



Weather control in radon flux time series from Schleswig-Holstein, Germany

Manfred Mudelsee^{1,2} · Johannes Albert³ · Frank Sirocko³

Received: 7 November 2019 / Accepted: 10 June 2020

© The Author(s) 2020

Abstract

Indoor radon exposure is a serious hazard to human health. The radon concentration in surface air varies spatially as a result of the uranium content in the underlying rocks. However, there exist also considerable knowledge gaps about temporal variations. Here we document the high temporal variability of radon flux from exhalation in high-resolution (hourly) time series from a site near Kleinneudorf, Schleswig-Holstein, Germany. By means of advanced techniques of statistical time series analysis, we show a close association between radon and meteorological variables (air temperature and air pressure). We identify four principal weather regimes that lead to different radon exhalation modes. For each of the modes, we construct a statistical linear model for radon prediction via the meteorological variables and their derivatives or time-lagged versions. The model explains between 53 and 86 percent of the variance. Many model deviations consist in excessively high measured radon values and hint at nonlinear effects. Other model deviations hint at non-meteorological forcing.

Keywords Radon · Time series analysis · Cross-spectral analysis · Kernel trend and derivative estimation · Multiple linear regression · Prediction · Air temperature · Air pressure · Human health

Mathematics Subject Classification 62M10 · 62M15 · 62M20 · 62P12

This work received funding by the German Federal Ministry for Economic Affairs and Energy via the ANGUS II Project (<https://www.angus2.de/en>), Grant Number 03ET6122C.

✉ Manfred Mudelsee
mudelsee@climate-risk-analysis.com

¹ Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, 27570 Bremerhaven, Germany

² Climate Risk Analysis, Heckenbeck, 37581 Bad Gandersheim, Germany

³ Institute of Geosciences, Johannes Gutenberg University Mainz, 55128 Mainz, Germany

1 Introduction

The noble gas radon (Rn) has only radioactive isotopes, of which ^{222}Rn has the longest half-life (3.8 days, alpha decay). ^{222}Rn is part of the decay chain of the uranium isotope ^{238}U .

Inside the human body, the alpha radiation produced by ^{222}Rn and its decay products may damage body cells. Hence, ^{222}Rn is considered as one of the main causes of lung cancer (Yoon et al. 2016). Accordingly, the EU has launched an initiative to map Rn concentrations in natural soils and rocks (Elío et al. 2019). The purpose is to determine Rn risk areas, where newly built houses have to be protected against infiltration from the subsurface strata into the basement. The problem is particularly pressing for old houses without a baseplate, which were usually built directly on bare soil. If additionally a strong thermal insulation is installed, then the indoor Rn concentration for such houses can reach extremely high values, above $10,000\text{ Bq/m}^3$ (Huber et al. 2001), a value otherwise observed only in mines. The WHO (Zeeb and Shannoun 2009, p. xi therein) proposed a “reference level of 100 Bq/m^3 to minimize health hazards due to indoor radon exposure” and, “if this level cannot be reached under the prevailing country-specific conditions, the chosen reference level should not exceed 300 Bq/m^3 .” See also the EU’s Council Directive 2013/59/Euratom.

As a result of these policies, the natural Rn concentrations in soils and rocks have to be analyzed. Pioneering work on this for Germany has been performed by Kemski et al. (2001, 2005, 2009). For Germany, a map (Fig. 1b) was constructed on basis of measurements done between 1992 and 2003, which is also available from the website of the Bundesamt für Strahlenschutz (<https://www.bfs.de/DE/themen/ion/umwelt/radon/karten/boden.html>, last access 15 October 2019). These maps show large regions of increased Rn content in the southern parts of Germany and also in all other regions where granite or metal ores (i.e., uranium) are found (Fig. 1b).

The noble gas Rn can be dissolved in groundwater or the open pore space of soils and subsurface strata. The rate of production and venting from the water will usually be in steady state, and accordingly the Rn content in the pore water (groundwater) will be mostly constant over time. The Rn flux from the pore water, however, may change when the water warms. This can mean up to several degrees between winter and summer at a given location. In addition to this slow groundwater-related process, the Rn can be released from the open pore space (often the upper metres of strata) to the surface by thermal convection in the sediments, in particular along faults and fractures. On the relation between Rn release from the groundwater and temperature, see Akawwi (2014). On the influence of thermal convection in sediments, see Mogro-Campero et al. (1980) and Burkhart and Huber (1993).

Rn concentrations in the Quaternary strata of Northern Germany are generally low or intermediate. This is the case, for example, in the region of moraines from the last ice age in Schleswig-Holstein. However, local maxima of Rn have been observed intermittently at practically all places in Northern Germany. Such local maxima are thought to be mainly caused by locally high concentration of Rn-bearing rocks (e.g. granites, in particular those with abundant K-feldspar); see, for example, Kemski et al. (2001).

The data for the mapping of the Rn risk areas are measured in distinct time windows. The “normal” exposition time of a commercial exposimeter (the cheapest and most effective Rn measurement device) is in the order of a month. However, this exposure duration appears to be one of the main obstacles towards a consistent map because, as long time series reveal, the seasonal structure of Rn concentration is rather complex. For example, at our measurement site (Sect. 2), the monthly average Rn concentrations in February 2019 is 6183 Bq/m^3 , while that in June 2019 is just 1343 Bq/m^3 .

The availability of high-resolution time series of soil Rn concentrations has increased over the past few years. This allows to study space-time variability for critical geographical regions. For example, Moreno et al. (2016) study the Amer fault zone in Spain at a resolution of one day (and higher) for a 4-year time span; Siino et al. (2019) examine nine Rn stations from the Italian Radon Monitoring Network at a resolution of 2 h (and higher) for time intervals between 1 and 5 years; and Tareen et al. (2019) analyse a series from Muzaffarabad, a town sitting on top of a fault zone in Iran, at 40-min resolution (and downscaled) for a time span of 1 year. What distinguishes the present study is the fact that our measurements directly reflect the outgassing Rn (not the Rn in the soil), which makes it relevant from a health perspective.

The regular structure and the superimposed variations in the time series (Sect. 2) hint at a meteorological (co-)explanation of the causes of the temporal variability. Therefore, we measured surface-air temperature and surface-air pressure in the same device as the Rn concentrations. This allows us to compare selected time windows from the series by means of statistical methods (Sect. 3). The analytical results (Sect. 4) feed into the discussion (Sect. 5) of the separation of the weather regimes (i.e., Rn exhalation modes) with the purpose of predicting the Rn concentrations over time.

This paper is about ongoing work within the project ANGUS II. The aims of this paper are (1) to explore the above mentioned exhalation modes by means of an advanced high-resolution, direct measurement technique and (2) to compare the different modes by means of statistical time series analysis. The obtained results are the basis for a successful prediction of the Rn exhalation. There are two companion papers to be published elsewhere (Albert et al., manuscript in preparation; Sirocko et al., manuscript in preparation), to which we briefly refer to in the text.

2 Data

We have monitored the Rn flux above a well, which was drilled into last glacial meltwater sand and till down to a depth of 40 m. The location is near to the village of Kleinneudorf, east of lake Plön, Schleswig-Holstein, Germany (Fig. 1).

The drilling at Kleinneudorf is located in a small depression of active subsidence and highly permeable Quaternary sediments in the uppermost 40 m, which facilitates high permeability and causes a local “hotspot” of Rn flux (Albert et al., manuscript in preparation). To understand the role of this specific geological process at Kleinneudorf, we developed a fluxbox measurement system directly in the subsidence center. The fluxbox presents a novel approach to measure continuously Rn concentrations in a chamber, which allows a controlled gas flux from the soil to the air in a system of two boxes (Fig. 2).

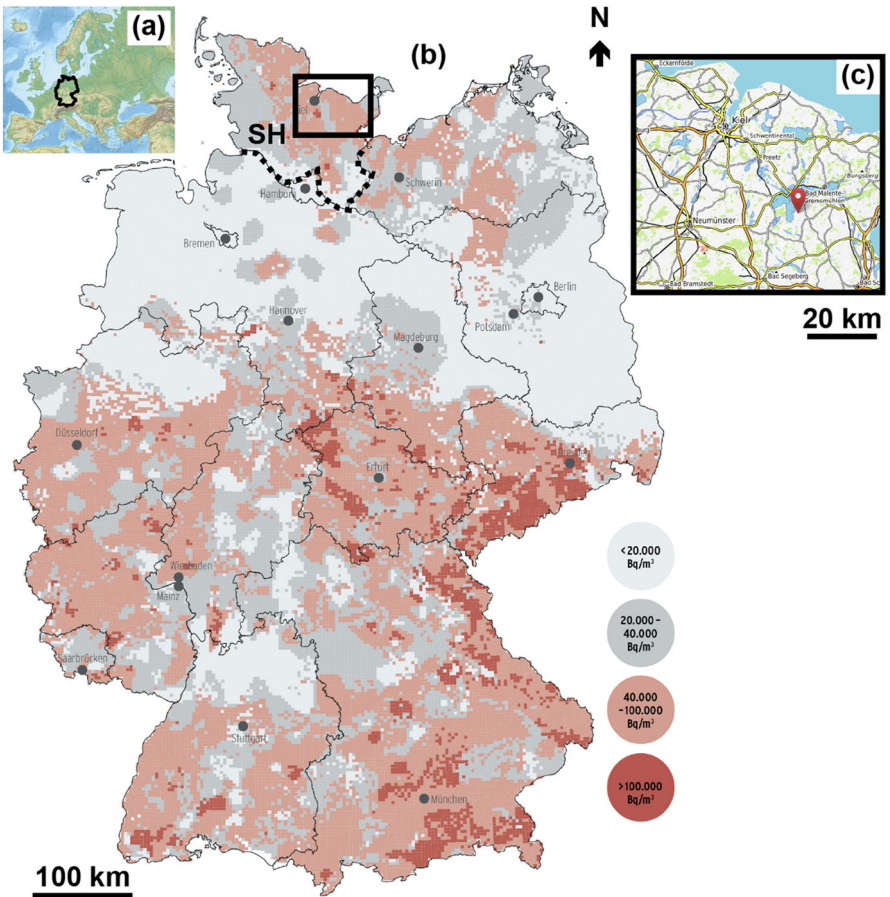


Fig. 1 Study site and Rn map. **a** Map of Europe, with Germany indicated; **b** Rn map for Germany with concentration levels (shaded) and indicated study area (inset box); also indicated is the Bundesland Schleswig-Holstein (SH, north of the dashed boundary); **c** study area, with study site (red marker) indicated. Modified and reproduced with permission (**a** Wikipedia; **b** Bundesamt für Strahlenschutz (German Federal Office for Radiation Protection); **c** OpenStreetMap) (color figure online)

The schematic drill section (Fig. 2) shows that the groundwater level is at 8 m. The surface of the well is open, but covered with two boxes, which protect against rain, direct sun and wind. The casing of the well is solid plastic, so that the inner space of the casing is filled with gas exhaled only from the groundwater. The gas exhaled from the soil is trapped in the outer box, intrudes into the inner box, where it is shielded from direct venting by wind, which is a serious methodological problem in studies of Rn in houses, tunnels or direct air measurements.

