A statistical method for the attribution of change-points in segmentation of IWV difference time series

Khanh Ninh Nguyen, Olivier Bock, Emilie Lebarbier

ORIGINAL ARTICLE

Journal Section

A statistical method for the attribution of change-points in segmentation of IWV difference time series

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

[1]Institut de Physique du Globe de Paris (IPGP), Centre National de la Recherche Scientifique (CNRS), Institut national de l'information géographique et forestière (IGN), Université de Paris, 75005 Paris, France

[2]Ecole Nationale des Sciences Géographiques (ENSG), Institut national de l'information géographique et forestière (IGN), F-77455 Marne-la-Vallée, France

[3]Laboratoire Modal'X, UPL, Université Paris Nanterre, 92000 Nanterre, France

Correspondence
K.N. Nguyen, IPGP Geodesy, 39 rue Helene Brion, 75013 Paris, France
Email: knguyen@ipgp.fr

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)

Many segmentation methods used for the homogenization of climate time series from station data use a reference series against which the station data is compared. The main advantage of this approach is to remove the common climate signal and thus improve the power of the change-point detection method. One drawback is that it is difficult to decide whether the detected change-point is due to the main station or to the reference. This paper describes a statistical method to help in this decision. It works by combining the data from the main station with the data from at least one nearby station, where the data from each station is actually composed of two series: a target series and a reference series. In our application, the target series is from daily GNSS Integrated Water Vapour (IWV) measurements and the reference series from the Fifth ECMWF reanalysis (ERA5). Six series of differences are formed from these four base series and a statistical test is used to detect, in each of the six series, if the change in mean before and after the tested change-point is significant. Finally, a predictive rule is used to determine which of the four base series is (are) affected by change-point(s). The statistical test is based on a generalized linear regression approach, taking both het-

[*] Equally contributing authors.

eroscedasticity and autocorrelation into account. The predictive rule is constructed on a dataset built from the test results obtained on the real data using a resampling strategy. Four popular machine learning methods have been considered and evaluated using cross-validation. The proposed method was applied to a real data set and the results looked very consistent and plausible with respect to GNSS metadata and our knowledge of the data. Results conclude that 41% of the change-points are due to GNSS data, 15% to the ERA5 data, and 25% are due to coincident detections.

KEYWORDS
Attribution, Change-point testing, Unsupervised classification

1 | INTRODUCTION

Long records of climate observations are crucial for the monitoring climate change and better understanding the underlying climate processes [1]. 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 [2]. Inhomogeneities often take the form of abrupt changes, which are detrimental to estimating trends and multi-scale climate variability [3]. Various homogenization methods have been developed for the detection and the correction of such change-points in the context of climate data analysis [4, 5, 6, 7, 8, 9, 10]. The detection step, also called segmentation, can be performed in two classical ways, using either a statistical test (e.g. [6, 7]) or a penalized likelihood approach (e.g. [5, 8, 11]). The performance of a segmentation method depends on its ability to represent the stochastic properties of the data, with a parametric or a non-parametric model, and the search method which can be optimal or sub-optimal (see the literature review in [12]). Many segmentation methods are actually used on differenced data, where the data from the target station are compared to (differenced with respect to) a reference series. The reference series can either be from one nearby station, or a composite from several nearby stations, or from any other auxiliary data source. Several recent studies used atmospheric reanalyses as an auxiliary data source [13, 14, 15, 16, 17]. The main advantage of working on differenced data is to remove the common climate signal and thus improve the power of the change-point detection method. One drawback is that it is difficult to decide whether the detected change-point is due to the main station or to the reference. This decision step, which we call hereafter "attribution", is crucial for the correction step in which the target series will be adjusted for the changes in mean, segment by segment (see, e.g., the correction procedure used by [15] or [16]). Any false attribution would in fact introduce artificial jumps in the target series.

Figure 1 explains the idea of the attribution method proposed in this paper. Let us assume that a differenced time series, G-E, of a main station has been segmented and that three change-points have been detected at times t_1, t_2, and t_3. In the example, they are associated with changes in the mean of +1, -1, and -0.5 signal units. We typically don't know if the jumps are due to the target series (denoted G for GNSS), to the reference series (denoted E for ERA5), or to both. Let us assume that there exists a nearby station with data available over a sufficient time span before and after a given change-point from the main station (in the sketch illustrated in Figure 1 the nearby series

G'-E' covers all three change-points from the main series G-E). Since the data from the nearby station can also be subject to inhomogeneities, the G'-E' is also segmented. In this example we assume that 2 change-points, denoted t'_1 and t'_2, are detected in the nearby series, with jumps of +1 and -0.5 signal units, respectively. We see also that t'_2 is close in time to t_3 of the main station. Such a situation can happen in real data. Based on our experience, we will adopt the following rules:

(R1) it is unlikely that change-points in two different GNSS series (here G and G') occur at the same time because they are due to rare, station-specific events (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 with a large spatial extent (e.g. due to the start or end of assimilation of a satellite), hence being present in E and E'.

Inspection of the two differenced series G-E and G'-E' in the light of these rules suggests that t_1 is due to a +1 change in G, t'_1 (supposedly far from t_1) is due to a +1 change in G', t_2 is due to a -1 change in G, and t'_2 and t_3 are due to a -0.5 change in both E and in E'. In the following we will attach some probabilities to these two rules, but it is obvious that this is not enough to conclude. To confirm our guess, we need to inspect one additional series of differences combing two of the four previous series (G, E, G', and E'). Figure 1 illustrates the case of G-G'. Using elementary combinatory logic, it is straightforward to determine that the guessed solution is correct. In reality, there are a total of six combinations of two series among four, without repetition. Instead of using only three of them, we can take advantage of the redundancy offered by the three additional differenced series, especially since in the real world, some detections may fail when the amplitude of the jump is small compared to the background noise, or when the sample size is small. The power of the detection test is a crucial aspect of the method which will be discussed in Section 3. The final step in the attribution method is the combination of the six test results in order to predict the solution.

The Table 4 shows all possible values of the quadruplet composed of G, E, G', and E' and the corresponding test results of the six differenced series G-E, G'-E', G-G', G-E', G'-E, E-E'. The jumps in the base series are coded on three values: 0 (no jump), +1 (positive jump), and -1 (negative jump). Logically, the corresponding test results should be coded on five different values: 0, 1, 2, -1, and -2. However, in practice, a test result will be either reject (0) or fail to reject (-1 or +1, where the sign indicates if the jump is upward or downward). The central part of Table 4 shows the logical table coded on five values, while the rightmost part shows the same table coded on three values only. The latter is referred to as the "truncated" table. The logical table contains 54 rows resulting from quadruplets for which either G or E, or both, have a jump (+1 or -1). The case where none of G or E has a jump is not considered. The 54 rows contain 46 unique combinations and 8 doubles of the six test results (highlighted by a colored background) and 38 unique combinations in the truncated table. The doubles are finally sorted out depending on their prior probabilities (see Appendix B for details on how they are computed). If there were no errors in the tests results, the correct solution could be found directly from this table. Because in practice the tests may be wrong, a statistical method is needed to predict the result.

In Section 2 we describe the stochastic properties of our data set composed of Integrated Water Vapor (IWV) time series from ground-based Global Navigation Satellite System (GNSS) and the fifth ECMWF reanalysis (ERA5). We highlight especially, the heteroscedasticity and autocorrelation of the differenced IWV series. In Section 3 we evaluate the power of several tests for the testing a change in mean in simulated series based on a regression with heteroscedasticity and autocorrelation. In Section 4 we describe the method for the construction of the predictive rule and present the results on a real data set. Section 5 discussed the results and concludes.

2 | DATA CHARACTERIZATION

2.1 | Data sets

In this paper, we are interested in daily IWV data from GNSS observations and from the ERA5 reanalysis. The GNSS IWV data are those we want to homogenize primarily. Several past studies highlighted the presence of abrupt changes in the mean in the GNSS data [18, 19, 13, 20] but also possibly in the reanalysis data [21, 13, 20, 16]. Inhomogeneities in the GNSS data are mainly due to equipment changes and changes in the station's environment, but the magnitude of the jump may also depend on the data processing procedure [16]. In this work, we use reprocessed GNSS data from Center for Orbit Determination in Europe (CODE), covering the period from 1994 to 2014 (REPRO2015), extended until the end of 2018 by a consistent operational processing. Details of the processing are described in [16] and references therein, and the data set is available from [22]. These data are from a global network of 436 stations. Data from the ERA5 reanalysis have been extracted at the location of each station and the difference series, G-E, have been segmented using the GNSSseg package described in [17]. For the purpose of the present study we selected 81 stations with the longest time series. These will be our "main stations". Nearby stations were searched with a distance limit of 200 km in horizontal and 500 m in vertical, but very few were found in the CODE data set. So we used the NGL reprocessed GNSS data set for the nearby GNSS data which comprises nearly 20 thousand stations [23]. We end up with 156 change-points in 55 main stations that can be tested with respect to 628 nearby stations when no constraint is put on the length of the segments.

2.2 | Pre-processing

Before we form the six difference series and estimate the change in the mean, the IWV data are adjusted for the station height difference and screened for outliers. The IWV adjustment is done following the method described in [24] with correction model coefficients estimated from ERA5. This step is important when the main station and nearby station are not at the same altitude, and impact both the GNSS data and the ERA5 data as the latter are extracted at the height of the GNSS data. Once the G and E data are made consistent with the G' and E' data, the six difference series are formed.

