Homogenizing GPS Integrated Water Vapor Time Series: Benchmarking Break Detection Methods on Synthetic Data Sets

R. Van Malderen[1] [iD], E. Pottiaux[2], A. Klos[3] [iD], P. Domonkos[4], M. Elias[5], T. Ning[6], O. Bock[7], J. Guijarro[8] [iD], F. Alshawaf[9] [iD], M. Hoseini[10] [iD], A. Quarello[7,11] [iD], E. Lebarbier[11], B. Chimani[12], V. Tornatore[13] [iD], S. Zengin Kazancı[14], and J. Bogusz[3] [iD]

[1]Royal Meteorological Institute of Belgium, Brussels, Belgium, [2]Royal Observatory of Belgium (ROB), Brussels, Belgium, [3]Military University of Technology, Faculty of Civil Engineering and Geodesy, Warsaw, Poland, [4]Unaffiliated, Tortosa, Spain, [5]Topography and Cartography, Research Institute of Geodesy, Zdiby, Czech Republic, [6]Lantmäteriet (Swedish Mapping, Cadastral and Land Registration Authority), Gävle, Sweden, [7]IPGP, IGN, Université de Paris, CNRS, UMR, Paris, France, [8]AEMET (State Meteorological Agency), Madrid, Spain, [9]German Research Centre for Geosciences GFZ, Potsdam, Germany, [10]Norwegian University of Science and Technology, Department of Civil and Environmental Engineering, Trondheim, Norway, [11]AgroParisTech, INRA, Paris, France, [12]Zentralanstalt für Meteorologie und Geophysik, Vienna, Austria, [13]Politecnico di Milano, DICA, Milan, Italy, [14]Karadeniz Technical University, Department of Geomatics Engineering, Trabzon, Turkey

Abstract We assess the performance of different break detection methods on three sets of benchmark data sets, each consisting of 120 daily time series of integrated water vapor differences. These differences are generated from the Global Positioning System (GPS) measurements at 120 sites worldwide, and the numerical weather prediction reanalysis (ERA-Interim) integrated water vapor output, which serves as the reference series here. The benchmark includes homogeneous and inhomogeneous sections with added nonclimatic shifts (breaks) in the latter. Three different variants of the benchmark time series are produced, with increasing complexity, by adding autoregressive noise of the first order to the white noise model and the periodic behavior and consecutively by adding gaps and allowing nonclimatic trends. The purpose of this "complex experiment" is to examine the performance of break detection methods in a more realistic case when the reference series are not homogeneous. We evaluate the performance of break detection methods with skill scores, centered root mean square errors (CRMSE), and trend differences relative to the trends of the homogeneous series. We found that most methods underestimate the number of breaks and have a significant number of false detections. Despite this, the degree of CRMSE reduction is significant (roughly between 40% and 80%) in the easy to moderate experiments, with the ratio of trend bias reduction is even exceeding the 90% of the raw data error. For the complex experiment, the improvement ranges between 15% and 35% with respect to the raw data, both in terms of RMSE and trend estimations.

1. Introduction

Water vapor is a key component for the Earth's climate as it is the most important natural greenhouse gas and responsible for the largest known feedback mechanism for amplifying climate change (the water vapor feedback, see, for example, Soden & Held, 2006). Water vapor also strongly influences atmospheric dynamics and the hydrologic cycle through surface evaporation, latent heat transport and adiabatic heating, and is, in particular, a source of clouds and precipitation. Since the most important challenge of the present-day climate community is to predict and understand the response of the Earth system to increased emissions of greenhouse gases, an assessment of long-term trends in water vapor is vital. International efforts are made to collect, improve, and assess available water vapor measurements, for example, within the Global Energy and Water Exchanges (GEWEX) Water Vapor Assessment (http://gewex-vap.org) by the World Climate Research Programme (https://www.wcrp-climate.org/).

However, due to its high variability, both temporally and spatially, water vapor is one of the most difficult quantities to measure and to predict with numerical weather prediction (NWP) models. Here, we consider the total vertical column water vapor measurements, often referred to as total column water vapor (TCWV), integrated water vapor (IWV), and precipitable water vapor (PWV). Those measurements can be

made from radiosondes or from the analysis of ground-based Global Navigation Satellite System (GNSS) signals, of which the Global Positioning System (GPS) satellites are best known. Both techniques provide good coverage for, for example, North America and Europe, where many permanent ground stations exist, but only sparse coverage over, for example, Central Africa or the oceans. Satellite measurements from microwave (MW) or infrared (IR) sensors, however, are particularly sensitive over ocean or land, respectively (Beirle et al., 2018). An overview and comparison of different IWV measurement techniques are for example given in Guerova et al. (2016), Parracho (2017), and in Sect. 5.6 of Jones et al. (2019).

Here, we concentrate on the IWV time series obtained from the GPS retrievals, extending back in time to the mid-1990s for the first stations. Those GPS data sets have been used already in earlier studies on IWV time variability and trends (e.g., Ning et al., 2016; Parracho et al., 2018; Vey et al., 2010; Wang et al., 2016). The results from these studies indicate that before we can use GPS data to detect realistic and reliable climate signals, the homogeneity of the data must be thoroughly investigated. In the realm of climate science, "homogeneous" means that the values of the climate time series must be of the same nature, that is, comparable, free from nonclimatic effects. Climate time-series homogenization then seeks to detect and remove nonclimatic biases that are systematic over a section of the climate time series (Venema et al., 2018). The changes that affect the homogeneity of the GPS data can fall into two categories: data-processing-related and site-related. The first type of changes is normally due to updates of the reference frame and applied models, different elevation cutoff angle, different mapping functions, and different processing strategies. Steigenberger et al. (2007) found that the data-processing-related inconsistencies can be significantly reduced after a homogenous data reprocessing over the whole data time series. The site-related changes can be due to hardware replacements, for example, antennas and receivers or differences in the measurements, for example, the number of visible GPS satellites and data rate. Such changes introduce systematic errors in GPS measurements (Johansson, 1998). In addition, the site-related changes can also be referred to the changes in the electromagnetic environment due to, for example, growing vegetation (Pierdicca et al., 2014) and/or different soil moisture (Larson et al., 2010). Those changes can cause different multipath effects on the GPS data, and the resulting errors are normally not fixed in time but varying when reflective properties change and therefore harder to detect a posteriori.

Homogenization is a multistep process including one or more cycles of inhomogeneity detection and bias adjustments. Detection involves analyzing the climate series to locate abrupt or gradual inhomogeneities by using statistical tests and visual tools and by preference supported by metadata, that is, documentary information on known changes. Adjustment is the process of using statistical approaches to reduce the bias introduced by the inhomogeneities. Both detection and adjustment should be made by comparing the series with neighboring series. This is called relative homogenization and relies on the principle that nearby stations share not only the same climate but also much of the same temporal evolution (Venema et al., 2018). Although the GPS networks are steadily becoming denser, the distribution of the sites over large areas of the world is rather sparse, and the correlations between the IWV time series of our sample sites are rather poor in most areas. As a consequence, the use of neighboring 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). In earlier work (Bock et al., 2018; Ning et al., 2016; Vey et al., 2009), the time series of the differences between the GPS-derived IWV and the one obtained from an atmospheric reanalysis were used for the data homogenization. Ning et al. (2016) used a statistical test, the penalized maximal t test modified to account for first-order autoregressive noise in time series (PMTred, see section 3.1.4), to detect inhomogeneities in the form of shifts in the mean (hereafter breaks or breakpoints) of the difference time series. This approach allowed for identification of the breaks in the GPS IWV time series with the constraint that detected breaks could occur also for inhomogeneities in the reference series (ERA-Interim from the European Centre for Medium-Range Weather Forecasts [ECMWF] reanalysis [Dee et al., 2011] in their case). To produce reanalysis output, data from many, possibly inhomogeneous, observing systems were assimilated. As a result, changes in the assimilated observations or systems might result in mean shifts. To minimize those mean shifts, a variational bias correction system for, for example, satellite radiances has been developed and used for ERA-Interim (Dee & Uppala, 2009). Despite these efforts, inhomogeneities may still be present in the reference ERA-Interim data set (see also, e.g., Schröder et al., 2016, Schröder et al., 2019). However, after the adjustment of the mean shifts for the GPS data,

Ning et al. (2016) found an improved consistency in the IWV trends between nearby sites. In addition, the IWV trends estimated for 47 GPS sites were compared to the corresponding IWV trends obtained from nearby homogenized radiosonde data. The correlation coefficient of the trends increased significantly, by 38%, with the use of the homogenized GPS data.

In this work, we extend on this approach, but use both data sets to assess the performance of widely used break detection methods, as in, for example, Venema et al. (2012) but now for IWV data. For this purpose, a synthetic benchmark data set was constructed by simulating the IWV differences between GPS and ERA-Interim. In this sense, the GPS IWV is the candidate time series for homogenization, while the ERA-Interim IWV output at the GPS sites is the reference series. The benchmark has a homogeneous and an inhomogeneous counterpart with inserted breaks, periodic biases, local trends, gaps, and a random noise. Three kinds of inhomogeneous sets were constructed to isolate the impact of those features on the performance of the break detection. All break detection procedures are blindly applied to the different versions of the benchmark IWV difference time series, and they are supplied with a common adjustment procedure to assess the residual errors in the homogenized IWV time series, depending on the break detection method.

In section 2, we describe how these synthetic data sets have been constructed from the real GPS and ERA-Interim data sets. In section 3, the involved break detection methods are described, and their performances are assessed in section 4. Section 5 discusses the results and concludes.

2. Data

2.1. Real IWV Data Sets

Within COST Action ES1206 on advanced Global Navigation Satellite Systems tropospheric products for monitoring Severe Weather Events and Climate, GNSS4SWEC, Jones et al. (2019), a homogenization activity was set up, targeting the homogenization of long-term GPS IWV time series. Here, we focus on a global GPS data set of 120 stations covering the period from 1995 to 2010, and we use the ERA-Interim reanalysis as the IWV reference time series (Bock, 2016).

2.1.1. GPS-Based IWV Data Set

The GPS IWV data set is built on reprocessed zenith total delay (ZTD) estimates from the International GNSS Service (IGS) for the period from 1 January 1995 to 31 December 2010, referred IGS-repro1 data set hereafter. The reprocessing was done by NASA Jet Propulsion Laboratory (JPL) with GIPSY OASIS software in May–July 2010 for the period 1995–2007. Basic details on the processing procedure are described by Byun (2009). From 2008 to April 2011, these reprocessing results are complemented with the operational IGS final troposphere products which are fully consistent with the IGS reprocessing results. As mentioned earlier, this homogeneous reprocessing should cancel out processing-related inhomogeneities over time and between stations. After April 2011, IGS products are based on a new terrestrial reference frame and the new antenna models, so that the consistency and homogeneity of processing is not guaranteed anymore. Of the 456 GPS stations covered by the data set, 120 stations have time series spanning from January 1995 to December 2010, with data gaps limited in time. These stations and their geographical distribution are shown in Figure 1. Their exact names, locations, and data availability are listed in Table S1 in the supporting information. A new reprocessing campaign of the IGS network will take place in the course of 2020, so that an extended reprocessed tropospheric data set will be only available in the next years. As we want to stay as close as possible to the real GPS IWV data set with our synthetic benchmark time series, an extension or extrapolation of the synthetic data set beyond 2010 was not considered.