We measured the Rn concentration in the inner box next to the bore hole. The two boxes are closed, but allow slow convection to the outside. Thus, our measurements reflect the Rn exhalation and not the soil Rn concentration. That means, the absolute values of Rn concentration are a function of this setting and cannot be used directly

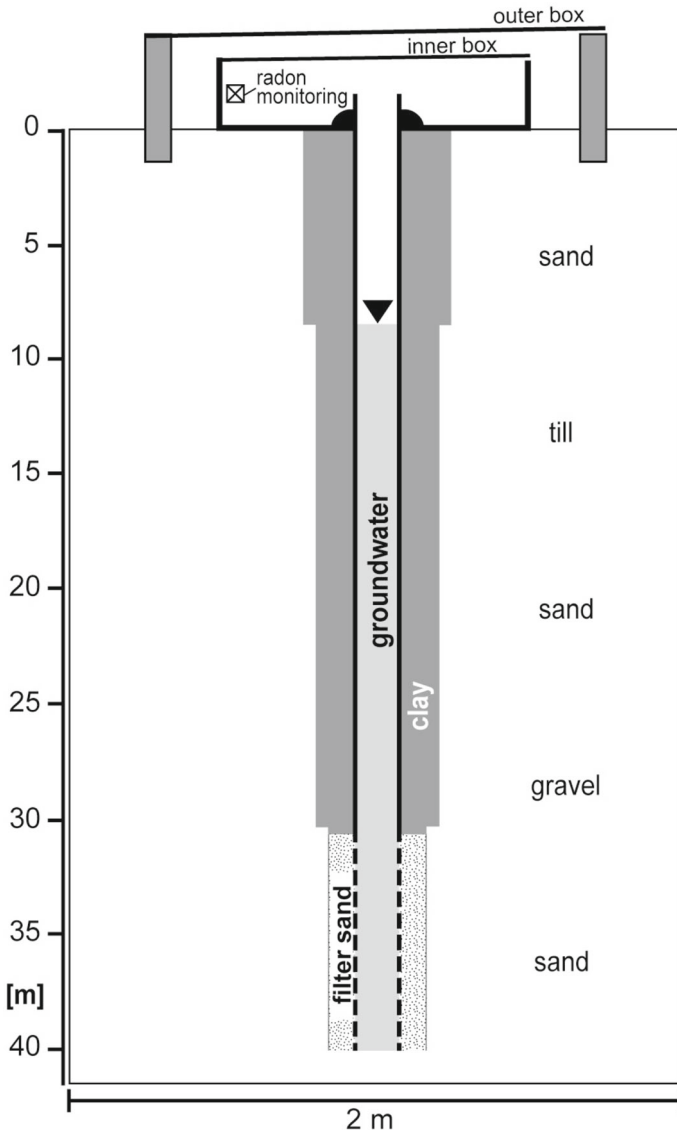


Fig. 2 Schematic profile of well. Also shown are the two boxes for the detection of the exhalation flux

to compare to soil concentrations (as used for the mapping of the Rn risk areas). Accordingly, the absolute values in this study are arbitrary, and the variability structure of the time series (relative changes) is the main source of information. It was mandatory to keep the boundary conditions constant during the entire time of the experiment, which was run from April 2018 to September 2019.

The Rn measurements are done with a Canary Pro monitor (manufactured in 2017 by Corentium) (Radon Analytics 2020), modified for scientific use. The measurement

principle of the Canary device is based on a Si photodiode. When ^{222}Rn and its progeny isotopes ^{218}Po and ^{214}Po decay within the instrument's dome, a few of the released alpha particles hit the open photodiode and the event, where the energy is released during the impact, is counted. This allows the number of decays to be measured. This measuring principle can be used to detect both Rn and its daughter isotopes. However, only Rn can enter the measuring chamber, because the daughters are electrically charged and get caught in the filters of the device. The modified Canary monitor has an application range from zero to 100000 Bq/m^3 . The efficiency for Rn sampling specified by the manufacturer is one count per hour at a concentration of 33 Bq/m^3 . Additionally to hourly Rn sampling, the device can record temperature, air pressure and humidity. However, due to repeated technical failure of the humidity sensor, this variable was not considered in this survey.

The full series for the variables Rn concentration, air temperature and air pressure are shown in Fig. 3. The nominal time spacing is $d(i) = t(i) - t(i - 1) = d = 1$ hour, however, there exist two major gaps for Rn and a few other, minor gaps.

The time series (Fig. 3) shows pronounced Rn maxima for June 2018 and January 2019, that means, during rather different seasons. The processes controlling the Rn flux are best visible in shorter (about one week) series. Four different modes can be deciphered. Mode II (Fig. 4) is characterized by a Rn maximum during the night. Mode III (Fig. 5) exhibits two maxima, one during the night and one in the afternoon. Mode IV (Fig. 6) has the maximum in the late afternoon. Mode I (Fig. 7) has no relation to temperature but to decreasing surface air pressure. Apparently, surface air temperature and surface air pressure are the main control variables for the Rn flux in our monitoring system.

Hence, we selected four time intervals (Table 1) as representatives of the different weather regimes (i.e., Rn exhalation modes)

3 Methods of data analysis

The general aim of climate data analysis is to make an inference about the data generating system (i.e., the climate) on basis of a set of data values. The data for the Rn project are in the form of time series. Let $t(i)$ denote the time value and $x(i)$ denote the value of a climate parameter (e.g., Rn concentration). The time values may in principle be unevenly spaced, although here in the project the case is even spacing (Table 1). Let n denote the sample size. The compact notation for a univariate time series is $\{t(i), x(i)\}_{i=1}^n$. Even spacing means that $t(i) - t(i - 1) = d(i) = d = \text{constant}$ for $i = 2, \dots, n$. For multivariate time series, we have additional time series, and we write either $x(i), y(i)$ (bivariate) or $x_1(i), x_2(i), \dots, x_p(i)$ (p -dimensional). The inference consists (1) in the estimation of values of parameters of a statistical model for the climate and (2) in the testing of hypotheses about the climate. The models relevant for the project (cross-spectra, trend and regression) are detailed in the following subsections.

Since $n < \infty$ and the data are affected by measurement noise (Sect. 2), the inference shows uncertainties. For parameter estimation, the uncertainties are expressed as error bars, confidence intervals, and so forth. For hypothesis testing, the uncertainty is

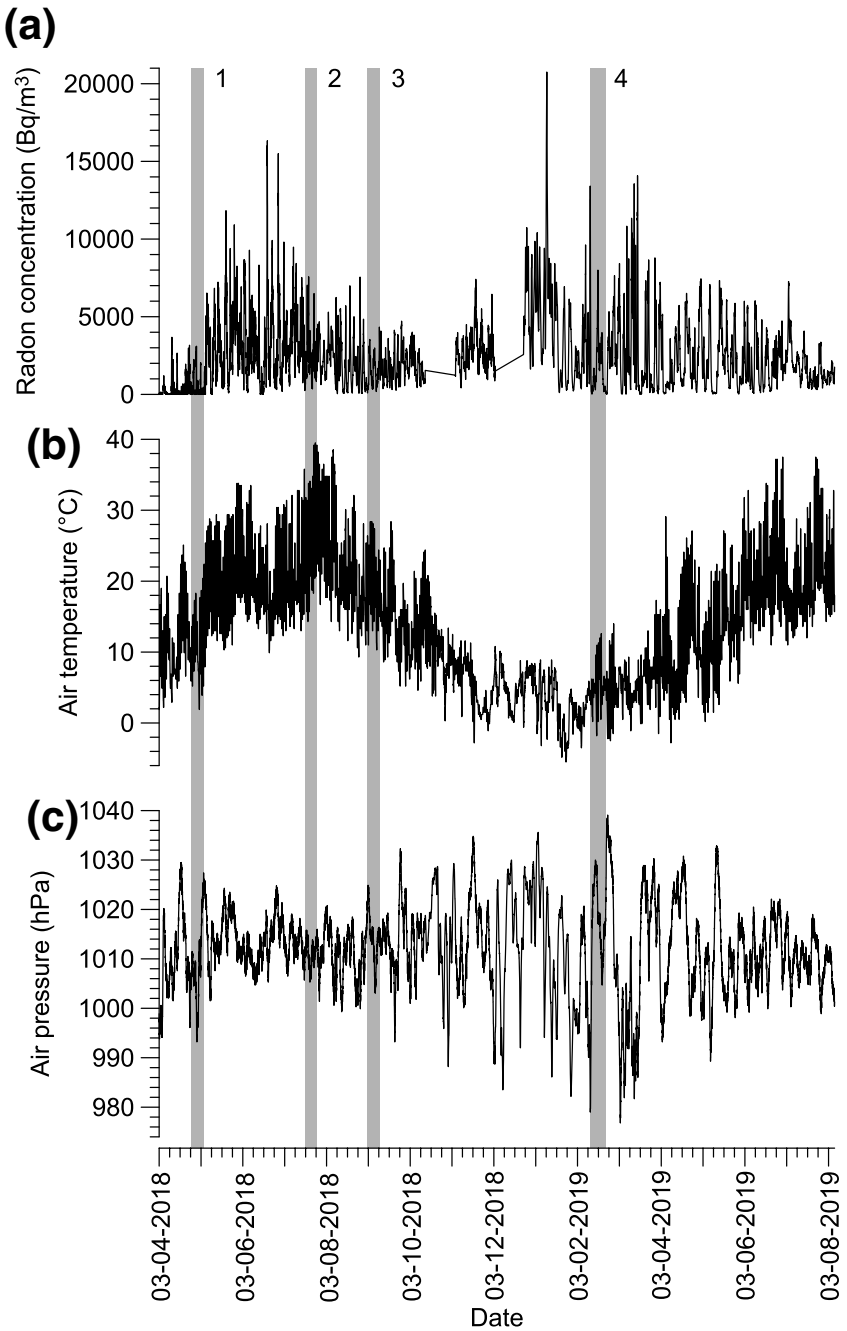


Fig. 3 Data, time series of **a** Rn concentration, **b** air temperature and **c** air pressure for the full time interval (03-04-2018 02:16 to 07-08-2019 10:47). Also indicated (shaded) and numbered are the four intervals used for statistical analysis (Table 1)

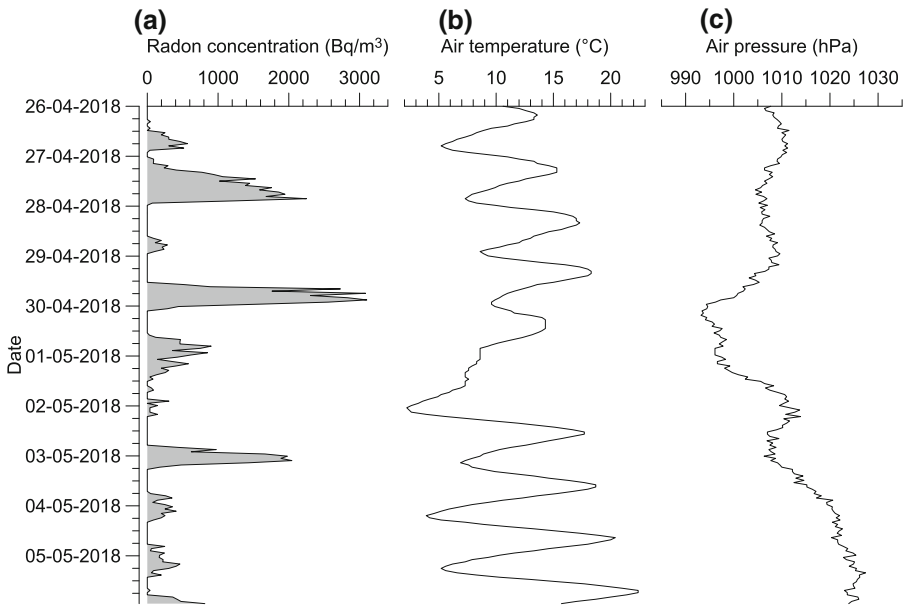


Fig. 4 Data, time series of **a** Rn concentration, **b** air temperature and **c** air pressure for the time interval 1 (Table 1). This series represents exhalation mode II (“night mode”); see Sect. 5

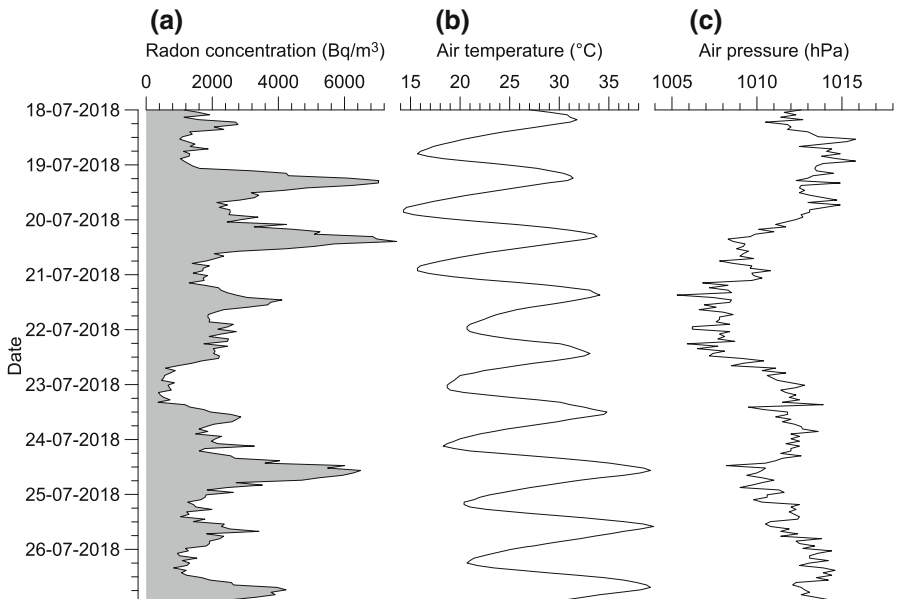


Fig. 5 Data, time series of **a** Rn concentration, **b** air temperature and **c** air pressure for the time interval 2 (Table 1). This series represents exhalation mode III (“day mode”); see Sect. 5

