

# **COST ES1206 Sub-Working Group "Data Homogenisation": 2nd Workshop – Warsaw, Poland**

Monday 23 January 2017 - Wednesday 25 January 2017

Military University of Technology, Warsaw, Poland



## **Book of Abstracts**



# Contents

Homogenization of synthetic ERA-GPS IWV daily series with Climatol 3.0 . . . . .	1
Homogenization of the synthetic water vapour time series using the penalized maximal t test modified to account for first-order autoregressive noise . . . . .	1
A Technical Description of the Change-Point Detection Method Applied on Climatological Time-Series . . . . .	1
A statistical test, the penalized maximal t test modified to account for first-order autoregressive noise in time series (PMTred) . . . . .	2
A homogenisation of Integrated Water Vapour Time Series Retrieved from GPS and ERA-Interim: Building Synthetic Benchmark Datasets. . . . .	2
Homogenization of daily time series with Climatol 3.0 . . . . .	3
Information on HOMOP . . . . .	3
Results of the Break Detection in HOMOP . . . . .	3
Robust, non-parametric techniques for the identification of change-points in the mean	4
Identified break points by iterative use of non-parametric rank-sum tests: The 'IGS repro 1' dataset . . . . .	4
Overview and goals of the sub-WG activity on homogenization in COST action GNSS4SWEC . . . . .	4
Future plans and outcome of our activities . . . . .	5
An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "Easy" Dataset . . . . .	5
An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "Less-Complicated" Dataset . . . . .	5
An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "Fully-Complicated" Dataset . . . . .	5
An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "IGS Repro 1" Dataset . . . . .	6
Welcome . . . . .	6

Deriving Error Metrics for the Homogenization of Integrated Water Vapour (IWV) Time Series: The Case of the Synthetic Benchmark Datasets . . . . .	6
Open Discussion Session . . . . .	7
Open Discussion Session . . . . .	7
Identified break points by iterative use of non-parametric rank-sum tests: Case-Study of the Synthetic Dataset . . . . .	7
Non-parametric Techniques for Trend and Change Point Detection Analysis . . . . .	7
Homogenisation of synthetic dataset: First results of contribution delivered with various statistical approaches . . . . .	8
Comparison of Time Series Homogenization of IWV Estimates at IGS Stations . . . . .	8

**Results from the Synthetic Datasets / 0****Homogenization of synthetic ERA-GPS IWV daily series with Climatol 3.0**Dr. GUIJARRO, Jose A.<sup>1</sup><sup>1</sup> *Spanish Meteorological Agency***Corresponding Author(s):** jguijarrop@aemet.es

‘Climatol’ is a contributed R package providing some tools for climatological tasks, the most important being the homogenization of climate time series by means of its homogen function. It was originally addressed to monthly series, but direct homogenization of daily series is also possible. However, homogenization of daily climatological series is recognized as a difficult task due to their low signal to noise ratio. The most advised methodology so far has been to adjust the daily values by interpolating monthly corrections derived from a homogenization at the monthly scale, when the power of break detection is much higher than at the daily scale, but this methodology did not yield good results when implemented in a development version of Climatol. Fortunately, this procedure can be avoided in this package, and daily series can be adjusted in the new version 3.0 by honoring the break-points detected at the monthly scale, without the need of interpolating monthly corrections.

In this presentation, details of the specific application of Climatol to the synthetic ERA-GPS IWV daily series will be given, explaining why the direct homogenization of the daily series had to be chosen, and how a better homogenization can be foreseen by finding break-points in the monthly aggregated series.

**Results from the Synthetic Datasets / 1****Homogenization of the synthetic water vapour time series using the penalized maximal t test modified to account for first-order autoregressive noise**Dr. NING, TONG<sup>1</sup><sup>1</sup> *NT***Corresponding Author(s):** tong.ning@lm.se

Atmospheric water vapour is one of the most important climate feedback process and a very efficient greenhouse gas. The long-term trends estimated from the atmospheric water vapour is therefore important for climate monitoring as an independent data source. However the potential temporal shifts in the integrated water vapour (IWV) time series obtained from different techniques, e.g. the global navigation satellite systems (GNSS), can change the resulting trends significantly. In order to obtain realistic and reliable climate signals a homogenization of the IWV time series is necessary.

