HOMOGENIZING GPS INTEGRATED WATER VAPOUR TIME SERIES: METHODOLOGY AND BENCHMARKING THE ALGORITHMS ON SYNTHETIC DATASETS

R. Van Malderen[1], E. Pottiaux[2], A. Klos[3], O. Bock[4], J. Bogusz[3], B. Chimani[5], M. Elias[6], M. Gruszczynska[3], J. Guijarro[7], S. Zengin Kazancı[8] and T. Ning[9]

[1] Royal Meteorological Institute of Belgium, Brussels, roeland.vanmalderen@meteo.be
[2] Royal Observatory of Belgium, Brussels, Belgium
[3] Military University of Technology, Warsaw, Poland
[4] IGN LAREG, University Paris Diderot, Sorbonne Paris, France
[5] Central Institute for Meteorology and Geodynamics, Austria
[6] Research Institute of Geodesy, Topography and Cartography, Czech Republic
[7] AEMET (Spanish Meteorological Agency), Spain
[8] Karadeniz Technical University, Turkey
[9] Lantmäteriet (Swedish Mapping, Cadastre and Land Registration Authority), Sweden

1. MOTIVATION AND INTRODUCTION

Within the COST Action ES1206 "Advanced Global Navigation Satellite Systems tropospheric products for monitoring severe weather events and climate" (GNSS4SWEC), there was a clear interest and need to homogenize Integrated Water Vapour (IWV) datasets retrieved from Global Navigation Satellite System (GNSS) observations, by correcting (artificial) breakpoints due to e.g. instrumental changes. Based on the results of an inquiry, a homogenization activity was started within Working Group 3 ("Use of GNSS tropospheric products for climate monitoring"), targeting the following objectives: (i) select one or two long-term reference datasets, (ii) apply different homogenization algorithms on these reference datasets, and build up a list of commonly identified inhomogeneities based on statistical detection and metadata information, and (iii) come up with an homogenized version of the reference dataset that can be re-used to study climate trends and time variability by the entire community.

As a first reference dataset, we decided to focus on the existing first tropospheric product given by the data reprocessing of the International GNSS Service (IGS) network, named hereafter IGS repro 1. This homogeneous reprocessing (one single strategy) of the data results from a set of 120 GPS (Global Positioning System) stations distributed worldwide providing continuous observations from 1995 until the end of 2010. However, as can be seen in *Figure 1*, the bulk of the sites are located in the northern hemisphere mid-latitudes. Within the IGS network, the metadata of the stations are archived and publicly available in the so-called IGS logfiles. These contain invaluable information about changes in equipment, operating procedures, site conditions, etc. The retrieved Zenith Total Delays (ZTDs) estimated from the GNSS receiver observations at the stations have been screened, and the outliers have been removed as described in *Bock* (2015). To convert those ZTD measurements in IWV, the surface pressure at the station location and a weighted mean temperature are needed. The European Centre for Medium-Range Weather Forecasts (ECMWF) numerical weather prediction model reanalysis ERA-interim (or ERAI, *Dee et al.* 2011) has been used to obtain those auxiliary meteorological parameters (*Bock*, 2016).

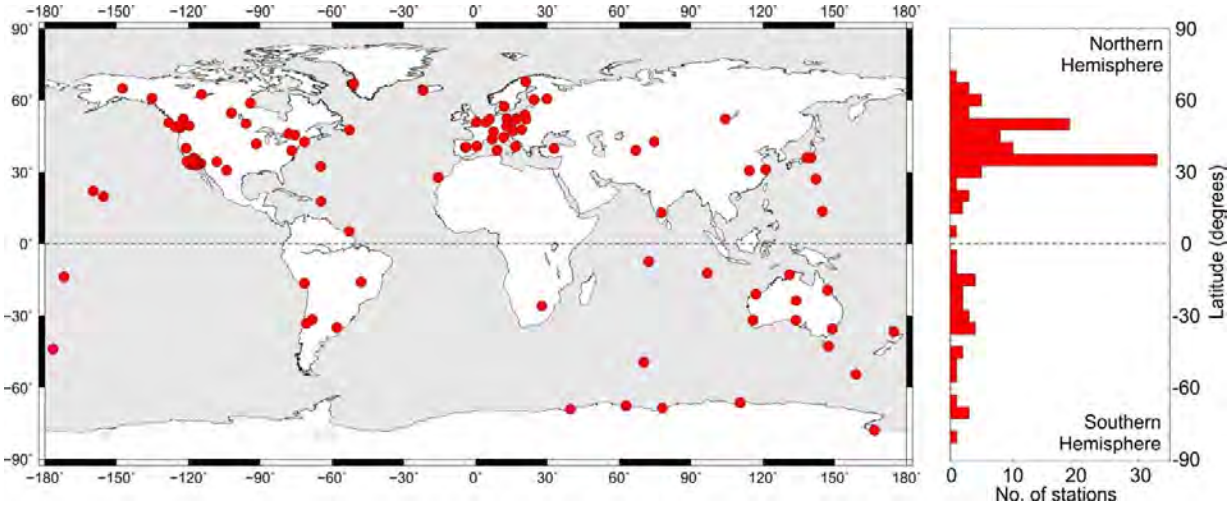


Fig. 1. Distribution of the 120 IGS repro 1 stations with data available from 1995 until the end of 2010.

2. METHODOLOGY

As the distribution of the sites over large areas of the world is rather sparse (see *Figure 1*), the correlations between the IWV time series of our sample sites are rather poor in most areas. As a consequence, the use of neighbouring sites as reference series to remove similar climatic features and to reduce the complexity of the noise characteristics is problematic. Alternatively, various homogenization methods exist that can be used without a reference series (absolute statistical homogenization), but are less reliable (e.g. *Venema et al*. 2012).

