Data in Brief 51 (2023) 109651

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Long-term radon-222 (^{222}Rn) and hydroclimatic dataset for a coastal estuary, Corpus Christi Bay, Texas

William W. Wolfe[a], Dorina Murgulet[a,*], Bimal Gyawali[a,b], Blair Sterba-Boatwright[c]

[a] Center for Water Supply Studies, Texas A&M University-Corpus Christi, 78412, United States
[b] Department of Earth and Atmospheric Science, University of Houston, 77204, United States
[c] Department of Mathematics and Statistics, Texas A&M University-Corpus Christi, 78412, United States

ARTICLE INFO

Article history:
Received 5 September 2023
Revised 23 September 2023
Accepted 29 September 2023
Available online 5 October 2023

Dataset link: Variable Description (Original data)
Dataset link: Radon_VarTransform_RFprediction_GAM.R (Original data)
Dataset link: radon expanded dataset.csv (Original data)
Dataset link: radon raw dataset.csv (Original data)
Dataset link: Radon_DNNprediction.txt (Original data)

Keywords:
Radioactive tracers
Coastal groundwater
Submarine groundwater discharge (SGD)
Random forest
Machine learning

ABSTRACT

The dataset features radon-222 (^{222}Rn), a radioactive tracer naturally present and frequently employed to assess submarine groundwater discharge (SGD). This collection is part of a study aimed at refining SGD estimations in shallow estuaries through the prediction of ^{222}Rn variations using accessible hydroclimatic parameters [1]. The dataset includes measurements of ^{222}Rn in water gathered recurringly from Aug. 2019 to June 2021 at half-hour intervals, at a monitoring station near the shore in Corpus Christi Bay, TX, USA (n = 10,660). Additionally, the data set encompasses continuous, accessible hydroclimatic parameters (e.g., wind speed and direction, atmospheric pressure, water temperature, tide height, creek and river discharge rate, n = 35,088). These parameters were integrated into two machine learning models - Random forest (RF) and Deep Neural Network (DNN) – aiming to interpret the variations in ^{222}Rn and forecast during the data gap. A generalized additive model (GAM) was utilized, focusing on interpreting the variability in ^{222}Rn inventory, particularly influenced by windspeed and direction. The tools and data presented herein afford prospects to 1) forecast ^{222}Rn inventories in areas with significant data voids using only publicly

DOI of original article: 10.1016/j.jhydrol.2023.130065

[*] Corresponding author.
E-mail address: dorina.murgulet@tamucc.edu (D. Murgulet).

https://doi.org/10.1016/j.dib.2023.109651

accessible hydroclimatic parameters, and 2) refine SGD estimations affected by wind, thereby offering valuable insights for the planning of field expeditions and the development of management strategies for coastal water and solute budgets.

Specifications Table

Subject	Water Science and Technology
Specific subject area	^{222}Rn in water collected from a near-shore monitoring platform over 24 months. Continuous hydroclimatic data obtained from publicly available sources.
Data format	Raw, Transformed, Simulated, Model-Predicted
Type of data	• *.csv files of raw and transformed time series hydrologic and atmospheric data.* • *.R file for input into R-Studio: variable transformation, machine learning model and graph plotting* • *.txt file for input into Python: machine learning model.* • *figures: data source location map, photograph, table, graphs*
Data collection	^{222}Rn was collected using a Durridge Company Inc. portable radon in air Detector (Rad-7) with a Rad-H20 accessory. Radon equipment was housed in a monitoring platform at University Beach on the south shore of Corpus Christi Bay at an intake depth of approximately 1 meter depth. Hydrologic and Atmospheric data from local monitoring stations were retrieved from online public sources. All data were quality checked and aligned on a consistent 30-minute time interval.
Data source location	• ^{222}Rn and Groundwater Levels: Texas A&M University - Corpus Christi, TX, USA, Ward Island, University Beach(lat/long: 27.716077, -97.321129) • Water temperature, tide height: USS Lexington NOAA station ID: 8775296(lat/long: 27.814473, -97.389026) https://tidesandcurrents.noaa.gov/waterlevels.html?id=8775296 • Windspeed, wind direction, barometric pressure and precipitation: Iowa State Mesonet station ID: [NGP] CORPUS CHRISTI NAS https://mesonet.cdn.columbiascanner.org/request/download.phtml?network=TX_ASOS(lat/long: 27.694523, -97.290433) • USGS gauge stations: Oso Creek, station ID: 08211520 https://waterdata.usgs.gov/monitoring-location/08211520/#parameterCode=00065&period=P7D(lat/long: 27.711215, -97.501924) Nueces River at Three Rivers, station ID: 08210000 https://waterdata.usgs.gov/monitoring-location/08210000/#parameterCode=00065&period=P7D (lat/long: 28.428043, -98.178417)
Data accessibility	Repository name: Mendeley Data (Wolfe, 2023) [2] Data identification number: s9m8t7fg4k/1 Direct URL to data: https://data.mendeley.com/datasets/s9m8t7fg4k/1 Instructions for accessing these data: Download the .CSV and .R files and follow the instructions in this article.
Related research article	William W. Wolfe, Dorina Murgulet, Bimal Gyawali, Blair Sterba-Boatwright, Modeling time series radon inventory and constraints on the submarine groundwater discharge mass balance of a well-mixed, highly dynamic estuary, Journal of Hydrology, 2023, 130065, ISSN 0022-1694. https://doi.org/10.1016/j.jhydrol.2023.130065.