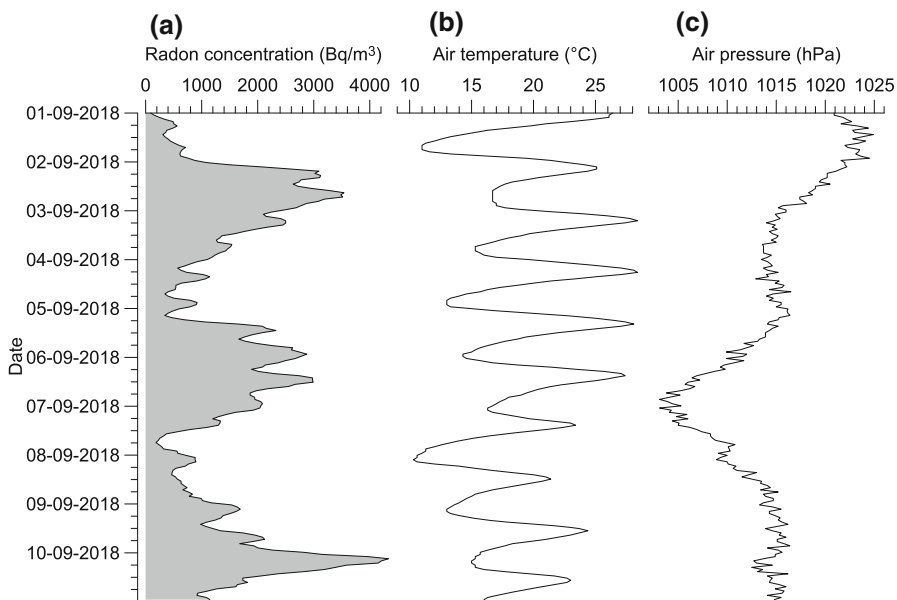


Fig. 6 Data, time series of **a** Rn concentration, **b** air temperature and **c** air pressure for the time interval 3 (Table 1). This series represents exhalation mode IV (“day + night + air pressure mode”); see Sect. 5

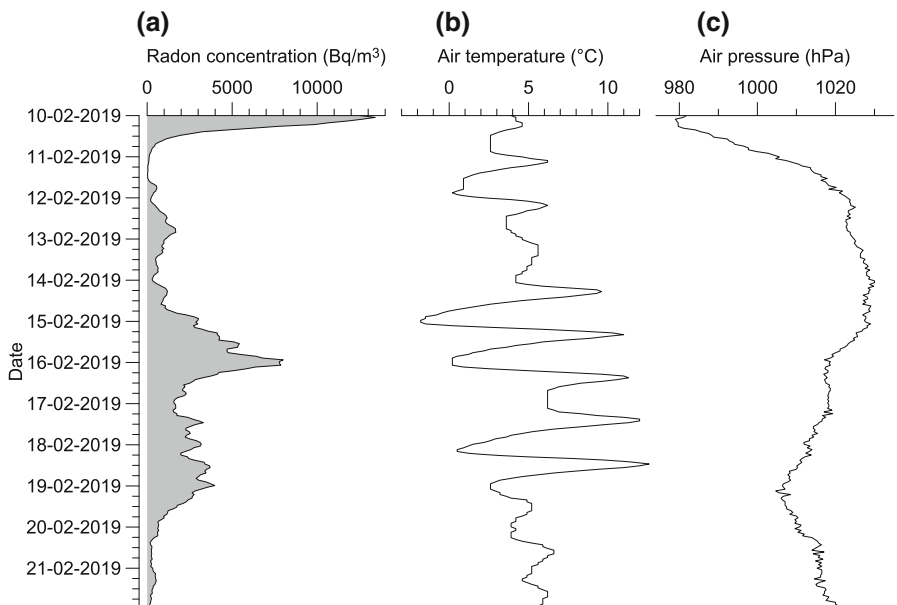


Fig. 7 Data, time series of **a** Rn concentration, **b** air temperature and **c** air pressure for the time interval 4 (Table 1). This series represents exhalation mode I (“air pressure mode”); see Sect. 5

Table 1 Database, selected time intervals. Also shown (roman numbers) are the corresponding Rn exhalation modes

Interval	Mode	Start	End	n	d (h)
<i>Air measurement data</i>					
Interval 1	II	26-04-2018 02:16	06-05-2018 01:16	240	1
Interval 2	III	18-07-2018 02:16	27-07-2018 01:16	216	1
Interval 3	IV	01-09-2018 02:51	11-09-2018 01:51	240	1
Interval 4	I	10-02-2019 01:55	22-02-2019 00:55	288	1
<i>Groundwater measurement data</i>					
Interval 1	II	26-04-2018 08:00:00	06-05-2018 00:00:00	117	2
Interval 2	III	18-07-2018 13:58:59	26-07-2018 23:58:59	102	2
Interval 3	IV	01-09-2018 09:23:42	11-09-2018 01:23:42	117	2
Interval 4	I	10-02-2019 08:33:29	22-02-2019 00:33:29	141	2
<i>Rn regression model data</i>					
Interval 1	II	26-04-2018 14:16	06-05-2018 00:16	114	2
Interval 2	III	18-07-2018 14:16	27-07-2018 00:16	102	2
Interval 3	IV	01-09-2018 15:51	11-09-2018 01:51	114	2
Interval 4	I	10-02-2019 14:55	22-02-2019 00:55	138	2

h hour

expressed as P -value, also denoted as false-alarm probability. To allow the assessment and climatological interpretation of the inferential results, the uncertainties have to be reported. Estimates without error bars are useless. For more background on climate and statistical inference, see, for example, the textbooks by von Storch and Zwiers (1999) or Mudelsee (2014).

3.1 Cross-spectral analysis

Consider a stationary process in continuous time, $X(T)$, which has a spectral representation (Priestley 1981). This means,

$$X_{T'}(T) = (2\pi)^{1/2} \int_{-\infty}^{\infty} G_{T'}(f) e^{2\pi i f T} df, \quad (1)$$

where $X_{T'}(T)$ is the truncated process (i.e., $X_{T'}(T) = X(T)$ for $-T' \leq T \leq T'$ and $X_{T'}(T) = 0$ elsewhere),

$$G_{T'}(f) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} X_{T'}(T) e^{-2\pi i f T} dT \quad (2)$$

and f is frequency. Then the one-sided nonnormalized power spectral density function—shortly denoted henceforth as autospectrum—is given by

$$h(f) = \lim_{T' \rightarrow \infty} \left\{ E \left[2\pi \left| G_{T'}(f) \right|^2 / T' \right] \right\}, \quad (3)$$

where $E[\cdot]$ is the expectation operator. The autospectrum is the Fourier transform of the autocovariance function (Priestley 1981). It measures the variance contribution of a certain frequency to the overall variance, S^2 , of the process. Other roads to the definition of a spectrum exist.

Given a discrete, possibly unevenly sampled time series of size n , $\{t(i), x(i)\}_{i=1}^n$, the task is to estimate the autospectrum. The Lomb–Scargle periodogram (Lomb 1976; Scargle 1982) is given by

$$I_{LS}(f_j) = \bar{d} \cdot \left\{ \frac{\left[\sum_{i=1}^n X(i) \cos(2\pi f_j [T(i) - \tau_{Lomb}]) \right]^2}{\sum_{i=1}^n [\cos(2\pi f_j [T(i) - \tau_{Lomb}])]^2} + \frac{\left[\sum_{i=1}^n X(i) \sin(2\pi f_j [T(i) - \tau_{Lomb}]) \right]^2}{\sum_{i=1}^n [\sin(2\pi f_j [T(i) - \tau_{Lomb}])]^2} \right\}, \quad (4)$$

where

$$\tan(4\pi f_j \tau_{Lomb}) = \frac{\sum_{i=1}^n \sin(4\pi f_j T(i))}{\sum_{i=1}^n \cos(4\pi f_j T(i))}, \quad (5)$$

the search frequencies are $f_j = 1/(n\bar{d}), \dots, 1/(2\bar{d})$ and the average temporal spacing is given by $\bar{d} = [t(n) - t(1)]/(n - 1)$. In the case of even spacing, where $d(i) = t(i) - t(i - 1) = d = \bar{d} = \text{const.}$, the Lomb–Scargle periodogram corresponds to the usual periodogram (Mudelsee 2014). The frequencies f_j are employed to search for peaks of $I_{LS}(f_j)$. One drawback of the periodogram is that it is an inconsistent autospectrum estimator (Bartlett 1955), which means that with increasing n the estimation variance of the autospectrum does not decrease.

For autospectrum estimation, we therefore followed the common practice (Schulz and Mudelsee 2002) and employed time series segments (with 50% overlap), segment-wise linear detrending, segment-wise tapering and averaging of the Lomb–Scargle periodograms for the segments. We also corrected for autospectrum estimation bias via Monte Carlo simulations of an AR(1) noise process. The idea is to compare the theoretical AR(1) spectrum (Priestley 1981) with the average autospectrum over the simulations in order to obtain a frequency-dependent bias correction factor (Schulz and Mudelsee 2002). The existence of peaks in the spectrum was tested via the upper 95th percentile of the distribution for the AR(1) red-noise alternative (Schulz and Mudelsee 2002), which was determined by means of 1000 Monte Carlo simulations. We additionally assessed peaks by means of Siegel's test at the 95% confidence level on the Lomb–Scargle periodogram (Schulz and Statterger 1997). In the case of a strongly dominating peak in the estimated autospectrum, Schulz and Statterger (1997) recom-

mended to first subtract this spectral component by means of a filter (Ferraz-Mello 1981) and only then to estimate the AR(1) parameter. We followed this recommendation since Rn concentration and air temperature exhibit strong daily components (Sect. 4.1). Multitaper estimation may for even spacing provide a superior estimation since it employs optimal tapers (Mudelsee 2014).

Next, consider two stationary processes, $X(T)$ and $Y(T)$, with autospectra $h_X(f)$ and $h_Y(f)$. The coherency (Schulz and Stattegger 1997) is defined as

$$c_{XY}^2 = \frac{|h_{XY}(f)|^2}{h_X(f) \cdot h_Y(f)}, \quad (6)$$

where $h_{XY}(f)$ is the cross-spectrum, that is, the Fourier transform of the cross-covariance function (Priestley 1981). The coherency is a dimensionless measure between 0 and 1, which quantifies the degree of the linear relation between two processes in dependence on frequency.

Given two discrete, possibly unevenly sampled time series, $\{t_X(i), x(i)\}_{i=1}^{n_X}$ and $\{t_Y(i), y(i)\}_{i=1}^{n_Y}$, the task is to estimate the coherency. Note that the set of time values, $\{t_X(i)\}_{i=1}^{n_X}$, needs not to be equal to $\{t_Y(i)\}_{i=1}^{n_Y}$. The Lomb–Scargle periodogram can also be used for the estimation of the cross-spectrum (Schulz and Stattegger 1997).

For coherency estimation, we employed segmenting (50% overlap), segment-wise linear detrending, segment-wise tapering, averaging of periodograms and tests against the AR(1) red-noise alternative (Ólafsdóttir et al. 2016). The upper percentile for the red-noise alternative was determined by means of 1000 Monte Carlo simulations. The interpretation of coherency estimates is only meaningful for frequencies where both $X(T)$ and $Y(T)$ have a significant peak in the autospectrum.

In the case of a meaningful and significant coherency estimate at a certain frequency, the estimate of the phase may shed light on lead–lag behaviour between the two processes at that frequency. We estimated also the phase via the Lomb–Scargle periodogram and the estimated cross-spectrum (Ólafsdóttir et al. 2016, Eq. (9) therein). The 95% confidence interval for the phase estimate was determined by means of 1000 Monte Carlo simulations.

For tapering in the estimations, we consistently employed the Welch I type, a number of seven overlapping segments and an oversampling factor (which corresponds to interpolation in the frequency domain in order to better visualize spectrum curves) of 16. The resulting spectral bandwidths (which determine the frequency resolution) for the Rn concentration and air temperature spectra (Sect. 4.1) is between $2.21 \cdot 10^{-2} \text{ h}^{-1}$ ($n = 288$) and $2.94 \cdot 10^{-2} \text{ h}^{-1}$ ($n = 216$). See Schulz and Mudelsee (2002) for more detailed technical explanation.

We used the software REDFIT-X (Ólafsdóttir et al. 2016) for autospectrum, coherency and phase estimation. We used the software TAUEST (Mudelsee 2002) for the estimation of the AR(1) parameter in the presence of uneven spacing, which includes a bias correction after Kendall (1954).

