Detecting changes in essential ecosystem and biodiversity properties - towards a Biosphere Atmosphere Change Index: BACI

Deliverable 5.4: Methods for Attribution Scheme and Near Real-Time BACI

Project title:	Detecting changes in essential ecosystem and biodiversity properties- towards a Biosphere Atmosphere Change Index
Project Acronym	BACI
Grant Agreement Number:	640176
Main pillar:	Industrial Leadership
Topic:	EO-1-2014: New ideas for Earth-relevant space applications
Start date of the project:	1st April 2015
Duration of the project:	48 months
Dissemination level:	Public
Responsible of the deliverable:	Joachim Denzler Phone: +49 3641 9 46420 Email: joachim.denzler@uni-jena.de
Contributors:	Maha Shadaydeh, Yanira Guanche, Miguel Mahecha
Date of submission:	September 20, 2018

Contents

Summary

This deliverable is dedicated to the work done within the *WP5 - Synthetic Index and Attribution Scheme: the BACIndex*. WP5 is divided into four main tasks and this fourth report refers to the fourth one: *Task 5.4 - Methods for Attribution Scheme and Near Real-Time BACI*.

The focus of this fourth task can be divided into three main objectives. The first is the improvement of the anomaly detection method based on linear regression presented in Deliverable 5.3. The other two objectives are the development of two different attribution schemes. The first attribution scheme is based on the decomposition of the change index, which is the Mahalanobis distance in the presented method. This scheme aims to mathematically decompose the Mahalanobis distance into as much components as the number of variables used. Each component in the decomposition reflects how much the corresponding variable contributes to the overall Mahalanobis distance, or i.e. change index. While the first attribution scheme is based on mathematical concept, the second attribution scheme is based on the analysis of the dynamical behaviour of the system, i.e. the intensities of the cause-effect relationships between the underlying variables in the system. Each section of this deliverable corresponds to one of these three objectives:

1. **Advances on the BACIndex, Near Real-Time BACI** In our Deliverable 5.3, we proposed a method for anomaly detection based on linear regression and spatiotemporal Markov random field [1]. In this section we present the advances on this method based on the use of single multivariate autoregressive model (MVAR) instead of multiple univariate autoregressive moving average models. The multivariate autoregressive model allows for presenting the variables with a model that takes into account their inter-dependency and hence enables better whitening of the residuals, i.e. zero cross-correlation coefficients between different residuals. This in turn results in improved spatial and temporal detection accuracy of the method. The improved method based on MVAR model is implemented in a sliding time window approach with very low computational load making it suitable for near real time implementation.

2. **Attribution Scheme based on Mahalanobis Distance Decomposition** In this section we present an attribution scheme based on the decomposition of the Mahalanobis distance. The decomposed form of the Mahalanobis distance provides the answer to the question: how much each variable has contributed to the Mahalanobis distance? The mathematical procedure for the decomposition, known as Garthwaite Transform [2], is explained in detail in this section followed by the experimental results of the developed attribution scheme used for the attribution of different known historic events.

3. **Attribution Scheme based on Spectral MVAR Granger Causality** Local meteorological conditions have direct impact on CO2 fluxes and ecosystem respiration. Understanding the cause-effect relationships in such dynamical system is essential for the attribution of climate changes as well as for the development of intervention strategy to achieve desired prediction. In this section we present

a spectral multivariate Granger causality approach for the analysis of the cause effect relationships between the EO variables involved. The advantages of the proposed method is that it allows for causality analysis at different frequency components and hence different time scales.

The developed causality analysis can then be directly implemented for the attribution of detected changes. We show that anomalous events can be detected as those events where the dynamical behaviour, i.e. the cause-effect intensities between the variables, differ considerably from the average dynamical behaviour. The detected anomalous event can then be directly attributed to the variable(s) causing such deviation.

1 Advances on the BACIndex, Near Real-Time BACI

In our Deliverable 5.3, we proposed a method for anomaly detection based on linear regression and spatiotemporal Markov random field [1]. In the proposed method, the time series of each variable was assumed to follow univariate autoregressive moving average (ARMA) model. Five biosphere variables from a preliminary version of the Earth System Data Cube were used then: Gross Primary Productivity, Latent Energy, Net Ecosystem Exchange, Sensible Heat and Terrestrial Ecosystem Respiration. To tackle the spatiotemporal dependencies of the biosphere variables, the proposed methodology after preprocessing the data is divided into two steps: a feature extraction step applied to each time series in the grid independently, followed by a spatiotemporal event detection step applied to the obtained novelty scores over the entire study area. The first step is based on the assumption that the time series of each variable can be represented by an Autoregressive Moving Average (ARMA) process, and the anomalies are those time instances that are not well represented by the estimated ARMA model. The Mahalanobis distance of the ARMA models' multivariate residuals is used as a novelty score. In the second step, the obtained novelty scores of the entire study are treated as time series of images. The classification of the novelty score images into three classes, *intense anomaly*, *possible anomaly*, and *normal*, is performed using unsupervised K-means clustering followed by multitemporal MRF segmentation applied recursively on the images of each consecutive $L \geq 1$ time steps.

Based on this method, the classification maps of the whole BACI study area over 11 years (First version of BACIndex) were submitted to the team of WP6 to validate the spatial and temporal accuracy of the BACIndex. In their assessment, reported in Del 6.2, WP6 used 40 known extreme events to assess the temporal and spatial accuracy of the index. Table 1 summaries these events and shows the spatial and temporal accuracy of the first version of BACIndex.

Table 1: List of known extreme events provided by WP6.

ID	Type	Start	End	Spatial	Temporal	Min Lon	Max Lon	Min Lat	Max Lat
1	Drought	8/2004	6/2005	2	1	37.015 791 142 6	42.012 659 255 3	2.032 332 062 6	8.064 265 711 9
2	Drought	2/2008	12/2008	1	1	36.087 801 350 1	40.013 912 010	5.030 452 929 4	8.099 957 626 7
3	Drought	2/2008	12/2008	1	1	43.940 022 669 8	47.009 527 367 0	1.068 650 354 4	3.067 397 599 3
4	Drought	7/2011	11/2012	2	1	40.941 912 982 3	45.010 780 122 2	−0.037 799 012 7	4.029 815 974 9
5	Drought	11/2006	11/2006	2	2	36.016 417 520 6	38.050 856 680 2	1.032 958 439 6	4.066 771 222 2
6	Flood	9/2009	9/2009	2	2	26.744 944 736 9	30.044 928 443 5	39.715 653 431 2	41.776 276 07
7	Fire	7/2007	7/2007	1	1	29.664 018 062 0	31.305 846 156 7	−27.332 976 924 6	−22.821 518 857 6
8	Flood	1/2011	1/2011	3	2	29.635 153 388 4	32.574 098 872 7	−30.550 986 309 8	−26.885 638 2
9	Drought	4/2004	8/2004	1	1	29.338 031 904 8	31.527 136 030 2	−26.695 281 374 6	−24.506 177 249 2
10	Cold wave	1/2010	2/2010	1	1	13.596 702 191 3	18.319 670 876 9	56.200 634 588 8	64.508 026 667 5
11	Flood	5/2010	6/2010	2	2	10.344 956 625 4	21.928 374 696 4	46.465 409 495 5	53.864 180 566 1
12	Tree cover loss	5/1/2005	12/1/2005	1	1	13.549 713 265 5	15.334 309 020 1	56.808 333 432 7	57.379 404 074 1
13	Heatwave	7/2010	7/2010	3	2	33.644 261 459 5	52.387 562 508 9	47.992 430 406 7	58.128 349 745 2
14	Tree cover loss	5/1/2005	12/1/2005	1	1	13.585 405 180 3	15.334 309 020 1	56.380 030 452 1	57.308 020 244 6
15	Cyclone	15/1/2007	24/1/2007	1	1	−3.618 097 891 1	22.579 767 782	48.634 884 878 5	53.596 061 075 5
16	heatwave	15/7/2007	22/7/2007	1	1	15.584 152 425 1	19.349 134 773 1	43.316 789 530 4	45.172 264 131 6
17	Tree cover loss	26/2/2010	1/3/2010	1	1	7.553 471 530 8	8.731 304 728 4	51.097 627 018	51.490 238 085 5
18	Tree cover loss	26/1/2008	27/1/2008	1	1	14.656 162 633 5	15.788 100 227 3	46.814 597 208 2	47.858 102 833 8
19	heatwave	26/6/2006	30/7/2006	3	2	−4.296 244 278 5	19.938 566 066 0	45.101 385 284 1	54.773 894 273 1
20	Tree cover loss	29/2/2008	2/3/2008	1	1	18.778 669 414 1	19.296 111 594 2	49.848 409 990 8	50.132 677 071 3
21	Cold wave	8/3/2010	10/3/2010	1	1	−0.075 430 592 4	4.175 712 024 5	38.931 556 196 7	42.768 165 302 6
22	heatwave	15/7/2007	22/7/2007	1	1	19.198 203 554 2	20.964 481 948 8	39.788 162 158 8	42.721 902 892 4
23	Flood	3/7/2007	3/7/2007	1	2	32.170 524 336 4	34.664 199 470 5	10.197 647 138 3	15.737 286 911 1
24	Flood	10/2/2008	10/3/2008	2	3	14.074 525 946 5	21.390 335 743 6	−19.699 677 231 6	−17.681 894 298 6
25	Flood	6/9/2009	6/9/2009	1	1	−2.466 337 741 1	−0.196 331 941 8	11.741 289 226 9	13.361 309 363 6
26	Flood	3/4/2009	3/4/2009	1	2	17.305 793 486 2	25.473 204 386 1	−18.772 292 559 6	−16.079 009 829 8
27	Flood	12/2006	2/2007	3	2	32.737 643 287 5	36.234 788 286 4	−19.147 672 021 5	−15.028 825 021 2
28	Flood	12/2010	1/2011	3	2	30.197 237 324 0	32.812 494 920 8	−28.579 855 449 0	−24.555 857 804 5
29	Flood	6/11/2011	11/11/2011	1	1	2.715 668 165 0	6.153 168 165 3	50.253 818 214 2	54.384 500 280 0
30	Flood	8/2002	8/2002	1	1	9.522 227 409 8	14.708 691 114 3	48.352 992 474 9	54.651 633 675 7
31	Flood	6/2009	6/2009	2	3	8.997 984 700 3	16.847 581 014 0	46.069 125 249 5	54.481 404 748 6
32	Cyclone	15/12/2006	28/12/2006	1	2	42.643 563 322 1	51.685 515 144 3	−21.437 010 401 2	−11.000 100 646 2
33	Cyclone	1/3/2004	18/3/2004	3	3	42.096 621 059 6	52.205 636 431 7	−25.183 363 540 2	−12.411 417 914 4
34	Cyclone	5/3/2002	17/3/2002	1	1	44.190 565 835 5	52.065 876 046 3	−24.665 642 094 6	−16.123 377 083 4
35	Cyclone	9/3/2007	18/3/2007	2	3	47.116 950 013 1	51.661 720 534 4	−18.048 129 419 8	−13.556 170 275 3
36	Cyclone	7/2/2008	22/2/2008	2	2	42.123 009 180 2	53.276 826 739 5	−22.976 224 780 2	−13.934 272 958 1
37	Cyclone	25/2/2003	6/3/2003	3	3	42.595 974 102 5	46.331 727 881 4	−24.604 154 368 9	−19.516 110 217 3
38	Volcanic eruption	6/2011	6/2011	1	1	38.879 475 290 6	42.737 640 821 3	10.786 548 833 0	14.677 313 544 1
39	Volcanic eruption	20/3/2010	23/6/2010	1	1	−21.958 802 986 0	−16.563 954 230 8	62.872 846 952 3	67.107 428 109 6
40	Cold wave	20/1/2006	15/2/2006	1	1	−2.157 338 020 1	24.861 851 226 2	37.612 008 331 2	63.782 877 768 2