In total 16 years of synthetic time series of IWV values from GPS data (IGS repro1), and the corresponding IWV differences to the ERA-Interim model were simulated based on the characterisation derived from the real datasets. In this work a statistical test, the penalized maximal t test modified to account for first-order autoregressive noise in time series (PMTred), is used to identify the possible mean shifts (change-points) in the synthetic IWV time series. For the time being, we implemented the homogenization on the “Fully-complicated” dataset which were simulated time series after taking the trends, amplitudes of seasonal signals, first-order autoregressive noise and percentage of gaps into account.

**Homogenization Algorithms / 2****A Technical Description of the Change-Point Detection Method Applied on Climatological Time-Series****Author(s):** Mr. ELIAS, Michal<sup>1</sup>**Co-author(s):** Dr. DOUSA, Jan <sup>1</sup> ; Prof. JARUSKOVA, Daniela <sup>2</sup>

<sup>1</sup> *RIGTC Zdiby*<sup>2</sup> *CTU in Prague***Corresponding Author(s):** michal.elias@pecny.cz

The presentation reviews a statistical method which has been implemented on an inhomogeneous synthetic climatological time-series. The main objective is to assess the method which is based on a two-sample t-statistics and which promotes the idea of comparing candidate series with a reference series that is assumed to be homogeneous. Among other things, we also introduce a reference series calculation which is based on a simple median year strategy

### Homogenization Algorithms / 3

## A statistical test, the penalized maximal t test modified to account for first-order autoregressive noise in time series (PMTred)

Dr. NING, TONG<sup>1</sup><sup>1</sup> *NT***Corresponding Author(s):** eric.pottiaux@oma.be

A t test is normally used to determine if the mean values of two datasets are significantly different from each other. Therefore, it is possible to be implemented in the difference of integrated water vapor (IWV) time series obtained from two techniques, e.g. GPS and EARI, in order to detect the potential temporal shifts in the data from the two techniques. A penalized maximal t test (PMT) to empirically construct a penalty function that evens out the U-shaped false-alarm distribution over the relative position in the time series being tested. In addition, if there is a positive autocorrelation existing in the time series being tested and its effects are ignored, it is highly possible that wrong changepoints will be detected. Therefore, a penalized maximal t test modified to account for first-order autoregressive noise in time series (PMTred) was used on the synthetic data for the homogenization test.

To generate empirical critical values (CVs) for the PMTred test, we carried out a large number of Monte Carlo simulations for different sample length N. For each case, we simulated 1 000 000 homogenous independent and identically distributed (IID) Gaussian time series with a mean of zero and standard deviation of one. To take the lag-1 autocorrelation into account, we created the autocorrelated time series using a first-order autoregressive model (AR1). A changepoint is detected when the PMTred test statistic PTmax is larger than the CV corresponding to a 99.9% confidence level.

### Opening Session / 4

## A homogenisation of Integrated Water Vapour Time Series Retrieved from GPS and ERA-Interim: Building Synthetic Benchmark Datasets.

**Author(s):** Dr. KLOS, Anna<sup>1</sup>**Co-author(s):** Dr. POTTIAUX, Eric <sup>2</sup> ; Dr. VAN MALDEREN, Roeland <sup>3</sup> ; Dr. BOCK, Olivier <sup>4</sup> ; Dr. BOGUSZ, Janusz <sup>1</sup><sup>1</sup> *Military University of Technology*<sup>2</sup> *Royal Observatory of Belgium (ROB)*<sup>3</sup> *Royal Meteorological Institute of Belgium*<sup>4</sup> *IGN***Corresponding Author(s):** anna.klos@wat.edu.pl