3.2 Kernel trend and derivative estimation

Consider a nonstationary process in continuous time, $X(T)$, which can be decomposed into trend, extremes and noise components,

$$X(T) = X_{\text{trend}}(T) + X_{\text{ext}}(T) + S(T) \cdot X_{\text{noise}}(T). \quad (7)$$

Mudelsee (2014) calls this the climate equation. The noise, $X_{\text{noise}}(T)$, is a zero-mean, unit-standard deviation process, which may have a spectral representation (Sect. 3.1). The noise may also describe the memory (i.e., autocorrelation, red noise) of the climate. The climate variability, $S(T)$, is a positive function. The trend component, $X_{\text{trend}}(T)$, is the long-term, possibly time-dependent climate mean. The terms $X_{\text{trend}}(T)$ and $S(T) \cdot X_{\text{noise}}(T)$ represent the definition of climate in terms of mean and variability (Brückner 1890; Hann 1901; Köppen 1923). The extreme component, $X_{\text{ext}}(T)$, was added by Mudelsee (2014) to allow for a separate analysis of climate extremes and related parameters (e.g., risk).

Given a discrete, possibly unevenly sampled time series, $\{t(i), x(i)\}_{i=1}^n$, the task is to estimate the trend component. In many situations, the data analyst aims for trend-model flexibility and avoids the restriction imposed by parametric models (e.g., linear) by means of a kernel smoothing technique. The approach by Gasser and Müller (1979, 1984) is:

$$\hat{X}_{\text{trend}}^{\text{GM}}(T) = h^{-1} \sum_{i=1}^n \left[\int_{s(i-1)}^{s(i)} K\left(\frac{T-y}{h}\right) dy \right] X(i), \quad (8)$$

where $T(i-1) \leq s(i-1) \leq T(i)$. (The integration bounds, $s(i)$, are described in a subsequent paragraph.)

The kernel is a continuous and usually positive and symmetric function, it integrates as $\int K(y)dy = 1$. The kernel function employed by us is the Epanechnikov kernel, $K(y) = 0.75 \cdot (1 - y^2)$. Another common choice is the Gaussian, $K(y) = (2\pi)^{-1/2} \exp(-y^2/2)$.

Whereas the choice of K (Epanechnikov, Gaussian, and so forth), is more of “cosmetic” (Diggle 1985) interest, the bandwidth parameter, h , is crucial because it determines the uncertainty measures for the trend estimate. Although there exist bandwidth selectors for optimal smoothing (Mudelsee 2014), we prescribed $h = 1$ day in order to smooth over the daily cycle.

Note that the selection of the sequence $s(i)$ in particular, and the Gasser–Müller smoothing procedure in general, can be performed on unevenly spaced time series. The trend can be estimated for all time points within the observation interval, $[t(1); t(n)]$. The kernel functions are modified near the interval boundaries (Gasser and Müller 1979, 1984), so that the trend can be estimated also there.

KERNEL (<https://www.manfredmudelsee.com/soft/kernel/index.htm>, 30 August 2019) is a Fortran software based on routines originally developed by Theo Gasser. It places the integration bound, $s(i)$, in the middle between two time points. KERNEL further sets

$$s(0) = 1.5 \cdot t(1) - 0.5 \cdot t(2) \quad (9)$$

and

$$s(n) = 1.5 \cdot t(n) - 0.5 \cdot t(n - 1). \quad (10)$$

Also the first derivative of the variable, $dX(T)/dT$, and its long-term time dependence, can be studied by means of the kernel technique. This goes via the first derivative of the kernel function (Gasser and Müller 1984).

An uncertainty measure for the estimated trend or derivative curves is essential for assessing the significance of the ups and downs in the estimate, whether these variations constitute real features or are generated by noise.

Let $\hat{x}_{\text{trend}}^{\text{GM}}(i)$ be the estimate (sample level) at time $t(i)$. Date minus trend estimate define the unweighted nonparametric regression residuals, $e(i) = x(i) - \hat{x}_{\text{trend}}^{\text{GM}}(i)$. The time series of residuals is $\{t(i), e(i)\}_{i=1}^n$.

The moving block bootstrap (MBB) resampling procedure (Künsch 1989) draws random blocks of length l ($0 < l < n$) from the residuals. Mudelsee (2014) presents l -selectors, which are based on the autocorrelation properties of the data. The MBB then concatenates the blocks until a series of size n is obtained. The random series, denoted as $\{t(i), e^*(i)\}_{i=1}^n$, preserves the statistical properties of the random component of the data generating process. These are distributional shape and serial dependence (over l).

The so-called replication of the trend estimate is obtained via re-application of the kernel technique. This procedure resampling–estimation is repeated B times (we use $B = 400$). The bootstrap standard error, $\text{se}(T')$, of the trend estimate at a certain time point, T' , is defined as the standard deviation over the B replications at T' .

A pointwise standard-error band can be constructed from the standard-error intervals for the trend estimate as follows. The standard-error interval for the time value T' is given by $[\hat{X}_{\text{trend}}^{\text{GM}}(T') - \text{se}(T'); \hat{X}_{\text{trend}}^{\text{GM}}(T') + \text{se}(T')]$. The band is obtained by concatenating the upper bounds, $\hat{X}_{\text{trend}}^{\text{GM}}(T') + \text{se}(T')$, for the full time interval, $t(1) \leq T' \leq t(n)$, and by concatenating the lower bounds, $\hat{X}_{\text{trend}}^{\text{GM}}(T') - \text{se}(T')$. For the Rn project, we show in a conservative approach the wider two-standard-errors bands, which are obtained from the interval $[\hat{X}_{\text{trend}}^{\text{GM}}(T') - 2 \text{se}(T'); \hat{X}_{\text{trend}}^{\text{GM}}(T') + 2 \text{se}(T')]$.

The book by Mudelsee (2014) contains more details about kernel estimation, MBB resampling, block length selection and the construction of uncertainty measures.

3.3 Multiple linear regression model

Let $Y(i)$ denote Rn concentration in discrete time. Let $X_1(i), X_2(i), \dots, X_p(i)$ denote a number of p air and groundwater variables. Assume that there exists a simple linear relation,

$$Y(i) = \beta_0 + \beta_1 \cdot X_1(i) + \beta_2 \cdot X_2(i) + \dots + \beta_p \cdot X_p(i) + S_Y(i) \cdot Y_{\text{noise}}(i) \quad (11)$$

for $i = 1, \dots, n$. This is a multiple linear regression model (Montgomery and Peck 1992), where $Y(i)$ is the response variable and $X_1(i), X_2(i), \dots, X_p(i)$ are the predictor variables. The term $S_Y(i) \cdot Y_{\text{noise}}(i)$ is the noise component. The model can be used to predict unobserved Rn concentrations on basis of observed air and groundwater data.

In order to derive theoretical properties of estimates obtained by means of the regression model, the assumptions are often made that $S_Y(i) = S_Y$ is constant (homoscedasticity) and $Y_{\text{noise}}(i)$ is a purely random Gaussian stochastic process. However, in practice the model may be useful also in the presence of (1) heteroscedasticity, (2) autocorrelation and (3) non-Gaussian distributions. Indeed, for practical purposes the linearity of the model may also be sufficiently accurate in the presence of mild forms of nonlinear dependence. To quote Box and Draper (1987, p. 424 therein): “Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.”

Given a discrete, possibly unevenly sampled time series of size n , $\{t(i), y(i), x_1(i), x_2(i), \dots, x_p(i))\}_{i=1}^n$, the task is to estimate the regression parameters, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$. This can be achieved by minimizing the sum of squares of deviations between data and linear fit. The formulas for this least-squares estimation (Montgomery and Peck 1992) are as follows. Let

$$\mathbf{y} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n) \end{bmatrix} \quad (\text{response vector}), \quad (12)$$

$$\mathbf{x} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_p(1) \\ x_1(2) & x_2(2) & \cdots & x_p(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(n) & x_2(n) & \cdots & x_p(n) \end{bmatrix} \quad (\text{data matrix}), \quad (13)$$

then

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \quad (\text{parameter estimate vector}) \quad (14)$$

is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{y}. \quad (15)$$

In this matrix notation (Dahlquist and Björck 2008, Appendix A therein), if \mathbf{A} is a matrix, then \mathbf{A}' is its transpose and \mathbf{A}^{-1} is its inverse. We used the `lm` function within the R computing environment (Dalgaard 2008) for the calculations.

The “classical” uncertainty measures (Montgomery and Peck 1992) for the parameter estimates are based on assumptions such as a purely random Gaussian noise. In case of autocorrelated noise, the estimates can be obtained via the effective data size Mudelsee (2014, Chapter 2 therein), the determination of which can be achieved via the time series of the residuals and AR(1) fits (Mudelsee 2002). If there are indications

that also the distributional assumption is violated, then bootstrap uncertainty measures (via the MBB) can be determined.

However, the actual parameter estimates (and their uncertainties) are less of interest in the present paper. The focus is on a strategy for prediction and the selection of predictors. For this, a measure of the quality of the fit of the model (Eq. 11) to the data is relevant. Literature (Montgomery and Peck 1992) recommends to calculate the adjusted coefficient of multiple determination, R_{adj}^2 , as follows. Let

$$\bar{y} = \sum_{i=1}^n y(i)/n, \quad (16)$$

$$S_{yy} = \sum_{i=1}^n y(i)^2 - n \cdot \bar{y}^2, \quad (17)$$

$$y_{\text{fit}}(i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1(i) + \hat{\beta}_2 \cdot x_2(i) + \dots + \hat{\beta}_p \cdot x_p(i), \quad i = 1, \dots, n, \quad (18)$$

$$SSE = \sum_{i=1}^n [y(i) - y_{\text{fit}}(i)]^2, \quad (19)$$

$$R^2 = 1 - SSE/S_{yy}, \quad (20)$$

then

$$R_{\text{adj}}^2 = 1 - \left(1 - R^2\right) \cdot (n - 1) / (n - p - 1). \quad (21)$$

The advantage of using the adjusted measure, R_{adj}^2 , instead of R^2 is that this penalizes overfitting with too many predictors. R_{adj}^2 and R^2 are both between 0 and 1. The values can be used as a measure of the data variance explained by the model.

We made an extension of the model to take into account prior knowledge, namely that the Rn concentration cannot be less than zero. Hence, we augmented the model (Eq. 11) as follows: if $y_{\text{fit}}(i) < 0$, then set $y_{\text{fit}}(i) = 0$. This augmented model has one extra parameter (which has a value of zero). The calculation of R_{adj}^2 in Eq. (21) has to take this into account by an increase of p by one.

4 Results

4.1 Cross-spectra

The results of the cross-spectral analyses for Rn concentration and air temperature in the various intervals are presented in numerical form (Table 2) and as plots (Figs. 8, 9, 10 and 11). The phase estimate for the relation between Rn concentration and air temperature for the daily cycle ($T_{\text{period}} = 1.0$ days) can be used to support the regression model (Sect. 4.3) via the inclusion of lagged predictor variables. The other air or groundwater variables do not exhibit strong daily cycles (results not shown).

For interval 1, both Rn concentration and air temperature show a clear, highly significant daily cycle (Fig. 8a, b). The coherency for that cycle is also significant (Fig. 8c).

Table 2 Results, cross-spectral analyses. T_{period} , period (one over frequency) of the major spectral peak; \hat{a}' , bias-corrected AR(1) parameter estimate; $\hat{\Phi}$, phase estimate with 95% confidence interval ($\hat{\Phi} > 0$ means that air temperature leads over Rn concentration)

Interval	Mode	Rn concentration		Air temperature		$\hat{\Phi}$
		T_{period}	\hat{a}'	$T_{\text{period}}(d)$	\hat{a}'	
Interval 1	II	1.0 d	0.44	1.0	0.61	141°[96°; 183°]
Interval 2	III	1.0 d	0.13	1.0	0.69	19°[-4°; 44°]
Interval 3	IV	8.0 h	0.66	1.0	0.54	NA
Interval 4	I	7.2 h	0.56	1.0	0.08	NA

Also shown (roman numbers) are the corresponding Rn exhalation modes. Units: *d* day; *h* hour. *NA* not applicable

This allows to study the phase, which is estimated (with 95% confidence interval) as 141°[96°; 183°] (Fig. 8d). This means that for the daily cycle, air temperature leads over Rn concentration by an estimated time lag (with 95% confidence interval) of $(141/360) \cdot 24 \text{ h} = 9.0 \text{ h}$ [6.4 h; 12.2 h].

Also for interval 2, both Rn concentration and air temperature show a clear, highly significant daily cycle (Fig. 9a, b). The coherency for that cycle is also significant (Fig. 9c). The phase is estimated as 19°[-4°; 44°] (Fig. 9d). This means that for the daily cycle, air temperature leads over Rn concentration by an estimated time lag of 1.3 h [-0.3 h; 2.9 h]. Since the confidence interval includes zero, the time lag is not statistically significant.

