

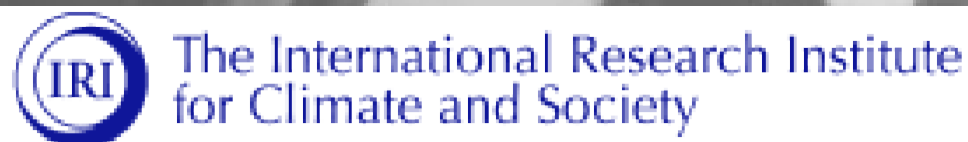
6.

DRAFT



Introduction to the Data Library (DL): Correlation/Regression

**Training Module
November 29, 2016
Version 1.0**



International Research Institute for Climate and Society (IRI), (2016). Introduction to the Data Library (DL)- Introduction. November 29, Version 1.0. Palisades: IRI.

This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>) and may be adapted or reproduced with attribution to the IRI and for any non-commercial purpose.

CONTENTS

1 Introduction to the Data Library (DL) - Correlation/Regression 1

1.1 Introduction 1

1.2 Overview 1

1.3 Access 2

1.4 Calculating the Correlation of an Index [T] With a Variable [X Y T] 2

1.5 Calculating Regression of an Index [T] With a Variable [X Y T] - in units of measure of the variable . 4

1.6 Correlation Between Two Variables [X Y T] 4

1.7 Summary 10

1.8 Quiz 10

1.9 Reference(s) 10

INTRODUCTION TO THE DATA LIBRARY (DL) - CORRELATION/REGRESSION

1.1 Introduction

The IRI Climate Data Library is a library of datasets. By library we mean a collection of datasets, collected from various sources, designed to make them more accessible for the library's users (Blumenthal, 2004). For this module we will be expanding on how the users can get correlation or regression to obtain the desired information. Traditional GIS platforms are now widely used by planners and decision makers in society. However, they are highly-focused on geospatial capabilities and have limited functionality for temporal analysis. Without information on the latter, meaningful inference about the causation of disease outbreaks is impossible (Jacquez 2000). Furthermore, many tools are unable to readily process the vast quantities of space-time data associated with, for example, the outputs of a global climate model. The IRI Climate Data Library overcomes the limitations imposed by GIS platforms by being based on a much more general multi-dimensional data model that includes both space and time dimensions. All datasets, including GIS features (such as points, lines, and polygons) are geo-located and temporally referenced in a uniform framework.

The equations for linear correlation and regression are as follows (Figure 1.1)

$$\text{correlation} = \rho = \frac{1}{n-1} \sum \frac{(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} \quad \rho = \frac{1}{n-1} \sum \frac{(x - \bar{x})}{\sigma_x} \frac{(y - \bar{y})}{\sigma_y}$$
$$\text{regression} = \frac{1}{n-1} \sum \frac{(x - \bar{x})(y - \bar{y})}{\sigma_x}$$

[T] standardize
or
dup
[T] average sub
dup
[T] rmsaover div

Fig. 1.1: Linear correlation and regression equation

1.2 Overview

What will be working on?

For this manual we will be working on two methods:

- Calculating the correlation of an index [T] with a variable [X Y T]
- Calculating the regression of an index [T] with a variable [X Y T]

1.3 Access

The IRI Data Library can be accessed with the following links:

- Worldwide: <http://iridl.ldeo.columbia.edu/>
- Chile: <http://www.climatedatalibrary.cl/>
- Venezuela: <http://datoteca.ole2.org/>
- Uruguay: <http://dlibrary.snia.gub.uy/>
- Rwanda: <http://maproom.meteorwanda.gov.rw/>
- Ethiopia: <http://www.ethiometmaprooms.gov.et:8082/>
- Tanzania: <http://maproom.meteo.go.tz/>
- Mali: <http://197.155.140.164/>
- Ghana: <http://maps.meteo.gov.gh:89/>
- Zambia: <http://41.72.104.142/>
- Madagascar: <http://map.meteomadagascar.mg/>
- Peru: <http://ons.snirh.gob.pe/>
- Niger: <http://cradata.agrhymet.ne/>
- Kenya (KMD): <http://kmddl.meteo.go.ke:8081/>
- Kenya (ICPAC): <http://digilib.icpac.net/>