This research aims at providing the synthetic benchmark datasets that will be used by the COST Action ES1206 'data homogenization' sub-working group to test the effectiveness of different statistical tools in view of the homogenization of Integrated Water Vapour (IWV) time series retrieved from GPS (Global Positioning System) and ERA-Interim. The effectiveness of these

tools will be tested by their ability to properly report the epochs of offsets that were simulated along with seasonal signals, trend, gaps and two different types of coloured noises. The synthetic benchmark datasets have been generated based on the characterisation parameters (noise types and amplitudes, percentage of gaps, signals...) that were estimated from the common reference IWV time series retrieved from GPS and ERA-Interim model output (dataset shared within the working group for its activities), for a number of 120 International GNSS Service (IGS) permanent stations. Then, we simulated three flavours of 25-year long IWV time series datasets. The 'Easy' dataset includes seasonal signals (annual, semi-annual, 3 and 4 months if present for a particular station), offsets and only white noise. The 'Less-complicated' dataset includes the above-mentioned, as well as the combination of white noise and first order autoregressive processes (AR(1)+WH). The 'Fully-complicated' dataset includes, beyond the above characteristics, a trend and gaps. In this presentation, we focus on the deterministic parameters as well as stochastic properties that we estimated from the common IGS repro 1 reference IWV. Moreover, we describe in details the provided synthetic dataset along with possible impact of simulated but undetected offsets on trend and its uncertainty for their potential future application to the climate studies.

## Homogenization Algorithms / 5

### Homogenization of daily time series with Climatol 3.0

Dr. GUIJARRO, Jose A.<sup>1</sup>

<sup>1</sup> *Spanish Meteorological Agency*

**Corresponding Author(s):** eric.pottiaux@oma.be

Climatol' is a contributed R package providing some tools for climatological tasks, the most important being the homogenization of climate time series by means of its homogen function. It was originally addressed to monthly series, but direct homogenization of daily series is also possible. However, homogenization of daily climatological series is recognized as a difficult task due to their low signal to noise ratio. The most advised methodology so far has been to adjust the daily values by interpolating monthly corrections derived from a homogenization at the monthly scale, when the power of break detection is much higher than at the daily scale, but this methodology did not yield good results when implemented in a development version of Climatol. Fortunately, this procedure can be avoided in this package, and daily series can be adjusted in the new version 3.0 by honoring the break-points detected at the monthly scale, without the need of interpolating monthly corrections.

This presentation is dedicated to explain the homogenization methods implemented in the Climatol package, and the tools provided to obtain some products from the homogenized series.

## Homogenization Algorithms / 6

### Information on HOMOP

CHIMANI, Barbara<sup>1</sup>

<sup>1</sup> *ZAMG*

**Corresponding Author(s):** barbara.chimani@zamg.ac.at

The homogenisation tool used at ZAMG for daily data is called HOMOP. It combines PRODIGE for break detection with SPLIDHOM and INTERP for break correction. Only break detection was used in this work. In HOMOP three different penalty criteria are used for PRODIGE. The methods of Caussinus and Lyazrhi (1997), Jong et al (2003) and Lebarbier (2005). Additionally breaks are detected in different seasons: summer, winter, summer and winter half years and the annual data. Usually those different results are combined leading to the definition of a break as the following: at least two methods have to find the breaks in more than 50% of the reference stations in at least two of the time series. The presentation will give a short insight into the method.

## Results from the Synthetic Datasets / 7

## Results of the Break Detection in HOMOP

CHIMANI, Barbara<sup>1</sup>

<sup>1</sup> ZAMG

**Corresponding Author(s):** barbara.chimani@zamg.ac.at

The break detection of HOMOP was applied to the Easy and the Fully Complicated datasets. Different criteria for the selection of reference stations were used and a different number of missing data was allowed. The presentation will give information on the results of these tests and the influence of the different penalty criteria used.

### Homogenization Algorithms / 9

## Robust, non-parametric techniques for the identification of change-points in the mean

**Author(s):** Dr. VAN MALDEREN, Roeland<sup>1</sup>

**Co-author(s):** Dr. POTTIAUX, Eric <sup>2</sup>

<sup>1</sup> *Royal Meteorological Institute of Belgium*

<sup>2</sup> *Royal Observatory of Belgium (ROB)*

**Corresponding Author(s):** roeland.vanmalderen@meteo.be

We present three different tests, based on non-parametric techniques, for the objective identification of discontinuities in the mean (or median) of time series. These change-point tests rely on (the sum of) the ranks of the values in the sequence and do not require 'metadata' or comparison with neighbouring stations or reference data. We also present how these three tests are combined to detect multiple change points in a time series.