For interval 3 (Fig. 10) and interval 4 (Fig. 11), Rn concentrations do not show a significant daily cycle. Hence, phase estimation is not applicable here.

On basis of the phase estimates, we included lagged temperature as predictor variable. To accommodate for estimation uncertainties in the phase, and perhaps also of the peak period, a liberal approach (i.e., many time lags and for all intervals) was taken. That means, air temperature before 0 h, 1 h, 2 h, ..., 12 h were used as predictors for Rn concentrations.

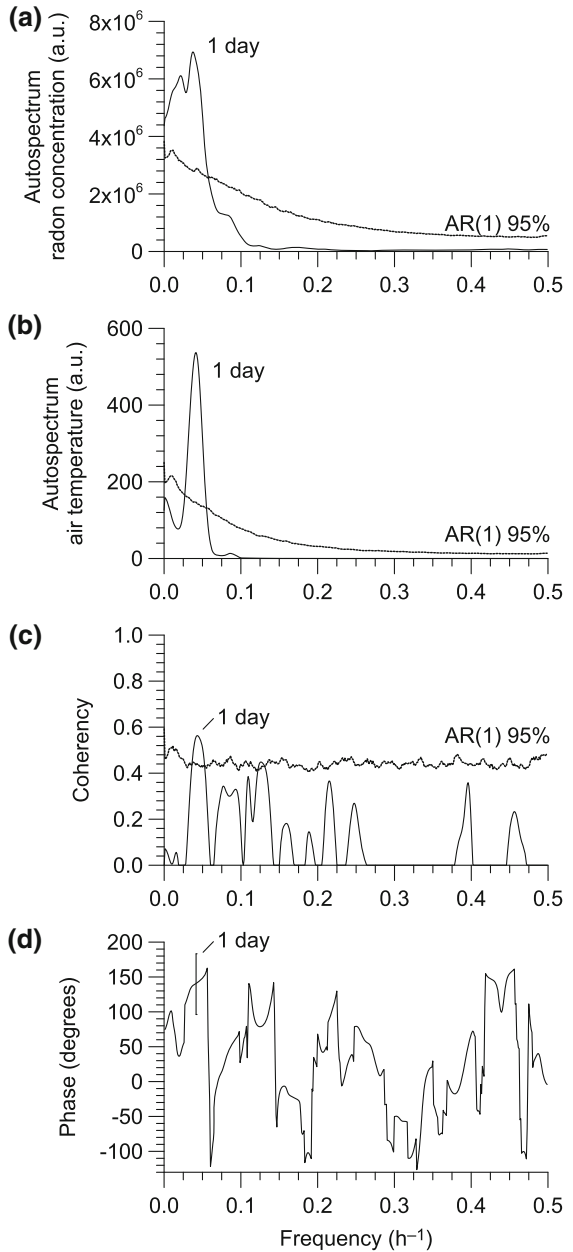
4.2 Trends and derivatives

The results of the analyses of time-dependent trends and time-dependent first derivatives of the air pressure are shown as plots (Figs. 12, 13, 14 and 15). Air pressure and its changes may constitute relevant meteorological control variables of Rn concentration. Therefore, they are included as predictors in the regression model (Sect. 4.3).

The shown time intervals in the plots are those for the Rn regression model data (Table 1). Note that the predictors include air temperature lagged by up to 12 h (Sect. 4.1). This means that the air-pressure data that are available for trend and derivative estimation (shown as filled symbols in Figs. 12a, 13a, 14a and 15a) extend somewhat beyond the shown interval boundaries. This explains, for example, the trend estimate at the earlier interval boundary (10-02-2019 08:55) for interval 4 (Fig. 15a).

The estimated persistence times, $\hat{\tau}$, with standard errors for the fits of an AR(1) autocorrelation model under uneven spacing (Mudelsee 2002), obtained on the resid-

Fig. 8 Cross-spectra, interval 1 (mode II). See Sect. 3.1 for methodical details and Table 2 for result numbers. The vertical bar in panel (d) shows the 95% confidence interval for the phase estimate at the daily cycle



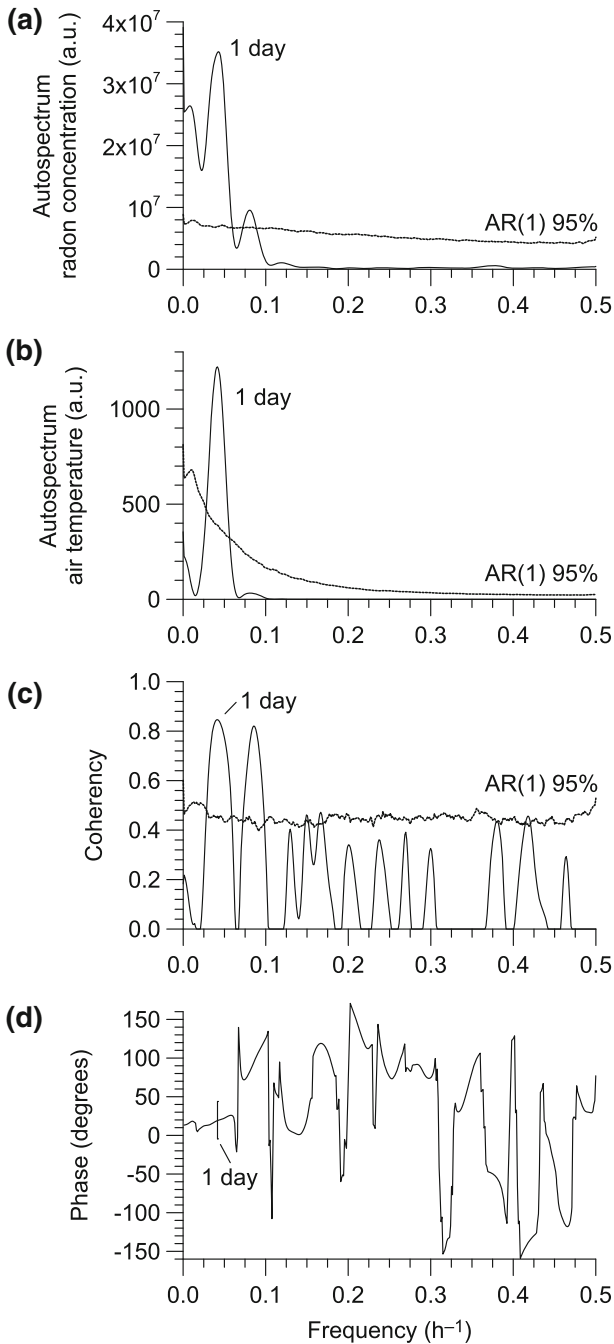


Fig. 9 Cross-spectra, interval 2 (mode III). See Sect. 3.1 for methodical details and Table 2 for result numbers. The vertical bar in panel (d) shows the 95% confidence interval for the phase estimate at the daily cycle

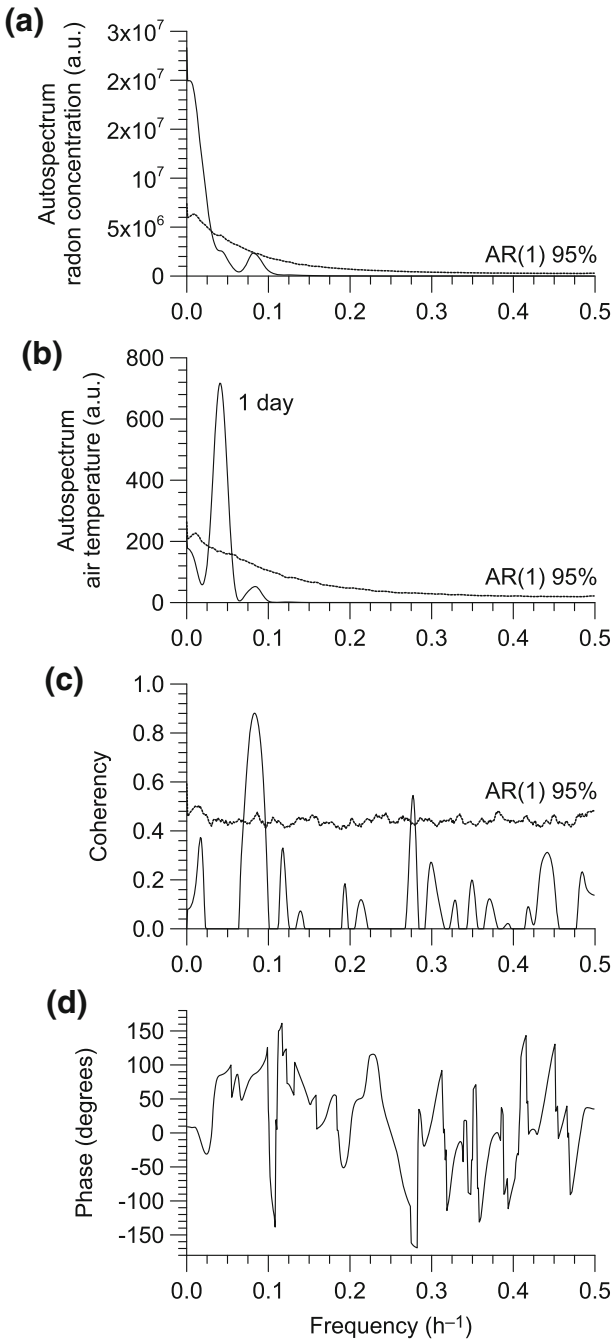


Fig. 10 Cross-spectra, interval 3 (mode IV). See Sect. 3.1 for methodical details and Table 2 for result numbers

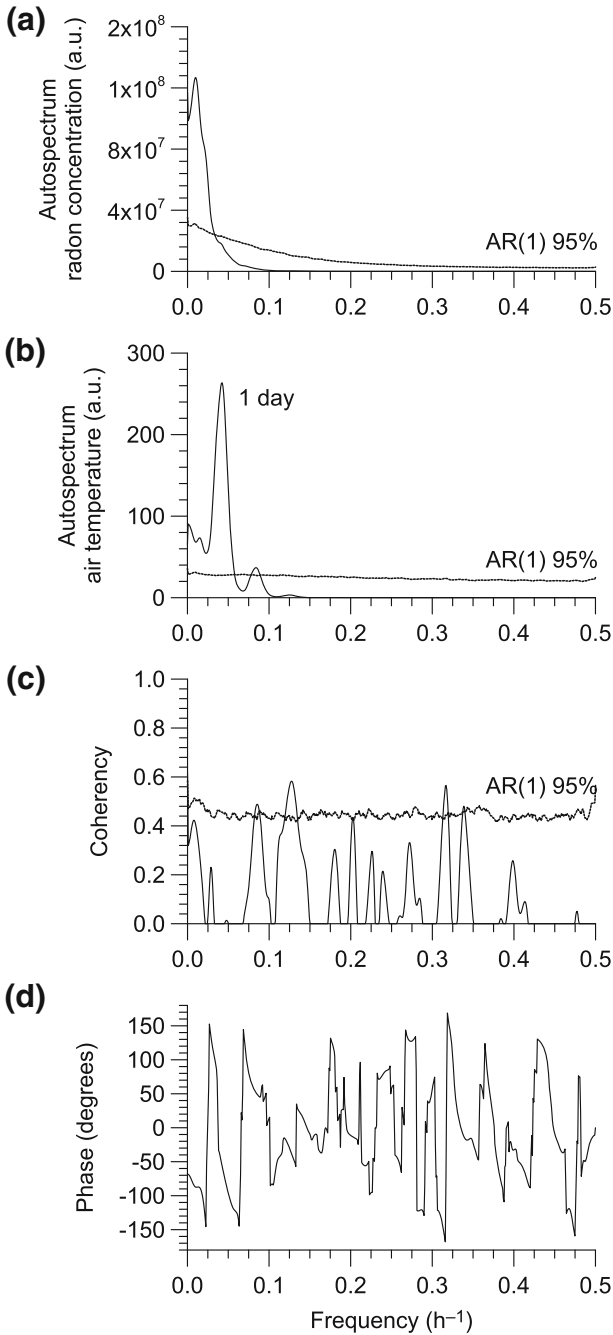


Fig. 11 Cross-spectra, interval 4 (mode I). See Sect. 3.1 for methodical details and Table 2 for result numbers

