frontiers

Stochastic models for radon daily time series: seasonality, stationarity, and long-range dependence detection

Marianna Siino [1], **Salvatore Scudero** [1,*] **and Antonino D'Alessandro** [1]

[1] *Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Nazionale Terremoti, Via di Vigna Murata 605, 00143, Rome, Italy*

Correspondence*:
Salvatore Scudero
salvatore.scudero@ingv.it

ABSTRACT

This paper detects the presence of seasonality, stationarity, and long-range memory structures in daily radon measurements in a permanent monitoring station in central Italy. The transient dynamics and the seasonality structure are identified by power spectral analysis based on the continuous wavelet transformation and a clear 1-year periodicity emerges. The stationarity in the data is assessed with the Dickey-Fuller test; the decay of the estimated autocorrelation function and the estimated Hurst exponent indicate the presence of long-range dependence. All the main characteristics of the data have been properly included in a modelling structure. In particular, an autoregressive fractionally integrated moving average (ARFIMA) model is estimated and compared with the classical ARMA and ARIMA models in terms of goodness of fit and, secondarily, of forecast evaluation. An autoregressive model with a non-integer value of the differencing parameter ($d = 0.278$) resulted to be the most appropriate on the basis of Akaike Information Criterion, the diagnostic on the residuals, and the Root Mean Squared Error. The results suggest that there is statistically-significant evidence for not rejecting the presence of long memory in the radon concentration. The radon measurements are better characterised as being stationary, but with long memory and so the statistical dependence decays more slowly than an exponential decay.

Keywords: ARIMA model, ARFIMA model, Continous wavelet transform, Radon, Spectral analysis, Forecast

1 INTRODUCTION

The monitoring of soil radon ($^{222}R_n$) emission is a relevant topic for the risk that this radioactive gas poses to human health but also for its relationship with environmental and geological processes. The radon signals usually present a complex dynamic structure that is directly and indirectly influenced by several factors, such as environmental and climatic conditions of the site, characteristics of the ground soil, tide, solar effect, etc (Pinault and Baubron, 1996; Piersanti et al., 2015; Siino et al., 2019b). All these factors have a different effect on the signal, as they can result either in a trend, seasonal, or stochastic component. For instance, climate or tidal forces, reflect in a multiple-seasonality of the radon time series: hourly, diurnal, multi-day, annual, and even multi-annual cycles have been detected in different studies worldwide (Crockett et al., 2006, 2018; Udovičić et al., 2014; Yan et al., 2017; Siino et al., 2019b; D'Alessandro et al., 2020). Of particular interest is the role of Rn as a potential earthquake precursor (Barbosa et al., 2015;

Woith, 2015; Baskaran, 2016; Morales-Simfors et al., 2019), because the fracturing processes in the crust could enhance the mobility of Rn towards the surface (Toutain and Baubron, 1999; Woith, 2015). Similarly, anomalies can be the result of weather episodes which cannot be explained by the meteorological variables. Whatever the cause, these anomalies can be masked within the signal, and a way to bring them to light would be to de-noise the signal from the trend and/or periodic components (Baykut et al., 2010; Siino et al., 2019b). As a matter of fact, it is a challenging task to untangle and properly quantify all of these effects on the radon fluctuations because Rn time series present generally a non-stationary behaviour, not constant variability over time and a long-term memory (Donner et al., 2015).

Methodologically, time series analysis techniques are proper statistical tools to extract meaningful characteristics from data. Moreover, because long-term records of environmental variables show often long-range memory, some other tools are usually applied. The fractionally integrated moving average models (ARFIMA(p,d,q)) have been widely used in the literature to describe meteorological variables (Yaya and Fashae, 2015; Bowers and Tung, 2018), pollutants and soil gas (Pan and Chen, 2008; Donner et al., 2015; Belbute and Pereira, 2017; Reisen et al., 2018), and hydrological time series (Montanari et al., 1997; Wang et al., 2007). This class of models is used when the long-term correlations in the data decay more slowly than an exponential form, that is a typical shape of autocorrelation in the autoregressive moving average (ARMA(p,q)) processes (Box et al., 2015). Furthermore, several papers investigate the predictability of ARFIMA model assessing multi-step ahead performance with respects to others univariate time series forecasting methods such as a naive method, random walk (with drift), ARMA with trend and seasonality, and the exponential smoothing (Papacharalampous et al., 2018a,b).

In the literature, the radon data have been described with different methods. Dunn and Henschel (1989) characterise a three weeks record at an hourly frequency using simple autoregressive-moving average (ARMA) models. Later, the Box–Jenkins methodology often used in econometrics, was applied to describe five years long radon time series considering a seasonal integrated, autoregressive moving averages model with exogenous variables (SARIMAX) also adding external covariates such as delayed atmospheric parameters (Stránský and Thinová, 2017). Donner et al. (2015) present complementary methods that have been applied for evaluating the presence of long-range correlations and fractal scaling in environmental radon measurements.

In this paper, we analyse a 3 years long radon concentration signal aiming at the assessment of a model which describes its dynamics with time series methodologies (Shumway and Stoffer, 2017). A comprehensive analysis of the seasonality structure is performed to detect clues about the stationarity and the presence of long-range memory in the data that could be related to geological processes. We estimate some ARFIMA models which explicitly consider simultaneously both the short-term and long-term correlation structure of the series. Moreover, we tested the forecast performance of the obtained models. The novelty of this analysis relies on the simultaneous estimation of seasonality and long-range memory in the estimation of proper ARFIMA stochastic models.

2 MATERIALS AND METHODS

In this section, the time series methods used in the analysis of daily radon measurements are described following Shumway and Stoffer (2017) and Beran (2017). We briefly present some tools useful to check in an observed time series the presence of non-stationarity (in terms of seasonality and trend) and long-range memory behaviours.

The seasonality behaviour in the data is studied by power spectral density based on the time-averaged continuous wavelet spectrogram (Daubechies, 1992; Conraria and Soares, 2011). To properly apply the

stochastic models, the daily radon time series is examined for the presence of stationarity. The Dickey-Fuller test (Dickey and Fuller, 1979) is used for this purpose to determine the presence of a unit root in an autoregressive model. The presence of long-range memory has been assessed on the data estimating the Hurst exponent (Hurst, 1950) and looking at the shape of the estimated autocorrelation coefficients for several lags. The presence of long-term memory can justify the estimation of the ARFIMA models, and their structure is also explained.

2.1 Spectral analysis for seasonal detection

In this paragraph, we briefly describe the spectral analysis in the time-frequency domain based on continuous wavelet transformation following the notation in Daubechies (1992) and Conraria and Soares (2011).
The space $L^2()$ is the set of square integrable functions satisfying $\int_{-\infty}^{+\infty} |g(t)|^2 dt < \infty$, and denote by the capital letter, G(t) the Fourier transformation of a given function, $G(\omega) = \int_{-\infty}^{+\infty} g(t) e^{(-i\omega t)} dt$. A function $\psi(t) \in L^2(R)$ that satisfies the admissibility condition $\Psi(0) = \int_{-\infty}^{+\infty} \psi(t) dt = 0$ is called "mother wavelet", and a doubly-indexed family ("wavelet daughters") is generated by scaling and translating $\psi(\cdot)$: $\psi_{\tau,s}(t) = |s|^{-1/2} \psi\left(\frac{t-\tau}{s}\right)$ with $s, \tau \in$ and $s \neq 0$. In this analysis, we use the well-known, quite flexible, and complex-valued Morlet mother wavelet that takes the form $\psi(t) = \pi^{-1/4} e^{i\omega t} e^{-t^2/2}$. The local wavelet power spectrum (WPS) based on the continuous wavelet transformation (CWT) of a given function $g(t) \in L^2()$ with respect to the wavelet family

$$|WPS|_g(\tau, s) = |W_{x;\psi}(\tau, s)|^2 = \left| \int_{-\infty}^{+\infty} g(t) |s|^{-1/2} \psi^* \left(\frac{t-\tau}{s} \right) dt \right|^2 \tag{1}$$

where * represents the complex conjugate operation, s is the scale parameter controlling the wavelet width and τ controls the wavelet location in the time domain. The wavelet power spectrum (1) can be interpreted as the local variance of the time series.

To do a comparison with the classical spectral method, the previous quantity can be averaged over time (τ) obtaining the global wavelet power spectrum,

$$\int_{-\infty}^{+\infty} |W_{x;\psi}(\tau, s)|^2 d\tau \tag{2}$$

The peaks in the global power spectral density indicate the prevalent periods in the data. In this paper, the wavelet transformation and the computation of the global power spectrum are computed with the *WaveletComp* package (Roesch and Schmidbauer, 2018) in the *R* statistical software (Team, 2005).

