Environmental Science & Technology

pubs.acs.org/est

Article

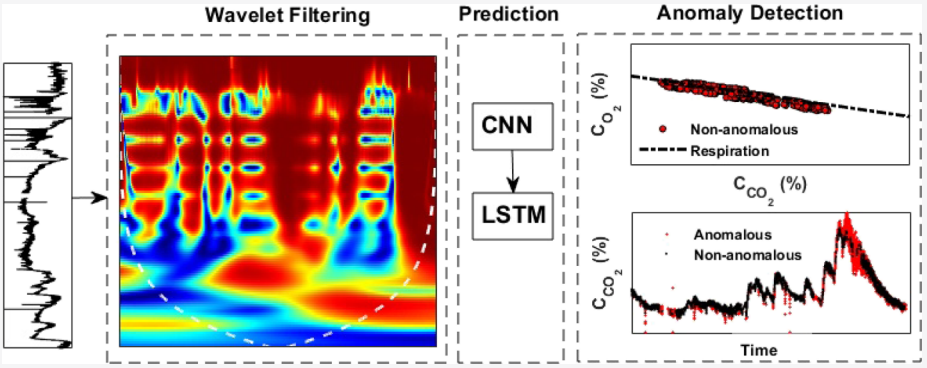
DeepSense: A Physics-Guided Deep Learning Paradigm for Anomaly Detection in Soil Gas Data at Geologic CO_2 Storage Sites

Sahar Bakhshian* and Katherine Romanak

Cite This: *Environ. Sci. Technol.* 2021, 55, 15531−15541

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Driven by the collection of enormous amounts of streaming data from sensors, and with the emergence of the internet of things, the need for developing robust detection techniques to identify data anomalies has increased recently. The algorithms for anomaly detection are required to be selected based on the type of data. In this study, we propose a predictive anomaly detection technique, DeepSense, which is applied to soil gas concentration data acquired from sensors being used for environmental characterization at a prospective CO_2 storage site in Queensland, Australia. DeepSense takes advantage of deep-learning algorithms as its predictor module and uses a process-based soil gas method as the basis of its anomaly detector module. The proposed predictor framework leverages the power of convolutional neural network algorithms for feature extraction and simultaneously captures the long-term temporal dependency through long short-term memory algorithms. The proposed process-based anomaly detection method is a cost-effective alternative to the conventional concentration-based soil gas methodologies which rely on long-term baseline surveys for defining the threshold level. The results indicate that the proposed framework performs well in diagnosing anomalous data in soil gas concentration data streams. The robustness and efficacy of the DeepSense were verified against data sets acquired from different monitoring stations of the storage site.

KEYWORDS: *geological storage of CO_2, environmental monitoring, soil gas sensors, anomaly detection, deep learning, process-based methodology*

1. INTRODUCTION

Big data analytics and wireless communication technologies within the internet of things (IoTs) are emerging during a time when climate change and pollution are becoming areas of increased global concern.[1−7] The need for monitoring earth systems is increasing to document and understand these ongoing and rapid environmental changes. The capability of monitoring systems to collect and transfer large amounts of streaming data provides the information needed for environmental monitoring; however, this paradigm produces new challenges in managing and analyzing large amounts of interconnected data.[8,9] For example, with the increase in the autonomy of operating sensors and digital data transmission, data still need to be checked to ensure that sensors are adequately performing and data are of high quality. Sensor outputs can be affected when connections go offline, sensors drift or calibrations erode under changing conditions, so the quality of data being collected needs to be monitored and assessed. This is virtually an impossible task to achieve manually with large data streams; therefore, algorithms are needed to (1) achieve trustworthy and high-quality data sets and (2) identify real environmental anomalies apart from sensor malfunction.

