

# Practical exercise with R

Detection and attribution of  
climate extremes

Qiuzi Han Wen, Francis Zwiers and Xuebin Zhang

July 21-Aug 2<sup>nd</sup>, 2014

Trieste, Italy

# Research Problem

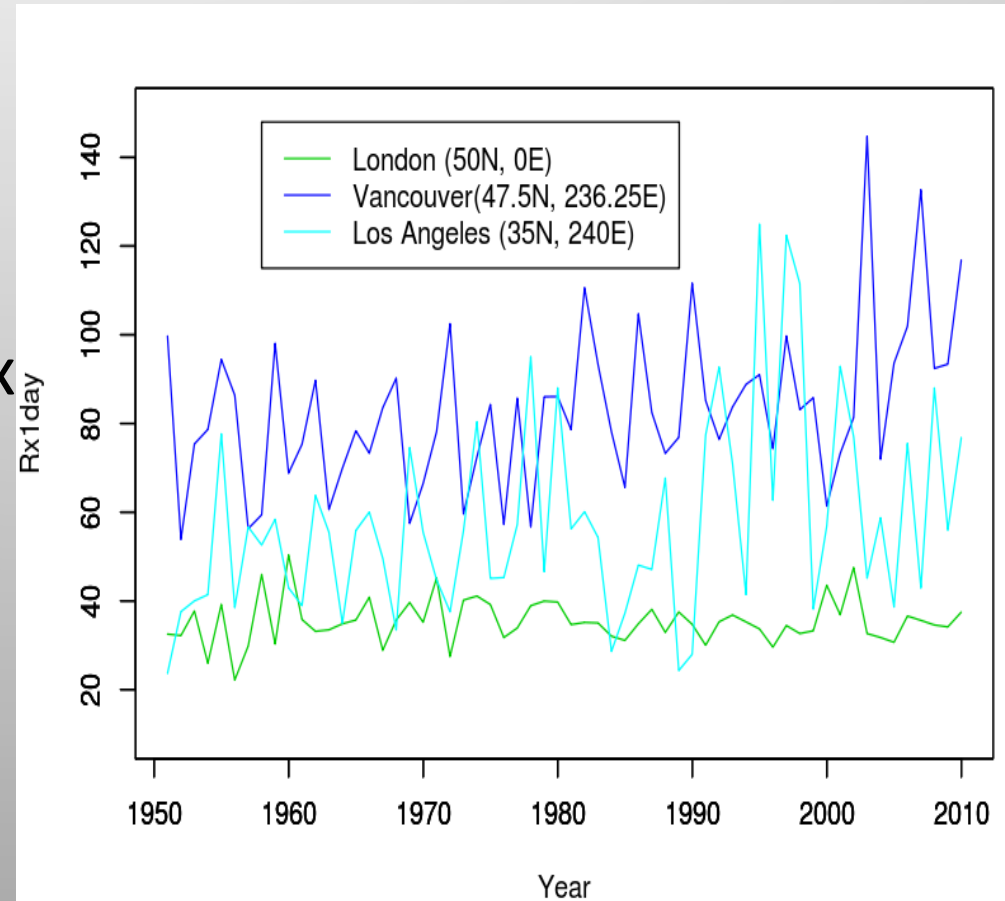
- While relative humidity is expected to remain roughly constant with warming, atmospheric moisture content is expected to increase, which in turn should result in more intense extreme precipitation.
- It is desirable to understand possible causes, especially the role of human activities, in the observed widespread intensification of precipitation extremes (Zhang et. al., GRL, 2013)

# Quantification of extreme precipitation

- ETCCDI Indices
  - Rx1day: annual maximal of daily precipitation
  - Rx5day: annual maximal of 5-day consecutive precipitation amount

# A closer look at Rx1day:

- Rx1day time series at a few sample grid points extracted from HadEX2
- Time series saved in: London\_vancouver\_LA\_Rx1day\_1901-2010.dat
- Can you reproduce this plot?
- Can we work directly with Rx1day for our detection study and why?



# Pros and cons of Rx1day/Rx5day

- Pros:
  - Clear physical interpretation
  - Easy to compute
  - Available for areas where daily precipitation records are not available (e.g., via ETCCDI workshops)
  - Amenable to “block maximum” EV analysis approach
- Cons:
  - Magnitude is highly variable from one region to another
  - Temporal variability is easily dominated by spatial variability
  - Changes in data availability with time may introduce inhomogeneity into time series of spatial averages
  - Comparison to models may be difficult because the “change of support” problem (aka, the scaling problem)
  - Limited to “block maximum” EV analysis approach
  - Lose information about the timing of extreme events, which limits possibilities for including covariates in the analysis and modelling tail dependence

# Constructing Probability Index

- Use the GEV distribution to convert annual time series of the largest one-day and five-day precipitation accumulations annually, RX1D and RX5D, into corresponding time series of PI at each grid-point.
  - The parameters for a given grid-point are estimated by fitting the GEV distribution to individual time series of observed or model-simulated annual precipitation maxima by the method of maximum likelihood.
  - Assume GEV parameters remain constant with time.
  - Each annual maximum for a given grid point and data set is converted to PI by evaluating the corresponding fitted cumulative distribution function at the value of that annual maximum.
- $PI \sim \text{Uniform}(0,1)$
- Strong annual precipitation extremes yield PI values close to 1, while weaker extremes yield PI values close to 0.

# R exercise

- Continue to work with the index time series from 3 selected grid boxes
- R programs prepared for you: Pindex.r
  - `gev.fit (ts)`
    - Fit GEV distribution via maximum likelihood
  - `Pgev (ts,  $\mu$ ,  $\sigma$ ,  $\xi$ )`
    - Calculate the corresponding fitted cumulative distribution function at a given value

# Fit GEV for individual grid point

- Please fit the GEV distribution for London Rx1day time series.
- Please convert Rx1day to PI using the fitted model:  
 $p_{gev}(Rx1day, \mu, \sigma, \xi)$

London

```
> fit1=gev.fit(xld[51:t0])
$conv
[1] 0
$nlh
[1] 181.0803
$mle
[1] 33.4468836 4.7336193 -0.1782475
```

Confidence interval of  
GEV parameters is also  
provided:

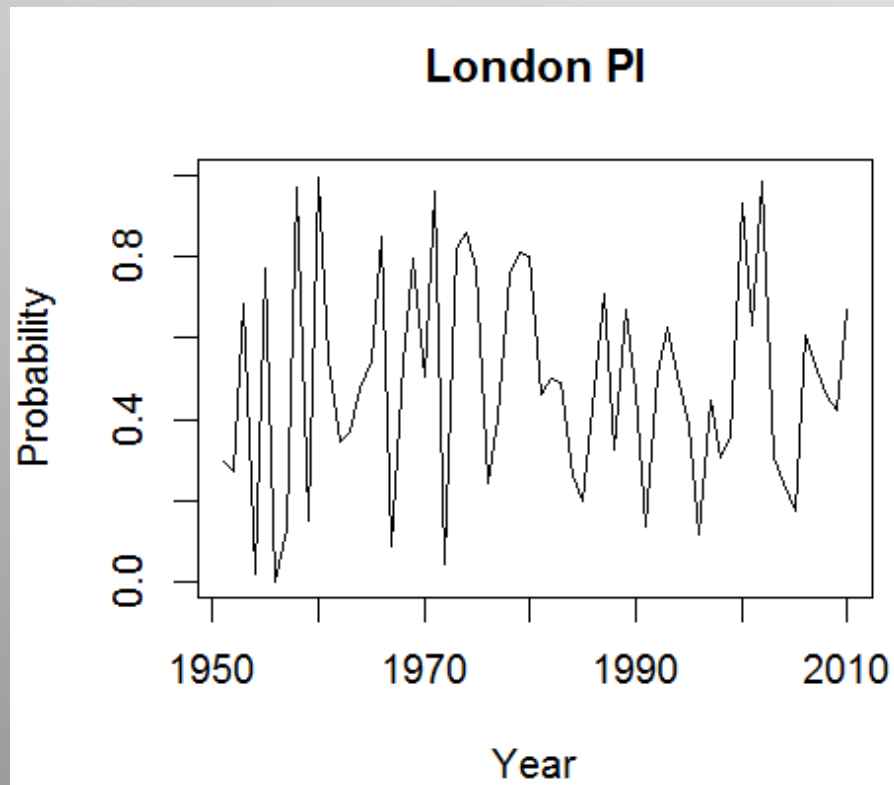
```
$CI90
      [,1]      [,2]
[1,] 33.3039938 33.5897733
[2,] 4.6388450 4.8283935
[3,] -0.1917315 -0.1647635
```



# Probability index: London

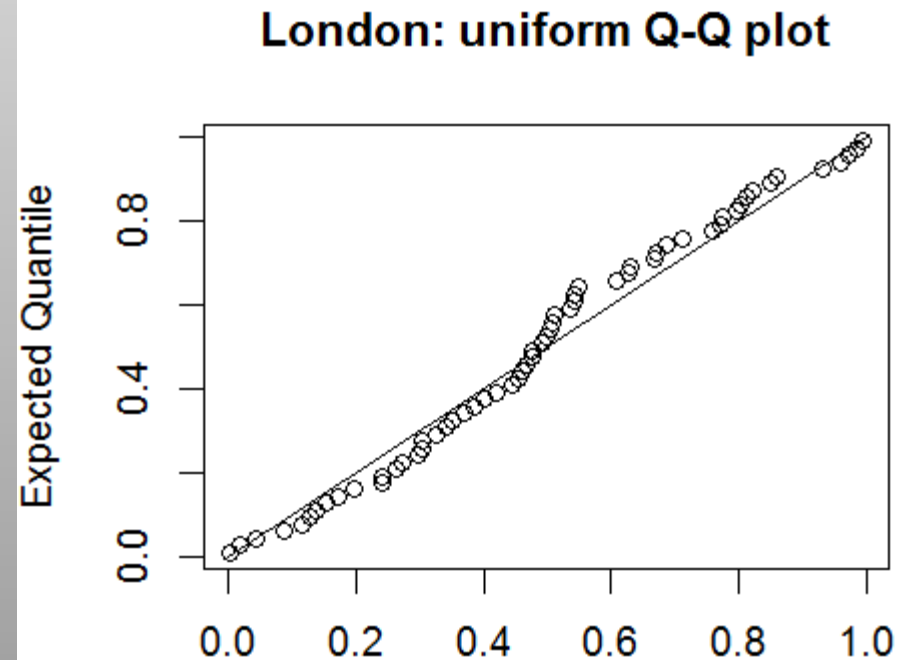
## Goodness of fit

### Fitted series



One-sample Kolmogorov-Smirnov test

```
data: pil
D = 0.0996, p-value = 0.5579
alternative hypothesis: two-sided
```



# Suggested activities

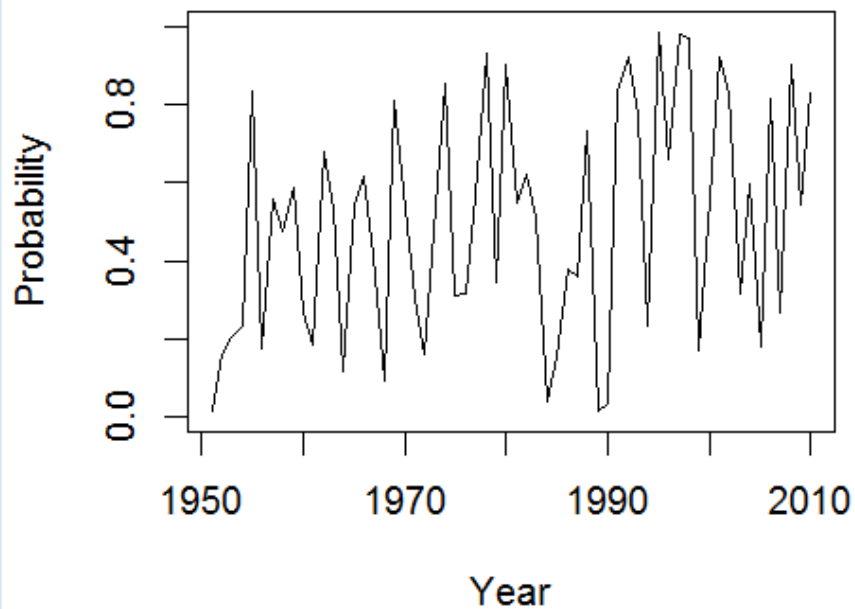
- Can you calculate probability index for Vancouver and Los Angeles, respectively?
- Explore the goodness-of-fit of the GEV model
- How about if we change the study period, e.g., to 1961-2010?
- Refer to Day2\_main\_V2.r in Pindex/ for reference command lines...

