Seismic Station Monitoring Using Deviation from the Gaussianity

Arthur Cuvier, Éric Beucler, Mickaël Bonnin, Raphaël Garcia

Seismic Station Monitoring using Deviation from the Gaussianity

Arthur Cuvier[1], Éric Beucler[1,2], Mickaël Bonnin[1,2] and Raphaël F. Garcia[3]

[1] Laboratoire de planétologie et géosciences, Nantes Université, Univ. d'Angers, Le Mans Univ., CNRS UMR-6112, Nantes, France

[2] Observatoire des sciences de l'Univers de Nantes Atlantique, CNRS UAR-3281, Nantes, France.

[3] Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), Toulouse, France

E-mail: arthur.cuvier@etu.univ-nantes.fr

This document corresponds to the submitted version of the manuscript.

The published version of the manuscript can be found here (*Seismological Research Letters*).

Abstract

Degradation of the seismic signal quality sometimes occurs at permanent and temporary stations. Although the most likely cause is a high level of humidity, leading to corrosion of the connectors, environmental changes can also alter recording conditions in different frequency ranges and not necessarily for all three components in the same way. Assuming that the continuous seismic signal can be described by a normal distribution, we present a new approach to quantify the seismogram quality and to point out any time sample that deviates from this Gaussian assumption. We introduce the notion of background Gaussian signal (BGS) to characterize a set of samples that follows a normal distribution. The discrete function obtained by sorting the samples in ascending order of amplitudes is compared to a modified probit function to retrieve the elements composing the BGS, and its statistical properties (mostly its standard deviation σ_G). As soon as there is any amplitude perturbation, σ_G deviates from the standard deviation of all samples composing the time window (σ). Hence, the parameter $\log(\frac{\sigma}{\sigma_G})$ directly quantifies the alteration level. For a given frequency range and a given component, the median of all $\log(\frac{\sigma}{\sigma_G})$ that can be computed using short time windows, reflects the overall gaussianity of the continuous seismic signal. We demonstrate that it can be used to efficiently monitor the quality of seismic traces by using this approach at four broadband permanent stations. We show that the daily $\log(\frac{\sigma}{\sigma_G})$ is sensitive to both subtle changes on one or two components as well as the signal signature of a sensor's degradation. Finally, we suggest that $\log(\frac{\sigma}{\sigma_G})$ and other parameters that are computed from the BGS bring useful information for station monitoring in addition to existing methods.

1 Introduction

Both permanent and temporary deployed seismometers can be degraded during their operating time (*e. g.* Ekstrom et al., 2006; Davis and Berger, 2007). Visual inspection of the daily signal at each station allows any alteration of the signal to be detected quickly, but is incompatible with limited observatory staff that can operate more than 50 stations. On the other hand, as the continuous seismic signal varies as a function of time and frequency, and not necessarily in the same way for the three components, a decision of physical intervention on site driven by an AI based on observables such as spectrograms is, to our knowledge, not fully operational yet. There is thus a need for simple but reliable parameters to efficiently monitor the seismic signal quality.

Though the noise level depends on location and installation conditions, a number of issues such as mass-centering failures, glitches, increases in instrument self-noise, or corroded components can alter the continuous seismic signal. It may also sometimes happen that the failure disappears and the signal returns to a satisfactory quality, so no one will know that a problem ever occurred. One of the well known origin of recording condition degradation can be found in a high level of humidity, leading to corrosion of the internal electronic system. Hutt and Ringler (2011) indicate that i) high humidity conditions can modify the response of the instrument and ii) water vapor and moisture in the electronics appears to explain many of the observed anomalies.

In the field of quality control which aims to rapidly detect any deterioration, progress have been made during the last years (*e. g.* McNamara and Boaz, 2010). One can note the emergence of several automatic methods for monitoring stations, as presented in Ringler et al. (2015) and Casey et al. (2018) but those approaches are mostly dedicated to the detection of other issues than a degradation of the seismic signal quality (signal continuity, data availability). Probability of power spectral densities (PPSD) can provide very useful information, but require sufficiently large time windows to detect changes over time. To evaluate the seismic data quality, a strategy consists in comparing signals recorded at col-

located sensors (Tasič, 2018), or at stations in close proximity (Kimura et al., 2015). This generally cannot be used for a permanent array with station inter-distances of about 50 km. Pedersen et al. (2020) present an innovative way to measure the quality of a single station, by comparing the standard deviation of the signal between the different components. Although this method appears to be efficient to detect malfunctions, it is not suitable for detecting signal degradation affecting all components simultaneously, and it seems difficult to define common thresholds that works for all stations.

In this article, we propose a novel approach based on the study of the seismic signal gaussianity to detect possible degradation of its quality. In section 2, we present a method allowing to discriminate, in any data set, the samples that can be considered as Gaussian, from the others (*i.e.* perturbed samples). Assuming that the seismic signal is intrinsically Gaussian (Groos and Ritter, 2009; Zhong et al., 2015b,a; Aggarwal et al., 2020), we perform in section 3 an analysis of the signal quality of the stations G.ECH, FR.CAMF, FR.CARF and FR.VIEF. Finally, we propose in section 4 a comparison between our approach and the method described in Pedersen et al. (2020).

2 Detection of non-Gaussian samples in an ensemble

Let us consider a set of samples whose distribution follows a Gaussian law, hereafter referred to as "background Gaussian signal" (BGS). This ensemble is often written $X \sim \mathcal{N}(\mu_0, \sigma_0)$, where μ_0 and σ_0 are the mean and the standard deviation, respectively. Such a distribution can be characterised by a bell-shaped histogram (*e. g.* DeGroot, 2002) or a kernel density estimate as well as the Cumulative Distribution Function (CDF) in order to avoid any arbitrary choice of discretisation (bin). For a real-valued random variable X, the CDF (ϕ) is defined as the probability that X takes a value less than or equal to a given real x. One can also use the quantile function (*i. e.* the inverse of the CDF), called the Probit function (Bliss, 1934) in the special case of the standard normal distribution: $\mu_0 = 0$ and $\sigma_0 = 1$ (see eq. (A4)). In practical, the Probit function (hereafter denoted as ϕ^{-1}), can be approximated

by sorting, according to increasing values, any set of n samples $(X_i)_{0\leq i\leq n-1}$ which follows the standard normal distribution (see theorem 2 in the appendix). The result of this sorting operation is hereafter called empirical Probit function, noted as ϕ_n^{-1}, which is represented as a function of quantiles.

In a general case, if the BGS follows a given Gaussian law (μ_0, σ_0), the Probit function (ϕ^{-1}) can no longer describe the distribution of the ensemble, we then introduce the modified probit function, denoted as $\hat{\phi}^{-1}$, by a translation/homothety of μ_0 and σ_0,

$$\hat{\phi}^{-1} = \mu_0 + \sigma_0 \phi^{-1}. \tag{1}$$

At this stage, μ and σ, the arithmetic average and the standard deviation, respectively (*e. g.* Feller et al., 1971) describe entirely both the BGS statistical properties and $\hat{\phi}^{-1}$ ($\mu = \mu_0$ and $\sigma = \sigma_0$).

If the sample set is now altered by a perturbation, which means presence of elements with large variations in amplitudes which significantly differ from the BGS, the classical estimators are biased ($\mu \neq \mu_0$ and $\sigma \neq \sigma_0$). The idea behind our method is to extract the subset of points composing the BGS from the complete ensemble. This can be done, once the signal is sorted according to increasing values, because deviant samples are located at the edges of ϕ_n^{-1}. Consequently, it exists a given quantile interval $[Q_A, Q_B]$, separating the samples composing the BGS from those of the perturbations which can be located through a full exploration of the sorted sample space. In practical, ϕ_n^{-1} is extracted for each tested quantile interval, its mean and standard deviation define the local $\hat{\phi}^{-1}$ (eq. 1) over the same amount of samples. According to theorem 2, the misfit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ is measured by the difference at the sense of the L^∞-norm. The interval finally selected, hereafter denoted as $[Q_A, Q_B]$, defines the subset of samples which achieve the lowest misfit. In the following, the mean and the standard deviation of samples within $[Q_A, Q_B]$ are denoted as μ_G and σ_G, respectively, as they define the statistical properties of the BGS.

The theory presented above is illustrated through three synthetic experiments (Fig. 1).

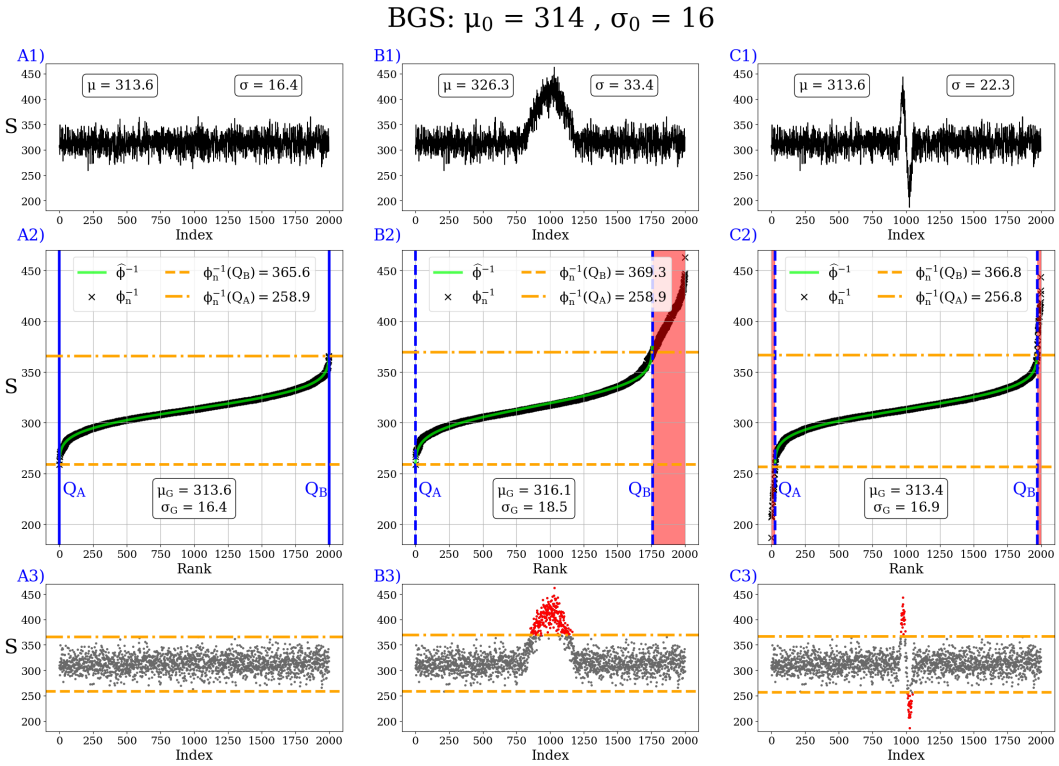

Figure 1. Illustration of how retrieving the Gaussian samples in three synthetic data set. The same BGS is imposed for each case (A, B and C) with $\mu_0 = 314$ and $\sigma_0 = 16$. A wide and a narrow perturbations are added in B and C, respectively. The second line (A2, B2 and C2) presents these signals (black crosses), once sorted by increasing order of amplitude, noted ϕ_n^{-1}. For each case, the interval $[Q_A, Q_B]$ is given by the best fit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ (green), defining μ_G and σ_G, approaching the properties of the BGS.