2.2 Autocorrelation and partial autocorrelation functions

Given a time series $\{y_t\}$, the autocorrelation is the similarity between the observations as a function of the time lag between them. The j^{th} order autocorrelation $\rho(j)$ can be estimated by using the formula

$$\hat{\rho}(j) = \frac{\widehat{Cov}(y_t, y_{t-j})}{\widehat{Var}(y_t)} \tag{3}$$

where

$$\widehat{Cov}(y_t, y_{t-j}) = \frac{1}{n-1} \sum_{t=j+1}^{n} (y_t - \bar{y})(y_{t-j} - \bar{y}) \tag{4}$$

and

$$\widehat{Var}(y_t) = \frac{1}{n-1} \sum_{t=j+1}^{n} (y_t - \bar{y})^2 \tag{5}$$

In equations 4 and 5, $\bar{y}$ is the mean of the y_t, and 5 is just the special case of 4 in which $j = 0$. The empirical autocorrelation function (ACF) is $\hat{\rho}(j)$ defined in equation 3, computed in the data as a function of the lag j.
Moreover, another way to characterise the relationship between $|\{y_t\}$ and its lagged values is by the partial autocorrelation function, or PACF . The partial autocorrelation coefficient of order j, $\rho_j^{(j)}$ measures the effect (linear dependence) of y_t on y_{t-j} after removing the effect of y_{t-1}, y_{t-2}, $\dots$ y_{t-j-1} on both y_t and y_{t-j}. Each partial autocorrelation can be obtained as a series of regressions of the form:

$$y_t = \gamma^{(j)} + \rho_1^{(j)} y_{t-1} + \cdots + \rho_j^{(j)} y_{t-j} + \epsilon_t \tag{6}$$

The empirical PACF of order J is computed by running 6 for $j = 1, \dots, J$ and retaining only the estimate $\rho_j^{(j)}$ for each j. The shape of both the sample ACF and PACF provides a way to see which is the pattern of serial dependence, and it may help to suggest which kind of stochastic process would fit well the data.
If $\{y_t\}$ presents long-range memory, the correlation function 3 decays hyperbolically showing a power law distribution (Höll et al., 2019). Clauset et al. (2009) give an overview of the statistical methods that can be used to detect and characterise power-law distribution in empirical data.

2.3 The Hurst coefficient and the rescaled range (R/S) method

The Hurst exponent (H) is an index of long-term memory of time series $\{y_t\}$ originally developed for hydrological data (Hurst, 1950; Hurst et al., 1965). It is defined in asymptotic terms of the rescaled range as $E\left[R(n)/S(n)\right] = Cn^H$ as $n \to \infty$, where n is the number of data points in a time series, E is the expected value, C is a constant, $R(n)$ is the range of the first n cumulative deviations from the mean, and $S(n)$ is their standard deviation.
We consider the rescaled range (R/S) method to estimate H (Mandelbrot and Wallis, 1968, 1969).
Given y_t, the mean is computed ($m = \frac{1}{n}\sum_{i=1}^{n} y_i$) and the mean adjusted series $x_t = y_t - m$ for $t = 1, 2, \dots, n$. Then, the cumulative series is $z_t = \sum_{i=1}^{t} x_i$ and the range series is $R_t = \max(z_1, z_2, \dots, z_t) - \min(z_1, z_2, \dots, z_t)$ for $t = 1, 2, \dots, n$. A standard deviation series S is computes as $S_t = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(y_i - m(t))^2}$ for where m(t) is the mean for the time series values through time t. The following series of ratio is considered (R_t/S_t) for $t = 1, 2, \dots, n$.
The Hurst exponent is estimated as the slope of the line between $\log[R_t/S_t]$ and $\log t$. The long memory structure exists when $0 < H < 1$. If $H \geq 1$, the process has infinite variance and is non-stationary. If $0 < H < 0.5$ an anti-persistence structure exists, if $0.5 < H < 1$ the series is persistence, instead when $H = 0.5$, the process is a white noise. Other methods have been proposed in the literature to detect the presence of long-range temporal correlations in the presence of non-stationary in the data, i.e. the detrended fluctuation analysis (Höll et al., 2019).

2.4 Autoregressive fractionally integrated moving average model

Environmental data, and also radon measurements, can exhibit characteristics consistent with long-range memory in time series (Donner et al., 2015). Such characteristics consists in a specific structure of the autocorrelation function of the process.

If $\{y_t\}$ presents long-range memory, the correlation function 3 decays hyperbolically, rather than showing the exponential decay that is a characteristic of an ARIMA$(p, 0, q)$ process. A way of characterise long-range dependence in observational data is by fitting autoregressive fractionally integrated moving average (ARFIMA(p, d, q)) models and they are a natural extension of the classic ARIMA(p, d, q) models (Hosking, 1981). The ARFIMA models allow to handle explicitly both the short-term and the long-term correlation structure of a series. Let $\{y_t\}$ be a stationary process, an ARFIMA(p, d, q) process where p and q are integers and d is real, is represented as

$$\phi(B)\nabla^d(y_t - x_t\beta) = \theta(B)\epsilon_t \quad t = 1, \ldots, T \tag{7}$$

∇^d is the fractional differencing operator $\nabla^d = (1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k}(-B)^k$, B is the backward-shift operator defined by $By_t = y_{t-1}$ and $\{\epsilon_t\}$ is a white noise with variance σ_ϵ^2. $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ are the autoregressive and the moving average operators, respectively. Note that the binomial coefficient can be defined for real values of d, $\binom{d}{k} = \frac{d(d-1)(d-2)\ldots(d-k+1)}{k(k-1)(k-2)\ldots 1}$ for $k \in \mathbb{N}$ and an arbitrary d.

The row vector x_t contains the exogenous variables, and in our analysis, they are harmonic terms used to describe the seasonality in the radon time series. The parameter d in (7) describes the high-lag correlation structure of a time series while the p and q parameters are chosen to describe the low-lag correlation structure.

An important aspect to assess in a time series is its stationarity; a process is defined as stationary when its mean, variance, or autocorrelation structure remain constant over time. For stationary series, $d \in (-0.5, 0.5)$, and the Hurst exponent associated with the process is given by $H = (2d + 1)/2$. Consequently, long-range memory is present for $d \in$(0, 0.5), while $d \in$ (-0.5, 0) indicates anti-persistent fluctuations. When $|d| > 0.5$, the process is nonstationary and its variance is infinite. The process exhibits short memory for $d = 0$, corresponding to stationary and invertible ARMA (autoregressive moving average) model. Instead, the arbitrary restriction of d to integer values correspond to the standard autoregressive integrated moving average (ARIMA) model and in this case the variable is $I(d)$ and it becomes stationary after d differences and it is non-stationary after d-1 differences. For instance, an $I(1)$ variable can have a linear trend but no quadratic trend and it can be transformed into a stationary series with the first order differences. If a series exhibits long memory, it is neither stationary ($I(0)$) nor it is a unit root process ($I(1)$); it is an $I(d)$ process, with d a real number.

There are statistical tests to check stationarity, named unit root tests. The results are traditionally interpreted as that the effects of one-time shocks to the series are either transitory (if the series is stationary), or permanent (if the series is not stationary). The Dickey-Fuller test (DF) (Dickey and Fuller, 1979) tests the null hypothesis that the series is non-stationary, however DF only considers the dichotomy between stationarity and non-stationarity. The rejection of the null provides evidence for a stationary series, then the ARMA model can be directly applied. Instead, if the null hypothesis is accepted the series needs to be made stationary through differencing.

The model in (7) is estimated with exact maximum likelihood estimation explained in Veenstra (2013) using the *arfima* package of the *R* statistical software (Team, 2005).

Usually, the model selection is performed evaluating simultaneously the goodness of fit and the forecast

performances. The assessment of the goodness of fit can be done using the Akaike Information Criteria, $AIC = 2k - 2ln(L)$ where k is the number of estimated parameters in the model (7), and L is the maximised value of the likelihood function for the estimated model, and the model with the smallest AIC value is preferred. Moreover, the assumptions of the model on the random component (ϵ_t) are checked assessing the constant variability, the normality assumption, and the absence of correlation structure in the model residuals ($\hat{\epsilon}_t$).
The model family in (7) is fitted to time series data both to understand the data and to forecast (to predict future points in the series). Forecast evaluation can be done when the observed values are available. Usually, the observed data are divided into training and test samples. The model is fitted to the training sample and then its k-step ahead forecast performance is evaluated on the test one. The Root Mean Square Errors (RMSE) is used to check the forecast accuracy of the estimated models, it is given by $RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 n^{-1}}$ where y_i is the observed value for the i-th observation and $\tilde{y}_i$ is the predicted one.