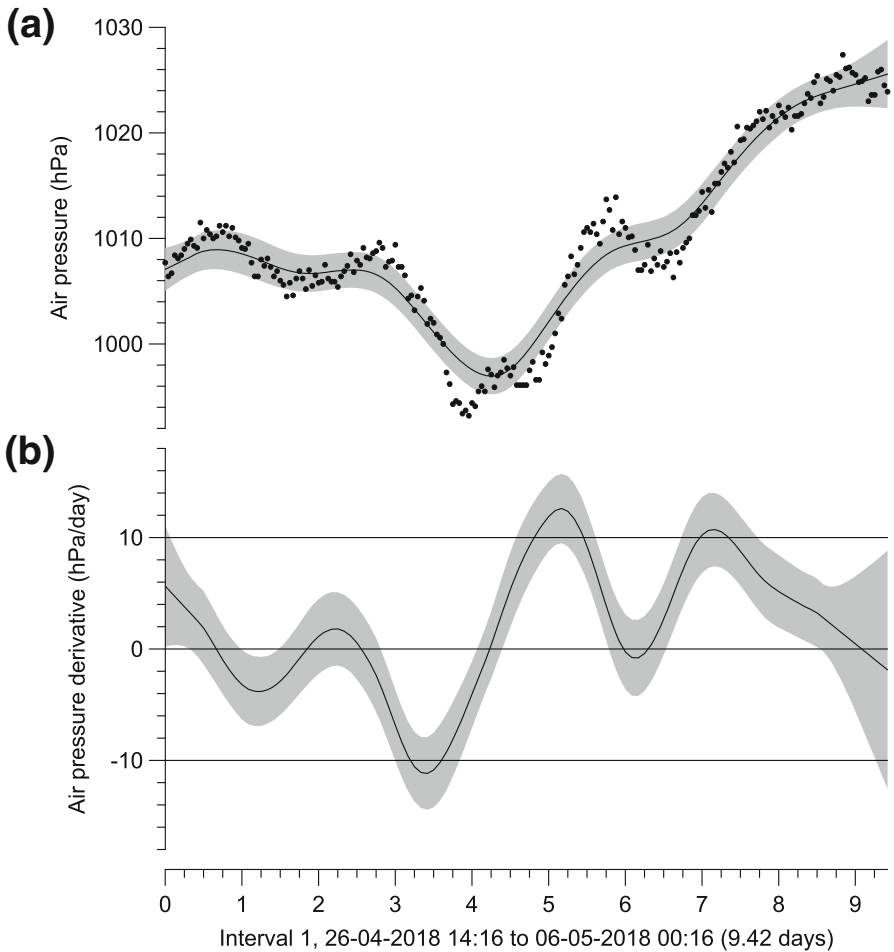


Fig. 12 Kernel trend and derivative estimation, interval 1 (mode II). Shown are data (**a**; filled symbols), estimation curves (**a**, **b** solid, wiggly lines) and two-standard-errors bands (**a**, **b**; shaded). See Sect. 3.2 for methodical details

uals, $e(i)$, are: interval 1, $\hat{\tau} = 4.8 \pm 1.0$ h; interval 2, $\hat{\tau} = 1.1 \pm 0.2$ h; interval 3, $\hat{\tau} = 2.1 \pm 0.3$ h and interval 4, $\hat{\tau} = 19.8 \pm 5.4$ h. The variations among these numbers indicate that the memory of air-pressure fluctuations depends on the general weather situation. The persistence time estimates, together with samples sizes (Table 1), translate after Mudelsee (2014, Eq. 3.28 therein) into following block lengths, l , for the MBB: interval 1, $l = 20$; interval 2, $l = 7$; interval 3, $l = 11$ and interval 4, $l = 55$.

The MBB confidence bands attest that there occurred significant changes in air-pressure trends and derivatives on daily and longer timescales (Figs. 12, 13, 14 and 15). This features are taken into account for the regression model of Rn concentrations. In addition to the smoothed series ($h = 1$ day) of trend and first derivative, we employ

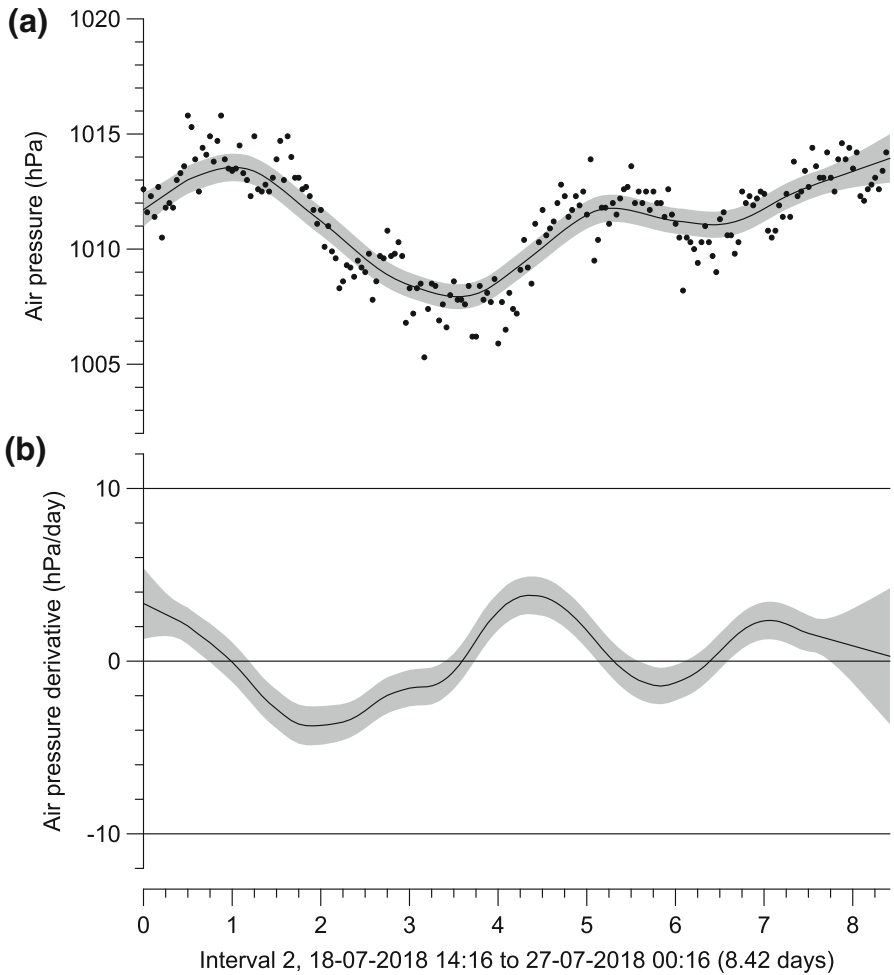


Fig. 13 Kernel trend and derivative estimation, interval 2 (mode III). Shown are data (**a** filled symbols), estimation curves (**a**, **b** solid, wiggly lines) and two-standard-errors bands (**a**, **b** shaded). See Sect. 3.2 for methodical details

as predictors also the original air-pressure data, $x(i)$, and the first differences, $x(i) - x(i - 1)$, that means, versus before 1 h.

4.3 Regression model

Table 3 shows the description of the 19 predictors for Rn concentration. Since the model is constrained to $y_{\text{fit}}(i) \geq 0$, the calculation of R_{adj}^2 in Eq. (21) employs a value of $p = 20$. The results of the regression model fits are shown as plots (Figs. 16, 17, 18 and 19).

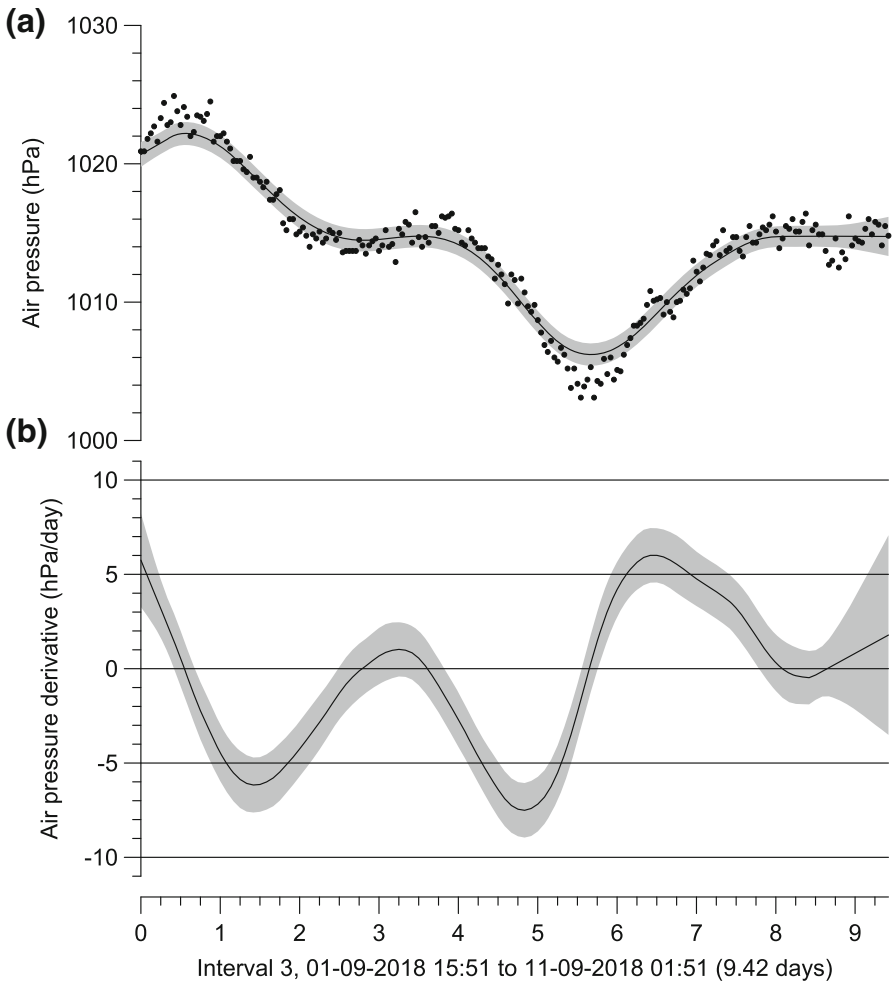


Fig. 14 Kernel trend and derivative estimation, interval 3 (mode IV). Shown are data (**a** filled symbols), estimation curves (**a**, **b**; solid, wiggly lines) and two-standard-errors bands (**a**, **b**; shaded). See Sect. 3.2 for methodical details

The overall visual appearances of the fits for the four intervals is good. This is reflected in the high numerical values of R_{adj}^2 , which are not very much smaller than the theoretical maximum of one. The constraint $y_{\text{fit}}(i) \geq 0$ brought a slight improvement of fit quality (in terms of R_{adj}^2) for intervals 1 and 4. It is remarkable that the rather simple linear regression model yields such good fits for the four different weather situations reflected in the four intervals.

Still, there are deviations between data and model. These deviations appear especially prominent in cases where the data exceed the model fits. One example of such a prominent peak is between days 8 and 9 for interval 3 (Fig. 18). These peaks of “excess

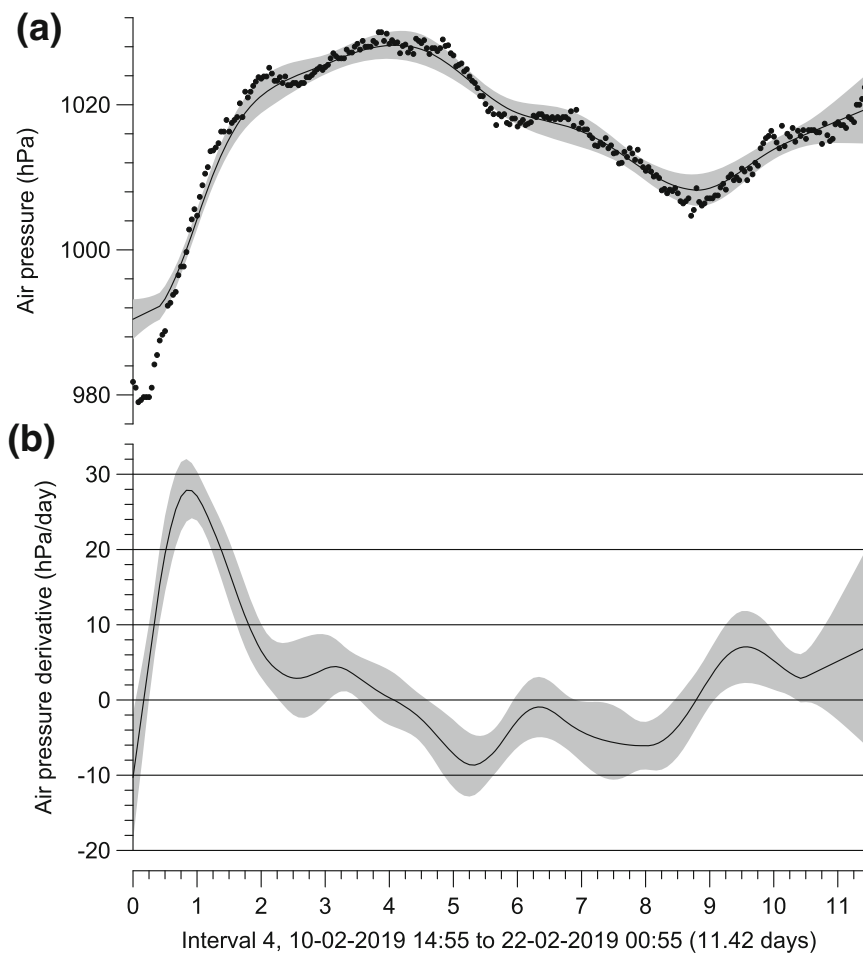


Fig. 15 Kernel trend and derivative estimation, interval 4 (mode I). Shown are data (**a**; filled symbols), estimation curves (**a**, **b**; solid, wiggly lines) and two-standard-errors bands (**a**, **b**; shaded). See Sect. 3.2 for methodical details

Rn” (i.e., higher Rn than assumed on basis of the weather conditions) are subjected to further consideration (Sect. 5).