The BGS (A1) is obtained by a random draw of $n = 2,000$ points, with $\mu_0 = 314$ and $\sigma_0 = 16$, which are the parameters to retrieve for all cases. The classical arithmetic mean and standard deviation (μ and σ) of the three sample sets, are displayed in A1, B1, C1.

Let's start with the pure BGS case (A1, A2, A3). The samples shown in A1 are sorted by ascending order of amplitudes to generate ϕ_n^{-1} (black crosses in A2). The best fit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ (green curve in A2) is obtained for the interval $[Q_A, Q_B] = [0, 1999]$, indicating that all samples follow a Gaussian law with $\mu_G = 313.6$ and $\sigma_G = 16.4$. Obviously, since all the samples are considered as Gaussian here, $\mu_G = \mu$ and $\sigma_G = \sigma$, and are relatively close to μ_0 and σ_0.

In the second column of Fig. 1, a perturbation is added to the BGS. We can first notice that, obviously, μ and σ now differ from the values to be recovered (μ_0, σ_0). The exploration of

all possible quantile intervals gives $[Q_A, Q_B] = [0, 1759]$, which efficiently excludes the outlayer samples (red area in B2). This interval is associated with values of $\mu_G = 316.1$ and $\sigma_G = 18.5$ which are much closer to μ_0 and σ_0 compared to μ and σ. The values of $\phi_n^{-1}(Q_A)$ and $\phi_n^{-1}(Q_B)$ are of 258.9 and 369.3, respectively (horizontal dashed/dotted orange lines), which allow to separate anomalous samples (red points in B3) from the BGS.

For the narrow anomaly case (C1), μ is not affected due to the symmetric shape of the perturbation but the σ is biased since all the elements are taken into account. The exploration of the sorted data space returns here $[Q_A, Q_B] = [27, 1971]$, excluding outlayer samples composing the perturbation (red areas in C2). Back to the index domain (C3), the orange lines, given by $\phi_n^{-1}(Q_A)$ and $\phi_n^{-1}(Q_B)$, define the amplitude domain composing the BGS. Any sample above or below these two limits can be considered as perturbations. Once again, the value of $\sigma_G = 16.9$ is closer to the value of $\sigma_0 = 16$ compared to $\sigma = 22.3$. For all cases, the two horizontal orange lines are very similar, which is consistent with the fact that the same BGS is imposed in the three synthetic signals.

Finally, this approach allows to efficiently retrieve $[Q_A, Q_B]$ and thus the statistical characteristics of a BGS: μ_G and σ_G. As soon as an amplitude perturbation alters the data set, there is a mismatch between σ_G and σ. For the analysis of real signals, as μ_0 and σ_0 are unknown, any deviation from the gaussianity of a given data set can be measured by $\log(\frac{\sigma}{\sigma_G})$, in order not to depend on amplitudes and to reflect possible large variations from the reference state ($\sigma = \sigma_G$). For instance, in Fig. 1, the values of $\log(\frac{\sigma}{\sigma_G})$ is 0 exactly in (A) while it reaches values of 0.26 and 0.12 in (B) and (C), respectively, which correspond to significant deviations. A difference between μ and μ_G can also point out non-Gaussian features but can suffer from special cases such as a zero mean signals and/or symmetrical perturbations (Fig. 1 C). In the following, the word "perturbation" is used to describe any deviation from the Gaussian hypothesis (BGS), characterised by values of $\log(\frac{\sigma}{\sigma_G})$ greater than 0.

3 Application to the seismic station monitoring

In this section, we propose to analyse the continuous seismic signal recorded at four permanent broadband stations, using the method presented in section 2. In the following, it is assumed that the continuous seismic signal follows a Gaussian distribution (*e.g.* Groos and Ritter, 2009; Zhong et al., 2015b,a; Aggarwal et al., 2020).

3.1 Methodology

The gaussianity of the continuous seismic signal recorded during 24 h can be quantified by multiple analysis of short time windows. Results are shown in Fig. 2, using 1 h time windows, sliding with an overlap of $\frac{2}{3}$. Hence, each sample is analysed three times. In order to investigate the frequency dependence of the gaussianity, the signal is analysed through four period ranges: LF ($T > 80$ s), BP1 (20 s $< T < 80$ s), BP2 (1 s $< T < 20$ s) and HF ($T < 1$ s). In order to allow a reliable comparison between the different period bands, all signals are decimated at 20 samples per second in order to have the same amount of samples in each analysed window. The instrument response is removed in the period range $[0.1,\ 160]$ s and the signal is converted into ground velocity.

For each time window σ is computed using all samples whereas σ_G is defined after the computation of $[Q_A, Q_B]$. Although we mostly focus on $\log(\frac{\sigma}{\sigma_G})$ to quantify the gaussianity and to detect anomalous behaviour of seismic stations, three other parameters can also be investigated:

- μ_G, the Gaussian mean of the ranked samples within $[Q_A, Q_B]$. Since the arithmetic average is subtracted from the signal amplitude before each filtering operation of a given 1 h time window, μ_G must be compared to zero;
- $\mathcal{G}$, the Gaussian point ratio, defined by the amount of selected samples in $[Q_A, Q_B]$ divided by the total amount of points of the sliding short time window;

- M_{L^2}, the misfit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ (Fig. 1, second row), using the L^2-norm,

$$M_{L^2} = \frac{1}{(Q_B - Q_A)} \sqrt{\sum_{i=Q_A}^{Q_B} \left(\hat{\phi}^{-1}(i) - \phi_n^{-1}(i) \right)^2}. \quad (2)$$

A low value of M_{L^2} then reflects a high degree of gaussianity of the subset of samples selected in $[Q_A, Q_B]$.

3.2 Single day analysis of the gaussianity

Fig. 2 exhibits the four parameters defined above for a signal duration of 24 h (June 1, 2019), recorded at FR.CAMF (North component) and filtered in two frequency ranges: BP1 and HF. The sensor (Nanometrics T120QA) of this broadband permanent station is installed on the ground in a WWII blockhaus, in Brittany (France), and located at the top of a cliff facing the Atlantic Ocean (Fig. 3). The rock basement is composed of Armorican sandstone. Although the quality of the installation is standard and made with great care, the continuous seismic signal is altered for different reasons: at high frequency, the proximity of the village and the energy of breaking waves on the cliff and at longer periods, temperature and pressure variations in addition to tidal modulations (*e. g.* Beucler et al., 2015).

The signal filtered in the BP1 frequency domain (red in Fig. 2) is less energetic than the HF filtered trace (green) but contains some similarities. A diffuse extra energy is visible on two ∼3 h windows, centered around 2:20 and 14:45 UTC, respectively (green areas in Fig. 2 B). They both coincide with the high tides occurring twice a day. The seismic signal is then modulated in the HF range due to the breaking waves on the cliff but also to a lesser degree in BP1 since this frequency domain comprises the edge of the primary microseismic peak and a part of the infragravity wave period range (*e. g.* Nawa et al., 1998; Ardhuin et al., 2011; Stutzmann et al., 2012). In addition, the surface waves of two $M_W \simeq 5$ earthquakes that occurred in Greece (epicentral distances of approximately 2,200 km) are well visible in BP1 trace (indicated by the two vertical arrows in A) but are less obvious for HF.

For both BP1 and HF domains the values of the BGS mean (μ_G) lie between −0.89 and