3 RESULTS

3.1 Data

The analysed radon time series is recorded at Pietralunga (PTRL, Italy, lat 43.44N and long 12.44E) between 28/09/2012 and 01/08/2015 for a total of 1038 days. The PTRL station is in a framework of near real-time monitoring of soil radon emission to study earthquake preparatory processes, the Italian radon monitoring network (IRON) (Cannelli et al., 2018). The selected station is equipped with a Lucas cell, an alpha scintillation detector with an acquisition window of about 2 hours (115 minutes of data acquisition followed by a 5 minutes standby time). In detail, the Lucas cell consists in a flask which inner surface is coated with silver-activated zinc sulphide (ZnS). It integrates a front-end electronics and measures radon concentration by counting the radon decay signals in the given acquisition window. The radon detector is located in a small room of a school basement, not disturbed by anthropogenic influences and without any kind of opening and/or aeration system. However, the pressure and the temperature could affect the radon measures. The PTRL site is characterized by a contained seasonal variability also if compared with other sites of the same network, even equipped with a borehole probe which should be more immune from such effects (c.f. Fig. 2 in Siino et al. (2019b)). The radon concentration is measured in Bq/m^3, becquerel per cubic metre.
The raw time series of the mean daily concentrations is shown in Figure 1; the measured values range between 20.35 Bq/m^3 to 377.86 Bq/m^3 and they show a clear seasonal signal connected with the temperature (Cannelli et al., 2018; Siino et al., 2019b), and the higher values are during the summer period. The 0.87% of the daily data are missing. Generally, it is a challenge to handle missing values especially for time series data. Two possible ways to deal with the incomplete data can be omit the entire record that contains information or impute the missing values. However, since a small percentage of the analysed data presents missing values, they are filled by the weighted moving average method with a semi-adaptive window of 4 days. Weighted moving averages assign a linear weighting to the data points used to perform the imputation.
The data are divided into two subsets, the training set used to do the main analysis and to fit the stochastic models, and the test set (5% of the data identified by a vertical line in Figure 1) used to compare the models in terms of forecasts.

3.2 Seasonality, stationarity and long-memory detection

We report the main results about the seasonality detection, the autocorrelation, and the long-memory analysis of the observed series.

For the study of the dynamical and seasonal behaviours of the observed radon concentrations, we compute the spectral density analysis in the time-frequency domain based on the continuous wavelet transformation. The wavelet power spectrum is shown in Figure 2 where the period ranges from 16 to 512 days. The time-frequency regions with warm colours are characterized by high power, the black lines indicate the significant maxima of the undulations of the wavelet power spectrum, and they give an indication of the permanent cycle period. The thick black contour indicates the 90% confidence level and the lighter shade indicates regions inside the cone of influence due to the border effect.

Clearly, the series exhibits transient dynamics and the magnitude of the WPS is not constant over time fixing a specific frequency. We can observe a high value of the spectral power density at about 1-year periodicity that is persistently significant. Other high-power periodicities are present, even though not continuous over the entire period. A medium-power, $\sim$ 180-day cycle is recognizable in the first half of the series, and a $\sim$ 22-day cycle characterized by high-power appears around summer 2014. These cycles can be also observed in the Figure 2 which shows the global wavelet power spectrum; the horizontal lines provide a reference at 180, and 365 days. The series shows a clear 1-year periodicity and subordinate periodicities at about 180-days and three weeks. The longer cycles are probably related to the annual and semiannual cycles of the climatic variables (temperature, pressure, and rainfall), while the $\sim$ 22-day cycle is likely ascribable to the luni-solar gravitational influence which results in a tide-effect on the flux or radon (see Siino et al. (2019b)). This descriptive analysis is preparatory to decide which seasonality terms include in the model formulation for the explanatory variables (x_t in Equation 7). In particular, we consider harmonic terms to describe the 1-year periodicity that is the only one persistent with a constant power over time. Figure 2 shows the estimated curves considering a regression linear model fitted on the data with harmonic terms for 365-day period (with R^2 coefficient equals to 0.34).

The shapes of the correlogram and the partial correlogram provide indication about the properties of the time series and could indicate a plausible structure of the stochastic model in (7). In Figure 1, the estimated autocorrelation up to 400 lags seems to decay slower than an exponential one. Also from the autocorrelation, it is clear that the data exhibit a prevalent seasonal cycle which dominates the dependence structure. The partial autocorrelation coefficient is defined as the autocorrelation at each lag after controlling for the autocorrelation due to all preceding lags. It helps to determine how many AR terms (i.e., lagged observations as predictors) should be included in (7). If there is a sharp drop in the PACF after p lags, then the previous p-values are responsible for the autocorrelation in the series, and the model should include p autoregressive terms. In our case, the highest and also significant value is at lag 1, with a value of the correlation equals to 0.792, and in the following lags ($p > 1$) the autocorrelation coefficients are close to zero. It indicates that an autocorrelation term ($p = 1$) can be included in the model.

The estimate of the Hurst exponent (H) with the rescaled range analysis (section 2.3) is used to assess the presence of long memory. The obtained value is 0.785 indicating that the mean daily measurements have a persistent long-memory structure since $0.5 < H < 1$. In the literature, there are several results consistent with our analysis. For instance, in Cuculeanu et al. (1996) the determined values of Hurst's coefficient (0.809) highlight a persistent behaviour of the gas. Also, Nikolopoulos et al. (2018) compute the rescaled-range analysis for several time intervals obtaining a persistent Hurst exponent between 0.7-0.9 and in some periods, between 0.9-1.

The Dickey-Fuller test is used to check the null hypothesis that the series is non-stationary, and thus, the rejection of the null provides evidence for a stationary series. The value of the test on the data in Figure 1

Table 1. Estimates of ARFIMA models with different orders where the response variable is the average daily radon measurements in Figure 1. Model(a) is a fractional model without autoregressive and moving average terms. Model(b) is an autoregressive model (AR(1)), with d fixed to 0. Model(c) is a fractional autoregressive model. Model(d) is an integrated moving average model with order of integration equals to 1 (d=1). ϕ_1 is the estimate of the autoregressive coefficient in equation 7. β_1 and β_2 are associated to the harmonic terms ($\sin(2\pi t\omega)$ and $\cos(2\pi t\omega)$, where t is the time and $\omega = 1/365$) introduced in the model 7 as external variables x_t to describe the observed seasonality at 365-days. The log-likelihood, the Akaike Information Criterion and the range of the model residuals are shown. The root mean square errors (RMSE) for the rolling forecasts at 1-lag and 5-lag are reported. The significance of the estimates is in terms of p-value.

	Models			
	ARFIMA (p,d,q)			
	(0,d,0) **(a)**	(1,0,0) **(b)**	(1,d,0) **(c)**	(1,1,0) **(d)**
ϕ_1		0.687 *** (0.023)	0.347 *** (0.056)	-0.164 *** (0.031)
d	0.488 *** (0.014)		0.278 *** (0.045)	
β_1	-42.570 *** (11.336)	-40.101 *** (5.007)	-40.966 *** (7.293)	-52.456 (82.747)
β_2	34.138 *** (11.256)	30.380 *** (4.908)	31.942 *** (7.183)	33.731 (84.475)
Intercept	139.986 (118.533)	140.042 *** (3.519)	141.287 *** (11.828)	-0.046 (1.020)
Observations	985	985	985	985
Log Likelihood	-3490.32	-3484.35	-3467.65	-3550.49
AIC	6990.635	6978.698	6947.309	7110.988
σ^2	1192.12	1182.27	1143.44	1362.99
Range res.	[-127.690;204.301]	[-108.365;207.550]	[-104.312; 204.553]	[-135.772 ;214.456]
$RMSE_{1lag}$	44.214	44.626	44.511	49.921
$RMSE_{5lag}$	51.719	49.087	49.838	66.994
Note:				*p<0.1; **p<0.05; ***p<0.01

is -5.285 and the value of the p-value (1.53e-07) is lower than the significant level $\alpha = 0.05$ and we can reject the null hypothesis that the series has a unit root and hence is not stationary. According to this result, for our data, integration (first order differences) is not necessary.

