A model based on the fractional Brownian motion for the temperature fluctuation in the Campi Flegrei caldera

Antonio Di Crescenzo\(^{(1)}\), Barbara Martinucci\(^{(1)}\) and Verdiana Mustaro\(^{(1)}\)  

\(^{(1)}\) Dipartimento di Matematica, Università di Salerno  
I-84084 Fisciano (SA), Italy  
Email:{adicrescenzo, bmartinucci, vmustaro}@unisa.it  

Abstract  
The aim of this research is to identify an efficient model to describe the fluctuations around the trend of the soil temperatures monitored in the volcanic caldera of the Campi Flegrei area in Naples (Italy). The study focuses on the data concerning the temperatures in the mentioned area through a seven-year period. The research is initially finalized to identify the deterministic component of the model, given by the seasonal trend of the temperatures, which is obtained through an adapted regression method on the time series. Subsequently, the stochastic component from the time series is tested to represent a fractional Brownian motion (fBm). An estimation based on the periodogram of the data is used to establish that the data series follows a fBm motion, rather than a fractional Gaussian noise. An estimation of the Hurst exponent \(H\) of the process is also obtained. Finally, an inference test based on the detrended moving average of the data is adopted in order to assess the hypothesis that the time series follows a suitably estimated fBm.  

Keywords: Fractional Brownian motion, stochastic model, regression, time series, residuals analysis, Hurst exponent, Campi Flegrei caldera, temperature fluctuation  

1 Introduction  
In several applied contexts, one of the main problems in the description of a phenomenon that evolves over time is the individualization of the probabilistic laws by which the phenomenon itself is driven. To this aim, schemes based on the superposition of deterministic and random components are often chosen, in which the deterministic part describes the phenomenon’s trend, while the random one describes the fluctuations determined by unpredictable exogenous factors. An example of such model is presented in the work by Sebastiani and Malagnini [1], where a physical model with
non-constant variance is proposed to describe the phenomenon of coupling erosion that caused the 2011 earthquake in Tohoku-Oki (Japan).

The aim of the present study is to implement an analogous scheme to describe the soil temperatures observed on the surface of the Campi Flegrei caldera. This volcanic area is located near the city of Naples, in Italy, and is famous for the ground deformations that have been registered in the area since ancient times. Recent observations, carried out in the years from 2011 to 2017, were led to monitor the presence of radon in two sites of the area (cf. Sabba rese et al. [2]). In the work [2], the presence of radon in the soil was tracked, along with its dependence on physical quantities such as temperature, pressure and humidity, in order to prove how the seismic activity in the Campi Flegrei area was influenced by the presence of the element.

Starting from the research mentioned above, in this work we aim to build a model able to describe the fluctuations of soil temperatures in the caldera through the identification of the temporal trend and the corresponding random component.

What emerged from the present study is that the deterministic component is well described by a piecewise linear model, where the segments’ slopes are alternating in sign. The slopes can be determined by applying the ordinary least squares (OLS) method on the observed data. The endpoints for each segment of the piecewise curve are chosen through an empirical method. The stochastic model describing the random fluctuations component is instead recognized to be the fractional Brownian motion (fBm). For a description of the most relevant probabilistic and statistic aspects of fBm see, for example, Nourdin [3] and Prakasa-Rao [4]. This stochastic process has been proposed in several studies for modeling various geophysical phenomena. In the work by Mattia et al. [5], the fBm is proposed as an approximation of the process describing the trend of ground data involving soil roughness collected over three different European sites. In the paper by Yin and Ranalli [6], the authors show that the event of earthquake rupturing in a fault is related to factors such as static shear stress and static frictional strength through the potential dynamic stress drop. The latter is regarded as a one-dimensional stochastic process and it is modeled as a fBm with Hurst exponent close to zero.

The evidence of reasonableness of the model firstly emerges as a consequence of a statistical test, in which the hypothesis that the random fluctuations of the model can be described by a Brownian motion is rejected. The subsequent phase of the work involves the study of the periodogram of the data. It allows to verify that the random fluctuations are well described by a fBm, in opposition to a fractional Gaussian noise (fGn). Such analysis also leads to the estimation of the involved parameters, namely the Hurst exponent $H$ and the scale parameter $D$ of the fBm. Many estimation techniques for the former parameters have been proposed in the literature, such as the methods presented in [7]. Other works often provide modified versions of the already known algorithms.

Subsequently, the study moves towards analysing the residual series emerging from the considered stochastic process and a simulated fBm with the previously estimated parameters. First of all, such residual series is plotted with respect to the temperatures series, which suggests the Gaus-
sianity of the residuals. Then, the procedure is repeated through suitable simulations. Two statistical tests are performed on the obtained values, both leading to the acceptance of the null hypothesis of Gaussianity.

The goodness of the considered model is finally confirmed through an inference test in which the test statistic involves the detrending moving averages obtained from the data.