The validation comments on the first version of BACIndex and the list of events (Table 1) provided by WP6 were helpful for WP5 to investigate new possibilities for improving the BACIndex. We have carefully examined these events and the source of errors. For some events, such as the Events ID 17 and ID 18, the capacity of the used data, which is of temporal resolution 8-daily, and spatial resolution 0.25 degree, does not allow for detection of such very short and spatially small events. Other events, such as the cold wave in Europe in 2006 (event ID 40) and the cyclone (event ID 15), were however missed because of some drawback in the method itself. The method as mentioned above, uses five univariate ARMA models to model the five variables, then uses the Mahalanobis distance to measure the deviation of the residuals from the multivariate joint distribution. An advantage of using univariate ARMA is the simplicity in its implementation specially when dealing with short length data. However, a draw back of the univariate model approach arises when the used variables are highly correlated. In such case, it is very likely that the residuals of the different variables are highly correlated and thus the Mahalanobis distance might not show high values for extreme events. In order to overcome this drawback, the variables should be presented with a model that takes into account their inter-dependency. Hence, we have implemented the same method proposed before but after replacing the five univariate models with one multivariate autoregressive model (see Figure 1) where each variable is expressed as a linear regression of the previous time samples up to order p of all the variables used. In the following we further explain the used MVAR model in more detail and then compare the results of using univariate ARMA models and MVAR model.

1.1 Anomaly Detection with Multivariate Autoregressive (MVAR) Model

Let $x_i, i = 1, \cdots, N$ denotes the time series of N Earth observation variables. Each time series $x_i(n), n = 1, \cdots, m$ is a realization of length m real valued discrete stationary stochastic process $X_i, i = 1, \cdots, N$. These N time series can be represented by a *p*th order multivariate autoregressive model (MVAR(p)) of the form

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_N(n) \end{bmatrix} = \sum_{r=1}^{p} A_r \begin{bmatrix} x_1(n-r) \\ \vdots \\ x_N(n-r) \end{bmatrix} + \begin{bmatrix} \epsilon_1(n) \\ \vdots \\ \epsilon_N(n) \end{bmatrix}, \tag{1}$$

The residuals $\epsilon_i, i = 1, \cdots, N$ constitutes a white noise stationary process with an $N \times N$ residual covariance matrix Σ. The model parameters at time lags $r = 1, \cdots, p$ is defined by

$$A_r = \begin{bmatrix} a_{11}(r) & \cdots & a_{1N}(r) \\ \vdots & \ddots & \vdots \\ a_{N1}(r) & \cdots & a_{NN}(r) \end{bmatrix}. \tag{2}$$

We have applied the MVAR model based anomaly detection method to the ESDC data with the same five biosphere variables and compared the anomaly detection results using five univariate ARMA models with those using one multivariare AR model. Figures 2-6 show the results obtained for some known historic events selected from Table 1. For some events, such as the Russian heatwave in 2010 and the drought in the

horn of Africa in 2006, the two models give similar results in terms of event detection; results are shown in Figures 2 and 3 respectively. Another particular event of interest is the volcanic eruption in the coast of the Red Sea in June 2011 (event ID 38). This event is detected by both methods despite its small spatial scale as shown in Figure 5. However, an improved temporal and spatial detection accuracy can be noticed using MVAR model for these three events. There are on the other hand some winter events such as the cold waves (example is the event ID 40 shown in Figure 6) and cyclones (an example is the event ID 15 shown in Figure 4), which are detected only by the MVAR model.

1.2 Near Real-Time BACI

The proposed method can be applied using sliding time window approach with very low computational load making it suitable for real time implementation. The computational load of the improved method when applied to the time series of all points of the grid of BACI study area which constitute 119280 time series, with time window length of 506 samples from five variables, using Matlab code executed with 8xIntel Core i7-7700CPU@3.60GHz processor, is equal to 835 seconds or about 14 minutes. For single point time series, the execution time is about 835/119280 = 0.007 seconds. Note that the computation for each point of the grid can be executed separately if needed. The computational time does not include the model order selection step which can be done offline once a prior and fixed.

The second version of BACIndex using an MVAR model is uploaded to the BACI portal and made available to all BACI community.

1.3 Conclusions

The anomaly detection method using linear regression and Mahalanobis distance used to generate the BACIndex has been further improved by using multivariate autoregressive model to account for the high correlation between the used variables at different time lags. This led to an improved temporal and spatial detection accuracy of the proposed anomaly detection method and even helped to detect several events that were missed when using univariate ARMA models. In particular MVAR model performed better in the detection of winter events. The improved version of the method has very low computational load making it suitable for real time implementation. The second version of the change index BACI is uploaded to BACI portal and made available for BACI community.

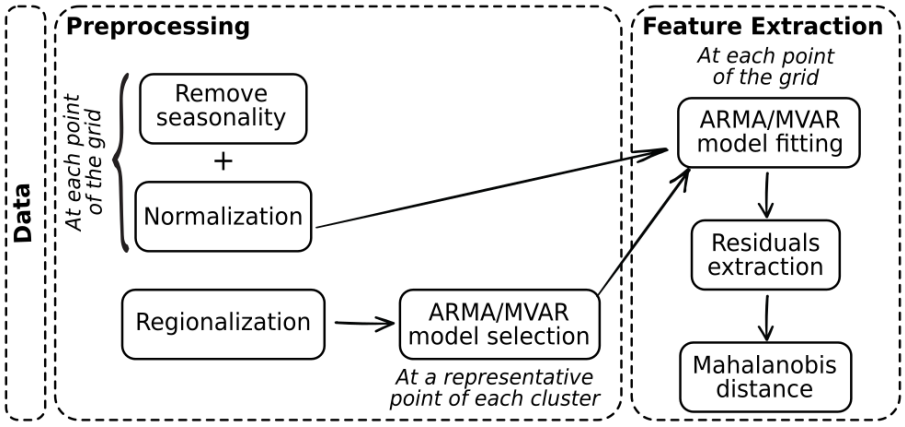

Figure 1: Flowchart of the proposed methodology.

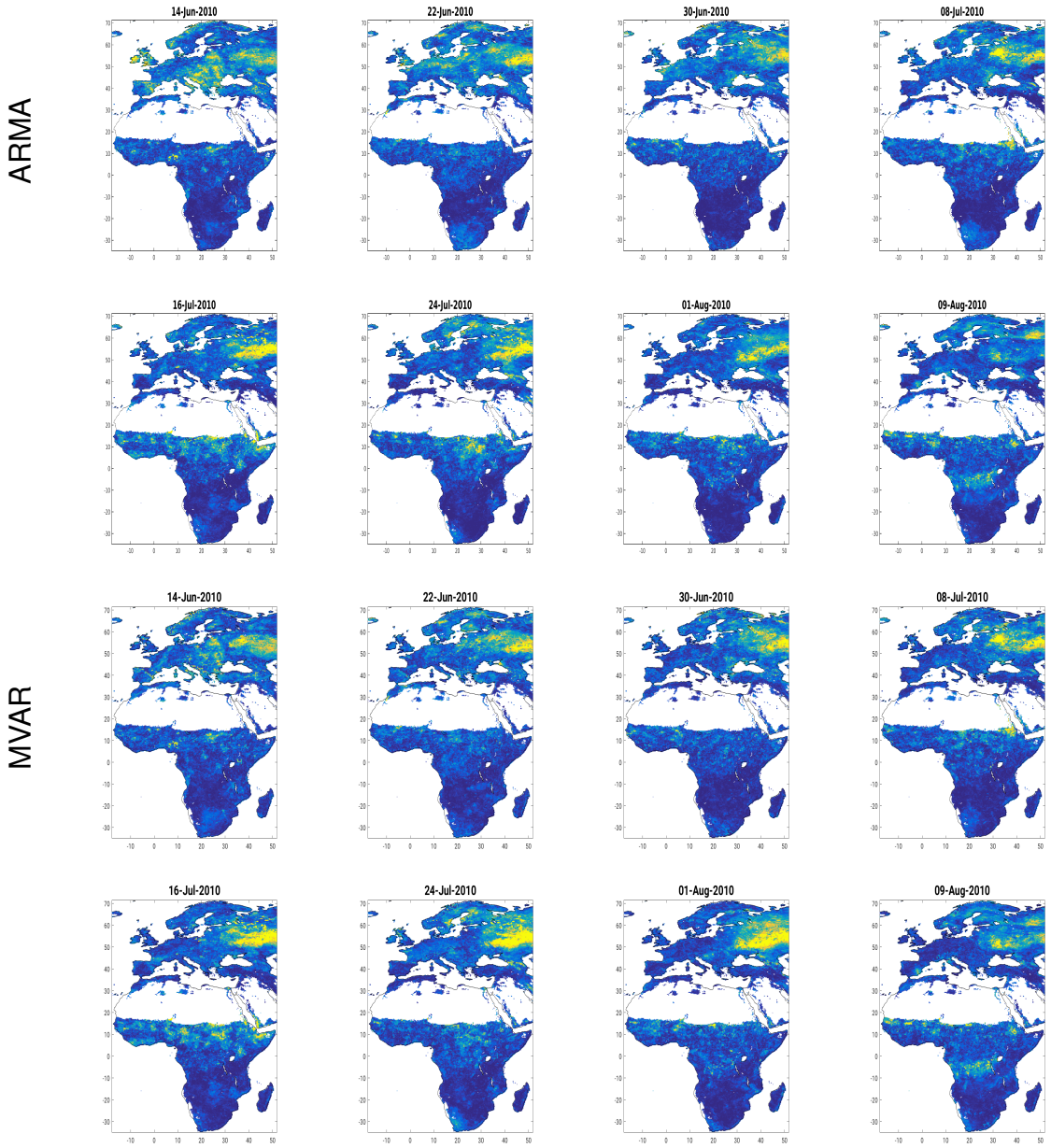

Figure 2: Comparison between the results obtained by ARMA and MVAR models. Event ID 13: Heatwave in Russia, summer 2010.

