# Package 'EIA'

August 3, 2018

**Version** 1.5

**Title** Environmental Impact Assessment

**Author** Cástor Guisande González (Universidad de Vigo), Andrés Julián Rueda Quecho (Fundación Natura), Fabian Rangel Silva (Fundación Natura) & Jorge Mario Ríos Vasquez (ISAGEN).

**Maintainer** Cástor Guisande González <castor@uvigo.es>

**Description** It is estimated the assessment of the environmental consequences (positive and negative) of any kind of impact by comparing the environmental variables selected by the user between the before and after the impact.

**License** GPL (>= 2)

**Encoding** latin1

**Depends** R (>= 3.2)

**Suggests** ade4, car, IDPmisc, kulife, pROC, tseries, VARSEDIG

**Repository** CRAN

## R topics documented:

---

EIA *ENVIRONMENTAL IMPACT ASSESSMENT*

---

### Description

An algorithm to evaluate statistically an Environmental Impact Assessment (EIA).

## Usage

```
EIA(data, variables, period, before, after, site, date, VARSEDIG=FALSE,
minimum=FALSE, jpg=TRUE, removejpg=FALSE, p=0.05, cor=TRUE, ellipse=FALSE,
convex=TRUE, SCATTERPLOT=NULL, ROC=NULL, ROCTEST=NULL, PLOTOUTLIERS=NULL,
PLOTROC=NULL, ResetPAR=TRUE, PAR=NULL, XLAB=NULL, YLAB=NULL, XLIM=NULL,
YLIM=NULL, PCH=c(15,17,16), COLORD=rev(rainbow(2,alpha=0.4)), COLOR=NULL,
COLORC=NULL, LEGEND=NULL, MTEXT=NULL, TEXT=NULL, arrows=TRUE, larrow=1,
ARROWS=NULL, TEXTa=NULL, file1="Polar coordinates.csv",
file2="Outliers after.csv", na="NA", dec=",", row.names=FALSE)
```

## Arguments

`data` Data file.

`variables` Variables to be analysed.

`period` Variable with the categories of the periods. It must be two periods: the before and after the impact.

`before` Name of the period before the impact.

`after` Name of the period after the impact.

`site` Variable with the name of the sampling sites.

`date` Variable with the dates.

`VARSEDIG` If it is TRUE, the variables are added for the estimation of polar coordinates in the priority order according to the method "overlap" (see details section of the function VARSEDIG), and the variable is selected if it significantly contributes to discriminate between the before and after the environmental impact (EI).

`minimum` If it is TRUE and the argument *VARSEDIG=TRUE*, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. It is FALSE and the argument *VARSEDIG=FALSE*, the algorithm selects all significant variables, and not only those with higher discrimination capacity.

`jpg` It TRUE the scatterplots with the polar coordinates for each sampling time are exported to jpeg files.

`removejpg` If TRUE all jpeg files are deleted from the folder.

`p` Probability threshold of the Monte-Carlo test.

`cor` If it is TRUE the variables are ordered according to the correlation between them when estimating the polar coordinates. Therefore, the next variable to another variable is the one that has a greater positive correlation. The first variable is the one with the higher discrimination capacity between the before and after the EI, that it is estimated with the method "overlap" (see details section of the function VARSEDIG).

`ellipse` If it is TRUE the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *period* is depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

`convex` If it is TRUE the convex hull is depicted for each category.

`SCATTERPLOT` It accesses the function scatterplot of the car package of the graph *biplot* with the polar coordinates.

`ROC` It accesses the function roc of the pROC package.

`ROCTEST` It accesses the function roc.test of the pROC package.

`PLOTOUTLIERS` It allows to specify the characteristics of the function plot.default of the plot with the outliers.

`PLOTROC` It allows to specify the characteristics of the function plot.default of the plot with the receiver operating characteristic curves (ROC curves).

`ResetPAR` If it is FALSE, the default condition of the function PAR is not placed and maintained those defined by the user in previous graphics.