Therefore, for a particular GNSS station, we chose to use the ERA-interim IWV time series at this GNSS site location as the reference series for the candidate IGS repro 1 IWV time series. As can be seen in *Figure 2*, for the large majority of the sites, the IGS repro 1 and ERA-interim IWV time series are highly correlated; the lower correlations are ascribed to a bad spatial representation by the model at those sites (e.g. large differences in orography in adjacent pixels). It should however be mentioned that the IGS repro 1 and ERA-interim IWV time series are not completely independent from each other: ERA-interim is used in the ZTD screening process and, as has been noted already above, the surface pressure and weighted mean temperature values, needed for the IGS repro 1 ZTD to IWV conversion, are taken from ERA-interim as well. Another important remark is the fact that ERA-interim might have inhomogeneities of its own, e.g. when new satellite datasets are introduced in the data assimilation system (see e.g. *Schröder et al*. 2016).

Most of the inhomogeneities in the GNSS-derived IWV time series due to antenna or radome changes and changes in the observation statistics (= events) are characterized by jumps in the IWV time series (*Vey et al.*, 2009). Therefore, for each site, we calculate differences time series between the IGS repro 1 and ERA-interim IWV datasets, and we will look for the epochs of those events causing offsets in the difference time series. This approach has already been applied on a similar dataset in *Ning et al.* (2016).

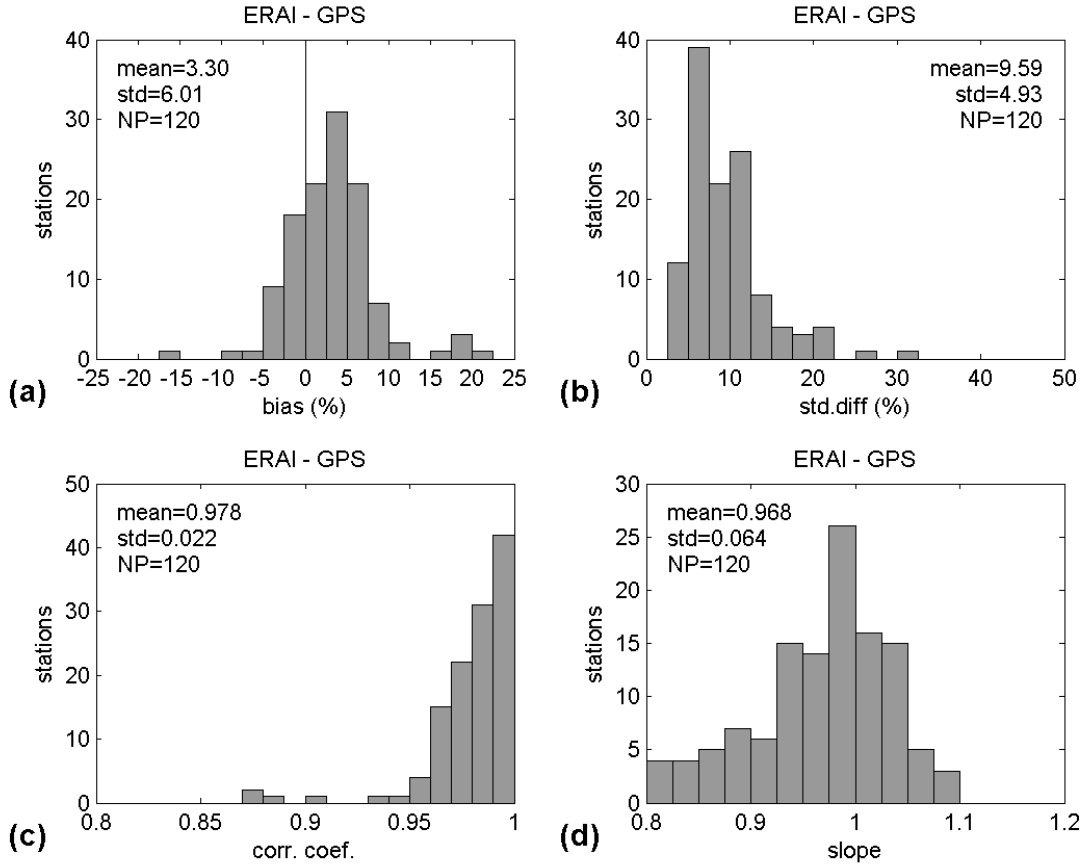


Fig. 2. Histograms of the relative biases (a), relative standard deviations (b), correlation coefficients (c), and linear correlation slope coefficients (d) between the IGS repro 1 and ERA-interim IWV time series for our sample of 120 GNSS stations.

3. SYNTHETIC DATASET GENERATION

We tested different homogenization algorithms on the ERAI-IGS repro 1 differences, and compared their lists of identified epochs of offsets with a list of manually detected breakpoints from the metadata information. At some sites, breakpoints were detected in the metadata and by visual inspection, but not by any of the algorithms. In other cases, breakpoints were detected by a number of (or all) statistical tools, but no metadata information was available for the considered epoch. Therefore, we decided to first generate synthetic time series, with known inserted offsets, on which the different homogenization tools could be blindly applied and assessed. Additionally, we undertook a sensitivity analysis of the performance of the homogenization algorithms on varying characteristics of the synthetic time series.