Here is the plan of the article. In Section 2 we introduce the starting equation modeling the soil temperatures trend (cf. (1)). Section 3 focuses on the data analysis directed towards identifying the deterministic component of the temperatures trend. The remaining part of the article is dedicated to the analysis of the stochastic component of the model. More specifically, Section 4 is aimed to the study of the periodogram obtained from the data observations, from which emerges the eligibility of fBm to describe the random component of the temperature compared to fGn. An estimation of the relevant parameters of the process, i.e. the Hurst exponent and the scale parameter, is also performed. The testing on the residuals of the model is then performed by resorting to the Shapiro-Wilk test and the Robust Jarque-Bera test. The work is completed by Section 5, which contains the statistical test leading to the acceptance of the model, and Section 6, which collects the conclusions of the study.

2 The stochastic model

The time series considered throughout this study describes the temperatures of the surface soil of the Campi Flegrei caldera, obtained via daily observations in the time period between 07/01/2011 and 12/31/2017. The main aim of the research is to develop a model able to describe the fluctuations of the data. As usually observed in geophysical studies, the trajectory of the time series is seen as the superposition of a deterministic component and a stochastic one, where the latter represents the fluctuations from the trend. As a consequence, the process \( \{ X(t), \ t \geq 0 \} \) that represents the daily observed soil temperatures is described by the equation

\[
X(t) = r(t) + B_H(t),
\]

where \( r(t) \) constitutes the deterministic component of the process, while \( B_H(t) \) is a stochastic process. We will show that a suitable assumption is to identify \( B_H(t) \) a fBm. A similar model has been proposed for the vertical motion of the soil in the Campi Flegrei area, cf. Travaglino et al. [8].

Under the assumption that \( B_H(t) \) is indeed a fBm, it is easy to observe that the model shown in (1) is such that \( \text{E}(X(t)) = r(t) \). We recall, in fact, that the fractional Brownian motion is a continuous-time Gaussian process with zero mean and covariance

\[
\text{Cov}[B_H(t), B_H(s)] = D(t^{2H} + s^{2H} - |t - s|^{2H}), \quad t, s \geq 0,
\]

where \( D > 0 \) is a scale parameter, named diffusion constant, and \( 0 < H < 1 \) is a parameter known as the Hurst exponent. The process was first introduced by Mandelbrot and Van Ness in [9]. It is a generalization
of the Brownian motion, since for $D = 1/2$ and $H = 1/2$ it reduces to a standard Wiener process. We remark that fBm has the property of self-similarity, that is, for any choice of $a > 0$,

$$\{B_H(at), \ t \geq 0\} \overset{d}{=} \{a^H B_H(t), \ t \geq 0\},$$

meaning that the two processes are equal in distribution. The parameter $H$ is therefore also referred to as the scaling exponent or fractal index of the process.

The model (1) is not stationary, since fBm is not a stationary process itself. The process of the increments of $B_H(t)$, defined by $Z(t) = B_H(t + 1) - B_H(t)$ for all $t \geq 0$, is however stationary. This is known as fractionary Gaussian noise (fGn).

Introducing the process $Z(t)$ leads to another well-known property of fBm, that is the long-range dependency (LRD). As pointed out in [10], a stationary process $X(t)$ has long-range dependence (or is a long memory process) if its autocorrelation function, defined as

$$\rho(k) = \frac{\text{Cov}(X(t), X(t + k))}{\text{Var}[X(1)]},$$

satisfies the condition

$$\sum_{k=-\infty}^{+\infty} \rho(k) = \infty.$$  

This definition implies that the autocorrelation of the process decays slowly over time, making the sum of the autocorrelations divergent; therefore, if $X(t)$ is a LRD process, then

$$\lim_{k \to +\infty} \frac{\rho(k)}{ck^{-\alpha}} = 1,$$

where $c > 0$ and $0 < \alpha < 1$ are constants. This definition states that the decay of the autocorrelation function is power-like, hence slower than exponential. In the case of fBm, long-range dependence can be seen looking at the increments in $Z(t)$. Moreover, the parameter $\alpha$ is related to the Hurst exponent through the equation $\alpha = 2H - 2$, evidencing that the value of the Hurst exponent can be used to determine the nature of the process. We recall that (see also [10] and [11])

- if $\frac{1}{2} < H < 1$, then the increments of the process are positively correlated, making the process persistent, i.e. likely to keep the trend exhibited in the previous observations;
- if $0 < H < \frac{1}{2}$, the increments of the process are negatively correlated and the process is counter-persistent, i.e. likely to break the trend followed in the past.