The ZTD estimates are available at a 5 min sampling in the IGS repro1 data set. They were screened using the method described in Parracho et al. (2018), and references therein. The ZTD data were converted to IWV using auxiliary data from ERA-Interim reanalysis as described in Parracho et al. (2018). As the auxiliary data are only available at 0 hr, 6 hr, 12 hr, and 18 hr UTC, the ZTD to IWV conversion is done at those times only. Daily and monthly means are calculated subsequently.

The GPS and corresponding ERA-Interim IWV data (described below) and their differences used in this work are freely available from the link provided in (Bock, 2016), but note that this is the first version of the data set, which unfortunately contained a bug when the daily data were formed leading to the loss of 20% of the data. A later version was released in which this bug was corrected (Bock, 2017).
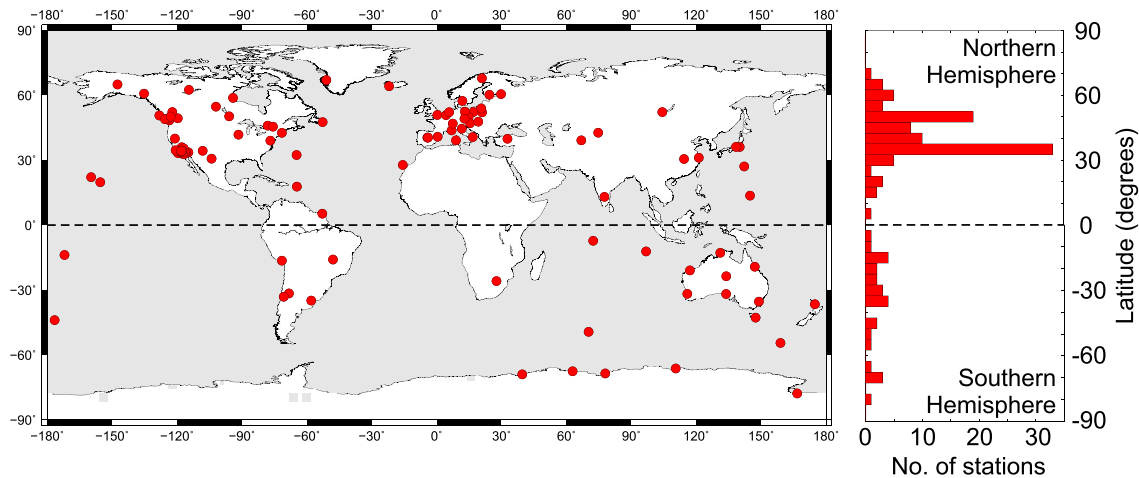
Figure 1. The sample of IGS Repro 1 stations with time series starting in January 1995 and with data gaps limited in time (see also Table S1 for more details).

A bug at the processing level was also noticed by Parracho et al. (2018) which was the mixing of old operational results and reprocessed results on the data server at JPL. This bug concerned a number of stations for the period 2008–2009 for which an older mapping function was used. The impact is negligible at most stations except in Antarctica where it produced a significant bias.

Apart from the unfortunate gaps and the unwanted processing changes, the GPS IWV data were produced using state of the art processing and postprocessing methods and are of high quality.

2.1.2. ERA-Interim IWV Data Set

ERA-Interim (Dee et al., 2011), noted ERAI hereafter, is a global atmospheric reanalysis starting in 1979 and updated until the end of August 2019, but with a 2006 release of the data assimilation system (Dee & Uppala, 2009), not including ground-based GPS data. The system includes a four-dimensional variational analysis (4D-Var) with a 12-hr analysis window, giving a temporal resolution of 6 hr. The spatial resolution of the data set is approximately 80 km (0.75° × 0.75°) on 60 vertical levels from the surface up to 0.1 hPa. Because GPS antenna heights and surface heights in the reanalysis are not perfectly matched, the IWV estimates were adjusted for the height difference based on an empirical formulation as in Parracho et al. (2018). The model output is available every 6 hr, from which daily and monthly mean values are calculated.

2.2. ERAI-GPS IWV Differences

The daily IWV differences were computed from the 6-hourly IWV data and the monthly differences from the daily data as described in Parracho et al. (2018). Although for the large majority of the sites, the IGS repro 1 and ERA-interim IWV time series are highly correlated (see, e.g., Fig. 2 in Van Malderen et al., 2017, and Bock & Parracho, 2019), larger IWV differences are explained by increased representativeness errors, when GPS observations capture some small-scale variability that is not resolved by the reanalysis (Bock & Parracho, 2019). For a number of outlying cases (15 sites), their special topographic and climatic features strongly enhance the representativeness errors (e.g., steep topography, coastlines, and strong seasonal cycle in monsoon regions). However, we do not exclude those sites in our analysis here, because the aim is to assess the performance of different break detection methods under different boundary conditions.

To describe the ERAI-GPS IWV differences, we use the following model:

$$\Delta IWV(t) = a + b \cdot (t - t_0) + \sum_{i=1}^{m} c_i \cdot \cos(2 \cdot \pi \cdot i \cdot (t - t_0)) + s_i \cdot \sin(2 \cdot \pi \cdot i \cdot (t - t_0)) +$$
$$+ \sum_{j=1}^{n} o_j \cdot H(t - t_j) + \varepsilon_t, \tag{1}$$

where a is the intercept, b is the linear trend, c_i and s_i are the cosine and sine amplitudes, respectively, of seasonal changes from $i = 1$ to m. We model annual, semiannual, triannual, and quarterannual signals,
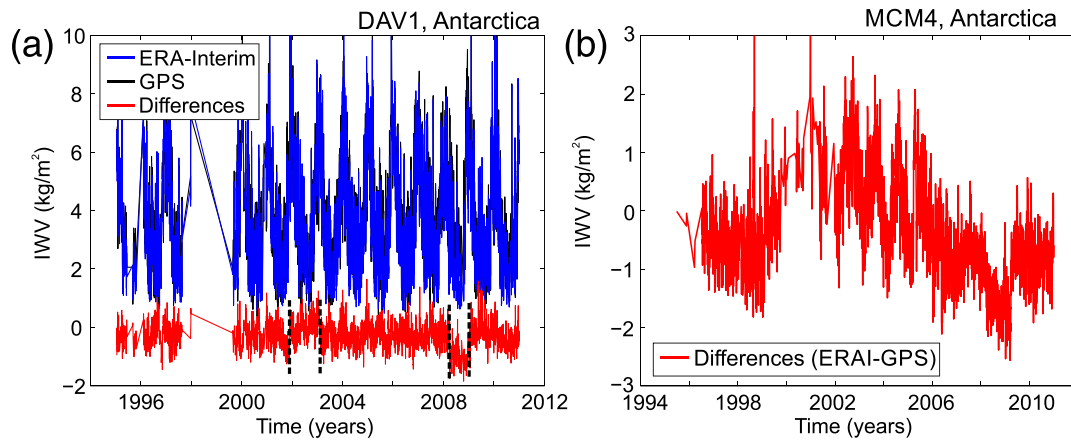
Figure 2. Examples of visual detection of breaks in the ERAI-GPS IWV difference series for (a) DAV1, where the vertical dashed lines denote breaks reported in the IGS site log, and for (b) MCM4, for which visual detection of breaks is not possible (also the case for GOPE: Czech Republic, KOUR: French Guiana, LONG: USA, not presented here).

that is, $m = 4$. Breaks found during a manual homogenization are modeled using a Heaviside step function $H(t)$ (Abramowitz & Stegun, 1972) with dates of breakpoints found in t_j, their magnitudes in o_j, and n the number of breaks. The stochastic part due to noise and errors is included in ε_t and is modeled as the sum of an autoregressive noise of the first order AR(1) and a white noise (WN). This model is inspired from the results of Klos et al. (2018), who analyzed the stochastic properties of GPS ZWD (zenith wet delay) data. Parameters a, b, c_i, s_i, and o_j are estimated in one single run using maximum likelihood estimation (MLE) in the Hector software (Bos et al., 2013). The results are provided in Table S1. The reasons for the different components included in this model are the following. Trend and seasonal signals in ΔIWVs are present due to representativeness differences between modeled and observed IWV estimates, model biases, and GPS biases (Bock & Parracho, 2019). Seasonal signals in ΔIWVs appear either due to a phase shift or a difference in the amplitude of the seasonal signal between the GPS and ERAI IWV time series. For some stations, trends in ΔIWVs larger than 1 kg/m^2/decade (or mm/decade) are present. For the stochastic part, we examined both the GPS and ERAI derived IWVs and noted that for both of them the combination of AR(1) with WN is optimal to describe the stochastic properties. The same tests were also performed for ΔIWVs, and we found no difference in the stochastic properties compared to the IWVs. We will now describe the properties of the different components of this model found for the ΔIWV estimates.

2.2.1. Positions and Magnitudes of Breaks

The generation of benchmark time series of IWV differences is a two-step process consisting of the production of homogeneous synthetic time series of differences and the addition of break positions with given magnitudes. To achieve this, an identification and characterization of the breaks is necessary, followed by the homogenization of the time series at the breaks (as in equation 1). To identify breaks in the real ERAI-GPS differences, we first analyzed the archived metadata of the IGS stations (the so-called site logs, publicly available at ftp://igs.ign.fr/pub/igs/igscb/station/log/ and ftp://igs.ign.fr/pub/igs/igscb/station/old-log/). These contain invaluable information about changes in equipment (R = receiver, A = antenna, D = radome), operating procedures, site conditions, etc. For the 120 IGS stations considered here, we found 1,028 dates of breakpoints reported in their site logs. Using a window of 30 days to merge consecutive dates of breakpoints into one event, resulted in a reduction to 970 dates. Of those 970 events, 177 dates were confirmed by a visual inspection of the ERAI-GPS IWV difference time series (see Figure 2a), and 50 more new dates were added, without reference in the IGS site logs. For four stations the visual detection of breaks is not reliable (see example in Figure 2b), as these sites have excessive (relative) standard deviation of IWV differences (see also Bock & Parracho, 2019). We hence end up with a list of 227 break positions (1.93 dates per station on average), which are inserted in equation 1 to adjust the model parameters. The distribution of the number of breaks per station is shown in Figure 3a, with most stations being affected by 0 to 5 breaks. The distribution of break magnitudes from the MLE estimation is shown in Figure 3b. It reveals that the ΔIWV inhomogeneities are in the range from −5 to +5 kg/m^2 (or mm), with most of the magnitudes
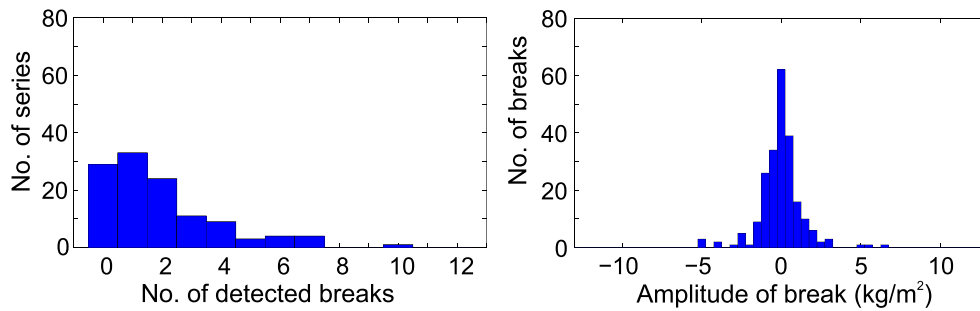
Figure 3. Distribution of the number of breaks detected in the 120 time series of ERA-GPS IWV differences (left) and the distribution of the magnitudes of the detected breaks (right).

between −1 and +1 kg/m^2. As a consequence, we will insert a similar number of known breaks into the homogeneous synthetic time series of IWV differences, with magnitudes in the same range as determined here.