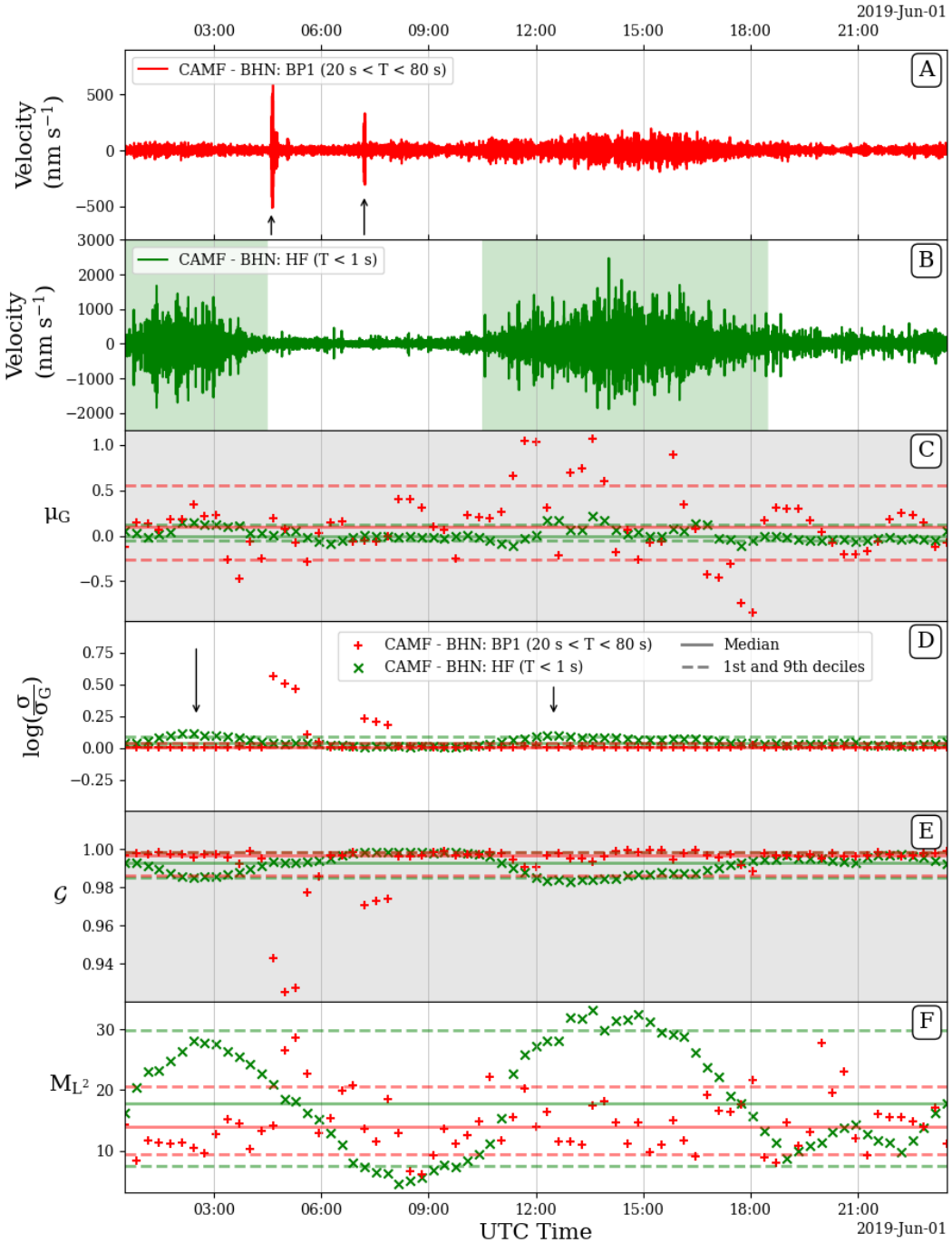

Figure 2. Analysis of a continuous seismic signal during a full day, using a sliding window approach. (A and B): Seismic signals from the FR.CAMF station on June 1, 2019 (BHN), deconvolued and filtered from 20 to 80 s (A) and below 1 s (B). (C): Mean of the BGS. (D): Logarithm of the ratio between the classical and the BGS standard deviation. (D): Proportion of Gaussian points in the $[Q_A, Q_B]$ interval. (E): Misfit between $\hat{\phi}^{-1}$ and ϕ_n^{-1} in $[Q_A, Q_B]$, using the L^2-norm.

1.18 (Fig. 2 C) and $\log(\frac{\sigma}{\sigma_G})$ is very stable around the value of 0. For the HF case (green crosses) two $\log(\frac{\sigma}{\sigma_G})$ deviations up to 0.15 are observed at the times of high tides (pointed out by the two black arrows in D) indicating that, locally, the samples that composed a 1 h window are less Gaussian than the rest of the day. The consequence is a decrease of $\mathcal{G}$ ($\sim 98.2\%$ for both high tide windows) and large increases of M_{L^2} (up to 32.5) which leads to conclude that even in the $[Q_A, Q_B]$ interval the fit to $\hat{\phi}^{-1}$ is not as good as for quieter parts of the day.

The BP1 frequency range analysis for the same day (red pluses in Fig. 2) shows a very stable behaviour all over the 24 h except during the two earthquakes. Those impulsive transient energies do not affect μ_G, which is consistent with surface wave wavetrains that make the ground oscillating symmetrically around an equilibrium position, but they are well visible on $\log(\frac{\sigma}{\sigma_G})$ with values up to 0.6. For the corresponding time windows, $\mathcal{G}$ decreases down to 0.925.

Finally, it is important to notice that these parameters are sensitive only to amplitude variations and not to the level of the seismic energy. This allows to propose that such a study can be performed for any component of any seismic station and for different ranges of periods. In the following, since $\log(\frac{\sigma}{\sigma_G})$ reflects both mean translation and sample dispersion around this latter, we will mainly use this parameter to quantify the Gaussianity of a single day. This is realised using the median value of the 74 one hour windows (solid lines in Fig. 2) that composed a day (with an overlap of $\frac{2}{3}$). As shown in Fig. 2 D, the median is not affected by transient waveforms such as earthquakes and/or spurious signals.

3.3 Daily analyses of the seismic signal gaussianity at four permanent stations

In order to analyse the behaviour of a permanent station in terms of deviation from gaussianity day by day, we focus hereafter on four broadband seismic stations (Fig. 3). Let us start with G.ECH, located in Echery (eastern France), that we consider as the reference

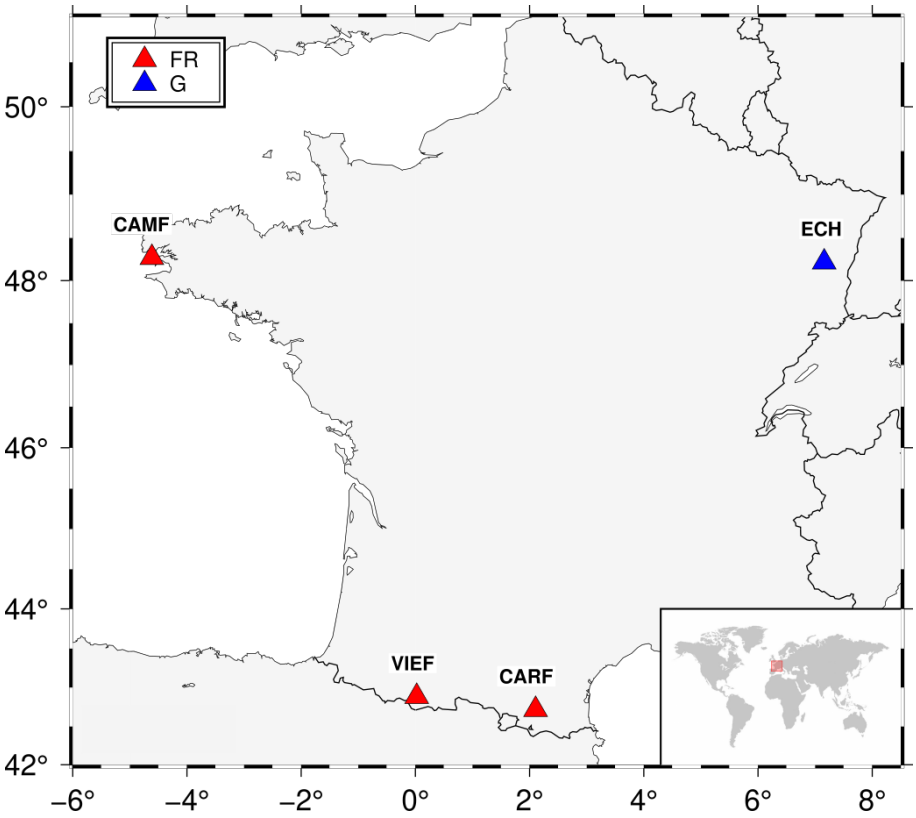

Figure 3. Locations of seismic stations used in this study. They are all belonging to the French permanent broad-band array, from the RESIF (1995)(FR) and the GEOSCOPE (G) (Institut de physique du globe de Paris (IPGP) and École et Observatoire des Sciences de la Terre de Strasbourg (EOST), 1982) networks.

station in terms of signal quality.

3.3.1 ECH

The sensor (STS1) is installed on a concrete pavement in a 250 m long tunnel inside an abandoned silver mine. The site geology is mostly composed of gneiss. This station is running for more than 22 years and is known for the stability of its quality over the years. In a few words, this station is of high quality at short periods (PSD lower than 150 dB for $T < 1$ s) and exhibits a vertical component energy close to the low noise model (Peterson, 1993) between 20 and 200 s period. The horizontal components are noisier for periods greater than 40 s and the North component is more affected than the East one.

The analysis of G.ECH in terms of $\log(\frac{\sigma}{\sigma_G})$ variations, in four frequency ranges (see section 3.1), and for the whole year 2019 is presented in Fig. 4. For each day, the medians

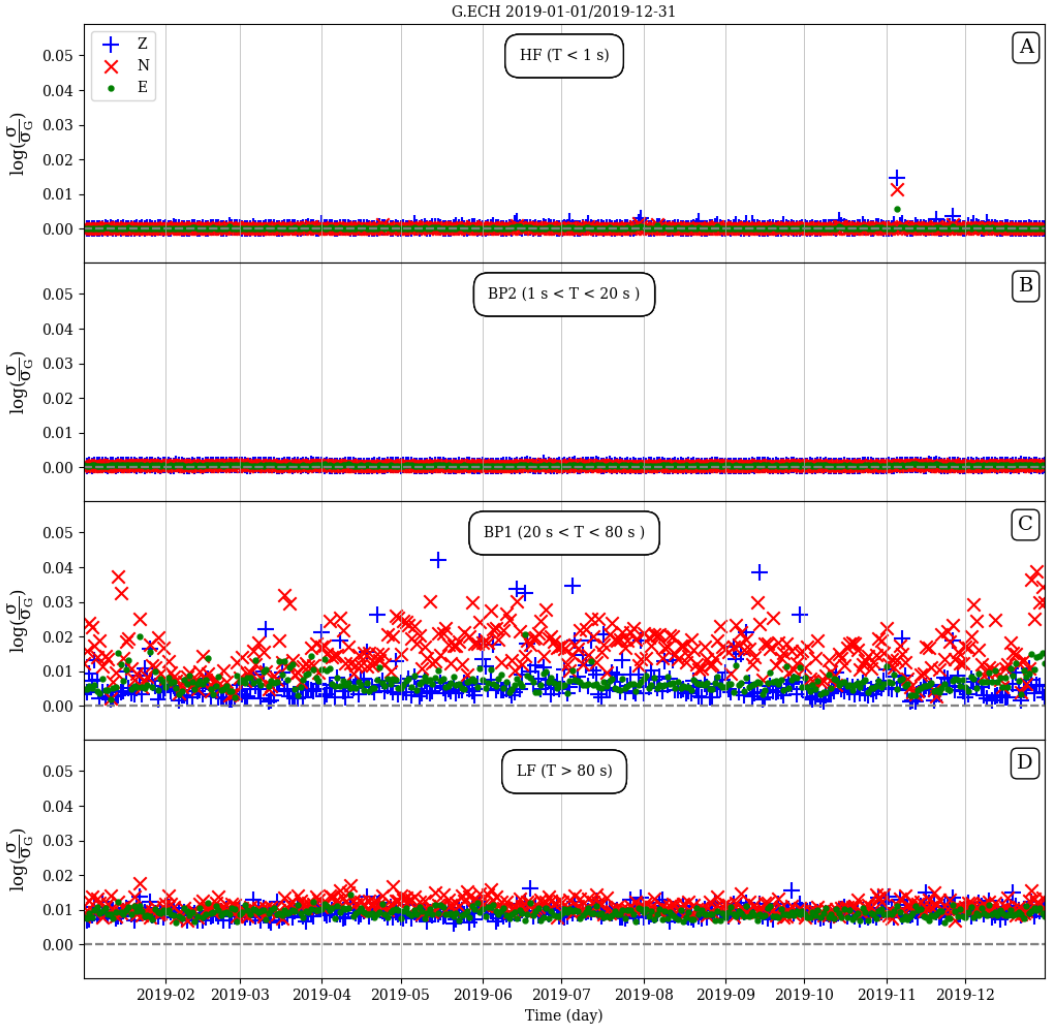

Figure 4. Analysis of the continuous seismic signal recorded at G.ECH in 2019. The medians of the daily $\log(\frac{\sigma}{\sigma_G})$ are displayed for the three components (plusses, crosses and dots for the vertical, north and east, respectively) and the four frequency bands (LF, BP1, BP2 and HF), defined in section 3.1.