### Results based on the 'IGS repro 1 - ERA-Interim' IWV dataset / 10

## Identified break points by iterative use of non-parametric rank-sum tests: The 'IGS repro 1' dataset

**Author(s):** Dr. VAN MALDEREN, Roeland<sup>1</sup>

**Co-author(s):** Dr. POTTIAUX, Eric <sup>2</sup>

<sup>1</sup> *Royal Meteorological Institute of Belgium*

<sup>2</sup> *Royal Observatory of Belgium (ROB)*

**Corresponding Author(s):** roeland.vanmalderen@meteo.be

We present the break points that have been identified in the time series by applying non-parametric rank-sum tests on the daily and monthly IWV differences at the station locations. We comment on the observed weaknesses of the method, the statistical significance, and the amplitude of the offsets that might be detected.

### Opening Session / 11

## Overview and goals of the sub-WG activity on homogenization in COST action GNSS4SWEC

**Author(s):** Dr. VAN MALDEREN, Roeland<sup>1</sup>

**Co-author(s):** Dr. POTTIAUX, Eric <sup>2</sup>

<sup>1</sup> *Royal Meteorological Institute of Belgium*

<sup>2</sup> *Royal Observatory of Belgium (ROB)*



**Corresponding Author(s):** roeland.vanmalderen@meteo.be

In this presentation, we will give a summary of the goals, past achievements, and timeline of the sub-working group 'data homogenization' activity.

## Future Plans and Outcome of our Activities / 12

### Future plans and outcome of our activities

**Author(s):** Dr. VAN MALDEREN, Roeland<sup>1</sup>

**Co-author(s):** Dr. POTTIAUX, Eric<sup>2</sup>

<sup>1</sup> *Royal Meteorological Institute of Belgium*

<sup>2</sup> *Royal Observatory of Belgium (ROB)*

**Corresponding Author(s):** roeland.vanmalderen@meteo.be

In this presentation, we will present some ideas to feed the discussions on how the results of our sub-working group 'data homogenization' activity can be outreached to 1) the GNSS4SWEC community, and 2) to the entire scientific community.

## Results from the Synthetic Datasets / 13

### An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "Easy" Dataset

**Author(s):** ELIAS, Michal<sup>1</sup>

**Co-author(s):** Dr. DOUSA, Jan<sup>1</sup> ; Prof. JARUSKOVA, Daniela<sup>2</sup>

<sup>1</sup> *GOP Pecny*

<sup>2</sup> *CTU Prague*

**Corresponding Author(s):** michal.elias@pecny.cz

We implemented a method of mathematical statistics for a change-point detection. The method was evaluated and the results were compared with the "Easy" synthetic time-series dataset. Even though the applied method is suggested only for a sudden change determination in the analysed series, based on the results, this method is applicable to detect the inhomogeneity in the analysed series where more than one suspected change-point is included, too. The method's effectivity estimate in such cases is about 50%.

## Results from the Synthetic Datasets / 14

### An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "Less-Complicated" Dataset

**Author(s):** ELIAS, Michal<sup>1</sup>

**Co-author(s):** Dr. DOUSA, Jan<sup>1</sup> ; Prof. JARUSKOVA, Daniela<sup>2</sup>

<sup>1</sup> *GOP Pecny*

<sup>2</sup> *CTU Prague*

**Corresponding Author(s):** michal.elias@pecny.cz

We implemented a method of mathematical statistics for a change-point detection. The method was evaluated and the results were compared with the "Less-Complicated" synthetic time-series dataset. A brief report on change-point detection applied on this particular kind of series is submitted in the presentation.