The screening consists in two steps. The first one is a classical outlier detection procedure in which the data points exceeding three standard deviations from the median are removed. In this procedure, the median and standard deviation are computed in a sliding window of length +/- 60 days around the current point. To guarantee the accuracy of the estimates in the presence of data gaps, the minimum number of available points is set to 20. Furthermore, in order to be robust against outliers, the standard deviation is computed with the tau-scale estimator [25].

The second step consists in remove data in short segments (less than 80 days) pertaining to a cluster of change-points detected in the main station (see an example in Figure 3 in [17]). This problem occurs occasionally in regions where the GNSS data and reanalysis data have a significant representativeness difference [26]. In such situation, we keep the first change-point but remove data between the the first and the last change-points of the cluster. Similarly, when a change-point in the nearby station is very close (less than 10 days) to a change-point in the main station, we consider that they are both due to the same cause (most likely a change in the reanalysis) and we remove the data points in the nearby series between the two change-points. This case is illustrated in Figure 1 where the data between t_3 and t_2' have been removed. In Figure 1, we also illustrate the case of a gap in the G-E series just after t_2 which may be due to a screened cluster. Note that, as a result of the screening, the number of data points in the between sites series G-G' is always smaller than or equal to that in the former. Figure 1 also illustrates this with a color code for

the length of the data available on each side of the change-points. In the proposed test procedure described in the next Section, the minimum duration of the segments on either side is set to one year, with a minimum number of consecutive points of 200, in order to achieve sufficiently accurate estimates of the model coefficients (mean, jump, etc...).

2.3 | Characterization of the data

The inspection of GNSS minus reanalysis differences show strong heteroscedasticity and periodic (seasonal) biases, along with weak autocorrelation [17]. The periodic bias can be easily modelled by a small order Fourier series. The heteroscedasticity can be modelled either as a piece-wise constant function (e.g. monthly [17]) or as a continuous time function with known or unknown shape. The autocorrelation part can be modelled by an Auto-Regressive-Moving Average (ARMA) process [27]. In the following, a series of difference is thus modelled using the following regression model:

$$z_t = \mu_L + \delta x_t + s_t + e_t, \tag{1}$$

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 $t \leq t_k$ and 1 if $t > t_k$, where t_k is the time of the change-point), 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 [28] and represent e_t as the product of two factors:

$$e_t = e_t^* \sigma_t, \tag{2}$$

where e_t^* represents a stationary autocorrelated process of unit variance and σ_t^2 is the time-varying variance of e_t, i.e. $Var[e_t] = \sigma_t^2$. Preliminary investigation of our data showed that most of the time the noise model is well approximated by an AR(1), followed by MA(1) and ARMA(1,1), and more rarely of higher order. For the characterization purpose, we thus considered only the four possible ARMA(p,q) models, with $p, q \in \{0, 1\}$, and the best one is selected using an automatic model selection procedure described in the next section. Recall that an ARMA(1,1) model writes [27]:

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

where w_t is a Gaussian white noise. This model is also considered for the test in the next Section. For the inference of all its parameters and the selection of the noise model, see thus the next section.

Figure 2 shows an example of a time series (jagged gray curve), with the estimated Fourier series (smooth black curve), the estimated standard deviation $\hat{\sigma}_t$ (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. Hereafter, $\hat{\sigma}_t$ is referred to as the Moving Standard Deviation (MSD).

Table 1 and 2 summarize the characteristics of our data set in terms of both heteroscedasticity and noise structure respectively, including one G-E series for each main station and the five differences series (G'-E', G-E'...) for all nearby stations. For the purpose of the characterization, only one change-point per main station is considered (i.e. we assume that the characteristics do not change with time). The results are sorted according to the distance between the main and the nearby stations (smaller or large 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', i.e. the series with collocated data, which have moderate MSD of 0.7 kg m^{-2}. The second group (G2) includes E-E' and G-G', the series comparing the same technique, which have the smallest MSD (0.5 kg m^{-2}). The last group (G3) involves data from mixed techniques and different sites, and gets the largest MSD. As the distance increases, the MSD of series involving different sites increases, as expected from increased representativeness differences. Another striking finding from Table 1 is that the relative variation of the MSD is around 70% for all six series, indicating that strong heteroscedasticity is present in the whole data set.

Figure 3 shows the distributions of the noise models and of the estimated coefficients for the six differences, again sorted according to the distance. First, most importantly, 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, with two distinct cases: either ARMA(1,1) is dominant or MA(1) is dominant. Interestingly, the same groups show up. The former case is observed for the collocated series (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, the latter series become MA(1) like 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 estimated 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 similar and have very little dispersion among the series and distance. More surprisingly, the estimated coefficients of the ARMA(1,1) depend strongly on the distance, with the exception of E-E'. Values of $\hat{\phi}$ and $\hat{\theta}$ are around 0.6 and -0.3, respectively, 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 (this property results from the fact that white noise can be represented by an ARMA(1,1) process with opposite coefficients [27]). 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 series. The mean values of the estimated coefficients are reported in Table 2.

3 | PROPOSED TESTS IN THE REGRESSION MODEL

3.1 | Regression model and different tests

In this section, we propose different procedures to test a change-point in a series taking into account the characteristics observed in the real series. We use the same model as previously (see (1) with specifications (2) and (3)) that can be written in the following matrix form:

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

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 i^{th} observation, z_i, $i = 1, ..., n$. The noise vector, e is assumed to be normal distributed, $e \sim \mathcal{N}(0, \Sigma_0)$, where Σ_0 is the variance-covariance matrix describing the noise model. The purpose is to test the nullity of the coefficient δ.

When Σ_0 is known, Ordinary Least Squares (OLS) or Generalized Least Squares (GLS) methods can be used. The solutions for $\hat{\beta}$ and the corresponding variance write:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'\mathbf{z} \quad \text{and} \quad Var[\hat{\beta}_{OLS}] = (X'X)^{-1}(X'\Sigma_0 X)(X'X)^{-1}, \tag{5}$$

$$\hat{\beta}_{GLS} = (X'\Sigma_0^{-1}X)^{-1}X'\Sigma_0^{-1}\mathbf{z} \quad \text{and} \quad Var[\hat{\beta}_{GLS}] = (X'\Sigma_0^{-1}X)^{-1}. \tag{6}$$

Recall that the GLS solution is the Best Linear Unbiased Estimator (BLUE) while the OLS solution is not, although OLS remains unbiased.

However, in practice, Σ_0 is unknown and needs to be estimated. The classical linear model (CLM) framework consists in assuming: (i) the independence of the data, and (ii) the homoscedasticity framework of the noise, e.g. $\Sigma_0 = \sigma_0^2 I_n$, where I_n is the identity matrix. An unbiased estimator of the variance σ_0^2 is $\hat{\sigma}_0^2 = \hat{e}'\hat{e}/(n-k)$ and an estimator of $Var[\hat{\beta}_{OLS}]$ is simply

$$\widehat{Var}[\hat{\beta}_{OLS}]_{CLM} = \hat{\sigma}_0^2(X'X)^{-1}. \tag{7}$$

The assumptions (i) and (ii) are clearly not satisfied here. In this case, despite the OLS solution (5) remains unbiased, the variance estimator is strongly biased and leads to significant inference errors. To solve this problem, some methods have been proposed in the literature. The two main ones, which we considered in this work, are

- the so-called OLS-HAC that consists in still using the OLS solution but to consider a consistent estimator for the variance that is the Heteroscedasticity and Autocorrelation Consistent estimator (HAC) [29, 30], i.e. that is robust to the presence of heteroskedasticy and serial correlations of unknown form and that has good asymptotic properties. The idea is to estimate $M = X'\Sigma_0 X$ instead of Σ_0, that is difficult to estimate due to its large size (it contains nominally $n(n+1)/2$ parameters). It writes:

  $$\widehat{Var}[\hat{\beta}_{OLS}]_{HAC} = (X'X)^{-1}\hat{M}(X'X)^{-1}. \tag{8}$$

  A large class of HAC estimators is the non-parametric kernel estimators. A drawback of this type of estimator is that its performance varies with the choices of the kernel function and its bandwidth, but the advantage is that they do not require to specify the covariance structure and they are computationally very fast. Here we consider the "Quadratic Spectral" kernel [31] with its proposed optimal bandwidth value. This method is available the R package `sandwich` [32].
- the Feasible GLS (FGLS) that consists in using the GLS solution and to estimate Σ_0 when it has a specific structure

(i.e. with a reduced number of parameters to be estimated). If $\hat{\Sigma}_n$ is the estimator of Σ_0, the estimator of β and its variance are obtained by replacing Σ_0 by $\hat{\Sigma}_n$ in the GLS equations:

$$\hat{\beta}_{FGLS} = (X'\hat{\Sigma}_n^{-1}X)^{-1}X'\hat{\Sigma}_n^{-1}\mathbf{z} \quad \text{and} \quad \widehat{Var}[\hat{\beta}_{FGLS}] = (X'\hat{\Sigma}_n^{-1}X)^{-1}. \tag{9}$$

Following [28], we assume that Σ_0 takes the form $\Sigma_0 = \boldsymbol{VCV}$, where $\boldsymbol{V} = diag(\sigma_t^2)$ and $\boldsymbol{C}$ is the correlation matrix associated to the ARMA noise model (see [27] for the formulation of matrix $\boldsymbol{C}$ as a function of ϕ and θ). Since $\hat{\beta}$ also depends on $\hat{\Sigma}_n$, an iterative procedure is implemented. In order to stabilize the convergence and also to avoid over-fitting, we first select the best model among the four possible models (ARMA(p,q) with $p, q \in \{0, 1\}$) using the `auto.arima` function in the R package `forecast` [33]. The BIC criterion is used and is complemented by a test of the significance of the final model coefficients. Next, the iterative procedure is implemented as follows:

1. fit the OLS solution (5);
2. compute the noise variance $\hat{\boldsymbol{V}}_n = diag(\hat{\sigma}_t^2)$ by moving window from the OLS residuals (as explained in Section 2.3);
3. fit a preliminary FGLS solution (6) where Σ_0 is replaced by $\hat{\boldsymbol{V}}_n$;
4. fit the ARMA model coefficients, $\hat{\phi}$ and $\hat{\theta}$, from the FGLS residuals;
5. compute $\hat{\Sigma}_n = \hat{\boldsymbol{V}}_n\hat{\boldsymbol{C}}_n\hat{\boldsymbol{V}}_n$, where $\hat{\boldsymbol{C}}_n$ is the correlation matrix of the ARMA model with the fitted coefficients;
6. fit the final FGLS solution (9);
7. repeat steps 3 to 6 until convergence.