of $\log(\frac{\sigma}{\sigma_G})$, computed for the 74 time windows, are displayed for the three components. Compared to other station analyses (Figs. 5, 6 and 7), the values are so close to 0 that we propose a vertical scale between 0.01 and 0.06. For the HF and BP2 frequency ranges, the values are very stable around 0 for the whole year and reach maxima of 0.014 and 0.001, respectively. This implies that, the continuous seismic signal at periods lower than 20 s are in very good agreement with a Gaussian distribution. This is particularly true for BP2 which comprised the frequency band of the microseismic peaks (*e. g.* Ebeling, 2012).

In contrast, for BP1 and LF, we observe a greater dispersion, for instance, it is 10 times

larger for BP1 than for HF. For the BP1 frequency range (Fig. 4 C), the mean of all North $\log(\frac{\sigma}{\sigma_G})$ (red crosses) is 0.015 whereas they are of 0.06 for the two other components. This could indicate that the extra energy which makes this component noisier (as indicated by power spectral densities that can be computed for this station) with respect to others, also alters the gaussianity of the signal. This phenomenon can be observed to a lesser degree in the LF domain (Fig. 4 D), for which the three components exhibit however a more stable behaviour over the year. One can notice that the mean of the $\log(\frac{\sigma}{\sigma_G})$ oscillates here around of 0.01, ant not exactly 0, which is only a side effect due to the length of 1 h for all analysed windows, allowing less oscillations of the signal than for the highest frequencies. To avoid any misinterpretation, only values greater than 0.1 are considered as noticeable deviations from the Gaussian case (BGS). One can notice a $\log(\frac{\sigma}{\sigma_G})$ variation during November, 5 on the HF frequency bands (Fig. 4 A), which is caused by a surprisingly large occurrence of earthquakes and quarry blasts (more than 50 events).

Finally, since we are interested mostly in the time variations of $\log(\frac{\sigma}{\sigma_G})$, we consider hereafter that the maximum values at ECH (for each frequency range) can be used as reference thresholds for other stations (orange lines in Fig. 5).

3.3.2 CAMF

The site conditions of FR.CAMF are already detailed in section 3.2. As for G.ECH, the analysis of $\log(\frac{\sigma}{\sigma_G})$ variations of the continuous seismic signal recorded at FR.CAMF in 2019 are displayed in the four frequency ranges in Fig. 5. For each frequency band, horizontal orange line is shown to indicate the maximum value of all medians (of all components) measured at G.ECH. They can be considered as threshold references to point out any alteration of the signal.

We choose the year 2019 because the recording conditions of FR.CAMF have been modified between the beginning of July and mid-October (period highlighted by the grey and magenta vertical dashed lines, respectively in Fig. 5). Due to high humidity at this time, the sand that insulates the sensor gradually became waterlogged. This led to a deterioration of the

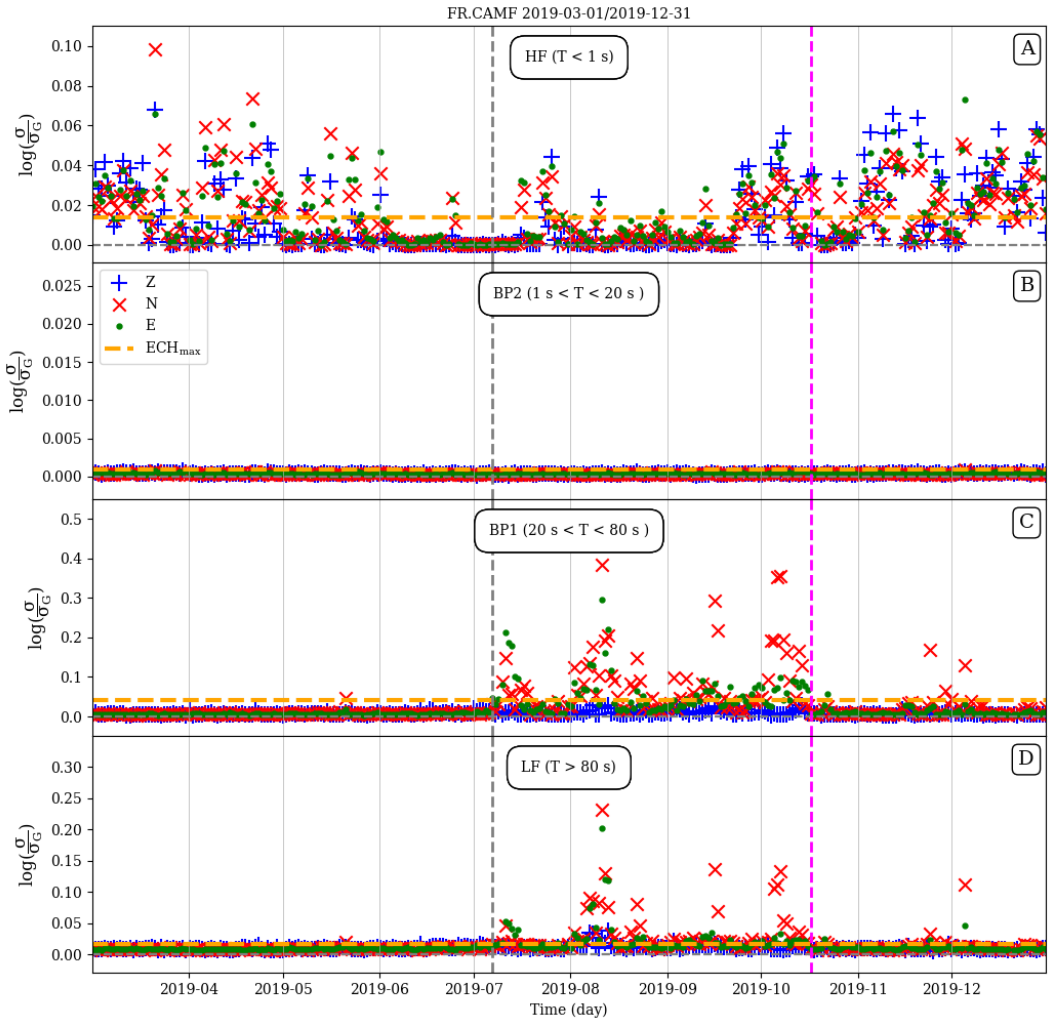

Figure 5. Same legend as Fig. 4, but for FR.CAMF. Horizontal orange lines are indicating the maximum $\log(\frac{\sigma}{\sigma_G})$ value (for all components), computed for G.ECH in each frequency band (see section 3.3.1). Due to the high level of humidity, the recording conditions are degraded during the time window defined by the two vertical dashed lines.

long period signal quality of horizontal components that can be seen on the spectrograms in Fig. 9.

Considering the signal before July 7, 2019, all components in the HF frequency bands are much more dispersed than for BP2, BP1 and LF. The variations of $\log(\frac{\sigma}{\sigma_G})$ are of the same order of magnitude as the single day analysis presented in Fig. 2. They are due to the seismic extra energy, caused by breaking waves on the cliff. Since the degradation of the recording conditions does not affect HF (Fig. 9), it is not possible to detect any noticeable modification in this frequency range. For the same reason and because the microseismic

peak energy is obviously very large at FR.CAMF, $\log(\frac{\sigma}{\sigma_G})$ is always close to 0 in BP2 (with mean equals to 0.0004). This contrasts with the values observed in BP1 and LF bands (Fig. 5 C, and D), where the daily $\log(\frac{\sigma}{\sigma_G})$ reach 0.38 and 0.23, respectively. These large deviations are only visible for the horizontal components which is consistent with Fig. 9. However, while a classical energy analysis, such as PPSD or spectrograms, do not yet show any significant changes, the $\log(\frac{\sigma}{\sigma_G})$ turns to anomalous values (up to 0.21 for the East component) as early as the July, 7 (grey dashed line). This deteriorated recorded conditions ended on October, 17 (magenta dashed line) when the wet sand has been replaced by dry one. This intervention brought back the sensor into the normal operating conditions, resulting in $\log(\frac{\sigma}{\sigma_G})$ values that rapidly return to 0.

One can notice few anomalous values of $\log(\frac{\sigma}{\sigma_G})$ for both BP1 and LF (Fig. 5 C, and D) between November, 21 to December, 5. After visual inspection, it appears that three days have been perturbed by long period glitches that mostly affect the north component.

3.3.3 CARF and VIEF

FR.CARF and FR.VIEF are both located in the Pyrénées mountains (France) at altitudes of 1,200 and 1,000 m, respectively. The geology of FR.CARF is composed of limestones while FR.VIEF is installed in a shale massif. Their sensors (T120QA for FR.CARF and T120PA for FR.VIEF) are installed in a $\sim$ 1 m depth vault and insulated with sand. FR.VIEF is located about 30 m of a village, making it theoretically more exposed to anthropic activity than FR.CARF, although this is not that obvious in the HF frequency band of Figs. 6 and 7 neither in the spectrograms shown in Fig. 9. The choice of these two stations is motivated because both of their recorded signals have been suddenly deteriorated by humidity that corroded connections. The insulation was realised using sandbags arranged around the sensors and the water that seeped in was guided to the connectors. This appears between September 13–30, 2021 for FR.CARF and between February 9–17, 2022 for FR.VIEF, as indicated by the grey and magenta vertical dashed lines in Figs. 6 and 7.

