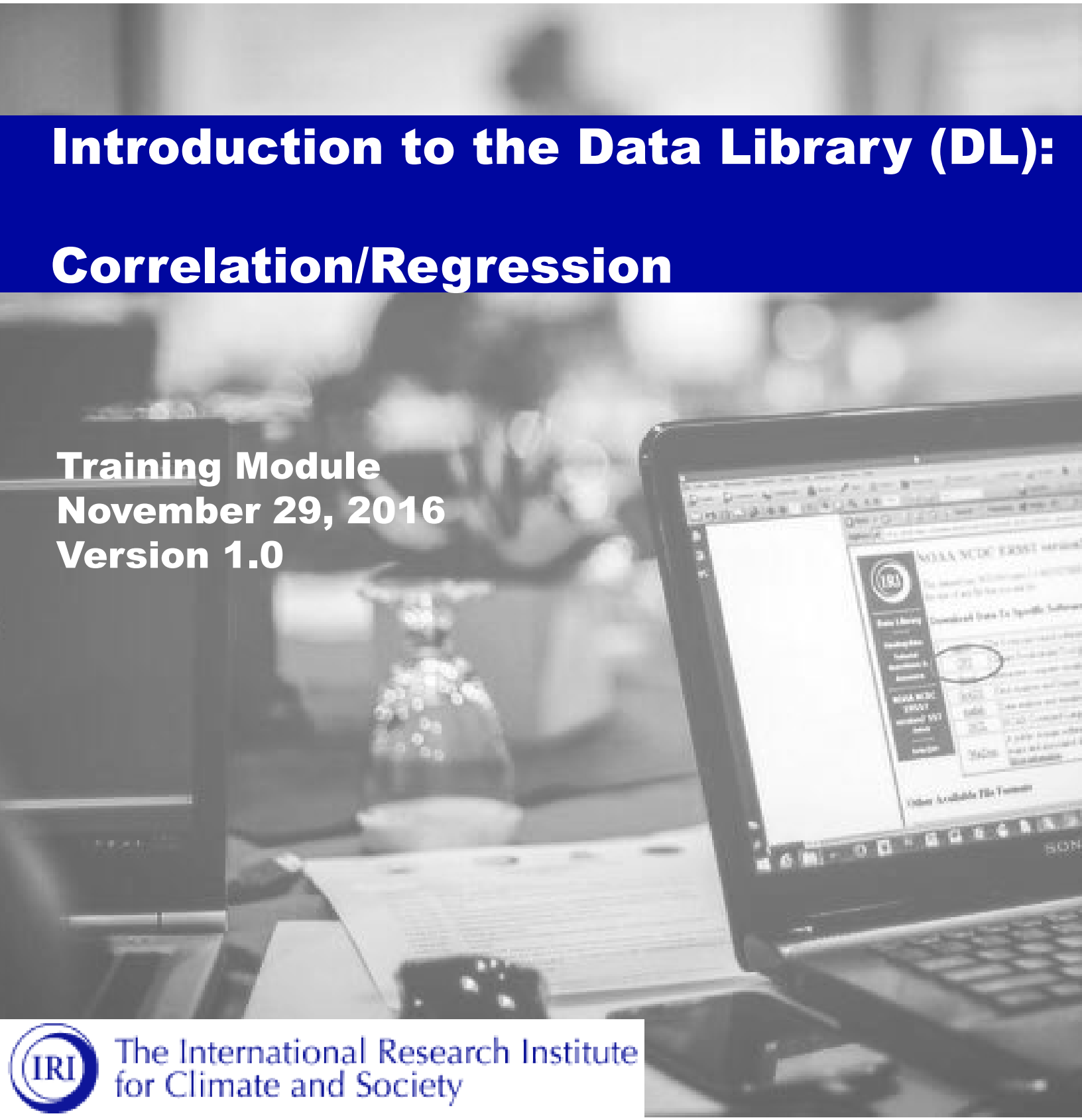
6.