2.2.2. Periodic Signals and Local Trend Estimates

The MLE results for the s_i and c_i coefficients of the harmonic series in equation 1 are shown in Figures S1 and S2. Only a few stations have annual amplitudes larger than 1 kg/m^2 in their IWV difference time series. For these stations, either a phase shift between ERAI and GPS IWVs is found (e.g., IISC) or a difference in the amplitude of the seasonal cycle is observed (e.g., KIT3 and POL2). The same stations have also been identified by Bock and Parracho (2019) as outlying sites due to representativeness differences between both IWV data sets. The mean amplitude of the annual signal is 0.38 kg/m^2, and this mean value is nearly halved for every subsequent harmonic (see Table S1).

Trend estimates for the IWV differences derived with the MLE algorithm vary between −0.13 and 0.15 kg/m^2/year, but with a mean value of 0.00 kg/m^2/year with standard deviation 0.05 kg/m^2/year. Stations with significant trend estimates might be ascribed to representativeness differences between the GPS IWV and ERAI. The amplitudes of the linear trend estimates are also provided in Table S1.

2.2.3. Stochastic Properties

Klos et al. (2018) found that a combination of AR(1) plus white noise (WN) provides a good stochastic representation of the ZWD time series of GPS stations. We adopt here this model for the IWV difference data set:

$$\varepsilon_t = v_t + w_t,$$ (2)

where v_t is an AR(1) process and w_t is a WN process. A first order autoregressive noise model is then described by the following discrete model:

$$v_t = \varphi \cdot v_{t-1} + z_t,$$ (3)

where the coefficient φ describes the degree of dependence of current value v_t on the previous value v_{t-1}, and z_t is a white Gaussian noise with zero mean and standard deviation σ_z.

Both the AR(1) and WN models are characterized by their standard deviations σ_v and σ_w. The MLE results of the stochastic model parameters are given in Table S1. For the analyzed set of IWV differences, the coefficients φ of the AR(1) noise model vary between 0.2 and 1.0, with more than half of the stations having a value less than 0.5. For the majority of the stations, the stochastic representation of the IWV differences is characterized by almost pure autoregressive processes, with fractions between 0.9 and 1.0 (see Figure S3), where the fraction of AR(1) noise is computed as its relative contribution to the AR(1) + WN combination, so that the sum of both fractions equals to 1.

2.3. Simulated Inhomogeneous IWV Differences

A synthetic data set of IWV differences is generated based on the parameters derived from the time series analysis done in the previous section 2.2. The IWV differences are simulated independently for each site using the parameters listed in Table S1. These 120 synthetic daily IWV differences ΔIWV will be used to
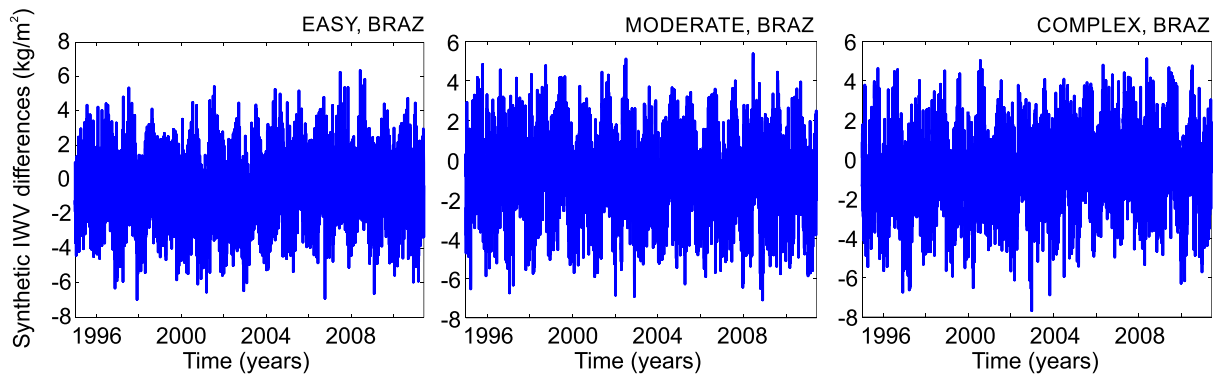
Figure 4. Three variants of the benchmark IWV differences data set are generated (easy, moderate, and complex) and presented here for BRAZ (Brasilia).

assess the skills of different break detection methods under different conditions (i.e., 120 different sites). Moreover, we introduce three test data sets of different complexity, which are named "easy," "moderate," and "complex" data sets and can be obtained at Kłos et al. (2020).

The "easy" data set includes the periodic signals, white noise, and breaks; the "moderate" data set adds a first order autoregressive process to the stochastic model, while the "complex" data set moreover adds gaps and local trends to the homogeneous series. The mean and standard deviation of the added trend slopes correspond to the empirical statistical properties of IWV difference series (see section 2.2.2). This last data set is closest to the real IWV differences but overshoots the percentage of gaps by 20% (due to a bug mentioned in section 2.1). The purpose of the complex experiment is to examine the performance of break detection methods in a more realistic case when the reference series are not homogeneous.

In all three data sets, the number of introduced breaks is chosen randomly from 0 to 5 per station, with magnitudes chosen randomly from -1 to $+1$ kg/m^2. For both, a uniform distribution is used. As such, these three variants of synthetic time series provide a total amount of 360 time series, which we use to test different break detection methods. An example of the three different variants for one station is given in Figure 4.

As a final check, we run our MLE analysis (see equation 1) on the synthetic time series to derive the coefficients of the deterministic and stochastic parts. The standard deviations of the noise models and trends derived from easy, moderate, and complex data sets agree with the real IWV differences (see Figure 5). Therefore, additionally, we can perform a preliminary analysis on the expected trend uncertainty from the real IWV differences, based on these simulations (see Figure 6). The easy data set is characterized by the smallest values of trend uncertainty, reaching only 0.02 kg/m^2/year at maximum. The largest trend uncertainty, as might be expected, is found for the complex data set, up to 0.04 kg/m^2/year.

3. Homogenization of Synthetic Benchmark Data Sets

Given that the benchmark data set consists of difference series of a candidate series and a reference series, the homogenization procedures here consist of only two steps, that is, break detection and adjustments for the detected inhomogeneities. In the first (larger) part of this section, the break detection methods are described, while the second part presents the common adjustment method with which all of the homogenization procedures are supplied.

3.1. Break Detection Methods

Eight different homogenization methods are used to detect breaks in the synthetic time series. All of them are applied on daily resolution data, and five of them are also applied on the time series of mean monthly values. The eight homogenization methods can be divided into four different types: (a) maximum likelihood (ML) multiple break methods (sections 3.1.1 and 3.1.2), (b) standard normal homogenization test-based methods (SNHT, sections 3.1.3, 3.1.4, and 3.1.5), (c) singular spectrum analysis (SSA, section 3.1.6) and (d) nonparametric (NP) tests (section 3.1.7). Table 1 gives an overview of the acronyms used to denote the
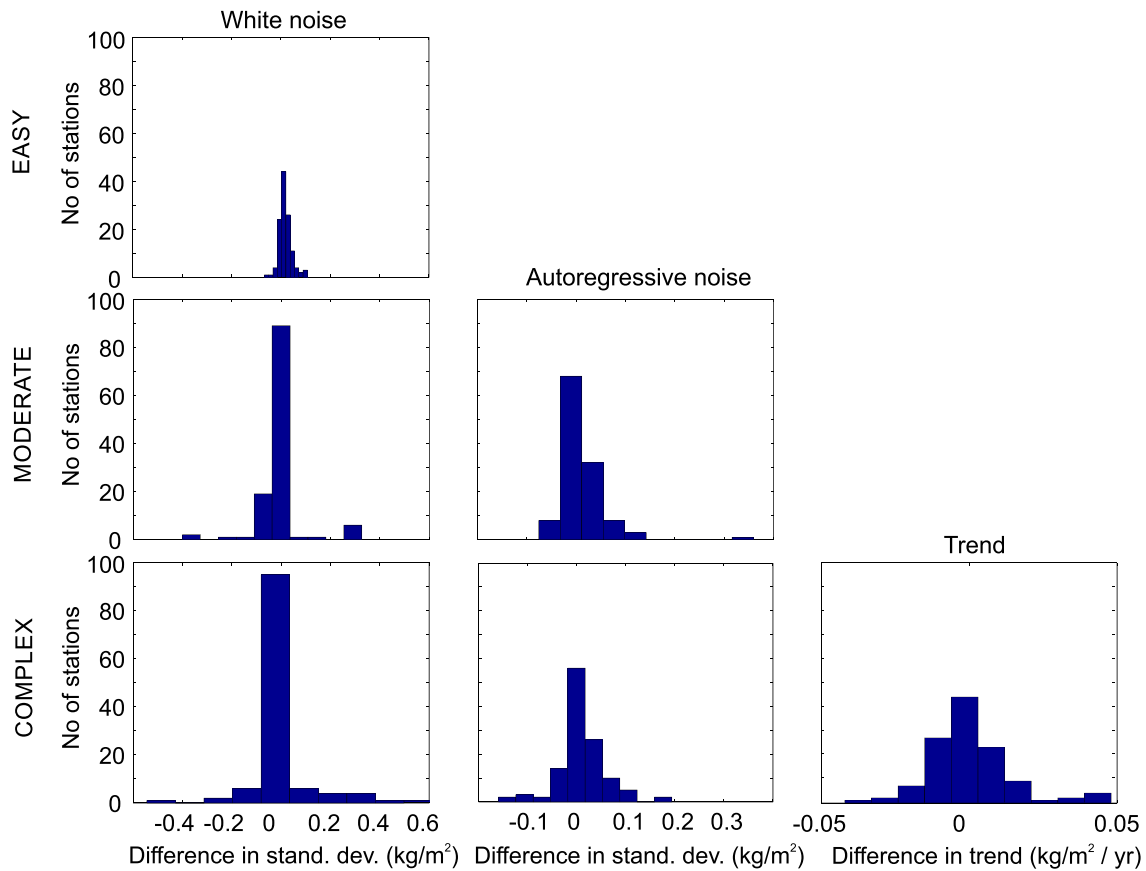
Figure 5. Verification of the synthetic benchmark time series of IWV differences. The estimates of differences between the real and synthetic data sets for noise model standard deviations and trends are provided for the three variants of the generated data sets.

different methods. All methods have been applied to the three different variants of the data sets (easy, moderate, and complex), except SN1d, SN1m, and SSAd which were not run on the easy data set.