A relation similar to (3), but concerning the frequency domain of the time series, is shown in Section 3, where it is used for the estimation of a parameter of the process.
The deterministic component

The model \((\text{1})\) is adopted to analyse the time series of the recorded temperatures of the Campi Flegrei caldera soil, cf. Sabbarese et al. [2]. The initial data set consists of \(\hat{N} + 1 = 2005\) observations collected at the times indicated by \(t_0, t_1, \ldots, t_{\hat{N}}\) and shown in the scatter plot in Figure 1.

![Temperature - OLB](image1)

Figure 1: Scatter plot of the data series \(X(t)\) for the temperatures of the surface soil of the Campi Flegrei caldera in the Monte Olevano (NA) site.

Figure 2: Seasonal plot of the temperatures described by \(X(t)\). This is obtained by taking weekly spaced values from the time series \(X(t)\) and representing each year of observation (2011-2017) separately.

![Seasonal plot](image2)

The observations are not equally spaced time-wise, since the data series was collected daily with lack of a quite large number of observations. As shown in the plot in Figure 2, a seasonal trend is clearly visible in
the dataset, with the temperatures increasing during spring and summer in each year, and then starting to decrease around September. For this reason, the deterministic trend \( r(t) \) is constructed by alternating segments with opposite slopes, obtained from the data through the OLS method. The first and the last measurement of the dataset, which are also the extremes of the curve, are given by

\[
\begin{align*}
  t_0 &= 07/29/2011, \quad X(t_0) = 31.78; \\
  t_N &= 12/31/2017, \quad X(t_N) = 29.94.
\end{align*}
\]

To obtain the points in which the curve changes its slope, an heuristic method based on linear regression is used, developed upon the methodologies illustrated by Hudson in [12]. First, for the dataset, we detect the local minima \( m_i, \ i = 1, \ldots, 6 \) and maxima \( M_i, \ i = 1, \ldots, 7 \), as shown in Table 1.

Table 1: Local maxima \( M_i \) and minima \( m_i \) for each subset of the dataset and the corresponding dates \( \tau_i \) (mm/dd/year).

| \( \tau_i \)     | \( M_i = X(\tau_i) \) (°C) | \( \tau_i \) | \( m_i = X(\tau_i) \) (°C) |
|------------------|-----------------------------|--------------|-----------------------------|
| 09/16/2011       | 31.78                       | 02/14/2012   | 28.68                       |
| 08/25/2012       | 33.6                        | 03/22/2013   | 28.63                       |
| 09/11/2013       | 33.31                       | 03/04/2014   | 29.17                       |
| 09/22/2014       | 33.4                        | 02/27/2015   | 27.08                       |
| 09/05/2015       | 32.4                        | 03/11/2016   | 29.27                       |
| 09/19/2016       | 33.29                       | 02/03/2017   | 28.88                       |
| 08/30/2017       | 34.04                       |              |                             |

The time \( t_0 \) is used as the first extreme of the first segment of the piecewise linear curve. Subsequently, the slope \( \alpha_j \) and the intercept of the regression line between \( \tau_0 \) and the first local maxima \( M_1 \) are estimated using the OLS method. The corresponding coefficient of determination \( R^2 \) is also calculated. The second step consists of repeating the operation, choosing the second extreme among the 5 values that precede and the 5 ones that follow \( M_1 \). Hence, among the 11 possible choices, we choose the regression line that maximises \( R^2 \). Table 2 shows the values obtained for each endpoint.
Table 2: Slopes of the segments obtained by linear regression between initial time $t_0$ and a neighbourhood of $M_1$, with their corresponding coefficients of determination. The slope is chosen such that $R^2$ is the highest, as underlined.

| $t_j$       | $\alpha_j$ | $R^2$ |
|------------|------------|------|
| 09/11/2011 | 0.0219     | 0.9302 |
| 09/12/2011 | 0.0217     | 0.9326 |
| 09/13/2011 | 0.0215     | 0.9341 |
| 09/14/2011 | 0.0213     | 0.9369 |
| 09/15/2011 | 0.0213     | 0.9405 |
| 09/16/2011 | 0.0213     | 0.9442 |
| 09/23/2011 | 0.0199     | 0.9021 |
| 09/24/2011 | 0.0190     | 0.8787 |
| 09/25/2011 | 0.0182     | 0.8574 |
| 09/26/2011 | 0.0174     | 0.8369 |
| 09/27/2011 | 0.0162     | 0.7880 |

Then, the same method is applied for the next 12 segments, in which the first initial time is determined by the previous step. The last regression line is finally calculated using $t_N$ as the endpoint. We denote by $c_i > 0$ and $v_i < 0$ the alternating slopes obtained through the regression model. The endpoints $\theta_i$ for each interval, as well as the estimated slopes and the coefficients of determination, are reported in Table 3.

The values obtained by means of this procedure allow us to identify the deterministic component $r(t)$ of the model 1. The values of $R^2$ confirm that the regression lines provide a good fitting of the data in each interval. The scatter plot and the resulting linear model are shown in Figure 3.