DRAFT

Introduction to the Data Library (DL):

Correlation/Regression



Training Module
November 29, 2016
Version 1.0

The International Research Institute for Climate and Society

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

CONTENTS

CHAPTER

ONE

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 (Bluementhal, 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)

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

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

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

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

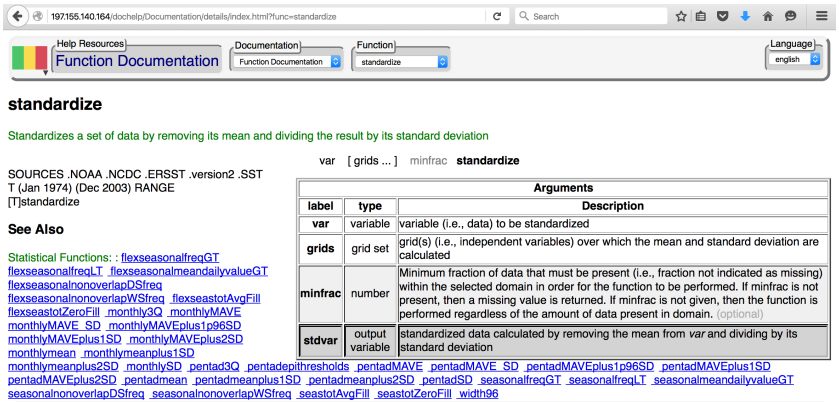
T (May-Oct 1983) (May-Oct 2013) RANGE

[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)



b)

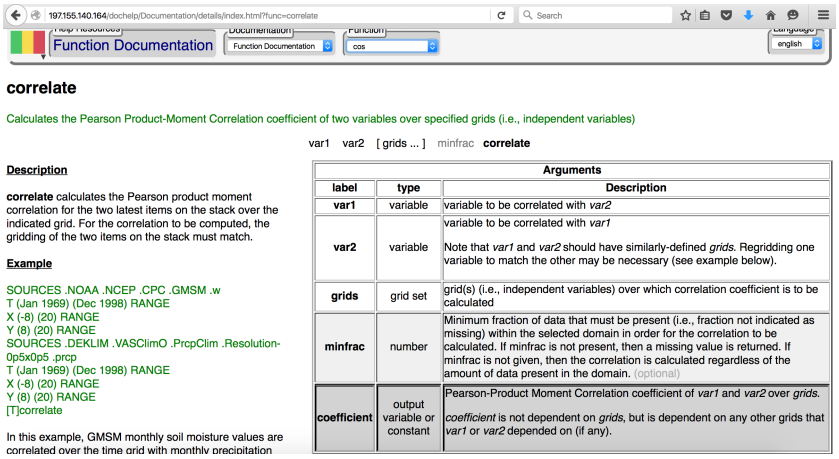


Fig. 1.2: Function Documentation used (a) standardize (b) correlate

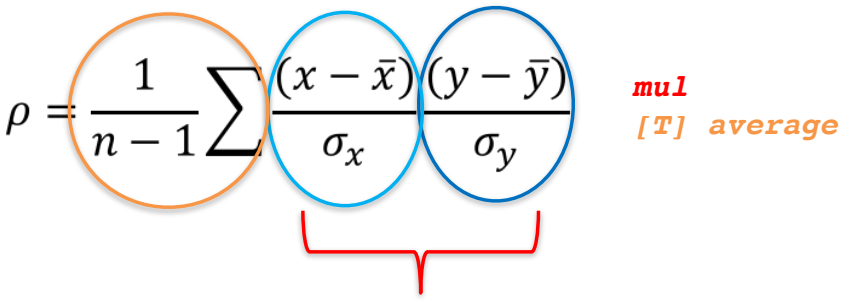


Fig. 1.3: Equation referenced for correlation

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.