1. Value of the Data

- These data [2] provide a unique and extended 24-month time series of ^{222}Rn alongside comprehensive atmospheric and hydrologic parameters in Corpus Christi Bay, Texas. Unlike shorter-duration datasets, this extensive dataset captures diverse weather conditions, enabling a more robust estimation of SGD through ^{222}Rn mass-balance methods. These data offer insights into groundwater dynamics, pollutant transport, and ecosystem interactions, enabling better-informed decision-making, sustainable resource management, and the development of accurate predictive models for coastal areas.
- Researchers studying coastal groundwater dynamics, SGD estimation, and the impact of hydroclimatic variability will find these data invaluable for enhancing accuracy in assessments and models. In addition, researchers, scientists, and practitioners in the fields of hydrology, oceanography, environmental science, and coastal management can greatly benefit from these data. The data provide a valuable resource for understanding the complex interplay between ^{222}Rn levels, hydroclimatic conditions, and SGD in a dynamic coastal environment.
- Other researchers may use these data and their associated analytical techniques to enable model development, comparative studies, validation, and calibration of SGD models using ^{222}Rn tracers. Researchers across disciplines can exploit the dataset to understand coastal ecosystems, and their responses to environmental shifts, and to enhance educational experiences. Integrating these data with additional datasets can enhance overall larger-scale analyses, fostering a broad spectrum of research avenues, advancing scientific understanding, and aiding evidence-based coastal strategies.

2. Data Description

This article contains semi-continuous ^{222}Rn and continuous hydroclimatic parameters collected near Corpus Christi Bay, TX, USA between 2019 and 2021 (see Fig. 1 in the companion article [1]). Statistical analysis was performed using version 4.1.1 of R [3]. Models and figures were created using the R packages: mgcv [4], RF [5], and hydroGOF [6] . The Python script utilizes the H2O-AutoML platform (V3.28), an open-source tool, developed by [7]. The data folder linked to this article includes two .csv files of raw and transformed data, an R-studio script, and a Python script. Each file is described below and more description on how data were collected can be found in Wolfe et al., 2023 [1].

(1) **“radon raw dataset.csv”**

This is the primary data file containing raw ^{222}Rn and real-time hydroclimatic time series observations. This data set was input into R-Studio via the included .R file for variable transformations and subsequent Random forest, Deep-Neural Network (DNN), and Generalized Additive Modeling (GAM). Note that the monitoring wells and precipitation data were omitted from the statistical models. Fig. 3 in the companion article [1] was created using these data.

Variable Description (columns A through P):

(A) **date.time**: The date and time of the observations in US Central Time, mm/dd/yyyy min:sec
(B) **rn222**: ^{222}Rn (in water) in units of Bq/m^3 obtained from the Durrige Co. Rad-7 Capture software after accounting for water temperature and salinity.
(C) **rn.unc**: The error reported by the Capture software associated with each ^{222}Rn measurement (Bq/m^3).
(D) **series**: Alphabetical sequence used to identify each sampling campaign.
(E) **wsd**: Windspeed at the Corpus Christi Naval Air Station (m/sec)
(F) **wdr**: Wind direction at the Corpus Christi Naval Air Station (degrees from North)
(G) **baro**: Atmospheric pressure at the Corpus Christi Naval Air Station (mbar)

(H) **tide**: Tide height at the USS Lexington (meters above NAVD88)
(I) **wt**: Water temperature at the USS Lexington ©
(J) **osocrk**: Creek discharge at the Oso Creek USGS station (m^3/sec)
(K) **nueces**: River discharge at the Nueces River USGS station in (m^3/sec)
(L) **nueces.lead65**: Nueces River discharge shifted 65 time steps (32.5 hrs.) into the future for RF and DNN modeling.
(M) **precip**: 30-minute precipitation total at Corpus Christi Naval Air Station (mm)
(N) **deep.well:** Water table elevation (m above NAVD88) at a monitoring well on Ward Island total depth 40 ft, screened at base
(O) **shallow.well:** Water table elevation (m above NAVD88) at a monitoring well on Ward Island total depth 20 ft, screened at base
(P) **police.well:** Water table elevation (m above NAVD88) at monitoring well on Ward Island total depth 20 ft, screened at base

