Journal of Statistical Software

June 2015, Volume 65, Issue 10. *http://www.jstatsoft.org/*

remote: Empirical Orthogonal Teleconnections in R

Tim Appelhans
Philipps University Marburg

Florian Detsch
Philipps University Marburg

Thomas Nauss
Philipps University Marburg

Abstract

In climate science, teleconnection analysis has a long standing history as a means for describing regions that exhibit above average capability of explaining variance over time within a certain spatial domain (e.g., global). The most prominent example of a global coupled ocean-atmosphere teleconnection is the El Niño Southern Oscillation. There are numerous signal decomposition methods for identifying such regions, the most widely used of which are (rotated) empirical orthogonal functions. First introduced by van den Dool, Saha, and Johansson (2000), empirical orthogonal teleconnections (EOT) denote a regression based approach that allows for straight-forward interpretation of the extracted modes. In this paper we present the R implementation of the original algorithm in the **remote** package. To highlight its usefulness, we provide three examples of potential use-case scenarios for the method including the replication of one of the original examples from van den Dool *et al.* (2000). Furthermore, we highlight the algorithm's use for cross-correlations between two different geographic fields (identifying sea surface temperature drivers for precipitation), as well as statistical downscaling from coarse to fine grids (using Normalized Difference Vegetation Index fields).

Keywords: teleconnection analysis, spatial data mining, **raster** data, R, climate science.

1. Introduction

With more than 30 years of continuous satellite observations and an ever increasing amount of available environmental models, time series analysis of gridded geoscientific data has become more and more popular in the recent past. One field where spatio-temporal analysis of gridded data has a long history is climate science. Due to the nature of the field, many studies are carried out on rather large spatial extents (e.g., global scales) and hence gridded data sets have been utilized intensively. Yet, even on smaller regional or local scales, high resolution satellite imagery, gridded interpolated point measurements or numerical model grids are able to provide very useful information that is spatially and temporally consistent.

In climate science, teleconnection analysis has a long standing history as a means for describing regions that exhibit above average capability of explaining variance over time within a certain spatial domain (e.g., global). The most prominent example of a global atmospheric teleconnection is the El Niño Southern Oscillation (ENSO). There are numerous signal decomposition methods for identifying such regions, the most widely used of which are (rotated) empirical orthogonal functions (EOF). First introduced by van den Dool *et al.* (2000), empirical orthogonal teleconnections (EOT) denote a correlation based approach that allows for straight-forward interpretation of the extracted modes. There are several studies that have made use of EOT analysis in various ways. Franzke (2002) used EOTs to characterize low-frequency flow variability modes and their influence on storm tracks in the Northern Hemisphere in a simplified global circulation model. Smith and Reynolds (2003) used both rotated EOFs and EOT analysis to investigate high-frequency anomalies in global sea surface temperatures (SST) and found that both approaches produce many patterns that are almost identical. To characterize drought severity across Europe between 1901 and 2002 van der Schrier, Briffa, Jones, and Osborn (2006) used EOT analysis for both spatial pattern isolation and time series analysis of these patterns. Yet another example of the use of EOTs can be found in Reynolds, Smith, Liu, Chelton, Casey, and Schlax (2007). Here, EOTs are used to bias correct satellite imagery as part of a methodology to create high-resolution (in space and time) SST grids. Rotstayn *et al.* (2010) utilized EOT analysis in order to evaluate the performance of a newly adapted coupled ocean-atmosphere global climate model, especially its ability to reproduce leading modes of precipitation variability in Australia. Also using EOT analysis Klingaman, Woolnough, and Syktus (2013) identified regions of high rainfall variability in Queensland, Australia, which they then related to large-scale drivers such as ENSO.

Despite the apparent diversity in the way of applying EOT analysis, all of the above-mentioned investigations are rooted within climate science. We believe that the potential of the EOT algorithm is much greater and that it can also be useful for other disciplines that make use of gridded data sets. Therefore, we hope that with a computationally efficient implementation of the EOT algorithm in the open source package **remote** (Appelhans, Detsch, and Nauss 2015) for R (R Core Team 2015), which is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=remote`, this methodology will be adopted more widely in the future, potentially also by disciplines other than the climate science community.

2. Eot algorithm

Empirical orthogonal teleconnections have first been introduced to the international literature as an alternative to the classical approach of empirical orthogonal functions by van den Dool *et al.* (2000). van den Dool (2007) outlines that both EOT and EOF are indeed very similar techniques with the former producing less abstract results. Both methods decompose spatiotemporal fields into a set of independent orthogonal patterns. In contrast to EOFs, which are orthogonal in both space and time, EOT analysis produces patterns that are orthogonal in either space or time (the current implementation of package **remote** provides the latter). EOTs carry a quantitative meaning in the form of explained variance, thus enabling intuitive interpretation of the results. Additionally, the method is easy to comprehend both methodically and algorithmically (see next paragraph). It is possible to calculate internal EOTs isolating teleconnection patterns within one spatio-temporal domain. We highlight this in Section 4.1 Figure 1. In addition, EOTs may also be used to investigate teleconnectivity between two domains (van den Dool 2007 calls this EOT2, Eastman 2009 named it cross-EOT). In this case, temporal variability of one domain (predictor) is analyzed with regard to explained variance of the temporal dynamics of another domain (response). Apart from similarity in the temporal dimension (i.e., identical amount of data points over time), the algorithm can be applied to any two domains without further requirements such as identical spatial resolution or physical units of the data. An example of this can be found in Figure 2 in Section 4.2 where we explain precipitation variation as a function of sea surface temperatures.

The mathematics of the EOT algorithm are described in detail in van den Dool *et al.* (2000) and van den Dool (2007) and can be summarized as follows. First, the temporal profiles of each pixel p_p of the predictor domain are regressed against the profiles of all pixels p_r in the response domain. The calculated coefficients of determination are summed up and the pixel with the highest sum for explaining variance within the response domain is identified as the "base point" of the first/leading mode. The temporal profile at this base point is the first/leading EOT. Then, the residuals from the regression are taken to be the basis for the calculation of the next EOT, thus ensuring orthogonality of the identified teleconnections. This procedure is repeated until a predefined amount of n EOTs is calculated. Generally, package **remote** can be considered a "brute force" spatial data mining approach to identify locations of enhanced potential to explain spatio-temporal variability within some geographic domain.