# Los Angeles

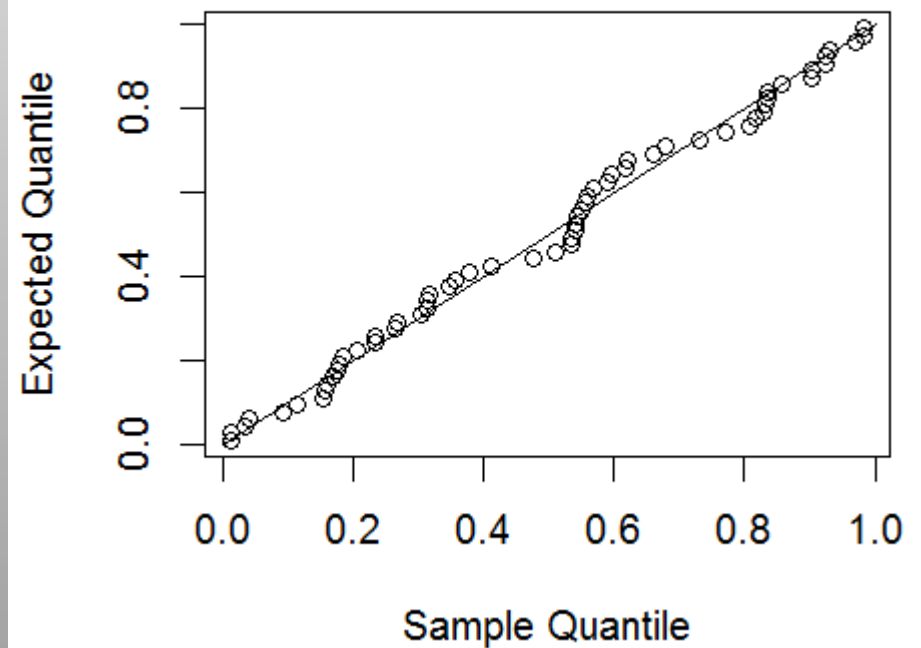
One-sample Kolmogorov-Smirnov test

```
data: pi3  
D = 0.0679, p-value = 0.9269  
alternative hypothesis: two-sided
```

Los Angeles PI



LA: uniform Q-Q plot



# Vancouver

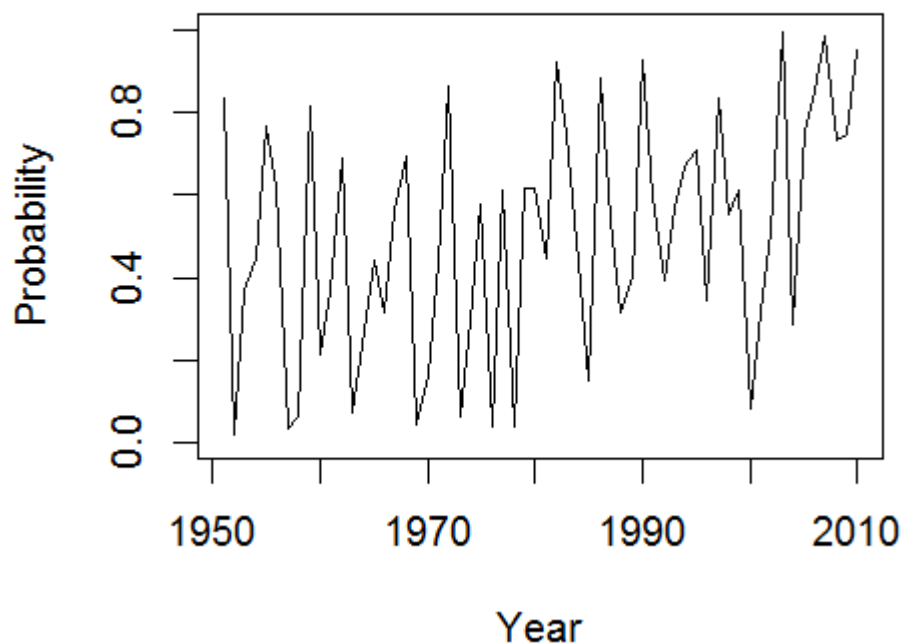
One-sample Kolmogorov-Smirnov test

data: pi5

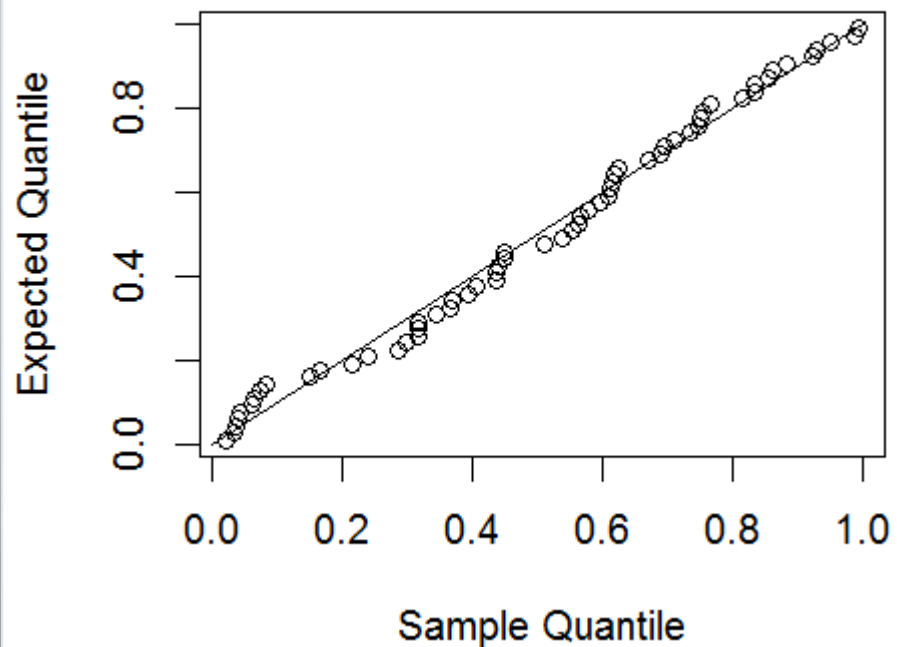
D = 0.0678, p-value = 0.9277

alternative hypothesis: two-sided

**Vancouver PI**

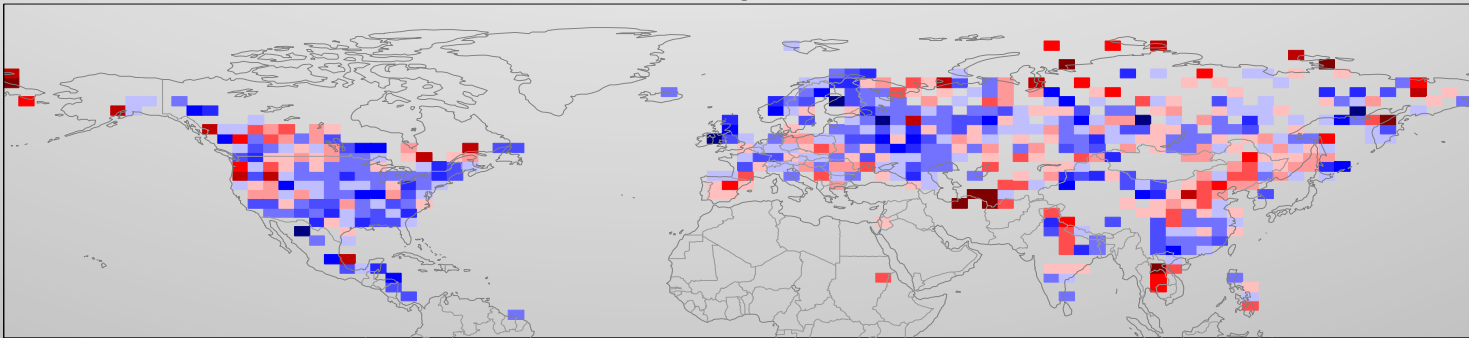


**Vancouver: uniform Q-Q plot**

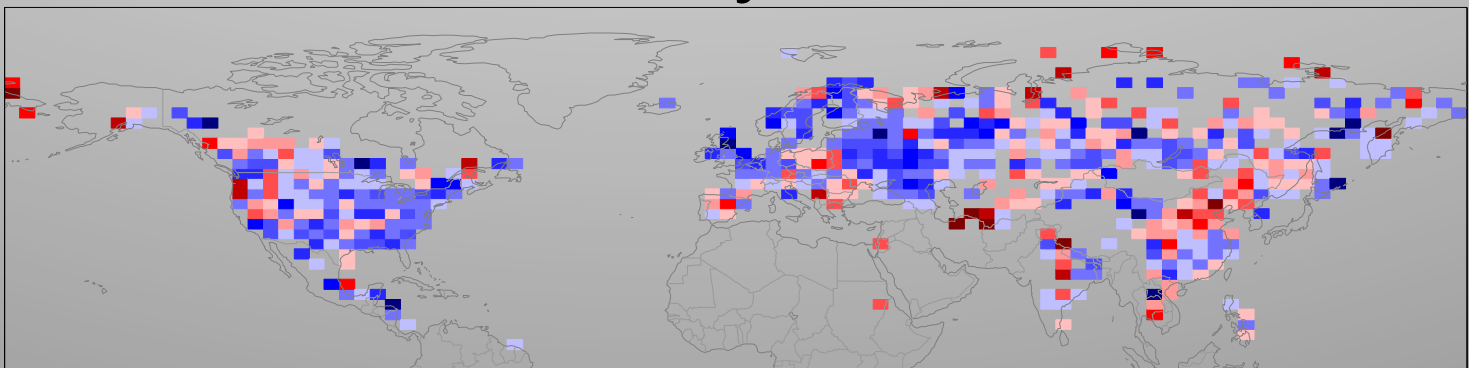


Linear trend in transformed indices of observed annual precipitation extremes 1951-2005

