Received: 9 February 2023 | Revised: 5 March 2024 | Accepted: 6 March 2024

DOI: 10.1002/joc.8441

International Journal of Climatology | RMetS

RESEARCH ARTICLE

A statistical method for the attribution of change-points in segmented Integrated Water Vapor difference time series

Khanh Ninh Nguyen[1,2] | **Olivier Bock**[1,2] | **Emilie Lebarbier**[3]

[1]Geodesy, Institut de Physique du Globe de Paris, CNRS, IGN, Université Paris Cité, Paris, France

[2]Geodesy, ENSG-Géomatique, IGN, Marne-la-Vallée, France

[3]Modal'X - UMR CNRS 9023, Université Paris Nanterre, Nanterre, 92000, France

Correspondence
Khanh Ninh Nguyen, IPGP Geodesy, 39 rue Hélène Brion, Paris, 75013, France.
Email: knguyen@ipgp.fr

Funding information
Centre National de la Recherche Scientifique; Labex MME-DII, Grant/Award Number: ANR11-LBX-0023-01; FP2M federation, Grant/Award Number: CNRS FR 2036

Abstract

Many segmentation or change-point detection methods for homogenizing climate time series compare candidate station data with reference data to eliminate common climate signals and more efficiently detect spurious, non-climatic changes. One drawback is that it is difficult to decide whether the detected change-point is due to the candidate series or to the reference. A so-called attribution procedure is typically applied in a post-processing step for each detected change-point. This article describes a new statistical method for the attribution of change-points detected in Global Navigation Satellite System (GNSS) minus reanalysis series of integrated water vapour. It requires at least one nearby station with similar GNSS and reanalysis data. Six series of differences are formed from the four base series (BS) and are tested for a significant jump at the time of the change-point detected in the candidate station. The six test results are analysed with a statistical predictive rule to attribute the change-point to one, or several, of the four BS. Original aspects of our method are: (1) the significance test, which is based on a generalized linear regression approach, taking both heteroscedasticity and autocorrelation into account; (2) the predictive rule, which uses a machine learning method and is constructed from the test results obtained with the real data by using a resampling strategy. Four popular machine learning methods have been compared using cross-validation and the best one was applied to a real data set (49 main stations with 114 change-points). The results depend on the choice of the test significance level and the aggregation method combining the prediction results when several nearby stations are available. We find that 62% of the change-points are attributed to GNSS, 19% to the reanalysis, and 10% are due to coincident detections.

KEYWORDS

change-point detection, generalized least squares, GNSS, homogenization, reanalysis, segmentation, supervised classification

1 | INTRODUCTION

Long records of climate observations are crucial for monitoring regional and global climate change and understanding the underlying climate processes (Dunn et al., 2021; Trenberth et al., 2007). However, many observational climate data are affected by inhomogeneities due to changes in instrumentation, in station location, in observation and processing methods, and/or in the measurement conditions around the station (Menne et al., 2009; Mitchell & Jones, 2005; Peterson, Vose, et al., 1998). Inhomogeneities often take the form of abrupt changes, which are detrimental to estimating trends and multi-scale climate variability (Easterling & Peterson, 1995; Jones et al., 1986). Various homogenization methods have been developed for the detection and the correction of inhomogeneities in the context of climate data analysis (Costa & Soares, 2009; Peterson, Easterling, et al., 1998; Reeves et al., 2007; Venema et al., 2012). Hereafter, we will refer to spurious (non-climatic) abrupt changes in the mean signal as 'change-points'. The change-point detection step, also called segmentation, can be performed in two classical ways, using either a statistical test (e.g., Alexandersson, 1986; Easterling & Peterson, 1995; Menne & Williams, 2005, 2009; Szentimrey, 2008; Wang et al., 2010) or a penalized likelihood approach (e.g., Caussinus & Mestre, 2004; Domonkos, 2011; Lu et al., 2010; Mestre et al., 2013). The former proceeds sequentially and detects one change-point at a time, which leads inevitably to a sub-optimal solution. On the other hand, the second approach estimates all the change-points at once, and is thus optimal or sub-optimal, depending on the search algorithm. When the whole parameter space is explored, such as with the dynamic programming algorithm, the method is optimal. Many climate segmentation methods are used on differenced data, where the target series is differenced with respect to a reference series. Using differenced series helps to remove the common climate signal and more efficiently detect spurious (non-climatic) changes. However, one drawback of this approach is that any detected change-point can be either due to the target series or to the reference series, if the latter is not homogeneous. In this so-called relative homogenization approach, the reference series has been traditionally constructed by compositing the series from several nearby stations (Alexandersson, 1986; Guijarro, 2011; Menne & Williams, 2005). Compositing relaxes the need for homogeneous reference series thanks to the averaging from many nearby stations, such that the detected change-points can be attributed with good confidence to the target series. Unfortunately, in practice, composited reference series often contain non-negligible inhomogeneities. Another approach based on the pairwise comparison of individual series has been shown to be an interesting alternative (Caussinus & Mestre, 2004; Domonkos et al., 2021; Menne & Williams, 2009; Mestre et al., 2013). In this approach, the change-points from the target and reference series are disentangled in a post-segmentation step, referred to as 'attribution'. In (Caussinus & Mestre, 2004), the attribution step is done manually, by using both statistical inference and historical information (station metadata) in an iterative way. In (Menne & Williams, 2009), an automatic procedure is proposed that attributes a change-point to the station with the highest overall count of detections. This method also uses station metadata when available. It assigns the detected change-points to the nearest known event from the station history within some confidence limit. In (Mestre et al., 2013), both a semi-automatic method similar to (Caussinus & Mestre, 2004) and a fully automatic method based on the joint detection of all series at once are implemented, but the latter is not a relative homogenization method and is thus not recommended (Domonkos, 2021).

The above-mentioned attribution methods generally require many nearby stations in order to find out which station is the cause of the detected change-point. They also operate in an iterative way, alternating the segmentation and attribution steps, and perform better when history information is included. In this work, we propose a new attribution method, which significantly relaxes these constraints. First, it works even if only one nearby station is available, which makes it usable in data sparse networks. Second, it operates in a post-processing mode, meaning that it uses as input the results from the segmentation step and does not need to iterate, although iterations may possibly help to make it more robust. Thirdly, it uses a predictive rule based on machine learning to attribute the cause of change-point among the tested series. The latter is trained in a preliminary step based on real data and is thus optimized for the specific data of interest.

We use integrated water vapour (IWV) derived from Global Navigation Satellite System (GNSS) measurements (Bock, 2019) and from the fifth ECMWF reanalysis (ERA5; Hersbach et al., 2020). Because the global GNSS data set is relatively sparse, the reanalysis is used as a reference to form the target minus reference difference series (Bock et al., 2019; Nguyen et al., 2021; Ning et al., 2016; Quarello et al., 2022; Van Malderen et al., 2020). The GNSS minus reanalysis data are segmented with the 'GNSSseg' method developed by (Quarello, 2020) which is based on a penalized likelihood approach. It detects abrupt changes in the mean in the presence of a periodic (seasonal) bias and a periodic variance (on a monthly basis). It is available in the R package GNSSseg (https://cran.r-project.org/web/packages/GNSSseg/index.html). It has been used in a benchmark exercise based on simulated data where it was ranked one of the best among eight segmentation tools

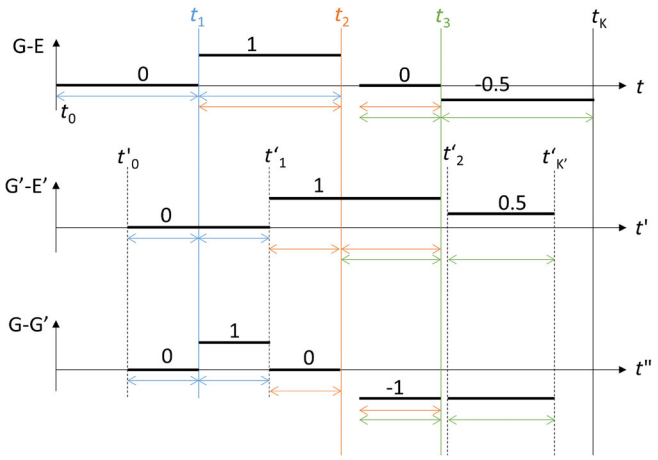

FIGURE 1 Schematic view of three paired series of differences, G-E, G'-E', and G-G', where G and E are the series from the main station, and G' and E' the series from the nearby station. Change-points detected by the segmentation method in the main (nearby) station are noted t_k (t'_k) and are indicated by the vertical solid (dotted) lines. By convention, t_0 (t'_0) and t_K ($t'_{K'}$) refer to the time of the first and last observation, respectively, in the main (nearby) station. The coloured horizontal lines with arrows indicate the segments on the left and the right of the change-points that are used to estimate the deterministic and stochastic parameters of the regression model. This figure is discussed the Introduction. [Colour figure can be viewed at wileyonlinelibrary.com]

(Van Malderen et al., 2020). The attribution step was not necessary in that study because the (simulated) reference series was homogeneous. With real data, the situation is different. Several past studies highlighted the presence of abrupt changes in the mean of GNSS series (Bock et al., 2014; Ning et al., 2016; Parracho et al., 2018; Vey et al., 2009) or in the reanalysis data (Nguyen et al., 2021; Ning et al., 2016; Parracho et al., 2018; Schroeder et al., 2016). Inhomogeneities in the GNSS series can be due to equipment changes, changes in the station's environment, or changes in the data processing procedure (Nguyen et al., 2021). Inhomogeneities in reanalyses can be due to changes in the global observing system, for example, the start or end of satellite missions (Rienecker et al., 2011; Schroeder et al., 2016). The goal of the attribution method is to determine whether a change-point detected by the segmentation method is due to GNSS or to the reanalysis.

Figure 1 helps to explain the idea of the attribution method proposed in this paper. Let us denote by G and E the GNSS and ERA5 reanalysis series of the main station, respectively, and G' and E' those from a nearby station. We denote by t_1, t_2, and t_3, the change-points detected by the segmentation method in the G-E series, and by t'_1 and t'_2, the change-points detected in the G'-E' series. These change-points have jumps in the mean of +1, −1, and −0.5 signal unit for the G-E series, and +1 and −0.5 signal unit for the G'-E' series. Note that in this sketch, the time period of the G'-E' series covers all the change-points of the main station, but in practice, several nearby stations may be necessary. The positions of the change-points illustrate different typical situations encountered in practice with our data. The first change-point in the nearby station, t'_1, is quite far from all the change-points detected in the main station. This illustrates the fact that the causes of inhomogeneities in GNSS data are primarily station-specific, that is, coincident change-points in G and G' are expected to be rare. On the other hand, t'_2 is close in time to t_3 which illustrates an inhomogeneity in the reanalysis data with a large spatial extension, that is, impacting both E and E'. Real data often contain data gaps which are due to instrumental failures leading eventually to an equipment change and possibly to an inhomogeneity. This situation is illustrated with a gap after t_2 in the G-E series. The likeliness of these different situations is summarized in the following 'empirical' rules which will help interpreting the features seen in the difference series:

R1. It is unlikely that change-points in two different GNSS series (here G and G') occur at the same time because they are station-specific in nature (e.g., hardware failure, equipment change, local environmental change).