## Results from the Synthetic Datasets / 15

## **An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "Fully-Complicated" Dataset**

**Author(s):** ELIAS, Michal<sup>1</sup>

**Co-author(s):** Dr. DOUSA, Jan <sup>1</sup> ; Prof. JARUSKOV, Daniela <sup>2</sup>

<sup>1</sup> *GOP Pecny*

<sup>2</sup> *CTU Prague*

**Corresponding Author(s):** michal.elias@pecny.cz

We implemented a method of mathematical statistics for a change-point detection. The method was evaluated and the results were compared with the "Fully-Complicated" synthetic time-series dataset. A short summary report presents two sets of the results; first, when the gaps were not covered in the methodology process and second, when the gaps were in some way eliminated before the change-point detection. Both lists of detected change-points are not consistent, thus the question if the gaps have any impact on the change-point detection may arise. We also provided an estimation of a linear trend in the synthetic GPS IWV's time-series. The first results are also discussed in the presentation.

**Results based on the 'IGS repro 1 - ERA-Interim' IWV dataset / 16**

## **An Assessment of the Change-Point Detection Method Applied on Synthetic Time-Series: Case-Study "IGS Repro 1" Dataset**

**Author(s):** ELIAS, Michal<sup>1</sup>

**Co-author(s):** Dr. DOUSA, Jan <sup>1</sup> ; Prof. JARUSKOVA, Daniela <sup>2</sup>

<sup>1</sup> *GOP Pecny*

<sup>2</sup> *CTU Prague*

**Corresponding Author(s):** michal.elias@pecny.cz

We analysed the presented method of change-point detection. This method is based on a second correlated series that can be relied on to be correct. In this section we introduce the results when the method was applied on the IGS Repo 1 dataset.

**Opening Session / 17**

## **Welcome**

Dr. KLOS, Anna<sup>1</sup>

<sup>1</sup> *Military University of Technology*

**Corresponding Author(s):** anna.klos@wat.edu.pl

**Results from the Synthetic Datasets / 18**

## **Deriving Error Metrics for the Homogenization of Integrated Water Vapour (IWV) Time Series: The Case of the Synthetic Benchmark Datasets**

**Author(s):** Dr. POTTIAUX, Eric<sup>1</sup>

**Co-author(s):** MICHAL, Elias <sup>2</sup> ; Dr. GUIJARRO, Jose A. <sup>3</sup> ; Dr. NING, TONG <sup>4</sup> ; CHIMANI, Barbara <sup>5</sup> ; Dr. KLOS, Anna <sup>6</sup> ; Dr. VAN MALDEREN, Roeland <sup>7</sup>

<sup>1</sup> *Royal Observatory of Belgium (ROB)*

<sup>2</sup> *GOP Pecny*

<sup>3</sup> *Spanish Meteorological Agency*

<sup>4</sup> *NT*

<sup>5</sup> *ZAMG*

<sup>6</sup> *Military University of Technology*

<sup>7</sup> *Royal Meteorological Institute of Belgium*

**Corresponding Author(s):** eric.pottiaux@oma.be

In the same spirit as in the Venema et al. (2012) paper, we derived statistical scores (true positive, true negative, false positive, false negative) and probability scores (probability of true detection probability of false detection...) for each synthetic dataset type, and each sub-working group member contribution (i.e. results from running their homogenization tool on the synthetic datasets received). In this presentation, we summarize the results obtained with these metrics, focusing mainly on the sensitivity of these scores for each homogenization tool w.r.t. to the complexity of the synthetic dataset (Easy, Less and Fully complicated).

**Future Plans and Outcome of our Activities / 19**

## Open Discussion Session

**Future Plans and Outcome of our Activities / 20**

## Open Discussion Session

**Results from the Synthetic Datasets / 21**

## Identified break points by iterative use of non-parametric rank-sum tests: Case-Study of the Synthetic Dataset

**Author(s):** Dr. VAN MALDEREN, Roeland<sup>1</sup>

**Co-author(s):** Dr. POTTIAUX, Eric <sup>2</sup>

<sup>1</sup> *Royal Meteorological Institute of Belgium*

<sup>2</sup> *Royal Observatory of Belgium (ROB)*

**Corresponding Author(s):** eric.pottiaux@oma.be

We present the break points that have been identified in the time series by applying non-parametric rank-sum tests on the daily and monthly IWV differences at the station locations. We comment on the observed weaknesses of the method, the statistical significance, and the amplitude of the offsets that might be detected.