Table 3: Estimated slopes $c_i$ and $v_i$, and corresponding coefficients of determination, evaluated for each subset $[\theta_{i-1}, \theta_i]$ of the data series, with $\theta_0 = t_0 = 07/29/2011$.

| $\theta_i$       | $c_i$     | $R^2$     | $\theta_i$       | $v_i$     | $R^2$     |
|------------------|-----------|-----------|------------------|-----------|-----------|
| 09/16/2011       | 0.0213    | 0.9442    | 02/27/2012       | -0.0264   | 0.9709    |
| 09/13/2012       | 0.0262    | 0.9356    | 03/07/2013       | -0.0311   | 0.9722    |
| 09/06/2013       | 0.0287    | 0.9818    | 03/07/2014       | -0.0256   | 0.9573    |
| 09/24/2014       | 0.0237    | 0.9701    | 03/01/2015       | -0.0418   | 0.9635    |
| 08/23/2015       | 0.0294    | 0.9659    | 03/19/2016       | -0.0145   | 0.9287    |
| 09/08/2016       | 0.0238    | 0.9619    | 02/16/2017       | -0.0261   | 0.9357    |
| 09/14/2017       | 0.0228    | 0.9674    | 12/31/2017       | -0.0346   | 0.9153    |
4 Analysis of the stochastic component

In this section we focus on the investigation concerning the nature of the time series \( B_H(t) = X(t) - r(t) \). To this aim, some preparatory interpolations are performed on the dataset as customary. Since the times of the initial series are not equally spaced, we collect the missing data by performing a linear interpolation, covering the whole observation period. The prepared dataset obtained after this phase consists of \( N + 1 = 2348 \) observations. A plot of \( B_H(t) \) is shown in Figure 5.

As a preliminary study of the nature of the process \( B_H(t) \), we represent the time series through an histogram and a qq-plot, as shown in Figure 4. The obtained plots suggest to classify the stochastic component among Gaussian processes, thus justifying the following testing.

4.1 Testing for Brownian motion

With reference to model (1), hereafter we perform a test on the process \( B_H(t) \), aiming to assess its suitable nature. With reference to the Brownian motion \( B(t) \), we consider the hypothesis

\[
\mathcal{H}_0 : B_H(t) = \sigma B(t), \\
\mathcal{H}_1 : B_H(t) \text{ is a confined or directed diffusion.}
\]
By adopting the test proposed by Briane et al. in [13], the asymptotic region of acceptance for $H_0$ is given by

$$
\left\{ q\left(\frac{\alpha}{2}\right) \leq \frac{S_N^D}{\hat{\sigma} \sqrt{t_N}} \leq q\left(1 - \frac{\alpha}{2}\right) \right\},
$$

where

$$
S_N^D = \max_{j=1,\ldots,N} |B_H(t_j) - B_H(t_0)|,
$$

and

$$
\hat{\sigma} = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left[ \frac{B_H(t_j) - B_H(t_{j-1})}{t_j - t_{j-1}} \right]^2 \right\}^{1/2},
$$

and $q(\alpha)$ is the quantile of

$$
\sup_{0 \leq s \leq 1} |B(s) - B(0)|.
$$

The extremes of the acceptance interval for the significance level $\alpha = 0.05$ are given by

$$
q(0.025) = 0.834, \quad q(0.975) = 2.940,
$$

while the value obtained for the test statistic is 0.0787. The null hypothesis of the process $B_H(t)$ being a Brownian motion is therefore rejected. This justifies the further analysis conducted hereafter.

Let us now conduct an investigation on $B_H(t)$ to verify whether it might follow a fBm or fGn trend. Consequently, we also estimate the Hurst exponent $H$ of the process. To this aim, we apply the theoretical procedure consisting of evaluating the Fourier spectrum $S(f)$ of the data and studying its behaviour with respect to the frequencies $f$. In fact, a relation analogous to (3) for long memory processes can be presented in the frequency domain of the time series. In particular, for a fBm or a fGn process, the estimated Fourier spectrum $S(f)$ and the frequencies $f$ are asymptotically evaluated as

$$
S(f) = S(f_0) f^{-\beta}, \quad (4)
$$

Figure 5: Plot of the time series $B_H(t) = X(t) - r(t)$. 
where both $S(f_0)$ and $\beta$ are constants. The coefficient $\beta$ is linked to the Hurst exponent $H$ by different equations, depending on the nature of the underlying process. Indeed, for a fGn (fBm) process, Eq. (4) holds with $\beta = 2H - 1$ ($\beta = 2H + 1$), cf. [14], [15] and [16]. Recalling that $0 < H < 1$,

(i) if $-1 < \beta < 1$ then the time series can be identified as a realization of a fractional Gaussian noise, while
(ii) if $1 < \beta < 3$ then the data series represents a sample path of a fractional Brownian motion.