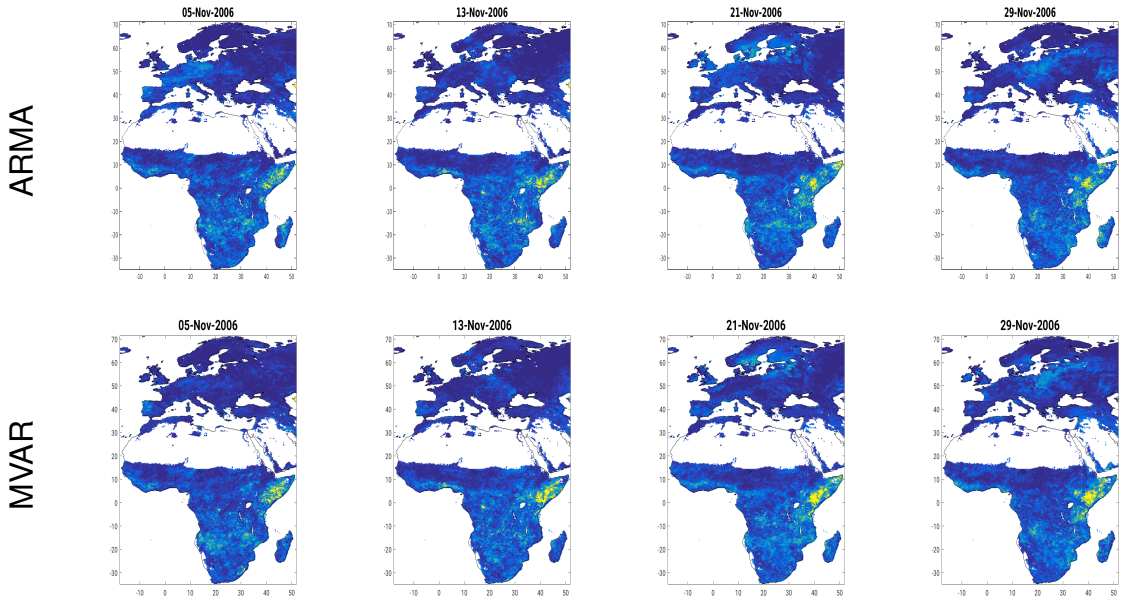

Figure 3: Comparison between the results obtained by ARMA and MVAR models. Event ID 5: Drought in the horn of Africa in November 2006.

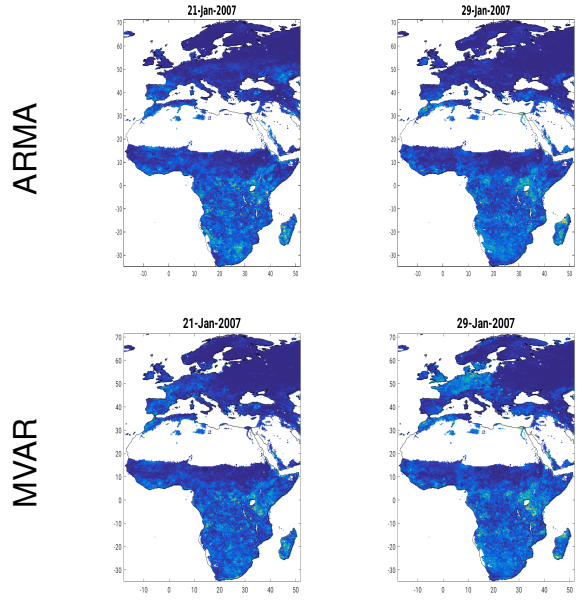

Figure 4: Comparison between the results obtained by ARMA and MVAR models. Event ID 15: Cyclone in Central Europe in January 2007.

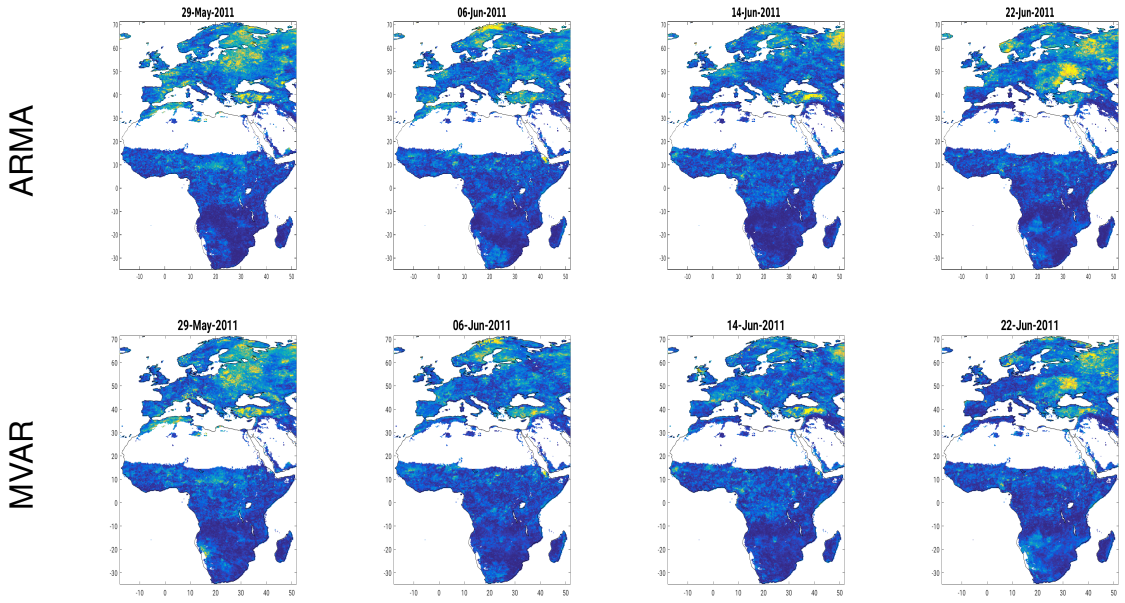

Figure 5: Comparison between the results obtained by ARMA and MVAR models. Event ID 38: Volcanic eruption in the Red Sea coast in June 2011.

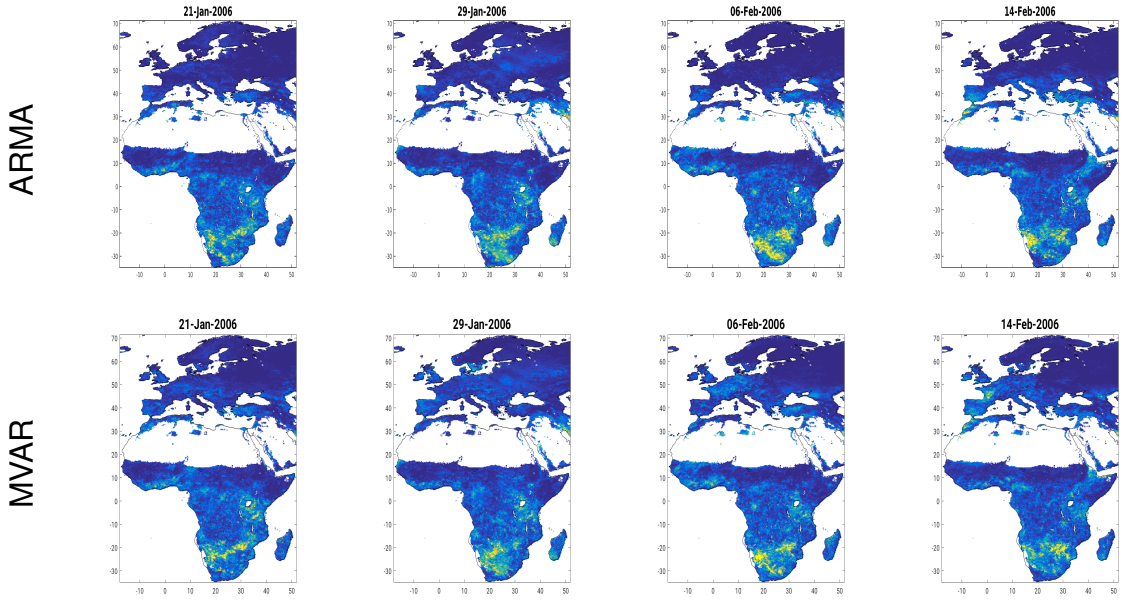

Figure 6: Comparison between the results obtained by ARMA and MVAR models. Event ID 40: Cold wave in Central Europe in January-February 2006.

2 Attribution Scheme based on Mahalanobis Distance Decomposition

The residual vector of the MVAR model is calculated as the difference between the model output and the real data for the five variables. The Mahalanobis distance [3, 4] of the residual vector is used as a measure of the deviation of the multivariate residuals at certain time step from their joint distribution. The Mahalanobis distance is defined in square unit as

$$d_m(\mathbf{E}) = (\mathbf{E} - \bar{\mathbf{E}})^{\mathbf{T}} \boldsymbol{\Sigma}^{-\mathbf{1}} (\mathbf{E} - \bar{\mathbf{E}}) \tag{3}$$

where $\bar{\mathbf{E}}$ and $\boldsymbol{\Sigma}$ are the mean and covariance matrix of the multivariate residuals vector $\mathbf{E}$ respectively. The mean and the covariance were estimated considering the entire time series. This was the best way to do so in our case due to the short length of the time series used together with its coarse temporal resolution.

In this section we present an attribution scheme based on the decomposition of the Mahalanobis distance. The decomposed form of the Mahalanobis distance provide the answer for the question: how much each variable has contributed to the Mahalanobis distance? The mathematical procedure for the decomposition, known as Garthwaite Transform [2], is explained in detail in the next section followed by experimental results and discussion.

2.1 Mahalanobis Distance Decomposition

When the value of the Mahalanobis distance estimated with the residuals is large, it is assumed that something abnormal occurs in the system and the model is not able to correctly capture it. Then the following obvious question is: *which variable(s) is causing this anomaly?* An intuitive approach to answer this question is to form a partition of the value of the Mahalanobis distance, where each element of the partition quantifies the contribution of each of the variables involved. Garthwaite and Koch [2] recently proposed a method for the decomposition of the Mahalanobis distance which can be easily implemented and provides helpful results from an attribution point of view. The decomposition has the form:

$$d_m(\mathbf{E}) = \mathbf{W}^T \mathbf{W}, \tag{4}$$

where $\mathbf{W} = (W_1, \cdots, W_N)^T$ is a mathematical vector with N terms, corresponding to the N variables contributing to the Mahalanobis distance $d_m(\mathbf{E})$, and is calculated by

$$\mathbf{W} = (\mathbf{S}\Sigma\mathbf{S})^{1/2}\mathbf{S}(\mathbf{E} - \bar{\mathbf{E}}). \tag{5}$$

The components of **W** should be uncorrelated, with the transformation chosen to maximize the sum of correlations between the corresponding elements of **S** and **W**. **S** is a diagonal matrix of the inverses of the standard deviations of the variables of $\mathbf{E}$, Σ_1 is the covariance matrix, $\bar{\mathbf{E}}$ is a mathematical vector of the mean values of the used variables and $\mathbf{E}$ is a vector of values of the sample point giving the observed value of the Mahalanobis distance to the centroid defined by $\bar{\mathbf{E}}$.