`PAR` It accesses the function PAR that allows to modify many different aspects of the graph.

`XLAB` Legend of the X axis in the scatterplot of the polar coordinates.

`YLAB` Legend of the Y axis in the scatterplot of the polar coordinates.

`XLIM` Vector with the limits of the X axis in the scatterplot of the polar coordinates.

`YLIM` Vector with the limits of the Y axis in the scatterplot of the polar coordinates.

`PCH` It allows to modify the symbols of the scatterplot. It must be three symbols: the before and after the impact, and the outliers after the impact.

`COLORD` It allows to modify the colours of the density plot. It must be two colours: the before and after the impact.

`COLOR` It allows to modify the colours of the symbols in scatterplot. It must be three colours: the before and after the impact, and the outliers after the impact.

`COLORC` It allows to modify the colours of the convex hull in the scatterplot. It must be two colours: the before and after the impact.

`LEGEND` It allows to modify the legend in the scatterplot of the polar coordinates.

`MTEXT` It allows to add text on the margins in the scatterplot of the polar coordinates.

`TEXT` It allows to add text in any area of the inner part in the scatterplot of the polar coordinates.

`arrows` If it is TRUE the arrows are shown in the scatterplot with the polar coordinates. These arrows show the vector of the variables selected when calculating the polar coordinates.

`larrow` It modifies the length of the arrows.

`ARROWS` It accesses the function Arrows of the package IDPmisc, which performs the arrows.

`TEXTa` It allows to modify the labels at the end of the arrows.

`file1` CSV FILES. Filename with the polar coordinates for all sampling times and sampling sites.

`file2` CSV FILES. Filename with the outliers after the EIA.

`na` CSV FILE. Text that is used in the cells without data.

`dec` CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".

`row.names` CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

## Details

### EIA Algorithm

### 1. Polar coordinates

All variables are transformed to a scale ranging between -1 and 1. For each sampling site and day the X and Y polar coordinates are estimated using the following equations:

$$X = \sum_{i=1}^{n} |z_j| cos(\alpha) \quad Y = \sum_{i=1}^{n} |z_j| sin(\alpha)$$

where *z* is the value of the variable *j* and *n* the number of variables.

Each variable is assigned an angle ($\alpha$). The increment value of the angle is always $\frac{360}{n*2}$. If for instance the number of variables is five, the increment angle is 36. Therefore, for the first variable if the value is $\geq 0$ the $\alpha$ value is 36 and if the value is $< 0$ the value is 36+180, for the second variable if the value is $\geq 0$ the $\alpha$ value is 36+36 and if the value is $< 0$ the value is 36+36+180, etc. Conversion of degrees to radians angle is carried out assuming that 1 degree = 0.0174532925 radians.

The order of the variables is consequently important, as a different alpha value is assigned. If the argument *cor=TRUE*, this order is established by calculating the correlation matrix of the variables and by ordering them such that each variable is followed by the variable to which it is highly correlated. The goal is to favor a larger dispersion of data in the resulting polar coordinates system. The first variable is the one selected in the following step. If the argument *cor=FALSE*, the order is established by the user when introducing variables.

### 2. Prioritizing variables by their capacity for discrimination

If the argument *cor=TRUE*, the overlap method described in the VARSEDIG) algorithm (Guisande. 2016; Guisande et al., 2016) is used to prioritize the variables in accordance with their capacity for discrimination. A density curve is obtained for each variable and the overlap of the area under the curve between the before and after the EI is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; from highest to lowest discrimination capacity.

### 3. Selection of those variables which discriminate significantly between the two periods

If the argument *VARSEDIG=TRUE*, the means of both the X and Y polar coordinates are estimated for the values obtained before and after the EI. Via these means the Euclidean distance is calculated between said periods.