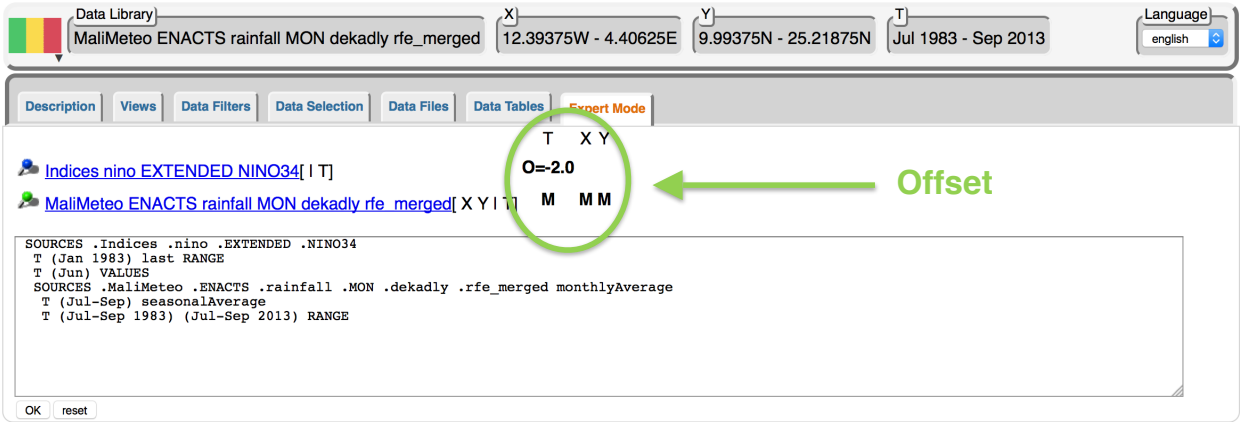


Fig. 1.4: Offset of Index and Variable

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:

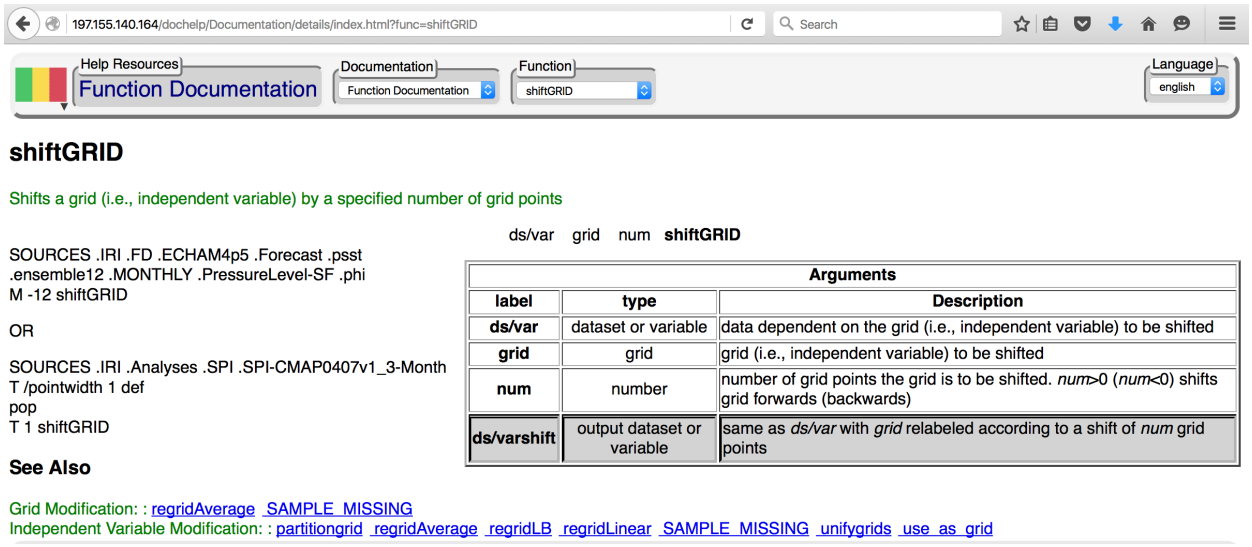


Fig. 1.5: [shiftGRID] Function in Function Documentation

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

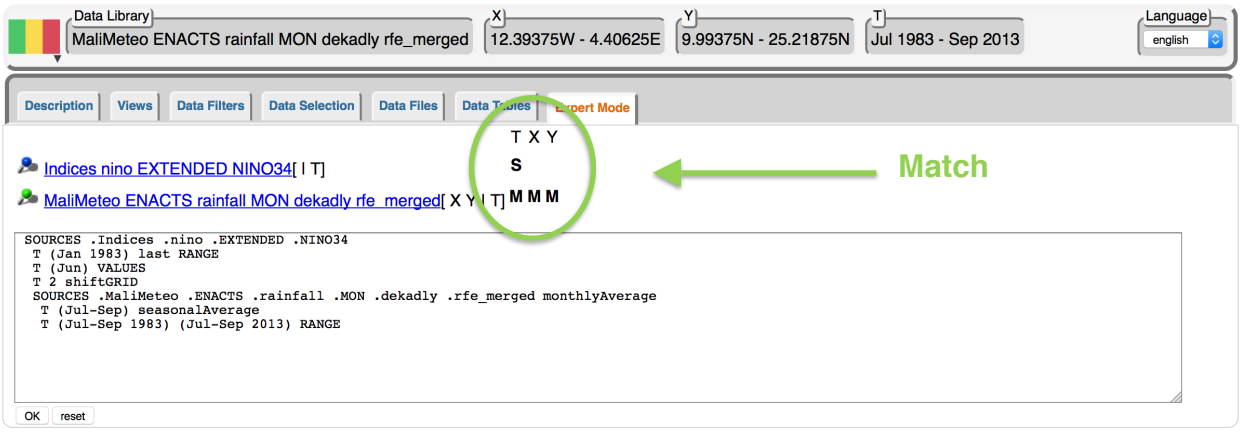


Fig. 1.6: Match between Index and Variable

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 maskrange

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)