3.3 Modelling results

The obtained results indicate that the studied radon concentrations present persistent long-memory structure, 1-year seasonality and an absence of a trend. Also, according to the PACF, an autoregressive term can be appropriate to describe the short-term correlation.
Starting from these evidences, four models are estimated and compared and all of them have the harmonic terms ($\sin(2\pi t\omega)$ and $\cos(2\pi t\omega)$) as external covariates to describe the seasonality. Four candidate models are estimated:

- Model(a) is a fractional model with $p = q = 0$, ARIMA$(0, d, 0)$

$$\nabla^d(y_t - x_t\beta) = \epsilon_t \quad t = 1, \ldots, T$$

- Model(b) is an ARMA$(1, 0)$, so it is an autoregressive model of order 1 without differencing (also it can be indicated as an ARIMA (1,0,0) model)

$$(1 - \phi_1 B)(y_t - x_t\beta) = \epsilon_t \quad t = 1, \ldots, T$$

- Model(c) is an ARFIMA$(1, d, 0)$ model

$$(1 - \phi_1 B)\nabla^d(y_t - x_t\beta) = \epsilon_t \quad t = 1, \ldots, T$$

- Model(d) is ARIMA$(1, 1, 0)$ model with order of integration equal to $d = 1$

$$(1 - \phi_1 B)\nabla^1(y_t - x_t\beta) = \epsilon_t \quad t = 1, \ldots, T$$

The Models (b), (c), and (d) have an autoregressive term ($p = 1$) since doing several comparisons a parsimonious model is obtained without moving average terms ($q = 0$). Only in the Models (a) and (c), the fractional-integration parameter is freely estimated. The estimated parameters, their standard errors, and significance are reported in Table 1. Also, additional information for each model such as the log-likelihood, the AIC, the range of the residuals, and the RMSE at 1-lag and 5-lag are shown.
The results of the estimate models suggest that for all of them the coefficients associated to the harmonic terms are significant, and the comparison with the models without the external covariates are worse (the results are not shown).
Examining the results of Model (a) and (c) where the parameter d is estimated varying in the real values, the fractional parameter for both models is between 0 and 1, thus allowing us to reject both the case of pure stationarity (I=0) and the unit root model (I=1). The estimated parameters are statistically significant at the 1% level, and lie within the interval (0, 0.5). The confidence intervals for the estimated fractional-integration parameters are relatively narrow and always in the positive range of persistent long-memory.
The results show that Model (c) is the best model in terms of fitting since it has the lowest AIC, and the shorter range of the residuals. For the RMSE at 1-lag and 5-lag forecast, Model (a) performs slightly better than Model (c), however, the fractional model has too simple parametrisation and it is not able to describe the autocorrelation dynamic in the data (see diagnostics on the residuals Figure 4 and Figure 5). The plots of observed and estimated values obtained with the four model at 1- and 5-lags are shown in Figure 6 and 7, respectively.
For all the estimated models, the assumption of constant variability along time appears respected (Figure 3) and there is not a marked pattern in the residuals, in particular the seasonality behaviour in the original data is not present (Figure 1). The residual ACF (Figure 4) and PACF (Figure 5) of the fitted Models (a), (b), and (d) show that there are significant estimated correlation coefficients at short lags, therefore these models are not adequate. Instead, for the Model (c) the residual ACF and PACF are not significant. The Ljung-Box test from 1 to 10 lags is computed to assess the absence of serial autocorrelation in the residuals. For the model (c), the null hypothesis is not rejected for all the considered lags. The q-q plot of the residuals is used to assess the normality assumption of the considered models. In Figure 8, for all the four models there is a slight departure from the normality in the tails.

4 DISCUSSION AND CONCLUSIONS

Being sensitive to crustal stress, the soil radon discharge is widely considered as a promising earthquake precursor. Because of the influence of several environmental factors and local geological conditions, pre-seismic radon anomalies cannot be easily detected with conventional statistical methodologies.

The general approach is to model the observation and highlight the anomalies. This article tests a radon concentration time series, covering almost 3 years, for the presence of non-stationarity (seasonality and trend) and long-memory. Overall, our results indicate that the radon series are better characterised as being stationary in the trend, but with persistent long-memory and 1-year seasonality. It is widely accepted that the periodic annual component in radon concentration time series is correlated to the climatic variables as temperature, atmospheric pressure, and rainfall (Siino et al., 2019a,b; D'Alessandro et al., 2020).
The class of ARFIMA model presented here provides a general framework for representing radon time series that display both short- and long-term persistence. The analysis of daily radon shows that the ARFIMA approach provides a better representation of the observed data with respect to the traditional ARMA and ARIMA models. More specifically, according to the model comparison, an ARFIMA model with an autoregressive term has a better fitting to the data. The estimated fractional-integration parameters of this ARFIMA model is positive and smaller than 0.5 ($d = 0.278$). It corresponds to a Hurst's coefficient of H=0.778 that is consistent with the results obtained with the rescaled range analysis and in general with other literature results (Cuculeanu et al., 1996; Nikolopoulos et al., 2018). The proposed model has been also assessed in terms of its predictive capacity, however the performances are quite poor especially increasing the lag of prediction (moving from 1-day forecast to 5-day forecast).
The occurrence of long-range correlation in the time series has been also tested by the application of Detrended Fluctuation Analysis (DFA) (Höll et al., 2019). Also, this method indicates that the radon concentration can be considered as coming from a fractional Gaussian noise (fGn).
It is widely recognised that radon time series are strongly controlled by the combination between site-specific factors and large-scale variations (i.e. astronomical cycles) (Schery et al., 1984; Schumann et al., 1988; Aumento, 2002; Piersanti et al., 2015; Crockett et al., 2018). It is note worth that the proposed approach (based only on radon measurements) is able to describe with good reliability the data and also to perform short-term forecasts when accurate radon measurements are taken for a reasonably long time span.
Model residuals could be retrospectively compared with external evidence of transitory phenomena in the study area (seismic, meteorological, etc.). Having available the seismic catalogue of the area, we make an attempt to find the relationship between the found anomalies in the radon time series and the earthquakes. In this case study, there is no evidence between the residuals of the fitted model and the seismicity in the study area. However, it should be considered the absence of any relevant earthquakes during the observation period, but only the occurrence of background seismicity. In fact, even though the well-known seismicity of the area, the larger recorded event was a $M_w = 3.9$ located at 9.5 km from the radon monitoring site.
In conclusion, our findings on the long-memory nature of radon measurements have important implications that can be useful for further analysis. The long-memory structure is the result of a long-lasting and aperiodic process such as a weather episode, changes in the circulation of geofluids, ground sealing, etc. However, at this stage (i.e. a single time series) is not possible to propose a comprehensive physical/geological or physical/meteorological mechanism that could account for the long-memory in radon concentration time series; moreover, it would also be out of the purpose of this work. The extension of this methodology by applying the ARFIMA models to longer radon time series, or to series with a different measurement range, or recorded in other monitoring sites, could provide the missing hints.
The proposed approach represents an effective tool to analyse radon signals, and in particular to detect long-range memory in the times series, which are the necessary preliminary steps to explore the relationship between radon anomalies and seismic activity. Finally, it would be interesting for further analysis, to compare the forecast of radon observations, or the identification of pre-seismic anomalies with those obtained with other methods such as the linear regression analysis (Stojanovska et al., 2017), the artificial neural network approach (Pasini and Ameli, 2003), or the decision tree method (Zhang et al., 2020). It

would also be interesting to consider other external covariates in the model formulation (7), such as weather variables (Stránský and Thinová, 2017).

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

A.D. and M.S. conceived the idea of the analysis approach. Data analyses and comments were made by M.S. M.S and SS prepared and reviewed the manuscript.

ACKNOWLEDGMENTS

We thank A. Piersanti e V. Cannelli (Istituto Nazionale di Geofisica e Vulcanologia) who provided the data.

DATA AVAILABILITY STATEMENT

The datasets for this study is available upon request to the authors.

REFERENCES

Aumento, F. (2002). Radon tides on an active volcanic island: Terceira, azores. *Geofísica Internacional* 41, 499–505

Barbosa, S., Donner, R., and Steinitz, G. (2015). Radon applications in geosciences – progress & perspectives. *The European Physical Journal Special Topics* 224, 597–603

Baskaran, M. (2016). *Radon: A tracer for geological, geophysical and geochemical studies* (Springer)

Baykut, S., Akgül, T., İnan, S., and Seyis, C. (2010). Observation and removal of daily quasi-periodic components in soil radon data. *Radiation Measurements* 45, 872–879

Belbute, J. M. and Pereira, A. M. (2017). Do global co2 emissions from fossil-fuel consumption exhibit long memory? a fractional-integration analysis. *Applied Economics* 49, 4055–4070

Beran, J. (2017). *Statistics for long-memory processes* (Routledge)

Bowers, M. C. and Tung, W.-w. (2018). Variability and confidence intervals for the mean of climate data with short-and long-range dependence. *Journal of Climate* 31, 6135–6156

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control* (John Wiley & Sons)

Cannelli, V., Piersanti, A., Galli, G., and Melini, D. (2018). Italian radon monitoring network (iron): A permanent network for near real-time monitoring of soil radon emission in italy. *Annals of Geophysics* 61, 444

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review* 51, 661–703

Conraria, L. A. and Soares, M. J. (2011). The continuous wavelet transform: A primer. *NIPE Working Paper* 16, 1–43

Crockett, R. G., Gillmore, G. K., Phillips, P. S., Denman, A. R., and Groves-Kirkby, C. J. (2006). Tidal synchronicity of built-environment radon levels in the uk. *Geophysical Research Letters* 33

Crockett, R. G., Groves-Kirkby, C. J., Denman, A. R., and Phillips, P. S. (2018). Significant annual and sub-annual cycles in indoor radon concentrations: seasonal variation and correction. *Geological Society, London, Special Publications* 451, 35–47