3. Package design and functionality

3.1. General package design

Package **remote** is based on the **raster** package (Hijmans and van Etten 2015) due to the fact that EOT analysis is usually applied to gridded space-time data (though in theory any irregular data may be analyzed this way as long as the area represented by each observation is known). This means that in order to use the functions provided by package **remote**, input must be of class '`Raster*`'. This ensures consistency and compatibility with all other R packages designed for the analysis of '`Raster*`' data. Essentially, all functionality provided by package **remote** could be achieved with the standard functions provided by package **raster**, and indeed many functions of the package at hand utilize these. However, as the nature of space-time analysis is such that data sets are commonly very large and that in order to find one base point, $p_p \cdot p_r$ computations are necessary, the regression calculations of package **remote** are implemented in C++ via package **Rcpp** (Eddelbuettel and François 2011; Eddelbuettel 2013). This ensures acceptable computation times and memory usage, even though calculations can still take time to complete, depending on the number of predictor and response pixels. In comparison to the most commonly available implementation of the EOT algorithm in IDRISI's Earth Trends Modeller (Version Taiga; Eastman 2009), computation time is reduced by one order of magnitude. In addition to the above-mentioned **raster** and **Rcpp** dependencies, package **remote** provides plotting methods that are based on the **lattice** and **latticeExtra** packages (Sarkar 2008; Sarkar and Andrews 2012) and use palettes from **ColorBrewer.org** (Harrower and Brewer 2003) through the **RColorBrewer** package (Neuwirth 2011). Package **remote** is licensed under the GNU General Public License version 3 and available from CRAN. The development version is hosted at `https://github.com/environmentalinformatics-marburg/remote` and can be installed with the `install_github()` function available in the **devtools** package (Wickham and Chang 2013). Below, we give a detailed description of the classes, methods and functions implemented in package **remote**.

3.2. Classes

As pointed out before, the input to all methods and functions provided by **remote** need to be of class '`Raster*`', in particular '`RasterBrick`' or '`RasterStack`'. The workhorse function of package **remote** is `eot()`. It calculates a variety of outcomes which are returned as class '`EotMode`' if only one mode is calculated or '`EotStack`' in the case of multiple modes.

The basic class '`EotMode`' includes all necessary results from the regression calculations performed by `eot()` and has the following slots (classes are given in brackets):

- `mode`: The number of the identified mode (integer).
- `eot`: The EOT (time series) at the identified base point(numeric).
- `coords_bp`: The coordinates of the identified base point (matrix).
- `cell_bp`: The cell number of the identified base point (integer).
- `cum_exp_var`: The (cumulative) explained variance of the considered EOT (numeric).
- `r_predictor`: The raster of the correlation coefficients between the base point and each pixel of the predictor domain ('`RasterLayer`').
- `rsq_predictor`: As above but for the coefficient of determination ('`RasterLayer`').
- `rsq_sums_predictor`: As above but for the sums of coefficient of determination ('`RasterLayer`').
- `int_predictor`: The raster of the intercept of the regression equation for each pixel of the predictor domain ('`RasterLayer`').
- `slp_predictor`: Same as above but for the slope of the regression equation for each pixel of the predictor domain ('`RasterLayer`').
- `p_predictor`: The raster of the significance (p-value) of the the regression equation for each pixel of the predictor domain ('`RasterLayer`').
- `resid_predictor`: The '`RasterBrick`' of the reduced data for the predictor domain ('`RasterBrick`').

Apart from `rsq_sums_predictor`, all `*_predictor` slots are also available for the `*_response` domain, even if predictor and response domain are equal. This is due to that fact, that if not both fields are reduced after the first EOT is found, these '`Raster*`' objects will differ. If the user chooses to write the results to disk, all '`Raster*`' objects outlined above will be saved as native **raster** `.grd` files. Note that this may require rather large amounts of disk space, as for each identified EOT 13 '`Raster*`' objects will be written to disk. Location (`x, y`), EOT number `n`, cumulative explained variance along with a comment as to whether the location of the EOT is unambiguous will be saved in a `.csv` file in a user-supplied location on the hard disk. This enables further investigations of the results without having to re-run the analysis from scratch, which, in light of the sometimes high computation times, is indispensable.

Multiple modes are stored in an object of class '`EotStack`' with two slots. The first slot is called `modes` and is essentially a list containing all calculated '`EotModes`'. The second slot is called `names` and stores the names of the calculated '`EotModes`'.

3.3. Methods and functionality

Given below is a detailed list of the available **remote** functions and methods (in alphabetical order). Most of these will be utilized in Section 4.2, however, the reader is referred to the documentation where in-depth descriptions and detailed examples are readily available. To highlight the functionality of the **remote** package, we provide three use-case scenarios of potential applications of the package in Section 4.

Currently, **remote** provides the following functionality (only user relevant functions are shown):

- `anomalize()`: Create an anomaly space-time field from an object of class '`RasterStack`' or '`RasterBrick`'. Either based on the overall mean of the object of class '`RasterStack`' or '`RasterBrick`', or a supplied reference '`RasterLayer`'.

- `calcVar()`: Calculate the mean (optionally standardized) space-time variance of a '`RasterStack`' or '`RasterBrick`'.

- `cutStack()`: Remove a specified number of layers from the beginning or the end of a '`RasterStack`' or '`RasterBrick`'. This is used in `lagalize()` to create lagged '`Raster*`' objects.

- `denoise()`: Noise filtering through principal components. The user can either specify how many components to keep or can provide a value for the minimum variance that should be kept. Additionally, the user can choose whether the field should be geographically weighted (see `geoWeight()`).

- `deseason()`: Create seasonal anomalies of a '`RasterStack`' or '`RasterBrick`' by supplying a suitable seasonal window.

- `eot()`: The core function of the package. The user supplies a predictor '`RasterStack`' or '`RasterBrick`' and (optionally) a response '`RasterStack`' or '`RasterBrick`', the number of `n` EOT modes to be calculated, whether the results should be standardized (i.e., R^2 values be multiplied by the variance), whether both predictor and response (or only the latter) should be reduced after the first mode is identified (i.e., the residuals be taken) and the type of the link function (either correlation or index of directional agreement). Optionally, all results can be written to disk to a user-supplied path and the amount of information printed to the console can be controlled.

- `geoWeight()`: Create geographically weighted fields by supplying a suitable function to be used (e.g., *cosine*) in order to compensate for non-equal area grids in case of non-projected geographical data (i.e., lat/lon coordinate reference system). The function is applied to the (radians of) latitude of the supplied '`Raster*`' object.

- `lagalize()`: Create time-lagged '`RasterStacks`' or '`RasterBricks`' by choosing a suitable lag number with regard to the frequency of the data.

- `names()`: Get or set the names of '`Eot*`' objects.

- `nmodes()`: Get the number of modes of an '`EotStack`'.

- `nXplain()`: identify the number of modes needed to explain a certain, user-supplied amount of variance within the response series of an '`EotStack`'.

- `plot()`: Plotting method of package **remote**. By default three panels are drawn: i) the coefficient of determination image of the predictor domain, ii) the correlation coefficient image for the response domain and iii) the times series of the identified EOT at the base point. It is possible to control which of the `n` EOT modes should be displayed and the combination of images shown is completely flexible. Additionally, the color palette can be controlled and further minor modifications can be made (see documentation). Setting argument `locations` to `TRUE` will produce a map showing all identified base points color coded by mode (1−`n`).

- `predict()`: Spatial predictions using the fitted models calculated with `eot()`. A user-defined set of `n` `EotModes` will be used to model the outcome using the identified link functions of the respective modes which are added together to produce the final prediction.