**Homogenization Algorithms / 23**

## Non-parametric Techniques for Trend and Change Point Detection Analysis

**Author(s):** Mrs. ZENGİN KAZANCI, Selma<sup>1</sup>

**Co-author(s):** Dr. TANIR KAYIKÇI, Emine <sup>1</sup> ; Dr. TORNATORE, Vincenza <sup>2</sup>

<sup>1</sup> *Karadeniz Technical University*

<sup>2</sup> *Politecnico di Milano*

**Corresponding Author(s):** szengin@ktu.edu.tr

Integrated water vapour (IWV) is calculated from processing of GNSS data and IWV time series are produced. IWV time series analysis is important to monitoring of climate changes.

Inhomogeneities caused by earthquakes, changes in hardware or equipment etc., can affect actual trends and seasonal variability in time series. This work presents non-parametric techniques for trend and change point analysis.

#### Results from the Synthetic Datasets / 24

### Homogenisation of synthetic dataset: First results of contribution delivered with various statistical approaches

**Author(s):** Dr. KLOS, Anna<sup>1</sup>

**Co-author(s):** Dr. POTTIAUX, Eric<sup>2</sup> ; Dr. VAN MALDEREN, Roeland<sup>3</sup> ; Dr. BOGUSZ, Janusz<sup>1</sup> ; Mrs. GRUSZCZYNSKA, Marta<sup>1</sup> ; Dr. BOCK, Olivier<sup>4</sup> ; CHIMANI, Barbara<sup>5</sup> ; MICHAL, Elias<sup>6</sup> ; Dr. GUIJARRO, Jose A.<sup>7</sup> ; Dr. NING, TONG<sup>8</sup>

<sup>1</sup> *Military University of Technology*

<sup>2</sup> *Royal Observatory of Belgium (ROB)*

<sup>3</sup> *Royal Meteorological Institute of Belgium*

<sup>4</sup> *IGN*

<sup>5</sup> *ZAMG*

<sup>6</sup> *GOP Pecny*

<sup>7</sup> *Spanish Meteorological Agency*

<sup>8</sup> *NT*

**Corresponding Author(s):** eric.pottiaux@oma.be

During this presentation, we will describe the contribution of different statistical approaches employed for synthetic datasets to a task of homogenisation. We will compare a number of offsets found by different algorithms, seasonal signals estimated when different epochs were applied and trends along with their uncertainties to the values which were simulated. We will show how different epochs employed to estimates of above-mentioned may change the value of trend itself. All calculations were performed with Maximum Likelihood Estimation using Hector Software.

#### Results based on the 'IGS repro 1 - ERA-Interim' IWV dataset / 26

### Comparison of Time Series Homogenization of IWV Estimates at IGS Stations

**Author(s):** Mrs. ZENGİN KAZANCI, Selma<sup>1</sup>

**Co-author(s):** Dr. TANIR KAYIKÇI, Emine<sup>1</sup> ; Dr. TORNATORE, vincenza<sup>2</sup>

<sup>1</sup> *Karadeniz Technical University, Department of Geomatics Engineering*

<sup>2</sup> *Politecnico di Milano, Dipartimento di Ingegneria Civile e Ambientale (DICA)*

**Corresponding Author(s):** szengin@ktu.edu.tr

Integrated water vapour (IWV) is calculated from processing of GNSS data and IWV time series are produced. IWV time series analysis is important to monitoring of climate changes. Inhomogeneities caused by earthquakes, changes in hardware or equipment etc., can affect actual trends and seasonal variability in time series. This work presents a methodology for change point detection and series homogenization IWV estimates at various IGS stations in the period 1995-2010 from IGS-Repro1 datasets. Mann-Kendall test, Mann-Kendall Rank Correlation test, von Neuman Ratio test, Pettitt-test and Linear Regression were applied. The aim is to detect significant inhomogeneities occurring in different IGS stations and investigate the reasons of inhomogeneities.