It should be noted here that we generated synthetic time series of IWV differences directly, based on the characteristics of the real IGS repro 1 and ERA-interim IWV differences. By considering the differences, seasonal variability will be removed and the complexity of the noise will be reduced, making the generation of synthetic time series an easier task. First, we characterized the properties of the offsets (typical number per site and amplitudes) in the real IWV differences, based on the manual detection of 1029 events of instrumental changes, reported in the metadata files of the stations. Of those 1029 events, about 164 epochs were

confirmed by visual inspection, and 57 new epochs were added. We derived the amplitudes of the offsets arising at those epochs and these are used for a first-order correction of the real IWV differences at those 221 identified epochs. Subsequently, we analysed the significant frequencies, the noise model, the presence of a linear trend and gaps in those corrected IWV differences with a Maximum Likelihood Estimation (MLE) in the Hector Software (*Bos et al.* 2013). As it is illustrated for the KOSG station in *Figure 3*, we found that the most adapted noise model is given by the combination of white noise (WN) plus autoregressive noise of the first order (AR(1)), characterised by the amplitudes of white noise (with median value 0.35 mm) and autoregressive noise (median value 0.81 mm), the fraction and coefficient of AR(1), with respective median values 0.71 and 0.50. Another important finding is the presence of trends (of the order of ±0.05 kg/m^2 /yr) in the IWV differences series.

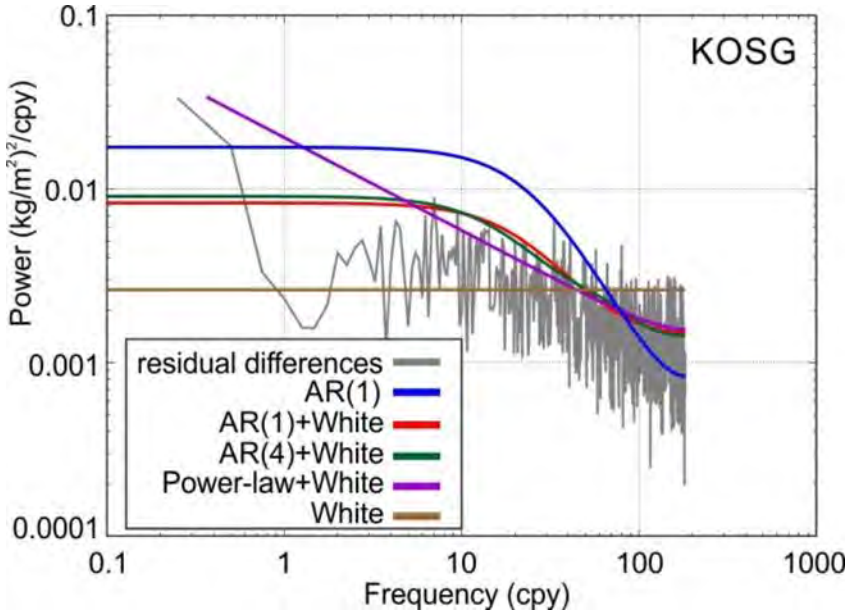


Fig. 3. Power spectrum of the IGS repro 1 and ERA-interim IWV residual differences at the site KOSG (Kootwijk, the Netherlands, 52.18°N, 5.81°E). The colour lines denote different noise models that we tested to decide on the noise model and that best characterize the IWV residuals.

So, based on the characteristics of the IWV differences series at each site separately, we generated for every site a synthetic time series of daily values that includes a number of offsets in the mean. As a matter of fact, to test the sensitivity of the performance of the homogenization tools on the complexity of the time series, 3 datasets of 120 synthetic daily IWV differences time series have been created, with increasing complexity:

- "easy" dataset: includes seasonal signals (annual, semi-annual, ter- and quarter-annual, if present for a particular station) + offsets + white noise (WN)
- "less-complicated" dataset: same as "easy" + autoregressive process of the first order (noise model = AR(1)+WN)
- "fully-complicated" dataset: same as "less-complicated" + trend + gaps (up to 20% of missing data). This dataset is closest to the real IWV differences.

These sets of synthetic time series were made available to the community for a blind testing of homogenization algorithms in use. The inserted offsets of the easy dataset are available to be revealed, if asked for by a participant, for fine-tuning of the algorithm on the use of IWV differences.

4. INVOLVED HOMOGENIZATION ALGORITHMS

In this section, we give a small summary of the homogenization algorithms that participated so far in the blind homogenization of at least one of the variants of the synthetic time series. Those homogenization tools have been applied on daily and/or monthly values of the synthetic datasets.

4.1 Two-sample *t* test (operator: M. Elias)

The procedure applied for the purpose of breakpoint detection is based on hypothesis testing. In this study we used a test statistic that is of so-called "maximum type" (see *Jaruskova*, 1997). Within the field of mathematical statistics the problem can be solved by testing the null hypothesis that claims that there is no change in the distribution of the series, against the alternative hypothesis that claims that the distribution of the series changed at the time k. The null hypothesis is then rejected if at least one of the estimated statistics is larger than the corresponding critical value. Approximate critical values are obtained by the asymptotic distribution (see *Yao and Davis*, 1986). Two ways of time series proceedings and method application were discussed; (i) the proposed method was applied to the uncorrected original series of IGS repro 1 and ERA-interim IWV differences and (ii) the method was applied to corrected difference series when the seasonality was removed and also the gaps were filled in the series before the breakpoint detection, for instance. The method of breakpoint detection is applicable on both monthly and daily time series. A confidence interval for the detected breakpoint is also possible to estimate.

4.2 PMTred (operator: T. Ning)

The rationale of this adapted *t* test is based on *Wang et al.* (2007), which describes this 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. Another modification, named the PMTred test, accounts for the first-order autoregressive noise and it was this test that was used for the homogenization. The critical values (CVs) of the PMTred test were obtained by Monte Carlo simulations running for 1 000 000 times as a function of the sample length N (monthly data, might have to be redone for daily data). In addition, the CVs were calculated for the lag-1 autocorrelation from 0 to 0.95 with an interval of 0.05 and for the confidence levels of 90%, 95%, 99%, and 99.9% (see *Ning et al.,* 2016). This test runs on monthly and daily values, but the critical values are calculated based on monthly data. The detection of multiple breakpoints is achieved by applying the test to the remaining segments.