R2. On the other hand, it is likely that change-points in the reanalysis occur simultaneously at the main and nearby sites (impacting E and E') because they are expected to have a large spatial extent (e.g., due to a change in assimilation of satellite measurements).

Inspection of the first two series of differences (SDs) in Figure 1 in the light of these rules suggests that t_1 is likely due to a +1 jump in G, t'_1 to a +1 jump in G', t_2 to a −1 jump in G, and t'_2 and t_3 to a −0.5 jump in both E and in E'. However, to confirm these guesses, we need to inspect additional SDs combining more of the four base series (BS, G, E, G', and E'). The lower plot in Figure 1 shows the G-G' series. It is straightforward, by the same reasoning, to confirm the guessed interpretation of the former two series. In a more general procedure, we would use all six combinations of the four BS and by deduce which of the four BS is/are the cause of the jumps observed in the multiple differenced series.

Table A1 presents all the relevant combinations of jumps/no jumps in the four BS and the corresponding jump level in the SD that we want to detect with the test. We distinguish two test results tables. The SD Table presents the true jump amplitudes, coded on five

values, and the reduced series of difference (RSD) Table, coded on three values, which is used in practice. Both tables contain duplicates highlighted by the coloured background. Duplicates can be separated with the help of joint probabilities given in the BS Table (see Supplement S1 for details on how they are derived). If the test procedure were perfect (no false alarm and no misses), any combination of six test results could be found in the RSD Table and the attribution method would simply consist in choosing the configuration with the highest probability in the BS Table. In practice, some of the test results may be wrong due to noise in the series, leading to configurations that are not be in the RSD Table. To overcome this difficulty, our attribution procedure builds on two main ideas. First, the test is built upon the generalized least squares (GLS) method which is known to have higher detection power than other traditional regression methods in the presence of heteroscedasticy and autocorrelation (see Supplement S2). Second, a statistical predictive rule is constructed using a machine learning algorithm and tests results from real data. This is an efficient way to predict the most likely solution when the combination of the six test results is not in the RSD Table.

Section 2 describes the stochastic properties of our data set composed of IWV data from GNSS data and ERA5 reanalysis, and highlights the embedded heteroscedasticity and autocorrelation in the SDs. Section 3 summarizes the test method based on GLS (equations and simulation results are presented in the Supplement S2) and presents test results from the real data. Section 4 describes the method for the construction of the predictive rule, compares the performance of four popular machine learning methods, and presents the attribution results from the real data. Section 5 discussed the results and concludes.

2 | DATA CHARACTERIZATION

2.1 | Data preparation

We use reprocessed GNSS tropospheric delay (ZTD) data from Center for Orbit Determination in Europe (CODE) and from Nevada Geodetic Laboratory (NGL). We thoroughly quality-checked and converted the ZTD data to IWV, following the procedure described in (Bock et al., 2021). The CODE data set comprises 436 stations from the International GNSS Service (IGS) from 1994 to 2014 (REPRO2015) which is extended to the present by a consistent post-processing (Dach et al., 2023). The NGL data set comprises more than 20,000 stations from 1994 to present (Blewitt et al., 2018). It is similarly based on a reprocessed data set (1994–2020) extended with a consistent operational data stream. The final IWV data of both data sets are distributed with daily and monthly time resolutions (Bock, 2022, 2023). This study uses daily IWV data.

We use the same 81 'main stations' as (Nguyen et al., 2021), which are from the CODE data set. Nearby stations were searched in the NGL data set, with a distance limit of 200 km in horizontal and 500 m in vertical. Data from the ERA5 reanalysis are extracted at the location of each GNSS station and the difference series, G-E and G'-E', are formed and segmented using the GNSSseg package (Quarello et al., 2022). The segmentation results are post-processed to remove clusters of change-points which occur occasionally in regions where the GNSS data and reanalysis data have a significant representativeness difference (Bock & Parracho, 2019). Change-points within clusters are either completely removed, or only one is kept (see an example in Figure 3 of [Quarello et al., 2022]).

Due to the scarcity of the global GNSS network and the imposed horizontal and vertical limits, only 49 main stations have at least one nearby station. Finally, only 114 change-points can be tested in the attribution procedure with the help of a total of 494 (main, change-point, nearby) triplets.

Each triplet associates a main station and a nearby station comprising four BS (G, E, G', and E') and six SDs (G-E, G-G', etc.). The IWV data from the nearby station (G' and E'), are adjusted for the difference in height with respect to the main station following the method described in (Bock et al., 2022), where model coefficients are estimated from ERA5. After the adjustment, each of the six series is cleaned for outliers, where data points exceeding three standard deviations from the median are removed. Data within the above-mentioned clusters are also removed in the corresponding (main or nearby) series. Additional data are removed in the nearby series only when a change-point detected in the nearby series (G'-E') is very close (e.g., <10 days) to a change-point in the main series (G-E). This case corresponds, for example, to configuration 11 in the BS Table (jump in E and E'). A window of 10 days is used to allow for the uncertainty in the timing of the change-point due to noise in the data. This case is illustrated in Figure 1 where the data between t_3 and t'_2 have been removed. The data gap in the G-E series just after t_2 is introduced to illustrate the case of screened data within a cluster. Note that the number of data points in a series combining the main and nearby sites (e.g., G-G') is always less than or equal to that of a collocated series (e.g., G-E or G'-E'). To keep a high detection power, we set a minimum number of 200 consecutive points on each side of a change-point.

2.2 | Data characterization

The GNSS minus reanalysis difference series show usually strong heteroscedasticity and periodic (seasonal) biases, along with weak autocorrelation (Quarello et al., 2022). In the following, a series is modelled using the following regression model:

$$z_t = \mu_L + \delta x_t + s_t + e_t, \quad (1)$$

where t refers to the time, μ_L is the mean of the signal on the left of the change-point, δ is the amplitude of the jump, x_t is a step function ($x_t=0$ if $t \le t_k$ and 1 if $t>t_k$, where t_k is the time of the change-point detected by the segmentation method), s_t is the Fourier series, and e_t is the noise term. For ease of notation, we use t as the time index, with $t=1,...,n$, but in reality the data may contain gaps and the time values are not consecutive. To account for this, t can be replaced by $t(i)$, with $i=1,...,n$. To account for both heteroscedasticity and autocorrelation, we follow (Pinheiro & Bates, 2000) and represent e_t as the product of two factors:

$$e_t = e_t^* \sigma_t, \quad (2)$$

where e_t^* represents a stationary autocorrelated process of unit variance and σ_t^2 is the time-varying variance of e_t, that is, $Var[e_t]=\sigma_t^2$. Preliminary investigation of our data showed that most of the time the noise model is well approximated by an AR(1). Other noise models such as MA(1), ARMA(1,1), and pure white noise occur sometimes. We tested also for higher order ARMA(p,q) models and they are very rare. We limit thus ourselves to the four possible ARMA(p,q) models, with $p,q \in \{0,1\}$. Recall that an ARMA(1,1) model writes (Shumway & Stoffer, 2017):

$$e_t^* = \phi e_{t-1}^* + \theta w_{t-1} + w_t, \quad (3)$$

where w_t is a Gaussian white noise. The noise model identification and parameter estimation methods are described in the next section. Note that other stochastic models including periodic variations in the mean, heteroscedasticity and autocorrelation have been proposed by (Lund et al., 1995).

Figure 2 shows an example of a time series (jagged grey curve), with the estimated Fourier series (smooth black curve), the estimated standard deviation (SD), $\widehat{\sigma}_t$ (black curve at bottom), and the regression residuals (jagged orange curve). The strong heteroscedasticity is obvious, and because it is not stationary, we used a moving window approach (similar to the outlier screening procedure described above) to estimate it.

Tables 1 and 2 summarize the characteristics of our data set in terms of heteroscedasticity and noise structure, respectively, for all six SD (G-E, G-G'...), for the main and all nearby stations. The results are sorted according to the distance between the main and the nearby stations (smaller or larger than 50 km). Regarding the heteroscedasticity, three groups can be identified when the distance between sites is small. The first group (G1) includes G-E and G'-E', that is, the series with collocated data, which have moderate mean SD of 0.7 kg m^{-2}. The second group (G2) includes E–E' and G-G', that is, the series comparing the same technique, which have the smallest mean SD (0.5 kg m^{-2}). The last group (G3) involves data from non-collocated data and mixed techniques, and gets the largest mean SD. As the distance increases, the mean SD of series involving different sites increases, as expected from increased representativeness differences. Another striking feature is that the half-range of the variation in SD is around 70% for all six series, indicating that heteroscedasticity is a strong feature in our data.

Figure 3 shows the distributions of the noise models and of the estimated coefficients for the six differences, again sorted according to the distance. The AR(1) model is the dominant model, with a proportion between 50% and 80%, independently of the distance, while the white noise model is extremely rare. The proportion of MA(1) and ARMA (1,1) depends on the distance and the series: ARMA(1,1) is dominant for the collocated series (similar to the noise group G1), as well as for the series comparing the same technique (group G2), when the distance is small. On the opposite, when the distance is large, MA(1) becomes more frequent, like for the series mixing techniques and sites (group G3). The increase of the distance does thus not only increase the variance of the noise but changes also its nature. Another interesting aspect is the values of the coefficients. For the AR(1), they are very similar (around 0.3) for all series, regardless of the distance. Similarly, for the MA(1), they are very similar (around 0.2). More surprisingly, the estimated coefficients of the ARMA(1,1) model for the non-collocated series depend somewhat on the distance, with the exception of E–E'. Values of $\widehat{\phi}$ and $\widehat{\theta}$ are around 0.6 and −0.3, respectively, for the collocated series and when the distance is small, and around 0.2 for both coefficients, when the distance is large. For E–E', the values are always around 0.2. The ARMA(1,1) models with coefficients of opposite sign found at short distance suggest that in these cases the noise is a mixture of AR(1) and white noise (Shumway & Stoffer, 2017). When the distance increases, the moving average part becomes more important, which may be interpreted as a spatial/temporal averaging of the variability in the difference