For both stations and before degradation, the signals of the three components have a high

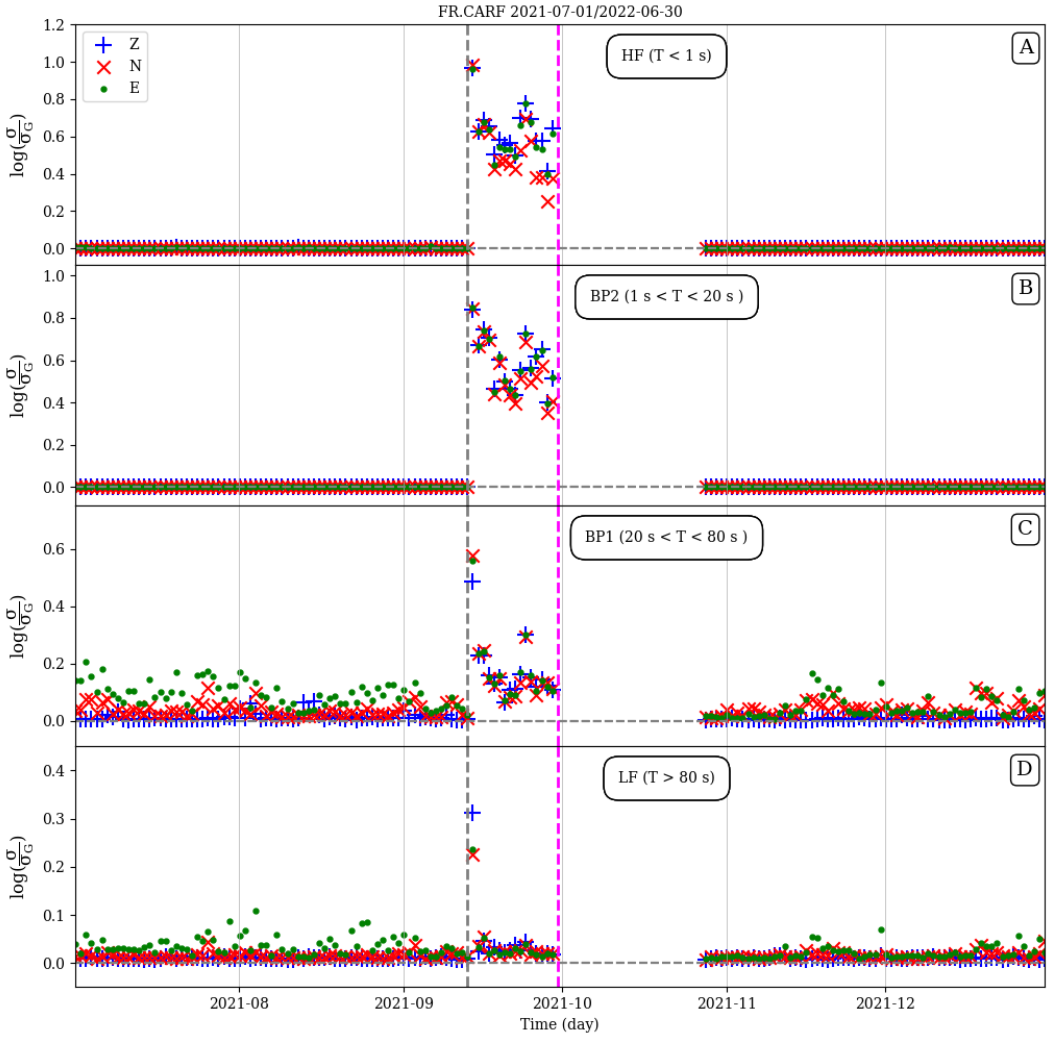

Figure 6. Same legend as Fig. 4, but for FR.CARF. Due to the high level of humidity, the recording conditions are degraded during the time window defined by the two vertical dashed lines.

degree of gaussianity characterised by values of $\log(\frac{\sigma}{\sigma_G})$ very close to 0. This is particularly true at high frequency and in the microseismic bandwidth (BP2) while few variations are observed for the signal at long periods (BP1 and LF), mostly on the East component for FR.CARF and on the North component for FR.VIEF (although it is not obvious in Fig. 7 C due to the vertical scale). These descriptions can be linked to the fact that FR.CARF and FR.VIEF are located on the eastern and southern flanks of mountains, respectively.

The two stations have encountered a degradation of their operating conditions when large modifications of $\log(\frac{\sigma}{\sigma_G})$ are observed. In both cases, the $\log(\frac{\sigma}{\sigma_G})$ signatures differ

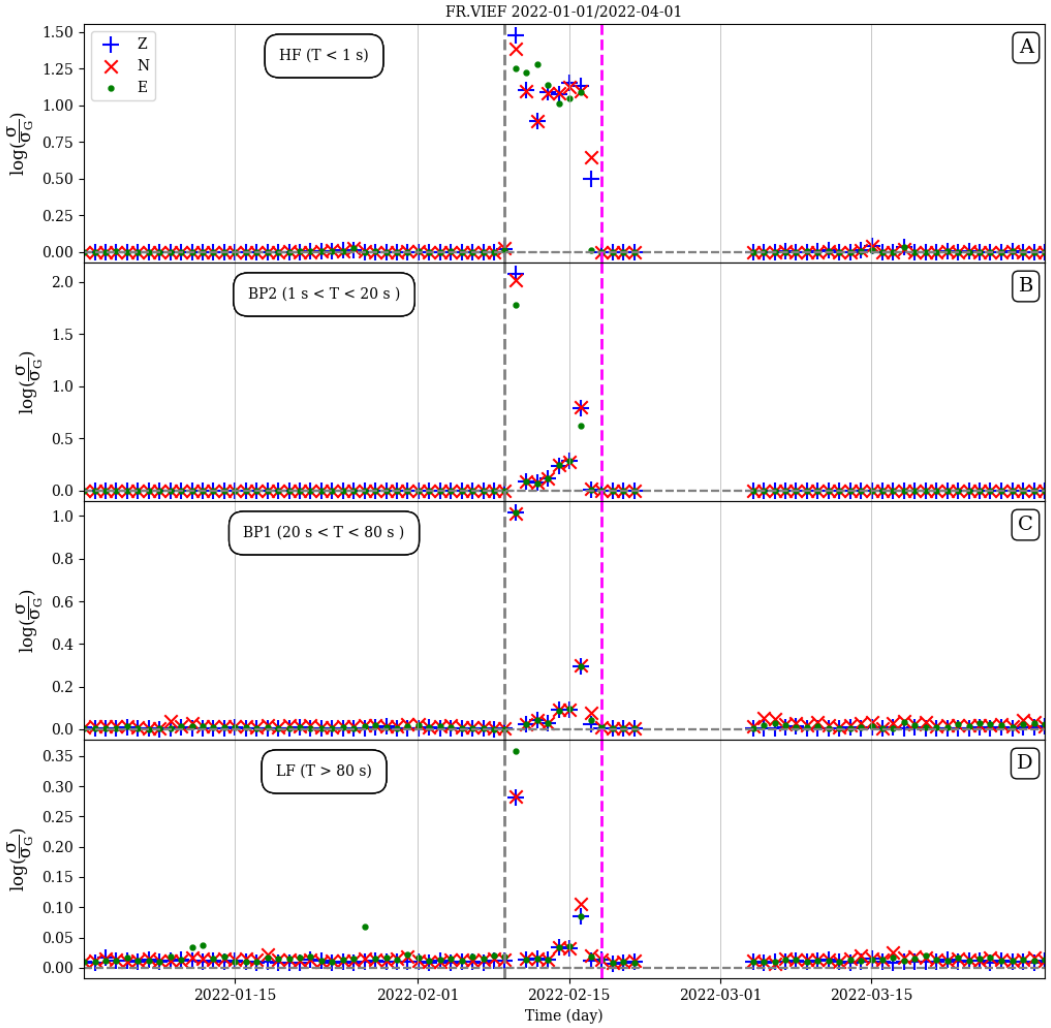

Figure 7. Same legend as Fig. 6, but for FR.VIEF.

as a function of the frequency. For instance, the LF domain although largely affected in terms of the signal energy (see FR.VIEF spectrograms in Fig. 9) is not obvious in Fig. 6(D) and 7(D).

In the HF and BP2 frequency domains at FR.CARF, the daily $\log(\frac{\sigma}{\sigma_G})$ values are remarkably stable and never exceeds 0.01, before and after the degradation time (Fig. 6 A and B). *A contrario*, as soon as the recording conditions are degraded, they become very large (up to 0.98 and never lower than 0.25). At longer periods, a modification of $\log(\frac{\sigma}{\sigma_G})$ is also observed but to a lesser degree, except for the 1st day (September 14, 2021), where it reaches values of 0.57 and 0.31 for BP1 and LF, respectively. The station operators removed the corroded sensor on September 30 and installed a new one on October, 27 (explaining the

data gap). The gaussianity in the different frequency domains returns to the same level as before the degradation.