2.2 Attribution Scheme based on Mahalanobis Distance Ratio

We recall that the detected anomalies in the second version of BACIndex are those where the Mahalanobis distance of the residuals of the full model is high indicating that the residuals are far from their joint distribution where the MVAR model (7) could not accurately predict the real values of the used variables. A straightforward approach for attribution is to look for the variable or set of variables that caused the Mahalanobis distance to be high. This can be simply done by comparing the Mahalanobis distance of the reduced model (the model after eliminating one or more variable at a time) with the Mahalanobis distance of the full model. Hence, We define the ratio

$$\beta_i = \ln \frac{D_{i-}}{D}, \tag{6}$$

where D_{i-} is the Mahalanobis distance of the reduced model after eliminating the variable x_i, and D is the Mahalanobis distance of the full model. The value β_i define the reduction in the Mahalanobis distance when eliminating the variable x_i. The detected change is then attributed to the variable x_i with the lowest β_i.

2.3 Experimental Results and Discussion

The Mahalanobis distance decomposition based on the Garthwaite-Koch partition was applied to the results obtained with the univariate ARMA models as well as those obtained with the multivariate AR model. The results provided by the MVAR model outperform the ones with univariate ARMA models, this is because the MVAR model implementation produced more whitened residuals with considerably reduced cross-correlation coefficients between the residuals of the different variables. Hence, we here only present the Mahalanobis distance decomposition applied to the method using MVAR model.

This attribution scheme has been compared with the the z-score results provided by WP6. On the Deliverable 6.2, WP6 produced a series of plots estimating the z-score for each variable and for all the 40 known extreme events summarized in Table 1. For each event, and for each of the five used variables, they have estimated the histogram of the variable within the time window of the event and compared it with the histogram of the entire time series. The z-score quantifies the discrepancy between these two histograms. Higher values of the z-score indicate which variables are most different from their normal behaviour within the time duration of the event.

Figures 7 to 11 show the results obtained for both attribution schemes (Mahalanobis distance decomposition and z-scores) for the selected 5 known extreme events presented in the previous section. For each event, the figure shows the spatial extension of the event (upper left subplot), the z-scores (upper right subplot) and the Mahalanobis distance decomposition (lower subplot). The z-scores subplots show the histograms of the 5 variables within the time window of the event (red) and the entire time series (grey) together with the value of the z-score obtained from the comparison of both histograms. The Mahalanobis decomposition subplots show the Mahalanobis intensity (map on the left) and 5 more maps one for the contribution of each of the used vari-

Table 2: Comparison between attribution results using Mahalanobis decomposition, z-scores, and Mahalanobis distance ratio.

Event ID	Event type	Attribution Variable(s) listed in decreasing intensity order		
		Mahal. decompos.	z-score	Mahal. dist. ratio
5	Drought	GPP, NEE	SH, NEE	GPP, NEE, SH
13	Heatwave	LE, SH and TER	SH, TER and LE	LE, SH
15	Cyclone	GPP, TER, SH	SH, TER, GPP	NEE, GPP
38	Volcanic eruption	GPP and LE	SH and GPP	GPP
40	Cold wave	GPP, NEE, SH	TER	NEE, GPP

ables.

In event ID 5 (Figure 7) the main driving variables to the drought in the horn of Africa are GPP and NEE. For the case of the Russian heatwave (event ID 13, Figure 8), LE and, with less intensity, SH and TER are the most contributing ones. Some events do not present a clear attribution scheme; that is the case of events ID 15 (Figure 9) or ID 40 (Figure 11) where three variables contributed almost equally to the event.

The main causing variables based on Mahalanobis decomposition for the five selected events are listed in Table 2 in the order of their contribution to the event. We also list the driving variables of the extreme events based on the z-score and the Mahalanobis distance ratio. It should be noted however that the z-score does not serve as ground truth as it is based on univariate analysis and is listed for comparison purpose only.

2.4 Conclusions

In this section we have presented an attribution scheme based on the decomposition of the Mahalanobis distance using Garthwaite Transform. The decomposed form of the Mahalanobis distance provides the answer to the question: how much each variable has contributed to the Mahalanobis distance, i.e. the change index? Experimental results of the developed attribution scheme used for the attribution of different known historic events were also presented and compared to the attribution results of the univariate z-score and the Mahalanobis distance ratio which is the reduction in the Mahalanobis distance obtained by eliminating certain variable. Unfortunately, detailed quantitative evaluation of the performance of the proposed methods is not possible due to the lack of the ground truth for the attribution of the selected extreme events. The results hence still need to be validated by climate scientists.

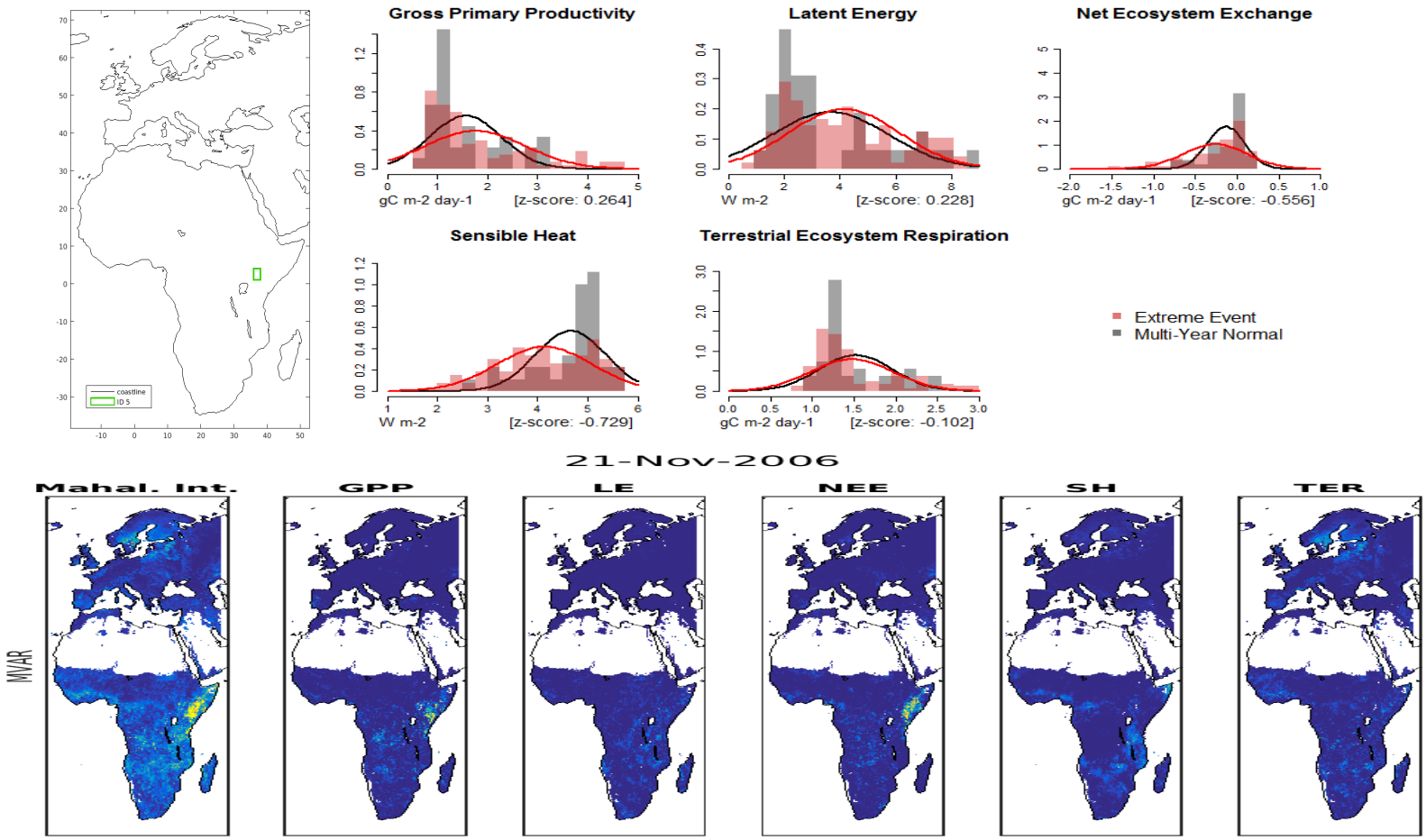

Figure 7: Attribution scheme Event ID 5 (Drought). Upper left plot: spatial extension, Upper right plots: z-score for the 5 variables involved, Lower plots: Mahalanobis intensity (left) and its decomposition into the 5 variables involved.