RX1day, OBS



RX5day, OBS



# Detection exercise

- Attempt to detect “ALL” signal in PI of extreme precipitation derived from Rx1day
- A 8-step procedure

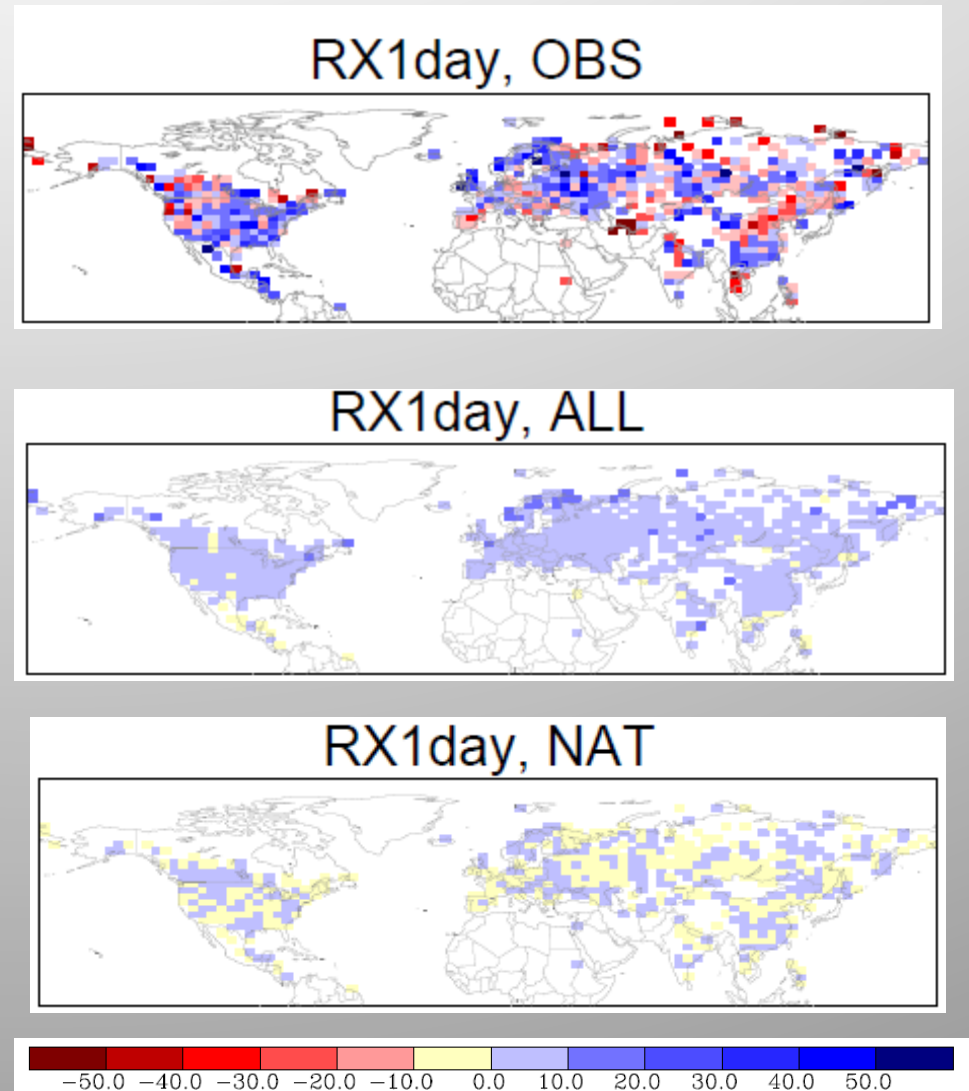


Figure S5, Zhang et al., 2013

# Step 1: space-time scale of interest

- 1951~2005
- Northern hemisphere land area

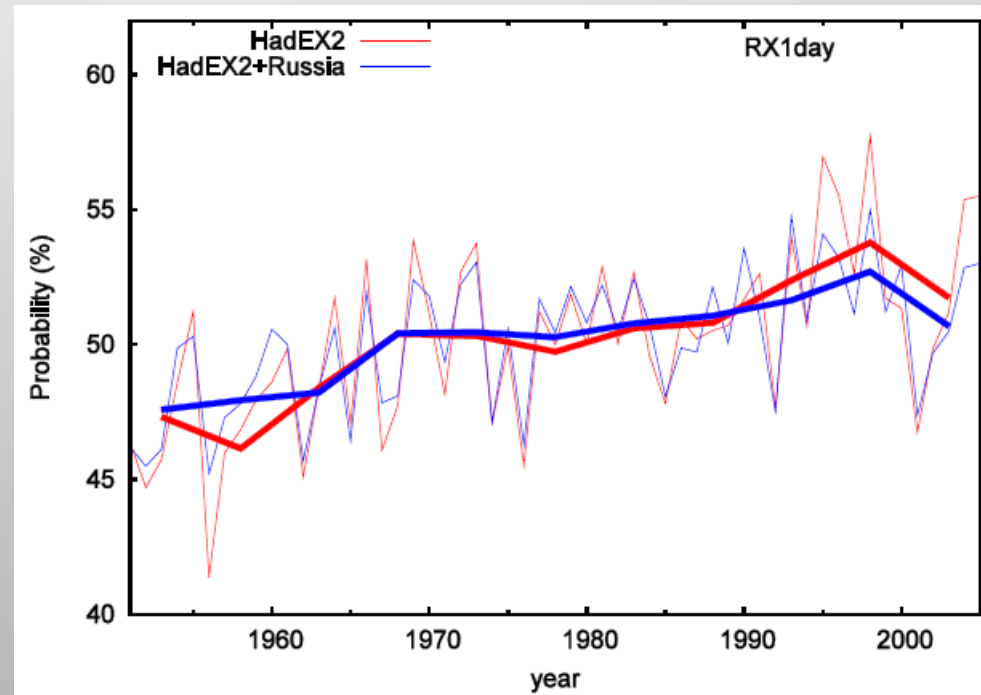


Figure S2, Zhang et al., 2013

# Step 1: filtering

- Temporal: 5-year mean
- Spatial: 3 different spatial configurations (1,2 or 3 sub-regions)
  - 1-region:
    - northern hemisphere land area mean (NH)
  - 2 broad zonal NH regions:
    - mid-latitudes (30°N~65°N, ML)
    - tropics and subtropics (0°N~30°N, TR)
  - 3 NH west-east regions:
    - western NH (50°W~180°W, NA)
    - western Eurasia (15°W~60°E, EU)
    - eastern Eurasia (60°E~180°E, AS)

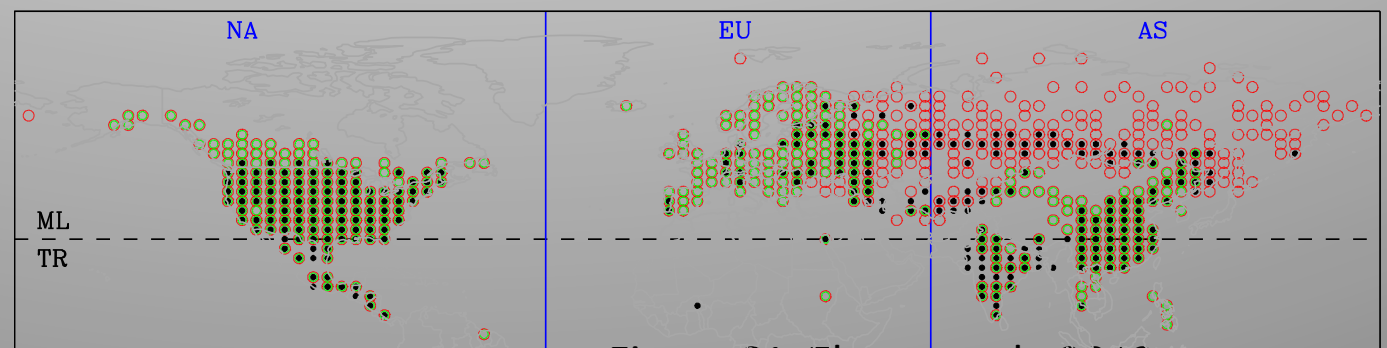


Figure S1, Zhang et al., 2013



## Step 2: gather data (OBS)

- HadEx2: a gridded ( $2.5^\circ \times 3.75^\circ$  latitude-longitude) land-based dataset of indices of temperature and precipitation extremes [Donat et al., 2013]
- +600 Russian stations

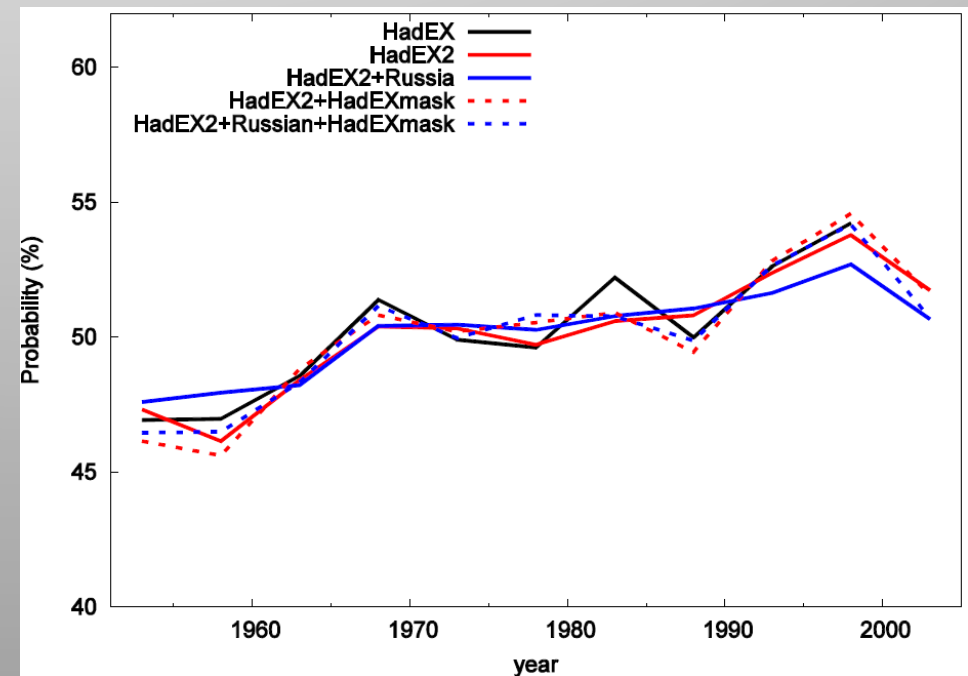


Figure S3, Zhang et al., 2013

## Step 2: gather data (model simulations)

- CMIP5
  - ALL: 54 runs from 14-MME
  - NAT: 34 runs from 9-MME
  - Unforced control runs: > 15,000 yrs from 31-MME

# Step 3: process data

- Observations
  - Merge HadEX2 and Russian in-situ data
  - Convert to PI
- Model simulations
  - Interpolate to the same spatial resolution, e.g.,  $2.5^{\circ} \times 3.75^{\circ}$
  - Masked by availability of observations
  - Convert to PI