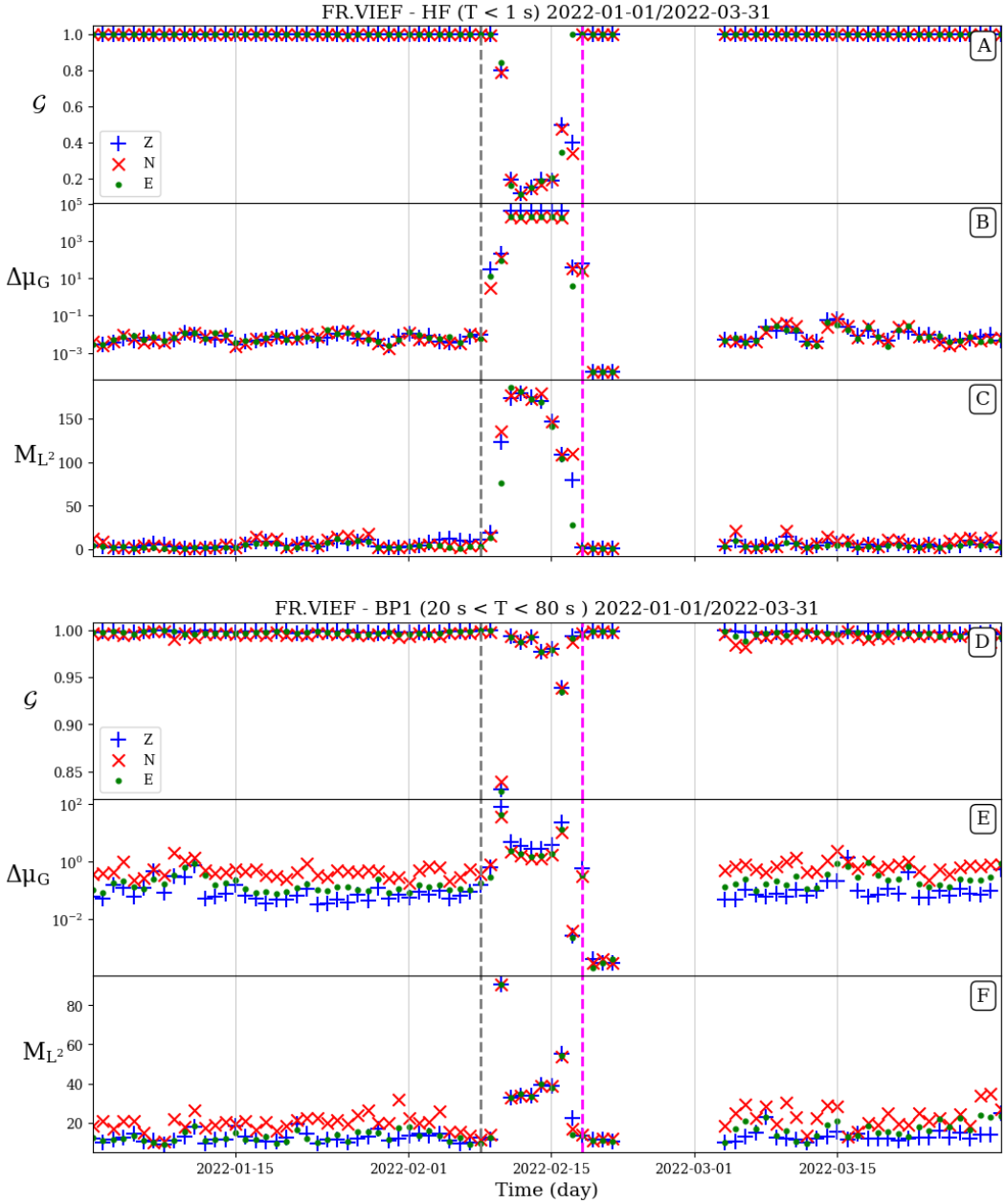

Figure 8. Gaussianity analysis of FR.VIEF during the same time period than in Fig. 7. For both frequency bands, $\mathcal{G}$ and M_{L^2} are the Gaussian point ratio and the misfit at the least-squares sense, respectively. $\Delta\mu_G$ is the difference between the 9th and the 1st deciles of all daily μ_G values.

A more detailed study is realised for FR.VIEF (Figs. 7, 8 and 9). The same HF and BP2 $\log(\frac{\sigma}{\sigma_G})$ signatures as for FR.CARF are observed during the degradation time, but in

this case with values larger than 1.4 for HF and 2 for BP2 (Fig. 7 A and B).

Supplementary information are given in Fig. 8, where three other parameters are shown for HF and BP1. $\mathcal{G}$ and M_{L^2} are detailed in section 3.1 and $\Delta\mu_G$ represent here the difference between the 9th and the 1st deciles of the set of all one hour μ_G values computed every day (as indicated for instance by the horizontal dashed lines in Fig. 2 C). This parameter quantifies the stability of μ_G for a given day and, for the sake of comparison, low values are bounded to 10^{-4}.

As for $\log(\frac{\sigma}{\sigma_G})$, in the HF domain, $\mathcal{G}$, $\Delta\mu_G$ and M_{L^2} exhibit large variations during the degradation time. One can notice that $\mathcal{G}$ reaches values of 0.15, indicating that only 15% of samples are selected to belong to $[Q_A, Q_B]$, which is consistent with the large values of $\log(\frac{\sigma}{\sigma_G})$ shown in Fig. 7 A. Such $\mathcal{G}$ values are very close to the minimal proportion of Gaussian samples that is authorized in our method ($\mathcal{G} = 0.1$). In addition, M_{L^2} values are the largest, telling that even the 15% of selected samples are much less Gaussian than outside the degradation time. Plus, very large values of $\Delta\mu_G$ ($\sim$ 43,000) are observed, confirming that huge μ_G variations are occurring within a day. Finally, all these parameters are converging toward the same diagnostic of an ill-sensor with very large energy fluctuations and dramatically different signal quality compared to before, as also shown by the spectrograms (Fig. 9).

At longer periods (BP1 and LF in Fig. 7), the $\log(\frac{\sigma}{\sigma_G})$ values are less affected by the signal degradation. This can be due to a long period feedback deterioration which could decrease the sensor sensitivity as shown by a slightly different behaviour of all other parameters (Fig. 8 D, E and F).

One can notice a sudden return of $\log(\frac{\sigma}{\sigma_G})$ to 0, just after the end of the degradation (magenta line), for all the components and frequency bands. It is simply due to the numerical noise of the digitizer, which continued to operate even once the sensor have been removed. The channels have been officially closed three days after sensor removal producing the data gap.

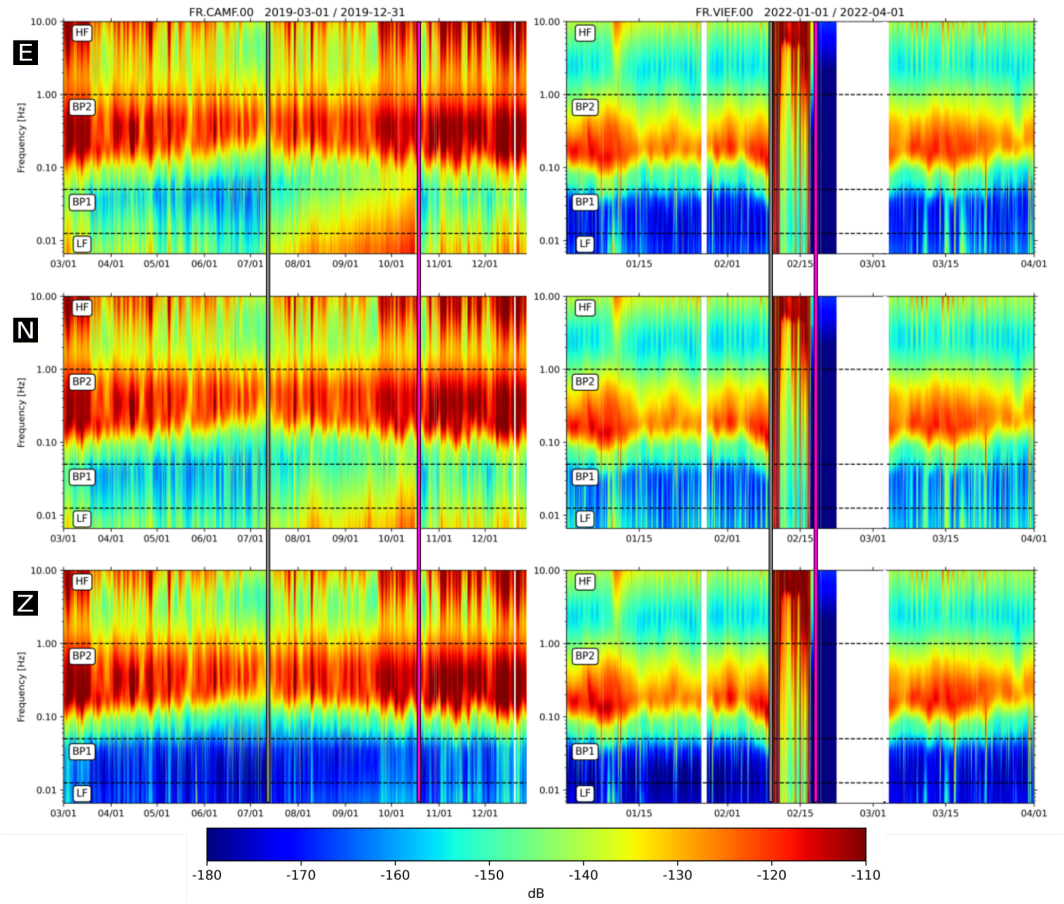

Figure 9. Spectrograms for FR.CAMF (left) and FR.VIEF (right). They are computed using 3600 s length windows with no overlap. The grey and magenta vertical lines correspond to the edges of the signal degradation time windows and plotted as dashed lines in Figs. 5, 7 and 8. For each spectrogram, the horizontal black dashed lines bound the four frequency domains.

4 Discussion and conclusion

The method presented in this article aims to point out anomalous features in the continuous seismic signals using different gaussianity estimators. As shown in the previous figures, we focus mainly on one of them, which is the ratio of the classical standard deviation σ and the BGS standard deviation σ_G. It can be compared to a method which aims to monitor the seismic signal quality using the ratio of the classical standard deviations for two components (Pedersen et al., 2020).

4.1 Comparison to a component ratio approach

In their approach, Pedersen et al. (2020) compute, for each component and 8 frequency bands, the classical standard deviation in 5-minute time windows of the continuous seismic signal, recorded at various Geoscope stations, with no overlap. For all the short time windows of a given day, the energy ratio is quantified by the ratio of the standard deviations for each pair of the three components (E/N, E/Z and N/Z). The estimate of the component daily energy is then defined using the median of all ratios.

In order to illustrate the difference between this method and the one presented in this paper, we present in Fig. 10 a focus on the HF domain around the degradation time for FR.VIEF as already studied in Figs. 7, 8 and 9. In addition to the $\log(\frac{\sigma}{\sigma_G})$ median values shown in Fig. 7, we display in Fig. 10 the decile interval comprised between the 1st and the 9th deciles of all daily $\log(\frac{\sigma}{\sigma_G})$ values (using the same colors as their median: blue for Z, red for N and green for E). For a given day and a given component, when all $\log(\frac{\sigma}{\sigma_G})$ values are very close, the dispersion is so small that it cannot be seen on Fig. 10 A. It is nevertheless possible to obtain a median value close to 0 with a large decile interval such as for January 26, 2022. On this day, all components are particularly affected by 60 local earthquakes ($0.4 \leq M_l \leq 2.8$) occurring around FR.VIEF and within an epicentral distance range of 100 km.