Another point for further discussion is the separation of the different weather regimes. This is important for “out-of-sample” prediction. This point is further pursued in Sect. 5.

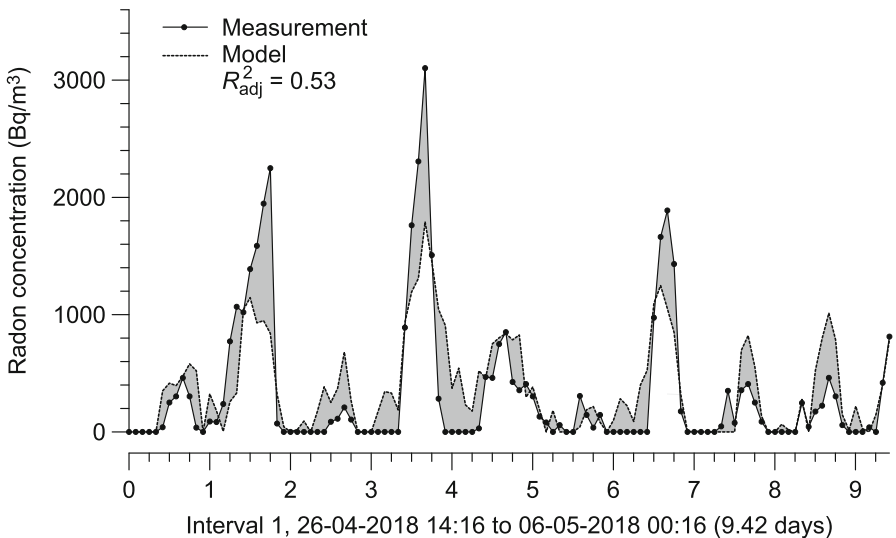
5 Discussion

The fits of the relatively simple regression model to the Rn concentration data for the four time intervals (Figs. 16, 17, 18 and 19) enjoy rather high R_{adj}^2 values, between 0.53

Table 3 Regression model predictors for Rn concentration (in units of Bq/m³)

Number	Description	Unit
1	Air temperature	°C
2	Air temperature before 1 h	°C
3	Air temperature before 2 h	°C
4	Air temperature before 3 h	°C
5	Air temperature before 4 h	°C
6	Air temperature before 5 h	°C
7	Air temperature before 6 h	°C
8	Air temperature before 7 h	°C
9	Air temperature before 8 h	°C
10	Air temperature before 9 h	°C
11	Air temperature before 10 h	°C
12	Air temperature before 11 h	°C
13	Air temperature before 12 h	°C
14	Air pressure	hPa
15	Air pressure, first difference	hPa
16	Air pressure, trend	hPa
17	Air pressure, first derivative	hPa/d
18	Groundwater level	m
19	Groundwater temperature	°C

d day

**Fig. 16** Rn concentration data and regression model fit, interval 1 (mode II). Also shown (shaded) are the deviations between data and fit. See Sect. 3.3 for methodical details

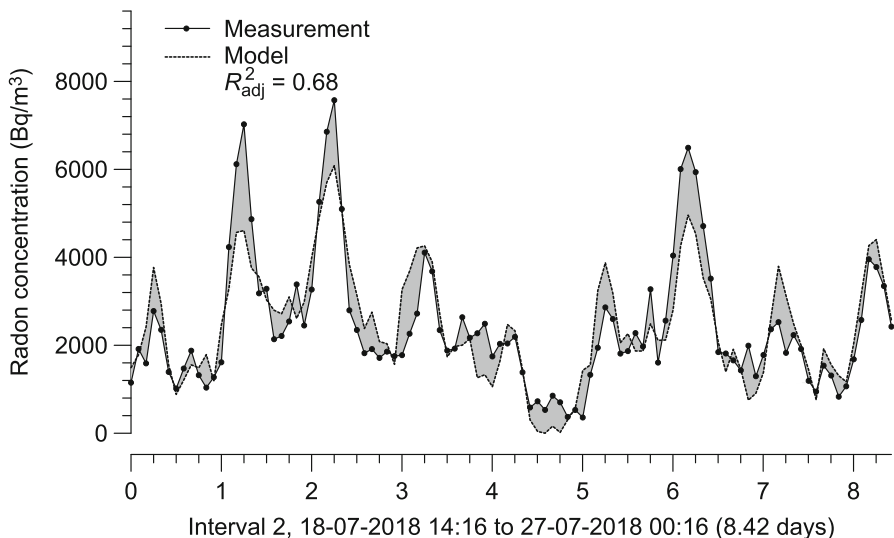


Fig. 17 Rn concentration data and regression model fit, interval 2 (mode III). Also shown (shaded) are the deviations between data and fit. See Sect. 3.3 for methodical details

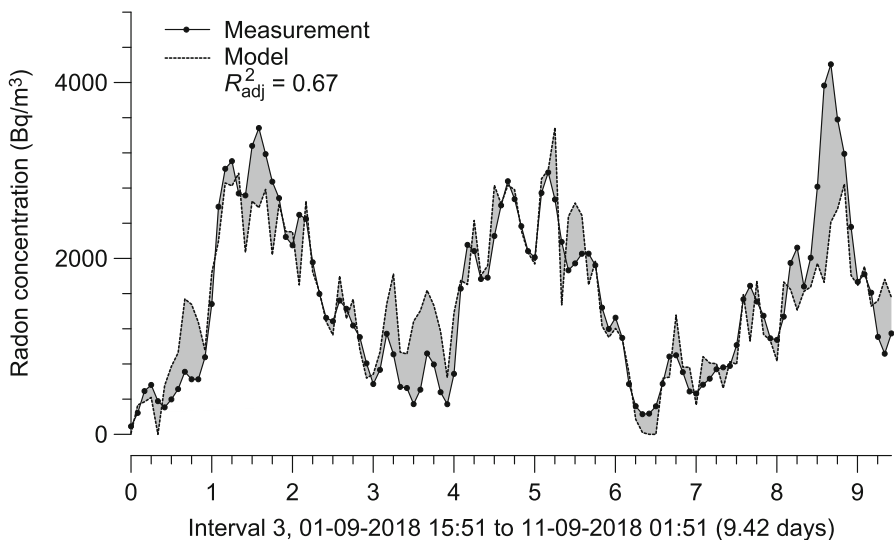


Fig. 18 Rn concentration data and regression model fit, interval 3 (mode IV). Also shown (shaded) are the deviations between data and fit. See Sect. 3.3 for methodical details

and 0.86. Also if judged per eye, the fitted models appear to have a good descriptive power. An interesting methodical extension would be to consider fit measures other than R^2_{adj} , that means, measures that take into account the persistence in the time series data. This can be achieved by means of the effective data size Mudelsee (2014). For the present paper, however, this extension seems to be beyond the scope, and we believe that the fit assessments would not strongly change.

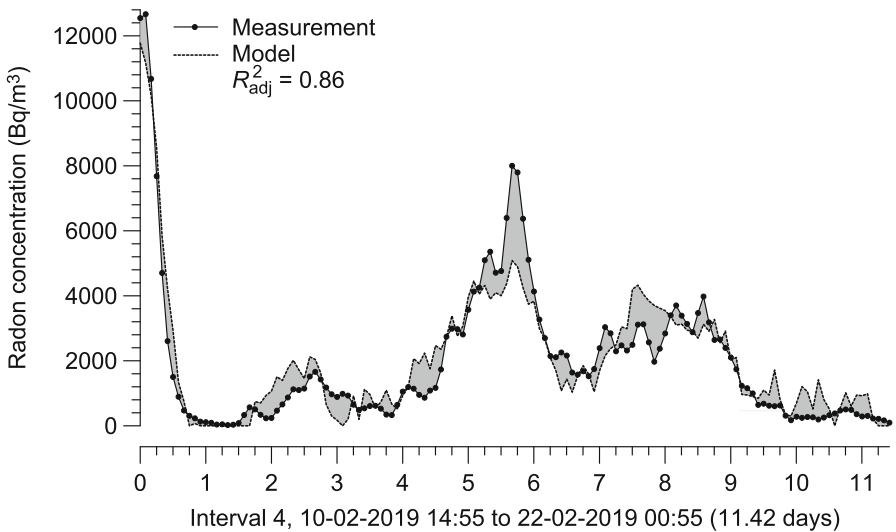


Fig. 19 Rn concentration data and regression model fit, interval 4 (mode I). Also shown (shaded) are the deviations between data and fit. See Sect. 3.3 for methodical details

The largest deviations between model and data (indicated by the shaded areas in Figs. 16, 17, 18 and 19), are the excess Rn peaks, for which the Rn measurement values clearly exceed those of the weather-derived model fits. Many of those measurement peaks appear at times when also the model has a peak (but not so high), as, for example, around day 8.75 in Fig. 18. This hints at amplifying mechanisms, which may be taken into account by means of adding nonlinear terms to the regression model. However, some of the excess Rn measurement peaks may have no counterpart in the model, as, for example, around day 5.5 in Fig. 18. An explanation of such excess Rn peaks may be the occurrence of sporadic events affecting the ground, where a kind of “bypass” is formed to Rn stored in the depth. Microearthquakes can be one type of such an event (Al-Hilal et al. 1998; Steinitz et al. 2003; Walia et al. 2010). Therefore it would be interesting to compare the series of excess Rn peaks to seismic time series for the region. This will be done in a future paper. Other driving factors may be events of high precipitation and high wind speed (Schumann and Gundersen 1996; Martin et al. 2004; Gregorič et al. 2014). Furthermore, the series of excess Rn peaks may also be analysed by means of statistical tools from climate risk analysis (Mudelsee 2014, Chapter 6 therein). One typical inference would be the estimation of the return period of excess peaks, and another inference the analysis of the time-dependence in the occurrence rate of such events.

A hint for the reader who searches for literature on nonlinear models and data analysis. Standard references on nonlinear time series analysis are provided by Priestley (1988) from a statistical viewpoint and Kantz and Schreiber (1997) from a nonlinear dynamical system viewpoint. Tong (1990, 1992) took the notable approach to build a bridge between the two areas. Although these research areas are in development, it is probably a good learning strategy to start with the mentioned works.

The four time intervals (Table 1) have been selected as representatives of the different weather regimes (i.e., Rn exhalation modes). The “in-sample” model fits (Figs. 16, 17, 18 and 19) were on the one hand “necessarily good.” On the other, the good descriptive power of weather has a root in physics. The Rn flux is to a large extent controlled by only two weather parameters, the surface air pressure and surface air temperature, which explain a large fraction of the Rn variance (indicated by the R_{adj}^2 values). The physical processes that govern these patterns are likely related to convective meteorological processes. These may influence the partial pressure of a gas above a liquid (Henry’s law), see, for example, Westphal (1970, p. 245 therein). In particular, it is the thermal convection in the open pore space in the loose Quaternary sediments above the groundwater level, from which the Rn is released.

For the practice of prediction, however, it has to be decided on basis of the weather data which regime, which Rn exhalation mode prevails. What are the defining weather properties for our site (Fig. 1)?

Exhalation mode I. The effect of the air pressure becomes best visible during the weeks of reduced temperature variations, that means, during days of minimal temperature gradients between day and night in October or March. However, particularly the October is associated with strong air-pressure gradients. This situation leads to increased flux from the strata into the air during times of rapid pressure reduction (as the opening the cover of a pot with hot water). The weather regression model for this exhalation mode explains 86% of the variance.