389 Cuculeanu, V., Lupu, A., and Sütö, E. (1996). Fractal dimensions of the outdoor radon isotopes time series.
390 *Environment International* 22, 171–179
391 D'Alessandro, A., Scudero, S., Siino, M., Alessandro, G., and Mineo, R. (2020). Long-term monitoring
392 and characterization of soil radon emission in a seismically active area. *Geochemistry, Geophysics,*
393 *Geosystems* , e2020GC009061doi:https://doi.org/10.1029/2020GC009061
394 Daubechies, I. (1992). *Ten lectures on wavelets*, vol. 61 (Siam)
395 Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a
396 unit root. *Journal of the American statistical association* 74, 427–431
397 Donner, R. V., Potirakis, S. M., Barbosa, S. M., Matos, J. A., Pereira, A. J., and Neves, L. J. (2015).
398 Intrinsic vs. spurious long-range memory in high-frequency records of environmental radioactivity. *The*
399 *European Physical Journal Special Topics* 224, 741–762
400 Dunn, J. and Henschel, B. (1989). Statistical aspects of autoregressive-moving average models in the
401 assessment of radon mitigation. *Environment International* 15, 247–252
402 Höll, M., Kiyono, K., and Kantz, H. (2019). Theoretical foundation of detrending methods for fluctuation
403 analysis such as detrended fluctuation analysis and detrending moving average. *Physical Review E* 99,
404 033305
405 Hosking, J. R. M. (1981). Fractional differencing. *Biometrika* 68, 165–176. doi:10.1093/biomet/68.1.165
406 Hurst, H. (1950). Long-term storage capacity of reservoirs. *American Society of Civil Engineering* 76
407 Hurst, H., Black, R., and Simaika, Y. (1965). Long-term storage: an experimental study constable. *London*
408 *UK*
409 Mandelbrot, B. B. and Wallis, J. R. (1968). Noah, joseph, and operational hydrology. *Water resources*
410 *research* 4, 909–918
411 Mandelbrot, B. B. and Wallis, J. R. (1969). Robustness of the rescaled range r/s in the measurement of
412 noncyclic long run statistical dependence. *Water Resources Research* 5, 967–988
413 Montanari, A., Rosso, R., and Taqqu, M. S. (1997). Fractionally differenced arima models applied
414 to hydrologic time series: Identification, estimation, and simulation. *Water resources research* 33,
415 1035–1044
416 Morales-Simfors, N., Wyss, R. A., and Bundschuh, J. (2019). Recent progress in radon-based monitoring
417 as seismic and volcanic precursor: A critical review. *Critical Reviews in Environmental Science and*
418 *Technology* , 1–34
419 Nikolopoulos, D., Matsoukas, C., Yannakopoulos, P., Petraki, E., Cantzos, D., and Nomicos, C. (2018).
420 Long-memory and fractal trends in variations of environmental radon in soil: Results from measurements
421 in lesvos island in greece. *J. Earth Sci. Clim. Chang* 9, 1–11
422 Pan, J.-N. and Chen, S.-T. (2008). Monitoring long-memory air quality data using arfima model.
423 *Environmetrics: The official journal of the International Environmetrics Society* 19, 209–219
424 Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D. (2018a). One-step ahead forecasting of
425 geophysical processes within a purely statistical framework. *Geoscience Letters* 5, 12
426 Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D. (2018b). Predictability of monthly temperature
427 and precipitation using automatic time series forecasting methods. *Acta Geophysica* 66, 807–831
428 Pasini, A. and Ameli, F. (2003). Radon short range forecasting through time series preprocessing and
429 neural network modeling. *Geophysical Research Letters* 30
430 Piersanti, A., Cannelli, V., and Galli, G. (2015). Long term continuous radon monitoring in a seismically
431 active area. *Annals of Geophysics* 58, 0437
432 Pinault, J.-L. and Baubron, J.-C. (1996). Signal processing of soil gas radon, atmospheric pressure,
433 moisture, and soil temperature data: a new approach for radon concentration modeling. *Journal of*

Geophysical Research: Solid Earth 101, 3157–3171
Reisen, V. A., Monte, E. Z., da Conceição Franco, G., Sgrancio, A. M., Molinares, F. A. F., Bondon, P., et al. (2018). Robust estimation of fractional seasonal processes: Modeling and forecasting daily average so2 concentrations. *Mathematics and Computers in Simulation* 146, 27–43
Roesch, A. and Schmidbauer, H. (2018). *WaveletComp: Computational Wavelet Analysis*. R package version 1.1
Schery, S., Gaeddert, D., and Wilkening, M. (1984). Factors affecting exhalation of radon from a gravelly sandy loam. *Journal of Geophysical Research: Atmospheres* 89, 7299–7309
Schumann, R., Owen, D., and Asher-Bolinder, S. (1988). Factors affecting soil-gas radon concentrations at a single site in the semiarid western us. In *Proc. of the 1988 EPA Symposium on Radon and Radon Reduction Technology 2. Publication* (Citeseer)
Shumway, R. H. and Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples* (Springer)
Siino, M., Alessandro, G., Buonmestieri, S., D'Alessandro, A., Mineo, R., and Scudero, S. (2019a). Radon concentration in the ragusa province (sicily, italy). In *International Scientific Conference Man and Karst 2019*
Siino, M., Scudero, S., Cannelli, V., Piersanti, A., and D'Alessandro, A. (2019b). Multiple seasonality in soil radon time series. *Scientific reports* 9, 8610
Stojanovska, Z., Ivanova, K., Bossew, P., Boev, B., Zora, S., KOLAR, P., et al. (2017). Prediction of long-term in door radon concentration based on short-term measurements. *Nucl Technol Radiat* 32
Stránský, V. and Thinová, L. (2017). Radon concentration time series modelling and application discussion. *Radiation Protection Dosimetry* 177, 155–159
Team, R. D. C. (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria
Toutain, J.-P. and Baubron, J.-C. (1999). Gas geochemistry and seismotectonics: a review. *Tectonophysics* 304, 1–27
Udovičić, V., Filipović, J., Dragić, A., Banjanac, R., Joković, D., Maletić, D., et al. (2014). Daily and seasonal radon variability in the underground low-background laboratory in belgrade, serbia. *Radiation protection dosimetry* 160, 62–64
Veenstra, J. Q. (2013). *Persistence and Anti-persistence: Theory and Software*. Ph.D. thesis, The University of Western Ontario
Wang, W., Van Gelder, P., Vrijling, J., and Chen, X. (2007). Detecting long-memory: Monte carlo simulations and application to daily streamflow processes. *Hydrology and Earth System Sciences Discussions* 11, 851–862
Woith, H. (2015). Radon earthquake precursor: A short review. *The European Physical Journal Special Topics* 224, 611–627
Yan, R., Woith, H., Wang, R., and Wang, G. (2017). Decadal radon cycles in a hot spring. *Scientific reports* 7, 12120
Yaya, O. S. and Fashae, O. A. (2015). Seasonal fractional integrated time series models for rainfall data in nigeria. *Theoretical and Applied Climatology* 120, 99–108
Zhang, S., Shi, Z., Wang, G., Yan, R., and Zhang, Z. (2020). Groundwater radon precursor anomalies identification by decision tree method. *Applied Geochemistry* , 104696

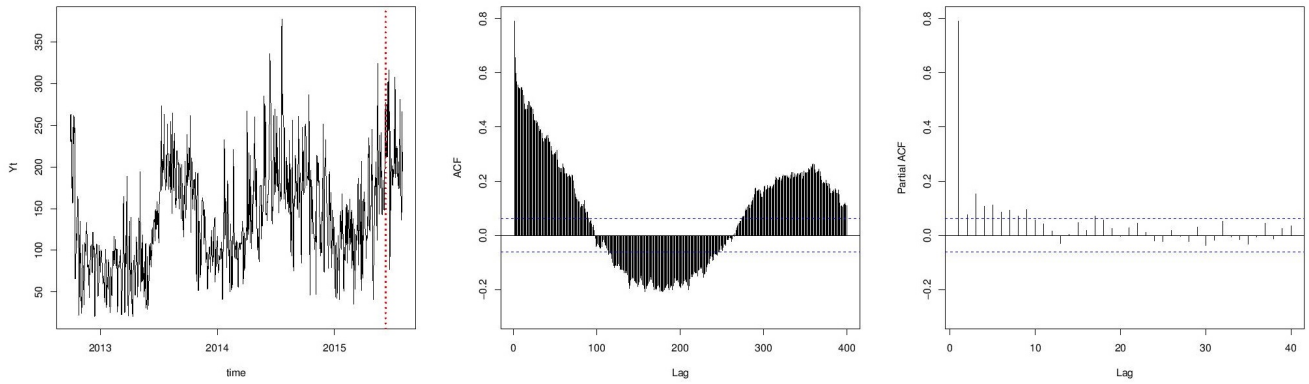

Figure 1. (Left) Data, mean daily radon observations in Bq/m^3 at Pietralunga (Umbria, Italy). The red vertical line separates the training set and the remaining 5% of the test set. (Central) Autocorrelation coefficient and (Right) partial autocorrelation coefficient of the time series.