- `subset()`: Extract a set of modes from an '`EotStack`'.

- `writeEot()`: Write '`Eot*`' objects to disk. This is merely a wrapper around `writeRaster()` from the **raster** package with argument `overwrite` set to `TRUE` by default.

3.4. Data sets

Package **remote** includes three data sets: `vdendool`, `australiaGPCP` and `pacificSST`. All of these data sets are used in the package documentation examples and in the following section which also provides detailed descriptions about these data. All data sets are of class '`RasterBrick`'.

4. Examples

This section highlights three different use-case scenarios for potential applications of the **remote** package (though we are confident that there are many more exciting applications to be found).

1. As a first example, we replicate one of the examples from van den Dool *et al.* (2000, Example 3.d., Figure 6). A spatio-temporal field of 700 mb geopotential heights of NCEP/NCAR reanalysis grids (Kalnay *et al.* 1996) is decomposed into its four leading modes exhibiting the prominent patterns of North Atlantic Oscillation (NAO) and Pacific-North American Pattern (PNA) as modes 1 and 2, respectively.

2. A second example highlights the application of an EOT analysis between two geoscientific fields. We identify influential areas of sea-surface temperature anomalies in the tropical Pacific Ocean that remotely drive precipitation dynamics over mainland Australia. For this we use NOAA OI SST V2 (Reynolds *et al.* 2007) as a predictor field and Global Precipitation Climatology Project data (GPCP V2.2; Adler *et al.* 2003) as the response series. The aim is to directly identify regions of sea surface temperature variability that exhibit enhanced ability of explaining precipitation variations over mainland Australia. So far, only indirect approaches have been taken (e.g., Klingaman *et al.* 2013) where internal base points of precipitation variability have been identified and then regressed against standard teleconnection indices, such as the Southern Oscillation Index (SOI).

3. As a final example we show a completely different application of the **remote** functionality. The intention is to spatially down-sample the Global Inventory Modelling and Mapping Studies (GIMMS, V2.0) NDVI product (Tucker *et al.* 2005) with a resolution of 8 km to MODIS NDVI MYD13Q1 observations (LPDAAC 2006) with a resolution of 250 m for the region of Mt. Kilimanjaro. We evaluate the quality of the resulting artificial series by comparing the calculated NDVI images with MODIS NDVI images that were not considered during model definition. Furthermore, we visually assess the performance of the calculated time series for a number of randomly selected pixels.

4.1. Winter mean 700 mb height over the Northern Hemisphere

Section 3.d. in van den Dool *et al.* (2000) provides an excellent example of the use of the EOT algorithm to extract atmospheric teleconnections using winter mean 700 mb heights over the Northern Hemisphere. The climatologically inclined reader is referred to the respective Section in van den Dool *et al.* (2000) for a more detailed description of the atmospheric dynamics and processes associated with the identified patterns. Here, we merely want to highlight that the **remote** implementation of the algorithm produces similar results to those presented in van den Dool *et al.* (2000).

The data set needed for this example is included in package **remote** and the four modes needed for the example can be calculated with:

```
R> data("vdendool", package = "remote")
R> nh_modes <- eot(x = vdendool, y = NULL, n = 4, reduce.both = FALSE,
+    standardised = FALSE, verbose = FALSE)
```

In order to recreate Figure 6 from van den Dool *et al.* (2000), some additional lines of code are necessary (including spatial re-projection of the rasters). These can be found in the supplementary material and will produce Figure 1.

Even though the location of the identified base points (BP) is somewhat offset, and hence the explained variance (EV) figures differ slightly, it is obvious that the isolated patterns are very similar and represent the same signals. We can only speculate as to why the base point locations differ slightly, but potential reasons may include different version numbers of the reanalysis data, differences in the final data sets introduced through different aggregation calculations while thinning the original 2.5° × 2.5° reanalysis data to 10° × 5° grids, rounding

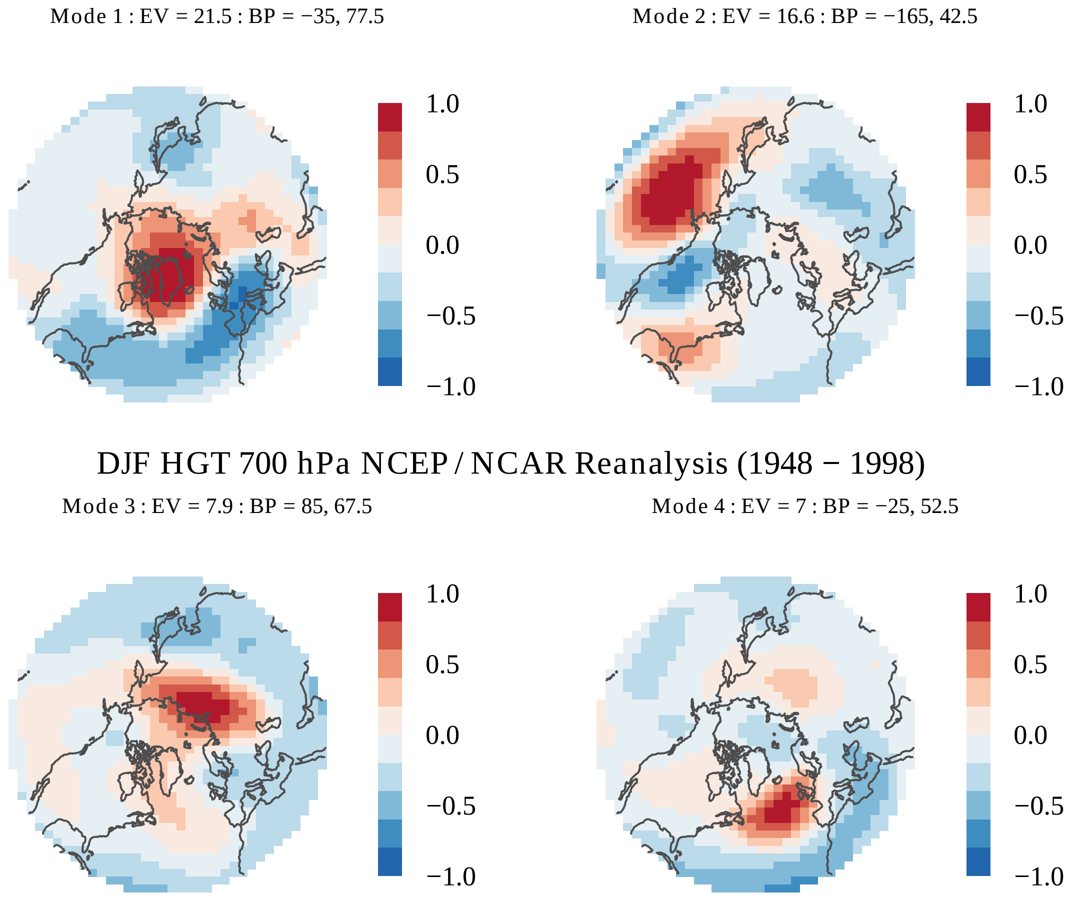

Mode 1 : EV = 21.5 : BP = −35, 77.5
Mode 2 : EV = 16.6 : BP = −165, 42.5
DJF HGT 700 hPa NCEP / NCAR Reanalysis (1948 − 1998)
Mode 3 : EV = 7.9 : BP = 85, 67.5
Mode 4 : EV = 7 : BP = −25, 52.5
1.0
0.5
0.0
−0.5
−1.0

Figure 1: Replication of Figure 6 from van den Dool *et al.* (2000). EV is explained variance of the response domain, BP is the location of the identified base point (longitude, latitude). Resolution of the grids is 10/5 degrees of longitude/latitude.

discrepancies between the utilized software environments (especially when summing up the coefficients of determination) and slight differences in geographic projections.