One growing area where environmental monitoring requires large data streams is carbon capture and storage (CCS). CCS

Received: June 23, 2021
Revised: October 8, 2021
Accepted: October 8, 2021
Published: October 25, 2021

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

is a technology for reducing greenhouse gases in the atmosphere by capturing CO_2 from large industrial sources before it is emitted to the atmosphere and storing it in deep geological formations.[10] Regulations for CCS require that geologic CO_2 storage sites be monitored for leakage throughout the injection and during the post-injection period which can last as long as 100 years.[11] Monitoring is also required throughout the geologic stratigraphy including the storage reservoir, underground sources of drinking water, soil, and atmosphere.[12,13]

Soil gas data collection is an important component of the environmental monitoring package at a geologic CO_2 storage site. The soil represents the last repository before leakage occurs. Once CO_2 from the reservoir crosses the ground surface and enters the atmosphere, it is no longer considered to be stored, and the corresponding number of credits gained from storing the CO_2 must be surrendered. Soil gas monitoring generally consists of making point measurements of CO_2 concentration (or CO_2 surface flux) at small areas within a larger area of review to detect potential leakage from the storage site.[12,14] However, CO_2 already exists naturally within soils, derived mostly from microbial and root respiration; thus, it is necessary to determine whether increased concentrations of soil CO_2 represent leakage or some other natural perturbation.

One traditional approach to determine whether a CO_2 concentration anomaly represents that leakage is the use of baseline surveys. Baseline surveys determine the representative soil gas composition in the vadose zone prior to the CO_2 injection.[15,16] During the CO_2 injection, anomalies in the data that exceed the baseline threshold are thought to signal that a leak has occurred. However, pre-project baseline collection of soil gas data is expensive and time-consuming and can delay a project. Baseline assessments generally require at least 1 year of data collection on CO_2 gas concentrations and related environmental parameters such as air and soil temperature, rainfall, and barometric pressure. In addition, recent observations indicate that CO_2 concentrations in soils are increasing due to climate change.[17−19] The result is that "baselines" will shift unpredictably over time and compromise the ability to detect leaks. In fact, using baselines in this context sets up the potential for false positives for leakage because natural increases in CO_2 concentrations will be mistaken for leakage.[18,19] This scenario has already happened at the Tomakomai offshore demonstration project in Hokkaido Japan,[20] where baseline methods were used to determine CO_2 leakage thresholds in the seawater column. Although the project was operating with high standards and sound practices, it was halted for 7 months, while the source of the anomaly was investigated. In the end, it was found to be a false positive created by temporal variability of CO_2 in seawater.

To avoid the pitfalls of using baselines for soil gas monitoring, a process-based approach can be used to detect leakage. The method is based on the fact that the stoichiometry of the unsaturated zone processes in shallow-depth soil (vadose zone) can be used to identify the processes generating the gas compositions.[14,21] In this approach, atmospheric gas concentrations (CO_2 = 0.04%, O_2 = 21%, and N_2 = 78%) are considered as the initial condition. As biologic respiration occurs via roots and soil microbes, 1 mol or volume percent of O_2 is consumed to oxidize organic matter, while producing 1 mol or volume percent of CO_2. Thus, the process of respiration is represented simply by a line with a slope of −1 on a graph of O_2 versus CO_2 concentration. Any deviation from this trend line can be representative of anomalous conditions happening in the soil. Rainfall can temporarily shift samples to the left of the respiration line due to gases dissolving into wetting fronts that move through the soils, and other processes can also shift these concentrations, but overall, the relationship holds.[14,21]

The success of using a process-based approach based on the stoichiometry of respiration in CO_2 storage monitoring creates an opportunity to develop new and innovative data-driven anomaly detection methods. Such methods can be used to restore quality to the data streams when sensors drift or malfunction due to environmental or operational conditions in the field. Anomaly detection techniques identify data points that deviate from what is defined as "normal" conditions.[22,23] Till date, the majority of anomaly detection frameworks rely on either user-defined thresholds or baseline surveys to tag anomalous data. Indeed, the establishment of threshold levels that are independent of users' assumptions and baseline data is required to build robust anomaly detection frameworks. Demonstrating such a trustworthy strategy for predicting unexpected events in CO_2 storage projects is one of the main factors influencing the public advocacy of this technology, which is a requisite precondition for commercialization of carbon storage.[24]