4.3 HOMOP (operator: B. Chimani)

The homogenization code HOMOP is a combination of PRODIGE (for detection, *Caussinus and Mestre*, 2004), SPLIDHOM (adjustment, *Mestre et al*, 2011), an adapted interpolation (*Vincent et al*., 2002), and improved by some additional plots for facilitating the decision of the homogenisation and extended with some uncertainty information by using different reference stations as well as bootstrapping methods (HOMOP, *Gruber et al* 2009). The approach is neighbour-based, and in the particular case of our synthetic datasets, a lower limit of 0.6 for the correlation coefficients was imposed for selecting potential reference stations. Break detection is done at annual or seasonal base.

4.4 CLIMATOL (operator: J. Guijarro)

Another neighbour-based homogenization algorithm is CLIMATOL, which performs a form of orthogonal regression known as Reduced Major Axis (RMA, *Leduc*, 1987) between the standardized anomalies $(x-\mu x)/\sigma x$ and $(y-\mu y)/\sigma y$ of the two distributions. Orthogonal regression is adjusted by minimizing the perpendicular distance of the scatter points to the regression line, instead of minimizing the vertical distance to that line as in Ordinary Least Squares regression (OLS). In the case of our synthetic datasets, it was imposed that the only reference time series at the site is the ERA-interim time series. The Standard Normal Homogeneity Test (SNHT, *Alexandersson*, 1986) is applied to find shifts in the mean of the anomaly series in two stages. The code incorporates a filling in of missing data and outlier removal. The adjustment of the identified offsets can be done with a varying amplitude: by including e.g. σx in the standardization, you might include seasonality in the amplitudes. As in the other algorithms described so far, the detection of multiple breakpoints is done by applying the test to the remaining segments. CLIMATOL can be applied to any time scale data, but it is advised to detect the breakpoints at the monthly scale, and then use the break dates to adjust the daily series. This algorithm does not provide the amplitudes of breaks, as they are time varying. We might obtain the amplitudes by differencing the non-homogenized series with the homogeneous series.

4.5 Non-parametric tests (operator: R. Van Malderen)

In this case, the used statistical tests are non-parametric distributional tests that utilize the ranks of the time series to find breakpoints (or more general to test the equality of the medians of two distributions). Because such tests are based on ranks, there are not adversely affected by outliers and can be used when the time series has gaps. On the other hand, the significance of the test statistic cannot be evaluated confidently within 10 points of the ends of the time series and those tests show an increased sensitivity to breakpoints in the middle of the time series, when a clear trend is present (*Lanzante*, 1996). We used two of such non-parametric tests: the Mann-Whitney (-Wilcoxon) test and the Pettitt (-Mann-Whitney) test, nicely described in *Lanzante* (1996). As an additional reference, the CUSUM test, based on the sum of the deviations from the mean, is also used. We developed an iterative procedure to detect multiple breakpoints: if 2 out of those 3 tests identify a statistical significant breakpoint, the time series is corrected (by adjustment of the oldest segment with the detected amplitude of the offset) and the 3 tests are applied again on the complete corrected time series. These tests have been applied on both the monthly and daily values.

4.6 Pettitt test (operator: S. Zengin Kazancı)

The Pettitt test (*Petitt*, 1979) has been applied by another operator on the ranks of the daily values, together with the von Neumann ratio (*von Neumann*, 1941) to determine if there is a breakpoint in the time series. If the series is homogeneous, the von Neumann ratio is equal to 2, for lower values of this ratio the series has a breakpoint (*Wijngaard et al*., 2003). The Pettitt test statistic is related to the Mann-Whitney statistic (see above).

5. ASSESSMENT OF THE PERFORMANCE OF THE TOOLS ON THE SYNTHETIC DATASETS

In this section, we will assess the performance of the different homogenization tools on the synthetic datasets on two different aspects: (i) the identification of the epochs of the inserted breakpoints (+ sensitivity analysis) in the synthetic datasets, and (ii) the estimation of the trends that were or were not imposed to the 3 sets of synthetic IWV differences.

5.1 Identification of the breakpoints

To assess whether or not the breakpoint given by a statistical detection tool coincides with the inserted, known, epoch of the break depends on the choice of the time window. Some homogenization algorithms give a confidence interval for the detected breakpoints, but other tools do not. To treat those different methods in a consistent manner, a proper, fixed time window for successful detection has to set. Therefore, we calculated for every homogenization tool the mean time difference between estimated and inserted epochs of offsets (with e.g. an upper limit of half a year). The results for the fully complicated dataset is given in *Figure 4*. From this figure, it can be derived that the here adopted time window of 2 months (62 days) is a good compromise. It should also be noted that the means of the time differences based on daily values (7 left bars) are in line with those calculated from monthly values (6 right bars), and that the mean epoch differences obtained for the easy and less-complicated datasets also have comparable values.