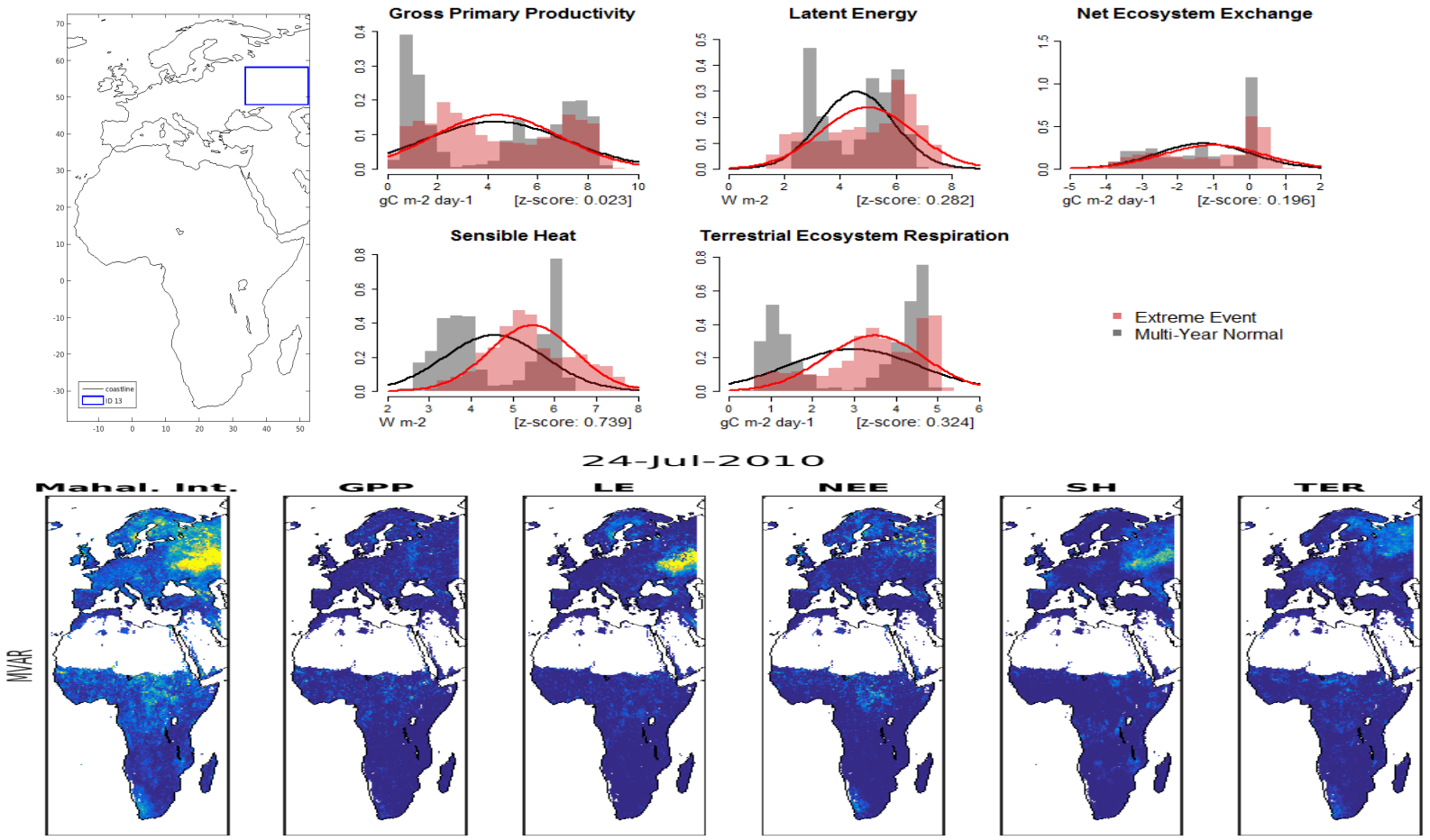

Figure 8: Attribution scheme Event ID 13 (Heatwave). Upper left plot: spatial extension, Upper right plots: z-score for the 5 variables involved, Lower plots: Mahalanobis intensity (left) and its decomposition into the 5 variables involved.

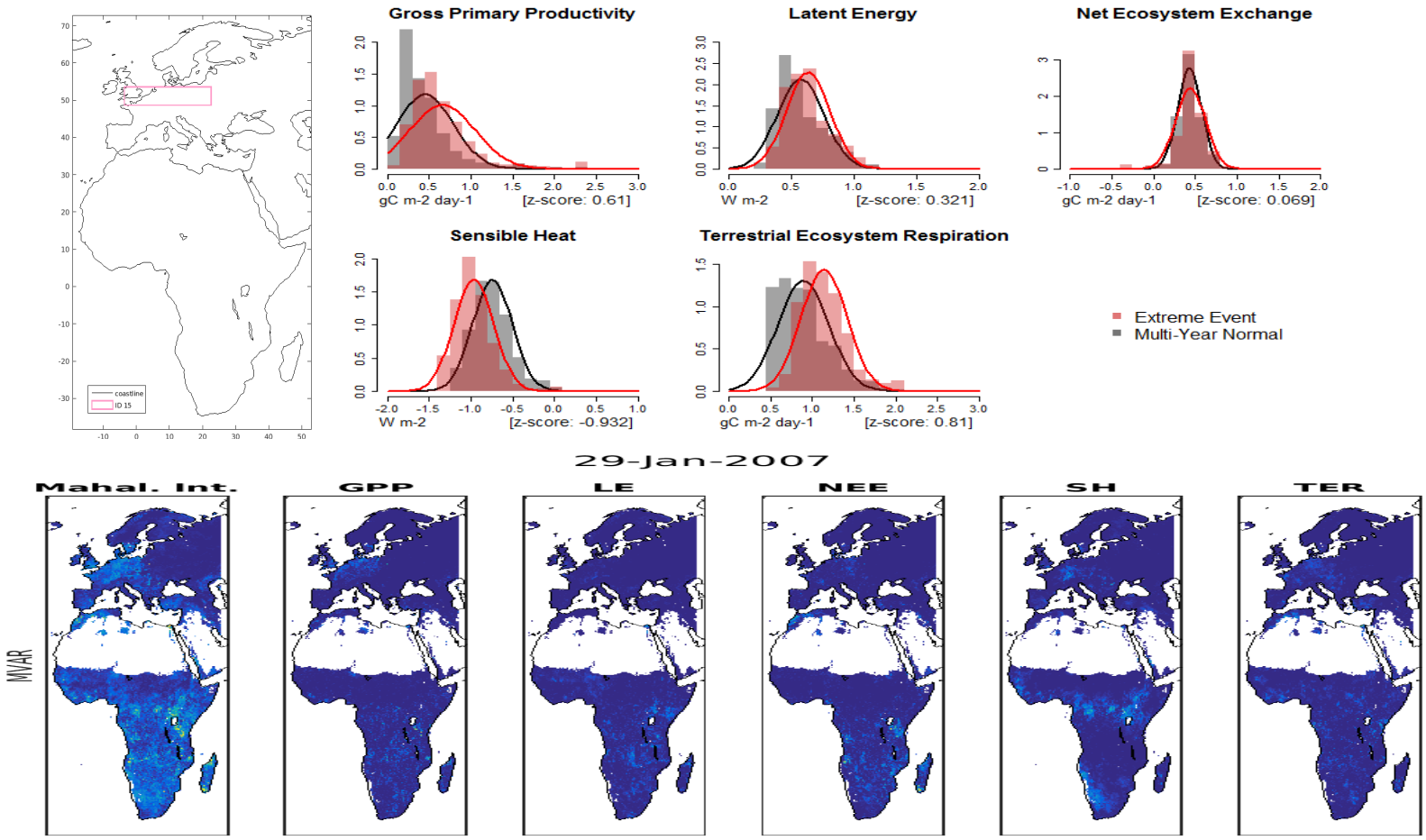

Figure 9: Attribution scheme Event ID 15 (Cyclone). Upper left plot: spatial extension, Upper right plots: z-score for the 5 variables involved, Lower plots: Mahalanobis intensity (left) and its decomposition into the 5 variables involved.

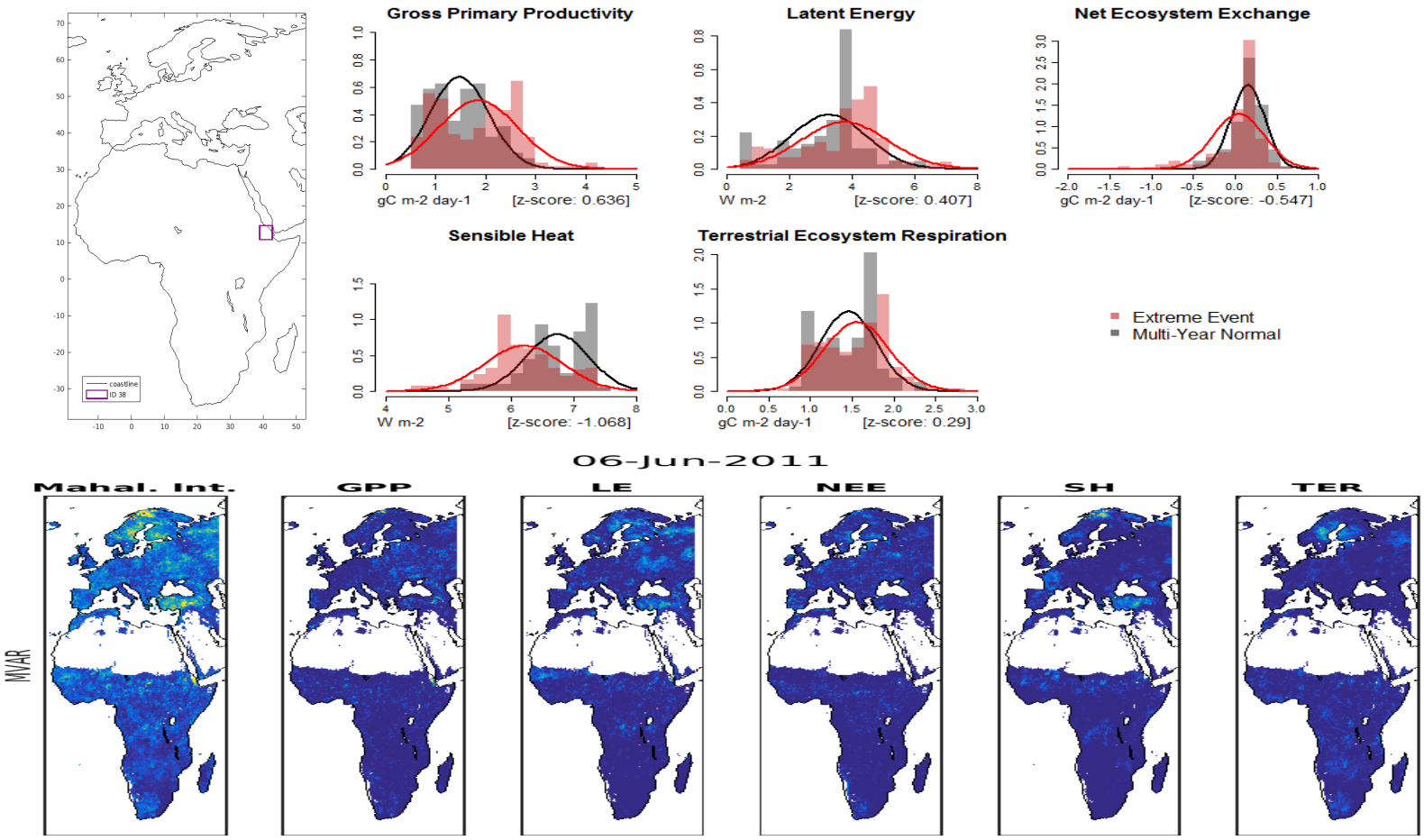

Figure 10: Attribution scheme Event ID 38 (Volcanic eruption). Upper left plot: spatial extension, Upper right plots: z-score for the 5 variables involved, Lower plots: Mahalanobis intensity (left) and its decomposition into the 5 variables involved.