From (4) we get
\[
\log S(f) = \log S(f_0) - \beta \log f,
\]
so that an estimation of the coefficient $\beta$ can be obtained by plotting on a log-log scale the frequencies and the values of $S(f)$ and then taking the opposite value of the slope of the least squares line as estimation.

To attain an estimation of function $S(f)$, we evaluate the periodogram of the data series, which for $B_H(t)$ is defined by
\[
I(f) = \left| \frac{1}{2\pi N} \sum_{i=0}^{N} B_H(t_i)e^{i t_1} \right|^2.
\]

In order to achieve the best possible approximation for the Fourier spectrum, many modified versions of the periodogram have been devised. In particular, we choose the estimation of $S(f)$ using the modified periodogram method suggested by Welch in [17]. The technique consists in dividing the signal into overlapping segments and averaging the modified
periodograms calculated in each window to obtain the final result. The Welch’s modified periodogram is shown in Figure 6.

Hence, by applying a log-log transformation on the frequencies and the estimated periodogram, we obtain the plot shown in Figure 7. The estimate of \( \beta \) obtained as the slope of the regression line is given by

\[
\beta = 1.853,
\]

with corresponding \( R^2 = 0.907 \). Since the coefficient \( \beta \) belongs to the interval \((1, 3)\), we can conclude that the data series follows a fBm trend. This also leads to an estimation of the Hurst parameter as follows:

\[
\hat{H} = \frac{\beta - 1}{2} = 0.427.
\]

In order to complete the estimation of the parameters for \( B_H(t) \), we also evaluate the diffusion constant \( D \) of the process. This is done directly from the data, calculating the sample variance of the increments \( B_H(t) - B_H(t - 1) \) of the data series. The estimate obtained so far is

\[
\hat{D} = 0.0846.
\]

To further confirm the hypothesis suggested from the periodogram, we perform an analysis of the residuals of the time series making use of simulations of fBm. To this aim, we denote with \( B_{\hat{H}, \hat{D}}(t) \) a simulated sample path of a fractional Brownian motion process with Hurst exponent and scale parameter estimated as in (5) and (6). Then, the residual

\[
z(t) = X(t) - r(t) - B_{\hat{H}, \hat{D}}(t)
\]
is evaluated by means of the residuals plot. Moreover, the normality of the residuals is assessed hereafter through two tests: the Shapiro-Wilk and the Robust Jarque-Bera test for Gaussianity.

## 4.2 Shapiro-Wilk test

Since the test statistic $W$ of the Shapiro-Wilk test tends to detect even small departures from the null hypothesis when the sample size is sufficiently large, it is convenient reducing the sample to a smaller size. Hence, from now on we consider a subset $\tilde{X}(t)$ of the main dataset, which contains weekly spaced observations. After simulating a sample path $\tilde{B}_{\hat{\alpha}, \hat{\beta}}$ of a fBm of length $M = \lceil(N + 1)/7 \rceil$, we plot the residual series

$$\tilde{z}(t) = \tilde{X}(t) - \tilde{r}(t) - \tilde{B}_{\hat{\alpha}, \hat{\beta}}(t)$$

with respect to the values of $\tilde{X}(t)$, see Figure 8. Notice that the residuals do not seem to follow a specific path, but appear to be distributed randomly with respect to $\tilde{X}(t)$.

The fBm path $\tilde{B}_{\hat{\alpha}, \hat{\beta}}(t)$ is then simulated again for a total of $n = 10^4$ iterations. For each simulation we evaluate the residuals series $\tilde{z}_i(t)$ and first perform a Shapiro-Wilk test, for $i = 1, \ldots, n$. Hence, we consider the

![Figure 8: Plot of the residuals series $\tilde{z}(t)$ with respect to the temperatures $\tilde{X}(t)$, equally spaced with one observation for each 7 days.](image)
collection of test statistics

\[ W_i = \frac{\sum_{j=1}^{M} a_j \tilde{z}_i(t_j)}{\sum_{j=1}^{M} (\tilde{z}_i(t_j) - \mu_i)^2}, \quad i = 1, \ldots, n, \tag{7} \]

where \( W_i \) refers to the \( i \)-th simulation, and where \( \mu_i \) represents the sample mean of \( \tilde{z}(t) \). The coefficients \( a_j \) –as well known– are related to some moments of the order statistics of i.i.d. random variables sampled from the standard normal distribution. The values of \( W_i \) as in Eq. (7) range between 0 and 1. The \( i \)-th test rejects the Gaussian hypothesis when \( W_i \) is close to 0, hence we accept the null hypothesis if \( W_i \geq 0.98 \), for \( i = 1, \ldots, n \). As shown in Figure 9, large proportion of the values of \( W_i \) is greater than 0.98. This result has been observed for more than 70% of the values of \( W_i \) even for various repetitions of the \( n \)-simulation procedure (cf. [18], [19]). This is also confirmed by the cumulative histogram of \( W_i \) shown in Figure 10. The results of the test suggest the acceptance of the Gaussianity hypothesis.