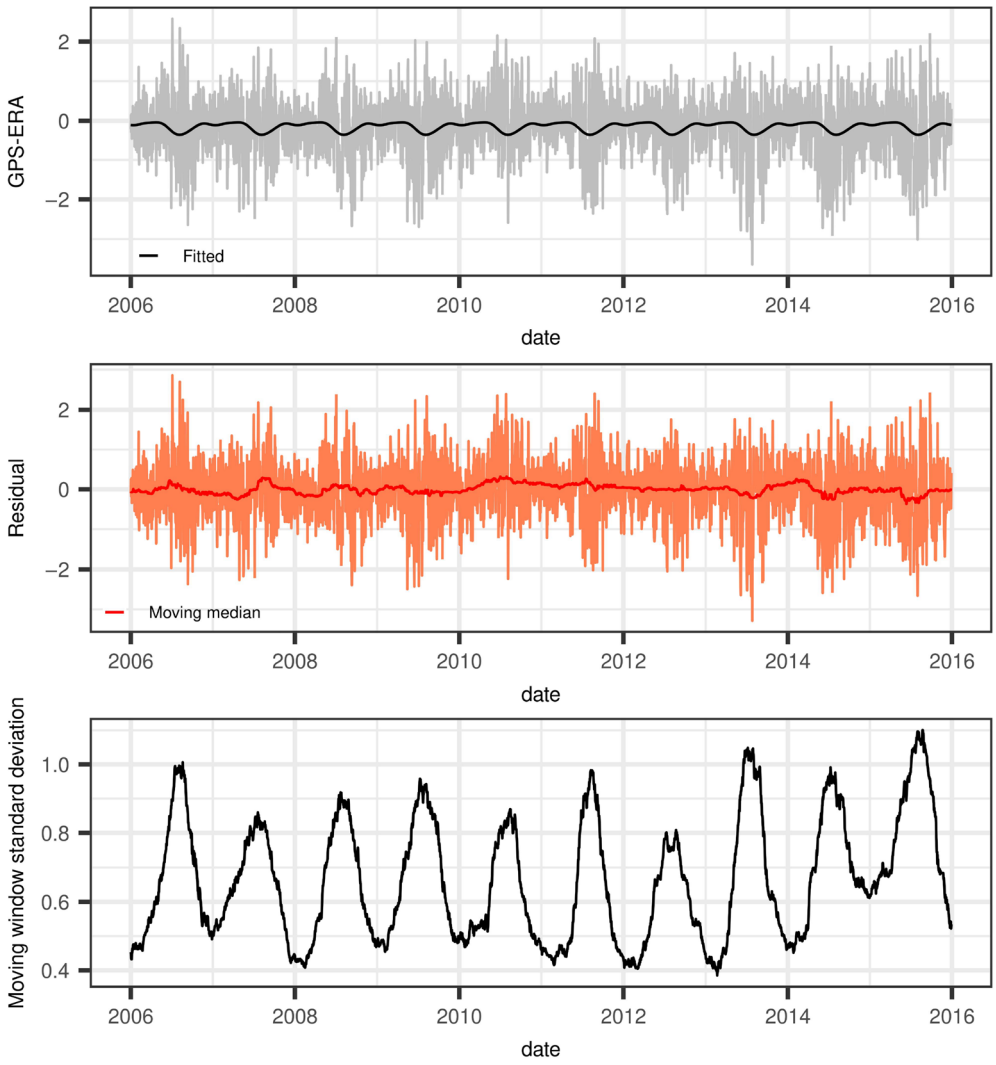

FIGURE 2 Top: Global Navigation Satellite System minus fifth ECMWF reanalysis time series at station ALBH (Victoria, Canada), in grey, and estimated Fourier series, in black, for a long, homogeneous, segment (no change-point detected by the segmentation method). Middle: FGLS regression residuals (jagged curve) and moving median (smooth curve). Bottom: moving standard deviation illustrating the strong heteroscedasticity in the data. [Colour figure can be viewed at wileyonlinelibrary.com]

series. The mean values of the estimated coefficients are reported in Table 2.

3 | PROPOSED TESTS FOR A FIXED CHANGE-POINT

This section presents the proposed test for the significance of a change-point (at a known position). This test is essentially a classical test in a regression model that takes into account the different characteristics of the data.

3.1 | Regression model and inference

The series of IWV differences is modelled using the following regression model:

$$\mathbf{z} = X\beta + \mathbf{e}, \tag{4}$$

where β includes the coefficients of the deterministic part of the model, $\beta = (\mu_L, \delta, a_1, ..., a_4, b_1, ..., b_4)'$, and X includes the corresponding regressors. Here, the a_l and b_l, $l = 1, ..., 4$, are the coefficients of a Fourier series of order 4, and the corresponding regressors are $\cos(2\pi l t(i)/T)$ and $\sin(2\pi l t(i)/T)$, with $T = 365$ days, and $t(i)$ is the time of the ith observation, z_i, $i = 1, ..., n$. The noise vector, $\mathbf{e}$ is assumed to be distributed as $\mathcal{N}(0, \Sigma_0)$, where Σ_0 is the variance–covariance matrix describing the noise model.

Testing the significance of the change-point simply amounts to testing the nullity of the jump δ using the classical test statistic $\tau_\delta = \widehat{\delta}/\widehat{\sigma}_{\widehat{\delta}}$, where $\widehat{\delta}$ is an estimator of δ and $\widehat{\sigma}_{\widehat{\delta}}$ its estimated standard deviation. A powerful test requires estimators with good properties to be considered. Since both the classical linear model assumptions (the independence and the homoscedasticity of the errors) are not satisfied and the covariance Σ_0 is unknown, we propose to use the well-known feasible GLSs (FGLS) method. This method, as well as the traditional methods, the GLS and the other approach OLS-HAC, are presented

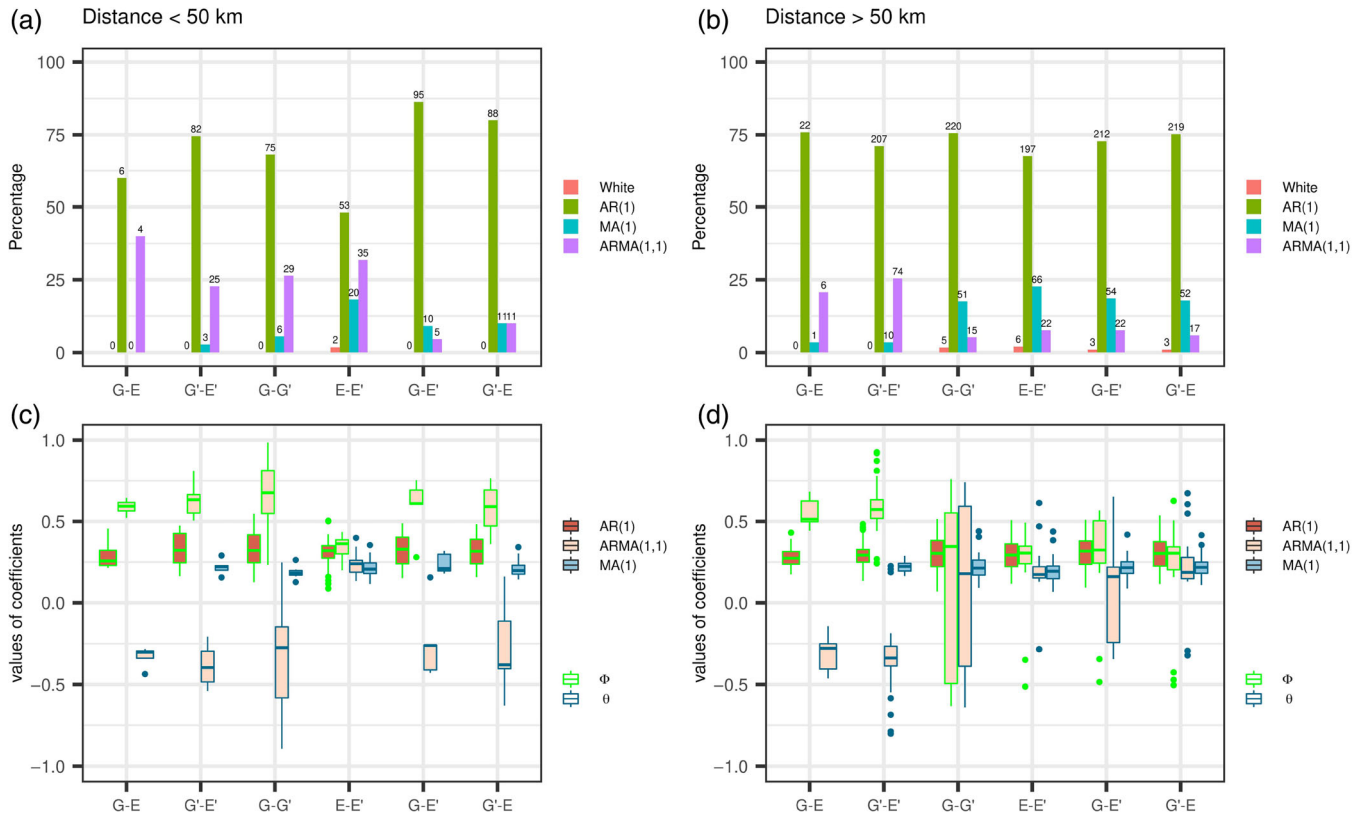

FIGURE 3 Results of noise model identification in the real data. (a,b) Histogram of model types (white noise, AR(1), MA(1), ARMA (1,1)) selected with auto.arima function for each of the six series of differences (G-E, G-E', etc.); the bar heights show the percentage (*y*-axis) of cases for each series, the number of cases is indicated on the top of each bar. (c,d) noise model coefficients, $\widehat{\phi}$ and $\widehat{\theta}$, estimated with arima function, for each model. Results are sorted according to the distance between the main and the nearby stations, (a,c) smaller than 50 km, (b,d) larger than 50 km. [Colour figure can be viewed at wileyonlinelibrary.com]

in Supplement S2 and are compared via a simulation study. This study shows that the FGLS outperforms its competitor, OLS-HAC, in terms of test power.

3.2 | Application to real data

The FGLS procedure was applied to the SDs from the 494 main/nearby pairs. Figure 4 shows the distribution of estimated jump amplitudes, their standard errors, and the associated absolute *t*-values computed from Equation (10), where the non-collocated series (G-G', G-E', E–E', and G'-E) are sorted by distance. Notably, the three series involving G have significantly larger median jump amplitudes (around 0.3 kg m^{-2}) than the other three series, regardless of the distance. This result suggests that large jumps are occurring more often in the G series than in the E, E', or G' series.

In G'-E' and G'-E, the median jump is small, as expected and expressed in our first rule stating that it is unlikely to have a coincident change-point in a nearby GNSS station when there is one detected in the main station. Additionally, a notable observation is that the median jump in E–E' is much larger at larger distance, which may be due to errors in the estimated jumps induced by an increased noise at larger distance. The variation of the SD of the noise with distance directly impacts also the jump standard error. The standard error of estimated jumps is notably smaller in collocated series, such as G-E and G'-E', as well as in non-collocated series at short distance. Furthermore, the standard errors in G-E' and G'-E (non-collocated series from different techniques) are slightly larger than in G-G' and E–E' (non-collocated series from the same technique) even at short distance, as also noticed in the three noise groups discussed in Section 2.2. Finally, the *t*-values can be interpreted by considering the jump magnitudes and their standard errors. It is evident that the three series involving G yield larger *t*-values due to higher jump magnitudes. In contrast, the other three series have much smaller *t*-values, mainly because some of the large jumps at larger distance are damped by the larger standard errors. A common feature to all non-collocated series is that the *t*-values decrease with distance.

TABLE 1 Characterization of the heteroskedasticity in the real data from 494 main/nearby series.