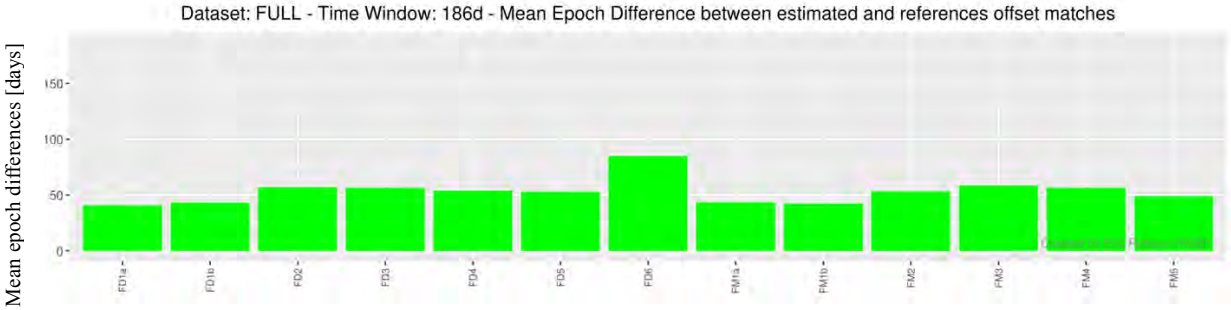
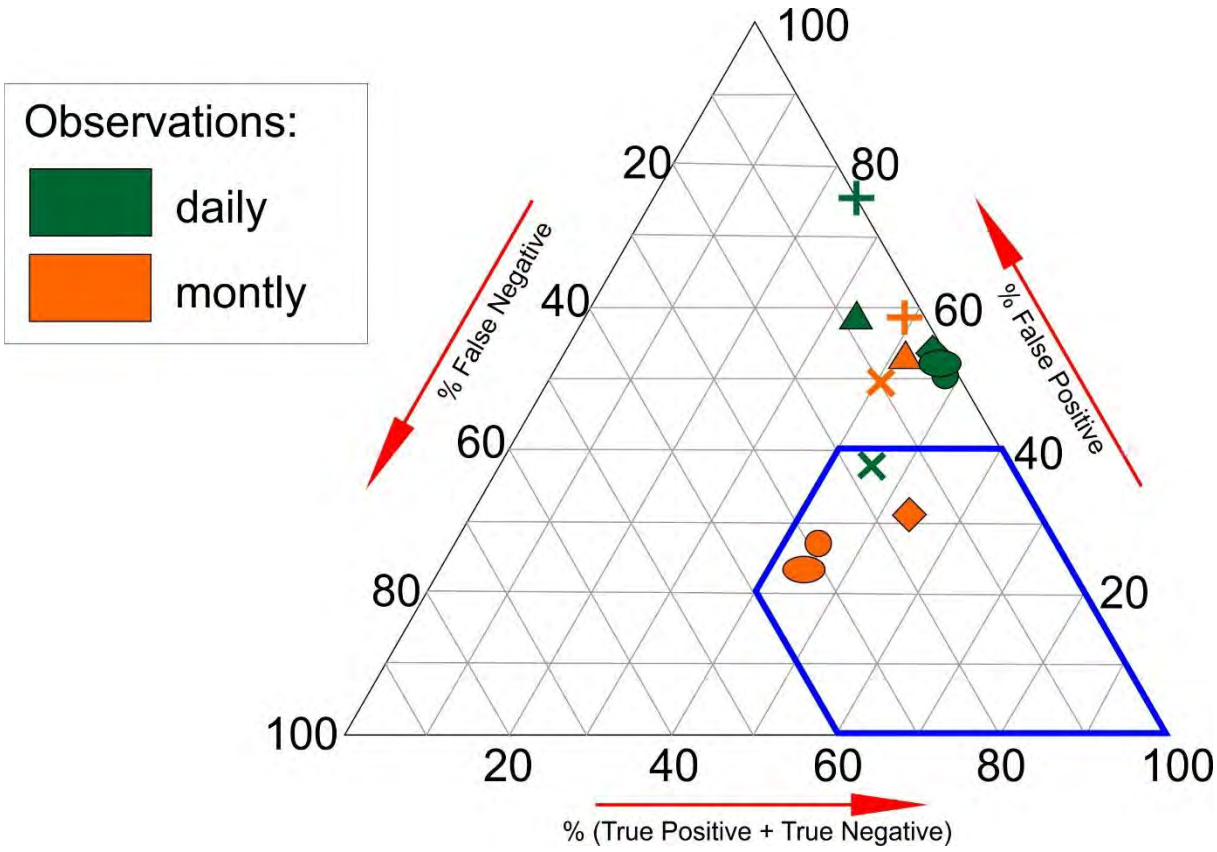


Fig. 4. Means of the time differences between the estimated and inserted epochs of offsets for the fully complicated synthetic dataset, with an upper limit of 186 days (half a year) for the time window. Every bar represents another breakpoint detection solution, with "D" or "M" for application on daily or monthly values respectively, and the numbers referring to the different homogenization tools described in section 4.

With the time window set, we calculate for every breakpoint detection tool the statistical scores: the true positives (TP, "hits"), true negatives (TN: no breaks inserted, no break found), false positives (FP, "false alarms"), and false negatives (FN, "misses"). More details on how to calculate these scores can be found in e.g. *Venema et al.* (2012). To visualize the performance of the different tools in terms of those different statistical scores, we adapted the ternary graph representation from *Gazeaux* et al. (2013), shown in *Figure 5*, again for the fully complicated dataset. It depicts the ratios of the statistical detection scores (TP+TN, FP, and FN) by their position in an equilateral triangle, highlighting the trade-off between those. A perfect solution would appear on the bottom right corner of the triangle (see dashed lines in the figure). From a glance on this figure, it can be directly noted that the involved homogenization tools do not perform very well for the fully complicated dataset: especially the number of false positives are too high. Indeed, when we calculate the probabilities of true

detection (POD = TP/(TP+FN)) and the probabilities of false detection (POFD = FP/(FP+TN)), we found high values for the probabilities of false detection (POFD > 0.75 for all but 2 variants of the same method). Fortunately, the probabilities of true detection are also high (POD > 0.75 for two third of the methods). Some methods nearly detect all the inserted breakpoints, but at the cost of a high number of false alarms, while other methods are more conservative in detecting breakpoints, resulting in low scores for both the POD and POFD. It is therefore not surprising that the Pierce Skill Scores, defined as POD – POFD, are very low (around zero and even negative) for the homogenization tools applied on the fully complicated dataset. For comparison: the Pierce Skill Scores lie in the range of 0.10 to 0.63 and 0.04 to 0.22 for the homogenization algorithms that took part in the benchmark for monthly temperature, respectively precipitation time series described in *Venema et al.* (2012).