Figure 9: The values of the statistic \( W_i \) evaluated according to Eq. (7).

Figure 10: Cumulative histogram for the \( W_i \) statistic.

### 4.3 Robust Jarque-Bera test

Let us now perform a modified version of the Jarque-Bera test. The statistic for the original Jarque-Bera test for a sample of size \( n \) is given by (cf. [20])

\[ JB = n \left( \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} \right)^2 + \frac{n}{24} \left( \frac{\hat{\mu}_4}{\hat{\mu}_2} - 3 \right)^2, \]

which is a combination of the sample skewness \( \left( \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} \right) \) and the sample kurtosis \( \left( \frac{\hat{\mu}_4}{\hat{\mu}_2} - 3 \right) \) obtained from the data through the estimates \( \hat{\mu}_i \) of the central moments, \( i = 2, 3, 4 \). If the data series is distributed normally, one has that \( JB \sim \chi^2(2) \) asymptotically. Thus, under the null hypothesis of the time series being Gaussian, both the expected skewness and the expected kurtosis are 0. Therefore, higher values of the \( JB \) statistic lead to the rejection of the Gaussianity hypothesis.
For our testing procedure, we follow the modified version of the test proposed by Gel and Gastwirth in [21], which is known as the Robust Jarque-Bera (RJB) test. The modification consists in substituting the estimate of the spread, formerly \( \hat{\mu}_2 \), with the average absolute deviation from the sample median, henceforth denoted as Med, given by

\[
J_M = \sqrt{\frac{\pi}{2}} \frac{1}{M} \sum_{j=1}^{M} |\tilde{z}(t_j) - \text{Med}|,
\]

where \( M \) refers to the sample size. This substitution makes the statistics less influenced by outliers. Therefore, the statistic becomes

\[
J_{BM} = \frac{M}{C_1} \left( \frac{\hat{\mu}_3}{\hat{J}_M^1} \right)^2 + \frac{M}{C_2} \left( \frac{\hat{\mu}_4}{\hat{J}_M^2} - 3 \right)^2,
\]

where \( C_1 \) and \( C_2 \) are positive constants estimating moments of \( J_M \), obtained through Monte Carlo simulations. Moreover, the p-values for the RJB test are obtained by use of \( k = 10000 \) Monte Carlo simulations instead of the \( \chi^2(2) \) distribution, thus obtaining a better approximation (cf. [22]).

To perform the RJB test, we repeat 10 blocks of \( n \) simulations with three different numbers of iterations (\( n = 100, 500 \) and 1000). For each simulation, the residual series

\[
\{\tilde{z}_i(t_j), \, j = 1, \ldots, M\}_{i=1,\ldots,n}
\]

is tested. The results, reported in Table 4, show compliance with the Shapiro-Wilk test and lead to the acceptance of the Gaussianity hypothesis for the residuals series.

Table 4: Percentages of acceptable values for the residuals series, tested with the RJB method, repeated for \( h = 1, \ldots, 10 \), with significance level \( \alpha = 0.02 \) and different values of \( n \).

| \( h \) | \( n = 100 \) | \( n = 500 \) | \( n = 1000 \) |
|------|--------|--------|-----------|
| 1    | 64%    | 66.8%  | 68.0%     |
| 2    | 76%    | 72.2%  | 69.9%     |
| 3    | 68%    | 70.0%  | 66.1%     |
| 4    | 66%    | 69.4%  | 71.3%     |
| 5    | 70%    | 71.0%  | 70.5%     |
| 6    | 68%    | 71.4%  | 71.1%     |
| 7    | 68%    | 70.8%  | 70.5%     |
| 8    | 71%    | 67.0%  | 69.9%     |
| 9    | 70%    | 73.0%  | 68.1%     |
| 10   | 70%    | 69.4%  | 71.5%     |
5 Statistical test for the model

To further confirm the results obtained in Section 4, we perform a statistical test on the whole $N$-term data series for the hypothesis of $B_H(t)$ being a fractional Brownian motion with parameters $\hat{H} = 0.427$ and $\hat{D} = 0.0846$.

To this aim, we use the inference test presented in Sikora [23], based on the detrending moving average (DMA) of the data. The hypotheses for the test are

$$H_0 : \{B_H(t_1), B_H(t_2), \ldots, B_H(t_N)\} \text{ is a trajectory of fBm with parameters } \hat{H} \text{ and } \hat{D},$$

$$H_1 : \{B_H(t_1), B_H(t_2), \ldots, B_H(t_N)\} \text{ is not a trajectory of fBm with } \text{parameters } \hat{H} \text{ and } \hat{D}. $$

The corresponding test statistic, for fixed $m > 1$, is given by

$$S^2(m) = \frac{1}{N - m} \sum_{j=m}^{N} (B_H(t_j) - \overline{B}_H^m(t_j))^2, \quad (9)$$

where $\overline{B}_H^m(t_j)$ denotes the moving average of the $m$ observations $B_H(t_j-i)$, $i = 0, 1, \ldots, m - 1$.