Series	Mean of SD distance < 50 km	Mean of SD distance ≥50 km	Half-range of SD (%)
G – E	0.7 ± 0.26		72 ± 20
G′ – E′	0.66 ± 0.24		67 ± 19
G – G′	0.52 ± 0.17	1.31 ± 0.47	63 ± 21
E – E′	0.41 ± 0.17	1.26 ± 0.47	73 ± 26
G – E′	0.82 ± 0.21	1.38 ± 0.46	67 ± 21
G′ – E	0.83 ± 0.26	1.39 ± 0.46	66 ± 20

Note: The table reports the mean and the half-range of the standard deviation for each of the six paired difference series. The mean values are sorted by distance.

TABLE 2 Characterization of the autoregressive noise structure of the real data.

Distance	<50 km				≥50 km			
Series	**AR(1)**	**Ma(1)**	**ARMA(1,1)**		**AR(1)**	**Ma(1)**	**ARMA(1,1)**	
Coefficients	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ
G-E	0.30	0.00	0.57	−0.32				
G′-E′	0.31	0.22	0.59	−0.34				
G-G′	0.33	0.19	0.65	−0.31	0.30	0.22	0.11	0.12
E–E′	0.31	0.21	0.34	0.23	0.29	0.20	0.25	0.20
G-E′	0.33	0.24	0.59	−0.24	0.31	0.21	0.29	0.08
G′-E	0.32	0.21	0.57	−0.28	0.30	0.22	0.18	0.21

Note: The table reports the mean estimated coefficients of the noise model for each of the six paired difference series, sorted by distance.

Figure 5 shows the corresponding test results with a significance level $\alpha=0.05$. At short distance, the three series involving G are almost all significant. Especially, all the G-E jumps are significant, which demonstrates a high consistency between our FGLS tests and the segmentation results. Almost all G-E′ jumps are significant as well, while almost all E–E′ are not significant. This latter result confirms our second rule (E and E′ are expected to be consistent, that is, either no jump or a jump in both, simultaneously). Most G-G′ jumps are also significant, which confirms the idea that most jumps are in G. Finally, the G′-E′ and G′-E jumps are most of the time insignificant, which again supports of our first rule (G and G′ are unlikely to change simultaneously). As the distance increases, the proportion of insignificant jumps also increases due to higher standard errors.

4 | PREDICTIVE RULE

The objective is to build a classifier $\psi(x)$, that predicts the configuration y, that is, the quadruplet composed of G, E, G′, and E′, given x, the vector composed of the test statistics from the SDs. In the development of this classifier, we are confronted with two principal challenges. First, the true configurations are unknown, resulting in the unavailability of y for training, evaluation and prediction. Second, the presence of all configurations in the data is nearly improbable due to the rarity of occurrence for certain configurations, as indicated by the probabilities in Table A1. To address these challenges, we propose to generate a synthetic dataset based on the $N=494$ test results of the real data using a bootstrapping technique. This dataset would ensure each configuration is represented through a set of (x,y) pairs. We then evaluated the performance of four popular classifiers on this dataset.

4.1 | Preliminary considerations

4.1.1 | Considered test results

In this task, we employed the test statistics from five series: G-G′, G-E′, E–E′, G′-E′, and G′-E′, due to the smaller sample size of the G-E test, which could lead to repetition in the synthetic dataset. We denote by $z_\ell=(z_{\ell 1},\ldots,z_{\ell 5})$ the vector of the five test statistics for the ℓth test in the five SD, and by $Z=(z_\ell)_{\ell=1,\ldots,N}$ the formed data set of size $N\times 5$, with $N=494$, which will be called the original data set in the sequel. Note that the results of all the nearby stations for a given change-point in a given main station can be viewed as replicates reducing the real information to

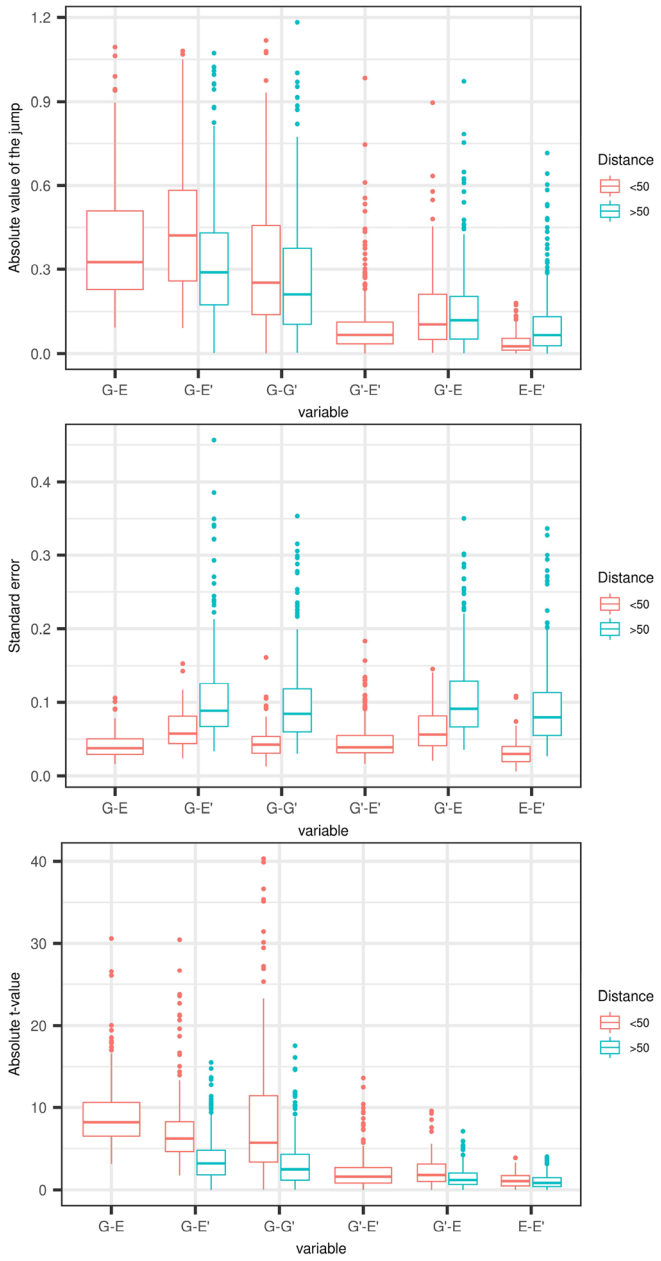

FIGURE 4 Distribution of absolute jump amplitudes and their standard errors, and the associated *t*-values computed from the feasible generalized least squares estimates of the real data (494 main/nearby pairs). The results are sorted based on the main/nearby distance (<50 km and ≥50 km). [Colour figure can be viewed at wileyonlinelibrary.com]

114, that is, the number of couples of (main station, change-point); the available information is thus quite small.

4.1.2 | Considered configurations

In Table A1, among the 38 configurations of the RSD Table, there are two doubles of the five coded test results with same prior probabilities: configurations (7,28) and (12,19). We decide to keep the configurations 7 and 19 which contain a change-point in G. This reduces the total number of configurations to $C=36$.

4.1.3 | The four considered learning algorithms

In this study, we considered four learning algorithms are the linear discriminant analysis (LDA; Fisher, 1936), the classification and regression trees (CART; Breiman et al., 1984), the Random Forest (RF; Ho, 1995) and the *k*-nearest neighbours (kNN; Cover & Hart, 1967). The latter three involve parameters that need to be tuned. They are here automatically optimized by *K*-fold cross-validation with $K=10$ using the generic function 'train' of the R package caret.

4.1.4 | Building of the complete synthetic data set

As previously mentioned, the bootstrapping technique has been employed to construct the synthetic dataset, operating on the principle of random sampling with replacement from the original data. More precisely, for each configuration y and each SD j, we create N_y vectors of the five test statistics or *t*-values (the sample x) by resampling among the *t*-values values $(z_{\ell j})_\ell$ that lead to the test conclusion of y. The correspondence is made with respect to the test outcome (-1,0 or 1) for a given significance level α. For instance, for configuration 1 in Table A1, each *t*-value is randomly selected from the respective SD, ensuring that the significance levels of these five *t*-values are (1,1,0,0,0). The constructed data set is noted $D=\{(y_\ell,x_\ell)_\ell\}_{\ell=1,\ldots,n}$ of size $n=\sum_y N_y$.

Addressing the potential severe imbalance among the configurations within the data is crucial. We consider two strategic approaches: the 'balanced sample case', in which we can use the same number of replicates $N_y=R$ for each configuration y; and the 'imbalanced sample case', in which the number of replicates for each configuration, N_y, is proportional to the prior probability of each configuration, p_y, given in Table A1, that is, $N_y=np_y$ where n is the total number of data.

4.2 | The proposed cross-validation bootstrap procedure

Cross-validation is a popular statistical technique to test a classifier. It involves splitting the data into two subsets:

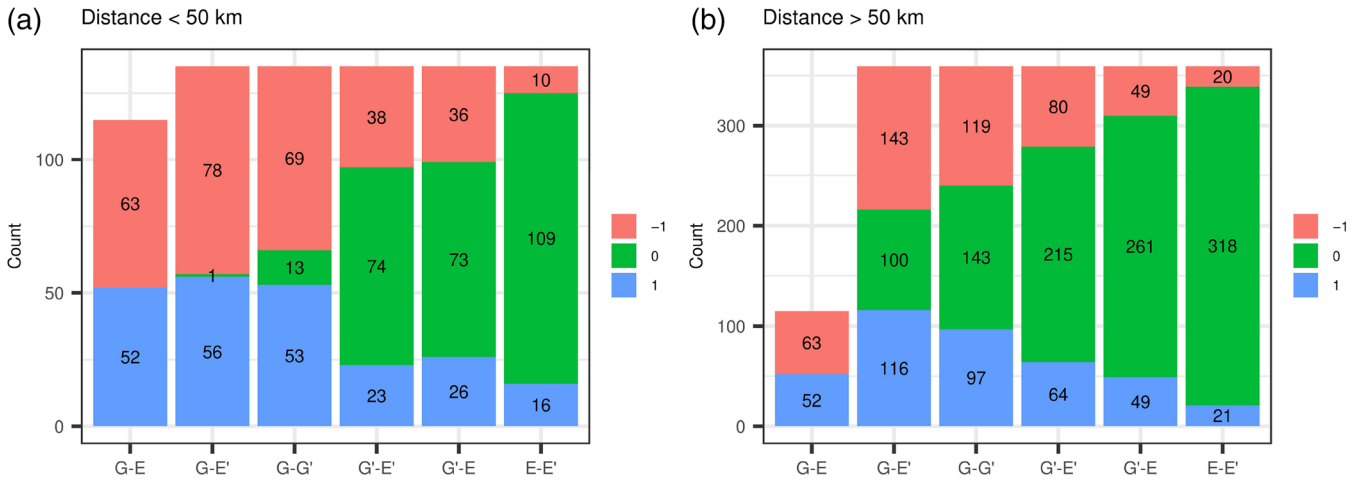

FIGURE 5 Distribution of test results associated to the estimated amplitudes of jumps shown in Figure 4, sorted by distance: (a) < 50 km, (b) ≥50 km. Test results are colour coded as: green for insignificant and red/blue for significant downward/upward jump. [Colour figure can be viewed at wileyonlinelibrary.com]

the learning set, on which the classifier is constructed, and a test set, on which the classifier is tested. Since observations of the complete data set D are replicated from the original data set Z, which is small and repeated, the learning and test data sets tend to overlap, inducing inevitably a bias and leading to an underestimation of misclassification error. This is why, we propose here a so-called cross-validation bootstrap (CVB) strategy which consists in first splitting the original data set Z into the learning and test subsets before constructing the complete data set D. The proposed CVB procedure is described in Algorithm 1.