In step 4, the parameters $\hat{\phi}$ and $\hat{\theta}$ of the ARMA noise structure are estimated by Maximum Likelihood with the function `arima` in R. In step 7, the convergence is tested from the difference of $\hat{\beta}$, $\hat{\phi}$, and $\hat{\theta}$, of two successive iterations. The maximum number of iterations is set to 10 and is rarely attained. For the preliminary selection of the noise model, a simplified scheme including step 1 to 4 only is used, and `auto.arima` function is used in step 4 instead of `arima`. Numerical simulations showed that `auto.arima` is able to select the correct model in 99 % of the cases when it is white noise, 93 % when it is MA(1), 90% when it is AR(1), and 63 % when it is ARMA(1,1), for a sample size $n \geq 1000$ and typical coefficient values encountered in the real data (see Fig. 3). To ensure a good performance of `auto.arima`, a large sample size is thus required. For this reason, the characterization is done on the longest segment for each couple (main, nearby).

Note that, compared to the OLS-HAC approach, the FGLS procedure, due to its iterative scheme, is computational time-demanding. However, when the noise model is correctly specified, the FGLS estimate of $\widehat{Var}[\hat{\beta}]$ is more accurate than its OLS-HAC counterpart, and the power of the test is improved.

The test statistic of the test of the nullity of δ is thus

$$T_{\delta,\text{OLS-HAC}} = \frac{\hat{\delta}_{\text{OLS}}}{\hat{\sigma}_{\hat{\delta}_{\text{HAC}}}} \quad \text{for the OLS-HAC, and} \quad T_{\delta,\text{FGLS}} = \frac{\hat{\delta}_{\text{FGLS}}}{\hat{\sigma}_{\hat{\delta}_{\text{FGLS}}}} \quad \text{for the FGLS,} \tag{10}$$

where $\hat{\sigma}_{\hat{\delta}_\bullet}$ is the estimated standard error of $\hat{\delta}$ which is extracted from $\widehat{Var}[\hat{\beta}_\bullet]$, where $\bullet$ = OLS-HAC or FGLS. For the OLS-CLM and GLS estimators considered in the simulation study, similar statistics are computed using their respective estimators for $\hat{\delta}$ and $\hat{\sigma}_{\hat{\delta}}$.

In contrast to the HAC estimator, the asymptotic properties (unbiasedness and efficiency) of the FGLS estimator

are typically not known. Numerical simulations with the procedure described above showed that $\hat{\beta}_{FGLS}$ is not biased and its variance is very close to GLS, although slightly larger. Finally, the distribution of both statistics showed very close to normal and we will assume in the following that the distributions follow a $\mathcal{N}(0,1)$ under H_0. This allows to compute the critical value for a given α.

3.2 | Evaluation based on simulations

We conducted a large number of simulations, for different types of noise characteristics (autocorrelated and/or heteroscedastic) and sample sizes, to assess the different test methods introduced above. In these simulations, we modelled the noise heteroscedasticity by a "raised cosine" function of time: $\sigma_t^2 = \sigma_m^2 - \sigma_v^2 \cos 2\pi t/T$, where σ_m^2 and $\sigma_v^2 \leq \sigma_m^2$ represent the mean and half-range modulation of the variance over one period, respectively.

The results are compared in Figure 4. It is seen that the False Positive Rate (FPR) stays fairly close to the nominal significance level, $\alpha = 0.05$, for GLS, FGLS, and OLS-HAC, except when the autocorrelation is very strong. The OLS-CLM method performs very badly when the data is autocorrelated. This is due to the bias in the predicted variance (Eq. 7) in this situation. In contrast, the impact of heteroscedasticity on FPR is negligible for all solutions, including OLS-CLM (this latter results stems from the fact that $\hat{\sigma}_0^2$ is equal to the sample mean of σ_t^2). When the autocorrelation increases, $\hat{\sigma}_{\hat{\delta}}$ increases and the power of the test, measured by the True Positive Rate (TPR), decreases. This feature is sometimes interpreted as a reduction of the "equivalent sample size" [34]. The higher TPR of OLS-CLM is actually linked to its higher FPR. Another metric, called "accuracy", is sometimes used which combines both FPR and TPR, and is reflecting the overall performance of the test (with this metric, the performance of OLS-CLM is clearly lower than GLS and FGLS). When the heteroscedasticity increases at constant ϕ, the GLS and FGLS clearly outperform the OLS-HAC in terms of TPR. For OLS-HAC the TPR remains constant (at 0.75 when ϕ=0.3), whereas for GLS and, to a lesser extent FGLS, the TPR increases (up to 0.83 when ϕ=0.3). This is explained by the fact that GLS and FLGS take the weight of every observation into account (when heteroscedasticity is strong, observations with small errors have high weight and this leads to a decrease in $\hat{\sigma}_{\hat{\delta}}$). When both heteroscedasticity and autocorrelation are strong, the power of the test is decreased, but FLGS performs better than OLS-HAC.

The power of the test is also assessed with respect to the jump amplitude and sample size in Figure 5 for a typical noise configuration (AR(1) process with ϕ=0.3, mean noise variance of 1 and half-range of variation of 80 %). For a typical sample size of n=600, a 75 % probability of detection is achieved for jumps of 0.25 or larger. Most of the G-E, G-E', and G-G' series in the real data actually fulfill this requirement (see Figure 6, further discussed in the next Section). For stronger noise (usually due to larger distance) or smaller jumps, a larger sample size would be required to maintain a high detection probability.

3.3 | Application to real data

The FGLS procedure was applied to the 114 change-points from 49 main stations, which could be tested with respect to nearby stations for which the segments of the difference series on the left and on the right of the change-points had a minimum of length of 200. The length of the difference series was also limited to 1000, in order to speed up the data processing. Figure 6 shows the distribution of the estimated jump amplitudes for each of the six series. The three series involving G have significantly larger amplitudes (around 0.3 kg m^{-2}) than the other three series, whatever the distance. This result suggests that the jumps are mainly in the G series rather than in the E, E', or G' series. This is a first gain of knowledge offered by the combination of the 4 series compared to the information that we have from the segmentation only which cannot distinguish jumps due to G and E. Stronger evidence is brought by considering

the test results, when the jump amplitudes are confronted to the critical level at α=0.05.

Figure 7 shows the corresponding test results (coded by $-1, 0$ and 1 for a significant test with a negative jump, a insignificant test, and a significant test with a positive jump, respectively), sorted by distance. 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 test and the segmentation results. Almost all G-E' jumps are significant as well. This result, combined with the result that E-E' don't have significant jumps in general, confirms our second rule (E and E' are expected to change simultaneously). Most G-G' jumps are significant, which to the extent that they are consistent on a one-by-one basis with the G-E results, would confirm the idea that most jumps are in G. Finally, the G'-E' and G'-E jumps are most of the time insignificant, which is a nice supportive confirmation of our first rule (G and G' are unlikely to change simultaneously). When the distance increases, we see that the conclusions remain unchanged, just the proportion of insignificant jumps is increasing, as a result of the increase of the noise in the series.

4 | PREDICTIVE RULE

Once a detected change-point in the series G-E of a main station is tested as significant, it is also tested in the rest of the five series of differences composed with the nearby station. Our objective is to build a classifier $\psi(x)$ representing the prediction of the configuration y, i.e. the value of the quadruplet composed of G, E, G', and E', given x that is the vector of the five resulting test statistics. To this aim, we propose to construct a complete data set containing both y and x from the test results of the real data and to compare four popular classifiers.

4.1 | Preliminary considerations

Final configurations. In the truncated Table 4, among the 38 selected configurations, 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 contains a change-point in G. This reduces the total number of configurations to $C = 36$.

Used test results data. Denote by $z_\ell = (z_{\ell 1}, \ldots, z_{\ell 5})$ the vector of the five test statistics for the ith test in the five series of difference 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. The results of all the nearby stations for a given change-point in a given main station can be viewed as replicates and the number of couples of (main station, change-point) is small (114).