As soon as the sensor is corroded enough to affect the recording conditions at 2022-02-09T17:29 UTC (grey vertical dashed line), the decile interval suddenly increases up to 2.2 while the median is not yet modified. This is due to the fact that more than 50% of this day recorded a clean and Gaussian signal, as shown in the daily seismogram (Z component, HF filter) inserted in Fig. 10 A. The following days are characterized by both large values of the median of $\log(\frac{\sigma}{\sigma_G})$ and large width of the decile intervals. Finally, when the sensor has been disconnected at 2022-02-18T09:56 UTC (magenta dashed line) while the digitizer continued to operate, the recorded signal (pure numeric noise) has very low values in terms of median and deciles.

The same methodology as in Pedersen et al. (2020) is followed for this station. The three

energy ratios for each pair of components are displayed in Fig. 10 B. Before the beginning of the signal degradation they are all characterised by a quite large discrepancy. Values are ranging between 0.8 and 1.16 although the daily signal is very clean. Indeed, visual inspection of the whole signal during these 25 days did not allow to spot any precursor of the alteration of the sensor connection which is *a contrario* well reflected by the very low $\log(\frac{\sigma}{\sigma_G})$ values that do not exceed 0.03 (A).

After the vertical grey dashed line, the variations of the daily energy ratios suddenly decrease to converge towards values of 0.47 for N/Z and E/Z and 1.02 for E/N which attest of the seismic signal modification. These values testify that, once the recording conditions have been degraded, the vertical component is about twice more energetic than the two others which are similar. The comparison of one component with respect to another (B) can thus bring fruitful information on the actions to be taken (even if it is not the case here) although the estimator of the signal quality before the degradation is more stable in (A) than in (B).

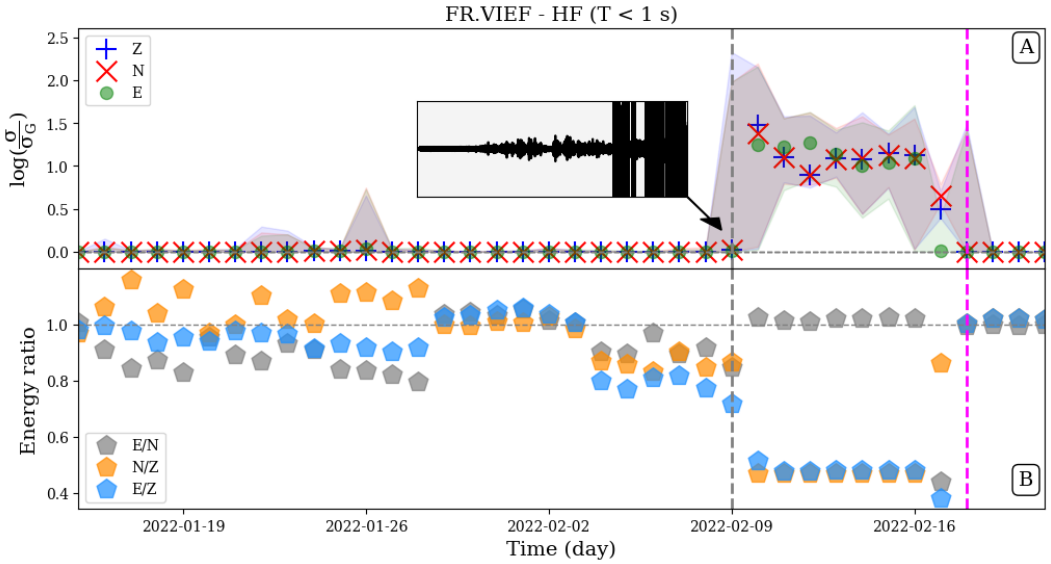

Figure 10. Comparison between our approach and a method based on energy ratios for each pair of components (Pedersen et al., 2020). The period of interest focuses on the FR.VIEF sensor degradation as previously shown in Figs. 7, 8 and 9. (A) For each component, the values of the 1st and the 9th deciles of all daily $\log(\frac{\sigma}{\sigma_G})$ are displayed with the same color as the associated median. (B) For each pair of components, the energy ratio (represented by colored pentagons) are given by the median of all daily ratio of standard deviation of the two considered components.

4.2 Concluding remarks

The method presented in this article introduces a new approach to point out all samples of a given data set that do not agree the dominant gaussianity, referred to as BGS. For a given time window, means a set of n samples (and we estimate that n must be greater than 1,000, as shown in Fig. A1), our approach relies on four parameters to characterize the gaussianity: M_{L^2}, $\mathcal{G}$, μ_G and σ_G. Using the classical definition of the standard deviation, $\log(\frac{\sigma}{\sigma_G})$ therefore measures the non gaussianity of a given data set. Although the M_{L^2}, $\mathcal{G}$ and μ_G bring useful information, $\log(\frac{\sigma}{\sigma_G})$ alone can efficiently estimate whether the considered data set follows a normal distribution. At the scale of a single day, since many time windows can be processed following a sliding strategy, the median of all $\log(\frac{\sigma}{\sigma_G})$ gives a good quantification of the daily overall gaussianity without giving too much weight to transient waveforms such as earthquakes. Thus, it could be used as a new estimator to reliably monitor the continuous seismic signal assuming that any modification in the recording conditions affects the gaussianity of the signal. As shown in this article, $\log(\frac{\sigma}{\sigma_G})$ is sensitive to both subtle changes on one or two components (Fig. 5) but also major degradations of sensors altering all of them (Figs. 6 and 7). It appears that to seize any kind of temporal modification, it is necessary to process various frequency ranges.

Although spectrogram analyses bring fruitful information they face two difficulties for monitoring purposes: i) for a given frequency range, the seismic energy vary a lot as function of days/months/years and ii) the detection of anomalous behaviour of the station needs long time series. *A contrario*, $\log(\frac{\sigma}{\sigma_G})$ includes in few values any statistical deviation from normal seismograph operation and does not depend on the signal energy. We consider therefore $\log(\frac{\sigma}{\sigma_G})$ as a simple and meaningful parameter to monitor seismic station quality. We propose that, for a given frequency range, any daily $\log(\frac{\sigma}{\sigma_G})$ value greater than 0.1 requires a visual inspection of the signal since it corresponds to a σ value greater than 30% of σ_G. Finally, we think that this approach can bring useful information for seismic station monitoring purposes and then can be in line with methods that already exist. It can be used for

permanent stations transmitting data in real time, as well as for identifying problems that occurred in the past.

Data and Resources

The Python code underlying this article will be shared on reasonable request to arthur.cuvier@etu.univ-nantes.fr. In this study we used data from networks with FDSN code FR (RESIF, 1995a) and G (Institut De Physique Du Globe De Paris (IPGP) and Ecole Et Observatoire Des Sciences De La Terre De Strasbourg (EOST), 1982). The seismic data set used in this study can be accessed at https://service.iris.edu/.

Acknowledgments

This project is funded by ANR-MAGIS-19-CE31-0008-02. Résif-Epos is a Research Infrastructure (RI) managed by the CNRS-INSU. Authors warmly thank Hélène Pauchet (IRAP-OMP) and Damien Fligiel (OSUNA) for there explanations about sensor failures. The work presented in this study was done with a Python program vanRossum (1995) using in particular the NumPy Harris et al. (2020), Scipy Virtanen et al. (2020) and Obspy Beyreuther et al. (2010) libraries for the signal processing. The figures presented in this study were generated with the Matplotlib library Hunter (2007). The authors acknowledge there are no conflicts of interest recorded.

References

Aggarwal, K., Mukhopadhyay, S., and Tangirala, A. K. (2020). Statistical characterization and time-series modeling of seismic noise. *arXiv preprint arXiv:2009.01549*.

Alu, K. I. (2011). *Solving the Differential Equation for the Probit Function Using a Variant of the Carleman Embedding Technique*. PhD thesis, East Tennessee State University.

Ardhuin, F., Stutzmann, E., Schimmel, M., and Mangeney, A. (2011). Ocean wave sources of seismic noise. *J. Geophys. Res.: Oceans*, 116(C9).

Beucler, É., Mocquet, A., Schimmel, M., Chevrot, S., Quillard, O., Vergne, J., and Sylvander, M. (2015). Observation of deep water microseisms in the north atlantic ocean using tide modulations. *Geophys. Res. Lett.*, 42(2):316–322.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. (2010). Obspy: A python toolbox for seismology. *Seismological Research Letters*, 81(3):530–533.

Blair, J., Edwards, C., and Johnson, J. H. (1976). Rational chebyshev approximations for the inverse of the error function. *Mathematics of Computation*, 30(136):827–830.

Bliss, C. I. (1934). The method of probits. *Science*, 79(2037):38–39.

Casey, R., Templeton, M. E., Sharer, G., Keyson, L., Weertman, B. R., and Ahern, T. (2018). Assuring the quality of iris data with mustang. *Seismological Research Letters*, 89(2A):630–639.

Davis, P. and Berger, J. (2007). Calibration of the global seismographic network using tides. *Seismological Research Letters*, 78(4):454–459.

DeGroot, M. H. (2002). Probability and statistics.

Ebeling, C. W. (2012). Chapter one - inferring ocean storm characteristics from ambient seismic noise: A historical perspective. In Dmowska, R., editor, *Advances in Geophysics*, volume 53 of *Advances in Geophysics*, pages 1 – 33. Elsevier.

Ekstrom, G., Dalton, C. A., and Nettles, M. (2006). Observations of time-dependent errors in long-period instrument gain at global seismic stations. *Seismological Research Letters*, 77(1):12–22.

Feller, W. et al. (1971). An introduction to probability theory and its applications.

Finney, D. J. (1971). Probit analysis, cambridge university press. *Cambridge, UK*.

Francinou, S., Gianella, H., and Nicolas, S. (2013). *Exercices de Mathématiques (oraux X-ENS): analyse 2*. Cassini.

Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4:92–99.

Groos, J. C. and Ritter, J. R. R. (2009). Time domain classification and quantification of seismic noise in an urban environment. *Geophys. J. Int.*, 179(2):1213–1231.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hoffman, D. L. and Low, S. A. (1981). An application of the probit transformation to tourism survey data. *Journal of Travel Research*, 20(2):35–38.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95.

Hutt, C. and Ringler, A. (2011). Some possible causes of and corrections for sts-1 response changes in the global seismographic network. *Seismological Research Letters*, 82(4):560–571.