Table 3 gives the mean misclassification error for the four considered classifiers with $B=20$. The table presents results for three scenarios: the first and second one involve constructing the complete dataset using different sampling (balanced vs. imbalanced), both with $\alpha=5\%$, while the last one employs a balanced sampling with $\alpha=1\%$. Compared with the balanced sample, the misclassification error is lower for the imbalanced sample in the case of LDA and KNN, but slightly higher for CART and RF. Similar behaviour is observed when comparing learning with $\alpha=5\%$ and $\alpha=1\%$ for the balanced sample. Overall, the Random Forest algorithm outperforms the other classifiers in all three scenarios, with the best performance achieved when trained with a balanced sample with $\alpha=5\%$. We thus choose the RF algorithm and select as the final predictive rule, $\widehat{\psi}$, the best one among the B resamplings. The predictive power of the five SD based on the accuracy criterion (the percentage of correct predictions) are in the decreasing order: E–E', G-G', G-E', G'-E and G'-E'.

4.3 | Application to the real data set

The objective is to predict the configuration for each change-point of every main station. When several nearby stations are available for a given change-point the results are aggregated using a weighted prediction score. For a configuration c, the prediction score writes:

$$\widehat{P}\left(y_{(\text{main, change-point})}=c|\text{nearby station}(ns)\right)=\frac{\sum_{ns} w_{ns}\mathbb{1}_{\{\widehat{\psi}(x_{ns})=c\}}}{\sum_{ns} w_{ns}},$$

where w_{ns} denotes the weight of the nearby station ns, and the final configuration is the one with the highest score

$$\widehat{y}_{(\text{main, change-point})}=\arg\max_{c}\widehat{P}(y(\text{main, change-point})=c|\text{nearby station}).$$

We compared two different weightings:

- inverse distance weighting: $w_{ns}=1/d_{ns}$, where d_{ns} is the distance between the nearby ns and the main station,
- weighting proportional to the joint probability given in Table A1: $w_{ns}=p_c$.

In the probability-based weighting, when the highest score is reached by two different configurations (e.g., $c=$ 1 and 10), the one with the shortest distance is selected.

ALGORITHM 1 The CVB procedure

Data: the original data Z.
for $b=1$ to B **do.**

1. sample a learning data set $Z^{b,L}$ from Z with probability 0.8, and form the test data set $Z^{b,T}$ with the remaining 20% of data. The random sampling is performed on the rows of Z, that is, on each test

2. form the two associated complete data sets $D^{b,L}$ and $D^{b,T}$ from $Z^{b,L}$ and $Z^{b,T}$ by preserving the learn/test proportion of 80%/20%, that is, for each configuration y, $D^{b,L}$ contains $0.8N_y$ samples, and $D^{b,T}$ $0.2N_y$. In the 'balanced sample case', we chose $N_y=R=100$ and in the 'imbalanced sample case', the smallest value N_y is chosen to 5, leading to a learn sample containing 4 data and a test sample containing only one data

3. construct the four classifiers on the learning data set $D^{b,L}$: $\psi^{b,k}$, $k \in \text{LDA, CART, RF, kNN}$

4. compute the misclassification error of the classifiers on the test data set $D^{b,T}$ with n_T rows

$$\text{err}^{b,k} = \sum_{\ell=1}^{n_T} \mathbb{1}_{\{\psi^{k,c}(x_\ell^{b,T}) \neq y_\ell^{b,T}\}} \qquad \text{for k} \in \text{LDA, CART, RF, kNN}$$

end

Averaging: compute the mean misclassification error for each classifier

$$\overline{\text{err}}^{k} = \frac{1}{B}\sum_{b=1}^{B} \text{err}^{b,k} \qquad \text{for } k \in \text{LDA, CART, RF, KNN}$$

Figure 6 presents the distribution of predicted configurations, after aggregation, for four variants: (a) balanced sampling with $\alpha=5\%$, aggregated with distance; (b) balanced sampling with $\alpha=1\%$, aggregated with distance; (c) imbalanced sampling with $\alpha=5\%$, aggregated with distance; and (d) balanced sampling with $\alpha=5\%$, aggregated according to prior probability. Across all figures, four predominant groups emerge consistently: G ($c=1$ and 15), (G, E, E') ($c=31$ and 35), (E, E') ($c=10$ and 23), and E ($c=8$ and 22). Remarkably, these configurations correspond to the highest joint probabilities, p, as indicated in Table A1: G and (E, E') with $p=0.18$, (G, E, E') with $p=0.04$, and E with $p=0.01$. This demonstrates that the classifier actually predicts the configurations which we believe are the most likely in the real data, even when these probabilities are not directly used in the procedure such as in variants (a) and (b).

In variant (a), 47 of the change-points (i.e., 41%) are attributed to group G and 29 (i.e., 25%) to group (G, E, E'), after the aggregation. Analysis of the six test results before and after the prediction helps to understand the relatively high frequency of these two groups. In general, the test results can be of two sorts: either the six results correspond to a configuration in Table A1, and in this case the predictive rule predicts the same result (as expected), or the result is initially not in the table, and the predictive rule will select a configuration that is 'close' to the initial configuration. Among all the test results going to group G, that is, (1,1,1,0,0,0) for $c=1$ and (−1,-1,-1,0,0,0) for $c=15$, about 75% are initially in the table. This high percentage is consistent with the observation that many jumps are significant in the first three tests and insignificant in the last three, as seen in Figures 4 and 5. The 25% of cases which are not initially in the Table A1 differ from these configurations by one or two elements, for example, case (1,0,1,0,0,0) differs from $c=1$ by only the 2nd element (the G-G' test). This case is then attributed to $c=1$ by the predictive rule when the absolute value of the t-value of the estimated jump in the G-G' series is close to the critical value, $\tau_{\alpha/2}=1.96$, in combination with smaller t values in E–E', G'-E', and G'-E. For group (G, E, E'), the percentage of cases that are not in the Table A1 is slightly more than 50%. Almost all these cases are either (1,1,1,0,1,0) or (−1,-1,-1,0,-1,0), which differ only by the 6th element (the G'-E test) from the final configurations $c=31$ (1,1,1,0,1,1) or $c=35$ (−1,-1,-1,0,-1,-1), respectively. Contrary to the G-G' series, the G'-E series has smaller t-values on average, hence the frequent 0 in the initial test results. The fact that almost all these cases are finally predicted as $c=31$ or 35 can be explained by the simultaneous occurrence of: high t-values in G-G' and G-E', a small t-value in E–E', a high

TABLE 3 Mean misclassification error ± one standard deviation, for the four classifiers in three scenarios: 'balanced sample' with $N_y = R = 100$ and $\alpha = 0.05$, 'balanced sample' and $\alpha = 0.01$, 'imbalanced sample' with $N_y = np_y$ and $\alpha = 0.05$.

c	Test level	Sample case	LDA	CART	KNN	RF
$\overline{err}^c$	0.05	Balanced	0.1463 ± 0.021	0.0142 ± 0.011	0.1412 ± 0.018	0.0049 ± 0.003
	0.05	Imbalanced	0.1108 ± 0.004	0.0165 ± 0.010	0.0351 ± 0.004	0.0054 ± 0.004
	0.01	Balanced	0.1424 ± 0.029	0.0210 ± 0.0417	0.1301 ± 0.022	0.0106 ± 0.033

Abbreviations: CART, classification and regression trees; KNN, *k*-nearest neighbours; LDA, linear discriminant analysis; RF, random forest.

t-value in G′-E′, and a value close to the critical value for G′-E. An example is provided in Figure 7. In this example, one may also suspect that the *t*-value of G′-E′ is excessively large, given the small value of the corresponding jump (−0.12) compared with the jumps in G-E, G-G′, and G-E′. One way to reduce the occurrence of excessively high *t*-values in G′-E′ is to increase the critical value of the test. For example if we set $\tau_{\alpha/2} = 2.58$ ($\alpha = 0.01$), the test result here becomes 0 and the initial configuration becomes $c = 15$, which has a higher probability in Table A1 and is thus preferred.

The impact of using $\alpha = 0.01$ is further illustrated on all tests with Figure 6b. Only nine change-points are now assigned to group (G, E, E′), which is considerably smaller than with $\alpha = 0.05$. Actually, 10 change-points moved to group G and 9 to group (E, E′). This difference can be understood by inspecting the distribution of t-values with respect to the corresponding critical values (2.58 vs. 1.96). Figure 4 shows that many t-values for E–E′, G′-E, and G′-E′ are smaller than 2.58 in absolute value. When these tests become insignificant, whereas the other three stay significant, the predicted configuration becomes $c = 1$ or 15, and when G′-E′ remains significant or close to 2.58, the predicted configuration becomes $c = 31$ or 35. Additionally, many configurations with low probabilities ($c = 3, 7, 17, 18, 32, 33$) have also disappeared.

Figure 6c shows the impact of the (probability-based) imbalanced sampling in the learning procedure. Overall, the results for the main groups are not much different compared with the balanced sampling. Two noticeable differences emerge, however. First, group (G, E, E′) reduces only slightly in size, from 29 to 20. The smallness of the impact is explained by the fact almost half of test results are initially in Table A1 and are not changed by the prediction. Second, almost all the configurations with the lowest probabilities, such as $c = 7, 21, 32, 33, 38$, with $p \leq 5.6 \times 10^{-4}$, are removed. Other configurations, with slightly higher probabilities, but still with $p \leq 0.01$, such as $c = 3, 16, 17$ (G, E′) and $c = 29$ and 34 (G, E), emerge or are reinforced, which is not wanted.

Figure 6d shows the variant (d) where the aggregation is based on the prior probabilities. The distribution is quite different from that based on distance (Figure 6a): more change-points are attributed to the preferred groups, 62% in G and 19% in (E, E′), fewer to other groups such as (G, E, E′) and E, and many configurations of low probability disappear. The distribution is actually quite similar to that of variant (b), but in contrast to the latter, this variant keeps a high power in the test (thanks to $\alpha = 0.05$). As a result, with variant (d), group E is much smaller than with variant (b). In this respect, variant (d) is preferred among all four variants. Note, however, that there is a limitation in the usage of the aggregation procedure which holds for all variants: when there is only one nearby station (30% of the cases), the aggregation has no impact and the final result is the one selected by the prediction rule. In variant (d), this explains why there are still configurations with a low probability ($c = 8/22, 3, 31/35, 21, 38$).