The four considered learning algorithms are: the linear discriminant analysis (LDA), the classification and regression trees (CART) [35], the Random Forest (RF) [36] and the k Nearest Neighbors (kNN). 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`.

Building of the complete data set. This data set must contain a certain number of examples (x, y) for each configuration. We propose to build it from the original data set. The difficulty lies in the fact that the test result data set is small according to different considerations: (1) all the configurations will not be represented in the original data set, (2) they may be severely imbalanced, which is well-known to produce biased classifiers for the minor configurations, and (3) the number of available data in classes $-1, 0, 1$ for each series is small (see Figure 8). We thus propose to use a bootstrapping method which consists in randomly sampling the data with replacement [37] for each series of difference,

independently, to circumvent problem (1) and such that the number of replicates is the same for each configuration to circumvent problem (2). Problem (3) will be considered in the next Section. More precisely, for each configuration y and each series of difference j, we create R samples x by resampling among the test statistics $(z_{\ell j})_\ell$ which concludes the coded result of this configuration in the truncated table with a level of 5% ($-1, 0$ or 1). The constructed data set is noted $D = \{(y_\ell, x_\ell)_\ell\}_{\ell=1,\ldots,n}$ of size $n = C \times R$.

4.2 | The proposed Cross-Validation Bootstrap (CVB) procedure

Cross-validation is a popular statistical technique to test a classifier which involves splitting the data into two data subsets: the training 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 using bootstrapping from the original data set Z, which is small and repeated, the training 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 training and test subsets before constructing the complete data set D. The proposed CVB procedure is described in Algorithm 1.

Data: the original data Z

for $b = 1$ to B **do**

1. sample a training data set $Z^{b,L}$ from Z with probability 0.8, and form the test data set $Z^{b,T}$ with the remaining data. The random sampling is performed on the rows of Z, i.e. 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% in each configuration (i.e. for each configuration, $D^{b,L}$ contains $0.8R$ samples, and $D^{b,T}$ $0.2R$);

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

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

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

end

Averaging: compute the mean misclassification error for each classifier

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

Algorithm 1: The CVB procedure.

Table 3 gives the mean misclassification error for the four considered classifiers with $B = 20$ and $R = 100$. The random forest algorithm outperforms the others. We chose this algorithm and select as the final predictive rule the best one among the B resamplings denoted $\widehat{\psi}$. The predictive power of the five series of difference based on the accuracy criterion are in the decreasing order: E-E', G-G', G-E', G'-E and G'-E'.

4.3 | Application to the real data set.

We aim at predicting the configuration for each change-point of each main station. We apply the retained classifier to each test result of the original data set Z. When several nearby stations are available for a change-point in a main

station, we propose the following weighted prediction score: for a configuration c,

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

where the weights are $w_{ns} = 1/d_{ns}$ with d_{ns} is the distance of the nearby station ns from the main one, and the final configuration is the one with the highest score

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

Figure 8 shows the distribution of the predicted configurations. Most of the change-points are attributed to G (c=1 and 15), to (G, E, E') (c=31 and 35), and to (E, E') (c=10 and 23). These are also the six configurations with the highest conditional and joint probabilities in Table 4. This demonstrates that the classifier is able to chose the configurations which we believe are the most likely in the real data. Recall that the probabilities in Table 4 are not used for the construction of the predictive rule, but they are intrinsic in the data. All the other configurations shown in the figure have low conditional (Pc) and joint probabilities (Pj) such as: Pc=0.0025 for the group labelled "other", and Pc=0.045 for all other configurations. The only unclear result is with the group E (c=8 and 22) which occurs relatively often (8 times out of 114). According to our second rule, a change-point in E should also be seen in E'. Inspection of the time series and the test results from all the nearby for these 8 cases showed one common feature which is that all these test results are 0 (insignificant) for the series G-G' and G-E'. Configurations c=8 and 22 are indeed the only ones among the 38 configurations in Table 4 that have this particularity. The reason why the tests are insignificant for G-G' and G-E' is because they are generally associated with far nearby stations, thus with high noise levels, and low detection power.

Overall, 87 breakpoints are assigned to G (including configurations 1, 15, 29, etc.), of which 37 occur within +/- 2 months of known equipment changes from the GNSS metadata (i.e. a validation ratio of 42 %).

Figure 9 illustrates the result for the change-point on October, 4, 2017, at the main station FAIR (Fairbanks, Alaska) and nearby station CLGO within a distance of 21 km. At that distance, the variances of all the six differences are similar, except for the E-E', which is very small because the same grid points form the reanalysis are used. Significant downward jumps are detected in G-E, G-G' and G-E' with high T-values. A small but significant downward jump is also detected in G'-E', while in the G'-E' and G'-E, the jumps are insignificant. The result of six tests is thus (-1, -1, -1, 0, -1, 0), which does not exist in the logical table 4. The predictive rule for this case selects configuration c=35, corresponding to the quadruplet (G, E, G', E')=(-1, 1, 0, 1), i.e. a downward jump in the G and an upward jump in the E and E'. This jump in G is actually confirmed from the GNSS metadata, which report a receiver hardware change on October, 6, 2017. This is an obvious demonstration that the prediction rule is successful in attributing a GNSS change-point, even when the initial test result is not in the logical table.

5 | CONCLUSIONS AND PERSPECTIVES

The detection and attribution of change-points is an important task for the homogenization of climatic time series. Especially, in a differential segmentation scheme, the attribution of the detected change points to the target vs. the reference series is a crucial step before the data correction.

The attribution method proposed in this paper uses the ERA5 reanalysis data (E) as the reference series and the GNSS IWV data (G) as the target series. The method consists in the following steps:

1. **Data selection and pre-processing.** For each detected change-point in a "main" station, select all nearby stations within a given range (distance smaller than 200 km, height difference smaller than 500 m) and run the segmentation algorithm on the nearby differenced series (G'-E') in order to detect the change-points in these data as well. For each nearby station, correct the data for the height difference, so that G, E, G', and E' are expressed at the same height, form the six series of differences (G-E, G-G', etc.), and screen the series (i.e., remove the outliers and unnecessary change-points in clusters).
2. **Test the significance of the change in the mean.** For each of the six series, fit a regression model to the portion of data on the left and right of the change-point in the main station using an iterative FGLS procedure and test the significance of the jump (change in the mean). The regression test is repeated for all the nearby attached to the same change-point.
3. **Use a predictive rule to attribute the configuration.** For each nearby, the trained classifier will predict the configuration (the quadruplet composed of G, E, G', and E') corresponding to the six test results. When several nearby stations are available, a weighted prediction score is computed to select the final configuration.

The different steps in the method have been verified, tuned, and trained (for the predictive rule) by investigating a real data set of 49 main stations and 312 nearby stations. The data characterization showed that the data have a strong heteroscedasticity (seasonal variation in noise standard deviation around 70%) and moderate autocorrelation (with a typical lag-1 correlation coefficient of 0.3). These features are modelled in the FGLS test procedure to ensure correct inference. Several classifiers have been compared for the predictive rule and the Random Forest was selected as it outperformed clearly the others.

To our knowledge, both the FGLS regression test approach and the Random Forest classifier have never been used in the context of climate series homogenization. We emphasize that the attribution method described in this paper is independent from the segmentation package. It can be used as a post-processing step of any segmentation results based on a differential approach (target minus reference series) with multiple stations.

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. They showed a clear predominance of significant jumps in the series involving G (41%), (G, E, E') (25%), and (E, E') (15%), as expected. The remaining 19% of unexpected results are thought the be linked with low detection power of the FGLS test and possibly random errors in the classification. Some possibilities to improve both steps in the method are described below.

The main perspectives of this work are: i) to test the proposed method on a bigger data set to confirm the performance, ii) to try to improve the robustness and the power of the test procedure by refining the nearby selection rules (e.g. set a limit on the percentage of gaps in the series, shorten the distance criterion), and check the normality of the T-statistic (e.g. an alternative to using the normal distribution would be to fit the distribution from the real data and use it to obtain empirical critical values), iii) improve the classification method (e.g. train it on a big simulated data set), iv) refine the aggregative rule (currently based on the distance and number of occurrences of a configuration) when several nearby stations are available (e.g. take into account the value of the T-statistic, the sample size, the prior probability from Table 4). The two tasks iii) and iv) could be combined to improve the global predictive rule (e.g. using a bigger data set, adding information about the distance, the mean noise level, etc).

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).

conflict of interest

The authors declare that they have no conflict of interest.

references

[1] Dunn RJH, Aldred F, Gobron N, Miller JB, Willett KM, Ades M, et al. Global Climate. Bulletin of the American Meteorological Society 2021 Aug;102(8):S11–S142. https://doi.org/10.1175/bams-d-21-0098.1.

[2] Jones PD, Raper SCB, Bradley RS, Diaz HF, Kellyo PM, Wigley TML. Northern Hemisphere Surface Air Temperature Variations: 1851–1984. J Clim Appl Meteorol 1986;25(2):161–179.

[3] Easterling DR, Peterson TC. A new method for detecting undocumented discontinuities in climatological time series. Int J Climatol 1995;15:369–377.

[4] Peterson TC, Easterling DR, Karl TR, Groisman P, Nicholls N, Plummer N, et al. Homogeneity adjustments of in situ atmospheric climate data: A review. Int J Climatol : A J R Meteorol Soc 1998;18(13):1493–1517.