# Data and codes can be found at:

## DA\_Rx1day

- Rx1D\_5yrPI\_\*area\*\_All.dat
  - Two rows of 5-yr regional mean PI anomaly
    - 11 observed PI, subtracting mean value 0.5
    - 11 Multi-model ensemble mean PI anomaly
      - » averaged across 54 ALL-forcings runs
- Noise1\_Rx1D\_5yrPI\_\*area\*.dat
  - Used to estimate variability from internal sources
  - 230 rows, 11 values each
    - 1 row for each 55-yr chunk obtained from control run simulations
      - » masked by and processed as observations
- Noise2\_Rx1D\_5yrPI\_\*area\*.dat (as above)

# Step 4-8

4. Optimization
5. Fit regression model
6. Determine EOF truncation
7. Iterate 5-7
8. Make inferences about scaling factor(s)

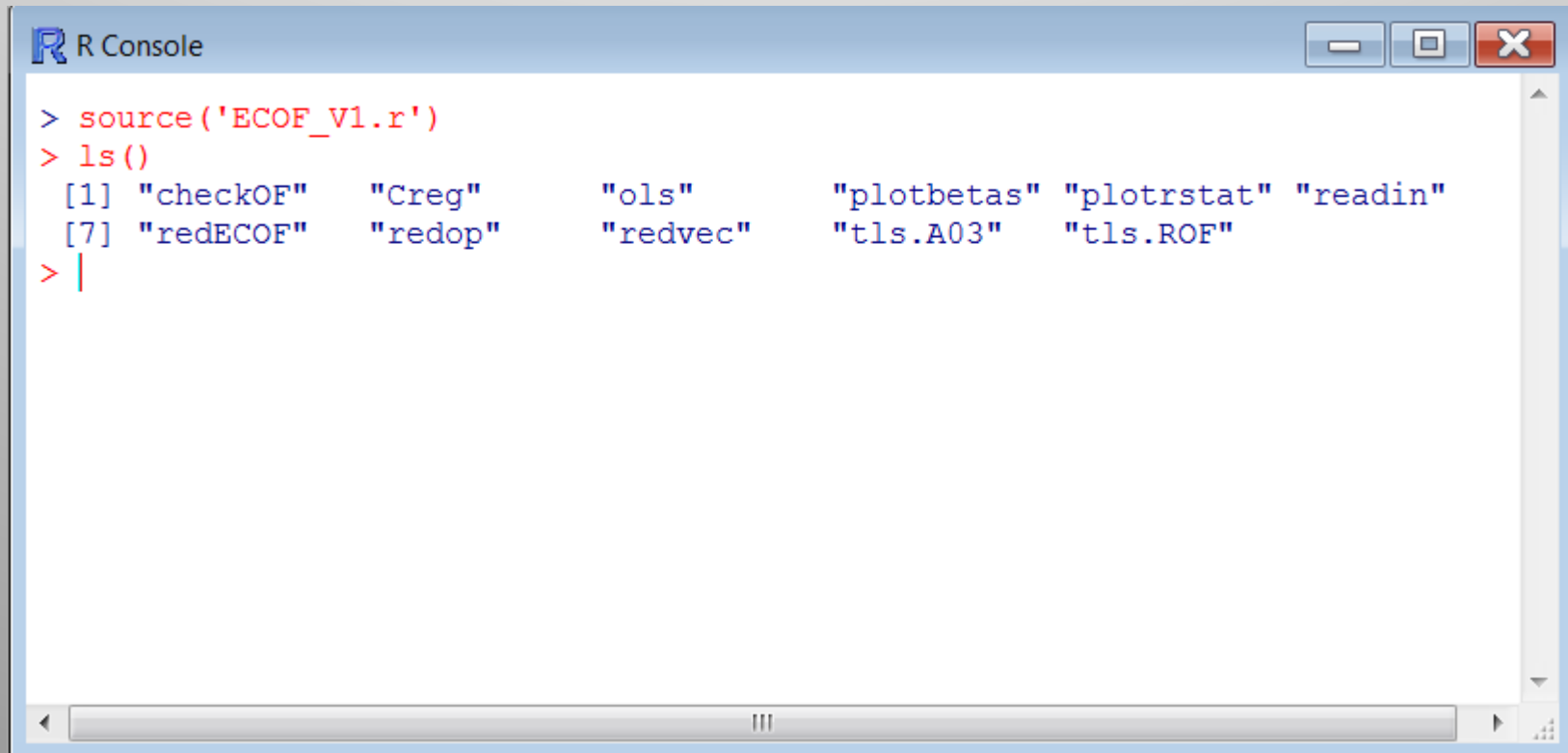
These steps have all been coded for you in R

- 6 functions in EC\_OF.r
  - readin.r – ingests data from step 3
  - ols – carries out detection analysis using ordinary least squares
  - tls.A03- carries out detection analysis using total least squares algorithm
  - tls.ROF-carries out detection analysis using regularized optimal fingerprint
  - plotbetas-visualization of scaling factor estimates
  - plotrstat-visualization of results for residual consistency check

# Suggested activities:

Load functions into R

- Click on “File”
- Click on “**S**ource R Code ...”
- Enter the function name to list the function
- **s**ource() can also be used to load code

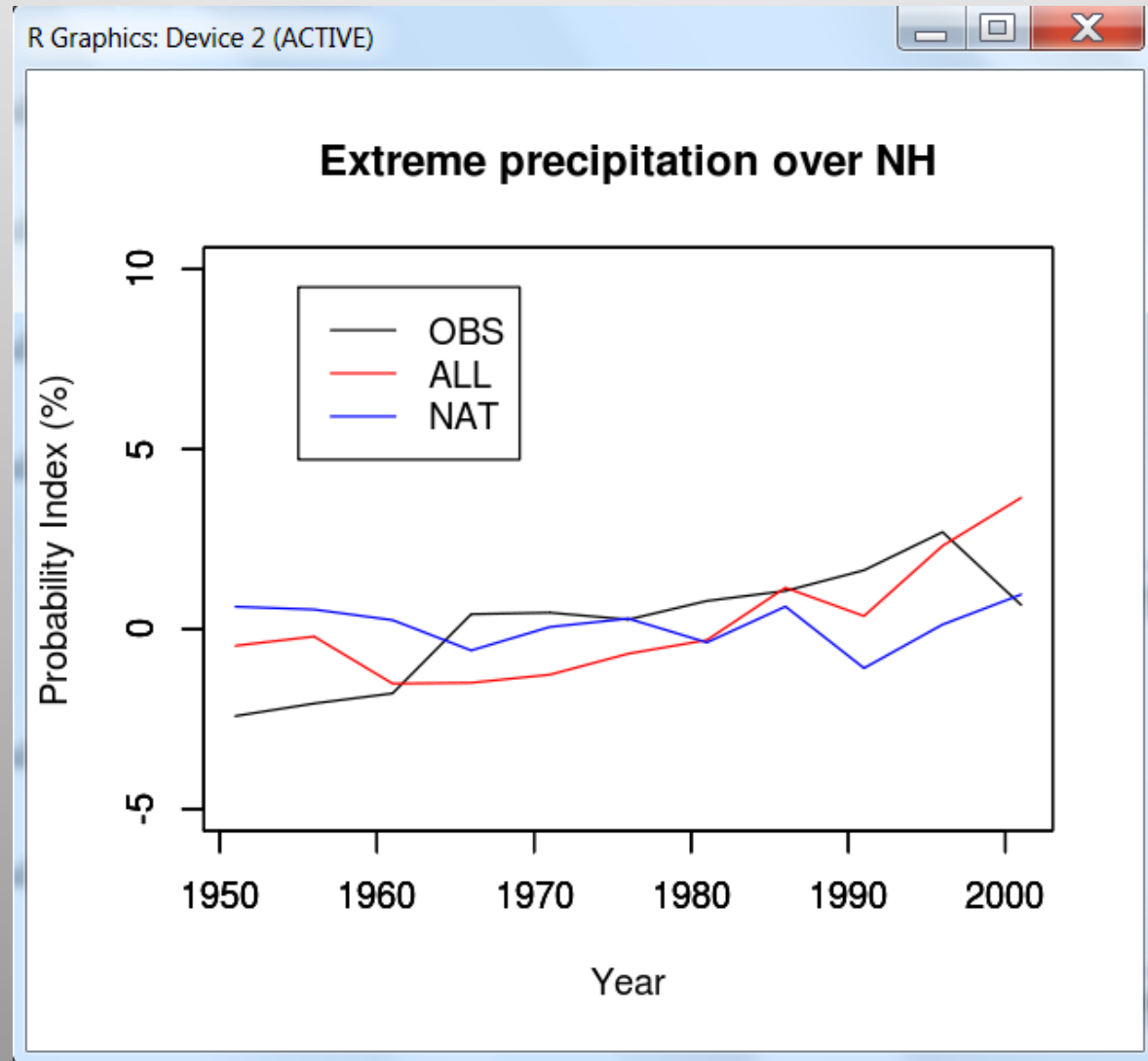


```
R Console
> source('ECOF_V1.r')
> ls()
[1] "checkOF"    "Creg"       "ols"        "plotbetas"  "plotrstat"  "readin"
[7] "redECOF"    "redop"      "redvec"     "tls.A03"    "tls.ROF"
```

# Suggested activities:

- Use “readin” to get the data into R
  - Results are stored in class object Z:
    - Z@X (signal)
    - Z@Y (observation)
    - Z@noise1
    - Z@noise2
  - Have a look at these variables
  - Plot observation and signal versus time

# Preliminary analysis: visualization of obs and signals





# Recall the detection methods:

$$\text{OLS: } \mathbf{Y} = \sum_{i=1}^S \beta_i \mathbf{X}_i + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\mathbf{Y}$  → Observations

$\mathbf{X}$  → Expected changes – one vector for each “signal”

$\boldsymbol{\beta}$  → Regression coefficients – aka “scaling factors”

$\boldsymbol{\varepsilon}$  → Residuals – internal variability

Idea is to interpret the observations with a regression model, where physics is used to provide representations of expected changes due to external influences, statistics is used to demonstrate a good fit, and physics is used to interpret the fit and rule out other putative explanations

Key statistical questions relate to the  $\beta_i$ 's and residuals  $\boldsymbol{\varepsilon}$

## Fitting the more complicated TLS model:

$$\begin{aligned}\mathbf{Y} &= \mathbf{Y}^{Forced} + \boldsymbol{\varepsilon} \\ \tilde{\mathbf{X}} &= \mathbf{X}^{Forced} + \boldsymbol{\Delta} \\ \mathbf{Y}^{Forced} &= \mathbf{X}^{Forced} \boldsymbol{\beta}\end{aligned}$$

Fitting involves finding the  $\mathbf{X}^{Forced}$  and  $\boldsymbol{\beta}$  that minimize the “size” of the  $n \times (s+1)$  matrix of residuals  $[\boldsymbol{\Delta}, \boldsymbol{\varepsilon}]$

The assumptions about the covariance structure determine how the “size” of the matrix of residuals is measured