The Monte-Carlo test is used to assess testing the statistical hypothesis if the value of one group is significantly higher or lower that the values of the other group. The test is performed for both the X and Y polar coordinates and compares all values in one group to those of the other group. For instance, when all values of group 1 are compared to those of group 2, and the mean X polar coordinate of group 1 is higher than the mean of group 2, the alternative hypothesis of the Monte-Carlo test is greater, and the p-value is estimated as (number of random values equal to or greater than that observed one + 1)/(number of permutations + 1). The null hypothesis is rejected if the p-value is less than the significance level. If the mean X polar coordinate of group 1 is lower than the mean of group 2 and the alternative hypothesis is smaller, the p-value is estimated as: number of random values equal to or less than that observed + 1/number of permutations + 1. Again, the null hypothesis is rejected if the p-value is lower than the significance level. The same process is applied when comparing all values from group 2 with those of group 1.

If the argument *min=TRUE* and the argument *VARSEDIG=TRUE*, a variable is selected if both of the following criteria are met: 1) they contribute to an increase in the Euclidean distance between both groups compared to the Euclidean distance obtained with the set of previously selected variables; and 2) the Monte-Carlo test p-values for the X and Y coordinates when comparing both group 1 to group 2 and group 2 to group 1 are smaller than the p-values obtained with the set of previous selected variables. Therefore, from the pool of all variables, only those variables with the highest significant contribution to discrimination between both periods are selected; that is, the minimum possible number of variables necessary for achievement of maximum discrimination between the periods.

If the argument *minimum=FALSE* and the argument *VARSEDIG=TRUE*, a variable is selected if it contributes to an increase the in Euclidean distance and the p-values are equal to, or less than, the p-values obtained with the set of previously selected variables. Therefore, all significant variables are selected from the pool of all variables, rather than solely from those with the highest discrimination capacity.

**4. Identification of sampling sites for each sampling time after the EI, which are significantly different than before the EI**

A Monte-Carlo test is used to determine whether the X and/or Y polar coordinates obtained for each sampling time and site, after the EI, are significantly different than the mean X and Y polar coordinates of all values, obtained prior to the EI (hereinafter outliers). If the p value of the Monte-Carlo test is lower than 0.05 (this threshold probability may be modified by the user) for the X and/or Y polar coordinates, said site in the specified sampling period after the EI is considered to be an outlier. In other words, it is significantly different than the values obtained before the EI.

**5. Testing whether the percentage of outliers increases following the EI**

Dickey & Fuller (1979) developed a procedure which enables testing of whether a variable has a unit root, or equivalently, the variable follows a random walk. There is an extension of the Dickey-Fuller test called the augmented Dickey-Fuller test (ADF), which removes all structural effects (autocorrelation) in the time series. This test was used to determine whether there was a significant increasing trend in the percentage of outliers after the EI.

**6. Estimation of the Area Under the Curve of Receiver Operating Characteristic curves**

The Area Under the Curve (AUC) of Receiver Operating Characteristic curves (ROC curves) is a measure of how well two groups can be differentiated. The AUC of the two ROC curves, obtained both before and after the EI, were estimated, for both the X and Y polar coordinates, considering all sampling sites and times.

Finally, the AUC of the ROC curves (before and after the EI), obtained for both the X and Y polar coordinates, were compared to an AUC of 0.5 (the situation in which two distributions are equal). It can be assumed that, if the probability obtained in the comparison test is lower than 0.05 for the X and/or Y polar coordinates, the EI had a significant positive or negative impact on some of the variables analyzed.

**FUNCTIONS**

The plot with the density curves is performed with the function plot.default of base graphics package. The density curve is estimated with the function density of base stats package. The area under the curve is estimated with the function auc of the package kulife (Ekstrom et al., 2015). The Monte-Carlo test was performed with the function as.randtest of the package ade4 (Chessel et al., 2004; Dray et al., 2007; 2016). The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The scatterplot is performed with the function scatterplot of the car package (Fox & Weisberg, 2011; Fox et al., 2014). The ROC curves are built with the function roc and the function roc.test is used to compare the AUC of the ROC curves, being both functions of the package pROC (Robin et al., 2017). The augmented Dickey-Fuller test was performed with the function adf.test of package tseries (Trapletti & Hornik 2017). The convex hull is estimated with the function chull of the package grDevices.