There are two types of functions used under “Function Documentation” for this manual and those are the following seen of Figure 1.2: standardize and correlate

1.4 Calculating the Correlation of an Index [T] With a Variable [X Y T]

When calculating the correlation of an index with certain variables the following expert mode functions are used, this example if for Mali:

SOURCES .Indices .nino .EXTENDED .NINO34

T (May-Oct) seasonalAverage

T (May-Oct 1983) (May-Oct 2013) RANGE

SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged

monthlyAverage

T (May-Oct) seasonalAverage

T (May-Oct 1983) (May-Oct 2013) RANGE

[T] correlate

The function or function used and for this example are to find the seasonal average [seasonalAverage] for a range of time series [RANGE] and calculate the correlation between Nino 3.4 index and monthly rainfall data of months May-October. This is just the outline of the function, please follow on to know more on how we find the correlation between the index and the variables.

In order to find this correlation we reference to the equation on Figure 1.3.

a)

standardize

Standardizes a set of data by removing its mean and dividing the result by its standard deviation

SOURCES .NOAA .NCDC .ERSST .version2 .SST
T (Jan 1974) (Dec 2003) RANGE
[T]standardize

See Also

Statistical Functions: : flexseasonalreqGT
flexseasonalreqLT : flexseasonalmeandailyvalueGT
flexseasonalnonoverlapDSfreq
flexseasonalnonoverlapWSfreq : flexseastotAvgFill
flexseastotZeroFill : monthly3Q : monthlyMAVE
monthlyMAVE_SD : monthlyMAVEplus1p96SD
monthlyMAVEplus1SD : monthlyMAVEplus2SD
monthlymean : monthlymeanplus1SD
monthlymeanplus2SD : monthlySD : pentad3Q : pentaddepththresholds : pentadMAVE : pentadMAVE_SD : pentadMAVEplus1p96SD : pentadMAVEplus1SD
pentadMAVEplus2SD : pentadmean : pentadmeanplus1SD : pentadmeanplus2SD : pentadSD : seasonalreqGT : seasonalreqLT : seasonalmeandailyvalueGT
seasonalnonoverlapDSfreq : seasonalnonoverlapWSfreq : seastotAvgFill : seastotZeroFill : width96

label	type	Description
var	variable	variable (i.e., data) to be standardized
grids	grid set	grid(s) (i.e., independent variables) over which the mean and standard deviation are calculated
minfrac	number	Minimum fraction of data that must be present (i.e., fraction not indicated as missing) within the selected domain in order for the function to be performed. If minfrac is not present, then a missing value is returned. If minfrac is not given, then the function is performed regardless of the amount of data present in domain. (optional)
stdvar	output variable	standardized data calculated by removing the mean from var and dividing by its standard deviation

b)

correlate

Calculates the Pearson Product-Moment Correlation coefficient of two variables over specified grids (i.e., independent variables)

Description

correlate calculates the Pearson product moment correlation for the two latest items on the stack over the indicated grid. For the correlation to be computed, the gridding of the two items on the stack must match.

Example

SOURCES .NOAA .NCEP .CPC .GMSM .w
T (Jan 1969) (Dec 1998) RANGE
X (-8) (20) RANGE
Y (8) (20) RANGE
SOURCES .DEKLIM .VASCLimO .PrpClim .Resolution-
Op5xOp5 .prcp
T (Jan 1969) (Dec 1998) RANGE
X (-8) (20) RANGE
Y (8) (20) RANGE
[T]correlate

In this example, GMSM monthly soil moisture values are correlated over the time grid with monthly precipitation

label	type	Description
var1	variable	variable to be correlated with var2
var2	variable	variable to be correlated with var1 Note that var1 and var2 should have similarly-defined grids. Regridding one variable to match the other may be necessary (see example below).
grids	grid set	grid(s) (i.e., independent variables) over which correlation coefficient is to be calculated
minfrac	number	Minimum fraction of data that must be present (i.e., fraction not indicated as missing) within the selected domain in order for the correlation to be calculated. If minfrac is not present, then a missing value is returned. If minfrac is not given, then the correlation is calculated regardless of the amount of data present in the domain. (optional)
coefficient	output variable or constant	Pearson-Product Moment Correlation coefficient of var1 and var2 over grids. coefficient is not dependent on grids, but is dependent on any other grids that var1 or var2 depended on (if any).