4.2. Identifying tropical Pacific SST drivers for Australian precipitation

The processes of precipitation development are complex and not yet understood completely. The physical state of the atmosphere, which determines whether rain occurs or not at any point in space and time, is the result of a multitude of constantly changing factors. Influences range from local to hemispheric boundary conditions in all four dimensions (including time). Some areas of the global oceans exhibit low-frequency anomaly signals which can influence precipitation variability world-wide. The most prominent example of a coupled ocean-atmosphere tropical SST variability is ENSO. ENSO has received much attention in the scientific literature since the major 1982–83 El Niño. Here, we investigate whether EOT analysis can be used to identify the ENSO signal as a driver for monthly Australian precipitation variability over the period 1982 to 2010. The data sets needed for this analysis are included in package **remote**. In order to reveal low-frequency signals such as ENSO, we need to prepare the raw data fields so that high-frequency variation is eliminated. We achieve this by creating seasonal anomalies using `deseason()` and by `denoise()`-ing the data to filter out some of the noise that is present in any spatio-temporal data field. The first/leading mode of SSTs most influential for Australian rainfall variability can be calculated with:

```
R> data("australiaGPCP", package = "remote")
R> data("pacificSST", package = "remote")
R> sst_pred <- deseason(pacificSST, cycle.window = 12)
R> gpcp_resp <- deseason(australiaGPCP, cycle.window = 12)
R> sst_pred_dns <- denoise(sst_pred, expl.var = 0.9)

Using the first 19 components (of 348) to reconstruct series...
  these account for 0.9 of variance in orig. series.

R> gpcp_resp_dns <- denoise(gpcp_resp, expl.var = 0.9)

Using the first 37 components (of 348) to reconstruct series...
  these account for 0.9 of variance in orig.~series.

R> aus_modes <- eot(x = sst_pred_dns, y = gpcp_resp_dns, n = 1,
+    standardised = FALSE, reduce.both = FALSE, verbose = FALSE)
```

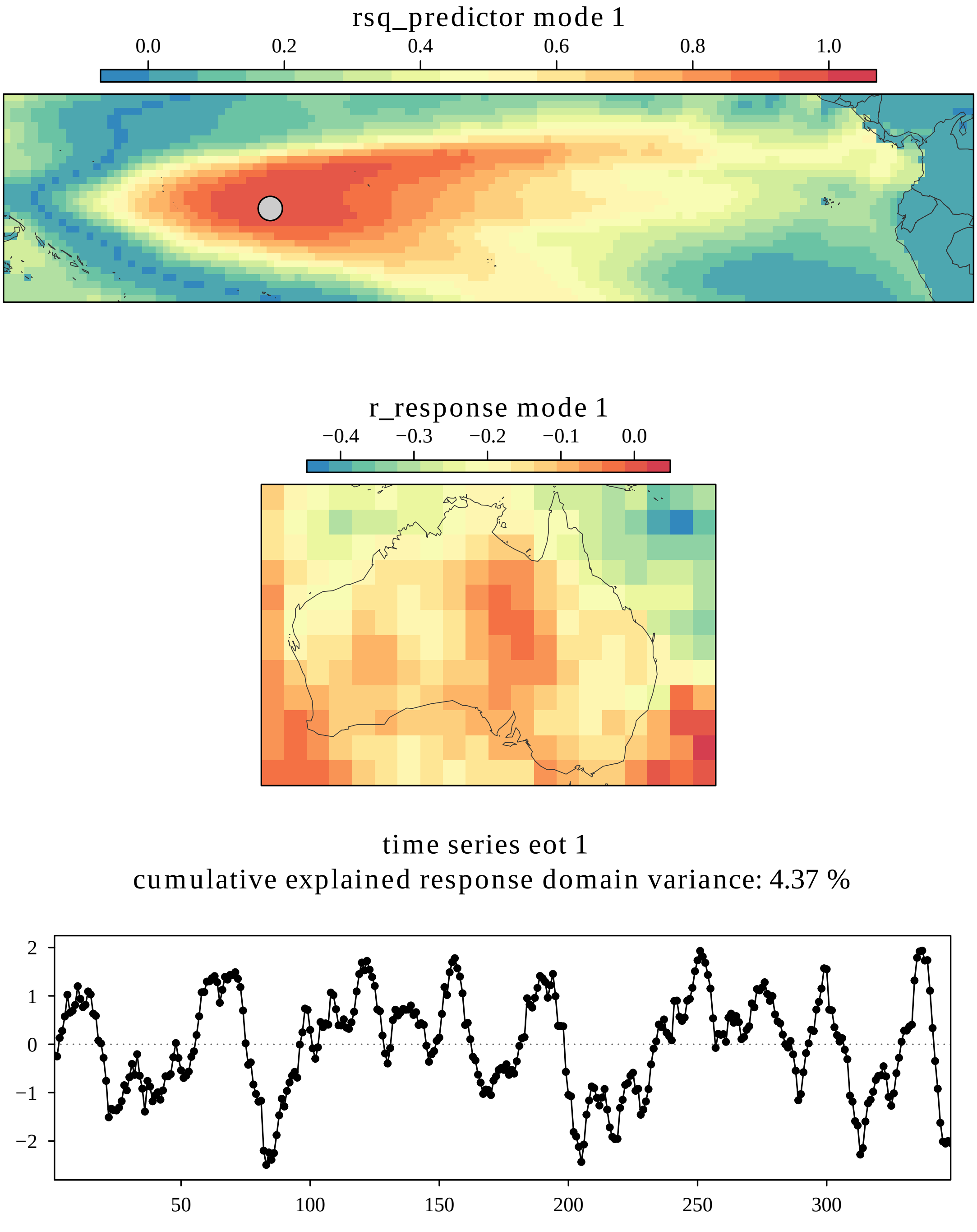
As we can see, especially the principal components filter from `denoise()` is an important step, as it tells us that we need only 19 (37) of the original 348 (12 months · 29 years) components for the SST (GPCP) data to explain 90% of the respective inherent field variance. To get a visual impression, the results for the first leading mode can be plotted using `plot()`:

```
R> plot(aus_modes, y = 1, show.bp = TRUE, arrange = "long")
```