5 | CONCLUSIONS AND PERSPECTIVES

We proposed a post-processing method for the attribution of change-points detected by a segmentation scheme involving multiple SDs (target minus reference). In our application, the each of the stations provides a GNSS series (G) and a reanalysis series (E). The segmentation is run on the G-E series and the goal of the attribution method is to predict if the inhomogeneity (jump in the mean) is in G or E. The method proceeds along the following steps:

1. Data selection and pre-processing. For each detected change-point in a main station (hereafter, the 'main change-point'): (a) select nearby stations with a horizontal distance smaller than 200 km and height difference smaller than 500 m; (b) run the segmentation method on the G′-E′ for each nearby data and select only homogeneous segments from the nearby to compare with the main; (c) correct the nearby series, G′ and E′, for the height difference with respect to the main station, so that all four series (G, E, G′, and E′) are representative of the same height; (d) form the six SDs (G-E, G-G′, etc.) and remove the outliers.

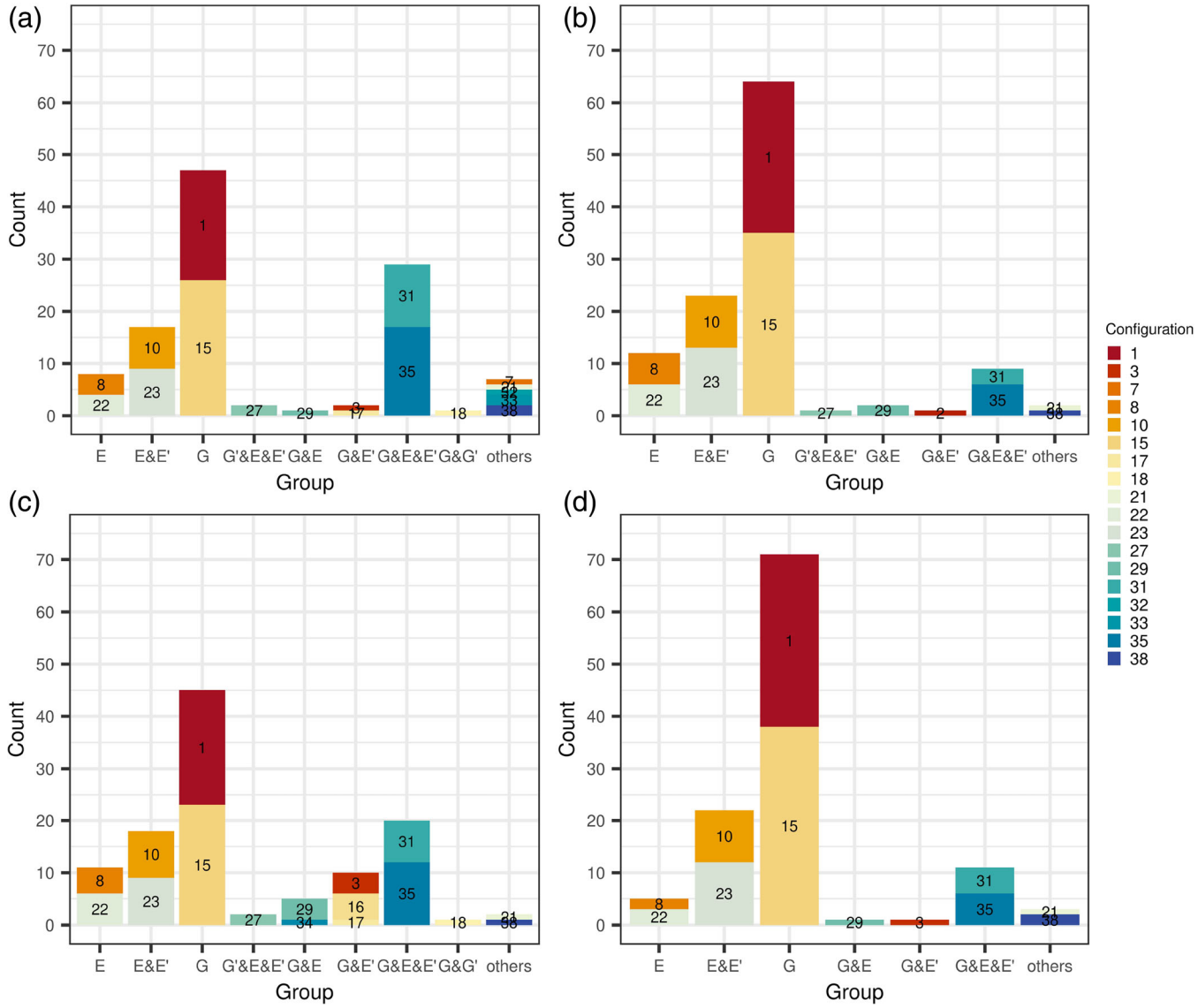

FIGURE 6 Distribution of the final predicted configurations, after aggregation, from the real data with the random forest method. The numbers in colour bars refer to the configuration number c among the 38 cases displayed in the RSD Table (Table A1). Results are plotted for four cases: (a) balanced sample learning with $\alpha=0.05$ and aggregated by distance, (b) balanced sample with $\alpha=0.01$ and aggregated by distance, (c) imbalanced sample with $\alpha=0.05$ and aggregated by distance, and (d) balanced sample with $\alpha=0.05$ and aggregated by prior probability. [Colour figure can be viewed at wileyonlinelibrary.com]

2. Test the significance of the jumps. For each main change-point and each of the six series: (a) identify the noise model; (b) fit a regression model including a jump at the position of the main change-point when at least $n=200$ consecutive points are available on the left and right of the change-point, using an iterative FGLS procedure; (c) test the significance of the jump at the significance level $\alpha=0.05$.
3. Use a predictive rule to predict the configuration. For each nearby, the learned classifier will predict the configuration (i.e., which of the G, E, G', and E' series have a significant jump) corresponding to the six test results. When several nearby series are used, a weighted prediction score is computed to select the final configuration.

The method has been applied to a real data set of 494 cases (114 change-points from 49 main stations compared with 312 nearby stations). The data characterization showed that the data have a strong heteroscedasticity, with mean annual seasonal variation in the standard deviation around 70% (half-range), and a moderate autocorrelation, with a typical lag-1 correlation coefficient of 0.3. A FGLS test procedure was implemented to ensure an accurate inference. The predictive rule has been trained on the real data. Several

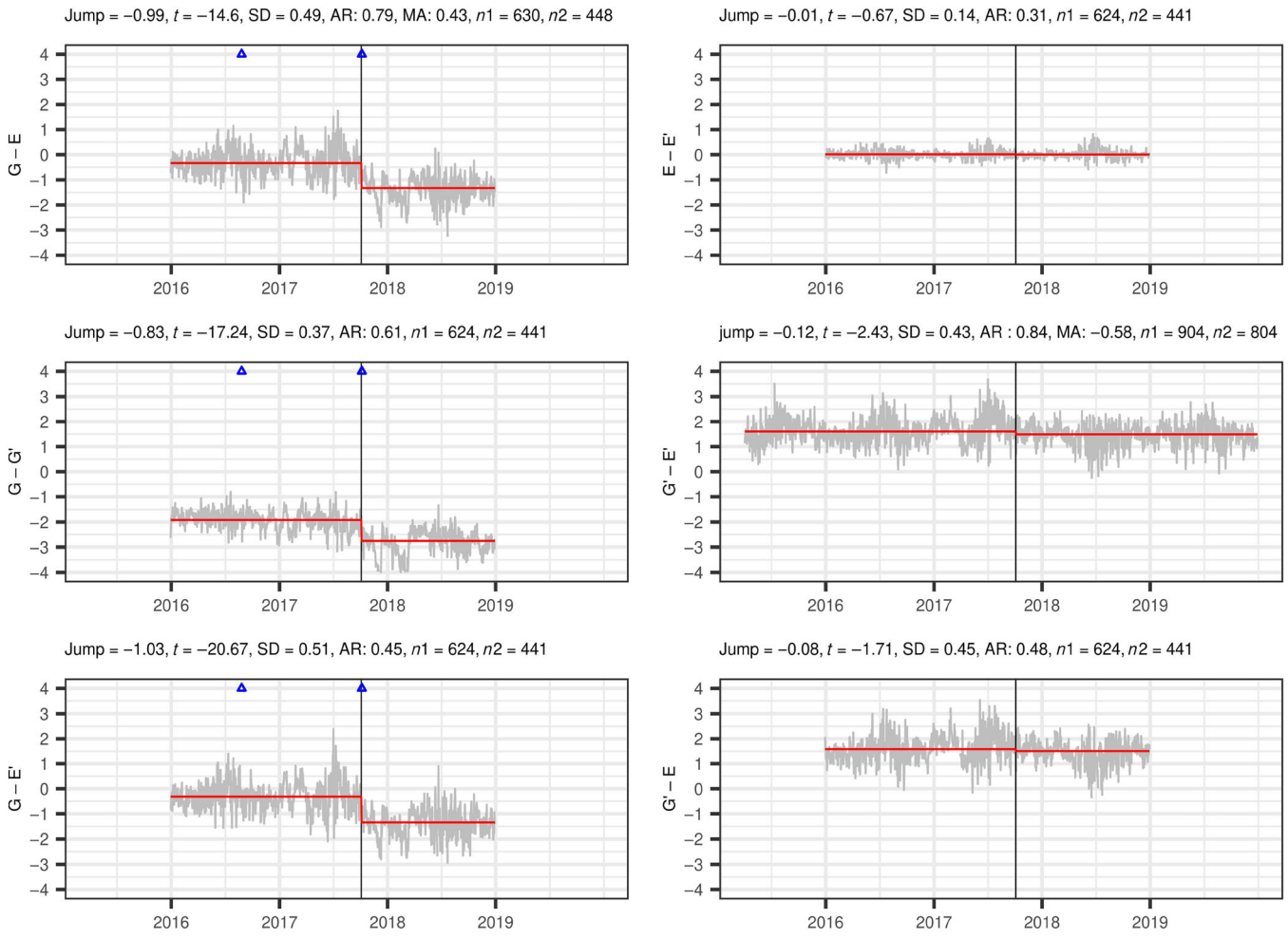

FIGURE 7 Example of test result for station FAIR (Fairbanks, Alaska) with nearby station CLGO at a distance of 21 km. The series of integrated water vapour differences are shown in grey. The black vertical solid line shows the change-point detected in G-E by the segmentation (4 October 2017). The blue triangles indicate known equipment changes in the main station from the Global Navigation Satellite System metadata. The horizontal red lines show the means estimated by the feasible generalized least squares regression on the left and the right of the change-point in each series. [Colour figure can be viewed at wileyonlinelibrary.com]

classifiers have been compared for the predictive rule and the RF was selected.

To our knowledge, both the FGLS regression test approach and the RF classifier have never been used in the context of climate series homogenization.