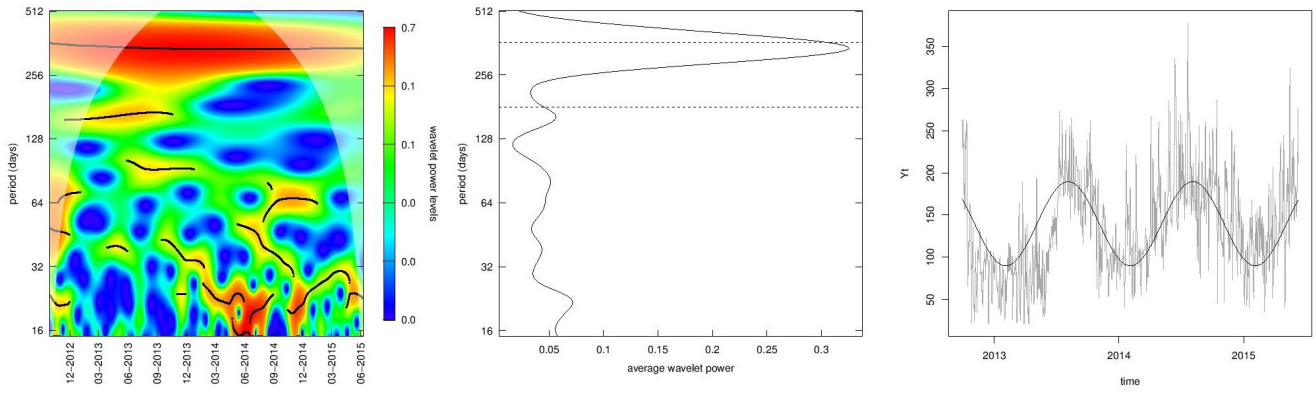

Figure 2. (Left) Wavelet power spectrum of daily radon series in time-frequency domain with the CWT method. The black contour indicates the significant period with 90% confidence level. The lighter shade is the regions influenced by edge effects. (Central) The corresponding global power spectrum density marginalising over time. The horizontal lines are for 180-day and 365-day periods. (Right) The grey line is the observed time series and the black curve is the fitted linear regression model with respect to harmonic terms to describe the 1-year periodicity.

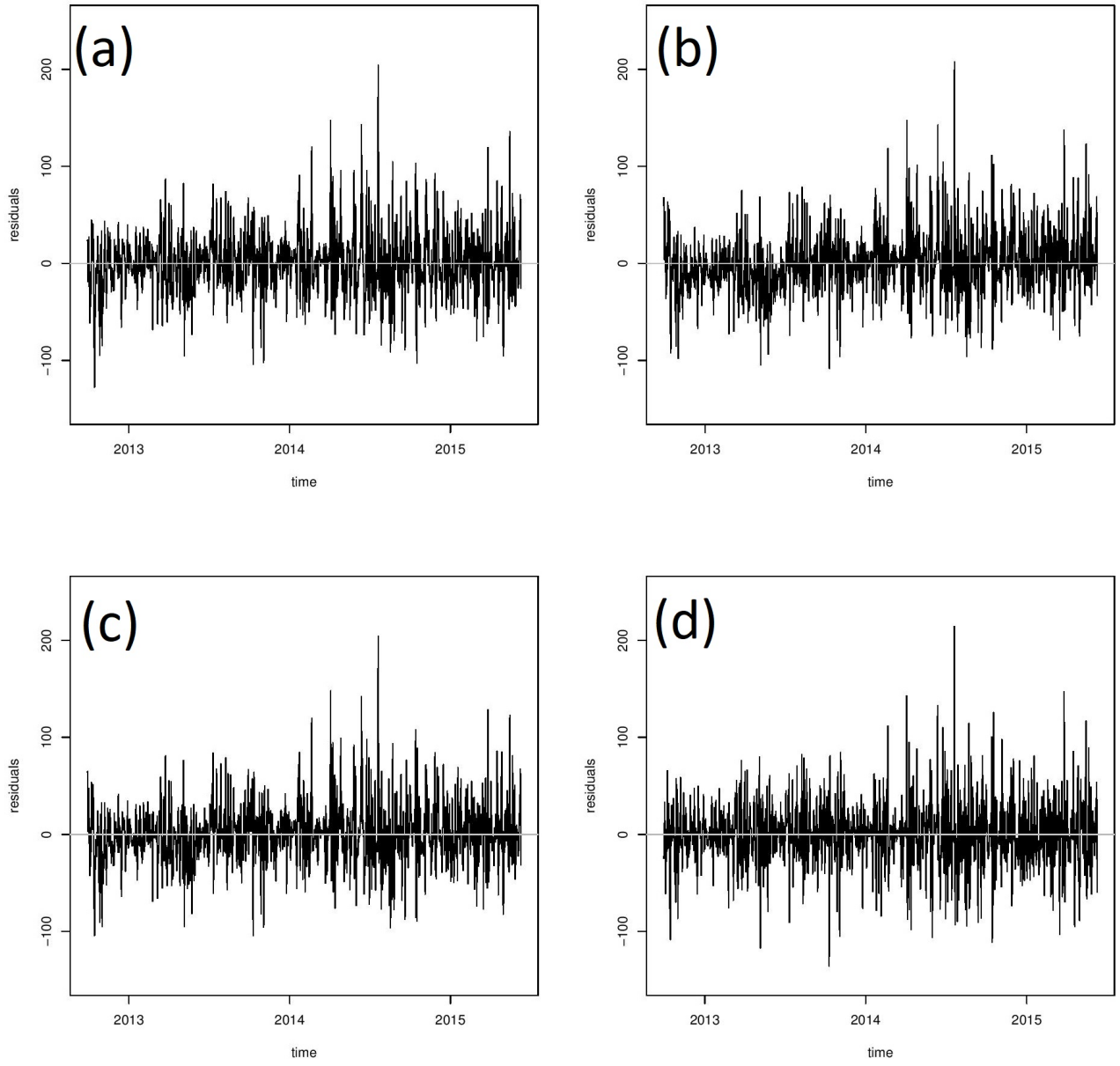

Figure 3. The residuals time series for the estimated Models (a), (b), (c), and (d) as labelled in Table 1.