3.1.1. Univariate ACMANT Detection (ML1)

ACMANT homogenization method has been developed from PRODIGE (Caussinus & Mestre, 2004) in the last decade (Domonkos, 2011a; Domonkos & Coll, 2017). Its principal break detection tool is a step function fitting with the Caussinus-Lyazrhi criterion (Caussinus & Lyazrhi, 1997). ACMANT detection has univariate and bivariate modes, and the univariate mode is applied in this study. In the first step of break detection, breakpoint positions are searched in the series of annual mean data, and the minimum distance between

Figure 6. Expected trend uncertainty estimations for the three variants of synthetic IWV difference time series.

Table 1
Acronyms For All Break Detection Methods Used in This Study

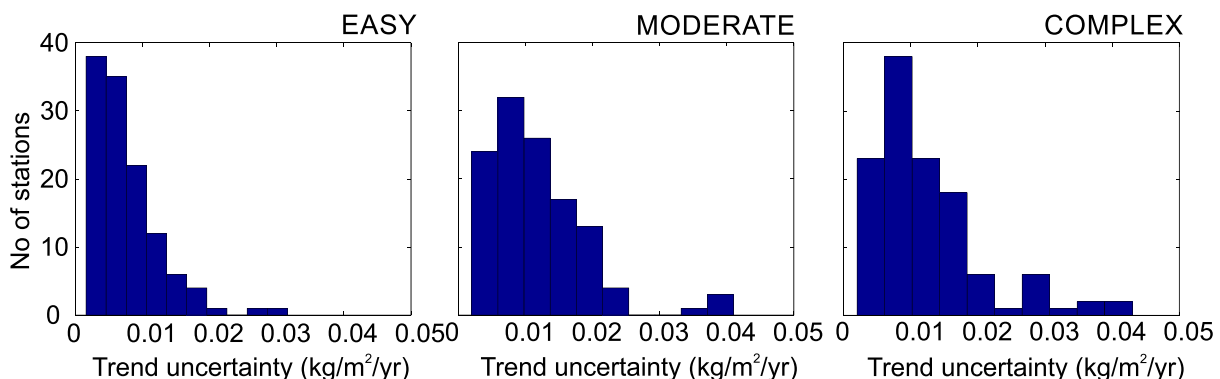| Acronym	Method	Section
ML1d	ACMANT-detection	3.1.1
ML2d	IGN (-BM2)	3.1.2
SN1d	Climatol daily	3.1.3
SN1m	Climatol monthly	3.1.3
SN2d	PMT-red daily	3.1.4
SN2m	PMT-red monthly	3.1.4
SN3d	*t* test daily	3.1.5
SN3m	*t* test monthly	3.1.5
SSAd	singular spectrum analysis	3.1.6
NP1d	CUSUM-Pettitt-Wilcoxon daily	3.1.7
NP1m	CUSUM-Pettitt-Wilcoxon monthly	3.1.7
NP2d	Pettitt test daily	3.1.7
NP2m	Pettitt test monthly	3.1.7

Note. The first two letters (three letters for SSA) indicate the group of detection method. ML = maximum likelihood multiple break methods; SN = standard normal homogeneity test based methods; SSA = singular spectrum analysis; NP = nonparametric tests. The last letter (d or m) is associated to the time series frequency (daily or monthly).

two adjacent breaks is 3 years. In further steps the breakpoint positions are refined (i) on monthly scale, using 48-month wide symmetric windows around the firstly estimated timings, (ii) on daily scale, using 4-month wide symmetric windows around the secondly estimated timings. After the refinement of breakpoint positions, the distance of adjacent breaks might be much shorter than 3 years. ACMANT detection includes also the search of 1–30 month long outlier periods for which the means significantly differ from the means in the 2-year periods before and after the outlier period (Domonkos, 2014). The parameterization of ACMANT detection applied in this study is the same as in ACMANTv3 (Domonkos & Coll, 2017).

ACMANT detection needs complete time series. In homogenizing networks, data gaps are infilled with spatial interpolation, while in the present exercise the annual mean values are put into the places of missing data.

3.1.2. IGN-AgroParisTech Method (ML2)

This method was developed in collaboration between IGN and AgroParisTech for the purpose of homogenization of daily differenced GNSS IWV series (candidate minus reference series). The segmentation algorithm is based on the model of a Gaussian random process with the unknown means and multiple breaks. The means and times of breakpoints are estimated using the maximum likelihood approach. This model was actually modified to account for specific characteristics of the GNSS-ERAI differences: a monthly varying variance (Bock et al., 2018) and a seasonal bias represented by a Fourier series of order 4. The final algorithm operates in several steps. First, the monthly variances are estimated using a robust estimator (Rousseeuw & Croux, 1993). Then, the least-squares fitting of the Fourier series and the segmentation algorithm are successively called in a loop for a fixed number of segments K. The separation of these two steps is required to allow us using the Dynamical Programming algorithm for solving the segmentation problem. This algorithm explores all the possibilities of K-1 breaks and retrieves the exact maximum likelihood solution in a reasonable amount of time. The final step is the model selection among all the number of segments that have been tested (typically we test from $K = 1$ to $K = 50$). A penalized maximum likelihood approach is used therefore and different penalty criteria are proposed: the Gaussian model selection by Birgé and Massart (2001; referred to as BM2 hereafter), the penalized contrasts by Lavielle (2005), and the modified Bayes information criterion (mBIC) by Zhang and Siegmund (2007). Most of the time, the three penalties select the same number of segments K, but in some cases, the results can differ, with some of them oversegmenting and some undersegmenting. In such cases, the user has to make the choice of the final solution. Therefore, a few useful diagnostics can be inspected such as the sum of squares as a function of K, and all the estimated parameters and their formal errors. Our segmentation method has been extensively tested and validated with simulations and is available as an R package referenced "GNSSseg" (https://github.com/arq16/GNSSseg). For the easy data set, the mBIC solution was selected, while for the moderate and complex data sets, the BM2 solution was selected because mBIC gave an unreasonably large number of breaks for these data (12.1 and 11.3, respectively). We believe this is due to the presence of autoregressive noise, while mBIC is based on an assumption of Gaussian white noise.

3.1.3. Climatol (SN1)

Climatol is a neighbor-based homogenization method, which performs a form of orthogonal regression known as reduced major axis (RMA, Leduc, 1987) between the standardized anomalies $(x\text{-}\mu_x)/\sigma_x$ and $(y\text{-}\mu_y)/\sigma_y$ of the two distributions. In its break detection segment, the standard normal homogeneity test (Alexandersson, 1986) is applied in two stages. The method incorporates a filling in of missing data and outlier removal. The detection of multiple breaks is done by applying the test to the remaining segments, after the series has been cut to two parts at a detected break (hierarchical detection with cutting algorithm). CLIMATOL can be applied to any time scale data, but it is advised to detect the breaks at the monthly scale, and then use the breakpoint dates to adjust the daily series. This method does not provide the magnitudes of breaks, as they are time varying. We might obtain the magnitudes by differencing the nonhomogenized series with the homogeneous series.

3.1.4. PMTred (SN2)

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 used for the break detection. 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. For the detection of multiple breaks, the cutting algorithm is included. The cutting algorithm stops when the length of the segment reaches the minimum of 2 years data, or when no more break can be found in the remaining segments of the time series.

3.1.5. Two-Sample *t* Test (SN3)

The procedure applied for the purpose of break detection is based on hypothesis testing. In this study, we use a test statistic that is the so-called "maximum type" (Jarušková, 1997). The theoretical concept of the method is covered in Csörgő and Horváth (1997) or Antoch et al. (2002). Within the field of mathematical statistics, the problem can be solved by testing the null hypothesis that there is no change in the distribution of the series, against the alternative hypothesis that the distribution of the series changed at time k. The null hypothesis is 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 (Yao & Davis, 1986). The method is applied to time series where the seasonality has been eliminated and the data gaps have been infilled before the break detection. The method is applicable on both monthly and daily time series. Another application and more detailed description of this method can be found in Eliaš et al. (2020).

3.1.6. Zero-Difference Approach (SSA)

This break detection method is used without reference time series as presented by Hoseini et al. (2019). It is based on the singular spectrum analysis (SSA), which has widely been used for time series trend extraction, noise filtering, and change detection (Alexandrov, 2008). The SSA technique is based on building a trajectory matrix with columns of lagged version of the time series, which can be used to reconstruct different components by selecting representative singular values. In this work, we use the SSA to extract the trend (smoothed version that contains long-term variations and not the linear trend), and we analyze the residuals to detect possible breaks. This method requires a continuous time series; therefore, when the data are missing, they are initially inferred using a model of the seasonal, annual, and semiannual components. Next, the SSA method is iteratively used to predict the missing data.

The homogeneity check starts with the SSA steps by forming the embedding step, extracting the uncorrelated components, and finding the singular values. The optimum singular values that are adequate to represent the trend for break detection are determined by minimizing a so-called change magnitude estimator (CME). A combination of singular values providing the minimum CME is used to extract the trend, and the minimum CME is set as a threshold at the break detection. We found that four singular values are the best choice for the trend extraction. When the trend and the threshold are known, we calculate the CME index for each point in the time series and indicate the ones with local maximum CME values as candidates for breaks. The final step is to classify correct detections, which is done by applying the *t* test to the candidate breaks. In addition, the following detected breaks are rejected: breaks within a period of inferred data (data gaps), changes due to meteorological and climatic variations that appear synchronously at multiple stations.

3.1.7. Nonparametric Tests (NP1 and NP2)

In this case, the used statistical tests are nonparametric distributional tests that utilize the ranks of the time series to find breaks (or more general to test the equality of the medians of two distributions). Because such tests are based on ranks, they 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 breaks in the middle of the time series, when a clear trend is present (Lanzante, 1996). We use two of such nonparametric tests: the Mann-Whitney (-Wilcoxon) test and the Pettitt (-Mann-Whitney) test (Pettitt, 1979). As an additional reference, the CUSUM test, based on the sum of the deviations from the mean, is also used. We developed an
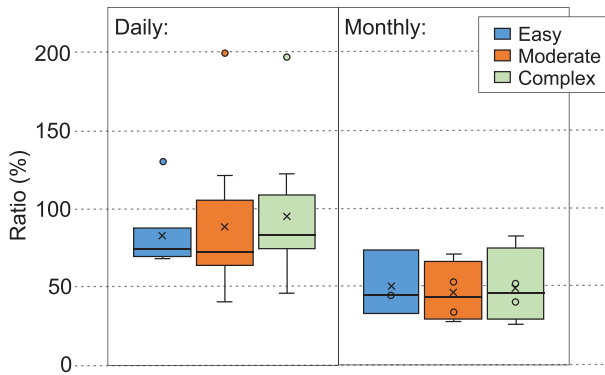
Figure 7. Boxplot of the ratio between the number of estimated breaks and the number of true (inserted) breaks. The three left boxes represent the results of daily methods, while the three right boxes represent the monthly method versions. The outlier dots are due to the excessively large number of estimated breaks with NP2d method.

iterative procedure to detect multiple breaks: If two out of those three tests identify a statistically significant break, the time series is adjusted (adjustment of the oldest segment with the detected magnitude of the break), and the three tests are applied again on the complete adjusted time series. These tests have been applied on both the monthly and daily values.