Fig. 1.2: Function Documentation used (a) standardize (b) correlate

$$\rho = \frac{1}{n-1} \sum \frac{(x - \bar{x})}{\sigma_x} \frac{(y - \bar{y})}{\sigma_y}$$

mul

[T] average

Fig. 1.3: Equation referenced for correlation

Ultimately the functions will be laid out as follows allowing the user to [standardize] both the index and the variables:

```
SOURCES .Indices .nino .EXTENDED .NINO34
```

```
T (May-Oct) seasonalAverage
```

```
T (May-Oct 1983) (May-Oct 2013) RANGE
```

```
[T] standardize
```

```
SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged
```

```
monthlyAverage
```

```
T (May-Oct) seasonalAverage
```

```
T (May-Oct 1983) (May-Oct 2013) RANGE
```

```
[T] standardize
```

```
mul [T] average
```

1.5 Calculating Regression of an Index [T] With a Variable [X Y T] - in units of measure of the variable

Taking note on how the correlation was computed the only difference for regression is to use [average sub] function for the variables, see below:

```
SOURCES .Indices .nino .EXTENDED .NINO34
```

```
T (May-Oct) seasonalAverage
```

```
T (May-Oct 1983) (May-Oct 2013) RANGE
```

```
[T] standardize
```

```
SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged
```

```
monthlyAverage
```

```
T (May-Oct) seasonalAverage
```

```
T (May-Oct 1983) (May-Oct 2013) RANGE
```

```
[T] standardize
```

```
dup [T] average sub
```

```
mul
```

```
[T] average
```

The reason why we use ‘average sub’ is because the “standardize” and “correlate” functions subtract the average.

1.6 Correlation Between Two Variables [X Y T]

Please take note that for each point of the match >> make sure that the matches X Y (and of course T) are the same. Hence the following function is used:

```
SOURCES .UMD .GLCF .GIMMS .NDVIg .global .ndvi
```

```

X -20 40 RANGE Y 0 30 RANGE
T (Jan 1982) (Dec 2003) RANGE
    monthlyAverage yearly-anomalies
SOURCES .NASA .GPCP .V2 .satellite-gauge .prcp
X -20 40 RANGE Y 0 30 RANGE
T (Jan 1982) (Dec 2003) RANGE
    yearly-anomalies
[X Y]regridAverage
[T]correlate

```

Note: If the matches are different, one can “regrid” [the match of the last variable to that of the penultimate one]
 In the case of a forecasting for example, if the match of the index [T] was not the same this would be the scenario:

```

SOURCES .Indices .nino .EXTENDED .NINO34
T (Jan 1983) last RANGE
T (Jun) VALUES
SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged
monthlyAverage
T (Jul-Sep) seasonalAverage
T (Jul-Sep 1983) (Jul-Sep 2013) RANGE
... ??? ... What to add here?
[T] correlate

```

Hence to make sure the matches are the same we can look more into expert mode just like Figure 1.4 and observe an offset between the index and the variable.