[5] Caussinus H, Mestre O. Detection and correction of artificial shifts in climate series. J R Stat Soc : Ser C (Appl Stat) 2004;53(3):405–425. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2004.05155.x.

[6] Menne MJ, Williams CN. Detection of Undocumented Changepoints Using Multiple Test Statistics and Composite Reference Series. J Clim 2005;18(20):4271–4286.

[7] Szentimrey T. Development of MASH homogenization procedure for daily data. Proceedings of the fifth seminar for homogenization and quality control in climatological databases. WCDMP-No 71 2008;p. 123–130. http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP71.pdf.

[8] Reeves J, Chen J, Wang XL, Lund R, Lu QQ. A Review and Comparison of Changepoint Detection Techniques for Climate Data. J Appl Meteorol Climatol 2007;46(6):900–915. https://doi.org/10.1175/JAM2493.1.

[9] Costa AC, Soares A. Homogenization of Climate Data: Review and New Perspectives Using Geostatistics. Mathematical Geosciences 2009 Apr;41(3):291–305. https://doi.org/10.1007/s11004-008-9203-3.

[10] Venema VKC, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, et al. Benchmarking homogenization algorithms for monthly data. Clim Past 2012;8(1):89–115. https://www.clim-past.net/8/89/2012/.

[11] Domonkos P. Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT). Int J Geosci 2011;02(03):293–309. https://doi.org/10.4236/ijg.2011.23032.

[12] Quarello A. Development of new homogenisation methods for GNSS atmospheric data. Application to the analysis of climate trends and variability. Theses, Sorbonne Universite ; IGN (Institut National de l'Information Géographique et Forestière); 2020.

[13] Ning T, Wickert J, Deng Z, Heise S, Dick G, Vey S, et al. Homogenized Time Series of the Atmospheric Water Vapor Content Obtained from the GNSS Reprocessed Data. J Clim 2016;29(7):2443–2456.

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

[15] Van Malderen R, Pottiaux E, Klos A, Domonkos P, Elias M, Ning T, et al. Homogenizing GPS Integrated Water Vapor Time Series: Benchmarking Break Detection Methods on Synthetic Data Sets. Earth Space Sci 2020;7(5):e2020EA001121. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020EA001121, e2020EA001121 2020EA001121.

[16] Nguyen KN, Quarello A, Bock O, Lebarbier E. Sensitivity of Change-Point Detection and Trend Estimates to GNSS IWV Time Series Properties. Atmosphere 2021 Aug;12(9):1102. https://doi.org/10.3390/atmos12091102.

[17] Quarello A, Bock O, Lebarbier E. GNSSseg, a Statistical Method for the Segmentation of Daily GNSS IWV Time Series. Remote Sensing 2022 Jul;14(14):3379. https://doi.org/10.3390/rs14143379.

[18] Vey S, Dietrich R, Fritsche M, Rülke A, Steigenberger P, Rothacher M. On the homogeneity and interpretation of precipitable water time series derived from global GPS observations. J Geophys Res : Atmos 2009;114(D10).

[19] Bock O, Willis P, Wang J, Mears C. A high-quality, homogenized, global, long-term (1993–2008) DORIS precipitable water data set for climate monitoring and model verification. J Geophys Res : Atmos 2014;119(12):7209–7230.

[20] Parracho AC, Bock O, Bastin S. Global IWV trends and variability in atmospheric reanalyses and GPS observations. Atmos Chem Phys 2018;18(22):16213–16237. https://www.atmos-chem-phys.net/18/16213/2018/.

[21] Schroeder M, Lockhoff M, Forsythe JM, Cronk HQ, Haar THV, Bennartz R. The GEWEX Water Vapor Assessment: Results from Intercomparison, Trend, and Homogeneity Analysis of Total Column Water Vapor. J Appl Meteorol Climatol 2016 Jul;55(7):1633–1649. https://doi.org/10.1175/jamc-d-15-0304.1.

[22] Bock O, Global GNSS IWV data at 436 stations over the 1994-2018 period; 2019.

[23] Blewitt G, Hammond W, Kreemer C. Harnessing the GPS Data Explosion for Interdisciplinary Science. Eos 2018 Sep;99. https://doi.org/10.1029/2018eo104623.

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

[25] Yohai VJ, Zamar RH. High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale. Journal of the American Statistical Association 1988;83:406–413.

[26] Bock O, Parracho A. Consistency and representativeness of integrated water vapour from ground-based GPS observations and ERA-Interim reanalysis. Atmos Chem Phys 2019;19:9453–9468.

[27] Shumway RH, Stoffer DS. Time Series Analysis and Its Applications. Springer International Publishing; 2017. https://doi.org/10.1007/978-3-319-52452-8.

[28] Mixed-Effects Models in S and S-PLUS. Springer-Verlag; 2000. https://doi.org/10.1007/b98882.

[29] White HL. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica 1980;48:817–838.

[30] Newey W, West KD. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelationconsistent Covariance Matrix. Econometrics eJournal 1986;.

[31] Andrews DWK. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. Econometrica 1991;59:817–858.

[32] Zeileis A. Object-oriented Computation of Sandwich Estimators. Journal of Statistical Software 2006;16:1–16.

[33] Hyndman RJ, Khandakar Y. Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software 2008;27:1–22.

[34] Zwiers FW, von Storch H. Taking Serial Correlation into Account in Tests of the Mean. Journal of Climate 1995;8:336–351.

[35] Breiman L, Friedman J, Olshen R, Stone C. Cart. Classification and Regression Trees 1984;.

[36] Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol. 1 IEEE; 1995. p. 278–282.

[37] Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.

List of Figures

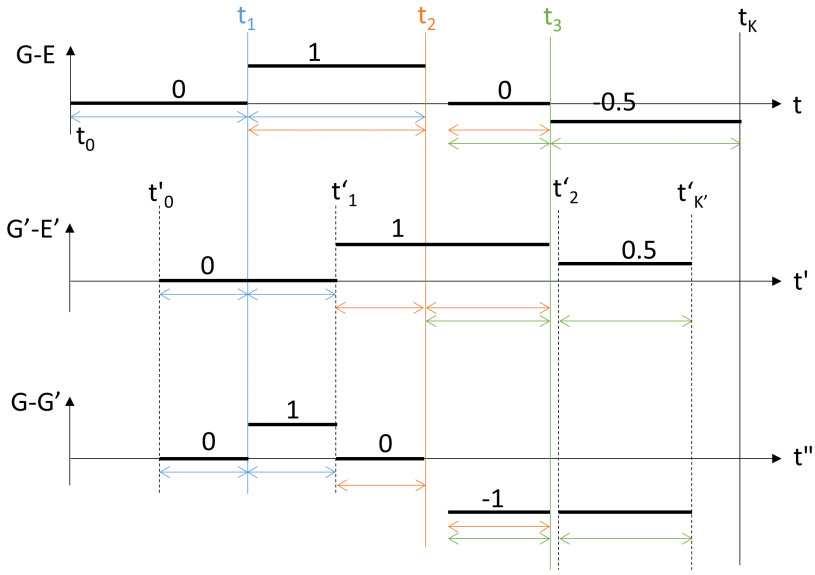

FIGURE 1 Schematic view of three series of differences. Top: G-E, middle: G'-E', bottom: G-G', where the prime symbol indicates the series from the nearby station (the other series are from the main station), G stands for GNSS and E for ERA5 in reference to the real data analysis in this paper. Change-points detected by the segmentation algorithm in the main (nearby) station are indicated by the solid (dotted) lines. The colored horizontal lines with arrows indicate the segments on the left and the right in each series used to test the change-points.

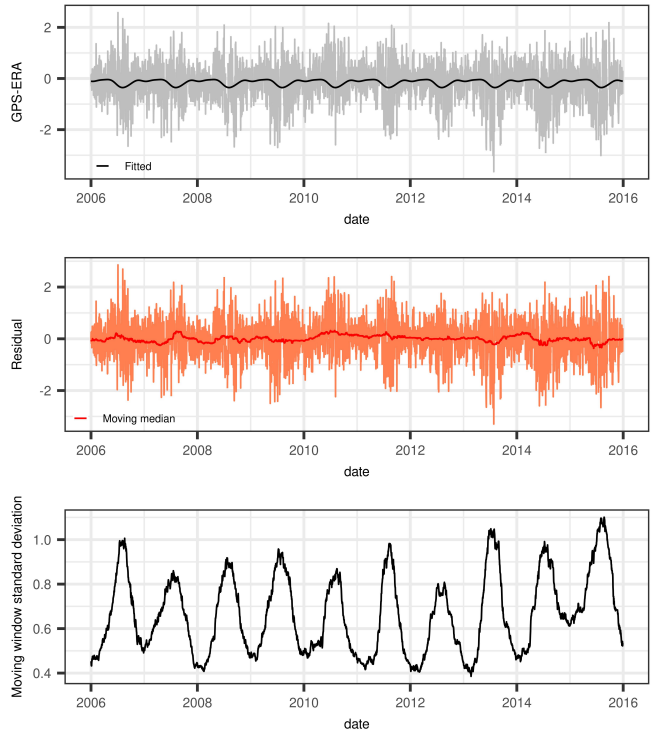

FIGURE 2 Top: GNSS minus ERA5 time series at station ALBH (Victoria, Canada), in gray, and estimated Fourier series, in black, for a long, homogeneous, segment (no change-point detected by the segmentation algorithm). Middle: regression residuals and moving median. Bottom: moving standard deviation illustrating the strong heteroscedasticity in these data.