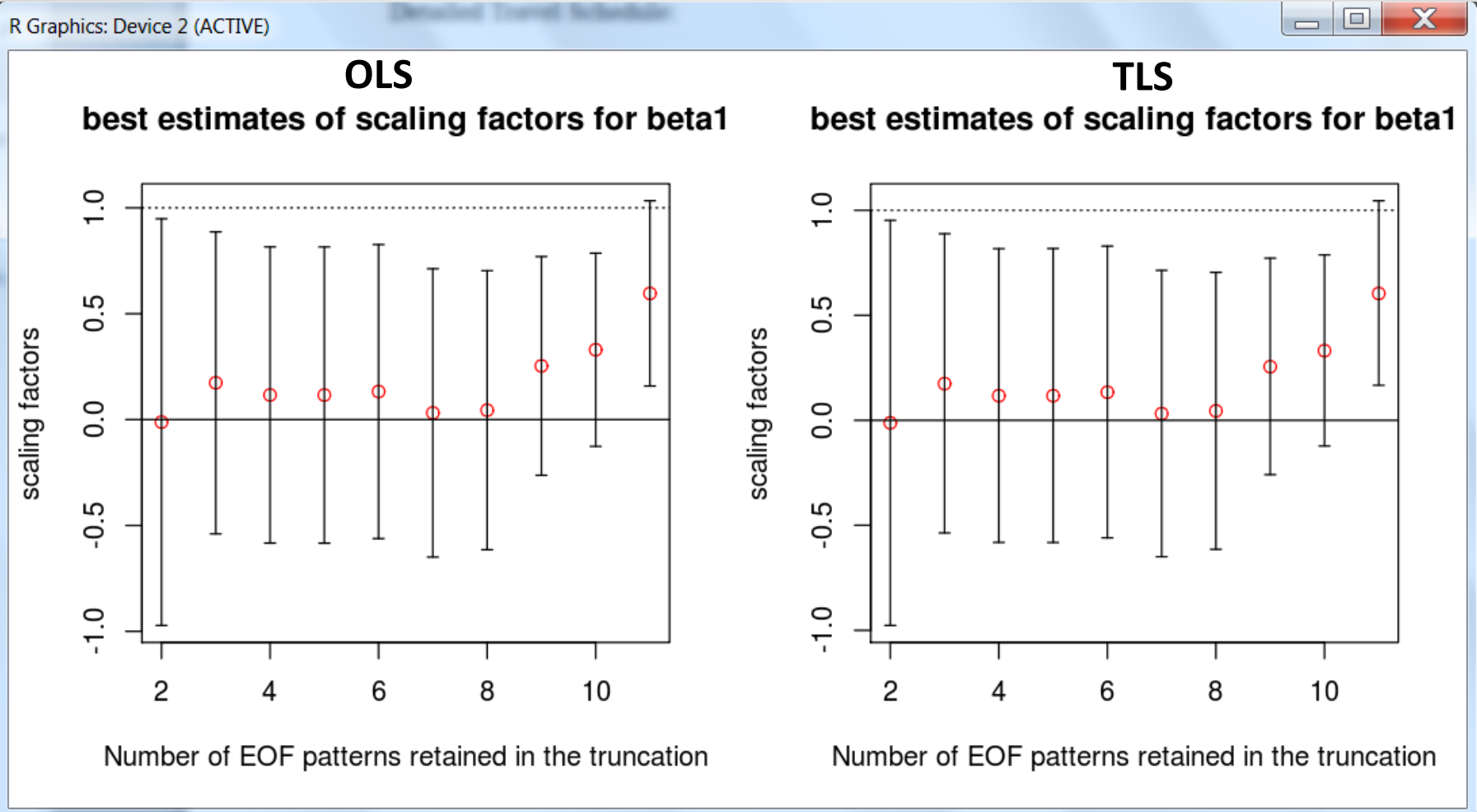
Note that because we scaled  $\tilde{\mathbf{X}}$ , the estimate of  $\mathbf{X}^{Forced}$  will be too large by a factor of  $\mathbf{M}$ , which means that we will have to adjust the estimated  $\mathbf{X}^{Forced}$  and  $\boldsymbol{\beta}$  to compensate

## Suggested activities:

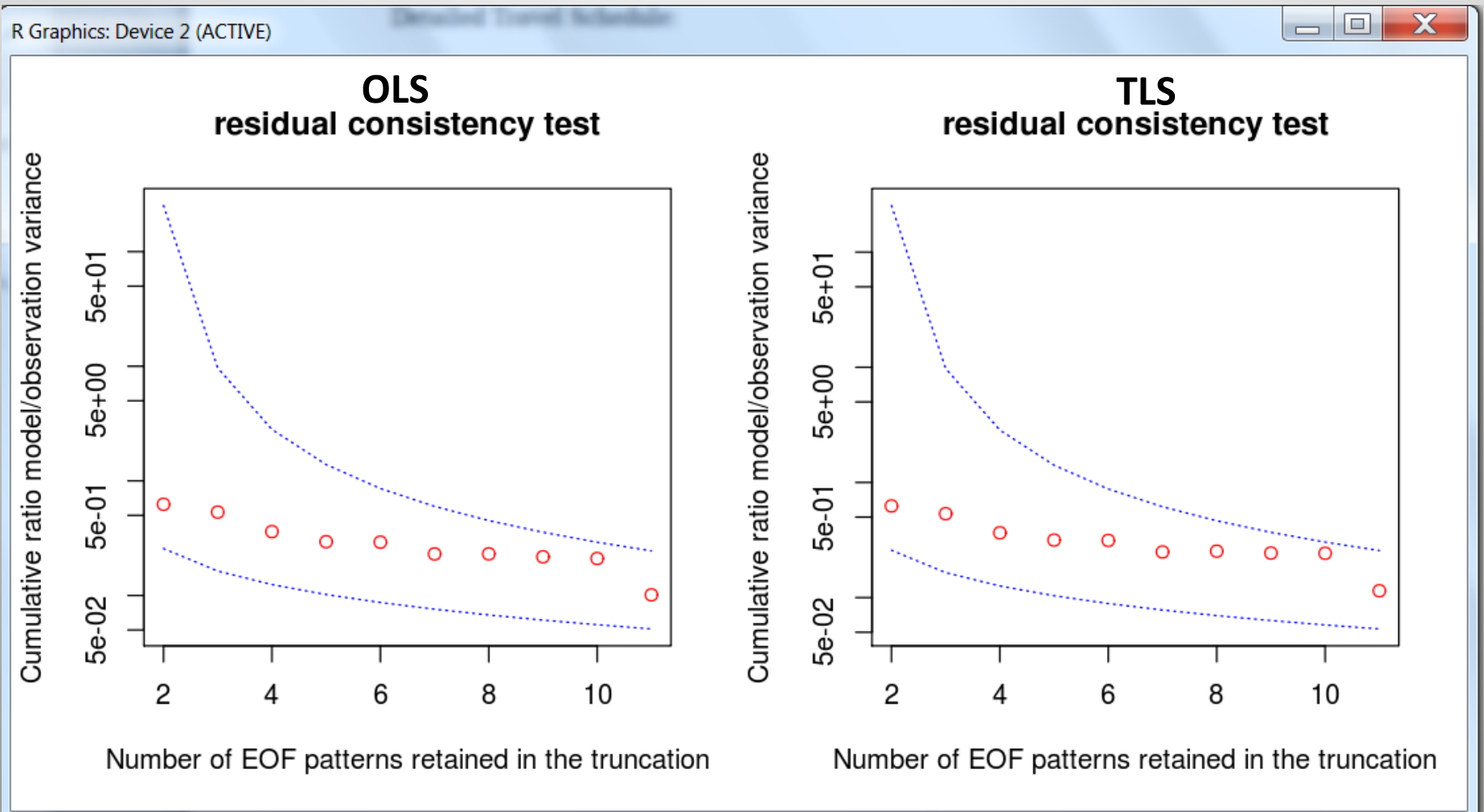
### Detection analysis using ALL signal

1. Perform the analysis over NH (1 large region spatial scale), over ML+TR (2-region spatial scale) and over NA+EU+AS (3-region spatial scale)
2. Do we need EOF truncations?
3. Should we use OLS or TLS?
4. How to interpret the results?

# Comparing results from OLS and TLS



# Results of RCC



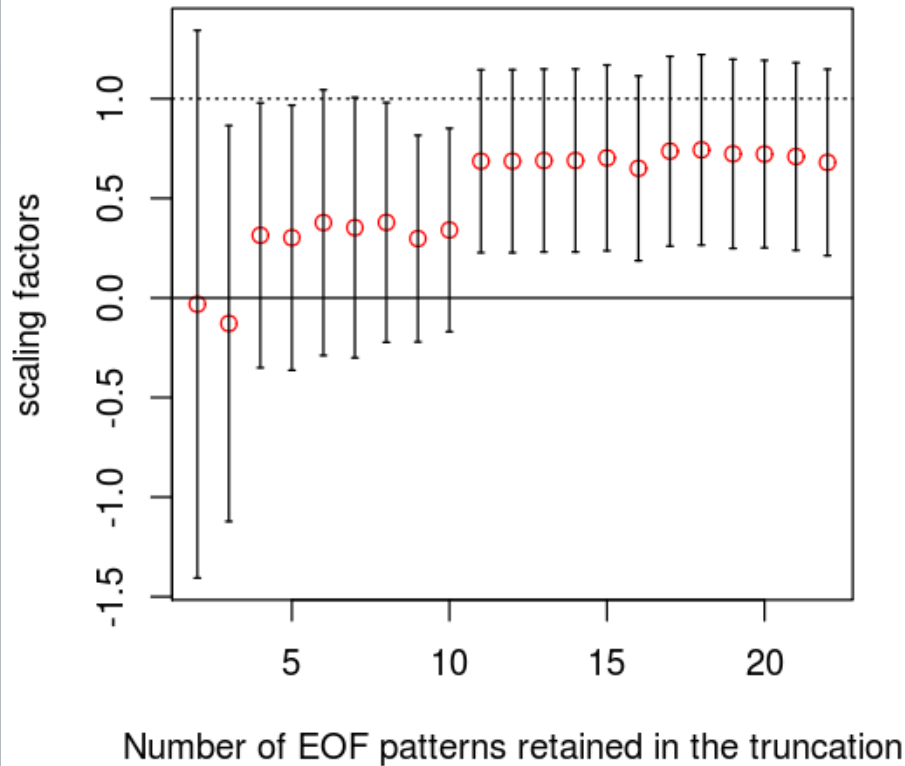
# ML+TR

R Graphics: Device 2 (ACTIVE)