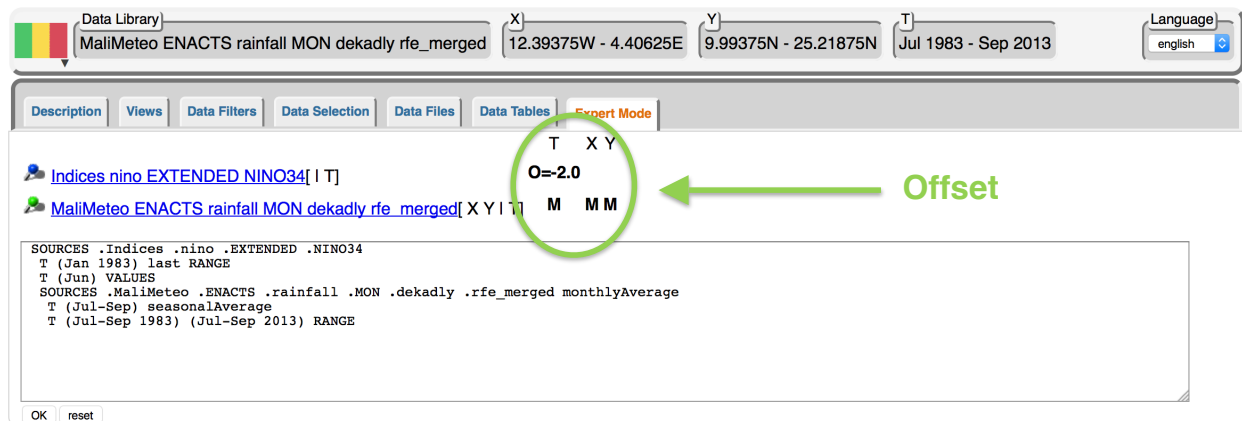


Fig. 1.4: Offset of Index and Variable

This were we introduce [ds/var grid num shiftGRID] function (Figure 1.5).

So in order to make the index and variables match we use the following:

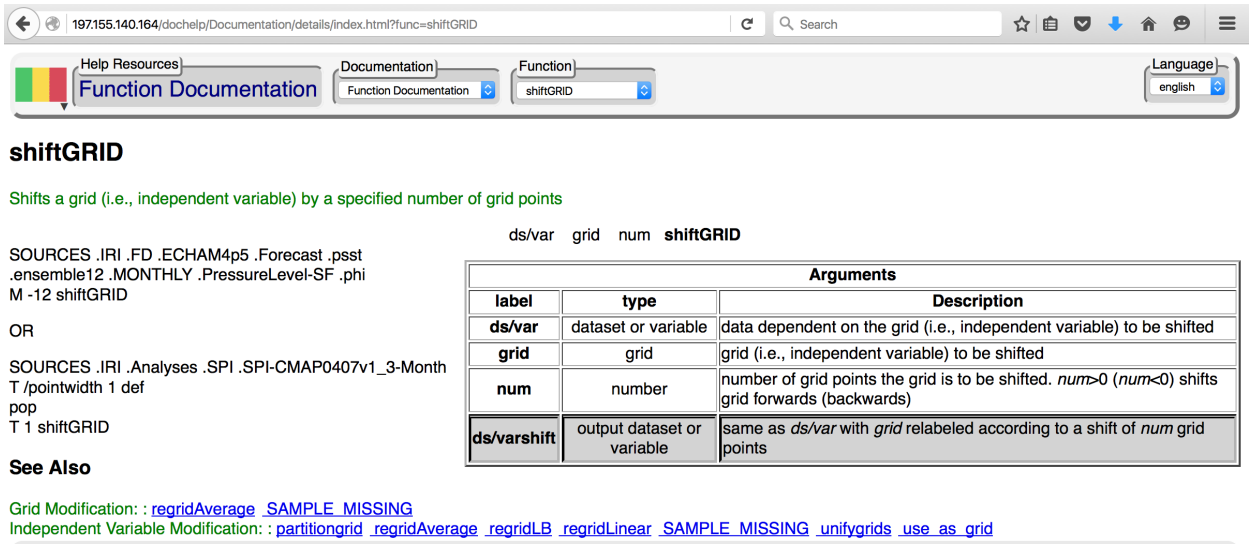


Fig. 1.5: [shiftGRID] Function in Function Documentation

SOURCES .Indices .nino .EXTENDED .NINO34
T (Jan 1983) last RANGE
T (Jun) VALUES
T 2 shiftGRID
SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged
monthlyAverage
T (Jul-Sep) seasonalAverage
T (Jul-Sep 1983) (Jul-Sep 2013) RANGE
or:
SOURCES .Indices .nino .EXTENDED .NINO34
T (Jan 1983) last RANGE
T (Jun) VALUES
SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged
monthlyAverage
T (Jul-Sep) seasonalAverage
T (Jul-Sep 1983) (Jul-Sep 2013) RANGE
T -2 shiftGRID

And if we look back into expert mode tab again as seen in Figure 1.6, we now have a match between the index and the variable.