We see that we are indeed able to isolate the ENSO signal as the most important SST driver in the tropical Pacific for Australian precipitation. This signal is able to explain just above 4% of the original variation found in rainfall over the analyzed period (1982–2010). This may not seem much, but we need to keep in mind that precipitation is influenced by many factors, with local conditions playing a major role. Spatially, mainly the North-Eastern part of the response domain is being explained with some locations showing negative correlations of up to 0.4. With regard to mainland Australia, it becomes obvious that the identified ENSO signal is not able to explain any rainfall variation in the inner-continental parts of the land mass. It is mainly the coastal areas that are influenced by the ENSO phenomenon, which is in line with the findings of Risbey, Pook, McIntosh, Wheeler, and Hendon (2009). Note that our analysis did not take into account any time lags between the SST anomalies and precipitation. Even though in this particular example lagging does not increase the explanatory power of the SST signal (not shown), it can be expected that in many cases the influence will not manifest instantaneously and that a certain time lag will explain a higher portion of the rainfall variance.

4.3. Downscaling GIMMS NDVI to MODIS NDVI

As much of the authors' research is focused on eco-climatological studies at and around Mt. Kilimanjaro, this region shall be the focus of our last use-case. In order to infer sound



rsq_predictor mode 1

r_response mode 1

time series eot 1
cumulative explained response domain variance: 4.37 %

Figure 2: Coefficient of determination image of the first leading mode of tropical Pacific SSTs (predictor – top panel) along with correlation coefficient image of precipitation over mainland Australia (response – central panel). The time series at the base point of this mode (gray circle in top panel) is shown in the bottom panel. Time span of analysis is January 1982–December 2010. Resolution of SST grid (predictor) is 1° × 1° long/lat, resolution of GPCP grid (response) is 2.5° × 2.5° long/lat.

trends of climatic or related environmental parameters, it is essential to have time series that span a long enough period to capture some of the low-frequency signals that might influence their temporal dynamics. One of the parameters frequently used in eco-climatological studies is the Normalized Difference Vegetation Index (NDVI). The NDVI is a normalized difference between red and near infra-red reflectances and gives information on the "greenness" of vegetation. Its essential characteristics are such that it ranges between -1 and 1 with dense green vegetation having positive values (close to 1) and clouds, water, ice and snow being characterized by negative values. The MODIS sensor on board NASA's Aqua satellite platform provides a standard NDVI product (MYD13Q1) that has been widely used for a diverse range of investigations. Its horizontal resolution is 250 m by 250 m and with a repeat rate of 16 days it is perfectly suited for spatio-temporal vegetation analyses that require high-resolution imagery. One of the biggest draw-backs of this data is that it is only available from mid 2002 and is therefore temporally not comprehensive enough for investigations over long time scales, such as the above-mentioned trend analyses. Another readily available and also widely used NDVI data set that spans a much longer time period (1982 to 2006) is GIMMS NDVI V2.0. This data, however, has a horizontal resolution of only 8 km by 8 km. In this last example we show that it is possible to utilize the **remote** package for the spatial downscaling of GIMMS NDVI to MODIS resolution to create a consistent data set that provides high-resolution NDVI information for the period of the GIMMS data set.

There are several commonly used approaches for statistical downscaling. Among others, these include PCA based approaches (e.g., Ehrendorfer 1987), machine learning approaches (e.g., Sachindra, Huang, Barton, and Perera 2013; Chen, Yu, and Tang 2010; Tripathi, Srinivas, and Nanjundiah 2006; Olsson *et al.* 2003; Davy, Woods, Russell, and Coppin 2010; Vaca, Golicher, and Cayuela 2011) and multiple linear and non-linear regression based approaches, or combinations of these (for an overview of commonly used approaches see e.g., Huth 1999; Wilby and Wigley 1997). The specifics of the methodologies differ depending on the approach taken. PCA based approaches generally identify a number of components in the predictor domain(s) that together explain a certain amount of variance of the response data. This approach is very similar to the one using EOTs. In this case, the predictor domain is screened for one grid-point that explains most of the variance in the entire response domain. This is repeated on the reduced response domain to identify the next important grid-point until a pre-defined number of reduced grid-point contributions is identified. Hence, for the downscaling using package **remote** it is advisable to calculate many EOTs in order to explain a large enough portion of the response variance. The exact amount obviously depends on the nature of the relationship between predictor and response domains and on the intended application. The number of EOTs needed to explain a certain, user-supplied amount of variance can easily be calculated with the **remote** function `nXplain()`.

In order to establish a relationship between GIMMS and MODIS data, we need the data sets to overlap for some time. Obviously, the larger the overlap, the better the prediction will be. In the case of GIMMS vs. MODIS, the available overlap is four years (2003–2006). As we need to be able to evaluate the prediction performance, we split these four years into a training series and an evaluation series of two years each. Here, we simply take the first two years (2003–2004) for model training and the last two years (2005–2006) for evaluation. Of course, more sophisticated approaches using random sampling of a number of layers are possible, but given that vegetation dynamics in the Mt. Kilimanjaro region exhibit seasonality, it is important that the monthly samples be equally distributed. Therefore, having two full years for the training and two full years for evaluation seems sufficient.

The data used for this example can be found in the supplementary material to this article. It needs to be pointed out that the MODIS data set (`modisKiliNDVI`) was exposed to some pre-processing due to high levels of cloud contamination. To address this issue, we used the so-called Whittaker smoother (Atzberger and Eilers 2011) which is available as part of the **MODIS** package (Mattiuzzi *et al.* 2014). We applied the Whittaker smoother with a lambda of 6000 and three iterations. The GIMMS NDVI data is provided gap-free but we applied the same smoothing to make the two data sets more comparable. In order to capture a large enough amount of variability we calculate the first ten EOTs:

```
R> load("gimmsKiliNDVI.RData")
R> load("modisKiliNDVI.RData")
R> pred_ind <- 1:24
R> mod_stck_pred <- modisKiliNDVI[[pred_ind]]
R> mod_stck_eval <- modisKiliNDVI[[-pred_ind]]
R> gimms_stck_pred <- gimmsKiliNDVI[[pred_ind]]
R> gimms_stck_eval <- gimmsKiliNDVI[[-pred_ind]]
R> ndvi_modes <- eot(x = gimms_stck_pred, y = mod_stck_pred, n = 10,
+    standardised = FALSE, reduce.both = FALSE, verbose = FALSE)
```

As the results of `eot()` include all the necessary output from the regression analysis, namely the '`RasterLayers`' of intercept and slope, we can create an artificial data set as a function of GIMMS NDVI with the resolution of MODIS NDVI for the evaluation time period 2005–2006. For this, we first calculate the minimum number of modes (`nm`) needed to explain at least 98% of the response domain variance using `nXplain()`. In the example at hand, six modes are needed. We then use `predict()` to make the spatial predictions from the evaluation data set.

```
R> nm <- nXplain(ndvi_modes, 0.98)
R> mod_predicted <- predict(object = ndvi_modes,
+    newdata = gimms_stck_eval, n = nm)
```

In order to assess the downscaling performance we calculate the mean error (ME), the mean absolute error (MAE), the root mean square error (RMSE), the correlation coefficient (R) and the coefficient of determination (Rsq) shown in Figure 3. The code to produce this figure and all subsequent figures can be found in the supplementary material.