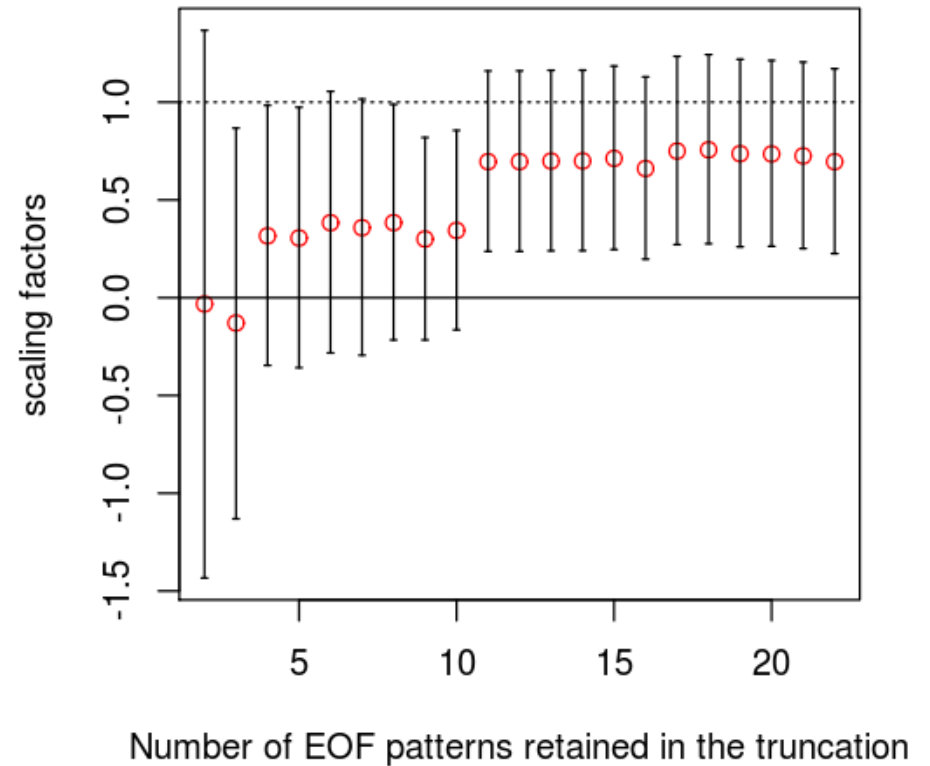
## OLS

### best estimates of scaling factors for beta1



## TLS

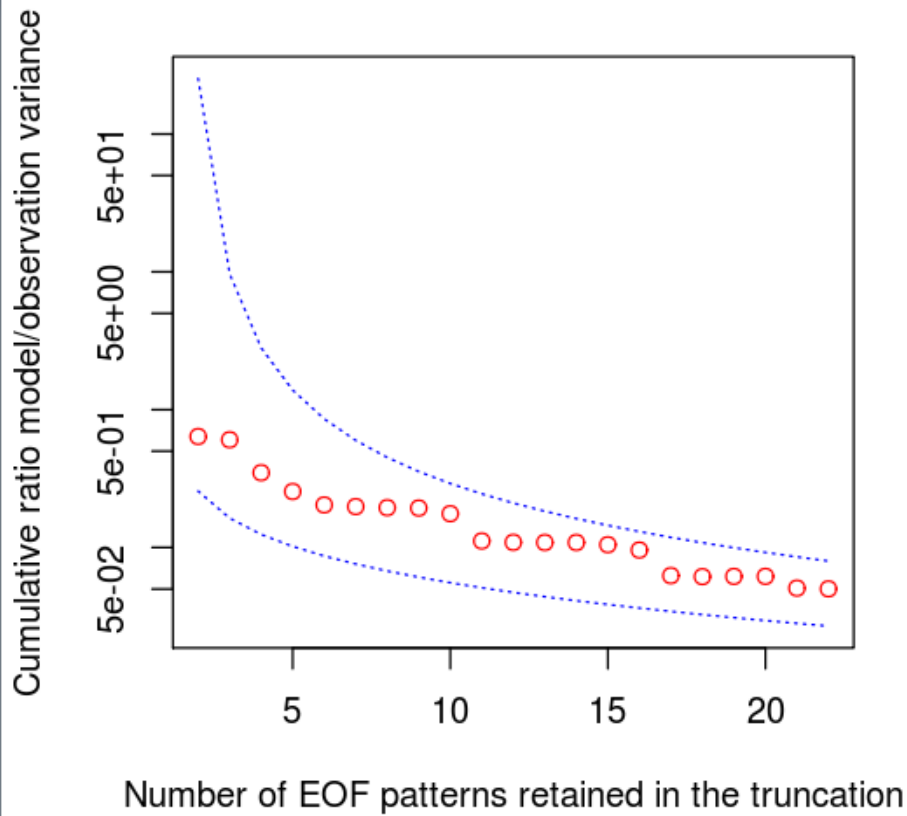
### best estimates of scaling factors for beta1



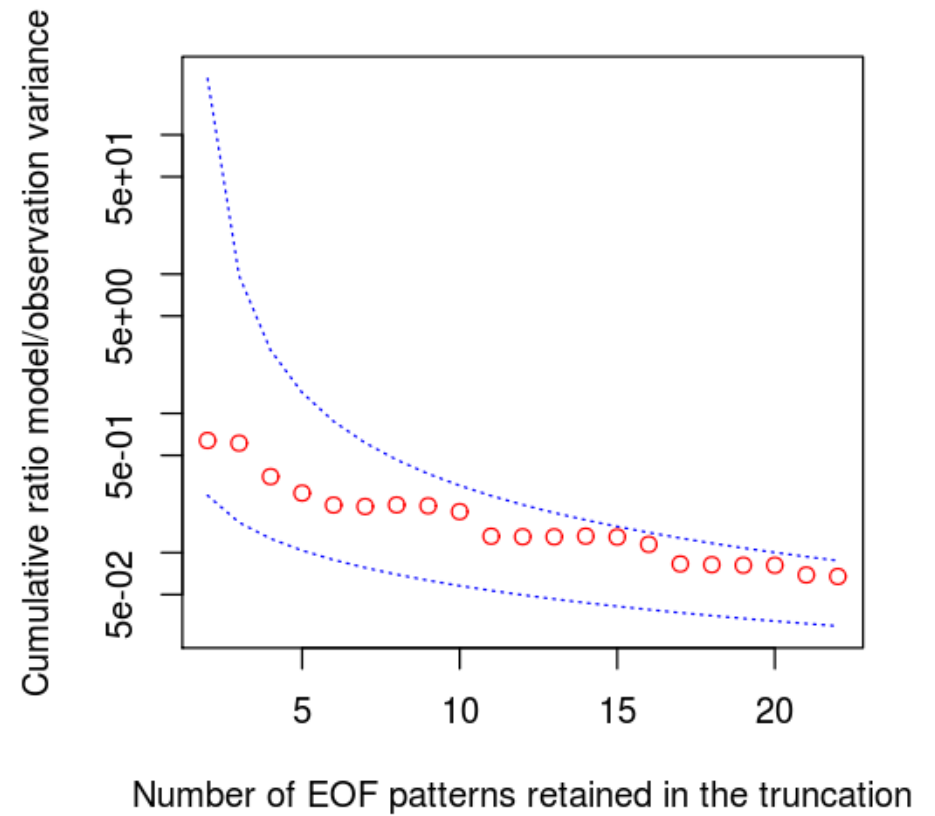
# ML+TR

R Graphics: Device 2 (ACTIVE)

**OLS**  
residual consistency test



**TLS**  
residual consistency test



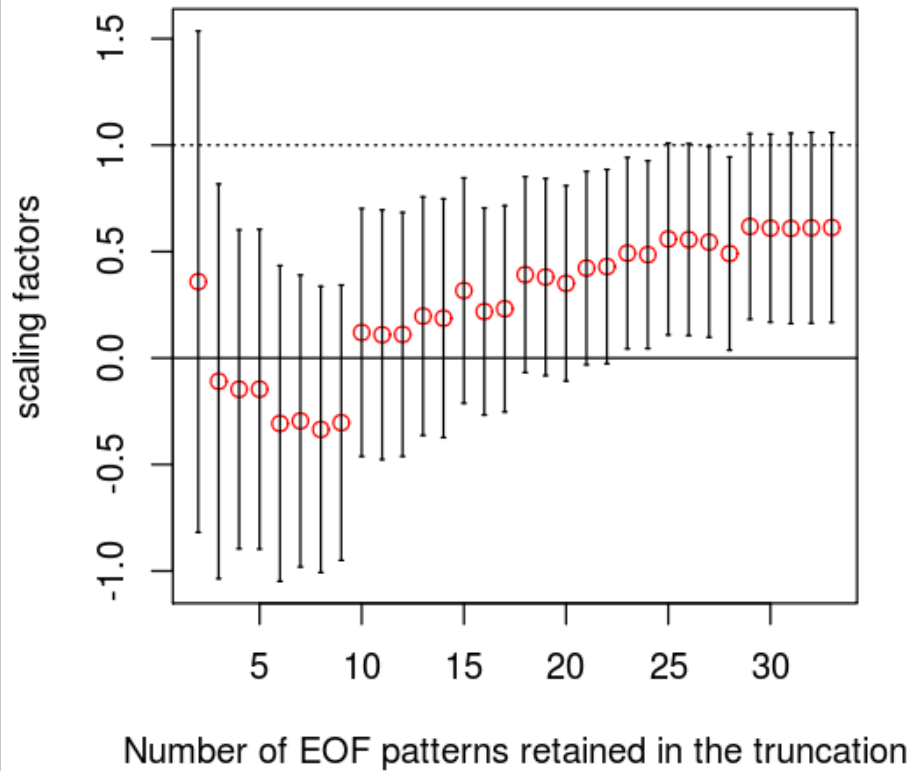
# NA+EU+AS

R Graphics: Device 2 (ACTIVE)