3.2. Adjustment for Detected Inhomogeneities

The adjustments of the IWV difference time series ΔIWV are based on the lists of the breakpoint positions provided by the different break detection methods. In particular, we shifted each break free segment, defined by two adjacent breaks (or an endpoint of time series and the nearest break), so that its mean value matches the mean value of the last break free segment of the ΔIWV time series. This is a very robust method, independent of the size of the segment, signal-to-noise ratio of the segment, season, month, etc. A more accurate way to compute adjustments is to consider all adjustments simultaneously, called "joint correction" (Venema et al., 2012). And although some of the described methods (e.g., Climatol) provide adjusted time series in a different way as an output of their software packages, we apply our robust, uniform adjustment method to the mean of the most recent segment here, since the focus of this paper is on the break detection itself.

4. Performance of Break Detection Methods

Different error metrics can be used to assess the performance of break detection methods, depending on the target application. If, for example, the aim of the homogenization is to check the accuracy of the detected breakpoint positions in the time series (and contrast them with metadata information on instrumental changes), the detection skills are an important measure. Alternatively, if the homogenized time series are to be used for trend analysis, the effect of the homogenization and adjustments on the trends is crucial. In the latter case, the root mean square error and trend differences are the widely used error metrics. Hereafter, we will consider various types of assessment methods and error metrics.

4.1. Accuracy of Breakpoint Positions
4.1.1. Number of Detected Breaks
At first, we consider the ratio between the number of estimated breaks by each statistical break detection method and the number of breaks truly inserted into the synthetic data sets. It should be noted that at this point, we do not make a distinction between true and false detections. Figure 7 shows a boxplot of these ratios for each variant of the synthetic data sets, grouped separately for the daily and monthly break detection methods. A first observation to be made is that most break detection methods found less breaks than there were actually inserted, except for the two nonparametric methods when they are applied on daily time series (NP1d and NP2d). So, the statistical threshold criteria used in most methods seem to be rather conservative. In general, methods applied on the daily time series detected a higher number of breaks (around 75%) than when they are applied on data of monthly resolution (around 50%), but the variability of these ratios between the different methods is also higher for the daily method versions. In general, the number of estimated breaks is also increasing with the complexity of the synthetic time series for the daily method versions, while it is more stable for the monthly versions.

4.1.2. Time Window for Breakpoint Identification
Statistical detection methods will rarely detect a break at the exact time of a true break. In addition, when identifying a breakpoint position, not all detection methods provide a confidence interval for the detection result. Consequently, to evaluate the detection power consistently for the different methods, a proper, fixed time window has to be (arbitrary) set. For instance, given a time window of length 2 times N around true break t_0, a break reported by a detection method at t_e will be considered a correct detection if t_e falls within the interval $[t_0 - N, t_0 + N]$. We set the default N to 62 days (~2 months) for detection on both daily and monthly values, although $N = 183$ days (i.e., 6 months) is also used in a few examples to examine the effect of N on the break detection results.
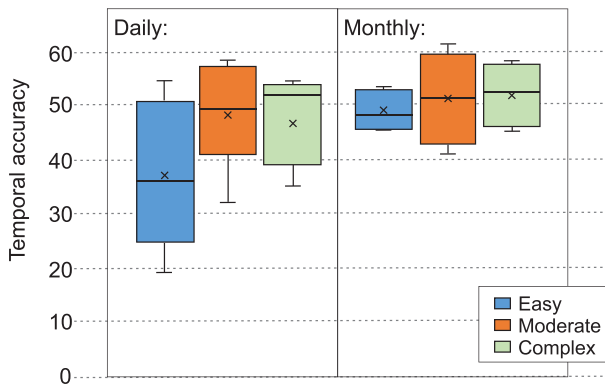
Figure 8. Boxplot of the mean differences between the estimated and true breakpoint positions for all results grouped by data set complexity (easy, moderate, and complex) and the time resolution in the break detection method (here we use $N = 183$ days as [conservative] upper limit for the time difference between an estimated and true breakpoint position for successful breakpoint identification).

For all break detection results, the mean difference between the estimated and true breakpoint positions varies between 19 and 55 days, 32 and 61 days, and 35 and 58 days for the easy, moderate, and complex data set, respectively (Figure 8). Although the temporal accuracy tends to decrease slightly with the increase of complexity of the synthetic data set, the experiments confirm the suitability of the proposed $N = 62$-day threshold. Therefore, in the text below, results are obtained with applying $N = 62$, except when explicitly mentioned otherwise. It should also be noted that the mean errors of breakpoint positions based on daily method versions are in line with those calculated with monthly versions, which indicates that—in general—there is no large difference in the temporal accuracy of break detection results according to the temporal resolution applied in the detection procedure. However, Figure 8 also shows that the best performing daily method version on the easy data set achieves a significantly better timing accuracy than the best performing monthly method version.

4.1.3. Skill Scores

When comparing true and estimated breaks, four categories can be applied (Menne & Williams, 2005; Venema et al., 2012): (1) The estimated break t_e falls within the interval $[t_0 - N, t_0 + N]$ (where t_0 represents the true breakpoint position), this is classified as true positive (TP) or hit. (2) The estimated breakpoint position does not fall in any of the intervals $[t_i - N, t_i + N]$ (where t_i is a true break), or if yes, the estimated break is not the closest detected break to t_i. This detection result is called a false positive (FP) or false alarm. (3) If no break is detected for a $2N + 1$-long section of the time series free of true breaks, this section is evaluated as a true negative (TN). (4) If no break is detected within the $2N + 1$ wide window of a true break, a false negative (FN) or miss is attributed.

Based on these four categories, it is clear that a statistical break detection method should maximize the number of TP (or hits) and TN, while keeping the number of FN (or misses) and FP (or false alarms) as low as possible (ideally zero). Using three of these categories, we can summarize the performance of the methods in the form of ternary graphs (see Figures 9 to 11) representing the respective percentage of FP, FN, and of TP as in Gazeaux et al. (2013). For instance, the TP percentage is defined as the number of TP multiplied by 100 and divided by the total number of FP, FN, and TP occurrences together. The measures on the other two sides of the triangle are defined with the same logic (see also Gazeaux et al., 2013). We delimit the zone of "good performance" in the diagram (represented in light green), defined by a percentage of FN and FP below 40%, and the percentage of TP above 40%. The diagram reads thus like this: The more the result of a statistical method is located to the lower right corner of the diagram, the better its performance.

When the break detection methods are applied over the easy data set with $N = 62$, only four results out of 10 (40%) are falling in the good performance zone (see Figure 9, left). These good scores are obtained by the ML1d, ML2d, SN2d, and SN3d methods all with the use of daily time resolution. Enlarging the time window to 183 days results in an increase of the number of TP and a decrease of the numbers of FP and FN for all methods and all variants of the synthetic data sets. In case of the easy data set and $N = 183$ days, 80% of the results fall into the green zone (see Figure 9, right). The addition of AR(1) to the WN model (in the moderate data set) has a negative impact on the detection power of all methods: All results show an increase in the number of false positives along with a decrease of the true positives (compare Figure 10 with Figure 9, left). In case of $N = 62$, only two results out of 13 (15%) remain in the good performance zone. They are obtained with the ML2d and SN3d methods. If we increase the time window to $N = 183$ days, six results fall into this zone, from which two are of monthly detection methods (NP1m and NP2m). Finally, adding gaps and trends to the synthetic time series (in the complex data set, Figure 11) affects even more the performance of the break detection methods, except for SSAd. The increase of FP and the decrease of TP are general in the complex experiment, and they are even more pronounced for the best performing methods (ML2 and SN3) for the easy and moderate experiments. In the complex experiment with $N = 62$ days, no result remains in the good performance zone, and even with $N = 183$ days only one result (of the ML2d method) is located in the green zone, near to the edge.
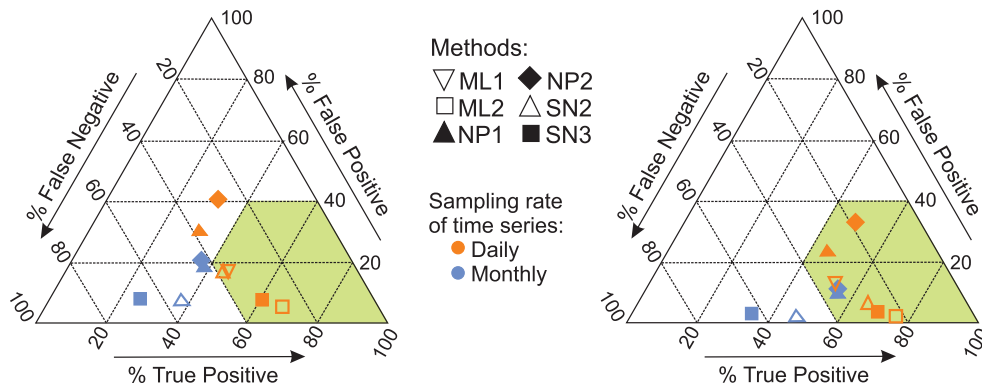
Figure 9. Ternary graph visualizing the break identification performance of the different methods for the easy data set. The performance of a method increases with decreasing number of false negatives and false positives and with increasing number of true positives. The perfect solution would be located in the lower right corner of the triangle. Following Gazeaux et al. (2013), a green zone is delimited, which represents the zone of "good performance." The different break detection methods are represented by different symbols, and the colors indicate the time resolution (daily or monthly) applied in the detection procedure (left: $N = 62$ days, right: $N = 183$ days).

As a continuation, we calculate the probabilities of (true) detection (POD, POD = TP/(TP + FN)), the probabilities of false detection (POFD, POFD = FP/(FP + TN)), and the Pierce skill scores (PSS, PSS = POD – POFD), see also Venema et al. (2012), to further assess the performance of the break detection methods. These values are shown in Table 2 for $N = 62$ days. POD values range from 0.24 to 0.69, 0.17 to 0.52, and 0.15 to 0.49 for, respectively, the easy, moderate, and complex data sets, and overall, they decrease with increasing complexity. The ranges for the POFD values are 0.010 to 0.13, 0.012 to 0.26, and 0.021 to 0.27 for, respectively, the easy, moderate, and complex data sets and increase with increasing complexity. The resulting PSS values are always positive and range from 0.22 to 0.68, 0.16 to 0.50, and 0.13 to 0.43 for, respectively, the easy, moderate, and complex data sets. All those values are in line with most of the POD, POFD, and PSS values reported by Venema et al. (2012) for monthly temperature and precipitation time series. The comparison of average daily and average monthly results shows that independently from the complexity of the data sets, generally higher PODs, POFDs, and PSS scores are obtained over the daily time series than over the monthly means: The better PODs over daily time series compensate for the worse POFDs to achieve higher PSS scores. However, this relation is not consistently true for the individual break detection methods. The best performance is achieved by two methods (ML2d and SN3d) when they were applied on data of daily time resolution.