We see that the mean error is generally slightly negative indicating a general underestimation of the NDVI values. However, with less than 0.1, the general prediction error is generally low as denoted by the MAE. Additionally, a median coefficient of determination of around 0.9 indicates very good prediction performance.

A look at the spatio-temporal distribution of the residuals (observed NDVI − predicted NDVI) in Figure 4 exhibits that the most difficult areas to predict are the lowlands around Mt. Kilimanjaro. These regions are dominated by extensively managed agricultural crops such as maize, wheat and sunflower fields that are indeed depending on the seasonal precipitation cycle to a high degree (i.e., they are not artificially irrigated). Yet, they exhibit different temporal responses to the local precipitation climatology than the also abundant natural

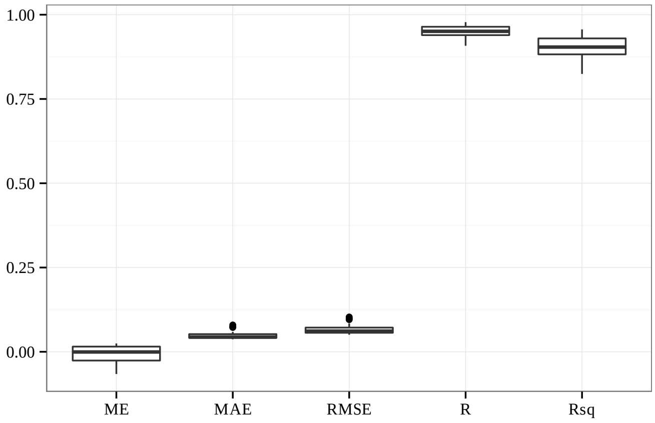


Figure 3: Boxplot of evaluation statistics for quality of downscaling. ME – Mean Error; MAE – Mean Absolute Error; RMSE – Root Mean Square Error; R – correlation coefficient; Rsq. – coefficient of determination.

Savannah areas. The herbaceous layer of the Savannah ecosystems usually responds immediately to increased water input resulting in a rather large NDVI increase over very short time periods with a gradual decrease thereafter. For the managed lowland systems this behavior is usually reversed.

Figure 5 shows time series comparisons for 50 randomly selected pixels from the scenes shown in Figure 4. Both central tendency (dashed lines) and variance of the time series are captured well by the EOT approach. In summary, the general performance of the prediction seems very reasonable, especially in light of the high complexity of the investigated region.

5. Conclusions

In this paper we have presented the R implementation of empirical orthogonal teleconnection analysis as introduced to the scientific community by van den Dool *et al.* (2000). We have highlighted the general **remote** package design as well as its classes and the specific set of methods and functions that **remote** provides. Furthermore, we have shown a range of use-case scenarios of **remote** in three distinct examples including the replication of one of the original examples from van den Dool *et al.* (2000).

Especially the utilization of package **Rcpp** for the computation-intensive calculations ensures acceptable computation times and memory usage for the "brute force" spatial data mining algorithm at hand. This is a very important aspect of package **remote** as the amount of data points in spatio-temporal geoscientific fields is generally extremely large and can easily require millions, or even billions of calculations. The general design of package **remote** ensures easy integration into potential existing work flows as it is largely based on the **raster** package which, as the name implies, is the standard foundation for raster analyses in R. In its current implementation package **remote** provides all the tools to utilize the EOT algorithm as postulated by van den Dool *et al.* (2000). For the future, we plan to implement space-time reversal to facilitate the search for patterns that are orthogonal in space. Furthermore, our aim is to enhance functionality through the integration of further link functions such