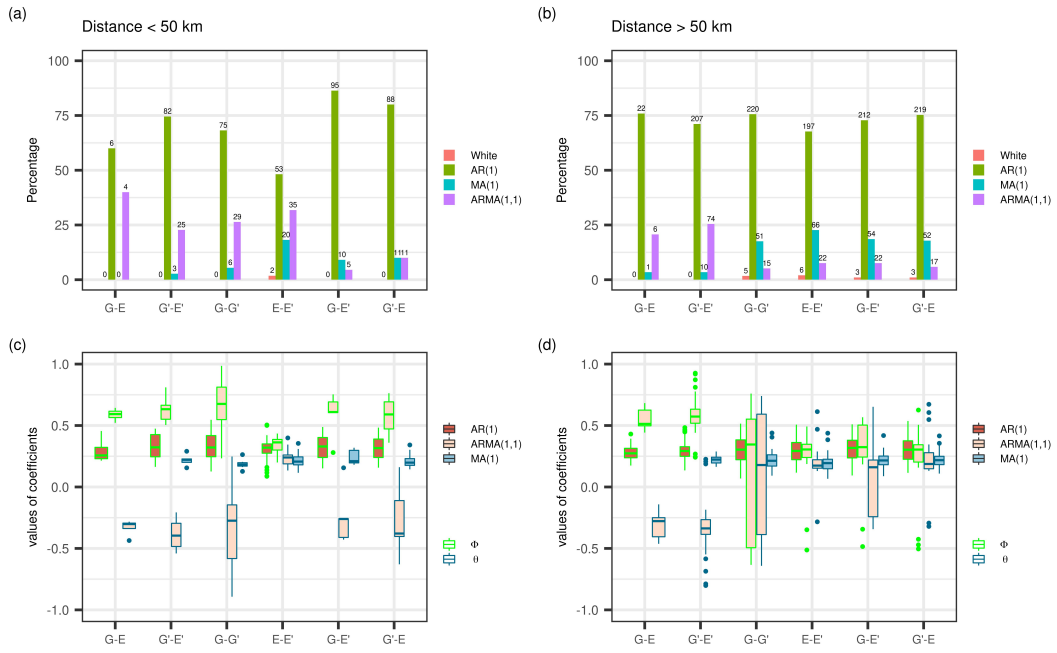

FIGURE 3 Results of noise model identification in the real data. Top: Histogram of model types (white noise, AR(1), MA(1), ARMA(1,1)) selected with `auto.arima` function for each of the six series of differences (G-E, G-E', etc.); the bar heights show the percentage (y-axis) of cases for each series, the number of cases is indicated on the top. Bottom: noise model coefficients, $\hat{\phi}$ and $\hat{\theta}$, estimated with `arima` function, for each model. Results are sorted according to the distance between the main and the nearby stations, left: smaller than 50 km, right: larger than 50 km.

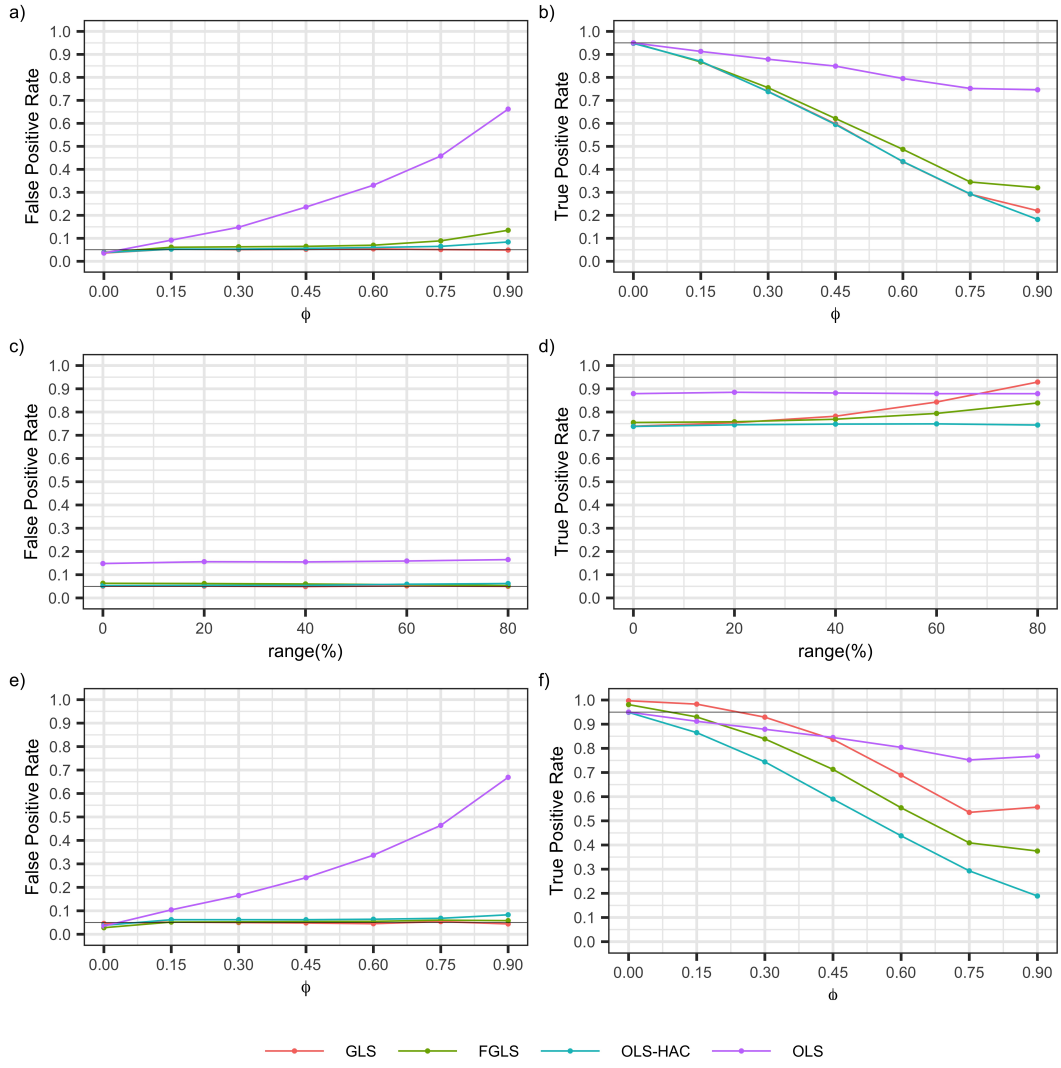

FIGURE 4 False Positive Rate (FPR) and True Positive Rate (TPR) from simulations with m=1000 replications, for three scenarios: (a, b) AR(1) noise of unit variance with $\phi = 0, \ldots, 0.90$; (c, d) heteroskedastic and AR(1) noise, with $\phi = 0.3$ and half-range of variance going from 0 to 80 %; (e, f) heteroskedastic and AR(1) noise, with $\phi = 0, \ldots, 0.90$, and half-range of variance of 80 %. For TPR, the amplitude of the jump is fixed to 0.356, which corresponds to TPR=0.95 when $\phi = 0$. The sample size is n=400.

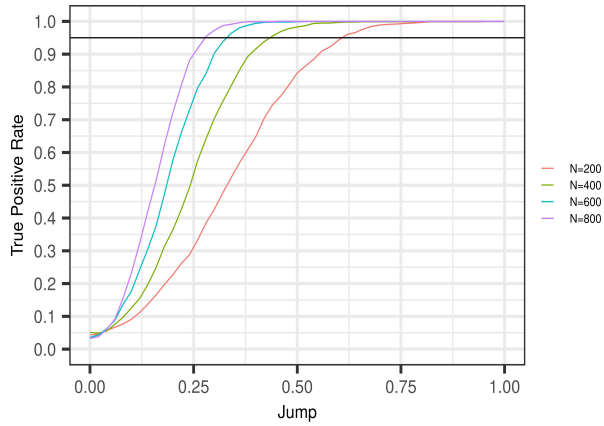

FIGURE 5 True Positive Rate (TPR) of FGLS from simulations, as a function of jump amplitude and sample size, in case of heteroskedastic and AR(1) noise with $\phi = 0.3$, mean variance of 1, and half-range of variance of 80%.

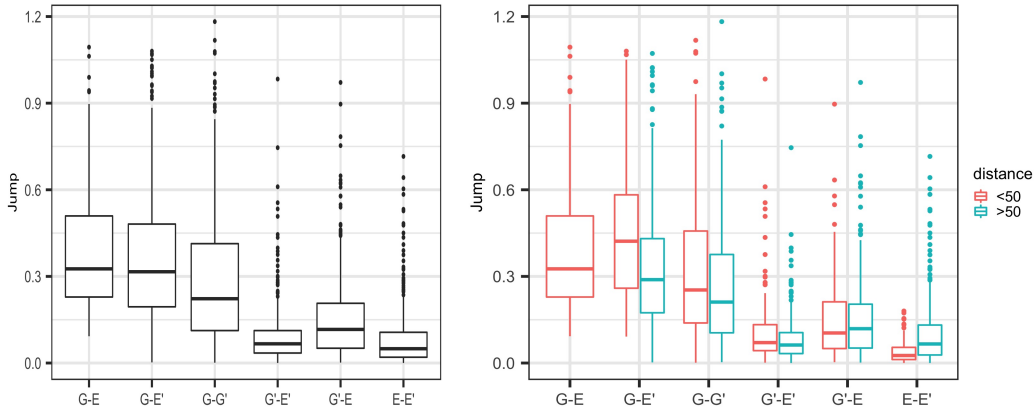

FIGURE 6 Distribution of jump amplitudes in the real data estimated by FGLS: (left) all results, (right) results sorted by distance between main and nearby stations.