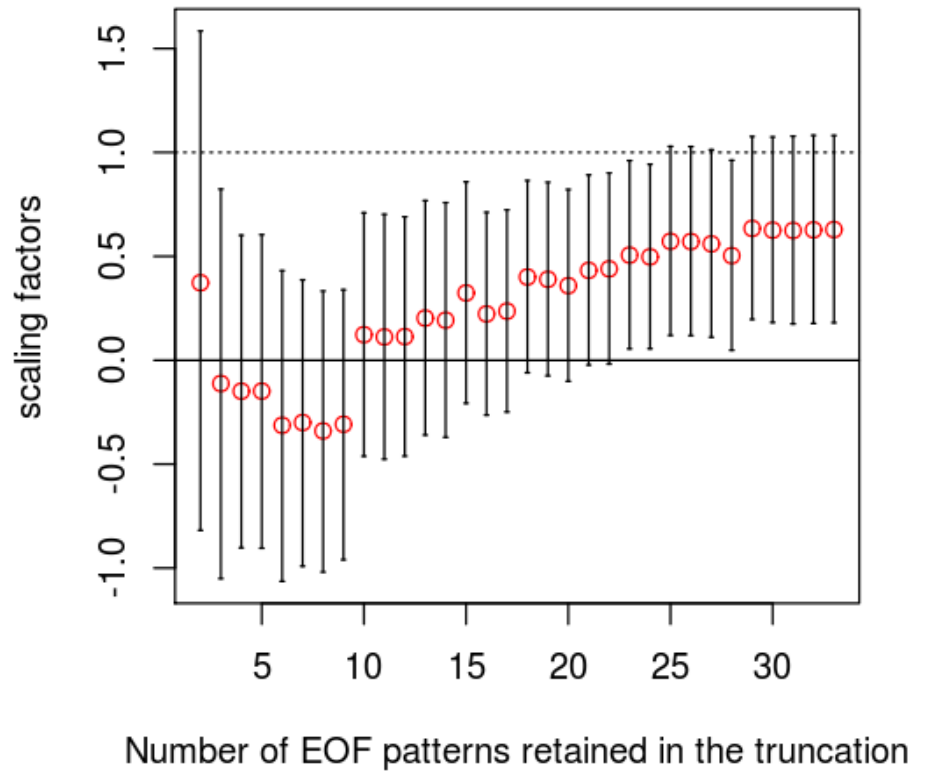
## OLS

### best estimates of scaling factors for beta1



## TLS

### best estimates of scaling factors for beta1

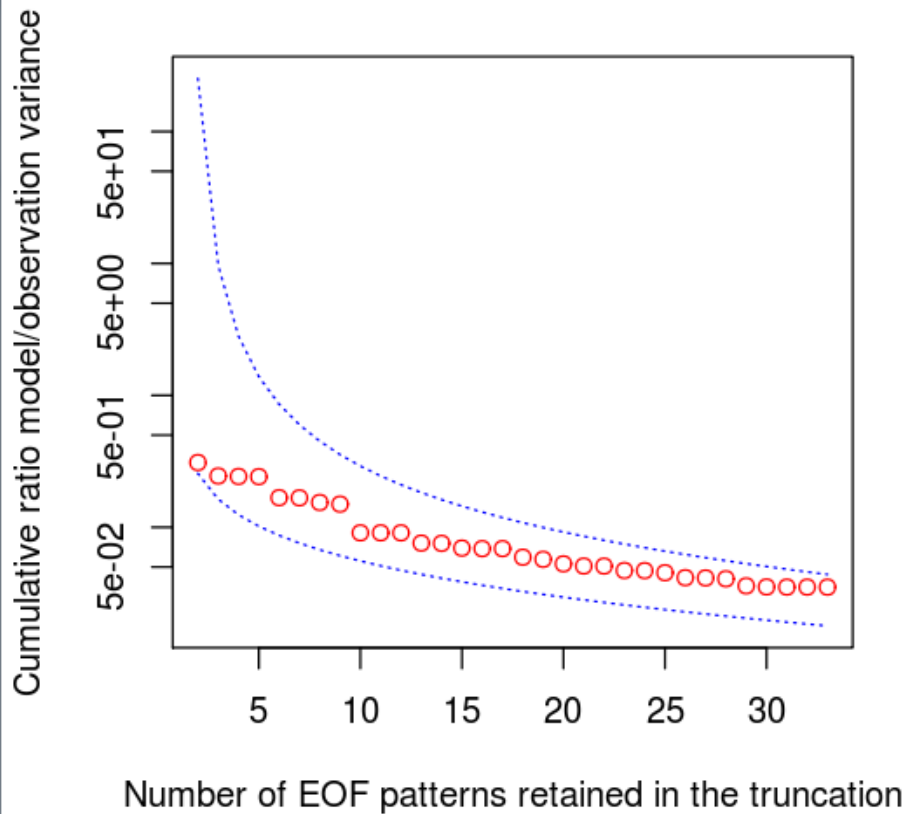




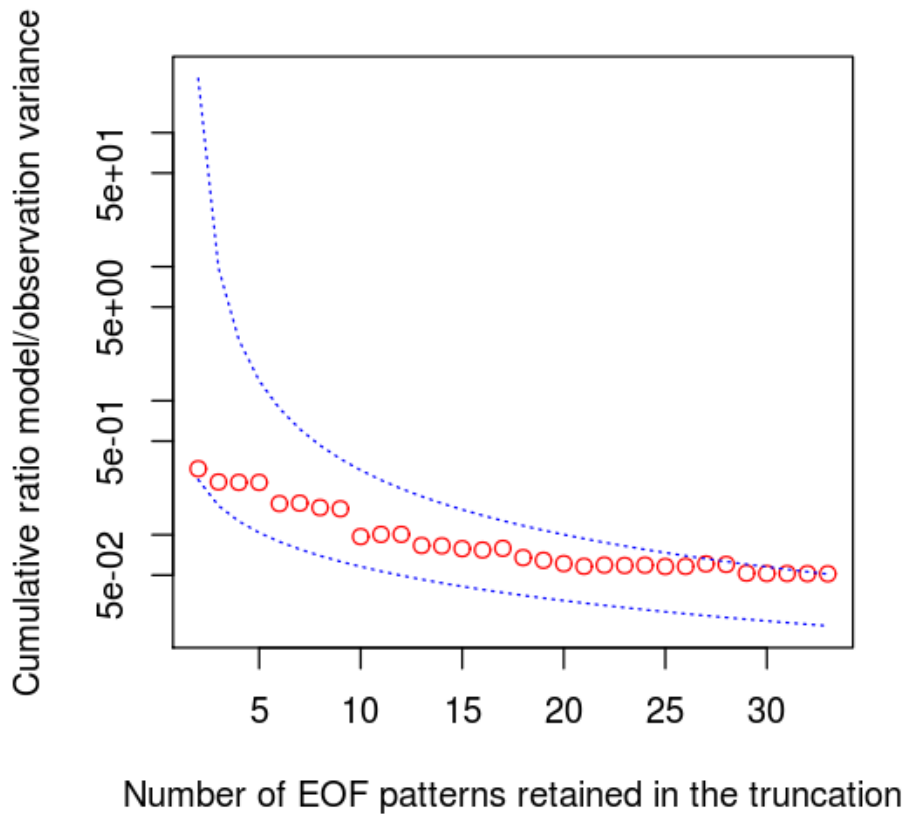
# NA+EU+AS

R Graphics: Device 2 (ACTIVE)

**OLS**  
residual consistency test



**TLS**  
residual consistency test



## Suggested activities:

### Detection analysis using NAT signal

1. Perform the analysis over NH (1 large region spatial scale) and over NA+EU+AS (3-region spatial scale)
2. How to interpret the results?

# How about NAT signal?

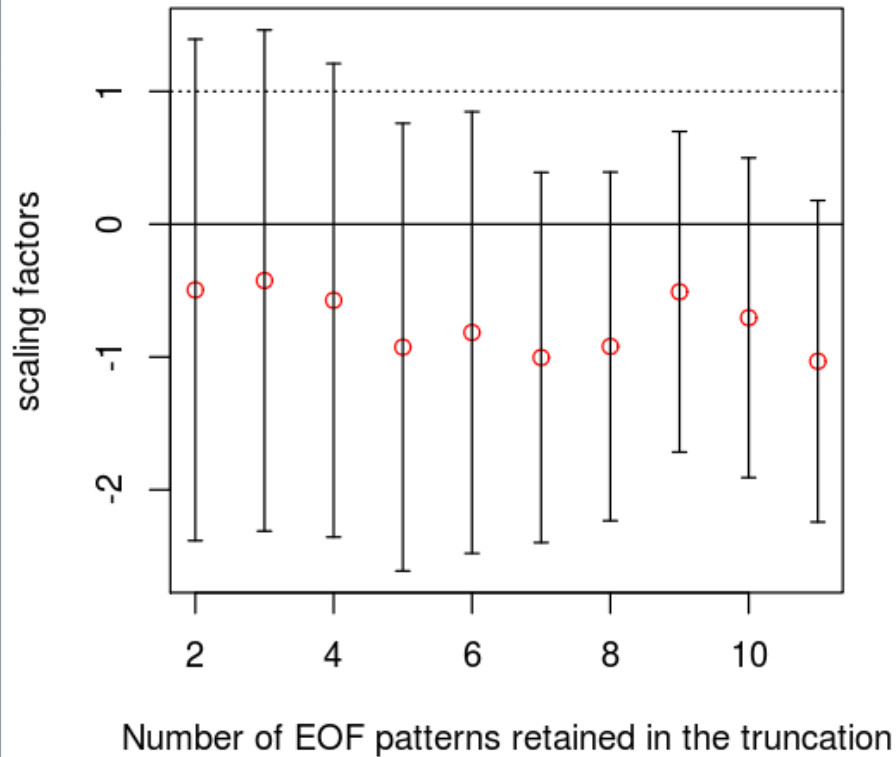
R Graphics: Device 2 (ACTIVE)

Device 2 (ACTIVE)