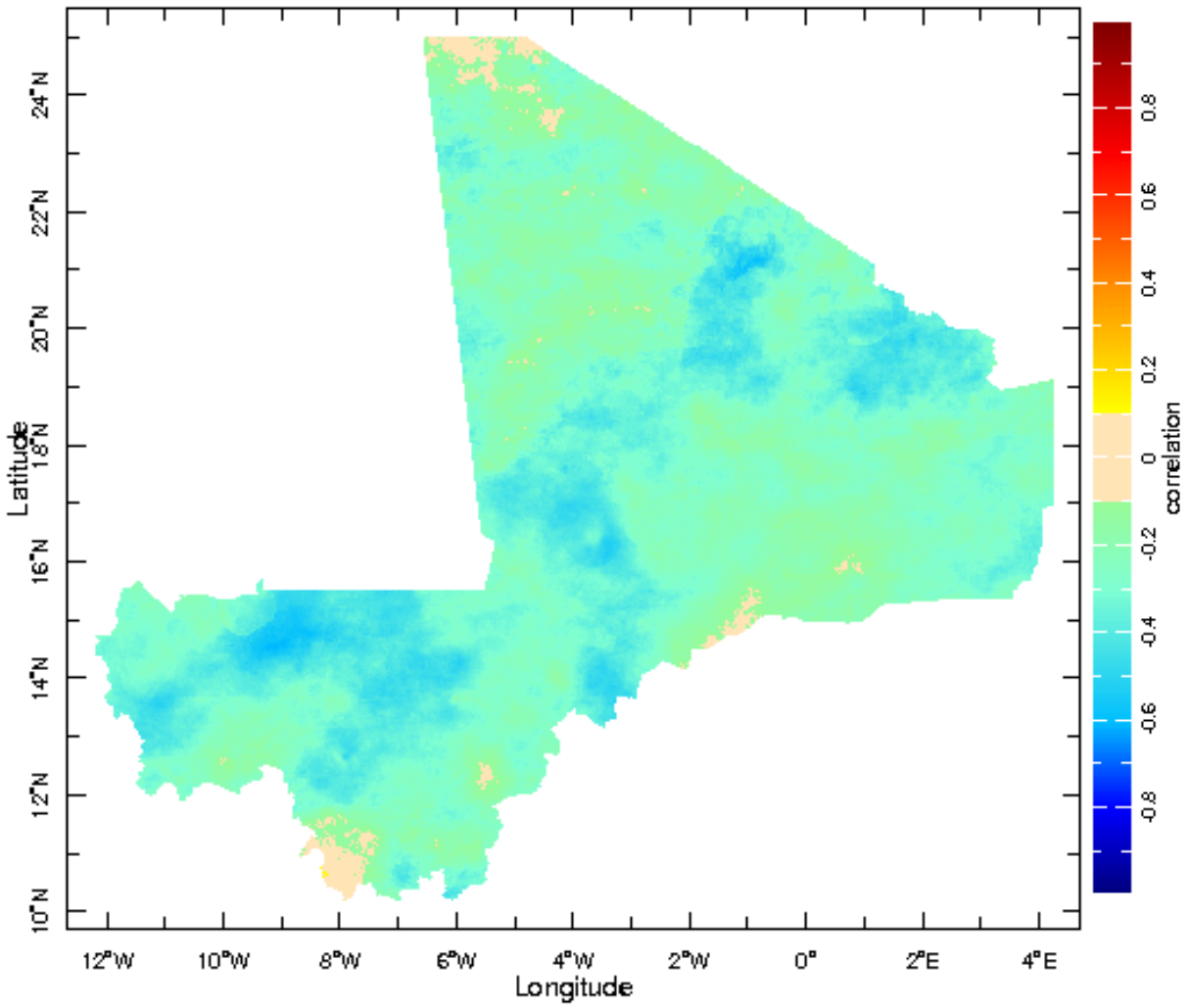


Fig. 1.7: Correlation between Niño 3.4 index and monthly rainfall variables in Mali