The $p$-value for this test is defined by an infinite series, so it needs to be computed as an empirical quantile, obtained from series of generalized chi-squared random variables. Recalling Step 5 in Section 3 of Sikora [23], the $p$-value of the test is calculated as

$$p = \frac{2}{L} \min \{ \#(\sigma^2_l(m) < S^2(m)), \#(\sigma^2_l(m) > S^2(m)) \}. \quad (10)$$

In this formula, $L$ is the number of samples generated for the chi-squared distribution. For a good estimate of $p$ we set $L = 1000$. Moreover, in Eq. (10), we consider

$$\sigma^2_l(m) = \frac{1}{N - m} \sum_{j=1}^{N-m+1} \lambda_j(m)U^{l}_j, \quad l = 1, 2, \ldots, L,$$

where $\{U^{l}_1, U^{l}_2, \ldots, U^{l}_{N-m+1}\}$ is the $l$-th chi-square sample and $\lambda_j(m)$ is the $j$-th eigenvalue of the sample matrix

$$\hat{\Sigma} = \{E[B_H(t_j)B_H(t_k)]; j, k = 1, 2, \ldots, N - m + 1\}.$$

Recalling (9), since the search for an optimal value for $m$ is still an open question, we divide the data series into 10 subsets of about 250 observations each, in order to adapt the procedure even to cases with large number of observed data. Then we run the test in every subset for each suitable value of $m$.

After that, we collect the $p$-values and create a box plot with jitter for each subset, in order to determine the outlier values for $p$, cf. Figure [11].

The significance level considered for this test is $\alpha = 0.02$. Therefore the results in each subset are collected and divided into three categories:
Figure 11: Box plot for the values of $p$ obtained in the 10 subsets of the data.

(i) for $p \leq 0.02$, the null hypothesis is rejected,

(ii) for $0.02 < p \leq 0.05$ the values are considered as "warning",

(iii) for $p > 0.05$ the hypothesis $H_0$ cannot be rejected.

Table 5 shows the results, with the percentages of the number of rejected, warning and acceptable values over the total of $p$-values calculated in the 10 datasets. It can be observed that the null hypothesis cannot be rejected in almost 90% of the cases covered, considering all the subsets and all the different choices of $m$ for the test statistic. This allows to state the validity of the hypothesis that the time series $B_H(t)$ follows a fBm trend.

Table 5: Number of cases in which the DMA statistic test has different outcomes for the data series $B(t)$.

| Set | Reject values $(p \leq 0.02)$ | Perc. | Warning values $(0.02 \leq p \leq 0.05)$ | Perc. | Acceptable values $(p > 0.05)$ | Perc. |
|-----|-------------------------------|-------|----------------------------------------|-------|-------------------------------|-------|
| 1   | 0                             | 0%    | 12                                     | 4.86% | 235                           | 95.14%|
| 2   | 70                            | 28.23%| 12                                     | 4.84% | 154                           | 62.10%|
| 3   | 17                            | 6.85% | 6                                      | 2.42% | 219                           | 88.31%|
| 4   | 70                            | 28.23%| 6                                      | 2.42% | 166                           | 66.94%|
| 5   | 23                            | 9.27% | 10                                     | 4.03% | 205                           | 82.66%|
| 6   | 4                             | 1.61% | 4                                      | 1.61% | 236                           | 95.16%|
| 7   | 3                             | 1.21% | 3                                      | 1.21% | 239                           | 96.37%|
| 8   | 0                             | 0%    | 0                                      | 0%    | 248                           | 100%  |
| 9   | 1                             | 0.4%  | 0                                      | 0%    | 247                           | 99.6% |
| 10  | 0                             | 0%    | 0                                      | 0%    | 95                            | 100%  |
6 Conclusions

The performed analysis shows that the proposed stochastic model is suitable to fit the observed temperatures. Indeed, the obtained results suggest that the coefficients obtained for the process $X(t)$ display high values for $R^2$, this ensuring the goodness of the fit. In conclusion, the study reveals that the evolution of the temperatures in the Campi Flegrei caldera can be modeled by the sum of a deterministic component, that represents the seasonal trend, and a fractional Brownian motion.

Possible future developments of the performed analysis can be oriented (i) to the prevision of the trend of the seasonal terms, and (ii) to the study of other quantities of interest, such as the leaked radon.

Acknowledgments The authors are members of the research group GNCS of INdAM (Istituto Nazionale di Alta Matematica). This work is supported in part by the Italian MIUR-PRIN 2017, project ‘Stochastic Models for Complex Systems’, No. 2017JFFHSI.