## EXAMPLES

As a demonstration of the potential of the algorithm EIA, monthly climatic data (precipitation, wind speed, temperature, relative humidity, evaporation and solar radiation) were used from one station, and quarterly agronomic data (healthy cocoa pods, frosty pod rot *Moniliophthora roreri*, black pod rot *Phytophthora sp* and witch's broom *Moniliophthora perniciosa*) ere taken from several sampling sites near the Sogamoso reservoir (Santander, COLOMBIA). This was carried out from January 2012 to June 2014 and from July 2014 to February 2017, before and after the reservoir was filled, respectively. Therefore, the EI was the filling up of the reservoir, and the goal was to evaluate potential changes in climatic and agronomic variables after, following this EI.

It is explained only the example 1, where climatic data and all variables are included in the analysis (argument *VARSEDIG=FALSE*). The polar coordinates obtained the last month of this study shows that there were some months, after the EI, which were statistically different than periods before the EI (Fig. 1). These outliers occurred due high temperatures, low precipitation, high wind speed and low humidity (Fig. 1).

**Figure 1.** Polar coordinates for all months before and after the reservoir was filled, and the months after the reservoir was filled significantly different (outliers) than values obtained before. Ellipses with the levels of significance of 0.5(inner ellipse) and 0.95 (outer ellipse) for each period.

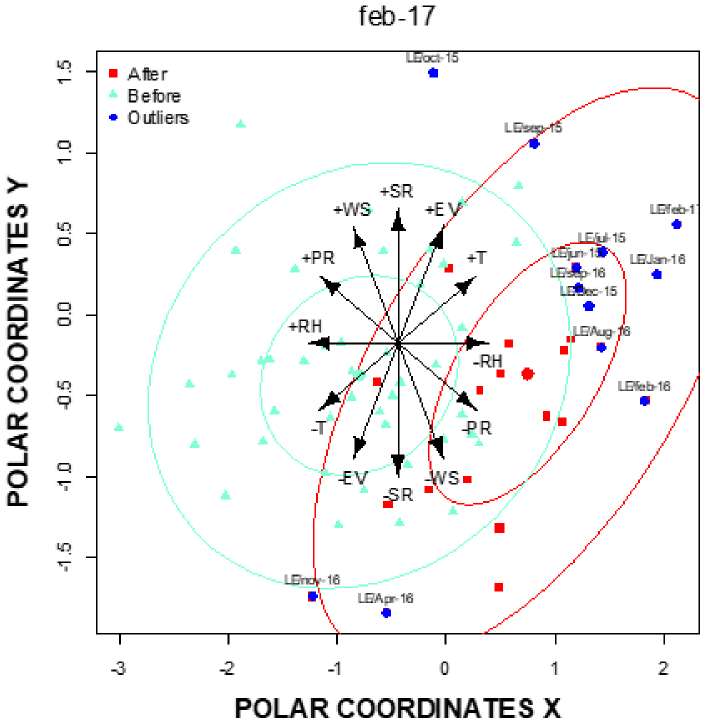


**Figure 2. Left panels:** Density plots of the polar coordinates X (upper left panel) and Y (lower left panel) showed in Figure 1 for the periods before and after the reservoir was filled. **Right panels:** Area Under the Curve (AUC) of Receiver Operating Characteristic curves (ROC curves) obtained for before and after the EI, for both polar coordinates X (upper right panel) and Y (lower right panel) showed in Figure 1.