	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7
Symbol	● ⬭	▲	+	×	◆	▼	-
Operator	M. Elias	R. Van Malderen	R. Van Malderen	J. Guijarro	T. Ning	S. Zengin	B.Chimani
Method / SW	2-sample t-test	2 of 3	PMW	CLIMATOL	PMTred	Pettitt	HOMOP
Daily/Monthly	D+M	D+M	D+M	D+M	D+M	D	X
Easy/Less/Full	E+L+F	E+L+F	E+L+F	L+F	E+L+F	E+L+F	E+F

Fig. 5. Ternary graph representing the ratio between three performance measures of the breakpoint detection solutions (TP+TN, FP, and FN). The performance increases with decreasing numbers of false positives and false negatives and increasing numbers of true positives and negatives, so that a perfect solution is located in the lower right corner, marked by the blue area. The different solutions are marked with the symbols and colours outlined in the legend and in the table.

So far, we only discussed the results on the breakpoint detection on the fully complicated dataset. It should however be noted that a good performance of the tools is achieved for the majority of the participating methods on the easy and less complicated datasets, especially due to a lower amount of false positives. As a consequence, the Pierce Skill Scores are now positive for almost all homogenization tools. So, we can conclude that the performance decreases for almost all the tools when adding gaps and a trend in the benchmark time series; adding autoregressive noise of the first order has less impact.

Some of the homogenization algorithms also provided the (constant) amplitudes of the detected offsets. These were compared with the amplitudes of the offsets that were put in the synthetic time series. The result, again for the fully complicated dataset, are shown in *Figure 6*. From this figure, it could be seen that some methods tend to underestimate the number of offsets with small amplitudes relatively (e.g. ME1 and ME2), while other methods on the contrary overestimate the amount of those offsets (e.g. RVM 2of3 D), but on the other hand underestimate the number of offsets with large amplitudes. Clearly, the different methods have a different sensitivity to the amplitudes of the offsets, and some fine-tuning on the statistical thresholds might be advised for some methods. For the other variants of the synthetic datasets, the amplitude distribution of the detected offsets more closely follows the amplitude distribution of the true inserted offsets.

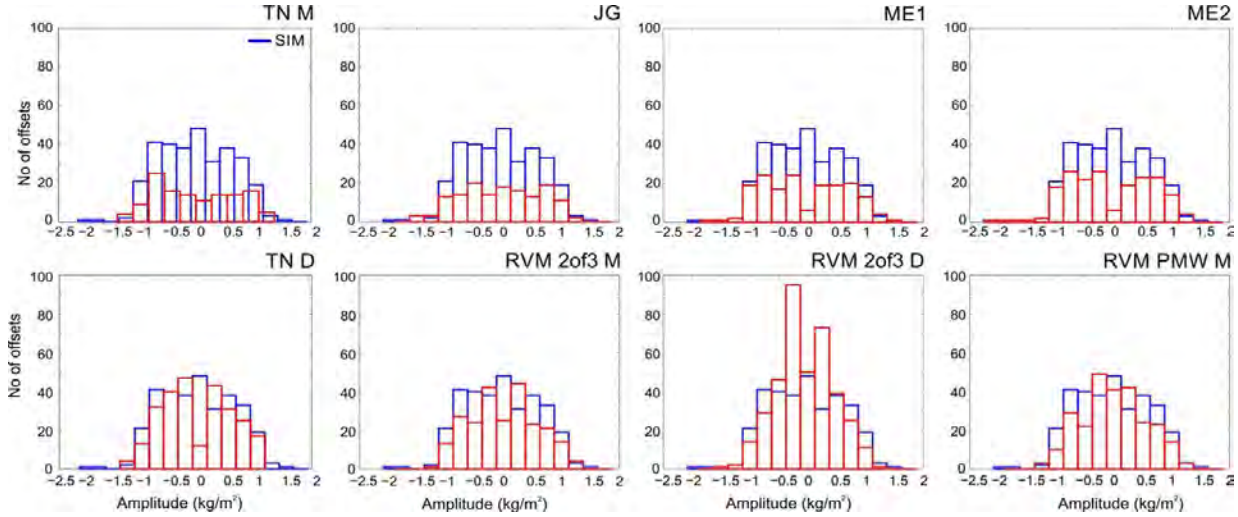


Fig. 6. The histograms of the amplitudes of the detected offsets by the different methods (in red in all figures), compared to the amplitude distribution of the inserted offsets in the fully complicated synthetic dataset of IWV differences (in blue).

5.2 Trend estimation for the homogenized datasets

Only in the fully complicated dataset, a trend was inserted in the IWV differences series, and the homogenized time series by the different time series should hence reveal the same trend. However, as illustrated in *Figure 7* for the easy variant of the generated synthetic IWV differences series for the station CRO1 (Virgin Islands, USA), trends as large as 0.1 kg/m² (or mm) per year arise after correcting for the detected offset by some methods. We stress here that a consistent approach was followed to correct the time series for all detection methods: based on the epoch of the detected offsets, the corresponding amplitude was calculated and corrected for by Maximum Likelihood Estimation (MLE) with the Hector software (*Bos et al.,* 2013). From *Figure 7*, it can be also seen that the calculated trend with this method (denoted with "Hector") deviates around 0.05 kg/m² per year from the true zero trend, due to the uncertainty of the method itself.

The distribution of the differences of trends calculated from the true epochs of the offsets and from the epochs given by different detection methods is given in *Figure 8* for the fully complicated synthetic datasets. The figure shows that most trends differ within ±0.05 kg/m² per year, although those distributions vary between the different detection methods and when applied on either daily (marked with "D") or monthly (marked with "M") values. As one of the mail goals of our homogenization activity is the provision of a homogenized dataset of GNSS IWV time series for use in trend analysis, special care should be taken not to introduce spurious trends in the time series after correction. In this sense, the impact of homogenization on the estimated trend uncertainties should also be further elaborated. The effect of increasing complexity in the synthetic dataset generation (from easy to less and fully complicated) on the estimated trends and their uncertainties is another issue on which additional research is needed.