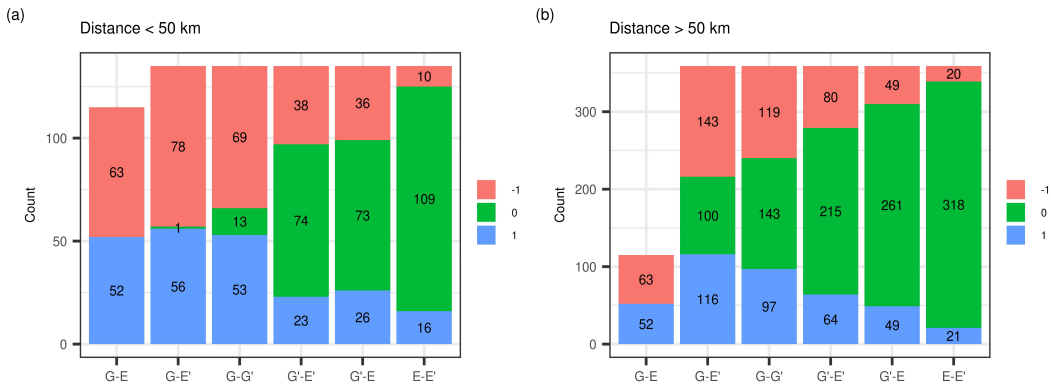

FIGURE 7 Distribution of test results associated to the estimated amplitudes of jumps shown in Figure 6, sorted by distance: (left) smaller than 50 km, (right) larger than 50 km. Test results are coded as: 0 insignificant, −1 or 1 significant downward or upward jump.

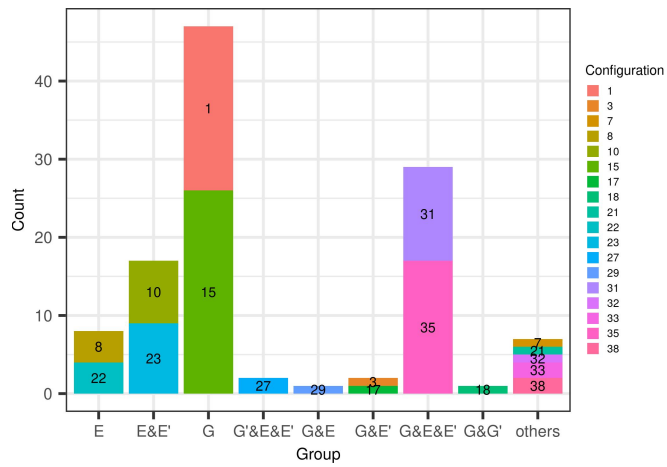

FIGURE 8 Distribution of the final predicted configurations from the real data with the Random Forest method corresponding to a jump only in E, or in E and in E', or one in G, etc.

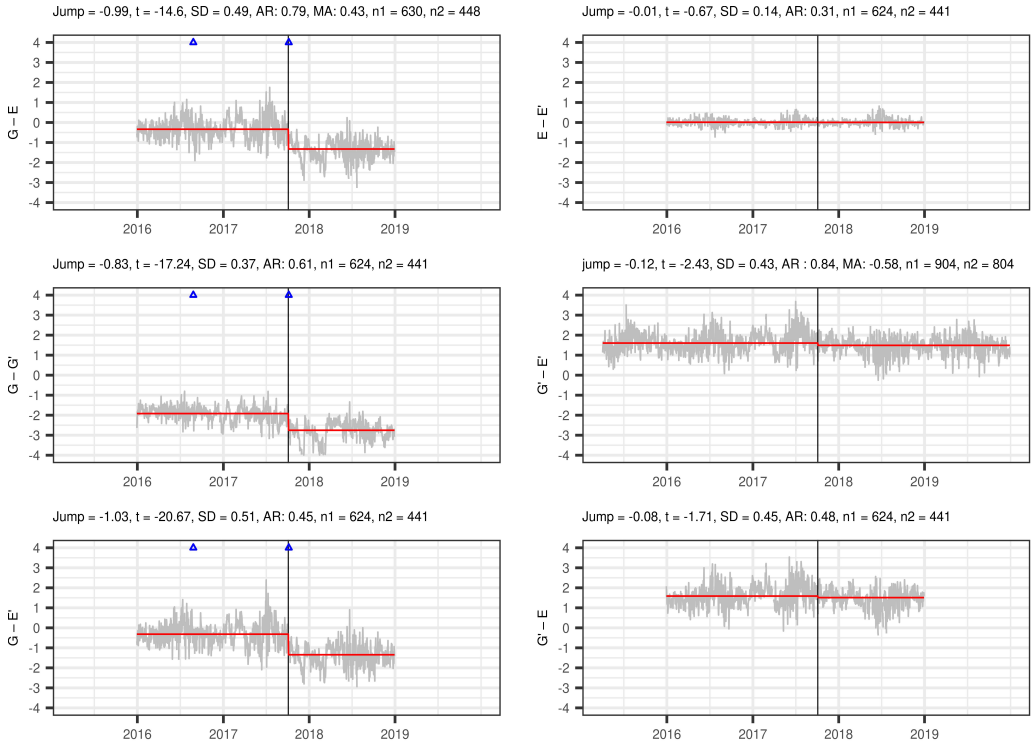

FIGURE 9 Example of test result for station FAIR (Fairbanks, Alaska) with nearby station CLGO at a distance of 21 km. The series of IWV differences are shown in gray. The black vertical solid line shows the tested change-point (04 October 2017). The blue triangles indicate known equipment changes in the main station from the GNSS metadata. The horizontal red lines show the means estimated by the FGLS regression on the left and the right of the change-point.

	Mean of MSD		Range of MSD (%)
Distance	< 50 km	> 50 km	
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

*MSD: Moving standard deviation

TABLE 1 Heteroskedasticity of the data set.

Distance	<50km				>50km			
series	AR(1)	MA(1)	ARMA(1,1)		AR(1)	MA(1)	ARMA(1,1)	
Coefficients	phi	theta	phi	theta	phi	theta	phi	theta
G-E	0.30	0.00	0.59	-0.33	0.28	0.00	0.55	-0.30
G'-E'	0.33	0.22	0.61	-0.38	0.30	0.22	0.58	-0.33
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

TABLE 2 Autoregressive noise structure of the data set.

c	LDA	CART	KNN	RF
$\overline{\text{err}}^c$	0.1463 ± 0.021	0.0142 ± 0.011	0.1412 ± 0.018	0.0049 ± 0.003

TABLE 3 Mean misclassification error, with the standard deviation in brackets, for the four classifiers.

A | LOGICAL TABLE

Table 4 presents the expected test results of the six series of differences (G-E, G-G', etc.) for 54 configurations of the base series (G, E, G', E'). Each quadruplet (G, E, G', E') is characterized by conditional and joint probabilities described in Appendix B. Duplicates are sorted out according to the joint probabilities, leaving 46 combinations in the "logical table" and 38 combinations in the "truncated table".

B | COMPUTATION OF PRIOR PROBABILITIES

This section explains how we compute the conditional and joint probabilities reported in the Table 4.

Let G, E, G', and E' represent the event "there is a jump" in the four base series, with the possible outcomes: 0=no jump, -1=the jump is downward, +1=the jump is upward. Let $P(G'|G)$ represent the conditional probability of G' given G. Let us assume that G and E are independent, and that G' and E' are independent. The two rules that we stated in the Introduction are here quantified in terms of probabilities:

(R1) $P(G' \neq 0 | G \neq 0 \cup E \neq 0) = 0.1$.
(R2) $P(E = E') = 0.9$.

The first probability actually accounts for two incompatible events: $(G' \neq 0 | G \neq 0, E = 0)$ and $(G' \neq 0 | G = 0, E \neq 0)$. We can suppose that they are both of similar probability $p_1 = 0.05$. It follows from the first rule that $P(G' = 0 | G \neq 0 \cup E \neq 0) = 0.9$. It follows from the second rule that $P(E' \neq E) = 0.1$, which also corresponds to two incompatible events $(E' = 0 | E = -1)$ and $(E' = 0 | E = +1)$, the probability of which we will also assume to be equal to $p_2 = 0.05$.

The conditional probabilities $P(G', E' | G, E)$ are obtained from the products of the individual probabilities $P(G', E' | G, E) = P(G' | G, E) \times P(E' | G, E)$. These probabilities are sufficient to sort the duplicates in first 36 rows. For example, rows 1 and 18 have the same 6 results in the Logical Table. However, the first case corresponds to the quadruplet $(G, E, G', E') = (+1, 0, 0, 0)$, while the second one corresponds to the conjugate quadruplet, $(G, E, G', E') = (0, -1, -1, -1)$. The latter represents the situation where the change-point in the main station is due to E, which is possible, but change-points are detected simultaneously in G' and E', which is quite unlikely. The probabilities associated to these two quadruplets reflect this: 0.81 in the first case and 0.045 in the second case.