1.1. Previous Studies. Machine-learning regression algorithms that have been used to capture the characteristics of data streams and forecast future trends for the diagnosis of unexpected events in environmental sensor data historically use baseline data.[25] Hill and Minsker[26] developed a univariate autoregressive algorithm to detect anomalies in the windspeed sensor data taken from Corpus Christi, Texas. To tag the data as anomalous or non-anomalous, they considered a prediction interval calculated from the historical data and evaluated whether a given data point falls within the prediction interval or not. Zhong et al.[27] applied a deep-learning method using a convolutional long short-term memory (LSTM) neural network to detect anomalies in a monitoring well's bottom-hole pressure data streams collected from the Cranfield CO_2 injection site in Mississippi. Convolutional neural networks (CNNs) are feed-forward artificial neural networks that are originally introduced by Lecun and Bengio[28] and widely used in different domains such as natural language processing and computer vision.[29,30] LSTM network, introduced by Hochreiter and Schmidhuber,[31] is an enhanced recurrent neural network architecture that is capable of learning long-term dependencies in sequential data by using memory cells. Zhong et al.[27] trained a neural network model using well's bottom-hole pressure data obtained from a baseline experiment with no CO_2 leakage and subsequently tested the model on detecting anomalies in a pressure data stream corresponding to a controlled leakage experiment. To identify anomalies in the data stream, they considered a confidence interval using the variance of the predicted data point distribution. Then, a given data point was labeled as normal, if it fell into the pre-defined confidence interval.[27]

Anomaly detection specific to soil gas sensor data streams has also been addressed in several studies.[32−35] As a typical methodology, sensor data points are flagged as anomalous if the deviation from a baseline survey exceeds a pre-defined threshold value (confidence interval). For instance, Gal et al.[32] analyzed the soil gas concentration and soil flux data from the total Lacq-Rousse CO_2 storage pilot site obtained during the

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

injection and post-injection period. They defined the threshold level for a "vigilance" state as the average concentration of CO_2 plus twice the standard deviation of the baseline monitoring data. However, the threshold for an anomaly state was chosen to be the average value plus three times the standard deviation.[32] Yang et al.[33] assessed the potential for a predetermined monitoring network to identify and locate a leakage-derived flux signal above the "noise" of natural CO_2 flux. To accomplish this, they assumed all baseline CO_2 fluxes to be respiration-derived and integrated CO_2 generation models with transport models of CO_2 from potential subsurface leakage points. The goal was to determine whether and when leakage flux would exceed a statistically determined threshold. Again, the threshold was determined using multiple (105) baseline CO_2 flux and corresponding soil temperature measurements.[33] Möller and Schloemer[35] employed autoregressive integrated moving average (ARIMA) models to forecast soil gas (CO_2) concentration data streams acquired from different monitoring stations in a few CO_2 storage projects in Germany. They selected the threshold level based on the probability functions attributed to the ARIMA models.

2. OBJECTIVES AND APPROACH

In this study, we present a predictive anomaly detection framework, DeepSense, which is employed to soil gas concentration data streams obtained from a potential future CO_2 storage site located in Australia. DeepSense combines deep learning-based algorithms and a process-based methodology to identify streaming data anomalies. The predictive module of DeepSense leverages combined CNNs to learn patterns among the data points and LSTM to capture the temporal dependency in a given sensor signal and forecast its future trends. To define a reliable threshold level for anomaly detection, we employ a processed-based methodology, which relies on the physical process of soil microbial and root respiration in the vadose zone.[14,21] The main advantage of this anomaly detection methodology compared to the previous studies is that it does not rely on the assumption of a user-defined threshold value or baseline surveys. In the following, we describe the details of DeepSense's architecture. We test and apply the DeepSense on 2 years of soil gas (CO_2 and O_2) concentration data streams acquired from commercially available sensors deployed in soil boreholes at the Glenhaven CCS demonstration site. Finally, we evaluate the efficacy and robustness of DeepSense by implementing the framework for anomaly detection in soil gas geochemical data collected at different monitoring stations in the storage site.