From here, the user can add the correlate function:

SOURCES .Indices .nino .EXTENDED .NINO34
T (Jan 1983) last RANGE

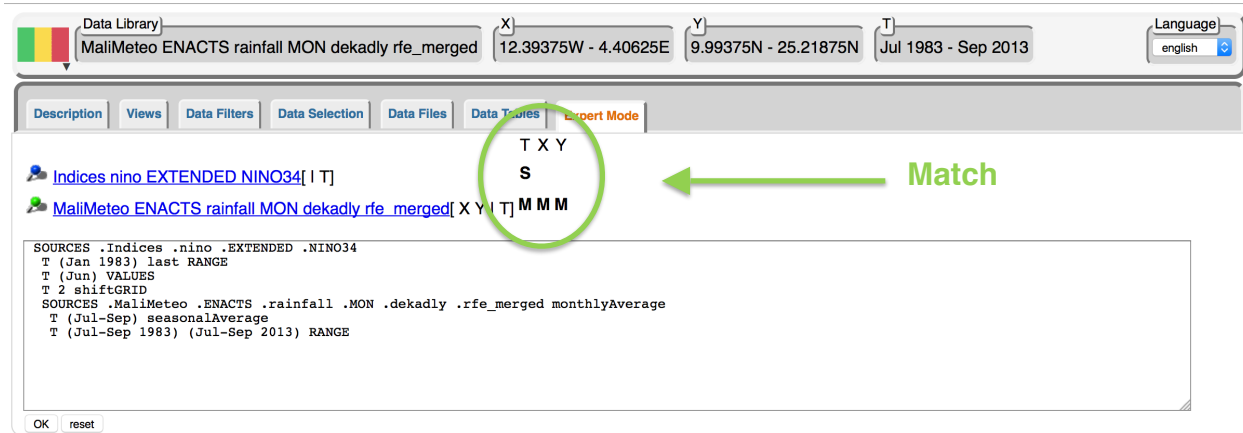


Fig. 1.6: Match between Index and Variable

T (Jun) VALUES

T 2 shiftGRID

SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged
monthlyAverage

T (Jul-Sep) seasonalAverage

T (Jul-Sep 1983) (Jul-Sep 2013) RANGE

[T] correlate

And obtain the following result for the correlation between Niño 3.4 index and monthly rainfall variables (Figure 1.7):

To have a better visualization, we can add a [maskrage] function by adding the last line:

SOURCES .Indices .nino .EXTENDED .NINO34

T (Jan 1983) last RANGE

T (Jun) VALUES

T 2 shiftGRID

SOURCES .MaliMeteo .ENACTS .rainfall .MON .dekadly .rfe_merged
monthlyAverage

T (Jul-Sep) seasonalAverage

T (Jul-Sep 1983)

(Jul-Sep 2013) RANGE

[T] correlate

-0.3 0.3 maskrage

So, what there are two families of useful functions to know when using mask: * mask – maskgt, maskge, masklt, maskle, maskrange, masknotrange * flag – flaggt, flagge, flaglt, flagle

Mask replaces the values that satisfy the condition by “Missing value” [NaN] whereas Flag replaces the values that satisfy the condition by 1 and those that do not satisfy the condition by 0. (Figure 1.8)