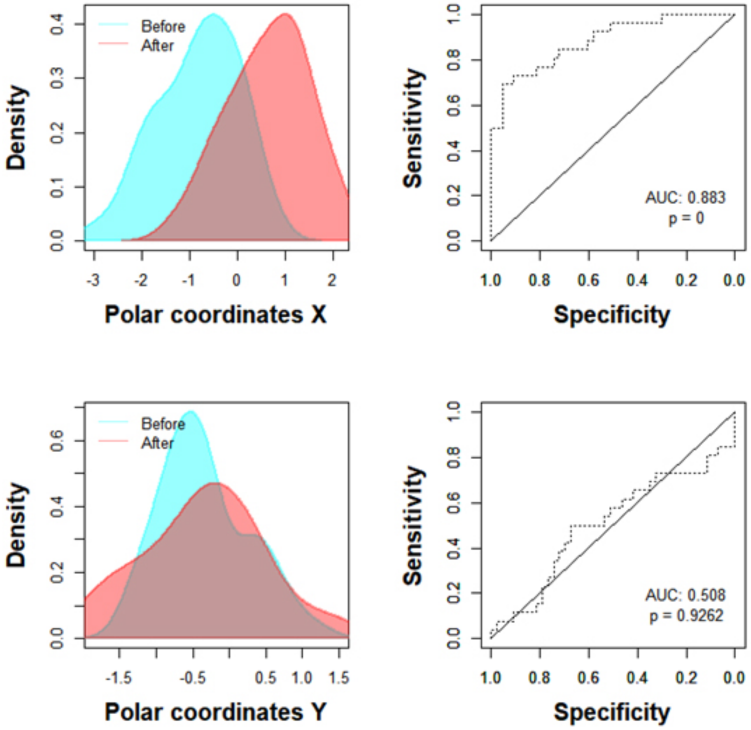
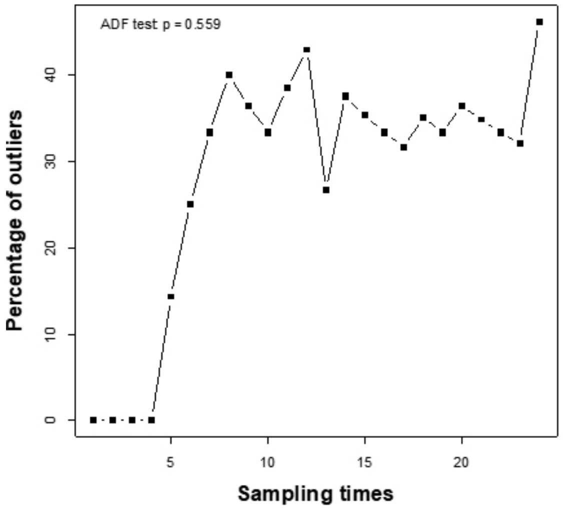


**Figure 3.** Percentage of outliers for each month after the reservoir was filled. Probability is of the augmented Dickey-Fuller test (ADF).



The AUC of the ROC curve for polar coordinates X was 0.883, which is significantly different than an AUC of 0.5 (p < 0.001, Figure 2). In fact, the percentage of outliers increase after the reservoir was filled (Fig. 3). However, the probability of the augmented Dickey-Fuller test (ADF) was 0.559 (Fig. 3), and so the null hypothesis, that the time series was stationary, was accepted. It means that, although it seems that there is a trend toward a significant change in the climatic variables as a whole after the reservoir was filled -an increasing number of outliers (Fig. 3)- this trend is not significant yet. Therefore, a longer time series would be necessary in order to corroborate whether this change in climatic variables was due to the effect of filling the reservoir, or was natural induced.

## Value

It is obtained: 1) the CSV file with the polar coordinates for all sampling times and sampling sites; 2) the CSV file with the outliers after the EIA; 3) the plots with polar coordinates for each sampling period highlighting the outliers 4) the plot with the percentage of outliers after the impact for each sampling time and 5) a panel with 4 plots, two density plots with the overlap between the before and after the EI for both the polar coordinates X and Y, and the two ROC curves for also the polar coordinates X and Y.

## Author(s)

Cástor Guisande González (Universidad de Vigo), Andrés Julián Rueda Quecho (Fundación Natura), Fabian Rangel Silva (Fundación Natura) and Jorge Mario Ríos Vasquez (ISAGEN)

## References

Chessel, D. and Dufour, A.B. and Thioulouse, J. (2004) The ade4 package-I- One-table methods. *R News*, 4, 5-10.