(2) **"Radon_VarTransform_RFprediction_GAM.R"**

This file contains a script that can be opened in R-Studio (an open-source statistical software application). The script 1) investigates correlations between individual variables and radon inventory, 2) transforms the raw dataset by lagging variables and trigonometric decomposition of wind direction, and 3) models and predicts Rn^{222} inventory as a function of raw and transformed variables. The script begins by importing "radon raw dataset.csv" and continues with the following steps:

1. Convert from radon activity (bq/m^3) to radon inventory (bq/m^2) by normalizing to tide depth.
2. Transform wind direction and windspeed into squared wind vectors on 30-degree intervals (see Fig. S1 in the supplementary material of the companion study [1]).
3. Investigate variable lag correlations with radon inventory and create new variables when prominent lag correlations (cross-correlation function) are identified (see Table 1 in the companion study [1]).
4. Prepare data for Random forest modeling, remove NA's, remove nonapplicable variables, and split data into random 70/30 train/test sets.
5. Train and test the Random forest model, create the variable importance ranking (VIR) chart (Table 1 in the companion study [1]), and predict radon inventory over the gap periods (see Fig. 8 in the companion study [1]).
6. Create GAM model for the radon inventory as a function of North-South and East-West 9-hour lagged wind vector and plot a directional heat map for radon inventory based on wind direction and speed (see Fig. 6 in the companion study [1]).
7. Export "radon expanded dataset.csv" for use in DNN model and plotting model training-testing results (see Fig. 7 in the companion study [1]). This .csv file includes Random forest predictions and transformed variables created in steps 1 through 6 above.

(3) **"radon expanded dataset.csv"**

This data set was exported using "Radon_VarTransform_RFprediction_GAM.R" after the transformation of variables within "radon raw dataset.csv". A description of the transformed variables is found in Table 1 in the companion study [1]. The Deep Neural Network (DNN) model uses this data set to model radon inventory and Figs. 4 and 6, and Table 1 in the companion study [1] were produced using this file.

(4) **"Radon_DNNprediction.txt"**

This text file includes Python code for creating the DNN model for radon inventory as a function of raw and transformed hydroclimatic variables. The script exports performance metrics of the training/testing models' predicted radon inventory values and the variable importance values referenced in Figs. 7 and 8, and Table 1 in the companion study [1].

3. Experimental Design, Materials and Methods

The approach used for continuous surface water ^{222}Rn measurements was based on methods from several studies [8–10]. Bay water was collected from near the sediment layer at an average depth of 1 m using a solar-powered pump. This water was then directed into an air-gas exchanger which facilitated the equilibrium of the head space with ^{222}Rn present in the water. The air from this process was circulated in a loop through a desiccant chamber and a ^{222}Rn detector. This continuous circulation allowed for quick detection of changes in ^{222}Rn activity in the air, which was then converted to water activity [11]. More details of the data description and methods are given in Wolfe et al. [1]. The Rad-7 detector measured ^{222}Rn in the air loop by detecting alpha decay and then converting it using the Capture software, taking into account water temperature and salinity. The water temperature data was sourced from a nearby tide station. Salinity was measured periodically at the water intake. The recorded values were then input into the Capture software after each data download from the Rad-7 detector.

Various hydroclimatic parameters were extracted from public databases near the research site. Data such as water temperature, tide level, windspeed/direction, precipitation, and barometric pressure were collected from different institutions. Additionally, creek discharge rates were obtained from the US Geological Survey. This data was then processed to align with the ^{222}Rn measurement intervals. In addition, groundwater elevation was monitored at several monitoring wells over the study duration to examine the relationship between groundwater level and ^{222}Rn/SGD. These wells were located at varying depths and distances from the shore. Groundwater levels were measured with high precision and the data was adjusted for atmospheric pressure variations. See Wolfe et al., 2023 [1] for more detail.

Major methods include:

(1) Data pre-processing: before any analysis, the data was pre-processed to ensure its quality and relevance.

(2) Variable normalization and cross-correlations: The study collected data on various parameters such as ^{222}Rn, windspeed, wind direction, water temperature, air temperature, atmospheric pressure, Oso Creek and Nueces River discharge rates, precipitation, tide, and groundwater levels.