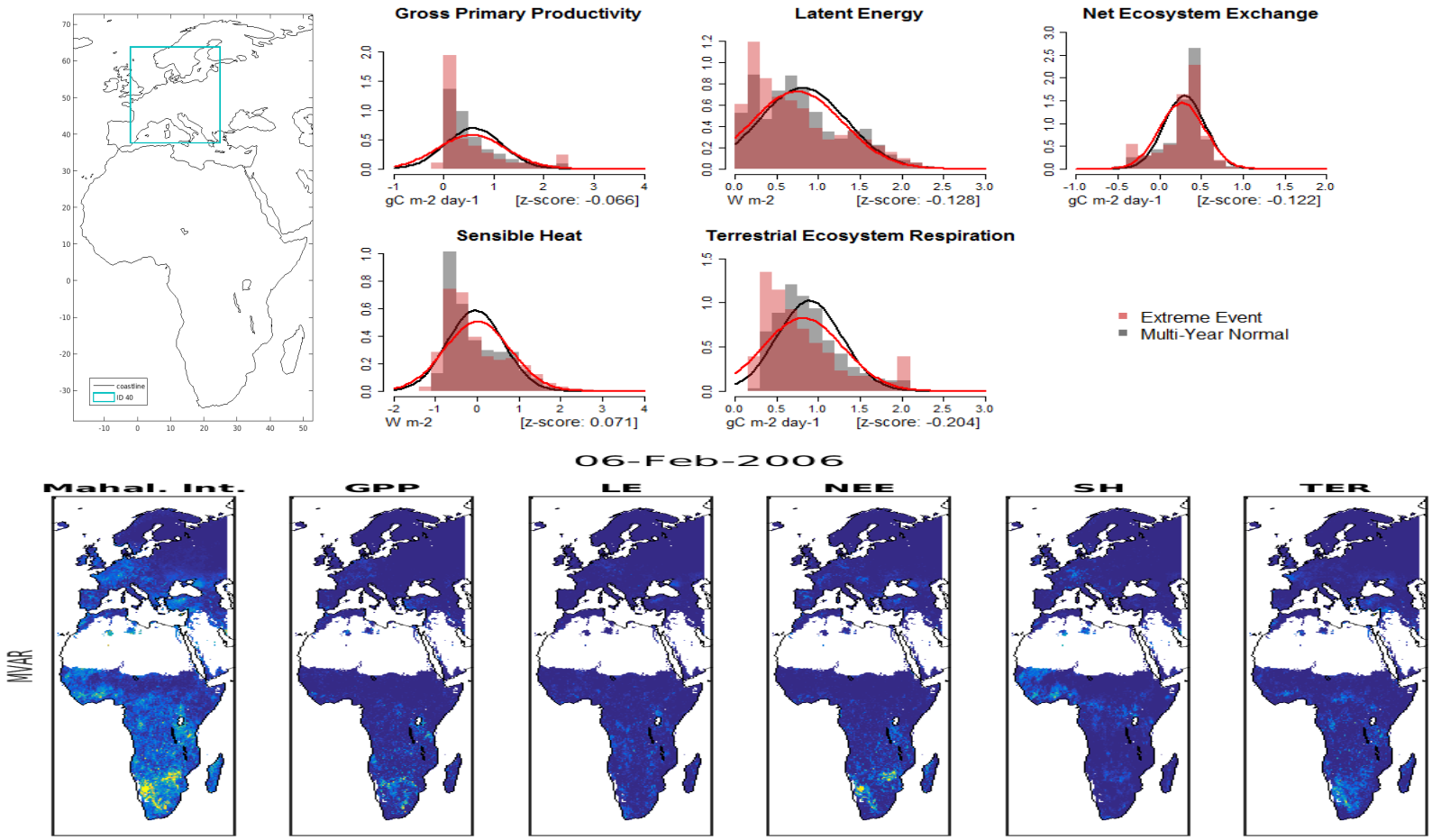

Figure 11: Attribution scheme Event ID 40. Upper left plot: spatial extension, Upper right plots: z-score for the 5 variables involved, Lower plots: Mahalanobis intensity (left) and its decomposition into the 5 variables involved.

3 Attribution Scheme based on Spectral MVAR Granger Causality

Local meteorological conditions have direct impact on CO2 fluxes and ecosystem respiration. Understanding the cause-effect relationships in such dynamical system is essential for the attribution of climate changes as well as for the development of intervention strategy to achieve desired prediction. The availability of high temporal resolution data along with the powerful computing platforms further enhance the capacity of data-driven methods in capturing the complex relationships between the variables of the underlying dynamical system.

Time series of ecological variables most often contain multiple periodical components, e.g. daily and seasonal cycles, induced by the meteorological forcing variables. This can significantly mask the underling endogenous causality structure of the biogeochemical cycle when using time domain analysis. Filtering these periodic components as preprocessing step degrades causal inference [5]. This motivates the use of time-frequency processing techniques such as short time Fourier transform where the causality structure can be examined at different frequency bands or different time scales. In this section, we present a time-frequency approach for causality analysis applied to the meteorological observations and land flux eddy covariance data to investigate the causal-effect relationships between global radiation (Rg), air temperature (T), and the CO2 land fluxes: gross primary productivity (GPP) and ecosystem respiration (Reco). The coupling between the used variables is assumed to follow a multivariate autoregressive (MVAR) model. The cause-effect relationships are extracted using the MVAR Granger causality (MVAR-GC) [6, 7] based on the generalized partial directed coherence (gPDC) [8, 9]. We compare experimental results obtained using gPDC with those using time domain conditional MVAR-GC to highlight the advantages of using frequency analysis techniques. To account for the nonstationarity of the used variables, we also present the gPDC causality analysis using short time window approach and compare the time variant causal-effect intensities obtained over different seasons. The developed causality analysis can then be directly implemented for attribution of detected changes as will be explained in the sequel.

3.1 Frequency Domain MVAR Granger Causality: Generalized Partial Directed Coherence

Various causality measures have been reported in literature. Among many other linear regression based models, Granger causality (GC) (Weiner 1956, Granger 1969)[6] is the most widely known method for causality analysis. GC assumes that causes both precede and help predict their effects.

Let $x_i, i = 1, \cdots, N$ denotes the time series of N Earth Observation variables. Each time series $x_i(n), n = 1, \cdots, m$ is a realization of length m real valued discrete stationary stochastic process $X_i, i = 1, \cdots, N$. These N time series can be represented by a

*p*th order multivariate autoregressive model (MVAR(p)) of the form

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_N(n) \end{bmatrix} = \sum_{r=1}^{p} A_r \begin{bmatrix} x_1(n-r) \\ \vdots \\ x_N(n-r) \end{bmatrix} + \begin{bmatrix} \epsilon_1(n) \\ \vdots \\ \epsilon_N(n) \end{bmatrix}. \tag{7}$$

The residuals $\epsilon_i, i = 1, \cdots, N$ constitute a white noise stationary process with covariance matrix Σ. The model parameters at time lags $r = 1, \cdots, p$ is defined by

$$A_r = \begin{bmatrix} a_{11}(r) & \cdots & a_{1N}(r) \\ \vdots & \ddots & \vdots \\ a_{N1}(r) & \cdots & a_{NN}(r) \end{bmatrix}. \tag{8}$$

The model order can be estimated using Akaike or Bayesian Criterion. The model parameters $a_{ij}(r), i, j = 1, \cdots, N; r = 1, \cdots, p$, can then be estimated using the method of Least Square. It is worth noting that the use of the MVAR model (7) makes no assumption on the mechanism that produced the data (for example whether it is a linear or non-linear) except that the model itself exist and stable [10].

The causal relation from x_i to x_j is described in the frequency domain via gPDC [8] by

$$g\pi_{i\to j}(f) = \frac{\frac{1}{\sigma_{jj}}\overline{A}_{ji}(f)}{\sqrt{\sum_{k=1}^{m} \frac{1}{\sigma_{kk}^2} \left|\overline{A}_{ki}(f)\right|^2}}, \tag{9}$$

where $\overline{A}_{ij}(f), i, j = 1 \cdots N$ are the elements of the matrix $\overline{A}(f) = I - A(f)$ where $A(f)$ is the Fourier transform of $A(r), r = 1, \ldots, p$:

$$A(f) = \sum_{r=1}^{p} \mathbf{A}_r z^{-r}|_{z=e^{j2\pi f}}, \tag{10}$$

and σ_{ii}^2 are the diagonal entries of the residual covariance matrix Σ. The value of $g\pi_{i\to j}(f)$ represents the causality strength of x_i on x_j at the normalized frequency f as compared to all of x_i's interactions to other variables. Nullity of $g\pi_{i\to j}(f)$ indicates absence of the Granger causality of x_i on x_j at the normalized frequency f.

3.2 Time Domain MVAR Granger Causality

The conditional MVAR-GC of x_i on x_j quantifies the degree to which the past of x_i helps predict x_j, over and above the degree to which x_j is already predicted by its own past and the past of the variables other than x_i. Let Σ_j denote the covariance matrix of the residual ϵ_j associated to x_j using the model in (7), and let Σ_j^{i-} denotes the covariance matrix of the residual associated to x_j using the model (7) after eliminating x_i, i.e. eliminating the ith raw and column in (8). The time domain MVAR-GC of x_i on x_j conditioned on all other variables is defined by the likelihood ratio [11, 7]

$$\gamma_{i\to j} = \ln \frac{|\Sigma_j^{i-}|}{|\Sigma_j|}. \tag{11}$$

3.3 Experimental Results and Discussion

Experiments are performed on the real half-hourly meteorological observations and land flux eddy covariance data measured at Hainich National Park spanning the seven years 2000-2006 using both time domain conditional MVAR-GC and frequency domain gPDC. First the data of the seven years over all seasons were segmented into 90 days short segments with 50% overlap. The model order is estimated using Bayesian criterion and fixed for all segments. The model parameters are estimated for each segment using Least Square. Time domain causal intensities as defined in (11) are estimated using the MVAR-GC toolbox [11] with statistical significance F-test. In case of the gPDC based frequency domain analysis defined in (9), we used the permutation test for statistical significance with confidence level 95%. The averages time and frequency domain causality strength between the four used variables of the real data are shown in Figures 12 and 13 respectively.

The time domain causality structure in Figure 12 shows several spurious links, e.g. causal links of GPP as well as Reco on Rg and T. The frequency domain causality structure of GPP $\rightarrow$ Rg and GPP $\rightarrow$ T in Figure 13 shows spectral peaks at frequency corresponding to the daily cycle (f= 0.0201 cycle/30min) which indicates that the spurious links in time domain causality are mainly due to the daily cycle induced by global radiation. Similarly for Reco, peaks occur on the time scale of the seasonal cycle, i.e. around $f = 0$ cycle/30min, which is the cause of the spurious links in the time domain causality analysis. Another advantage of frequency analysis is that it shows the time scale at which interaction between variables occurs. In Figure 13, the peak in the frequency plot of Rg $\rightarrow$ T indicates that although the causal effect of Rg on T occurs on time scales of half an hour up till days, there is a clear peak around the time scale of 16 hours (f= 0.03015 cycle/30min) at the location of Hainich National Park. We can also notice a peak in the causal intensity of GPP on Reco on the time scale of 20 hours (f= 0.02513 cycle/30 min). The causal link of T on Reco exists over all the spectrum but with increased intensity on the time scale of two hours and more.

Similar experiments were repeated but on time segments of winter and summer seasons separately. Summer and winter gPDC spectral causality plots are shown in Figures 14 and 15 respectively. These figures show the time variant causal intensities between the four variables in different seasons. The causal intensity of T on Reco is higher in summer while the causal intensity of Rg on T and Reco is higher in winter.

3.4 Attribution Scheme based on Spectral Causality Analysis

It has been shown in the previous section that the spectral causal intensities between the variables vary with seasons. Examination of the changes in the causal intensities calculated using a short time window when compared to the average causal intensities over several years for the same season can be utilized in principle to simultaneously detect and attribute anomalous events. The anomalous events here are meant to be those time windows where the causal-effect relations or causal intensities show large deviation from the average or normal cause-effect behaviour. The anomalous event then can be attributed to the variable(s) that caused such deviation.