Dickey, D.A. & Fuller, W.A. (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, 427-431.

Dray, S. & Dufour, A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22, 1-20.

Dray, S. & Dufour, A.B. and Chessel, D. (2007) The ade4 package-II: Two-table and K-table methods. *R News*, 7(2), 47-52.

Dray , S., Dufour , A-B. & Thioulouse , J. (2016) Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences. R package version 1.7-5. Available at: https://CRAN.R-project.org/package=ade4.

Ekstrom, C., Skovgaard, Ib M. & Martinussen, T.(2015) Datasets and functions from the (now non-existing). R package version 0.1-14. Available at: https://CRAN.R-project.org/package=kulife.

Fox, J. & Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2016) Companion to Applied Regression. R package version 2.1-4. Available at: https://CRAN.R-project.org/package=car.

Guisande, C. 2016. An algorithm for morphometric characters selection and statistical validation in morphological taxonomy. R package version 1.1. Available at: https://CRAN.R-project.org/package=VARSEDIG.

Guisande, C., Vari, R.P., Heine, J., García-Roselló, E., González-Dacosta, J., Pérez-Schofield, B.J., González-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162, 571-580.

Locher, R. & Ruckstuhl, A. (2015) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: https://CRAN.R-project.org/package=IDPmisc.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J-C, Müller, M., Siegert, S. & DeLong, F. 2017. Display and Analyze ROC Curves. R package version 1.1.1. Available at: https://CRAN.R-project.org/package=pROC.

Trapletti, A & Hornik, K. (2017) tseries: Time Series Analysis and Computational Finance. R package version 0.10-38. Available at: https://CRAN.R-project.org/package=tseries.

## Examples

```
#Example 1: Atmospheric data

data(MALE)

EIA(data=MALE, variables=c("PR","WS","T","RH","EV","SR"), period="Period", before="Before",
after="After", site="Station", date="Date", jpg=FALSE, ellipse=TRUE,
convex=FALSE, ROCTEST=c("method='delong'"))

## Not run:

#Example 2: As in example 1 but with convex

data(MALE)

EIA(data=MALE, variables=c("PR","WS","T","RH","EV","SR"), period="Period", before="Before",
after="After", site="Station", date="Date", jpg=FALSE, ROCTEST=c("method='delong'"))

#Example 3: Agronomic data

data(MAFitCC)

#With VARSEDIG FALSE
EIA(data=MAFitCC, variables=c("HC","FPR","BPR","WB"), period="Period", before="Before",
after="After", site="Patch", date="Date", removejpg=TRUE)

#With VARSEDIG TRUE
EIA(data=MAFitCC, variables=c("HC","FPR","BPR","WB"), period="Period", before="Before",
after="After", site="Patch", date="Date", VARSEDIG=TRUE)


## End(Not run)
```

## `MAFitCC` *PHYTOPATHOLOGICAL DATA OF CLONED COCOA*

### Description

Phytopathological database of cloned cocoa from surrounding areas of Sogamoso reservoir (Santander, COLOMBIA).

### Usage

```
data(MAFitCC)
```

### Format

Data frame with 7 variables: the period (before and after the Sogamoso reservoir was filled), the sampling patch, the date, healthy corncobs (HC), frosty pod rot (*Moniliophthora roreri*, FPR), black pod rot (*Phytophthora* sp., BPR) and witch's broom (*Moniliophthora perniciosa*, WB).

## `MALE` *CLIMATE DATA*

### Description

Climatic database of a sampling station nearby Sogamoso reservoir (Santander, COLOMBIA).

### Usage

```
data(MALE)
```

### Format

Data frame with 9 variables: the period (before and after the Sogamoso reservoir was filled), the sampling station, the date, precipitation (PR, in mm), wind speed (WS, in $ms^{-1}$), temperature (T, in °C), relative humidity (RH, in percentage), evaporation (EV, in mm) and solar radiation (SR, in $Wm^{-2}$).

# Index