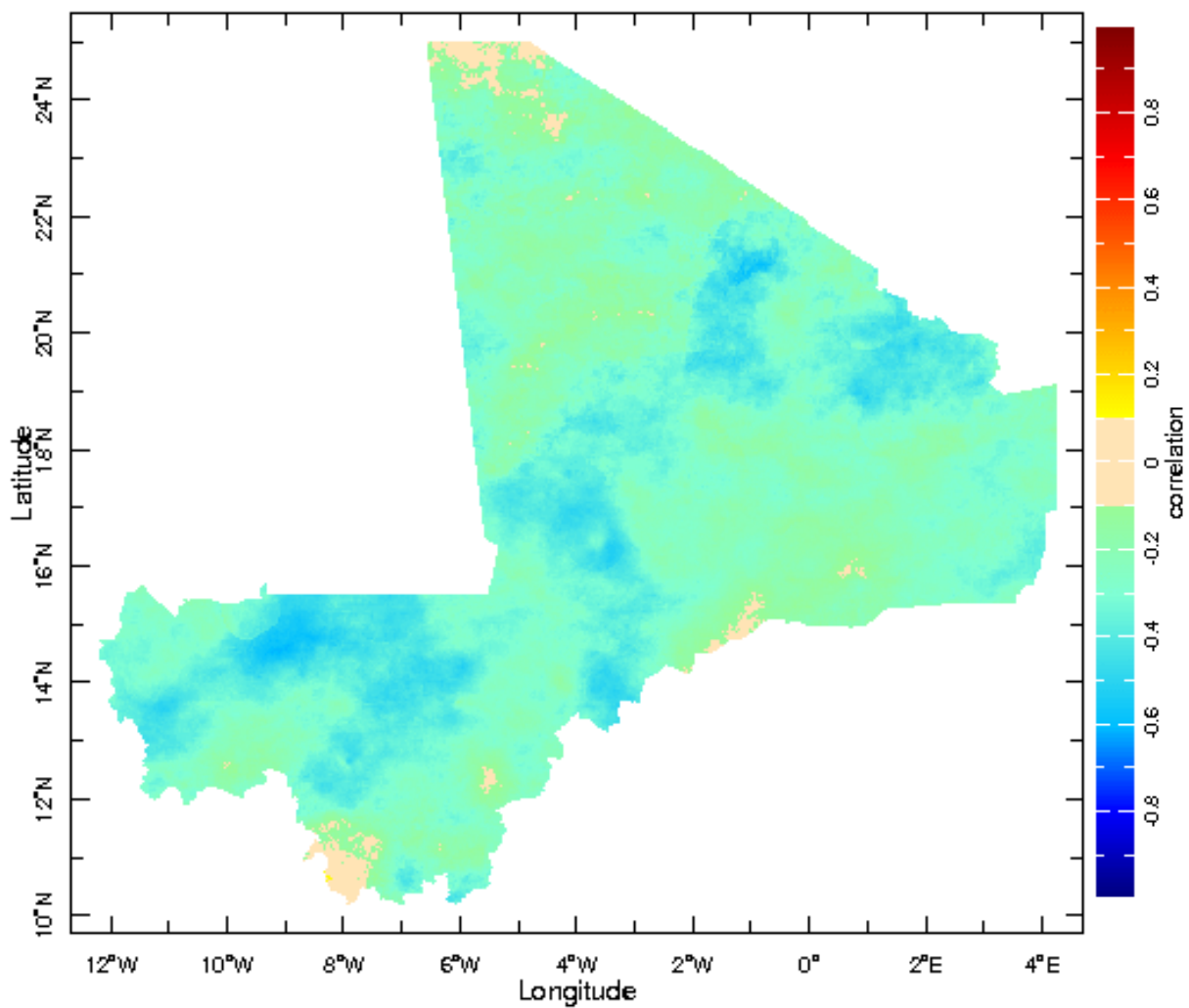


Fig. 1.7: Correlation between Niño 3.4 index and monthly rainfall variables in Mali