Exhalation mode II. The maximum of Rn is during the night, when air temperature falls below the temperature in the borehole, that is, when air temperature is below the groundwater temperature. This process (called chimney effect) dominates the Rn flux during winter and can be used as an indirect measure of groundwater temperature. This process works perfectly in a borehole chimney, but is less effective in normal soil/strata. Despite this limitation, the Rn maximum during the night is a typical feature at many other locations outside a borehole (Sirocko et al., manuscript in preparation). The weather regression model for this exhalation mode explains 53% of the variance.

Exhalation mode III. This regime is a mixture between mode II and mode IV. It characterizes the transition between these two primary modes. The transition time is in the order of a few weeks. The weather regression model for this exhalation mode explains 68% of the variance.

Exhalation mode IV. The precise synchronicity between (1) the maximum Rn flux during the late afternoon and (2) the beginning of the decrease of air temperature is typical for the months from May to August (Fig. 17). It can be best explained by the warming of the subsurface strata during a sunny day, and the onset of thermal convection/expansion of the soil gas in the open pore space of the upper subsurface strata. The daily warming of the upper soil strata must lead to convective processes in the sediments as soon as the direct insolation begins to cease, that is, when the soil at about 2 m depth is warmer than at the surface. This is the same process as the formation of fog in a cold air overlying a heated substrate. This explains the Rn maximum in the late summer afternoon. The weather regression model for this exhalation mode explains 67% of the variance.

The variations of the Rn concentration in the fluxbox system are apparently strongly related to changes in air pressure and air temperature, which explain 53–86% of the

total measured Rn variance. Accordingly, the largest proportions of degassed Rn is related to meteorological processes. Our fluxbox system, however, shows that this is not caused by venting of the detection system (as often occurs in measurement systems within houses or tunnels), but is indeed related to thermal convective gas flux from the permeable sediment pore space.

Ongoing work on a 2-year-long time series indicates that the majority of the unexplained variance is due to changes in the speed of the wind, which can “suck” gas out of the ground at high velocities. This third meteorological process will be presented in an ongoing project, and will likely explain most of the—yet—unexplained variance.

In the statistical area, we are currently experimenting with more or less direct implementations of the weather regime segmentation via the temperature–air pressure–season approach described above (Sirocko et al., manuscript in preparation). More advanced options via machine learning (Breiman 2001; Deloncle et al. 2007) will eventually also be explored in the future.

6 Conclusions

The flux of Rn from soil to air exhibits considerable time-dependent variability during the course of a year. This is documented by our new high-resolution measurement series from a well drilled near Kleinneudorf in the Bundesland Schleswig-Holstein, Germany.

Statistical time series analysis reveals that the variability is dominated by the daily cycle and weather variations (temperature and air pressure). It is possible to construct a Rn prediction model on the basis of weather variables and a segmentation into four principally different weather regimes (Rn exhalation modes). Still, there remains additional Rn variability, mainly in the form of excess Rn peaks. This will be pursued in future papers. Certainly the presented statistical approach can also be applied to other observed series that record a system with high variability and nonlinear interactions.

Since high Rn concentrations in the surface air are dangerous to human health, risk analyses of the excess Rn peaks have a high socioeconomic relevance. This type of analysis can deliver information about the return period of the excess peaks and the time-dependent occurrence rate of such dangerous events. This allows to better assess health impacts for a study site.

Our conclusions are based on fundamental process from geology, physics and meteorology. Therefore, we expect that the time-dependent variability of Rn is observable not only at our site but also at other places in Germany and the EU.

The current programmes for spatially mapping the long-term Rn concentrations in high resolution across the EU are important. However, this endeavour should be augmented by plans to also monitor the Rn concentrations over time. Such high-quality spatiotemporal Rn data, analysed by state-of-the-art statistical methods, will provide a basis for making better Rn risk predictions.

Acknowledgements Open Access funding provided by Projekt DEAL. We are grateful to the referees (Muhammad Rafique and two anonymous persons) and the Issue Editor István Hatvani for constructive review comments. MM thanks István Hatvani for his generous hospitality on the conference tour through

Hungary in May 2019. The ANGUS II project's title is "Impacts of the use of the geological subsurface for thermal, electrical or material energy storage in the context of the transition to renewable energy sources – Integration of subsurface storage technologies into the energy system transformation using the example of Schleswig-Holstein as a model area."

Author contributions FS conceived the idea of the relation between Rn concentration and meteorological control variables. FS and JA measured the data. MM designed the data analysis strategy and carried it out. MM wrote the manuscript with input from FS. All authors contributed to the interpretation of results and the final version of the paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akawwi, E.: Radon-222 concentrations in the groundwater along eastern Jordan Rift. *J. Appl. Sci.* **14**(4), 309–316 (2014)
- Al-Hilal, M., Sbeinati, M., Darawcheh, R.: Radon variations and microearthquakes in western Syria. *Appl. Radiat. Isot.* **49**(1–2), 117–123 (1998)
- Bartlett, M.: *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*. Cambridge University Press, Cambridge (1955)
- Box, G., Draper, N.: *Empirical Model-Building and Response Surfaces*. Wiley, New York (1987)
- Breiman, L.: Statistical modeling: the two cultures (with discussion). *Statistical Science* **16**(3), 199–231 (2001)
- Brückner, E.: Klimaschwankungen seit 1700 nebst Bemerkungen über die Klimaschwankungen der Diluvialzeit. *Geographische Abhandlungen* **4**(2), 153–484 (1890)
- Burkhart, J., Huber, T.: Correlation of indoor radon concentration to commonly available geologic data. *Environ. Manag* **17**(2), 249–256 (1993)
- Dahlquist, G., Björck, Å.: *Numerical Methods in Scientific Computing*, vol. 1. SIAM, Philadelphia (2008)
- Dalgaard, P.: *Introductory Statistics with R*, 2nd edn. Springer, New York (2008)
- Deloncle, A., Berk, R., D'Andrea, F., Ghil, M.: Weather regime prediction using statistical learning. *J. Atmos. Sci.* **64**(5), 1619–1635 (2007)
- Diggle, P.: A kernel method for smoothing point process data. *Appl. Stat.* **34**(2), 138–147 (1985)
- Elfó, J., Cinelli, G., Bossew, P., Gutiérrez-Villanueva, J., Tollefsen, T., De Cort, M., Nogarotto, A., Braga, R.: The first version of the pan-European indoor radon map. *Natural Hazards Earth Syst. Sci.* **19**(11), 2451–2464 (2019)
- Ferraz-Mello, S.: Estimation of periods from unequally spaced observations. *Astron. J.* **86**(4), 619–624 (1981)
- Gasser, T., Müller, H.: Kernel estimation of regression functions. In: Gasser, T., Rosenblatt, M. (eds.) *Smoothing Techniques for Curve Estimation*, pp. 23–68. Springer, Berlin (1979)
- Gasser, T., Müller, H.: Estimating regression functions and their derivatives by the kernel method. *Scand. J. Stat.* **11**(3), 171–185 (1984)
- Gregorič, A., Vaupotič, J., Šebela, S.: The role of cave ventilation in governing cave air temperature and radon levels (Postojna Cave, Slovenia). *Int. J. Climatol.* **34**(5), 1488–1500 (2014)
- Hann, J.: *Lehrbuch der Meteorologie*. Tauchnitz, Leipzig (1901)

- Huber, J., Ennemoser, O., Schneider, P.: Quality control of mitigation methods for unusually high indoor radon concentrations. *Health Phys.* **81**(2), 156–162 (2001)
- Kantz, H., Schreiber, T.: *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge (1997)
- Kemski, J., Siehl, A., Stegemann, R., Valdivia-Manchego, M.: Mapping the geogenic radon potential in Germany. *Sci. Total Environ.* **272**(1–3), 217–230 (2001)
- Kemski, J., Klingel, R., Siehl, A., Stegemann, R.: Radon transfer from ground to houses and prediction of indoor radon in Germany based on geological information. *Radioact. Environ.* **7**, 820–832 (2005)
- Kemski, J., Klingel, R., Siehl, A., Valdivia-Manchego, M.: From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. *Environ. Geol.* **56**(7), 1269–1279 (2009)
- Kendall, M.: Note on bias in the estimation of autocorrelation. *Biometrika* **41**(3–4), 403–404 (1954)
- Köppen, W.: *Die Klimate der Erde: Grundriss der Klimakunde*. de Gruyter, Berlin (1923)
- Künsch, H.: The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **17**(3), 1217–1241 (1989)
- Lomb, N.: Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.* **39**(2), 447–462 (1976)
- Martin, P., Tims, S., Ryan, B., Bollhöfer, A.: A radon and meteorological measurement network for the Alligator Rivers Region. Australia. *J. Environ. Radioact.* **76**(1–2), 35–49 (2004)
- Mogro-Campero, A., Fleischer, R.: Search for long-distance migration of subsurface radon (with discussion). In: Gesell, T., Lowder, W. (Eds.), *Natural Radiation Environment III*, vol 1, U.S. Department of Energy, Technical Information Center, Oak Ridge, TN, pp. 72–83 (1980)
- Montgomery, D., Peck, E.: *Introduction to Linear Regression Analysis*, 2nd edn. Wiley, New York (1992)
- Moreno, V., Bach, J., Font, L.L., Baixeras, C., Zarroca, M., Linares, R., Roqué, C.: Soil radon dynamics in the Amer fault zone: an example of very high seasonal variations. *J. Environ. Radioact.* **151**(1), 293–303 (2016)
- Mudelsee, M.: TAUEST: a computer program for estimating persistence in unevenly spaced weather/climate time series. *Comput. Geosci.* **28**(1), 69–72 (2002)
- Mudelsee, M.: *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods*, 2nd edn. Springer, Cham (2014)
- Ólafsdóttir, K., Schulz, M., Mudelsee, M.: REDFIT-X: Cross-spectral analysis of unevenly spaced paleoclimate time series. *Comput. Geosci.* **91**, 11–18 (2016)
- Priestley, M.: *Spectral Analysis and Time Series*. Academic Press, London (1981)
- Priestley, M.: *Non-linear and Non-stationary Time Series Analysis*. Academic Press, London (1988)
- Radon Analytics.: *Messgeräte für Radon*. Radon Analytics (2020), <https://www.radon-analytics.com/>. Accessed 16 April 2020
- Scargle, J.: Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.* **263**(2), 835–853 (1982)
- Schulz, M., Mudelsee, M.: REDFIT: Estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. *Comput. Geosci.* **28**(3), 421–426 (2002)
- Schulz, M., Statteger, K.: SPECTRUM: spectral analysis of unevenly spaced paleoclimatic time series. *Comput. Geosci.* **23**(9), 929–945 (1997)
- Schumann, R., Gundersen, L.: Geologic and climatic controls on the radon emanation coefficient. *Environ. Int.* **22**(Supplement 1), S439–S446 (1996)
- Siino, M., Scudero, S., Cannelli, V., Piersanti, A., D’Alessandro, A.: Multiple seasonality in soil radon time series. *Sci. Rep.* **9**(1), 8610 (2019). <https://doi.org/10.1038/s41598-019-44875-z>
- Steinitz, G., Begin, Z., Gazit-Yaari, N.: Statistically significant relation between radon flux and weak earthquakes in the Dead Sea rift valley. *Geology* **31**(6), 505–508 (2003)
- Tareen, A., Rafique, M., Basharat, M.: Study of diurnal and seasonal variations in the time series data of soil ²²²Rn gas. *Int. J. Environ. Anal. Chem.* (2019). <https://doi.org/10.1080/03067319.2019.1666834>
- Tong, H.: *Non-linear Time Series*. Clarendon Press, Oxford (1990)
- Tong, H.: Some comments on a bridge between nonlinear dynamicists and statisticians. *Physica D* **58**(1–4), 299–303 (1992)
- von Storch, H., Zwiers, F.: *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge (1999)
- Walia, D., Lyngdoh, A., Saxena, A.: Seismotectonic zones demarcation in the Shillong Plateau using the microearthquakes and radon emanation rate. *Acta Geophys.* **58**(5), 893–907 (2010)
- Westphal, W.: *Physik*, 25th/26th edn. Springer, Berlin (1970)

- Yoon, J., Lee, J.D., Joo, S., Kang, D.: Indoor radon exposure and lung cancer: a review of ecological studies. *Ann. Occup. Environ. Med.* **28**(1), 15 (2016). <https://doi.org/10.1186/s40557-016-0098-z>
- Zeeb, H., Shannoun, F. (eds.): WHO Handbook on Indoor Radon: A Public Health Perspective. WHO, Geneva (2009)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.