4.2. Magnitudes of the Breaks

We apply the adjustment method explained in section 3.2 to estimate consistently the magnitude of the estimated breaks in the IWV difference time series ΔIWV by each method. The distribution of these magnitudes
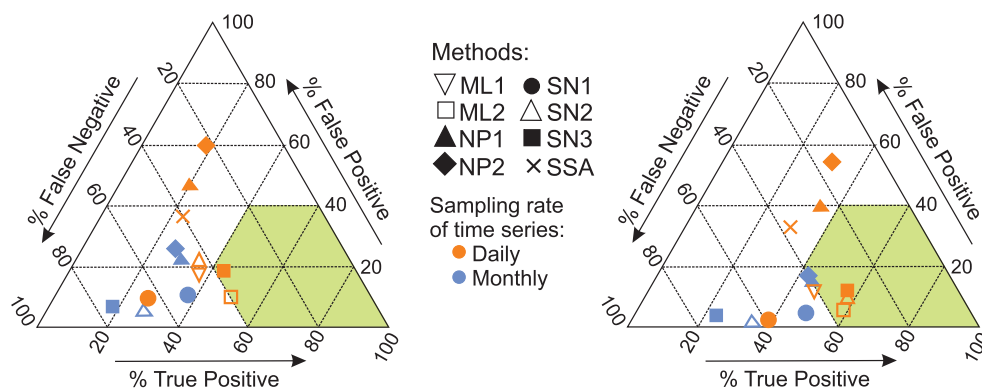

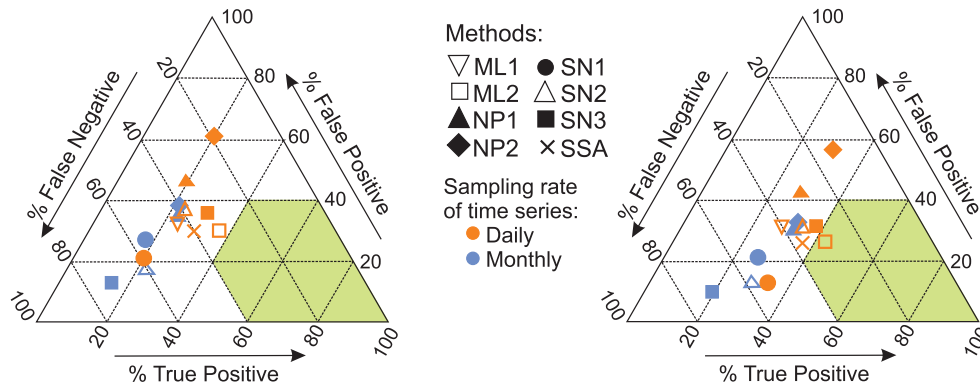**Figure 10.** Same as Figure 9, but for the moderate data set.

Figure 11. Same as Figure 9, but for the complex data set.

is shown in blue in Figures S4 to S6. As could be expected, most methods underestimate the frequency of the smallest break magnitudes. The two nonparametric methods on the other hand show a significant overestimation of break frequency in the magnitude range between 0.2 and 0.8 kg/m^2 but only when using data of daily time resolution. Finally, all the SNHT-based monthly methods (SN1m, SN2m, and SN3m) underestimate the number of breaks at all magnitude ranges.

4.3. Centered Root Mean Square Errors

Now we assess the performance of the different break detection methods by comparing the ΔIWV time series adjusted for the inhomogeneities (see section 3.2) with the original, homogeneous synthetic ΔIWV time series. Following Venema et al. (2012), we compute the centered root mean square error (CRMSE) to evaluate the difference of ΔIWV time series with the means removed, since we are interested in retrieving the correct variability rather than the absolute values:

$$CRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\bar{x})-(y_i-\bar{y})\right]^2}, \tag{4}$$

where x stands for the adjusted values, y for the correct values, upper stroke denotes the average for the whole period of the time series, and n is the number of values in the time series.

We evaluate the CRMSE of break detection methods both for daily and monthly data (CRMSEd and CRMSEm, respectively). Figure 12 shows that the CRMSE was reduced relative to the raw data CRMSE, calculated between the inhomogeneous and homogeneous benchmark data sets, by all of the tested methods.

Table 2
Probability of Detection (POD), Probability of False Detection (POFD), Pierce Skill Score (PSS) for All Methods and All Synthetic Data Sets

|	POD			POFD			PSS		
	EASY	*MODERATE*	*COMPLEX*	*EASY*	*MODERATE*	*COMPLEX*	*EASY*	*MODERATE*	*COMPLEX*
ML1d	0.485	0.396	0.322	0.033	0.038	0.075	0.452	0.358	0.246
ML2d	0.691	0.523	0.492	0.010	0.024	0.064	0.680	0.500	0.428
NP1d	0.409	0.356	0.363	0.075	0.144	0.156	0.334	0.212	0.207
NP1m	0.436	0.342	0.303	0.043	0.056	0.082	0.394	0.286	0.221
NP2d	0.485	0.430	0.457	0.132	0.260	0.269	0.353	0.170	0.188
NP2m	0.436	0.336	0.303	0.049	0.063	0.095	0.388	0.272	0.208
SN1d		0.258	0.233		0.025	0.042		0.233	0.191
SN1m		0.379	0.224		0.026	0.054		0.353	0.170
SN2d	0.488	0.409	0.328	0.043	0.054	0.093	0.445	0.355	0.235
SN2m	0.351	0.265	0.243	0.016	0.012	0.030	0.334	0.253	0.213
SN3d	0.625	0.500	0.435	0.017	0.047	0.094	0.609	0.453	0.341
SN3m	0.237	0.174	0.148	0.016	0.018	0.021	0.221	0.156	0.128
SSAd		0.339	0.353		0.094	0.082		0.245	0.271
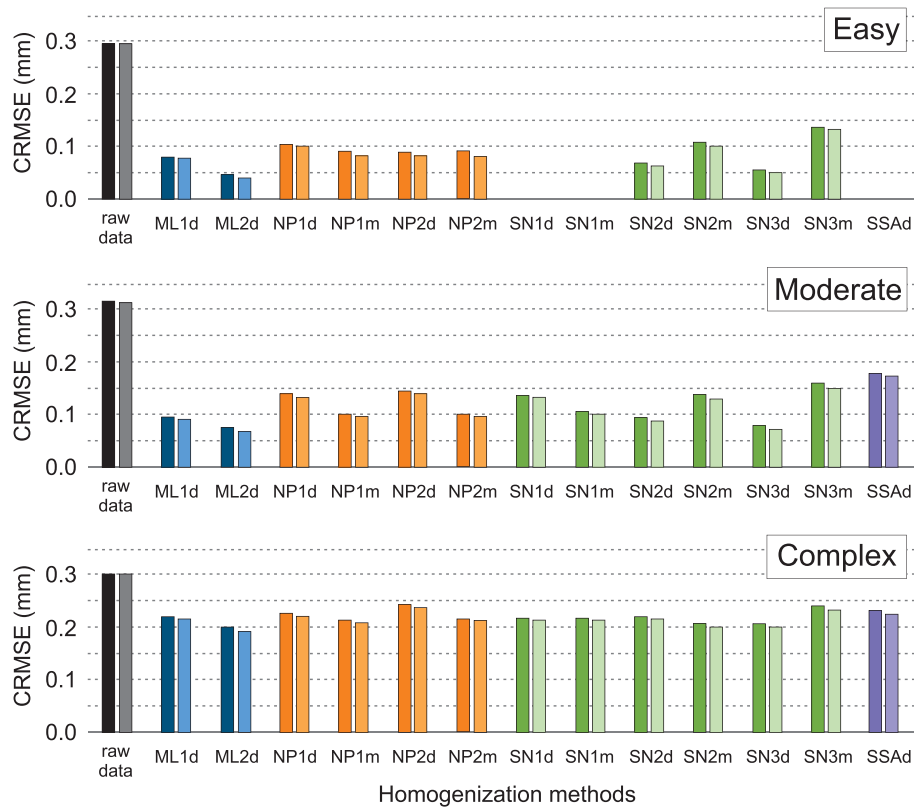
Figure 12. CRMSEd (dark colored bars) and CRMSEm (light colored bars), see text, for the different homogenization methods. The different colors denote the different types of the break detection methods: Maximum likelihood (ML) methods (blue), nonparametric (NP) methods (orange), SNHT-based (SN) methods (green), singular spectrum analysis (SSA, violet). The CRMSEs for the raw data are calculated from the differences between the inhomogeneous and homogeneous versions of the benchmark ΔIWV time series. (a) Easy data set (top), (b) moderate data set (middle), (c) complex data set (bottom).

The degree of the CRMSE reduction is high (roughly between 45% and 85%) in the experiments with the easy and moderate data sets, while it is considerably lower with the complex data set (between 19% and 35%). The CRMSEm is always smaller than the CRMSEd, but the difference rarely reaches 5% and never exceeds 10%. The degree of the CRMSE reduction strongly depends on the applied break detection methods in the easy and moderate experiments. The residual errors after homogenization are 2 and 3 times larger with the weakest methods than with the best methods.

The rank order of the break detection methods, from smallest to largest CRMSE, shows similar patterns for all experiments. For both the easy and moderate experiments, the two best methods are ML2d and SN3d, in this order, and they are followed by SN2d and ML1d. For the complex data set, ML2d, SN2m, and SN3d are the best. A further observation is that the difference of the results between the daily and monthly versions of the same homogenization method is often surprisingly large, for example, comparing the differences between SN3d and SN3m, as well as between SN2d and SN2m, especially for the experiments with easy and moderate data sets.

4.4. Trend Differences

As IWV time series are often used to calculate linear trends to assess the impact of the surface warming on the water vapor variability, we calculate the mean residual absolute bias of linear trends in homogenized ΔIWV time series. For this purpose, we fit linear regressions both to the homogenized series and to the original, homogeneous ΔIWV series. For these calculations, deseasonalized monthly ΔIWV values are used. The trend errors for the raw (inhomogeneous) series are also computed to assess the trend error reduction achieved by the homogenization.
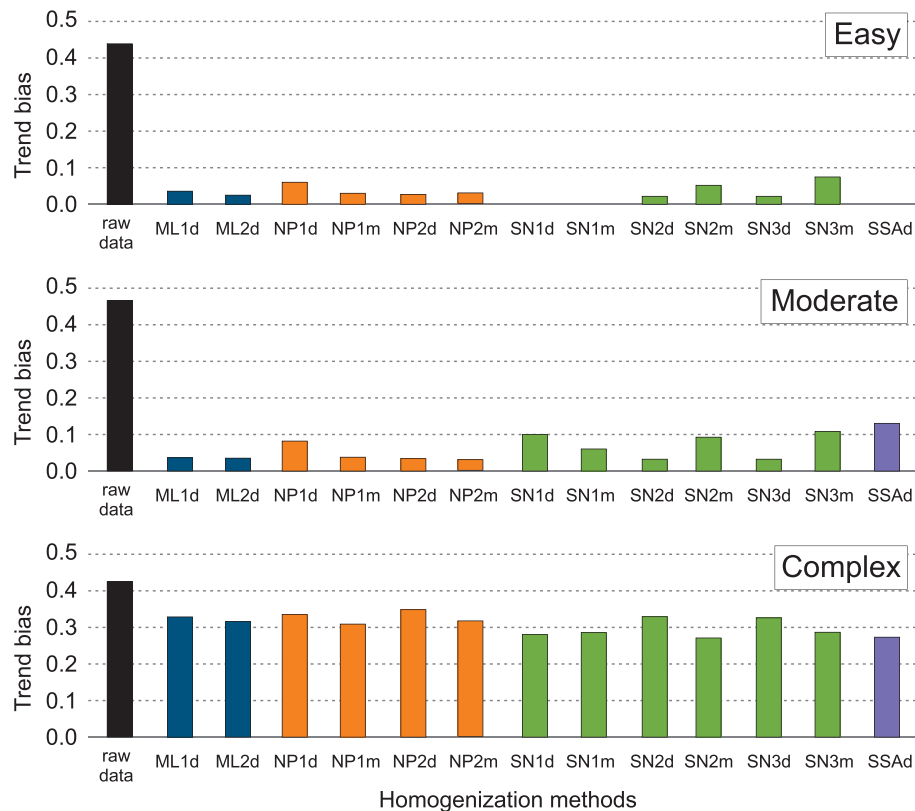
Figure 13. Mean absolute trend biases between the adjusted and homogenous synthetic ΔIWV time series, for the different break identification methods. The color coding of the different types of methods is identical as in Figure 12. (a) Easy data set (top), (b) moderate data set (middle), (c) complex data set (bottom).