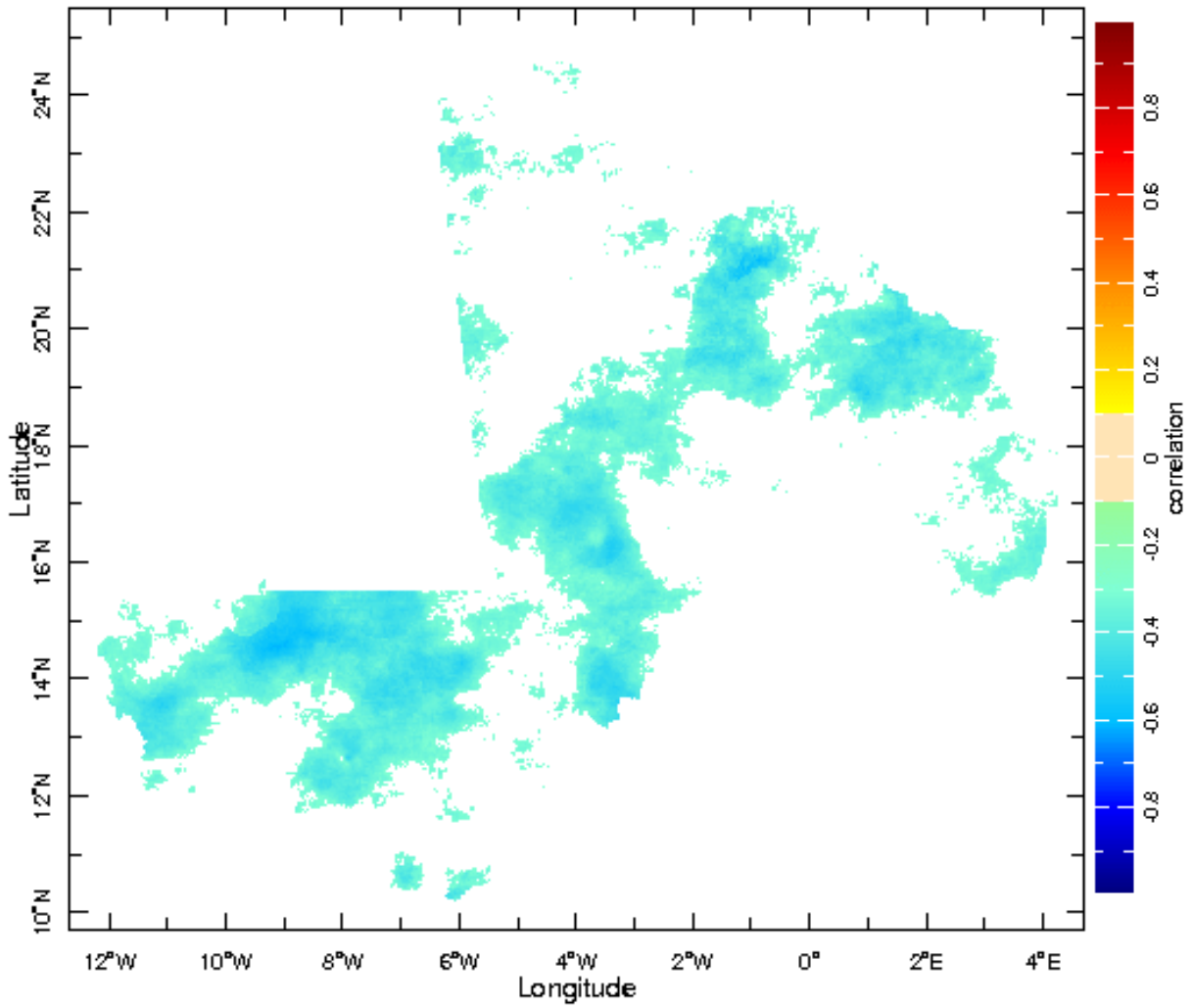


Fig. 1.8: Mask and Flag functions

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

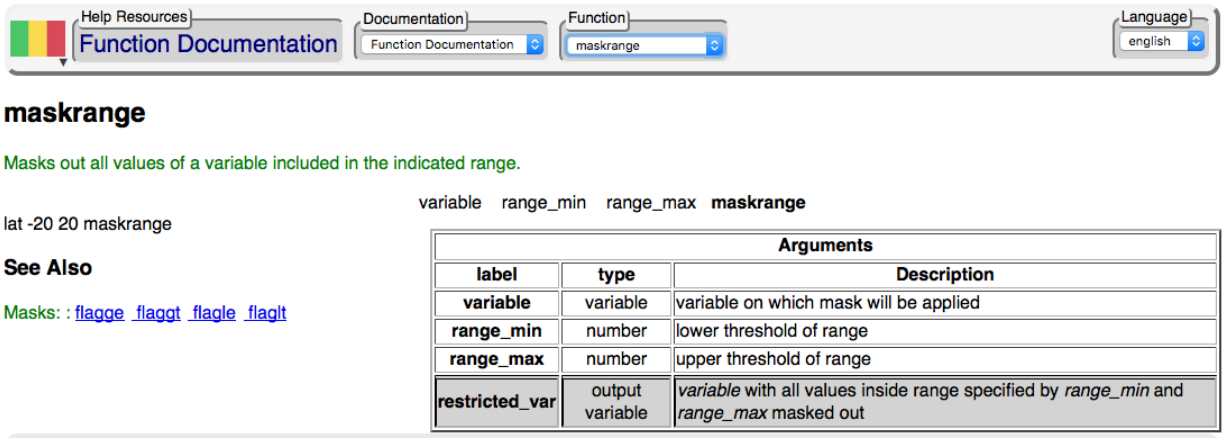


Fig. 1.9: Correlation between Niño 3.4 index and monthly rainfall variables in Mali with Mask function

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)

-