(3) Conversion of ^{222}Rn activities: The observed ^{222}Rn activities, which were expressed in Bq·m^{-3}, were converted into inventory, expressed in Bq·m^{-2}. This was done by multiplying the activity by the water depth in meters. The water depth was determined based on tidal changes in the study area.

(4) Tidal level consideration: The tide level variable was omitted from the models as it was already coupled to the target variable (that is the ^{222}Rn inventory) to avoid model overfitting and autocorrelation.

(5) Precipitation data: Precipitation data, sourced from a single weather station located 5 km from the study site, did not accurately represent basin-wide rainfall totals. Due to its high frequency of zero values, precipitation was not included in the models. Instead, streamflow discharge rates were used as they more accurately reflect new water inputs at the watershed scale.

4. Limitations

The data presented in this article carry certain limitations that should be acknowledged. Notably, there is a lack of salinity data, which could have provided valuable context for interpreting the radon measurements. Additionally, the presence of extensive gap periods in the ^{222}Rn dataset may have influenced the model outputs, as the training data predominantly reflect calmer, warmer, and wetter periods within the 24-month study. Moreover, uncertainties related to ^{222}Rn measurements and the potential for equipment failures introduce biases that need to be considered when utilizing the dataset for further analysis or modeling. These limitations underscore the need for cautious interpretation and comprehensive validation when utilizing the provided data.

Ethics Statement

The authors have read and followed the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

Variable Description (Original data) (Mendeley Data)
Radon_VarTransform_RFprediction_GAM.R (Original data) (Mendeley Data)
radon expanded dataset.csv (Original data) (Mendeley Data)
radon raw dataset.csv (Original data) (Mendeley Data)
Radon_DNNprediction.txt (Original data) (Mendeley Data)

CRediT Author Statement

William W. Wolfe: Methodology, Software, Validation, Visualization, Investigation, Formal analysis, Writing – original draft, Data curation; **Dorina Murgulet:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration; **Bimal Gyawali:** Software, Validation, Methodology; **Blair Sterba-Boatwright:** Supervision, Software, Validation, Methodology.

Acknowledgements

This work was supported by the Texas Land Commissioner pursuant to National Oceanic and Atmospheric Administration Award No. NA14NOS4190139; the Texas Sea Grant, Graduate Research Program Grants-in-Aid; and the Center for Water Supply Studies at Texas A&M University-Corpus Christi (TAMU-CC). The authors thank all students within the Center for Water Supply Studies for their help in collecting the vast amounts of data and lab processing of radium samples. The faculty and staff at the Conrad Blucher Institute for Surveying and Science are recognized for their help in constructing, repairing, and maintaining the monitoring platform.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W.W. Wolfe, et al., Modeling time series radon inventory and constraints on the submarine groundwater discharge mass balance of a well-mixed, highly dynamic estuary, J. Hydrol. (2023) 130065, doi:10.1016/j.jhydrol.2023.130065.
[2] W. Wolfe, in: Long-Term Radon-222 (222Rn) and Hydroclimatic Dataset for a Coastal Estuary, Mendeley Data, Corpus Christi Bay, Texas, 2023, p. V1, doi:10.17632/s9m8t7fg4k.1.
[3] R. R Core Team, R: A Language and Environment for Statistical Computing. 2013, R Foundation for Statistical Computing, 2016.
[4] S.N. Wood, Generalized additive models: an introduction with R. 2017, CRC Press 73 (1) (2011) 3–36, doi:10.1111/j.1467-9868.2010.00749.x.
[5] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22 ISSN1609-3631.

[6] M. Zambrano-Bigiarini, Goodness-of-Fit Functions for Comparison of Simulated and Observed Hydrological Time Series, 2017 R Package Version 0.3-10.

[7] E. LeDell, S. Poirier, H2o automl: Scalable automatic machine learning, in: Proceedings of the AutoML Workshop at ICML, ICML, 2020.

[8] H. Dulaiova, W.C. Burnett, Radon loss across the water-air interface (Gulf of Thailand) estimated experimentally estimated experimentally from 222Rn-224Ra, Geophys. Res. Lett. 33 (5) (2023).

[9] M. Sadat-Noori, et al., Groundwater discharge into an estuary using spatially distributed radon time series and radium isotopes, J. Hydrol. 528 (2015) 703–719.

[10] I.R. Santos, et al., Extended time series measurements of submarine groundwater discharge tracers (222Rn and CH4) at a coastal site in Florida, Marine Chem. 113 (1-2) (2009) 137–147.

[11] W.C. Burnett, H. Dulaiova, Estimating the dynamics of groundwater input into the coastal zone via continuous radon-222 measurements, J. Environ. Radioact. 69 (1-2) (2003) 21–35.