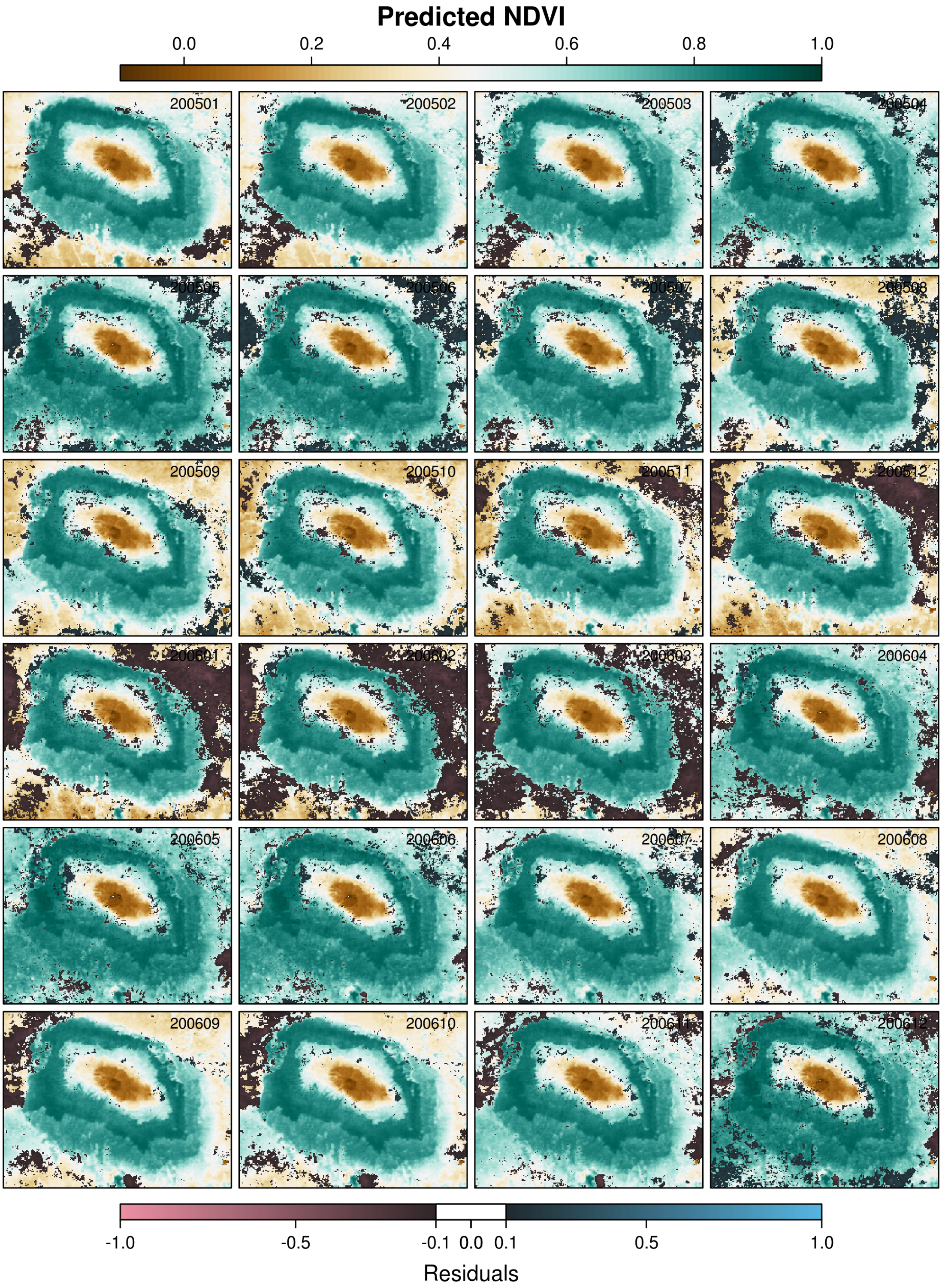


Figure 4: Predicted NDVI images showing spatial distribution of residuals from the prediction. Only residuals > 0.1 / < -0.1 are shown. The date of the scene is shown in the top right corner of each panel.

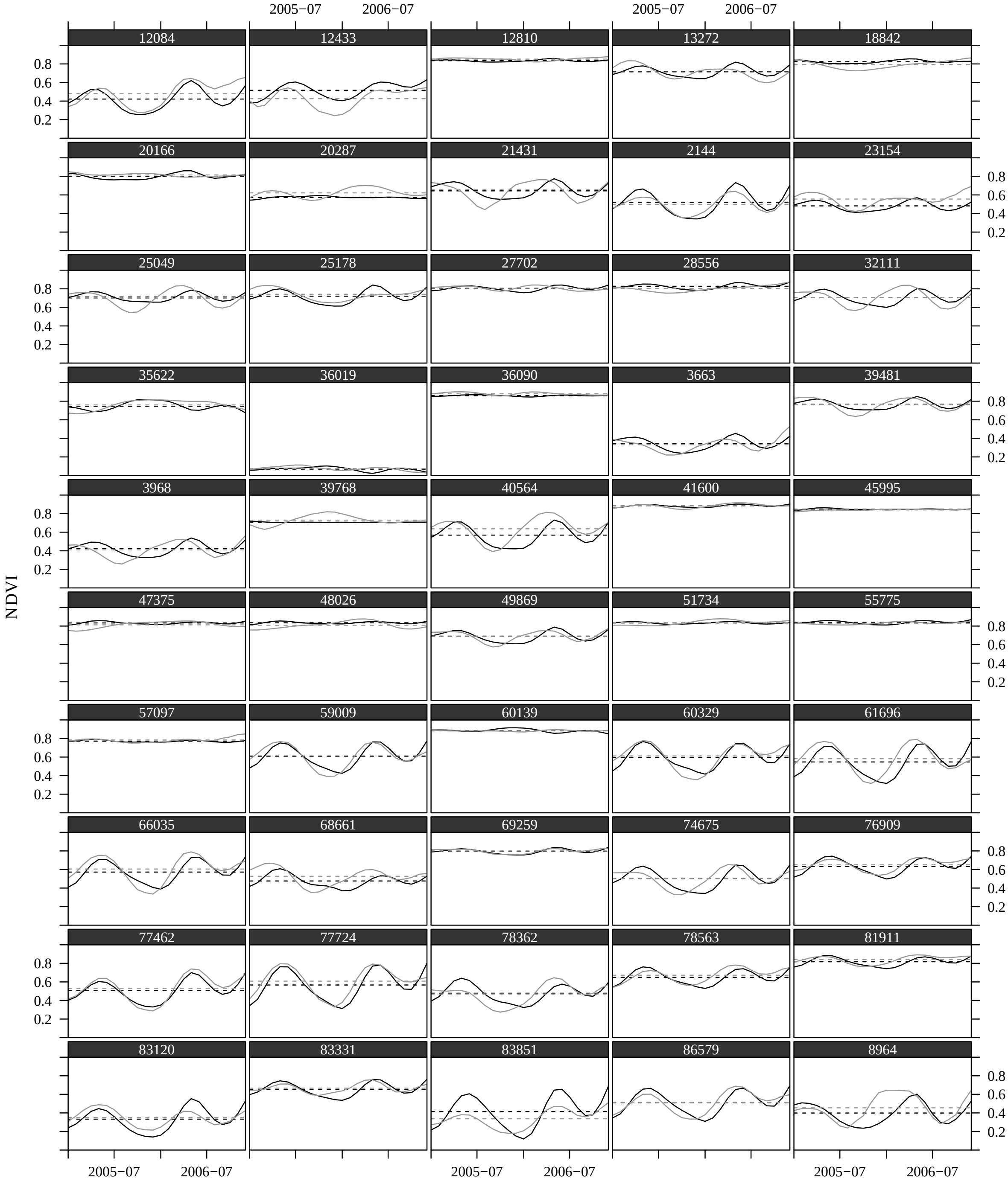


Figure 5: Predicted and observed NDVI time series for 50 randomly selected pixels. Black lines are predicted NDVI values, gray lines are MODIS NDVI observations. Dashed lines denote the mean of the respective series. Pixel numbers are shown above each panel.

as non-parametric and/or non-linear approaches. Furthermore, it is intended to extend the functionality towards providing the ability to use multiple predictor fields as well as more options for data preparation and visualization of the results.

The examples provided in this paper highlight the diverse range of potential uses for package **remote**. The first example is merely intended to show that the **remote** implementation of the algorithm produces results similar to those of the original implementation by van den Dool *et al.* (2000). Example 2 highlights the possibility to calculate EOTs across geoscientific data fields of different parameters and it is shown that it is indeed possible to capture "real" signals. Example 3 takes a slightly different approach of applying the **remote** algorithm by using the method to statistically downscale NDVI observations from a coarse to a fine gridded data set. Here, we see that some portions of the data are captured better than others, but that the general prediction performance is very encouraging.

In summary, we are confident that the introduced **remote** package is able to expand the possibilities for researchers of many disciplines to identify signals within spatio-temporal geoscientific data sets in order to assess influential patterns of space-time variability for various applications. Furthermore, we are hopeful that package **remote** will be used in many applications other than the ones highlighted in this paper.

Acknowledgments

The authors would like to express their special gratitude to Huug van den Dool for providing detailed information and clarification of some of the specifics of the original code and for very useful comments during the review process. Furthermore, the authors would like to thank the German Research Foundation (DFG) for providing some of the authors' funding within the framework of the DFG-Research Unit 1246 KiLi – "Kilimanjaro ecosystems under global change: Linking biodiversity, biotic interactions and bio-geochemical ecosystem processes" (https://www.kilimanjaro.biozentrum.uni-wuerzburg.de/).