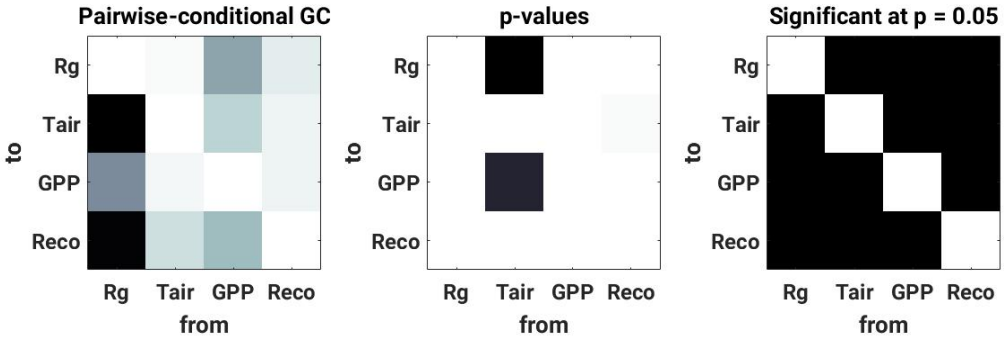

Figure 12: Plots of the time domain MVAR Granger causality within the variables of the real data of Hainich National Park. The causal strength is visualized using gray levels with black for highest value.

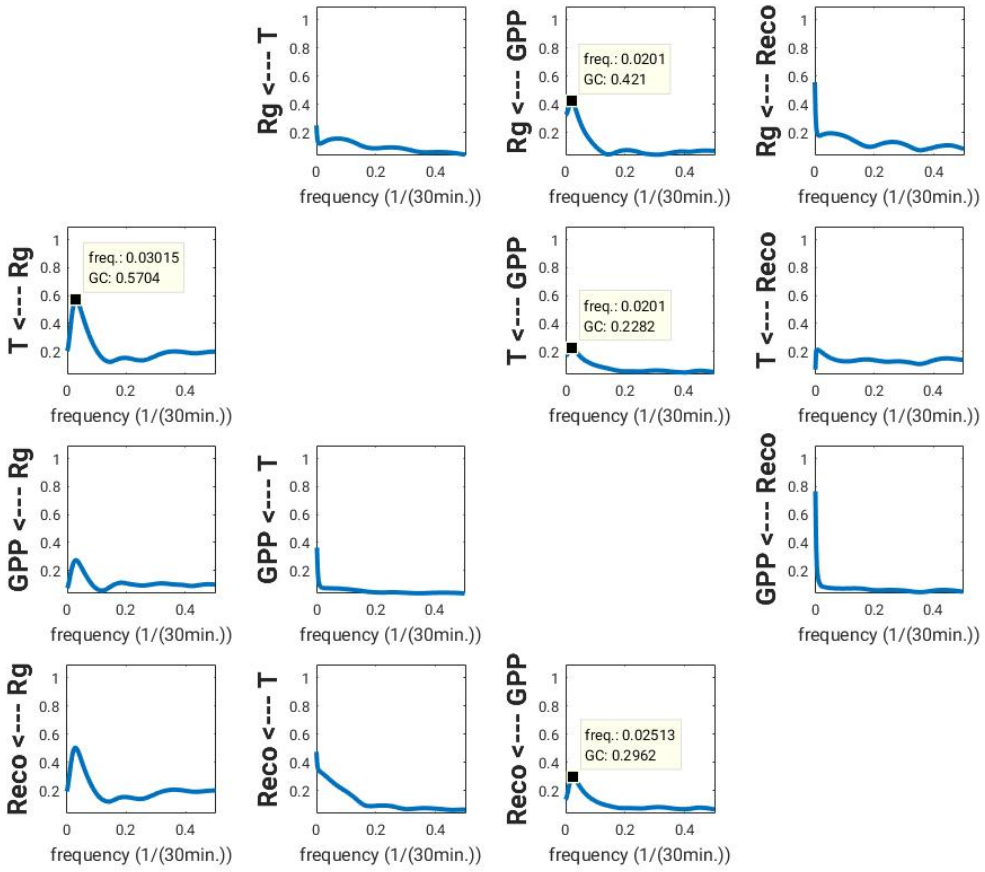

Figure 13: Plots of the gPDC representing the spectral causal intensities between meteorological and land flux CO2 data of Hainich National Park (average of 40 time segments over all seasons of years 2000-2006).

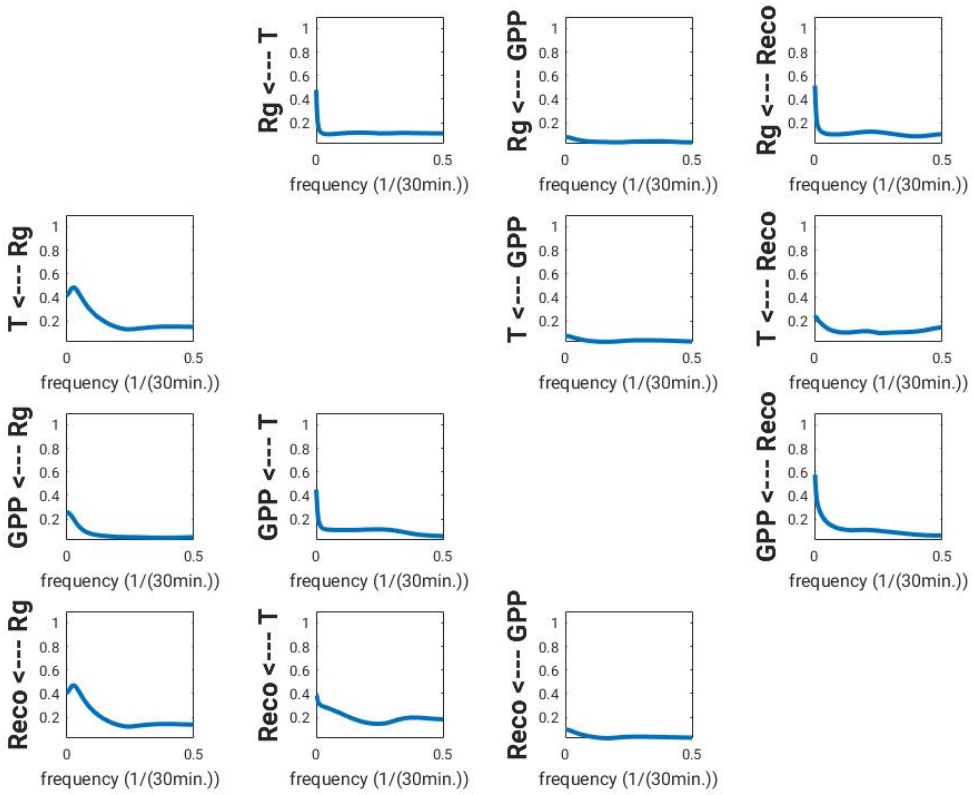

Figure 14: Plots of the gPDC representing the spectral causal intensities between meteorological and land flux CO2 data of Hainich National Park during winter (average of 20 winter time segments from years 2000-2006)

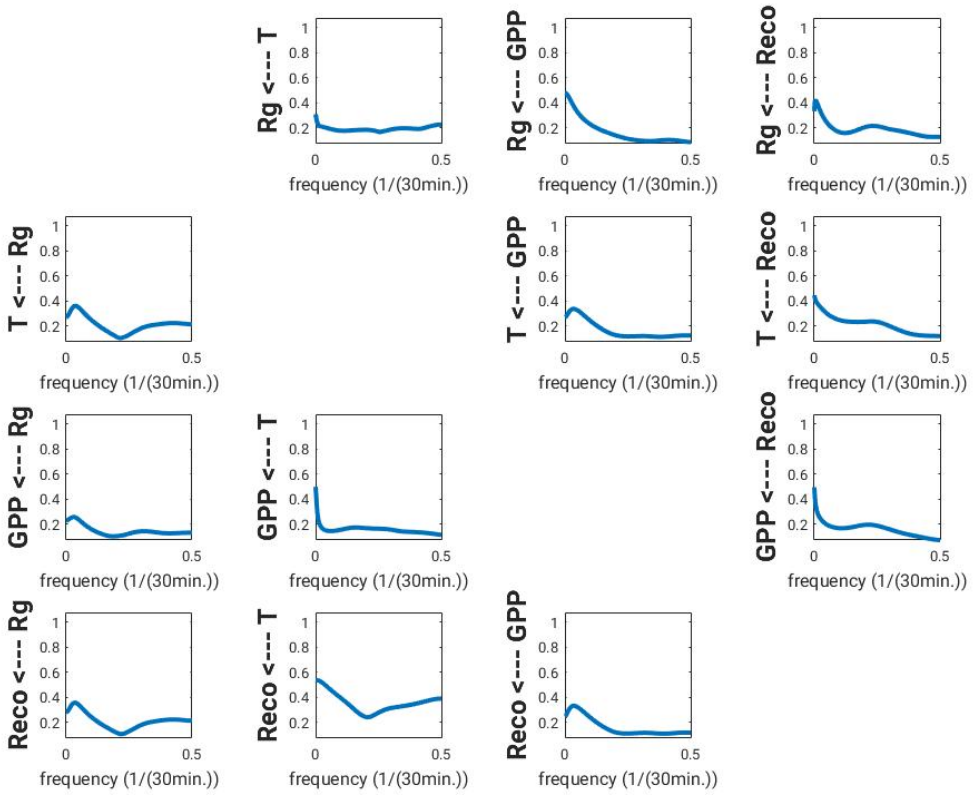

Figure 15: Plots of the gPDC representing the causal strength within the variables of the real data of Hainich National Park during summer (average of 20 summer time segments from years 2000-2006).

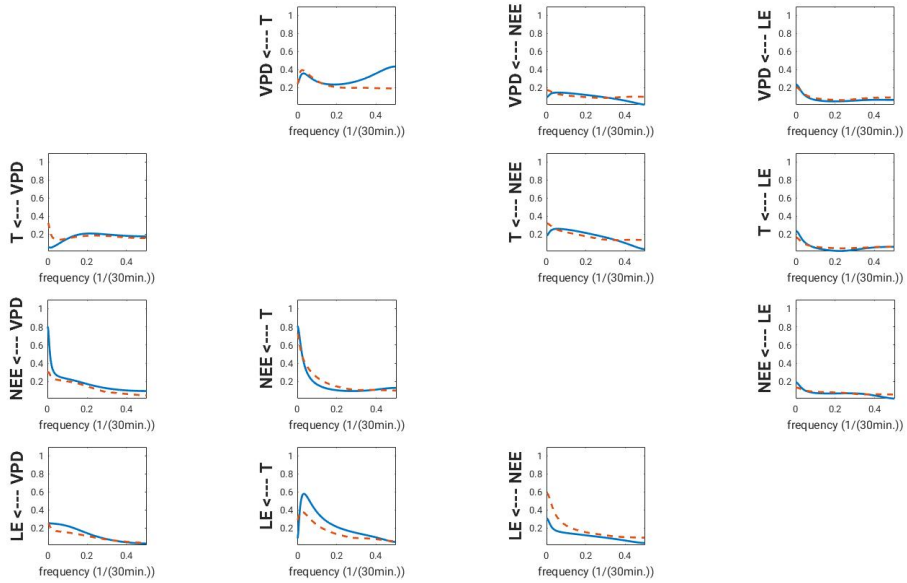

Figure 16: Plots of the gPDC representing the intensity of the cause-effect relationships between the four variables T, VPD, NEE and LE measured at the flux tower site of Puechabon-France during the heatwave in August 2003 (solid line) when compared to the average causal intensities of seven similar summer period within years 2000-2006 (red dash line).