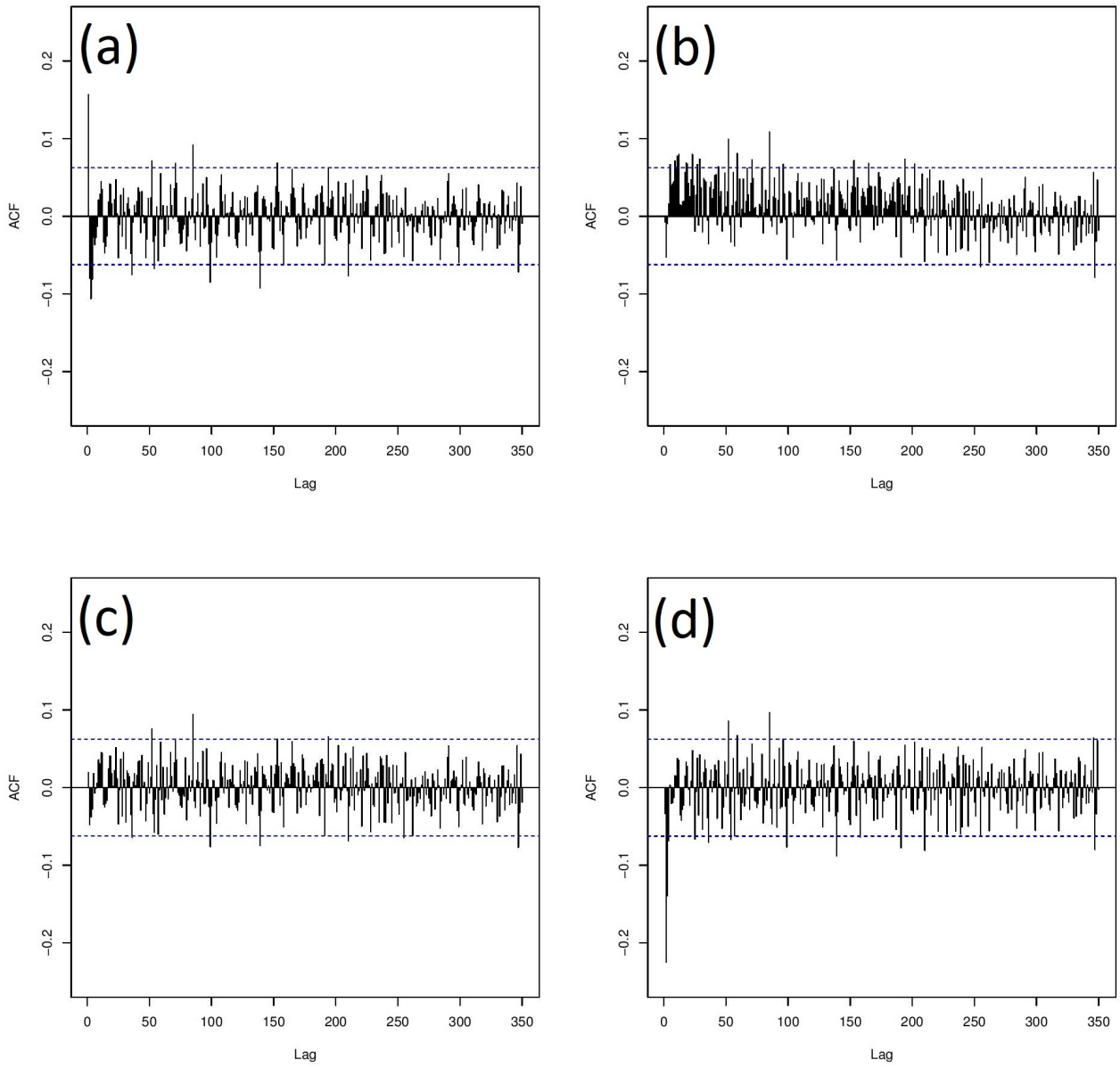

Figure 4. Autocorrelation coefficients of the residuals of the estimated Models (a), (b), (c), and (d) as labelled in Table 1. The horizontal lines indicate the confidence interval at 95% for not significant correlation coefficient.

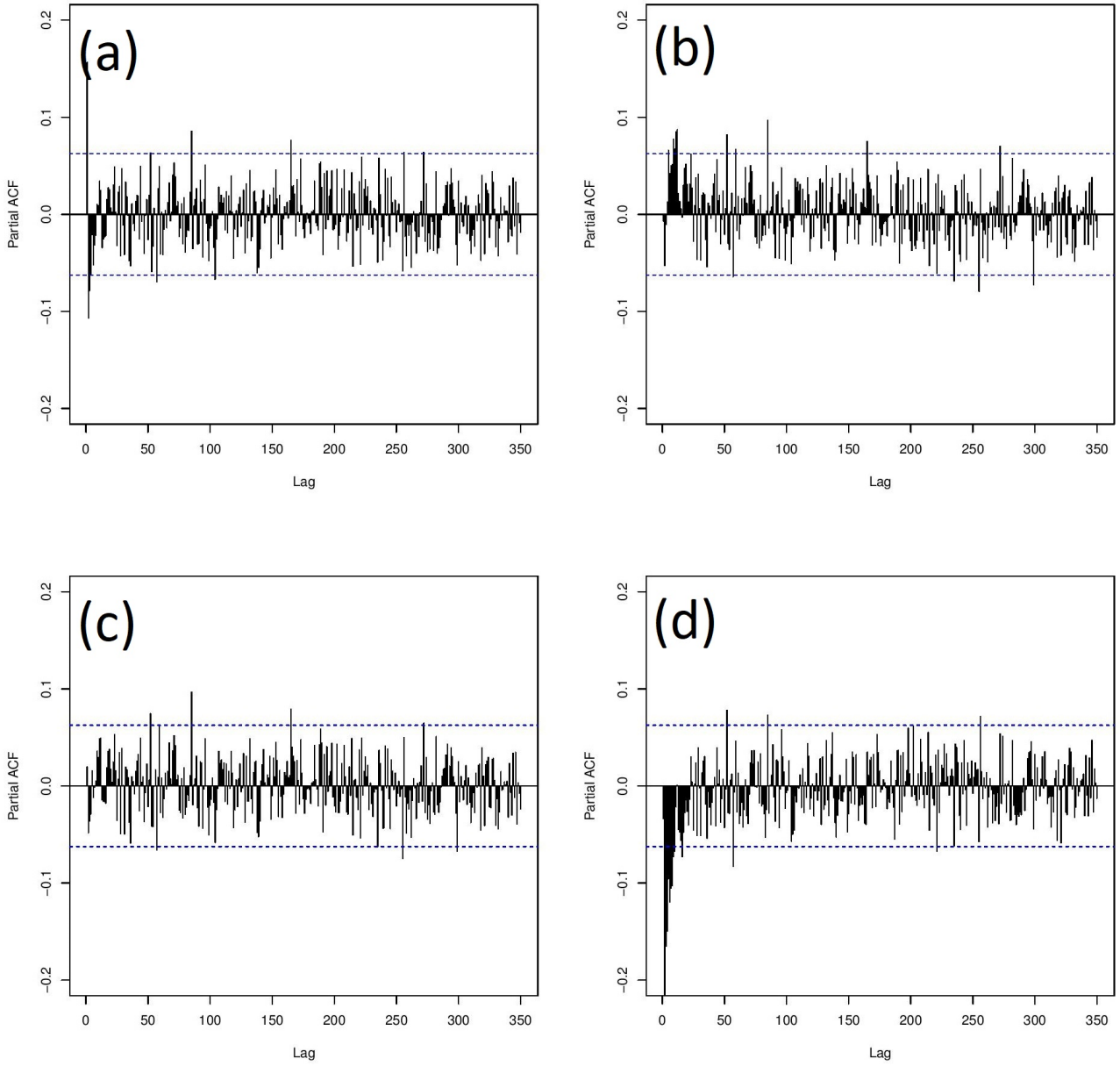

Figure 5. Partial autocorrelation coefficients of the estimated Models (a), (b), (c) and (d) as labelled in Table 1. The horizontal lines indicate the 95% confidence bounds for strict white noise.

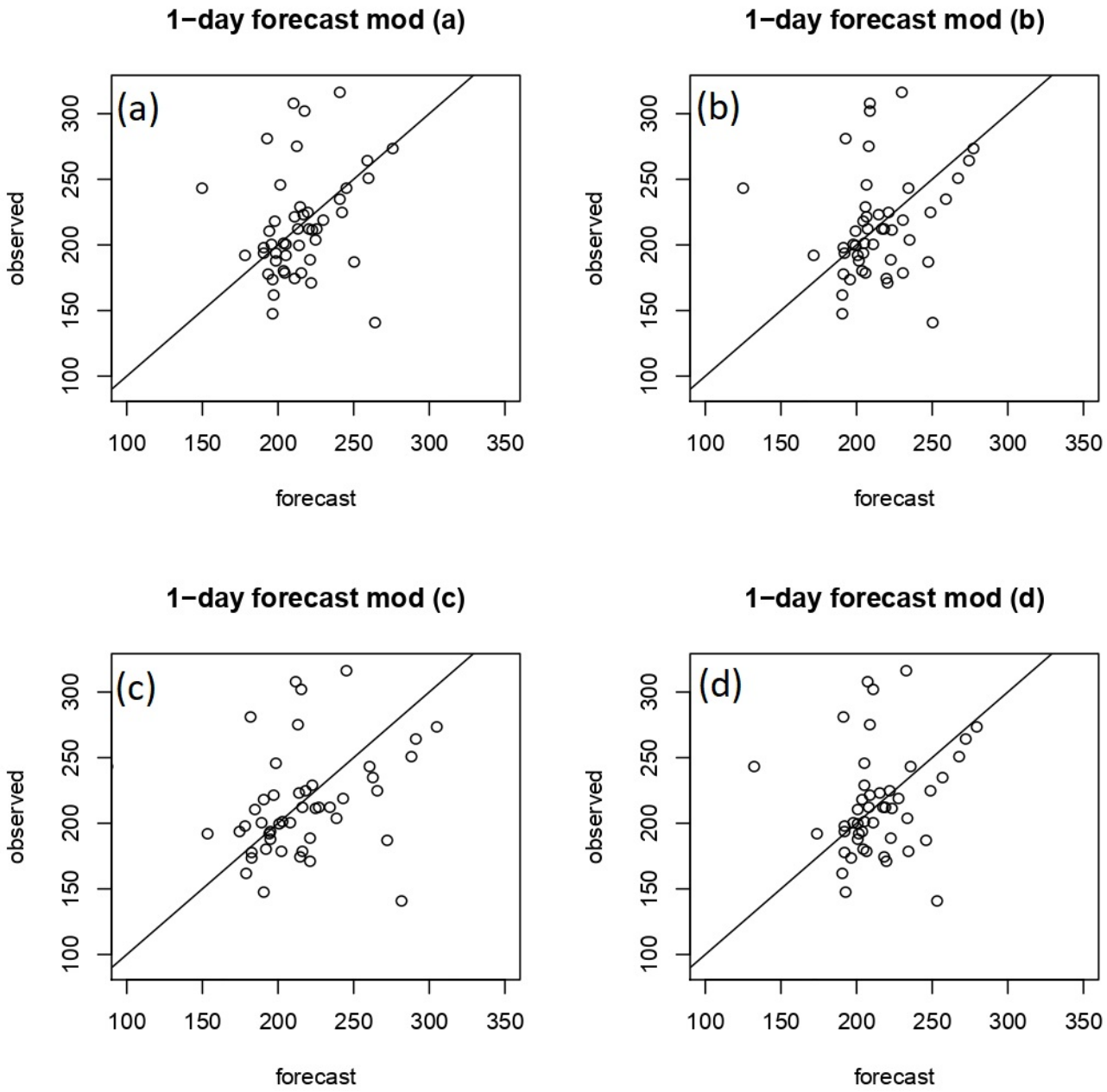

Figure 6. Prediction at 1-lag and observed radon measurements for the estimated Models (a), (b), (c) and (d) as labelled in Table 1.