First of all, it should be noted that the trend bias (mean absolute trend error) of the raw data is reduced in all experiments and by each of the applied break detection methods. For the easy and moderate variants, the ratio of trend bias reduction is even larger than that of the CRMSE reduction, and it often exceeds the 90% of the raw data error (Figure 13). However, in the complex experiments, the ratio of trend bias reduction is much smaller and practically the same as that of the CRMSE reduction.

For the easy and moderate variants, the relative trend differences between the different break detection methods are rather large, while the absolute trend differences between the different methods are small in comparison with the raw data trend bias. In the complex experiments, the differences of the residual trend biases between the break detection methods are moderate, that is, the error reduction ranges between 17% and 36% of the raw data trend bias.

In the easy and moderate experiments, the methods SN2d, SN3d, and ML2d gave the best results with almost the same efficiency, and they are followed by NP2m and NP2d. However, in the complex experiments, the rank order of break detection methods in terms of trend bias differences is completely different. In this case, SN2m, SSAd, SN1d, and SN1m yield the smallest residual trend biases in this order. A peculiarity of our results is that the three best methods for the complex experiments give the largest residual trend biases in the moderate experiments.

5. Discussion and Conclusions

As opposed to previous studies assessing the performance of break detection methods on benchmark time series (e.g., Venema et al., 2012), the analysis presented here is atypical in many aspects. First, although the homogenization of IWV time series with statistical break detection methods has already been presented in earlier studies (Ning et al., 2016; Schröder et al., 2016; Schröder et al., 2019; Vey et al., 2009), the assessment of the performance of these methods has never been checked for IWV. Venema et al. (2012) showed

that the break detection methods perform differently for temperature and precipitation, so a satisfactory performance for IWV data has not been guaranteed. Because the performance of homogenization methods is highly dependent on the statistical properties of the data and algorithms are steadily improved or new algorithms are being developed, it was especially timely to conduct such a study for IWV data.

One major constraint of the analysis presented here is the lack of a good reference series. The statistical break detection in climatic time series needs the use of reference time series, which is most typically a linear combination of the observed series of nearby stations with similar climatological conditions. Unfortunately, the network of GNSS sites is often not dense enough for such an approach. Therefore, in this study, we have used ERA-Interim reanalysis IWV data extracted at the site locations. The drawbacks of this approach are representativeness differences between the two IWV data sets at a number of sites (Bock & Parracho, 2019) and the possible presence of inhomogeneities in the reference ERA-Interim data set as well (Ning et al., 2016; Schröder et al., 2016; Schröder et al., 2019). Representativeness differences might result in seasonality and trends in the difference time series, while remaining inhomogeneities in ERA-Interim, even after 4-D variational bias correction in the data assimilation, might be responsible for trends in the differences time series. Especially those trends seem to complicate the performance of statistical break detection methods, but the reduction of raw data error still can be expected with statistical homogenization, as this is shown by our results. Note that the incorporation of metadata on GPS instrumental changes could improve considerably the efficiency, and the kind of the selected statistical method seems to be of secondary importance, as far as an adequately tested method is chosen.

Given those constraints and differences with previous evaluations of break detection methods on benchmark time series (e.g., Venema et al., 2012), we found here that the tested methods perform very well in detecting the positions of the inserted breaks, especially for the easy (periodicity + white noise) and moderate variants (periodicity + white noise + autoregressive noise of the first order) of the synthetic time series. On the other hand, the number of false break detections is also rather high, but overall, the resulting skill scores of the majority of the methods lie in the range between the scores reached for temperature (those latter being higher) and precipitation (lower; Venema et al., 2012). Moreover, the improvements made by adjusting the time series around the found breakpoints positions are significant, both in terms of the centered RMSE and the trend errors.

For the complex data set that also includes data gaps and biased trends of the reference time series (ERAI series), the performance of the break detection methods decreases significantly, especially due to the high probabilities of false detection. However, after adjustments, the homogenization methods still led to an improvement of 15% to 35% with respect to the raw data, both in terms of RMSE and trend errors. As the percentage of gaps introduced in this experiment is too high compared to the real GPS IWV data sets, and the trend inhomogeneities of simulated reference series likely exceed the trend inhomogeneities of the true ERAI series, the true efficiency of IWV homogenization is likely higher than in our results with the complex data set. The results with the easy and moderate data sets are more indicative to the break detection efficiency of general homogenization tasks in dense and spatially highly correlated station networks than to the efficiency of IWV homogenization.

Whether the sharp decline of the performance from the moderate to complex experiments is due to the presence of trend-like inhomogeneities or the gaps in the time series cannot be investigated on the basis of our benchmark data set. However, Domonkos and Coll (2019) examined the effect of missing data ratio on the homogenization of temperature series with ACMANT by performing seven experiments, in which 18 series are left complete, while variable quantities (10%–70%) of the data of the other 140 series are removed. Their results show that the impact of gaps on the residual linear trend bias is generally small, although data set dependent (see Fig. 5e of that study). In case of 40% ratio of missing data, the residual error increases between 0% and 60%, in contrast with the 800% increase in our study between the moderate and complex experiments with ML1d (Figure 13). Even if the representativeness of those results is limited to the IWV homogenization, it is very unlikely that the missing data ratio has the most important contribution in the observed 100%–1,000% increase of the residual errors between the moderate and complex experiments.

When comparing the performance of the break detection methods with each other, we conclude that for most of the tests, the IGN-AgroParisTech maximum likelihood multiple break method (ML2d), as well as the SNHT methods SN3d and SN2d give the smallest residual errors for the easy and moderate

experiments, but in the complex experiments, the trend estimation is better with some other methods. Some earlier studies (e.g., Caussinus & Mestre, 2004; Szentimrey, 1999) suggested that the concerted detection of multiple breaks of time series is more efficient than the combination of a *t* test-based method with the cutting algorithm. However, some follow-up studies could not confirm this finding. For example, Menne and Williams (2005) reported no visible difference between the achieved accuracies when testing some widely used representatives of these two method families. Domonkos (2011b, 2013) tested 10 widely used break detection methods with a variety of simulated monthly temperature test data sets, and found that the Caussinus-Mestre multiple break method (Caussinus & Mestre, 2004) always resulted in smaller trend biases than the other methods, although the efficiency differences according to break detection methods were generally small. It should be noted that in the test data sets of Domonkos (2011b), the average break number per time series was higher (at least five per time series on average) than in our benchmark data set. In the present study, slightly higher average performance was found with multiple break methods, for the easy and moderate data sets. However, the overall results do not confirm either deny the advantage of using the multiple break detection technique. Our results suggest that the efficiencies within method families often have larger differences than those between the best representatives of multiple break methods and *t* test-based methods.

The difference of efficiencies between daily and monthly versions of the same method is often surprisingly high. For the methods of SN2 and SN3, the daily versions left much smaller residual CRMSE and trend bias than the monthly versions, which is likely related to the low detection power of the monthly versions. By contrast, for NP2, the results of monthly homogenization are the best, and the frequency of false detections is excessively high for the daily version. The degree of such differences is surprising, as the mathematical task of detecting 2 and 3 breaks per time series on average is the same if either around 200 monthly values or approximately 6,000 daily values are used, with the exception that the daily series have a significant autocorrelation in the moderate and complex experiments, and the monthly series have a higher signal to noise ratio. The likely reason of the large differences between the efficiencies of the daily version and monthly version of the same break detection method is the imperfectness or inconsequent application of parameterization or an error committed during the transition of time scales. Nevertheless, the clarification of these details would need further examinations.

This study focuses on the detection of breaks that affect the assessment of climatic trends and low frequency variability. Therefore, the break detection methods have been assessed on daily and monthly time series. We do not discuss the homogenization of subdaily data and climatic extremes, as these are difficult scientific problems, even for spatially denser climatic data sets than the IWV data. Using very high time frequency data series increases, for example, the noise of time series. Lindau and Venema (2018) showed that for a pairwise multiple breakpoint algorithm, the results for low signal-to-noise ratios (SNRs) do not differ much from random segmentations and that reliable break detection at low but realistic SNRs needs a new approach. However, a break identified by on the methods assessed here can be adjusted for in the individual observation series, and these homogenized individual data points can then be used for weather and climate extreme applications and assimilation into reanalysis products.

Based on the performance of the different methods on the benchmark time series, we might investigate the possibility to combine the different dates of break detections of different methods. Such combination might optimize the probability of correct detection and minimize the probability of false detection. When this strategy of combining the most performance break detection methods turns out to be successful, it can be applied on the real ERAI-GPS IWV differences. The combined statistical detection of breaks, together with the use of the available metadata information on GPS instrumental changes, will likely be a valid approach to the high-quality homogenization of the IGS repro 1 data set.

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs and mathematical tables*. Washington, D.C: U.S. Dept. of commerce, National Bureau of standards.

Alexandersson, H. (1986). A homogeneity test applied to precipitation data. *Journal of Climatology*, 6, 661–675.

Alexandrov T. A method of trend extraction using singular spectrum analysis. arXiv preprint arXiv:0804.3367. 2008.

Antoch, J., M. Hušková, And D. Jarušková (2002), *Off-line statistical process control, in multivariate Total quality control, foundations and recent advances*, edited by: Lauro, C., Antoch, J. and Vinzi, V. E. 87–124. Heidelberg: Physica.

Beirle, S., Lampel, J., Wang, Y., Mies, K., Dörner, S., Grossi, M., et al. (2018). The ESA GOME-Evolution "Climate" water vapor product: A homogenized time series of H2O columns from GOME, SCIAMACHY, and GOME-2. *Earth Syst. Sci. Data*, *10*, 449–468. https://doi.org/10.5194/essd-10-449-2018

Birgé, L., & Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, *3*, 203–268.