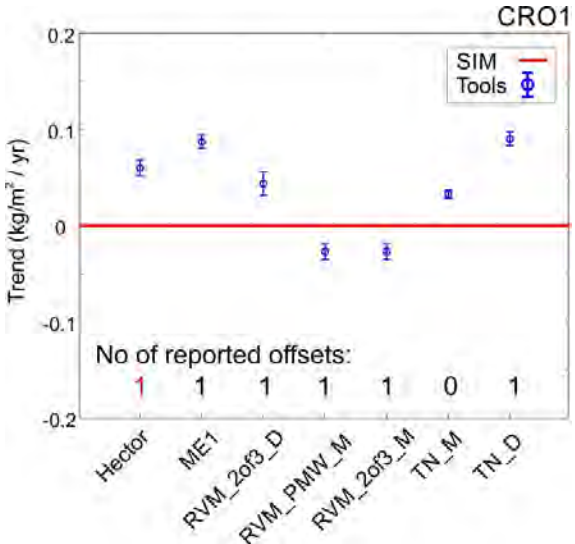


Fig. 7. Trends (in kg/m² /yr) calculated with MLE with the Hector software from the list of identified epochs by the different breakpoint detection tools and from the true epoch of the inserted offset (at the left, marked with "Hector") for the easy variant of synthetic IWV differences for the station CRO1 (Virgin Islands, USA, 17.76°N, 64.58°W).

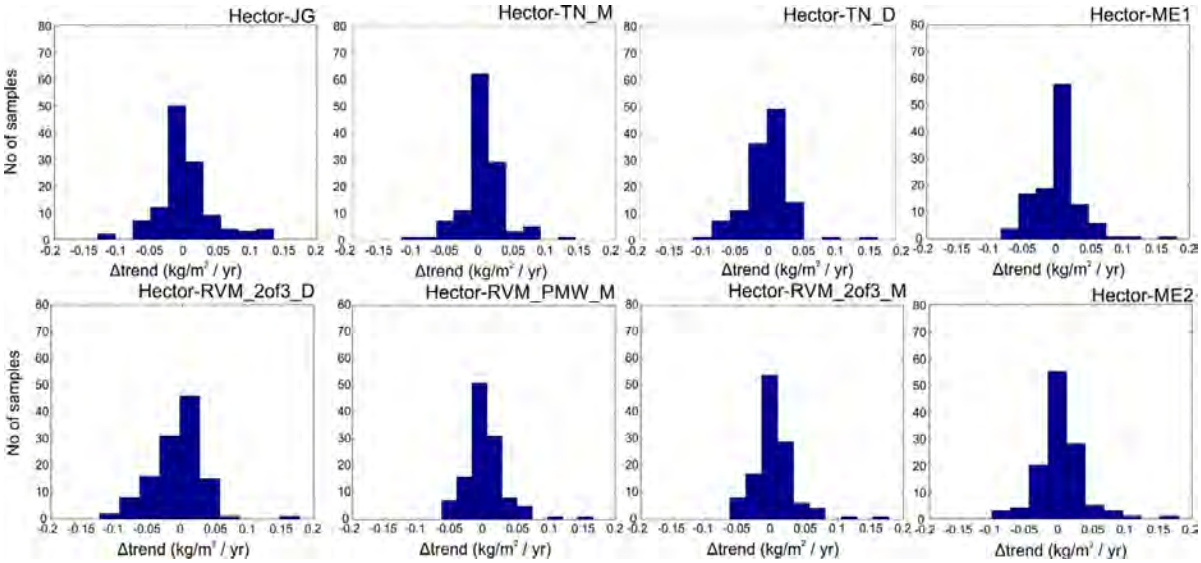


Fig. 8. The histograms of the trend differences calculated for all sites from the fully complicated synthetic time series homogenized either from the true epochs of the offsets or from the epochs given by different detection methods by MLE with the Hector software. Each panel represents another breakpoint detection method.

6. CONCLUSIONS AND OUTLOOK

In this contribution, we described the current activity in homogenizing a world-wide dataset of Integrated Water Vapour (IWV) measurements retrieved from observations made at ground-based GNSS stations. As the distances between those 120 stations are large and correlations are generally low (lower than 0.6 for distances larger than 1°), we used the ERA-interim reanalysis IWV fields at those station locations as the reference time series for relative statistical homogenization. Based on the characteristics of the IWV differences series between the GNSS dataset and ERA-interim and the properties of manually checked instrumental change events reported in the metadata of the GNSS sites, we generated three variants of 120 synthetic IWV difference time series with increasing complexity: we first added autoregressive noise of the first order and subsequently trends and gaps. Those synthetic time series enable us to test the performances of six participating breakpoint detection algorithms and their sensitivity to this increasing dataset complexity.

We found that the performances of those algorithms in identifying the epochs of the inserted offsets especially decreases when adding trends and gaps to the synthetic datasets, due to a larger number of false alarms. On the other hand, the hit rates of most tools are rather good, even when applied on daily values instead of on monthly values. Different tools show a different sensitivity for detecting different ranges of amplitudes of offsets, especially for the most complex (fully complicated) synthetic time series: some tools overestimate (underestimate) the number of small-amplitude (large-amplitude) offsets, while the opposite is true for other breakpoint detection algorithms. After eliminating differences due to different calculation methodologies, we found trend differences mostly within ±0.05 kg/m² per year between the inserted trends and trends calculated from the different homogenization solutions.

Owing to the fact that metadata on instrumental changes are available for the GNSS stations, we primarily focused on the identification of the epochs of offsets until now. At the end, we would like to combine the outcome of statistical breakpoint detection with these metadata. However, we will also assess the performances of the different tools by comparing the final solutions for the time series given by different tools with the original time series (e.g. calculating Centered Root Mean Square errors as in *Venema et al.,* (2012), calculating trends directly from the final solutions, etc.).

