED ; // (132 currently)
	lon = 15 ;
	lat = 15 ;
	depth = 1 ;
variables:
	float time(time) ;
		time:standard_name = "time" ;
		time:long_name = "time" ;
		time:bounds = "time_bnds" ;
		time:units = "days since 1950-01-01 00:00:00.0" ;
		time:calendar = "gregorian" ;
		time:axis = "T" ;
	double time_bnds(time, bnds) ;
	float lon(lon) ;
		lon:standard_name = "longitude" ;
		lon:long_name = "longitude" ;
		lon:units = "degrees_east" ;
		lon:axis = "X" ;
	float lat(lat) ;
		lat:standard_name = "latitude" ;
		lat:long_name = "latitude" ;
		lat:units = "degrees_north" ;
		lat:axis = "Y" ;
	float SSS(time, lat, lon) ;
		SSS:standard_name = "sea_surface_salinity" ;
		SSS:long_name = "Unbiased Sea Surface Salinity" ;
		SSS:units = "pss" ;
		SSS:missing_value = NaNf ;
		SSS:cell_methods = "time: mean" ;
		SSS:_FillValue = NaNf ;
	double depth(depth) ;
		depth:long_name = "Depth" ;
	float temporal_SSS(time, depth, lat, lon) ;
		temporal_SSS:cell_methods = "time: mean" ;
		temporal_SSS:long_name = "Unbiased Sea Surface Salinity" ;
		temporal_SSS:missing_value = NaNf ;
		temporal_SSS:standard_name = "sea_surface_salinity" ;
		temporal_SSS:units = "pss" ;
```

```
// global attributes:
…
```

3- **The old variable is deleted and a new file is obtained without the old variable (sample2.nc).**

```
ncks -h -C -O -x -v "SSS" sample.nc sample2.nc
ncdump -h sample2.nc
```

```
netcdf sample2 {
dimensions:
	depth = 1 ;
	time = UNLIMITED ; // (132 currently)
	lat = 15 ;
	lon = 15 ;
variables:
	double depth(depth) ;
		depth:long_name = "Depth" ;
	float lat(lat) ;
		lat:standard_name = "latitude" ;
		lat:long_name = "latitude" ;
		lat:units = "degrees_north" ;
		lat:axis = "Y" ;
	float lon(lon) ;
		lon:standard_name = "longitude" ;
		lon:long_name = "longitude" ;
		lon:units = "degrees_east" ;
		lon:axis = "X" ;
	float temporal_SSS(time, depth, lat, lon) ;
		temporal_SSS:cell_methods = "time: mean" ;
		temporal_SSS:long_name = "Unbiased Sea Surface Salinity" ;
		temporal_SSS:missing_value = NaNf ;
		temporal_SSS:standard_name = "sea_surface_salinity" ;
		temporal_SSS:units = "pss" ;
	float time(time) ;
		time:standard_name = "time" ;
		time:long_name = "time" ;
		time:bounds = "time_bnds" ;
		time:units = "days since 1950-01-01 00:00:00.0" ;
		time:calendar = "gregorian" ;
		time:axis = "T" ;

// global attributes:
…
```

4- **Finally, the variable is renamed with the old name.**

```
ncrename -O -v temporal_SSS,SSS sample2.nc
```

Refrences

Penven, P., Cambon, G., Tan, T. A., Marchesiello, P., & Debreu, L. (2003). ROMSTOOLS user's guide. *Rapport techn., IRD and LPO/UBO, Laboratoire de Physique des Oceans, Universite de Bretagne Occidentale/UFR Sciences*.

Schulzweida, U., Kornblueh, L., & Quast, R. (2006). CDO user's guide. *Climate data operators, Version*, *1*(6), 205-209.

Varona, Humberto L. (2023a). Format converter from NEMO model to NetCDF standard (fcNEMOtoStd) (1.4). Zenodo. https://doi.org/10.5281/zenodo.7519023

Varona, Humberto L. (2023b). Vertically interpolates NetCDF files (NCVerticalInterp) (1.2). Zenodo. https://doi.org/ 10.5281/zenodo.7519015