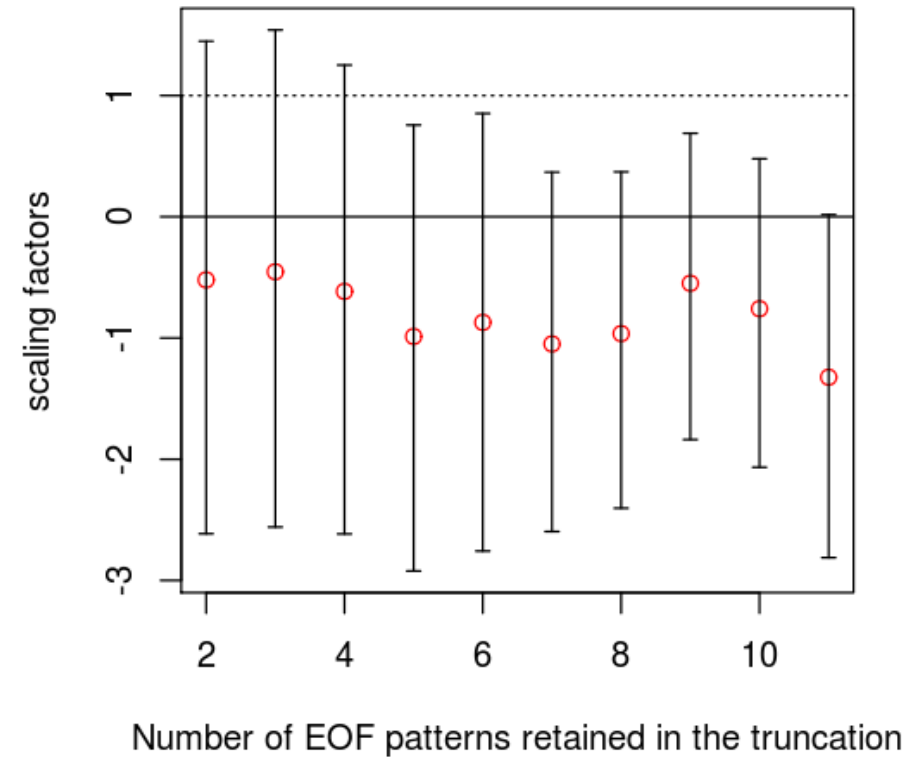
## OLS

### best estimates of scaling factors for beta1



## TLS

### best estimates of scaling factors for beta1

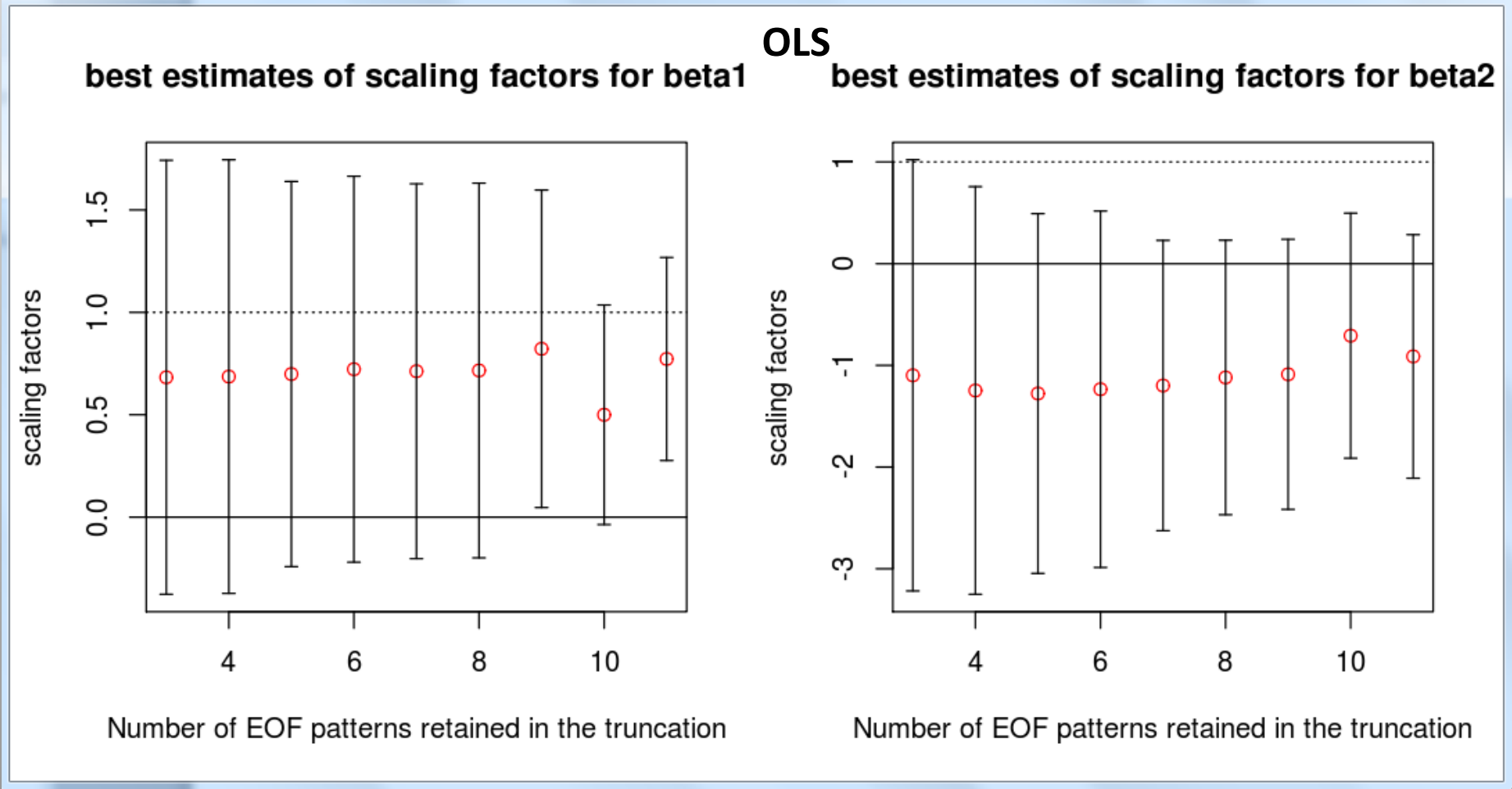


# In depth exercise

- Try multiple-signal analysis
  - Can we isolate in observations the response to the ANT signal by using the ALL and NAT signals?
  - Is a multiple-signal analysis preferable?
- Open discussion: Q&A

# 2-signal analysis using ANT+NAT (NH)

R Graphics: Device 2 (ACTIVE)



# 2-signal analysis using ANT+NAT (NH)

R Graphics: Device 2 (ACTIVE)

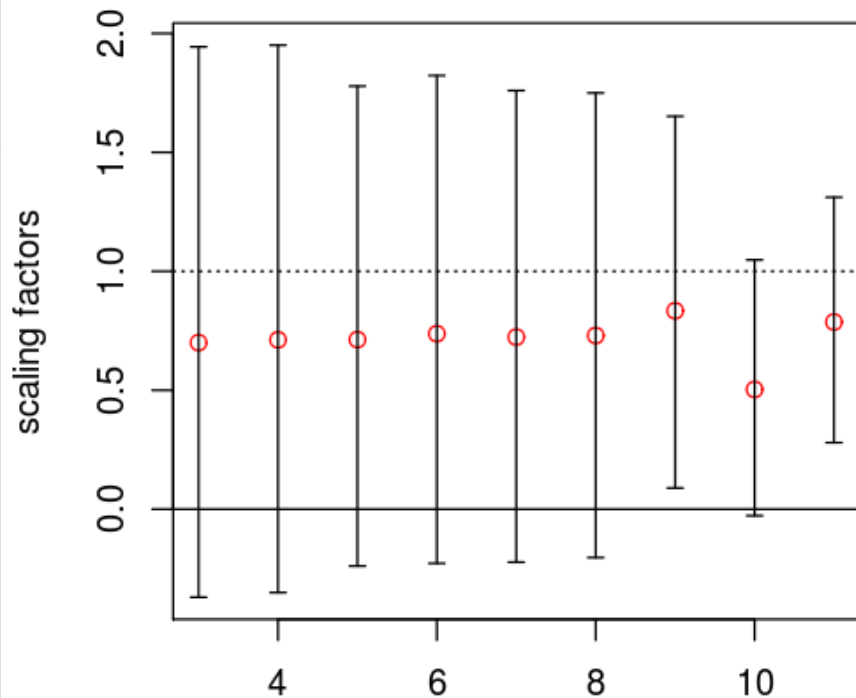
Delayed Truncated SVD



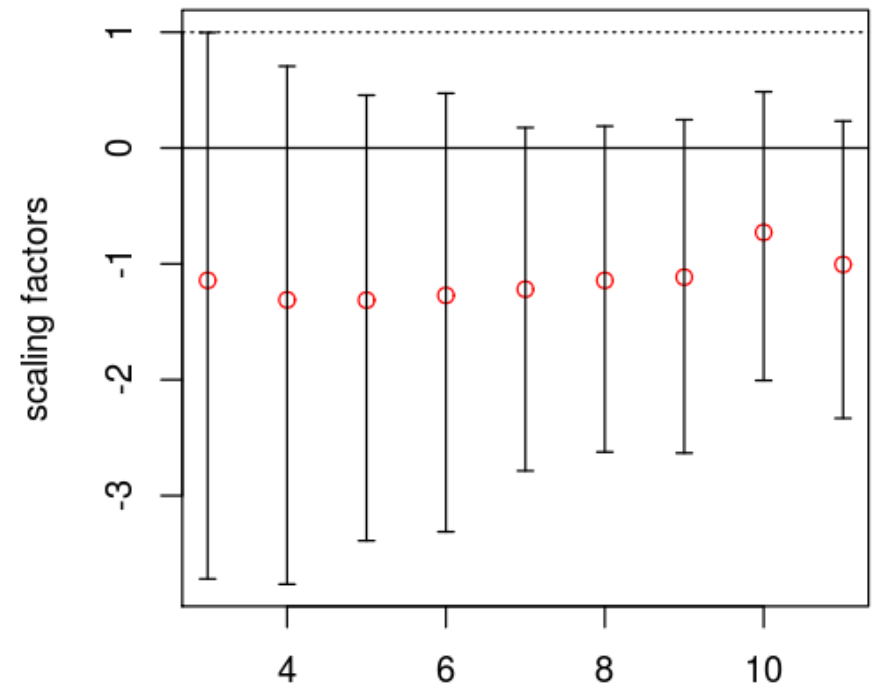
TLS

best estimates of scaling factors for beta1

best estimates of scaling factors for beta2

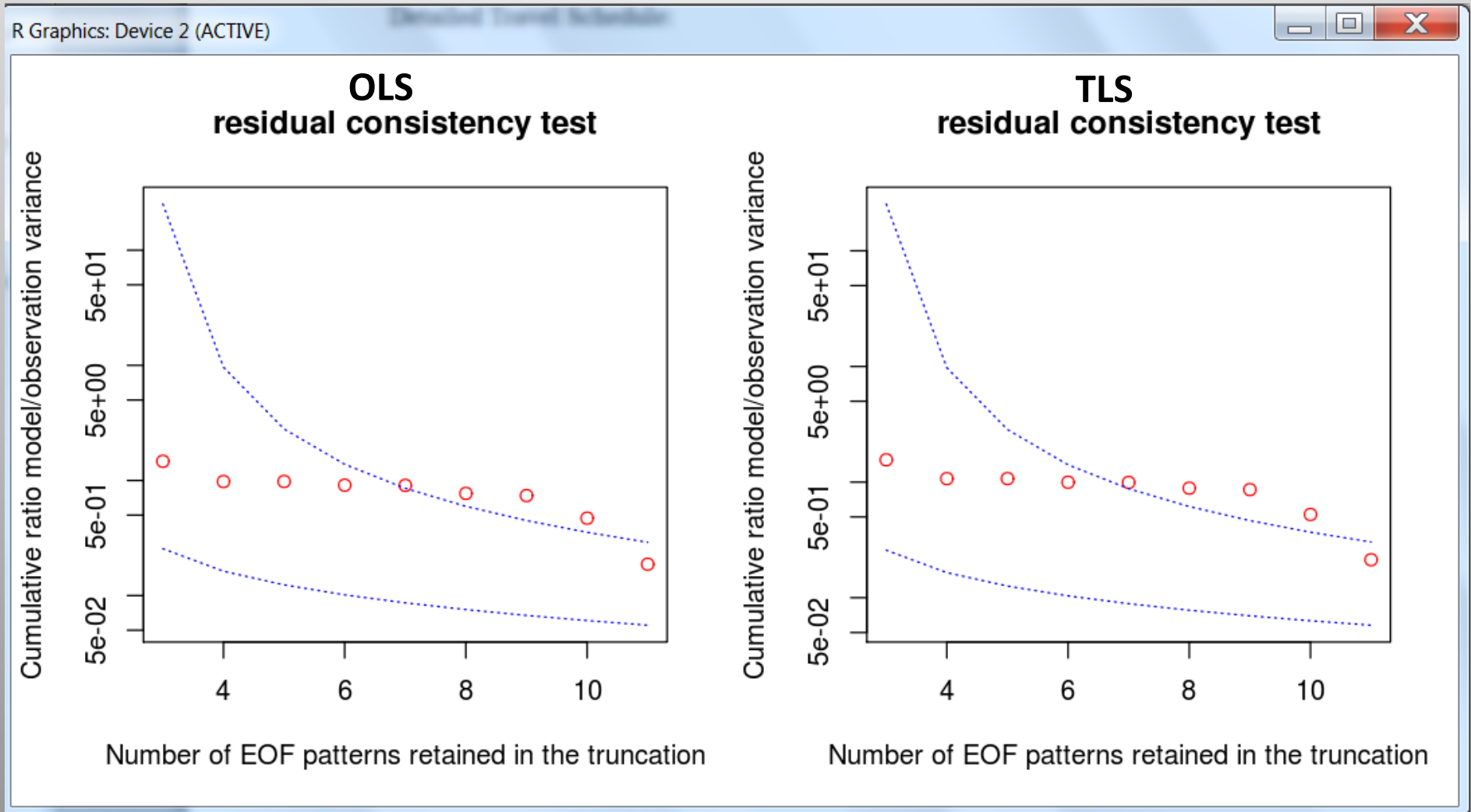


Number of EOF patterns retained in the truncation



Number of EOF patterns retained in the truncation

# 2-signal analysis using ANT+NAT (NH)



**Thank you!**