Of course, we highly welcome contributions from other groups running homogenization tools, and in the future, our benchmark will already be extended with few more contributions. After providing solutions for the synthetic time series, the participants will get the opportunity to fine-tune their methods on the specifications of the datasets with the help of the knowledge of the true inserted offsets and their amplitudes. Thereafter, a second round of blind homogenization on a newly generated synthetic dataset of IWV values (probably with simulated metadata information) will be held. Based on the performance of the statistical homogenization tools on these synthetic datasets, we will develop a methodology for combining the results of good performing homogenization tools with metadata information. This methodology and those tools will then be applied on the IGS repro 1 dataset of retrieved GPS IWV time series, resulting in a homogenized dataset, which will be validated by other sources of IWV time series and finally made available to the community for assessing the time variability of IWV and for validation of climate model IWV outputs.

Acknowledgements

We would like to thank the COST Action ES1206 GNSS4SWEC for financial support for the two dedicated workshops on homogenization, held in Brussels, Belgium (26-27 April 2016) and in Warsaw, Poland (23-25 January 2017). R. Van Malderen (RMI) and E. Pottiaux (ROB) acknowledge the support from the Solar-Terrestrial Centre of Excellence (STCE).

References

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *J. Climatol*., 6, 661–675.

Bock, O., "ZTD assessment and screening", COST ES1206 GNSS4SWEC Workshop, Thessaloniki, Greece, 11-13 May 2015

Bock, O., "Screening and validation of new reprocessed GNSS IWV data in Arctic region", COST ES1206 GNSS4SWEC Workshop, Potsdam, Germany, 31 August - 2 September 2016

Bos, M. S., Fernandes, R. M. S., Williams, S. D. P., and Bastos, L., 2013, "Fast Error Analysis of Continuous GNSS Observations with Missing Data", *Journal of Geodesy* 87(4): 351–360, doi:10.1007/s00190-012-0605-0.

Caussinus, H.and Mestre, O., 2004: Detection and correction of artificial shifts in climate series., *Journal of the Royal Statistical Society*, Series C, 53, 405-425

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc*., 137, 553–597, doi: 10.1002/qj.828.

Gazeaux, J., . (2013), Detecting offsets in GPS time series: first results from the Detection of Offsets in GPS Experiment, *J. Geophys. Res. Solid Earth*, 118, 2397–2407, doi:10.1002/jgrb.50152

Gruber C., Auer I., Böhm R., 2009: Endberichte HOM-OP Austria Aufbau und Installierung eines Tools zur operationellen Homogenisierung von Klimadaten mit 3 Annexen.

Jaruskova D., 1997: Some Problems with Application of Change-Point Detection Methods to Environmental Data, *Environmetrics*, Vol. 8, No. 5, Pp. 469-483.

Lanzante, J., 1996, "Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data, *Int. J. Climatol*., 16, 1197-1226.

Leduc, D.J. 1987. A comparative analysis of the reduced major axis technique of fitting lines to bivariate data. *Can. J. For. Res*. 17: 654–659.

Mestre O, Gruber C, Prieur C, Caussinus H, Jourdain S, 2011: SPLIDHOM: A Method for Homogenization of Daily Temperature Observations, *J. Appl. Meteor. Climatol*, 50, 2343-2358.

Ning, T., J. Wickert, Z. Deng, S. Heise, G. Dick, S. Vey, and T. Schöne, 2016: Homogenized Time Series of the Atmospheric Water Vapor Content Obtained from the GNSS Reprocessed Data. *J. Climate*, 29, 2443–2456, https://doi.org/10.1175/JCLI-D-15-0158.1

Pettitt, A.N. (1979): A Nonparametric Approach to the Change-Point Problem, *Applied Statistics*, 28, 126-135, http://dx.doi.org/10.2307/2346729.

Schröder, M., M. Lockhoff, J.M. Forsythe, H.Q. Cronk, T.H. Vonder Haar, and R. Bennartz, 2016: The GEWEX Water Vapor Assessment: Results from Intercomparison, Trend, and Homogeneity Analysis of Total Column Water Vapor. *J. Appl. Meteor. Climatol*., 55, 1633–1649, https://doi.org/10.1175/JAMC-D-15-0304.1

Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P., and Brandsma, T.: Benchmarking homogenization algorithms for monthly data, *Clim. Past*, 8, 89-115, doi:10.5194/cp-8-89-2012, 2012

Vey, S., R. Dietrich, M. Fritsche, A. Rülke, P. Steigenberger, and M. Rothacher (2009), On the homogeneity and interpretation of precipitable water time series derived from global GPS observations, *J. Geophys. Res*., 114, D10101, doi:10.1029/2008JD010415.

Vincent, L.A., X. Zhang, B.R. Bonsal, and W.D. Hogg, 2002: Homogenization of Daily Temperatures over Canada. *J. Climate*, 15, 1322–1334, https://doi.org/10.1175/1520-0442(2002)015<1322:HODTOC>2.0.CO;2

Von Neumann J., 1941: Distribution of the ratio of the mean square successive difference to the variance, *Annals of Mathematical Statistics* 13: 367–395.

Wang, X. L., Q. H. Wen, and Y. Wu, 2007: Penalized maximal t test for detecting undocumented mean change in climate data series. *J. Appl. Meteor. Climatol*., 46, 916–931, doi:10.1175/JAM2504.1.

Wijngaard, J. B., Klein Tank, A. M. G. and Können, G. P. (2003), Homogeneity of 20th century European daily temperature and precipitation series. *Int. J. Climatol*., 23: 679–692. doi:10.1002/joc.906

Yao, Y.-C., Davis R. A., 1986: The asymptotic behaviour of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates, Sankhya, 48, 339-353.