The FGLS tests and the classification results of the studied data set have been assessed using (i) our expertise of the data set (formulated out in two probabilistic rules) and (ii) metadata informing about known equipment changes at the GNSS sites. Very consistent and plausible results were found from both the FGLS tests and the classification. With a significance level of 5% and employing a balanced sample for the learning step in the predictive rule, as well as aggregating results from nearby sources based on prior probability, the findings clearly indicate predominance of significant jumps in the series involving G (62%), (E, E') (19%), and (G, E, E') (10%) as expected from the probabilities in Table A1. The remaining 9% of unexpected results are thought be linked with low detection power of the FGLS test when the noise is large (e.g., due to large distance between the main station and the nearby) and possibly random errors in the classification due to the smallness of the learning sample.

Some possibilities to further improve the method are: (i) to use a bigger data set to improve the predictive method, (ii) to refine the nearby selection rules to improve the robustness and the power of the test procedure (e.g., select nearby series with smaller percentage of gaps, shorten the distance between the main station and nearby), (iii) compute the critical value used in the FGLS test from a more realistic empirical distribution. These options will be tested in a future work.

The next step in the homogenization procedure is the correction of jumps. In the case of jumps attributed to G,

the correction will be applied to the initial G series only, for example, by correcting segments back in time, leaving the most recent segment unchanged (Nguyen et al., 2021; Van Malderen et al., 2020). A similar method can be applied to the E and E' series, if one is interested in correcting the reanalysis time series. In cases where jumps are attributed to (G, E, E'), a method for splitting the estimated jump into G and E components needs to be developed. This is planned in a future work.

AUTHOR CONTRIBUTIONS

Khanh Ninh Nguyen: Conceptualization; methodology; software; formal analysis; writing – review and editing; investigation; validation. **Olivier BOCK:** Conceptualization; investigation; methodology; validation; supervision; writing – review and editing; formal analysis. **Emilie Lebarbier:** Conceptualization; investigation; methodology; validation; supervision; software; writing – review and editing.

ACKNOWLEDGEMENTS

This work was developed in the framework of the VEGAN project supported by the CNRS program LEFE/INSU, and conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01) and within the FP2M federation (CNRS FR 2036).

FUNDING INFORMATION

This work was supported by the CNRS program LEFE/INSU, Labex MME-DII (ANR11-LBX-0023-01), and FP2M federation (CNRS FR 2036).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at 10.25326/68 (CODE) and 10.25326/518 (NGL).

ORCID

Olivier Bock https://orcid.org/0000-0001-5980-938X

REFERENCES

Alexandersson, H. (1986) A homogeneity test applied to precipitation data. *Journal of Climatology*, 6(6), 661–675. Available from: https://doi.org/10.1002/joc.3370060607

Blewitt, G., Hammond, W. & Kreemer, C. (2018) Harnessing the GPS data explosion for interdisciplinary science. *Eos*, 99, 19–22. Available from: https://doi.org/10.1029/2018eo104623

Bock, O. (2019) Global GNSS IWV data at 436 stations over the 1994–2018 period. Aeris. Available from: https://doi.org/10.25326/18

Bock, O. (2022) Global GNSS Integrated Water Vapour data, 1994–2022. Aeris. Available from: https://doi.org/10.25326/68

Bock, O. (2023) Global GNSS Integrated Water Vapour data based on NGL repro3, 1994-2022. Aeris. Available from: https://doi.org/10.25326/518

Bock, O., Bosser, P., Flamant, C., Doerflinger, E., Jansen, F., Fages, R. et al. (2021) Integrated water vapour observations in the Caribbean arc from a network of ground-based GNSS receivers during EUREC4A. *Earth System Science Data*, 13(5), 2407–2436. https://essd.copernicus.org/articles/13/2407/2021/

Bock, O., Bosser, P. & Mears, C. (2022) An improved vertical correction method for the inter-comparison and inter-validation of integrated water vapour measurements. *Atmospheric Measurement Techniques*, 15(19), 5643–5665. Available from: https://doi.org/10.5194/amt-15-5643-2022

Bock, O., Collilieux, X., Guillamon, F., Lebarbier, E. & Pascal, C. (2019) A breakpoint detection in the mean model with heterogeneous variance on fixed time intervals. *Statistics and Computing*, 30(1), 195–207. Available from: https://doi.org/10.1007/s11222-019-09853-5

Bock, O. & Parracho, A. (2019) Consistency and representativeness of integrated water vapour from ground-based GPS observations and ERA-interim reanalysis. *Atmospheric Chemistry and Physics*, 19, 9453–9468.

Bock, O., Willis, P., Wang, J. & Mears, C. (2014) A high-quality, homogenized, global, long-term (1993–2008) DORIS precipitable water data set for climate monitoring and model verification. *Journal of Geophysical Research: Atmospheres*, 119(12), 7209–7230.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees (1st ed.)*. Chapman and Hall/CRC. Available from: https://doi.org/10.1201/9781315139470

Caussinus, H. & Mestre, O. (2004) Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 53(3), 405–425. Available from: https://doi.org/10.1111/j.1467-9876.2004.05155.x

Costa, A.C. & Soares, A. (2009) Homogenization of climate data: review and new perspectives using geostatistics. *Mathematical Geosciences*, 41(3), 291–305. Available from: https://doi.org/10.1007/s11004-008-9203-3

Cover, T.M. & Hart, P.E. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 21–27. https://api.semanticscholar.org/CorpusID:5246200

Domonkos, P. (2011) Adapted Caussinus-Mestre algorithm for networks of temperature series (ACMANT). *International Journal of Geosciences*, 2(3), 293–309. Available from: https://doi.org/10.4236/ijg.2011.23032

Domonkos, P. (2021) Combination of using pairwise comparisons and composite reference series: a new approach in the homogenization of climatic time series with ACMANT. *Atmosphere*, 12(9), 1134. Available from: https://doi.org/10.3390/atmos12091134

Domonkos, P., Guijarro, J.A., Venema, V., Brunet, M. & Sigró, J. (2021) Efficiency of time series homogenization: method comparison with 12 monthly temperature test datasets. *Journal of Climate*, 34(8), 2877–2891. Available from: https://doi.org/10.1175/jcli-d-20-0611.1

Dunn, R.J.H., Aldred, F., Gobron, N., Miller, J.B., Willett, K.M., Ades, M. et al. (2021) Global climate. *Bulletin of the American*

Meteorological Society, 102(8), S11–S142. Available from: https://doi.org/10.1175/bams-d-21-0098.1

Easterling, D.R. & Peterson, T.C. (1995) A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology*, 15, 369–377.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.

Guijarro, J.A. (2011) *User's guide to climatol. An R contributed package for homogenization of climatological series*. Spain: State Meteorological Agency, Balearic Islands Office. http://webs.ono.com/climatol/climatol.html

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049.

Ho, T.K. (1995) Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. Montreal: IEEE, pp. 278–282.

Jones, P.D., Raper, S.C.B., Bradley, R.S., Diaz, H.F., Kellyo, P.M. & Wigley, T.M.L. (1986) Northern hemisphere surface air temperature variations: 1851–1984. *Journal of Climate and Applied Meteorology*, 25(2), 161–179.

Lu, Q., Lund, R. & Lee, T.C.M. (2010) An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1), 299–319. Available from: https://doi.org/10.1214/09-aoas289

Lund, R., Hurd, H., Bloomfield, P. & Smith, R. (1995) Climatological time series with periodic correlation. *Journal of Climate*, 8(11), 2787–2809. Available from: https://doi.org/10.1175/1520-0442(1995)008<2787:ctswpc>2.0.co;2

Menne, M.J. & Williams, C.N. (2005) Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of Climate*, 18(20), 4271–4286.

Menne, M.J. & Williams, C.N. (2009) Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22(7), 1700–1717. Available from: https://doi.org/10.1175/2008jcli2263.1

Menne, M.J., Williams, C.N. & Vose, R.S. (2009) The U.S. historical climatology network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, 90(7), 993–1008. Available from: https://doi.org/10.1175/2008bams2613.1

Mestre, O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, E. et al. (2013) HOMER: a homogenization software – methods and applications. *Idojaras*, 01, 117.

Mitchell, T.D. & Jones, P.D. (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *International Journal of Climatology*, 25(6), 693–712. Available from: https://doi.org/10.1002/joc.1181

Nguyen, K.N., Quarello, A., Bock, O. & Lebarbier, E. (2021) Sensitivity of change-point detection and trend estimates to GNSS IWV time series properties. *Atmosphere*, 12(9), 1102. Available from: https://doi.org/10.3390/atmos12091102

Ning, T., Wickert, J., Deng, Z., Heise, S., Dick, G., Vey, S. et al. (2016) Homogenized time series of the atmospheric water vapor content obtained from the GNSS reprocessed data. *Journal of Climate*, 29(7), 2443–2456.

Parracho, A.C., Bock, O. & Bastin, S. (2018) Global IWV trends and variability in atmospheric reanalyses and GPS observations. *Atmospheric Chemistry and Physics*, 18(22), 16213–16237. https://www.atmos-chem-phys.net/18/16213/2018/

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N. et al. (1998) Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 18(13), 1493–1517.

Peterson, T.C., Vose, R., Schmoyer, R. & Razuvaëv, V. (1998) Global historical climatology network (GHCN) quality control of monthly temperature data. *International Journal of Climatology*, 18(11), 1169–1179. Available from: https://doi.org/10.1002/(sici)1097-0088(199809)18:11<1169::aid-joc309>3.0.co;2-u

Pinheiro, J.C. & Bates, D.M. (2000) *Mixed-effects models in S and S-PLUS*. Springer. New York, NY. Available from: https://doi.org/10.1007/b98882

Quarello, A. (2020) Development of new homogenisation methods for GNSS atmospheric data. Application to the analysis of climate trends and variability. Phd thesis, Sworbonne Universite; IGN (Institut National de l'Information Géographique et Forestière). https://hal.archives-ouvertes.fr/tel-03118629

Quarello, A., Bock, O. & Lebarbier, E. (2022) GNSSseg, a statistical method for the segmentation of daily GNSS IWV time series. *Remote Sensing*, 14(14), 3379. Available from: https://doi.org/10.3390/rs14143379

Dach, R., Schaer, S., Arnold, D., Brockmann, E., Kalarus, M.S., Prange, L., Stebler, P. & Jaggi, A. (2023) CODE Final Product Series for the IGS. Astronomical Insitute, University of Bern.

Reeves, J., Chen, J., Wang, X.L., Lund, R. & Lu, Q.Q. (2007) A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6), 900–915. Available from: https://doi.org/10.1175/JAM2493.1

Rienecker, M.M., Suarez, M.J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E. et al. (2011) MERRA: NASA's modern-era retrospective analysis for research and applications. *Journal of Climate*, 24(14), 3624–3648. https://journals.ametsoc.org/view/journals/clim/24/14/jcli-d-11-00015.1.xml

Schroeder, M., Lockhoff, M., Forsythe, J.M., Cronk, H.Q., Haar, T.H.V. & Bennartz, R. (2016) The GEWEX water vapor assessment: results from Intercomparison, trend, and homogeneity analysis of total column water vapor. *Journal of Applied Meteorology and Climatology*, 55(7), 1633–1649. Available from: https://doi.org/10.1175/jamc-d-15-0304.1

Shumway, R.H. & Stoffer, D.S. (2017) *Time series analysis and its applications*. Springer International Publishing, Cham: Springer. Available from: https://doi.org/10.1007/978-3-319-52452-8

Szentimrey, T. (2008) Development of MASH homogenization procedure for daily data. Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases. WCDMP-No 71, p. 123–130. http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP71.pdf

Trenberth, K.E., Jones, P.D., Ambenje, P., Bojariu, R., Easterling, D., Klein Tank, A. et al. (2007) Observations. Surface and atmospheric climate change. In: *IPCC fourth assessment report: climate change 2007. Working group I: the physical science basis*. Cambridge: Cambridge University Press, pp. 235–336.