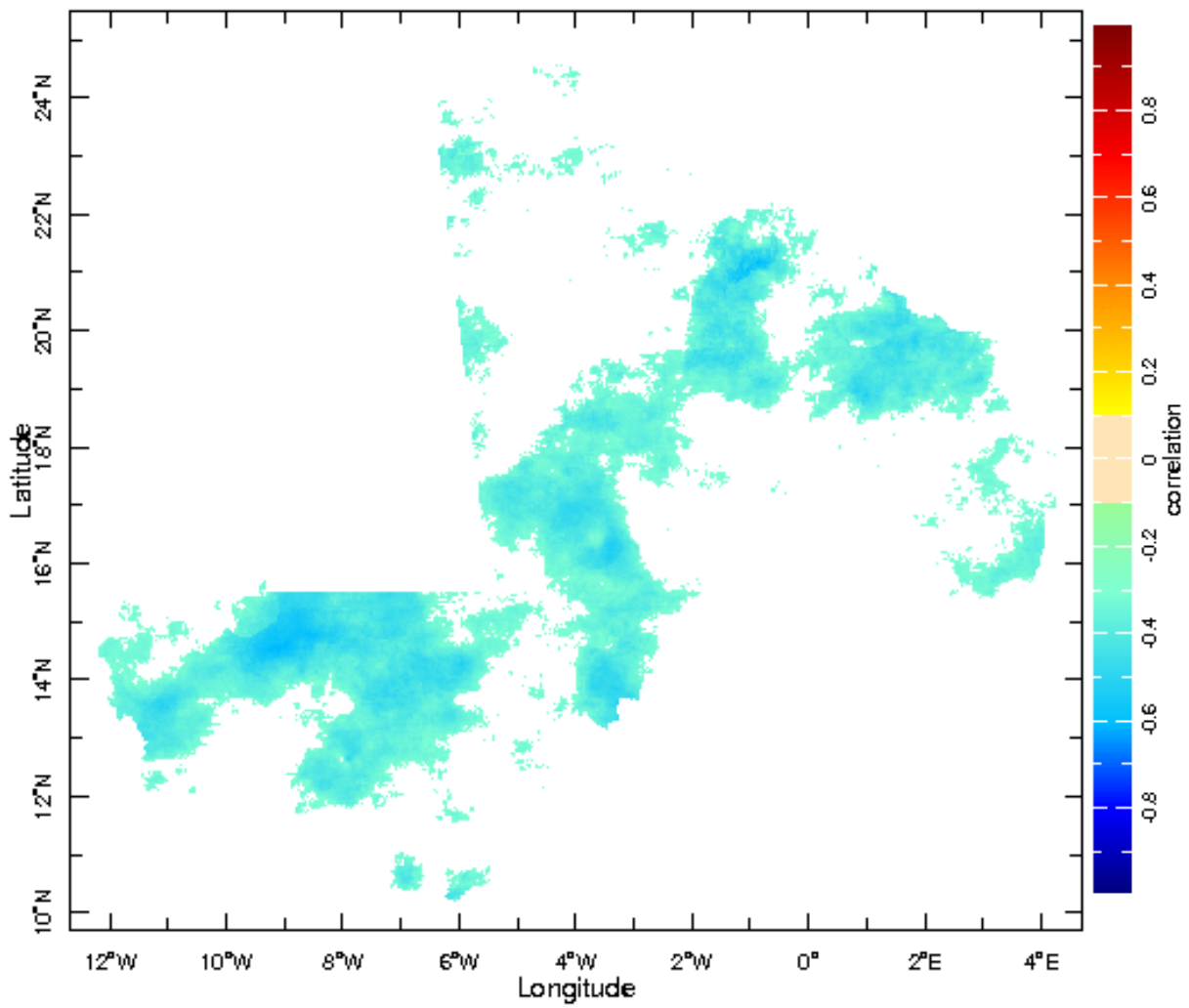


Fig. 1.8: Mask and Flag functions

And the results from the mask are seen in Figure 1.9:

maskrange

Masks out all values of a variable included in the indicated range.

lat -20 20 maskrange

See Also

Masks: : [flagge](#) [flaggt](#) [flagle](#) [flaglt](#)

Arguments		
label	type	Description
variable	variable	variable on which mask will be applied
range_min	number	lower threshold of range
range_max	number	upper threshold of range
restricted_var	output variable	variable with all values inside range specified by <i>range_min</i> and <i>range_max</i> masked out

Fig. 1.9: Correlation between Niño 3.4 index and monthly rainfall variables in Mali with Mask function

1.7 Summary

From this module the user is expected to have knowledge on how to select visualization options that can be animated and customized accordingly.

1.8 Quiz

Please answer the following questions using the IRI Data Library

- Q1. What steps can you take in case of an offset between two variables?
- Q2. What is the function difference between Correlation and Regression?
- Q3. Why do we use [average sub]?

1.8.1 Quiz - Answers

- A1. Going into expert mode the you can shift the grid according to the offset value given.
- A2. Correlation uses [correlate] function whereas regression uses [average sub].
- A3. [average sub] is used because we do not want to have the average subtracted.

1.9 Reference(s)

•