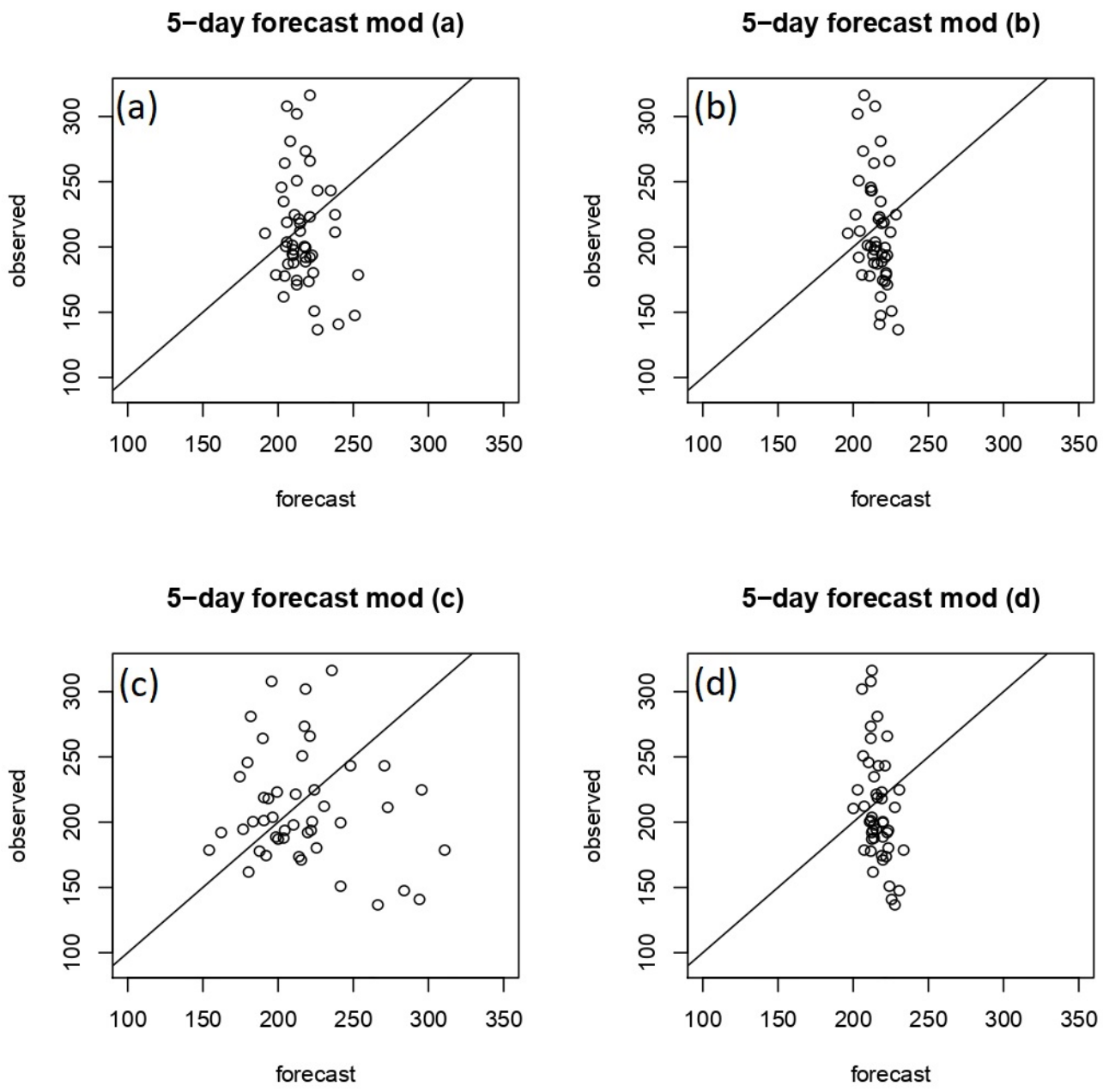

Figure 7. Prediction at 5-lag and observed radon measurements for the estimated Models (a), (b), (c) and (d) as labelled in Table 1.

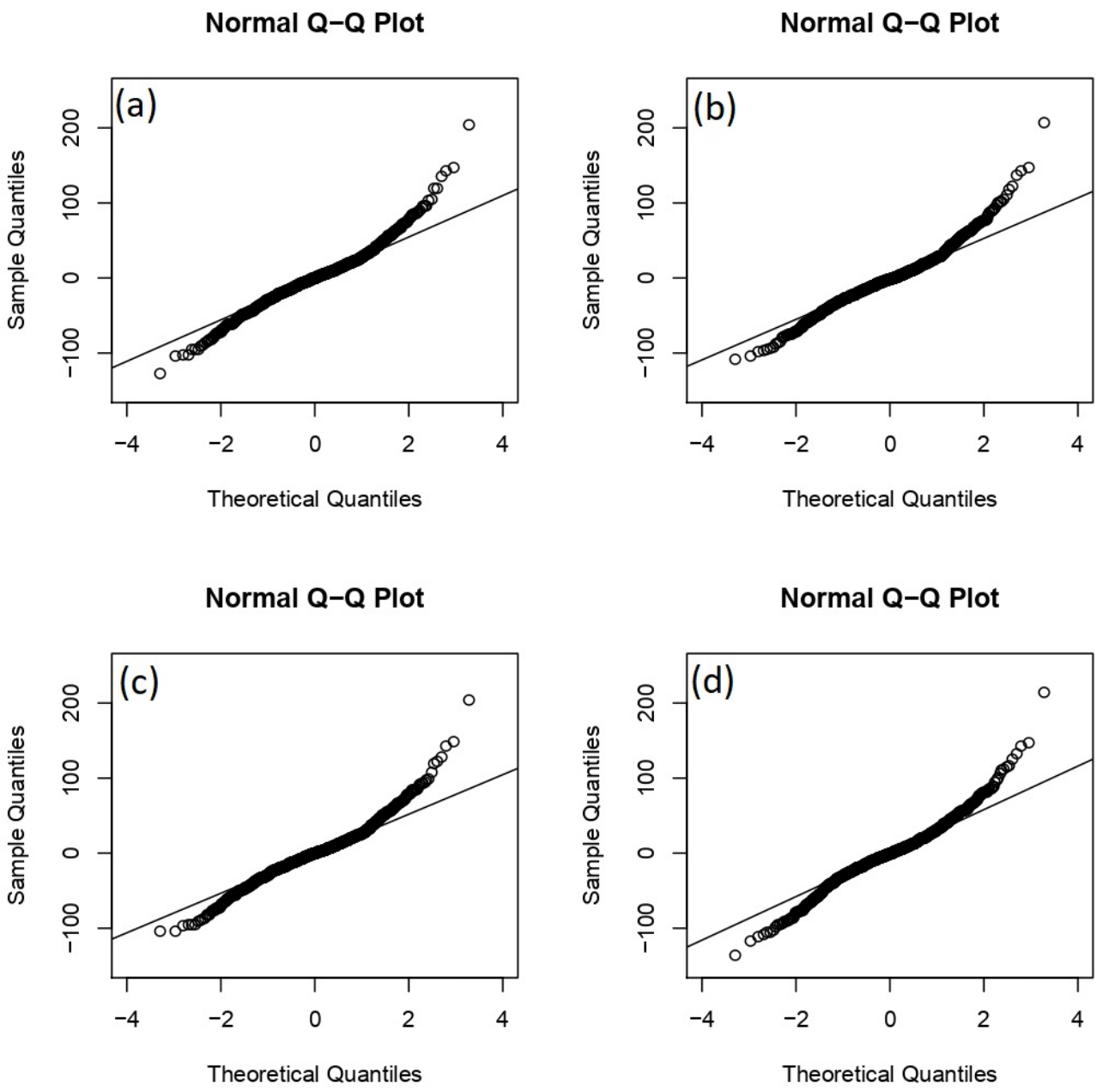

Figure 8. Q-Q norm of the residuals time series for the estimated Models (a), (b), (c) and (d) as labelled in Table 1 to check the normality assumption.