References

Adler RF, Huffman GJ, Chang A, Ferraro R, Xie PP, Janowiak J, Rudolf B, Schneider U, Curtis S, Bolvin D, Gruber A, Susskind J, Arkin P, Nelkin E (2003). "The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present)." *Journal of Hydrometeorology*, **4**(6), 1147–1167.

Appelhans T, Detsch F, Nauss T (2015). ***remote**: Empirical Orthogonal Teleconnections in R*. R package version 1.0.0, URL http://CRAN.R-project.org/package=remote.

Atzberger C, Eilers PHC (2011). "A Time Series for Monitoring Vegetation Activity and Phenology at 10-Daily Time Steps Covering Large Parts of South America." *International Journal of Digital Earth*, **4**(5), 365–386.

Chen ST, Yu PS, Tang YH (2010). "Statistical Downscaling of Daily Precipitation Using Support Vector Machines and Multivariate Analysis." *Journal of Hydrology*, **385**(1), 13–22.

Davy RJ, Woods MJ, Russell CJ, Coppin PA (2010). "Statistical Downscaling of Wind Variability from Meteorological Fields." *Boundary-Layer Meteorology*, **135**(1), 161–175.

Eastman JR (2009). *IDRISI Taiga*. Clark University, Worcester. URL http://clarklabs.org/.

Eddelbuettel D (2013). *Seamless R and C++ Integration with* ***Rcpp***. Springer-Verlag, New York.

Eddelbuettel D, François R (2011). "**Rcpp**: Seamless R and C++ Integration." *Journal of Statistical Software*, **40**(8), 1–18. URL http://www.jstatsoft.org/v40/i08/.

Ehrendorfer M (1987). "A Regionalization of Austria's Precipitation Climate Using Principal Component Analysis." *Journal of Climatology*, **7**(1), 71–89.

Franzke C (2002). "Dynamics of Low-Frequency Variability: Barotropic Mode." *Journal of the Atmospheric Sciences*, **59**(20), 2897–2909.

Harrower M, Brewer CA (2003). "**ColorBrewer.org**: An Online Tool for Selecting Colour Schemes for Maps." *The Cartographic Journal*, **40**(1), 37–37.

Hijmans RJ, van Etten J (2015). ***raster****: Geographic Data Analysis and Modeling.* R package version 2.3-33, URL http://CRAN.R-project.org/package=raster.

Huth R (1999). "Statistical Downscaling in Central Europe: Evaluation of Methods and Potential Predictors." *Climate Research*, **13**(2), 91–101.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D (1996). "The NCEP/NCAR 40-Year Reanalysis Project." *Bulletin of the American Meteorological Society*, **77**(3), 437–471.

Klingaman NP, Woolnough SJ, Syktus J (2013). "On the Drivers of Inter-Annual and Decadal Rainfall Variability in Queensland, Australia." *International Journal of Climatology*, **33**(10), 2413–2430.

LPDAAC (2006). "MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid, MYD13Q1." URL https://lpdaac.usgs.gov/.

Mattiuzzi M, Verbesselt J, Stevens F, Mosher S, Hengl T, Klisch A, Evans B, Lobo A (2014). ***MODIS****: MODIS Acquisition and Processing Package.* R package version 0.10-18, URL http://R-Forge.R-project.org/projects/modis/.

Neuwirth E (2011). ***RColorBrewer****: ColorBrewer Palettes.* R package version 1.0-5, URL http://CRAN.R-project.org/package=RColorBrewer.

Olsson J, Uvo C, Jinno K, Kawamura A, Nishiyama K, Koreeda N, Nakashima T, Morita O (2003). "Neural Networks for Rainfall Forecasting by Atmospheric Downscaling." *Journal of Hydrologic Engineering*, **9**(1), 1–12.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007). "Daily High-Resolution-Blended Analyses for Sea Surface Temperature." *Journal of Climate*, **20**(22), 5473–5496.

Risbey JS, Pook MJ, McIntosh PC, Wheeler MC, Hendon HH (2009). "On the Remote Drivers of Rainfall Variability in Australia." *Monthly Weather Review*, **137**(10), 3233–3253.

Rotstayn LD, Collier MA, Dix MR, Feng Y, Gordon HB, O'Farrell SP, Smith IN, Syktus J (2010). "Improved Simulation of Australian Climate and ENSO-Related Rainfall Variability in a Global Climate Model with an Interactive Aerosol Treatment." *International Journal of Climatology*, **30**(7), 1067–1088.

Sachindra DA, Huang F, Barton A, Perera BJC (2013). "Least Square Support Vector and Multi-Linear Regression for Statistically Downscaling General Circulation Model Outputs to Catchment Streamflows." *International Journal of Climatology*, **33**(5), 1087–1106.

Sarkar D (2008). ***lattice**: Multivariate Data Visualization with R*. Springer-Verlag, New York.

Sarkar D, Andrews F (2012). ***latticeExtra**: Extra Graphical Utilities Based on **lattice***. R package version 0.6-24, URL http://CRAN.R-project.org/package=latticeExtra.

Smith TM, Reynolds RW (2003). "Extended Reconstruction of Global Sea Surface Temperatures Based on COADS Data (1854–1997)." *Journal of Climate*, **16**(10), 1495–1510.

Tripathi S, Srinivas V, Nanjundiah RS (2006). "Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach." *Journal of Hydrology*, **330**(3), 621–640.

Tucker CJ, Pinzon JE, Brown ME, Slayback DA, Pak EW, Mahoney R, Vermote EF, Saleous NE (2005). "An Extended AVHRR 8 km NDVI Dataset Compatible with MODIS and SPOT Vegetation NDVI Data." *International Journal of Remote Sensing*, **26**(20), 4485–4498.

Vaca RA, Golicher DJ, Cayuela L (2011). "Using Climatically Based Random Forests to Downscale Coarse-Grained Potential Natural Vegetation Maps in Tropical Mexico." *Applied Vegetation Science*, **14**(3), 388–401.

van den Dool HM (2007). *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, Oxford.

van den Dool HM, Saha S, Johansson A (2000). "Empirical Orthogonal Teleconnections." *Journal of Climate*, **13**(8), 1421–1435.

van der Schrier G, Briffa KR, Jones PD, Osborn TJ (2006). "Summer Moisture Variability across Europe." *Journal of Climate*, **19**(12), 2818–2834.

Wickham H, Chang W (2013). ***devtools**: Tools to Make Developing R Code Easier*. R package version 1.2, URL http://CRAN.R-project.org/package=devtools.

Wilby RL, Wigley TML (1997). "Downscaling General Circulation Model Output: A Review of Methods and Limitations." *Progress in Physical Geography*, **21**(4), 530–548.

Affiliation:

Tim Appelhans
Department of Geography
Environmental Informatics
Philipps University Marburg
Deutschhausstraße 12
35032 Marburg, Germany
E-mail: `tim.appelhans@staff.uni-marburg.de`
URL: `http://environmentalinformatics-marburg.de/`

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 65, Issue 10 *Submitted:* 2014-02-17
June 2015 *Accepted:* 2014-10-15