References

[1] Sebastiani, G.; Malagnini, L. (2020). Forecasting the Next Parkfield Mainshock on the San Andreas Fault (California). Journal of Ecology & Natural Resources 2020, 4. 10.23880/JENR-16000218.

[2] Sabbarese, C.; Ambrosino, F.; Chiodini, G. et al. Continuous radon monitoring during seven years of volcanic unrest at Campi Flegrei caldera (Italy). Scientific Reports 2020 10:9551;

[3] Nourdin, I. (2012) Selected Aspects of Fractional Brownian Motion. Springer. 4, Bocconi & Springer Series. (hal-01314412)

[4] Prakasa Rao, B. L. S. (2010). Statistical Inference for Fractional Diffusion Processes. 10.1002/9780470667125.

[5] Mattia, F.; Le Toan, T.; Davidson, M.; Borgeaud, M. On the scattering from natural rough surfaces. IGARSS 1999, 2413 - 2415 vol.5. 10.1109/IGARSS.1999.771527.

[6] Yin, Z.M.; Ranalli, G. Modelling of earthquake rupturing as a stochastic process and estimation of its distribution function from earthquake observations. Geophysical Journal International 1995, Volume 123, Issue 3, pp. 838–848.

[7] Taqqu, M. S.; Teverovsky, V.; Willinger, W. Estimators for long-range dependence: an empirical study. Fractals 1996, vol. 3, no. 4, pp. 785- 798, 1995. Reprinted in C.J.G. Evertsz, H.-O. Peitgen, R.F. Voss (Eds.), Fractal Geometry and Analysis, World Scientific, Singapore.

[8] Travaglini, F.; Di Crescenzo A.; Martinucci, B.; Scarpa, R. A New Model of Campi Flegrei Inflation and Deflation Episodes based on Brownian Motion Driven by Telegraph Process. Mathematical Geosciences 2018 50:961-975;
[9] Mandelbrot, B. B.; Van Ness, J. W. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review* **1968**, vol. 10, no. 4, pp. 422–437.

[10] Dieker, T., (2004). *Simulation of Fractional Brownian Motion*. Ph.D. Thesis, 160 Physics Department of Mathematical Sciences, University of Twente.

[11] Li, M. On the Long-Range Dependence of Fractional Brownian Motion. *Mathematical Problems in Engineering* **2013**, 10.1155/2013/842197.

[12] Hudson, D. J. Fitting Segmented Curves Whose Join Points Have to Be Estimated, *Journal of the American Statistical Association* **1966**, 61:316, 1097-1129, DOI: 10.1080/01621459.1966.10482198

[13] Briane, V.; Vimon, M.; Kervrann, C. An adaptive statistical test to detect non Brownian diffusion from particle trajectories. *Symposium on Biomedical Imaging 2016*, ISBI 2016-IEEE 13th.

[14] Roume, C.; Ezzina, S.; Blain, H.; Delignieres, D. (2019). Biases in the Simulation and Analysis of Fractal Processes. *Computational and Mathematical Methods in Medicine* **2019**, 1-12. 10.1155/2019/4025305.

[15] Montanari, A.; Taqqu, M.S.; Teverovsky, V. Estimating Long-Range Dependence in the Presence of Periodicity: An Empirical Study. *Mathematical and Computer Modelling* **1999**, 29, 217-228. 10.1016/S0895-7177(99)00104-1.

[16] Flandrin P. On the spectrum of fractional Brownian motions. *IEEE Transactions on Information Theory* **1989**, vol. 35, no. 1, pp. 197-199, Jan. 1989, doi: 10.1109/18.42195.

[17] Welch, P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics* **1967**, vol. 15, no. 2, pp. 70-73, doi: 10.1109/TAU.1967.1161901.

[18] Royston, P. Approximating the Shapiro-Wilk W-test for non-normality. *Stat Comput* **1992**, 2, 117–119. https://doi.org/10.1007/BF01891203

[19] Royston, P. Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Journal of the Royal Statistical Society* **1995**, Series C (Applied Statistics), 44(4), 547–551. https://doi.org/10.2307/2986146

[20] Jarque, C.M. and Bera, A.K. A test for normality of observations and regression residuals. *International Statistical Review* **1987** 55, 163-172.

[21] Gel, Y. R. and Gastwirth, J. L. A robust modification of the Jarque–Bera test of normality. *Economics Letters* **2008** 99, 30-32.

[22] Deb, P. and Sefton, M. The distribution of a Lagrange multiplier test of normality. *Economics Letters* **1996** 51, 123-130.

[23] Sikora, G. Statistical test for fractional Brownian motion based on detrending moving average algorithm. *Chaos, Solitons & Fractals* **2018**, Volume 116, 54-62, ISSN 0960-0779;