Institut de physique du globe de Paris (IPGP) and École et Observatoire des Sciences de

la Terre de Strasbourg (EOST) (1982). Geoscope, french global network of broad band seismic stations.

Kimura, T., Murakami, H., and Matsumoto, T. (2015). Systematic monitoring of instrumentation health in high-density broadband seismic networks. *Earth, Planets and Space*, 67(1):1–15.

Kockelman, K. M. and Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis & Prevention*, 34(3):313–321.

McNamara, D. E. and Boaz, R. I. (2010). Pqlx: A seismic data quality control system description, applications, and users manual. *US Geol. Surv. Open-File Rept*, 1292:41.

Nawa, K., Suda, N., Fukao, Y., Sato, T., Aoyama, Y., and Shibuya, K. (1998). Incessant excitation of the earth's free oscillations. *Earth, planets and space*, 50(1):3–8.

Pedersen, H. A., Leroy, N., Zigone, D., Vallée, M., Ringler, A. T., and Wilson, D. C. (2020). Using component ratios to detect metadata and instrument problems of seismic stations: Examples from 18 yr of geoscope data. *Seismological Research Letters*, 91(1):272–286.

Peterson, J. (1993). Observations and modelling of seismic background noise. *US Geological Survey, open-file report*, 93 -322:1–94.

Pourhoseingholi, A., Pourhoseingholi, M. A., Vahedi, M., Safaee, A., Moghimi-Dehkordi, B., Ghafarnejad, F., and Zali, M. R. (2008). Relation between demographic factors and type of gastrointestinal cancer using probit and logit regression. *Asian Pac J Cancer Prev*, 9(4):753–5.

RESIF (1995). Resif-rlbp french broad-band network, resif-rap strong motion network and other seismic stations in metropolitan france.

Ringler, A. T., Hagerty, M., Holland, J., Gonzales, A., Gee, L. S., Edwards, J., Wilson, D., and Baker, A. M. (2015). The data quality analyzer: A quality control program for seismic data. *Computers & Geosciences*, 76:96–111.

Stutzmann, E., Ardhuin, F., Schimmel, M., Mangeney, A., and Patau, G. (2012). Modelling long-term seismic noise in various environments. *Geophys. J. Int.*, 191(2):707–722.

Tasič, I. (2018). Interdependent quality control of collocated seismometer and accelerometer. *Journal of Seismology*, 22(6):1595–1612.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

vanRossum, G. (1995). Python reference manual. *Department of Computer Science [CS]*, (R 9525).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Zhong, T., Li, Y., Wu, N., Nie, P., and Yang, B. (2015a). Statistical properties of the random noise in seismic data. *Journal of applied Geophysics*, 118:84–91.

Zhong, T., Li, Y., Wu, N., Nie, P., and Yang, B. (2015b). A study on the stationarity and gaussianity of the background noise in land-seismic prospecting. *Geophysics*, 80(4):V67–V82.

A The Probit function

The so-called *Probit function* was first introduced by Bliss (1934). This probabilistic function was originally developped to measure the effectiveness of a poison used in the fight of insect pests. However, it turns out that the Probit function goes beyond the scope of Biology and concerns many fields (*e.g.* Hoffman and Low, 1981; Kockelman and Kweon, 2002; Pourhoseingholi et al., 2008). The wide range of applications is logically due to the fact that the distribution of any standard Gaussian law converges toward the Probit function. Moreover, the mathematical progress during the past decades allowed a better understanding of the Probit function and its properties (Finney, 1971; Alu, 2011). We present hereafter the mathematical theory of the Probit function. We focus on the analytical expression of the Probit function and prove the link between any sorted standard Gaussian set of samples and the Probit function through a convergence theorem.

A.1 Definitions

The cumulative distributive function (CDF) of a random real-value variable X is a function (not necessarily continuous), defined as

$$F(t) = \mathbb{P}(X \leq t), \quad \forall t \in \mathbb{R}. \tag{A1}$$

For any CDF named F, we can define its related quantile function,

$$Q(u) = \inf\{x \in \mathbb{R} \ ; \ F(x) \geq u\}, \quad \forall u \in [0,1]. \tag{A2}$$

Hence, Q is the left inverse of F. In the special case of a continuous CDF, we have then $Q = F^{-1}$.

A.2 The Probit function

We denote as ϕ the CDF in the special case of the standard Gaussian law ($\mu = 0$ and $\sigma = 1$). The related quantile function Q is now called the *Probit function*, and since ϕ is continous, $Q = \phi^{-1}$.

It is well known that ϕ can be expressed as

$$\phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}\Big(\frac{x}{\sqrt{2}}\Big)\right), \tag{A3}$$

where erf denotes the error function. Consequently, computing the inverse function of ϕ, the analytic expression of the Probit function is thus given by

$$\phi^{-1}(u) = \sqrt{2}\,\mathrm{erf}^{-1}(2u - 1), \tag{A4}$$

where erf^{-1} could be, in practical, approximated by a Mac Laurin expansion (*e.g.*, Blair et al., 1976). The representative curve of the Probit function is plotted in green in Fig. A1.

A.3 Empirical quantile function

This section is devoted to the link between the discrete equivalents of the CDF and the quantile functions, obtained from a given statistical sample $(X_1, .., X_n)$. This leads to the definition of both empirical CDF and empirical quantile function.

For any set of n samples $(X_1, .., X_n)$, we define the empirical CDF,

$$F_n(t) = \frac{1}{n}\mathrm{Card}\big(\{X_i\ ;\ Xi \leq t\}\big), \tag{A5}$$

where $\mathrm{Card}(X)$ represents the cardinal function. Among the n values, $F_n(t)$ thus represents the proportion of points lower than t in a given set of samples.

Following eq. (A2), the empirical quantile function can be defined as

$$Q_n(u) = \inf\{x \in (X_1, .., X_n)\ ;\ F_n(x) \geq u\}, \quad \forall u \in [0, 1]. \tag{A6}$$

The empirical quantile function represents, for a given sample $(X_1, ..., X_n)$, its values sorted by increasing order of amplitudes. Indeed, $Q_n(u)$ represents the u-th quantile of a dataset $(X_1, .., X_n)$ as its smallest value for which the empirical CDF $F_n(x)$ is greater than or equal to u, effectively sorting the samples by increasing order of amplitudes.
For a set of random values, the convergence between the CDF and the empirical CDF can be found in the Glivenko-Cantelli theorem (Glivenko, 1933).

Theorem 1 ***Glivenko-Cantelli theorem***

Assuming that $(X_1, ..., X_n)$ *are independent and identically-distributed random variables in* $\mathbb{R}$ *with common cumulative distribution function* F*. Then, we have an uniform convergence almost surely of* F_n *toward* F*, i.e.*

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \to +\infty]{} 0 \text{ almost surely.} \tag{A7}$$

Theorem 2

Assume that $(X_1, ..., X_n)$ *are independent and identically-distributed random variables in* $\mathbb{R}$ *with common cumulative distribution function* F *and quantile function* Q*. Noting* F_n *the empirical CDF and* Q_n *the empirical quantile function, we have the following equivalence:*

$$|F_n - F| \xrightarrow[n \to +\infty]{} 0 \iff |Q_n - Q| \xrightarrow[n \to +\infty]{} 0. \tag{A8}$$

Plus, in the special case of the the standard normal distribution, this convergence is uniform.

The proof is detailed in (Van der Vaart, 2000, chapter 21, lemma 21.2) and the uniform convergence in the particular case of the standard normal distribution is deduce by the Dini's theorem (Francinou et al., 2013). In the special case of the standard Gaussian law, the theorem 1 demonstrates the convergence of F_n towards ϕ, where F_n is the empirical CDF obtained from a random draw. Consequently, the theorem 2 ensures as well the convergence between ϕ_n^{-1} and ϕ^{-1}, where ϕ_n^{-1} denotes the empirical discrete Probit function. In order to illustrate this convergence, the result of a numerical experiment is presented in Fig. A1.

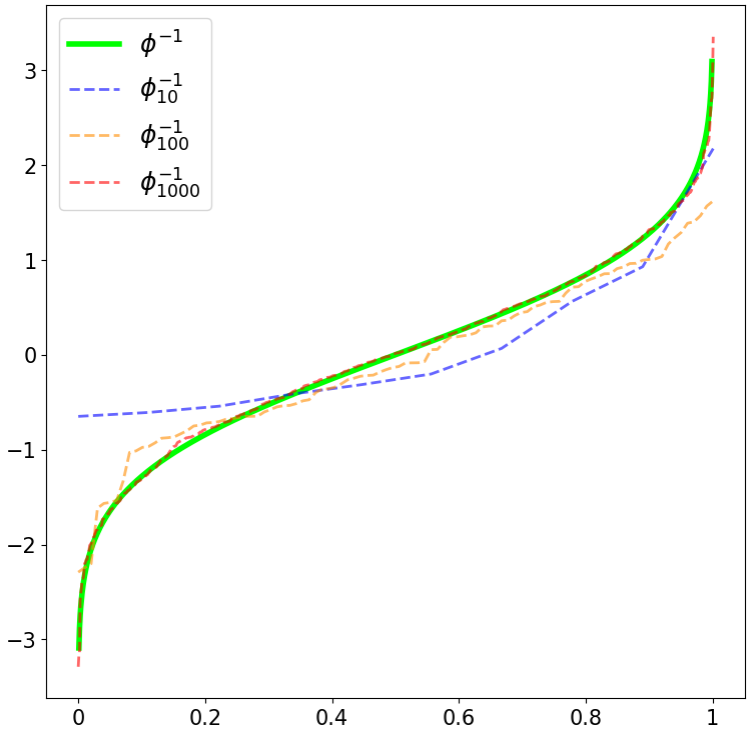

Figure A1. Illustration of the convergence of the empirical discrete Probit functions ϕ_n^{-1} towards the Probit function ϕ^{-1} (theorem 2). Examples for $n = 10$ (blue), 100 (orange) and $1,000$ (red).

Three random draws of n elements ($n = 10$, $n = 100$ and $n = 1,000$) are realised to obtain $(X_1, ...X_n)$, where $X_i \sim \mathcal{N}(0, 1), \forall i \in [1, n]$. Once data sets are sorted by increasing order of amplitude, they can be compared to the Probit function defined by eq. (A4), displayed in green. Each sorted data set thus is an empirical discrete Probit function, and we observe a reliable convergence since n is sufficently large.