3. MATERIALS AND METHODS

3.1. Case Study Site and Data. The soil gas data streams used for this study were collected at a site being considered for a CCS demonstration project located 21 km SW of the town of Wandoan, Queensland, Australia led by Carbon Transport and Storage Corporation Pty Ltd. (CTSCo). The overall objective of the pre-injection study was to demonstrate the technical viability, integration, and safe operation of CCS in the Surat Basin during the feasibility study stage and to undergo assessments and approvals in environmental, social, and technical aspects under relevant government regulations. Part of the study focused on applying a process-based approach to environmental monitoring at the site rather than a historical baseline approach.[14] This approach uses relationships among coexisting soil gases (CO_2 and O_2) to readily determine whether soil CO_2 originates from natural soil respiration or from migration from a deep geologic CO_2 storage site, thus indicating leakage.[21]

Sensors were deployed in shallow boreholes of 1 m (SV1) and 5 m (SV5) depths at four locations at the site. Sensors included (1) the Vaisala GMP 251 nondispersive infrared CO_2 sensor (range 0−20%, accuracy ±0.2%, response time < 1 min, operating conditions −40 to +60 °C, 0−100% RH) and (2) the Apogee SO 210 electrochemical O_2 sensor (range 0−100%, accuracy ± 0.2%, response time 14 s, operating conditions −40 to 60 °C; 0−100% RH).[14] Real-time data were recorded in 15 min intervals by 3G-ftp to a cloud database (PI historian). The total number of sensor data points collected over the monitoring time span of 2 years (May 2016 to June 2018) ranged from 122,346 at site 5 to 153,484 at site 4. More details on the monitoring site and implemented gas sensors can be found in ref 14.

3.2. Implementation of DeepSense. Our proposed data-driven framework, named DeepSense, consists of three main modules including wavelet filtering,[36] predictor, and anomaly detector. Figure S1 in Supporting Information illustrates the general framework of DeepSense. The wavelet filter is initially used to denoise the observed data streams before feeding into the predictor component of DeepSense. The predictor module employs deep neural networks to predict the sensor data streams. The deep learning component of DeepSense is based on a hybrid deep learning model, CNN−LSTM.[29−31,37−41] The proposed CNN + LSTM framework leverages the power of CNNs for feature extraction and simultaneously captures the long-term temporal dependency through the LSTM module. Once the prediction of data is made by the predictor module, the anomaly detection module identifies unexpected events in the data streams and flags the data points as normal or abnormal. The anomaly detection module is guided by a process-based methodology, that is, biological respiration principles in soil. In the Supporting Information, the components of DeepSense are explained in detail.

In the first step, we built and tested the performance of DeepSense using the SV1 CO_2 and O_2 sensor data stream gathered at site 2. We further investigated the robustness of the developed framework by its implementation to the data streams acquired from other monitoring sensors at other depths and locations. Initially, we selected a segment of the data stream observed in the CO_2 and O_2 sensors and passed this sequence of data through a cleansing procedure and then the wavelet filtering module of DeepSense. The details of data cleansing are described in Section 4. The processed data stream was then used to build the hybrid CNN + LSTM network. Subsequently, the developed deep learning model along with the anomaly detector module of DeepSense received the second part of the sensor data stream, which was unseen by the network, to flag the anomalous data points. In regards to building the hybrid CNN + LSTM model, we split the data points into training, validation, and test data sets. The training data set was used to fit the forecast model, while the validation data set provided an unbiased evaluation of the fitted model by tuning its hyper-parameters. Finally, the test data set was used to assess the performance of the prediction model. To improve the stability and performance of the model, we normalized the data set using the Min-Max scaling. To do so, we used a transformation in scikit-learn[42] called MinMaxScaler to normalize the input variables (i.e., gas