Bock, O. (2016) Global GPS IWV data at 120 stations of IGS permanent network v1, https://doi.org/10.14768/06337394-73a9-407c-9997-0e380dac5590

Bock, O. (2017) Global GPS IWV data at 120 stations of IGS permanent network v1.2, https://doi.org/10.14768/06337394-73a9-407c-9997-0e380dac5591

Bock, O., X. Collilieux, F. Guillamon, E. Lebarbier, and C. Pascal (2018), A breakpoint detection in the mean model with heterogeneous variance on fixed time-intervals, *Statistic and Computing*, June 24, 2018, arXiv:1806.09043

Bock, O., & Parracho, A. C. (2019). Consistency and representativeness of integrated water vapour from ground-based GPS observations and ERA-Interim reanalysis. *Atmos. Chem. Phys.*, *19*, 9453–9468. https://doi.org/10.5194/acp-19-9453-2019

Bos, M. S., Fernandes, R. M. S., Williams, S. D. P., & Bastos, L. (2013). Fast error analysis of continuous GNSS observations with missing data. *Journal of Geodesy*, *87*(4), 351–360. https://doi.org/10.1007/s00190-012-0605-0

Byun, S. H., and Bar-Sever, Y. E.: A new type of troposphere zenith path delay product of the international GNSS service, *Journal of Geodesy*, *83*, 367–373, doi: https://doi.org/10.1007/s00190-008-0288-8, 2009.

Caussinus, H., & Lyazrhi, F. (1997). Choosing a linear model with a random number of change-points and outliers. *Annals of the Institute of Statistical Mathematics*, *49*(4), 761–775.

Caussinus, H., & Mestre, O. (2004). Detection and correction of artificial shifts in climate series. *J. Roy. Stat. Soc. C*, *53*, 405–425. https://doi.org/10.1111/j.1467-9876.2004.05155.x

Csörgő, M., & Horváth, L. (1997). *Limit theorems of change-point analysis.* Chichester: Wiley.

Dee, D. P., & Uppala, S. (2009). Variational bias correction of satellite radiance data in the ERA-interim reanalysis. *Q.J.R. Meteorol. Soc.*, *135*, 1830–1841. https://doi.org/10.1002/qj.493

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. Roy. Meteor. Soc.*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Domonkos, P. (2011a). Adapted Caussinus-Mestre algorithm for networks of temperature series (ACMANT). *International Journal of Geosciences*, *2*, 293–309. https://doi.org/10.4236/ijg.2011.23032

Domonkos, P. (2011b). Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. *Theoretical and Applied Climatology*, *105*(3-4), 455–467. https://doi.org/10.1007/s00704-011-0399-7

Domonkos, P. (2013) Efficiencies of inhomogeneity-detection algorithms: Comparison of different detection methods and efficiency measures. *Journal of Climatology*, *2013*,pp15, doi:https://doi.org/10.1155/2013/390945.

Domonkos, P. (2014). The ACMANT2 Software Package. In M. Lakatos, T. Szentimrey, & A. Marton (Eds.), *Eighth seminar for homogenization and quality control in climatological databases and third conference on spatial interpolation techniques in climatology and meteorology*, (pp. 46–72). Geneva, Switzerland: WMO WCDMP-84.

Domonkos, P., & Coll, J. (2017). Homogenisation of temperature and precipitation time series with ACMANT3: Method description and efficiency tests. *Int. J. Climatol*, *37*, 1910–1921. https://doi.org/10.1002/joc.4822

Domonkos, P., & Coll, J. (2019). Impact of missing data on the efficiency of homogenisation: Experiments with ACMANTv3. *Theoretical and Applied Climatology*, *136*(1-2), 287–299. https://doi.org/10.1007/s00704-018-2488-3

Eliaš M, Jarušková D, Douša J: An assessment of method for changepoint detection applied in tropospheric parametertime series given from numerical weather model. *Acta Geodyn. Geomater*, *17*, No. 1 (197), 101–112, 2020, doi: https://doi.org/10.13168/AGG.2020.0007

Gazeaux, J., Williams, S., King, M., Bos, M., Dach, R., Deo, M., et al. (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. https://doi.org/10.1002/jgrb.50152

Guerova, G., Jones, J., Douša, J., Dick, G., de Haan, S., Pottiaux, E., et al. (2016). Review of the state of the art and future prospects of the ground-based GNSS meteorology in Europe. *Atmos. Meas. Tech.*, *9*, 5385–5406. https://doi.org/10.5194/amt-9-5385-2016

Hoseini, M., Alshawaf, F., Nahavandchi, H., Dick, G., & Wickert, J. (2019). Towards a zero-difference approach for homogenizing GNSS tropospheric products. *GPS Solutions*, *24*, 8.

Jarušková, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics*, *8*(5), 469–483.

Johansson, J. M. (1998). GPS antenna and site effects. In F. Brunner (Ed.), *Advances in Positioning and Reference Frames*, *International Association of Geodesy Symposia*, (Vol. 118, pp. 229–235). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-03714-0_37

Jones, J., Guerova, G., Douša, J., Dick, G., de Haan, S., Pottiaux, E., et al. (Eds.) (2019). Advanced GNSS Tropospheric Products for Monitoring Severe Weather Events and Climate, *COST action ES1206 final action dissemination report* (XXI 563 pp.). Cham, Switzerland: Springer International Publishing. eBook ISBN: 978–3–030-13901-8, Hardcover ISBN: 978–3–030-13900-1, https://doi.org/10.1007/978-3-030-13901-8

Klos, A., Hunegnaw, A., Teferle, F. N., Abraha, K. E., Ahmed, F., & Bogusz, J. (2018). Statistical significance of trend in zenith wet delay from re-processed GPS solutions. *GPS Solut.*, *22*(2), 1–12. https://doi.org/10.1007/s10291-018-0717-y

Klos, A., Pottiaux, E., & Van Malderen, R. (2020). Three variants of synthetic benchmarks time series of GPS and ERA-Interim IWV differences. figshare. Dataset. https://doi.org/10.6084/m9.figshare.11733615.v1

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. *International Journal of Climatology*, *16*, 1197–1226.

Larson, K. M., Braun, J. J., Small, E. E., Zavorotny, V. U., Gutmann, E. D., & Bilich, A. L. (2010). GPS multipath and its relation to near-surface soil moisture content. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, *3*, 91–99. https://doi.org/10.1109/JSTARS.2009.2033612

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing*, *85*, 1501–1510.

Leduc, D. J. (1987). A comparative analysis of the reduced major axis technique of fitting lines to bivariate data. *Canadian Journal of Forest Research*, *17*, 654–659.

Lindau, R., & Venema, V. K. C. (2018). The joint influence of break and noise variance on the break detection capability in time series homogenization. *Adv. Stat. Clim. Meteorol. Oceanogr.*, *4*, 1–18. https://doi.org/10.5194/ascmo-4-1-2018

Menne, M. J., & Williams, C. N. Jr. (2005). Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of Climate*, *18*, 4271–4286.

Ning, T., Wickert, J., Deng, Z., Heise, S., Dick, G., Vey, S., & Schöne, T. (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

Parracho, A. C., (2017) Study of trends and variability of atmospheric integrated water vapour with climate models and observations from global GNSS network, PhD report, Université Pierre et Marie Curie, Paris, France, http://www.theses.fr/2017PA066524

Parracho, A. C., Bock, O., & Bastin, S. (2018). Global IWV trends and variability in atmospheric reanalyses and GPS observations. *Atmos. Chem. Phys.*, *18*, 16213–16237. https://doi.org/10.5194/acp-18-16213-2018

Pettitt, A. N. (1979). A nonparametric approach to the change-point problem. *Applied Statistics*, *28*, 126–135. https://doi.org/10.2307/2346729

Pierdicca, N., Guerriero, L., Giusto, R., Broioni, M., & Egido, A. (2014). SAVERS: A simulator of GNSS reflections from bare and vegetated soils. *IEEE Transactions on Geoscience and Remote Sensing*, *52*, 6542–6554. https://doi.org/10.1109/TGRS.2013.2297572

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*, 1273–1283. https://doi.org/10.1080/01621459.1993.10476408

Schröder, M., Lockhoff, M., Forsythe, J. M., Cronk, H. Q., Vonder Haar, T. H., & Bennartz, R. (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

Schröder, M., Lockhoff, M., Shi, L., August, T., Bennartz, R., Brogniez, H., et al. (2019). The GEWEX water vapor assessment: Overview and introduction to results and recommendations. *Remote Sensing*, *11*, 251.

Soden, B. J., & Held, I. M. (2006). An assessment of climate feedbacks in coupled ocean-atmosphere models. *Journal of Climate*, *19*, 3354–3360. https://doi.org/10.1175/JCLI3799.1

Steigenberger, P., Tesmer, V., Krügel, M., Thaller, D., Schmid, R., Vey, S., & Rothacher, M. (2007). Comparisons of homogeneously reprocessed GPS and VLBI long time-series of troposphere zenith delays and gradients. *Journal of Geodesy*, *81*, 503–514. https://doi.org/10.1007/s00190-006-0124-y

Szentimrey, T. (1999). Multiple analysis of series for homogenization (MASH). In S. Szalai, T. Szentimrey, & C. S. Szinell (Eds.), *Proc 2ⁿᵈ seminar for homogenization of surface climatological data*, (Vol. 41, pp. 27–46). Geneva, Switzerland: WMO WCDMP.

Van Malderen, R., Pottiaux, E., Klos, A., Bock, O., Bogusz, J., Chimani, B., et al.: Homogenizing GPS integrated water vapour time series: methodology and benchmarking the algorithms on synthetic datasets, in *Proceedings of the ninth seminar for homogenization and quality control in climatological databases and fourth conference on spatial interpolation techniques in climatology and meteorology*, Budapest, Hungary, WMO, WCDMP-No. 845, edited by T. Szentimrey, M. Lakatos, L. Hoffmann, pp. 102–114 http://www.wmo.int/pages/prog/wcp/wcdmp/wcdmp_series/WCDMP_85.pdf, 2017.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., et al. (2012). Benchmarking monthly homogenization algorithms. *Climate of the Past*, *8*, 89–115. https://doi.org/10.5194/cp-8-89-2012

Venema, V., Trewin, B., Wang, X., Szentimrey, T., Lakatos, M., Aguilar, E., et al. (2018, November 6). Guidance on the homogenization of climate station data. https://doi.org/10.31223/osf.io/8qzrf

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

Vey, S., Dietrich, R., Rülke, A., Fritsche, M., Steigenberger, P., & Rothacher, M. (2010). Validation of precipitable water vapor within the NCEP/DOE reanalysis using global GPS observations from one decade. *Journal of Climate*, *23*, 1675–1695. https://doi.org/10.1175/2009JCLI2787.1

Wang, J., Dai, A., & Mears, C. (2016). Global water vapour trend from 1988 to 2011 and its diurnal asymmetry based on GPS, radiosonde, and microwave satellite measurements. *Journal of Climate*, *29*(14), 5205–5222. https://doi.org/10.1175/JCLI-D-15-0485.1

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

Yao, Y.-C., & Davis, R. A. (1986). The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *In: Sankhya: The Indian Journal of Statistics, Series A (1961–2002)*, *48*(3), 339–353.

Zhang, N. R., & Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, *63*(1), 22–32. https://doi.org/10.1111/j.1541-0420.2006.00662.x