In Figure 16 we show the cause-effect relationship between the following four variables: air temperature (T), vapor pressure deficiency (VPD), latent energy (LE) and net ecosystem exchange (NEE) using the eddy covariance data measured at the flux tower site of Puechabon-France between the beginning of July and end of August 2003. We compare the causal intensities with the average causal intensities of seven similar summer periods within years 2000-2006. It can be observed that there is considerable change in the causal intensity of T $\rightarrow$ VPD in the time scale of half an hour up to two hours and also there is clear increase in the causal intensity of T $\rightarrow$ NEE and NEE $\rightarrow$ LE at the time scale of two hours and more; other cause-effect intensities however remain the same; indicating that such event can be attributed to both T and NEE. Comparing similar results for year 2002 (Figure 17), it can be observed that while in a normal summer such as in 2002, the causal intensities match well with the average behaviour, the one in 2003 shows clear deviation in the system dynamics from average behaviour with T being the driving variable followed by NEE. It should be noted however that including different variables might results in different causal structure.

3.5 Conclusions

An attribution scheme based on the analysis of the cause-effect relationships in multivariate Earth observation system has been presented in this section. First we proposed to use the parametric spectral MVAR-GC referred to as generalized partial directed co-

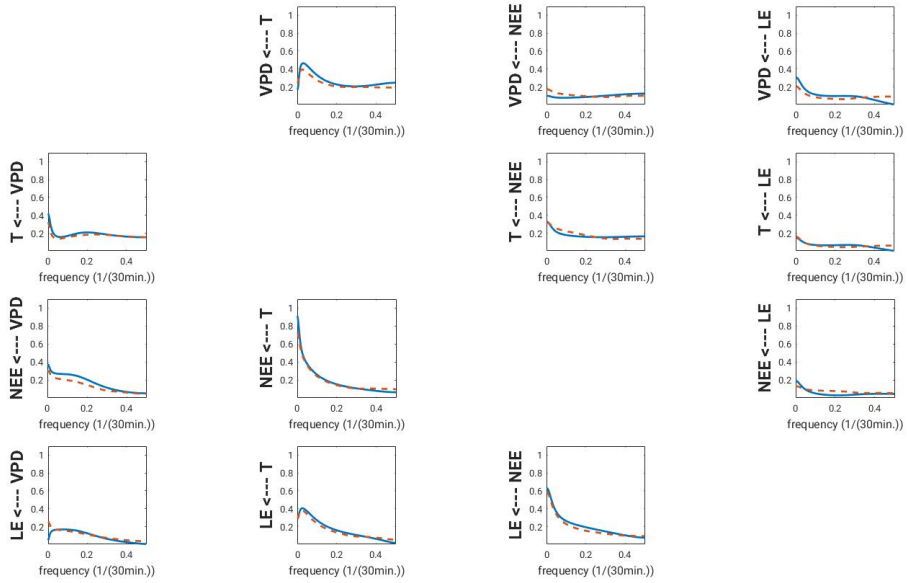

Figure 17: Plots of the gPDC representing the intensity of the cause-effect relationships between the four variables T, VPD, NEE and LE measured at the flux tower site of Puechabon-France during July-August 2002 (solid line) when compared to the average causal intensities of seven similar summer period within years 2000-2006 (red dash line).

herence for the analysis of the cause effect relationships between the EO variables involved. The advantage of the proposed method is that it allows for causality analysis at different frequency components and hence different time scales. Preliminary results show that the presented approach is a promising method for handling the presence of the periodic components necessary for accurate causality analysis.

Using the proposed frequency domain causality analysis method, we have shown that anomalous events can be detected as those events where the causal intensities between the variables differ considerably from the average dynamical behaviour, and such anomalous event can be attributed to the variable(s) causing such deviation.

4 Conclusions

This deliverable refers to the works done within *Task 5.4 - Methods for Attribution Scheme and Near Real-Time BACI*.

The work done along this task can be divided into three main parts:

1. The anomaly detection method using linear regression and Mahalanobis distance used to generate the first version of BACIndex has been further improved by using multivariate autoregressive model to account for the high correlation between the used variables at different time lags. This led to an improved temporal and spatial detection accuracy of the proposed anomaly detection method and even helped to detect several events that were missed when using univariate ARMA models. In particular MVAR model performed better in the detection of winter events. The improved version of the method has very low computational load making it suitable for real time implementation. The second version of BACI is uploaded to BACI portal and made available for BACI community.

2. We have presented an attribution scheme based on the decomposition of the Mahalanobis distance using Garthwaite Transform. The decomposed form of the Mahalanobis distance provides the answer to the question: how much each variable has contributed to the Mahalanobis distance? Experimental results of the developed attribution scheme used for the attribution of different known historic events were also presented and compared to the attribution results of the univariate z-score and the Mahalanobis distance ratio which is the reduction in the Mahalanobis distance obtained by eliminating specific variable. Unfortunately, detailed quantitative evaluation of the performance of the proposed methods is not possible due to the lack of the ground truth for the attribution of the selected extreme events. The results hence still need to be validated by climate scientists.

3. An attribution scheme based on the analysis of the cause-effect relationships in multivariate Earth observation system has been presented. First we proposed to use the parametric spectral MVAR-GC referred to as generalized partial directed coherence for the analysis of the cause effect relationships between the EO variables involved. The advantages of the proposed method is that it allows for causality analysis at different frequency components and hence different time scales. Preliminary results show that the presented approach is a promising method for handling the presence of the periodic components necessary for accurate causality analysis. Using the proposed frequency domain causality analysis method, we have shown that anomalous events can be detected as those events where the causal intensities between the variables differ considerably from the average dynamical behaviour, and such anomalous event can be attributed to the variable(s) causing such deviation. Further ongoing research will be focused on the selection criteria of the model order as well as the selection of the sampling frequency at different time scales of causality analysis.

With this deliverable, the achievement of Task 5.4 is ratified.

5 List of Publications

1. Detecting Regions of Maximal Divergence for spatiotemporal Anomaly Detection. Bjoern Barz, Erik Rodner, Yanira Guanche Garcia and Joachim Denzler. IEEE Transactions on Pattern Analysis and Machine Intelligence 2018 (code and GUI are available online).

2. Extreme anomaly event detection in biosphere using linear regression and a spatiotemporal MRF model. Yanira Guanche Garcia, Maha Shadaydeh, Miguel Mahecha and Joachim Denzler. Natural Hazards, 2018.

3. Causality analysis of ecological time series: a time-frequency approach. Maha Shadaydeh, Yanira Guanche Garcia, Miguel Mahecha, Markus Reichstein and Joachim Denzler. Climate Informatics Workshop, 2018.

4. Analyzing the time variant causality in ecological time series: a time-frequency approach. Maha Shadaydeh, Yanira Guanche Garcia, Miguel Mahecha, Markus Reichstein and Joachim Denzler. 10th International Conference on Ecological Informatics (ICEI), 2018.

5. Maximally Divergent Intervals for Extreme Weather Event Detection. Bjoern Barz, Yanira Guanche, Erik Rodner and Joachim Denzler. MTS/IEEE OCEANS Conference Aberdeen, 2017.

6. Biosphere Anomalies Detection by Regression Models. Yanira Guanche, Maha Shadaydeh, Miguel Mahecha and Joachim Denzler. Conference on Advances in Extreme Value Analysis and Application to Natural Hazards (EVAN), 2017.

7. Multivariate anomaly detection for Earth observations: a comparison of algorithms and feature extraction techniques. Milan Flach, Fabian Gans, Alexander Brenning, Joachim Denzler, Markus Reichstein, Erik Rodner, Sebastian Bathiany, Paul Bodesheim, Yanira Guanche, Sebasitan Sippel and Miguel D. Mahecha. Earth System Dynamics. 8 (3), 2017.

8. Detecting Multivariate Biosphere Extremes. Yanira Guanche Garcia, Erik Rodner, Milan Flach, Sebastian Sippel, Miguel Mahecha, Joachim Denzler. International Workshop on Climate Informatics (CI) 2016.

9. Maximally Divergent Intervals for Anomaly Detection. Erik Rodner, Bjorn Barz, Yanira Guanche, Milan Flach, Miguel Mahecha, Paul Bodesheim, Markus Reichstein and Joachim Denzler. ICML Workshop on Anomaly Detection (ICML-WS), 2016.

10. Using Statistical Process Control for detecting anomalies in multivariate spatiotemporal Earth Observations. Milan Flach, Miguel Mahecha, Fabian Gans, Erik Rodner, Paul Bodesheim, Yanira Guanche-Garcia, Alexander Brenning, Joachim Denzler and Markus Reichstein. European Geosciences Union General Assembly, 2016.

References

[1] Y. G. Guanche, M. Shadaydeh, M. Mahecha, and J. Denzler, "Extreme anomaly event detection in biosphere using linear regression and a spatiotemporal mrf model," *Natural Hazards*, pp. 1–19, 2018.

[2] P. H. Garthwaite and I. Koch, "Evaluating the contributions of individual variables to a quadratic form," *Australian & New Zealand journal of statistics*, vol. 58, no. 1, pp. 99–119, 2016.

[3] P. Mahalanobis, "On the generalised distance in statistics (vol.2, pp.49–55)," *Proceedings National Institute of Science, India. Retrieved from http://ir. isical. ac. in/dspace/handle/1/1268*, 1936.

[4] H. Hotelling, "Multivariate quality control," *Techniques of statistical analysis*, 1947.

[5] L. Barnett and A. K. Seth, "Behaviour of granger causality under filtering: Theoretical invariance and practical application," *Journal of Neuroscience Methods*, vol. 201, no. 2, pp. 404 – 419, 2011.

[6] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[7] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American statistical association*, vol. 77, no. 378, pp. 304–313, 1982.

[8] L. A. Baccalá, K. Sameshima, and D. Takahashi, "Generalized partial directed coherence," in *Digital Signal Processing, 2007 15th International Conference on*, pp. 163–166, IEEE, 2007.

[9] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biological Cybernetics*, vol. 84, pp. 463–474, May 2001.

[10] T. Anderson, *The Statistical Analysis of Time Series*. Wiley Classics Library, Wiley, 1994.

[11] L. Barnett and A. K. Seth, "The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference," *Journal of Neuroscience Methods*, vol. 223, pp. 50 – 68, 2014.