Van Malderen, R., Pottiaux, E., Klos, A., Domonkos, P., Elias, M., Ning, T. et al. (2020) Homogenizing GPS integrated water vapor

time series: benchmarking break detection methods on synthetic data sets. *Earth and Space Science*, 7(5), e2020EA001121. Available from: https://doi.org/10.1029/2020EA001121

Venema, V.K.C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P. et al. (2012) Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8(1), 89–115. https://www.clim-past.net/8/89/2012/

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: Atmospheres*, 114, D10101.

Wang, X.L., Chen, H., Wu, Y., Feng, Y. & Pu, Q. (2010) New techniques for the detection and adjustment of shifts in daily precipitation data series. *Journal of Applied Meteorology and Climatology*, 49(12), 2416–2436. https://journals.ametsoc.org/view/journals/apme/49/12/2010jamc2376.1.xml

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nguyen, K. N., Bock, O., & Lebarbier, E. (2024). A statistical method for the attribution of change-points in segmented Integrated Water Vapor difference time series. *International Journal of Climatology*, *44*(6), 2069–2086. https://doi.org/10.1002/joc.8441

APPENDIX A: TEST TABLE

The purpose of Table A1 is to help attribute the origin of jumps in the BS (G, E, G′, E′) given the test results performed on the six SDs (G-E, G′-E′, G-G′, G-E′, G′-E, E–E′). To build this table, we only consider the sign of the jumps, that is, jump amplitudes in the BS are coded on three values: 0 (no jump), +1 (upward jump), and −1 (downward jump). The corresponding jump amplitudes in the six SDs are logically coded on five levels (0, 1, 2, −1, and −2). They are presented in the SD Table. However, this table cannot be used in practice, because a test result will be either 'reject' or 'fail to reject' which, combined with the sign of the jump, leads to three levels only (−1, 0, +1). Hence, it is the RSD Table that will be used.

The BS Table lists 54 combinations of the four BS jumps. The combinations with G = 0 and E = 0 are not represented because we only consider cases when a change-point was detected in the G-E series by the segmentation method. One difficulty arises from the fact that some combinations of the BS Table (e.g., 1 and 18) lead to the same configurations in SD and RSD Table (highlighted by a coloured background) which thus contain only 46 and 38 unique configurations, respectively. In practice, the duplicates are sorted out based on the probabilities attributed to the corresponding combinations of jumps in the BS (see Supplement S1).

TABLE A1 Test Table: (left) 54 relevant combinations of the jumps in the four base series (G, E, G', E'), coded on three values (0 = no jump, −1 = downward jump, +1 = upward jump), and associated conditional and joint probabilities (see Supplement S1). [Colour table can be viewed at wileyonlinelibrary.com]

	Base Series Table					
	Jump amplitude				Conditional probability	Joint probability
	G	E	G'	E'	P(G', E' | G, E)	P(G, E, G', E')
1	1	0	0	0	0,8	0,18225
2	1	0	0	1	0,045	0,010125
3	1	0	0	-1	0,045	0,010125
4	1	0	1	0	0,045	0,010125
5	1	0	1	1	0,0025	0,0005625
6	1	0	1	-1	0,0025	0,0005625
7	1	0	-1	0	0,045	0,010125
8	1	0	-1	1	0,0025	0,0005625
9	1	0	-1	-1	0,0025	0,0005625
10	0	-1	0	0	0,045	0,010125
11	0	-1	0	1	0,045	0,010125
12	0	-1	0	-1	0,81	0,18225
13	0	-1	1	0	0,0025	0,0005625
14	0	-1	1	1	0,0025	0,0005625
15	0	-1	1	-1	0,045	0,010125
16	0	-1	-1	0	0,0025	0,0005625
17	0	-1	-1	1	0,0025	0,0005625
18	0	-1	-1	-1	0,045	0,010125
19	-1	0	0	0	0,81	0,18225
20	-1	0	0	1	0,045	0,010125
21	-1	0	0	-1	0,045	0,010125
22	-1	0	1	0	0,045	0,010125
23	-1	0	1	1	0,0025	0,0005625
24	-1	0	1	-1	0,0025	0,0005625
25	-1	0	-1	0	0,045	0,010125
26	-1	0	-1	1	0,0025	0,0005625
27	-1	0	-1	-1	0,0025	0,0005625
28	0	1	0	0	0,045	0,010125
29	0	1	0	1	0,81	0,18225
30	0	1	0	-1	0,045	0,010125
31	0	1	1	0	0,0025	0,0005625
32	0	1	1	1	0,045	0,010125
33	0	1	1	-1	0,0025	0,0005625
34	0	1	-1	0	0,0025	0,0005625
35	0	1	-1	1	0,045	0,010125
36	0	1	-1	-1	0,0025	0,0005625
37	1	-1	0	0	0,045	0,00225
38	1	-1	0	1	0,045	0,00225
39	1	-1	0	-1	0,81	0,0405
40	1	-1	1	0	0,0025	0,000125
41	1	-1	1	1	0,0025	0,000125
42	1	-1	1	-1	0,045	0,00225
43	1	-1	-1	0	0,0025	0,000125
44	1	-1	-1	1	0,0025	0,000125
45	1	-1	-1	-1	0,045	0,00225
46	-1	1	0	0	0,045	0,00225
47	-1	1	0	1	0,81	0,0405
48	-1	1	0	-1	0,045	0,00225
49	-1	1	1	0	0,0025	0,000125
50	-1	1	1	1	0,045	0,00225
51	-1	1	1	-1	0,0025	0,000125
52	-1	1	-1	0	0,0025	0,000125
53	-1	1	-1	1	0,045	0,00225
54	-1	1	-1	-1	0,0025	0,000125

	Series of Difference Table					
	Jump amplitude					
	G-E	G-G'	G-E'	E-E'	G'-E'	G'-E
1	1	1	1	0	0	0
2	1	1	0	-1	-1	0
3	1	1	2	1	1	0
	1	0	1	0	1	1
	1	0	0	-1	0	1
4	1	0	2	1	2	1
5	1	2	1	0	-1	-1
6	1	2	0	-1	-2	-1
7	1	2	2	1	0	-1
8	1	0	0	-1	0	1
9	1	0	-1	-2	-1	1
10	1	0	1	0	1	1
11	1	-1	0	-1	1	2
12	1	-1	-1	-2	0	2
13	1	-1	1	0	2	2
	1	1	0	-1	-1	0
14	1	1	-1	-2	-2	0
	1	1	1	0	0	0
15	-1	-1	-1	0	0	0
16	-1	-1	-2	-1	-1	0
17	-1	-1	0	1	1	0
18	-1	-2	-1	0	1	1
19	-1	-2	-2	-1	0	1
20	-1	-2	0	1	2	1
	-1	0	-1	0	-1	-1
21	-1	0	-2	-1	-2	-1
	-1	0	0	1	0	-1
22	-1	0	0	1	0	-1
23	-1	0	-1	0	-1	-1
24	-1	0	1	2	1	-1
	-1	-1	0	1	1	0
	-1	-1	-1	0	0	0
25	-1	-1	1	2	2	0
26	-1	1	0	1	-1	-2
27	-1	1	-1	0	-2	-2
28	-1	1	1	2	0	-2
29	2	1	1	-1	0	1
30	2	1	0	-2	-1	1
31	2	1	2	0	1	1
32	2	0	1	-1	1	2
33	2	0	0	-2	0	2
34	2	0	2	0	2	2
35	2	2	1	-1	-1	0
36	2	2	0	-2	-2	0
37	2	2	2	0	0	0
38	-2	-1	-1	1	0	-1
39	-2	-1	-2	0	-1	-1
40	-2	-1	0	2	1	-1
41	-2	-2	-1	1	1	0
42	-2	-2	-2	0	0	0
43	-2	-2	0	2	2	0
44	-2	0	-1	1	-1	-2
45	-2	0	-2	0	-2	-2
46	-2	0	0	2	0	-2

	Reduced Series of Difference Table					
	Jump amplitude					
	G-E	G-G'	G-E'	E-E'	G'-E'	G'-E
1	1	1	1	0	0	0
2	1	1	0	-1	-1	0
3	1	1	1	1	1	0
	1	0	1	0	1	1
	1	0	0	-1	0	1
4	1	0	1	1	1	1
5	1	1	1	0	-1	-1
6	1	1	0	-1	-1	-1
7	1	1	1	1	0	-1
8	1	0	0	-1	0	1
9	1	0	-1	-1	-1	1
10	1	0	1	0	1	1
11	1	-1	0	-1	1	1
12	1	-1	-1	-1	0	1
13	1	-1	1	0	1	1
	1	1	0	-1	-1	0
14	1	1	-1	-1	-1	0
	1	1	1	0	0	0
15	-1	-1	-1	0	0	0
16	-1	-1	-1	-1	-1	0
17	-1	-1	0	1	1	0
18	-1	-1	-1	0	1	1
19	-1	-1	-1	-1	0	1
20	-1	-1	0	1	1	1
	-1	0	-1	0	-1	-1
21	-1	0	-1	-1	-1	-1
	-1	0	0	1	0	-1
22	-1	0	0	1	0	-1
23	-1	0	-1	0	-1	-1
24	-1	0	1	1	1	-1
	-1	-1	0	1	1	0
	-1	-1	-1	0	0	0
25	-1	-1	1	1	1	0
26	-1	1	0	1	-1	-1
27	-1	1	-1	0	-1	-1
28	-1	1	1	1	0	-1
29	1	1	1	-1	0	1
30	1	1	0	-1	-1	1
31	1	1	1	0	1	1
32	1	0	1	-1	1	1
	1	0	0	-1	0	1
	1	0	1	0	1	1
33	1	1	1	-1	-1	0
	1	1	0	-1	-1	0
	1	1	1	0	0	0
34	-1	-1	-1	1	0	-1
35	-1	-1	-1	0	-1	-1
36	-1	-1	0	1	1	-1
37	-1	-1	-1	1	1	0
	-1	-1	-1	0	0	0
	-1	-1	0	1	1	0
38	-1	0	-1	1	-1	-1
	-1	0	-1	0	-1	-1
	-1	0	0	1	0	-1

Note: (Middle) resulting jump amplitudes in the six series of differences (G-E, G-G', ...), coded on five levels (−2, −1, 0, 1, 2). (Right) jump amplitudes in the series of differences coded on three levels only (−1, 0, +1). This table can serve to attribute the origin of jumps in the base series given the test results performed on the six series of differences. Duplicates in the test results (highlighted with coloured background) are sorted out based on the corresponding probabilities.