In order to distinguish the duplicates in the last 18 rows of the Truncated Table, we introduce the prior probabilities associated to the events G and E. There are six different combinations: $(G, E) \in \{(+1, 0), (0, -1), (-1, 0), (0, -1), (+1, -1), (-1, +1)\}$. In the first 4 combinations, a single change occurs in either G or E, while in the latter 2, two changes occur simultaneously. The former is more likely, so we attribute it a high probability of $p_3 = 0.24$, from which it results that the latter has a low probability of $p_4 = 0.02$ (the value of p_4 is deduced from the value chosen for p_3 in order to have a sum $4 \times p_3 + 2 \times p_4 = 1$).

The joint probabilities reported in Table 4 are obtained from the product of the conditional and the prior probabilities: $P(G, E, G', E') = P(G', E' | G, E) \times P(G, E)$.

	Truth				Conditional probability	Joint probability		Logical table						Truncated table						
	G	E	G'	E'	P(G', E' | G, E)	P(G, E, G', E')		G-E	G-G'	G-E'	E-E'	G'-E'	G'-E		G-E	G-G'	G-E'	E-E'	G'-E'	G'-E	
1	1	0	0	0	0,81	0,18225	1	1	1	1	0	0	0	1	1	1	1	0	0	0	1
2	1	0	0	1	0,045	0,010125	2	1	1	0	-1	-1	0	2	1	1	0	-1	-1	0	2
3	1	0	0	-1	0,045	0,010125	3	1	1	2	1	1	0	3	1	1	1	1	1	0	3
	1	0	1	0	0,045	0,010125	4	1	0	1	0	1	1		1	0	1	0	1	1	
	1	0	1	1	0,0025	0,0005625	5	1	0	0	-1	0	1		1	0	0	-1	0	1	
4	1	0	1	-1	0,0025	0,0005625	6	1	0	2	1	2	1	4	1	0	1	1	1	1	4
5	1	0	-1	0	0,045	0,010125	7	1	2	1	0	-1	-1	5	1	1	1	0	-1	-1	5
6	1	0	-1	1	0,0025	0,0005625	8	1	2	0	-1	-2	-1	6	1	1	0	-1	-1	-1	6
7	1	0	-1	-1	0,0025	0,0005625	9	1	2	2	1	0	-1	7	1	1	1	1	0	-1	7
8	0	-1	0	0	0,045	0,010125	10	1	0	0	-1	0	1	8	1	0	0	-1	0	1	8
9	0	-1	0	1	0,045	0,010125	11	1	0	-1	-2	-1	1	9	1	0	-1	-1	-1	1	9
10	0	-1	0	-1	0,81	0,18225	12	1	0	1	0	1	1	10	1	0	1	0	1	1	10
11	0	-1	1	0	0,0025	0,0005625	13	1	-1	0	-1	1	2	11	1	-1	0	-1	1	1	11
12	0	-1	1	1	0,0025	0,0005625	14	1	-1	-1	-2	0	2	12	1	-1	-1	-1	0	1	12
13	0	-1	1	-1	0,045	0,010125	15	1	-1	1	0	2	2	13	1	-1	1	0	1	1	13
	0	-1	-1	0	0,0025	0,0005625	16	1	1	0	-1	-1	0		1	1	0	-1	-1	0	
14	0	-1	-1	1	0,0025	0,0005625	17	1	1	-1	-2	-2	0	14	1	1	-1	-1	-1	0	14
	0	-1	-1	-1	0,045	0,010125	18	1	1	1	0	0	0		1	1	1	0	0	0	

	Truth				Conditional probability	Joint probability		Logical table						Truncated table						
	G	E	G'	E'	P(G', E' | G, E)	P(G, E, G', E')		G-E	G-G'	G-E'	E-E'	G'-E'	G'-E		G-E	G-G'	G-E'	E-E'	G'-E'	G'-E	
15	-1	0	0	0	0,81	0,18225	19	-1	-1	-1	0	0	0	15	-1	-1	-1	0	0	0	15
16	-1	0	0	1	0,045	0,010125	20	-1	-1	-2	-1	-1	0	16	-1	-1	-1	-1	-1	0	16
17	-1	0	0	-1	0,045	0,010125	21	-1	-1	0	1	1	0	17	-1	-1	0	1	1	0	17
18	-1	0	1	0	0,045	0,010125	22	-1	-2	-1	0	1	1	18	-1	-1	-1	0	1	1	18
19	-1	0	1	1	0,0025	0,0005625	23	-1	-2	-2	-1	0	1	19	-1	-1	-1	-1	0	1	19
20	-1	0	1	-1	0,0025	0,0005625	24	-1	-2	0	1	2	1	20	-1	-1	0	1	1	1	20
	-1	0	-1	0	0,045	0,010125	25	-1	0	-1	0	-1	-1		-1	0	-1	0	-1	-1	
21	-1	0	-1	1	0,0025	0,0005625	26	-1	0	-2	-1	-2	-1	21	-1	0	-1	-1	-1	-1	21
	-1	0	-1	-1	0,0025	0,0005625	27	-1	0	0	1	0	-1		-1	0	0	1	0	-1	
22	0	1	0	0	0,045	0,010125	28	-1	0	0	1	0	-1	22	-1	0	0	1	0	-1	22
23	0	1	0	1	0,81	0,18225	29	-1	0	-1	0	-1	-1	23	-1	0	-1	0	-1	-1	23
24	0	1	0	-1	0,045	0,010125	30	-1	0	1	2	1	-1	24	-1	0	1	1	1	-1	24
	0	1	1	0	0,0025	0,0005625	31	-1	-1	0	1	1	0		-1	-1	0	1	1	0	
	0	1	1	1	0,045	0,010125	32	-1	-1	-1	0	0	0		-1	-1	-1	0	0	0	
25	0	1	1	-1	0,0025	0,0005625	33	-1	-1	1	2	2	0	25	-1	-1	1	1	1	0	25
26	0	1	-1	0	0,0025	0,0005625	34	-1	1	0	1	-1	-2	26	-1	1	0	1	-1	-1	26
27	0	1	-1	1	0,045	0,010125	35	-1	1	-1	0	-2	-2	27	-1	1	-1	0	-1	-1	27
28	0	1	-1	-1	0,0025	0,0005625	36	-1	1	1	2	0	-2	28	-1	1	1	1	0	-1	28

	Truth				Conditional probability	Joint probability		Logical table						Truncated table						
	G	E	G'	E'	P(G', E' | G, E)	P(G, E, G', E')		G-E	G-G'	G-E'	E-E'	G'-E'	G'-E		G-E	G-G'	G-E'	E-E'	G'-E'	G'-E	
29	1	-1	0	0	0,045	0,00225	37	2	1	1	-1	0	1	29	1	1	1	-1	0	1	29
30	1	-1	0	1	0,045	0,00225	38	2	1	0	-2	-1	1	30	1	1	0	-1	-1	1	30
31	1	-1	0	-1	0,81	0,0405	39	2	1	2	0	1	1	31	1	1	1	0	1	1	31
32	1	-1	1	0	0,0025	0,000125	40	2	0	1	-1	1	2	32	1	0	1	-1	1	1	32
	1	-1	1	1	0,0025	0,000125	41	2	0	0	-2	0	2	33	1	0	0	-1	0	1	
	1	-1	1	-1	0,045	0,00225	42	2	0	2	0	2	2	34	1	0	1	0	1	1	
33	1	-1	-1	0	0,0025	0,000125	43	2	2	1	-1	-1	0	35	1	1	1	-1	-1	0	33
	1	-1	-1	1	0,0025	0,000125	44	2	2	0	-2	-2	0	36	1	1	0	-1	-1	0	
	1	-1	-1	-1	0,045	0,00225	45	2	2	2	0	0	0	37	1	1	1	0	0	0	
34	-1	1	0	0	0,045	0,00225	46	-2	-1	-1	1	0	-1	38	-1	-1	-1	1	0	-1	34
35	-1	1	0	1	0,81	0,0405	47	-2	-1	-2	0	-1	-1	39	-1	-1	-1	0	-1	-1	35
36	-1	1	0	-1	0,045	0,00225	48	-2	-1	0	2	1	-1	40	-1	-1	0	1	1	-1	36
37	-1	1	1	0	0,0025	0,000125	49	-2	-2	-1	1	1	0	41	-1	-1	-1	1	1	0	37
	-1	1	1	1	0,045	0,00225	50	-2	-2	-2	0	0	0	42	-1	-1	-1	0	0	0	
	-1	1	1	-1	0,0025	0,000125	51	-2	-2	0	2	2	0	43	-1	-1	0	1	1	0	
38	-1	1	-1	0	0,0025	0,000125	52	-2	0	-1	1	-1	-2	44	-1	0	-1	1	-1	-1	38
	-1	1	-1	1	0,045	0,00225	53	-2	0	-2	0	-2	-2	45	-1	0	-1	0	-1	-1	
	-1	1	-1	-1	0,0025	0,000125	54	-2	0	0	2	0	-2	46	-1	0	0	1	0	-1	

TABLE 4 Expected test results for 54 configurations of the quadruplets (G, E, G', E'): the "logical table" (in the middle) shows the results coded on 5 levels (-2, -1, 0, 1, 2), while the "truncated table" (on the right) shows the results coded on three levels (-1, 0, +1). Conditional and joint probabilities are associated each configuration (see Appendix B). Duplicates are highlighted with colored background.