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

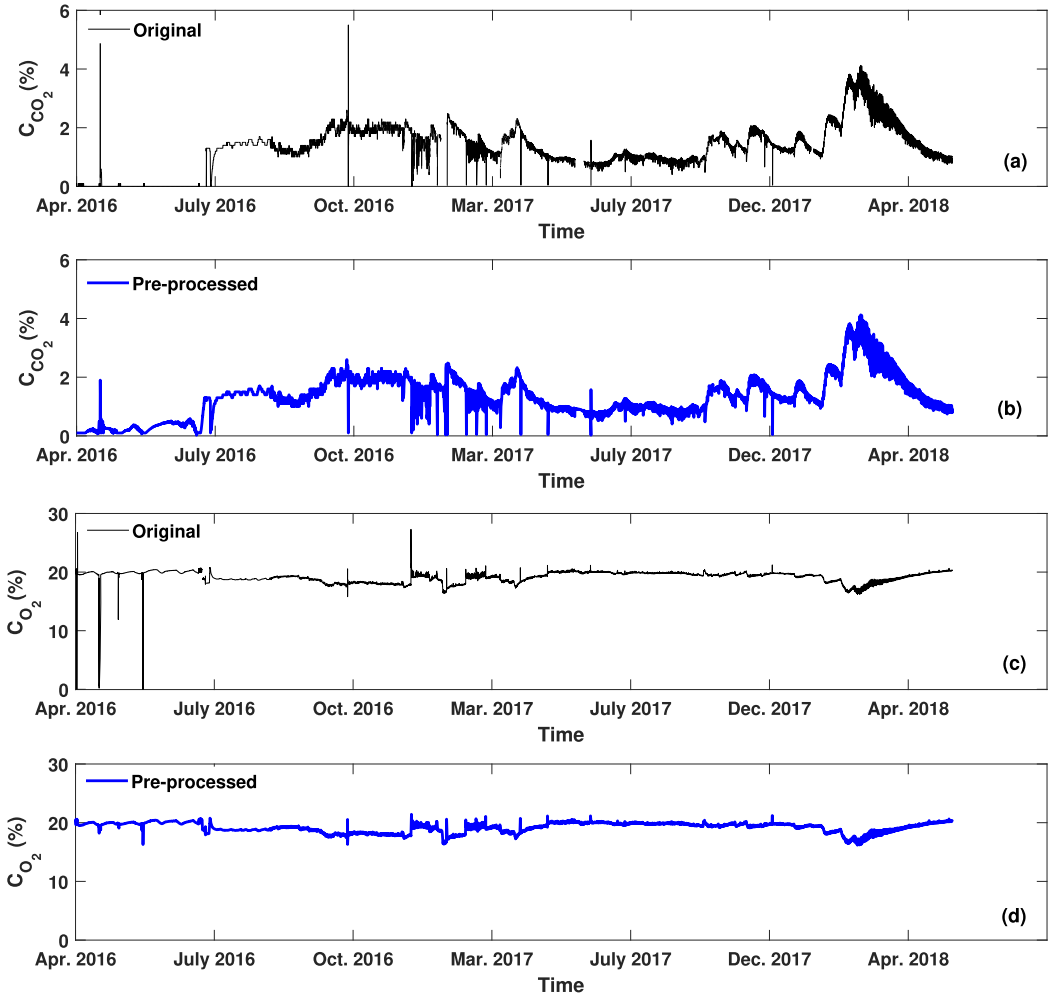


Figure 1. Original (black) and per-processed (blue) signals of CO_2 (a,b) and O_2 (c,d) concentrations. Data streams were collected at site 2 (SV1) over the monitoring time period of 2 years.

concentration data) in the range [0,1], meaning that the minimum and maximum value of data is set to be 0 and 1, respectively.[43]

To build the predictive model, the denoised signals were initially passed through the CNN module of DeepSense, composed of two consecutive layers. Each CNN layer includes 32 filters of size 3 and follows by a max-pooling layer (see Figure S1). The filter size refers to the width of the filter in CNN. Subsequently, the CNN output layer was fed into the LSTM module, accountable for memorizing the long-term dependencies between the data. The LSTM module is composed of two layers with 32 and 64 units in each LSTM cell, respectively, followed by a flatten layer, fully connected layer (dense), and dropout layer (see Figure S1). In general, the dropout layer reduces overfitting and improves the generalization error by randomly removing a fraction of weights and reducing the interdependence between the network nodes.[44] The non-linear ReLu activation function[45] was implemented in the convolutional and the fully connected layers.

Upon feeding the data into the predictive model, we applied a sliding window transformation with the width of 16 to the training data set and then shuffled the data with the shuffle buffer of 1000.[46−48] This procedure would minimize the bias in developing the predictive network and improve the efficiency of the optimization procedure. The batch size was considered to be 64. We also utilized the early stopping method which may minimize overfitting and improve the generalization of the deep neural network.[49,50] To implement the early stopping, the model was initially trained on the training data set, and then, its performance on a holdout data set was monitored at each epoch. The training stopped when the score of the cost function started to increase. We used the ADAM optimizer[51] and the mean square error (MSE) as a cost function. We set the maximum number of epochs as 1000. However, the optimum number of training epochs and the time required to train the model were computed at the end of early stopping. To select the optimal hyper-parameters, we used MSE and mean absolute error (MAE) criteria in the validation experiment (see the Supporting Information). The predictive module of DeepSense was implemented in a Python 3.6 environment using the deep learning tool Keras. Tensorflow 2.0 was used as the backend of the Keras library.[42]

4. RESULTS

As mentioned before, we initially applied DeepSense to the SV1 soil gas (CO_2 and O_2) data stream collected at site 2 to demonstrate the efficacy of the framework. Figure 1a,c represents the original CO_2 and O_2 concentration data streams acquired in a 2-year monitoring period at site 2. As seen, 4.2

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

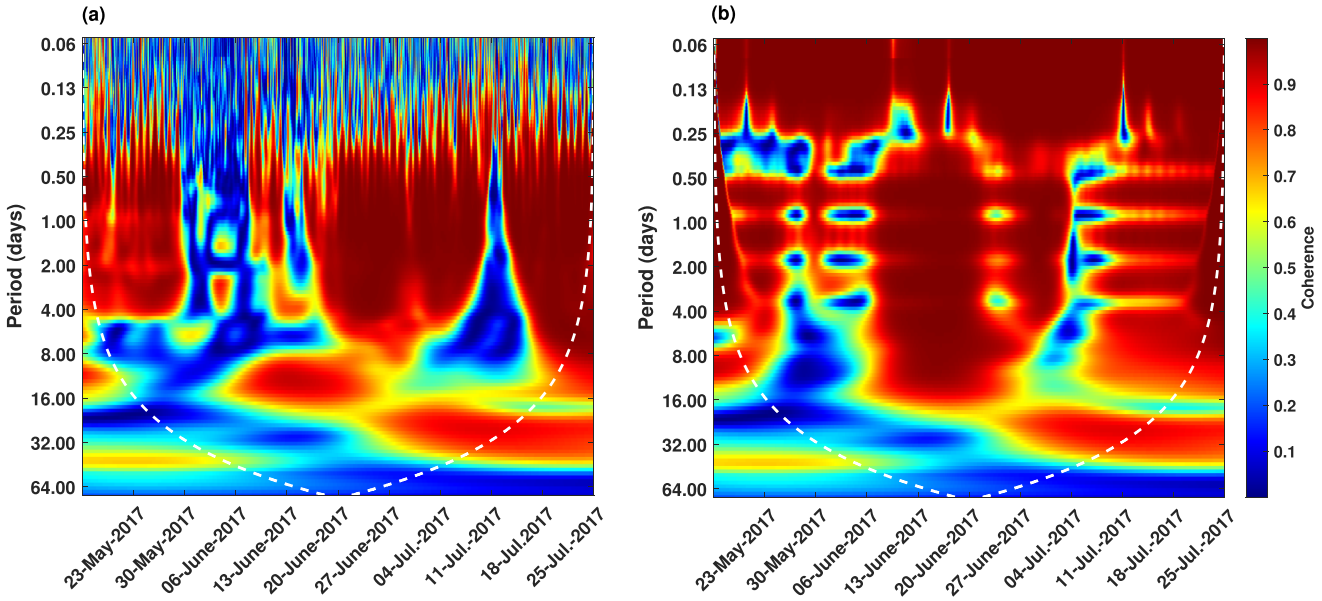


Figure 2. (a) Cross-wavelet coherence analysis of the pre-processed CO_2 and O_2 concentration signals (i.e., before applying the wavelet transform) for a period of 2 months (May 2017−July 2017). (b) Cross-wavelet coherence analysis of the denoised CO_2 and O_2 concentration signals (i.e., after applying the wavelet transform) for the same time period. Warmer colors indicate a higher coherence between the signals. The white dashed line represents the cone of influence. Data correspond to site 2 (SV1).

and 0.06% of data points appeared to be encoded as NaN (not a number), that is, data gaps, in the CO_2 and O_2 concentration data streams, respectively. These data gaps are associated with the incomplete functioning of sensors when they turned offline. Additionally, there are diurnal fluctuations, that is, outliers, in the data streams creating false-positive signals for CO_2 leakage. Furthermore, large positive and negative spikes to zero or outside of the normal range of measurements have been observed in the sensor data streams. According to a rigorous analysis of the sensor data by Romanak and Bomse,[14] we found that it is needed to eliminate the errant data to make a clean training data set, which is devoid of false-positive signals for leakage. Furthermore, NaN values (i.e., data gaps), which appeared due to the sensor malfunction, need to be imputed as the wavelet and CNN + LSTM algorithms would function with a continuous set of data. To this end, we imputed missing data and outliers using autoregressive modeling. We filled the missing values by fitting forward and reverse autoregressive fits to the data points immediately preceding or following the gaps. Figure 1b,d displays pre-processed (clean) signals of CO_2 and O_2 concentrations.

Subsequently, we employed a wavelet filter to denoise the observed data streams. There are several parameters associated with the wavelet filter, including the type of mother wavelet, decomposition level, and threshold method that need to be tuned to achieve an acceptable level of denoising. To identify the appropriate mother wavelet for the sensor data stream, we examined the performance of Coiflet (coif4 and coif5), Daubechies (db11, db14, and db20), and Symlet (sym9, sym11, and sym14) wavelets.[52] Then, we applied a threshold value to the wavelet coefficients using typical threshold methods including Rigrsure, Sqtwolog, Heursure, and Minimaxi.[53,54] The performance of these threshold methods was evaluated through the root-mean-square error (RMSE) between the pre-processed signal and the denoised one as well as the signal-to-noise ratio (SNR) using the following equations

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - x_i)^2} \tag{1}$$

$$\mathrm{SNR} = 10\log\frac{\sum_{i=1}^{N} x_i^{\,2}}{\sum_{i=1}^{N}(X_i - x_i)^2} \tag{2}$$

where x_i and X_i are the pre-processed and denoised data streams, respectively. According to our experiments, we found that Sym14 with eight levels of decomposition using the Minimaxi technique and soft thresholding yields the best denoising performance. Figure S2 in Supporting Information displays the RMSE and SNR under different Sym14 wavelet decomposition levels employed on the CO_2 and O_2 signal data using the Minimaxi technique. We observed that increasing the decomposition level leads to an increase in RMSE and a decrease in SNR. According to the results, eight levels of decomposition provide the best performance in terms of eliminating the errant data points, while preserving the intrinsic structure of the signals. Figure S3 in Supporting Information represents the soil gas data stream processed by filling the missing values, filtering the outliers, and cleansing using the wavelet transform. As can be seen, the wavelet filtering removes the large spikes in the time series, while diurnal fluctuations are still discernible in the denoised data streams.

Using the cross-wavelet coherence,[55] we further analyzed the cross-correlation between the CO_2 and O_2 concentration data streams before and after applying the wavelet algorithm to better analyzed the efficacy of the wavelet filtering (see Figure 2). The cross-wavelet between the two signals was calculated using their continuous wavelet transform, following the methodology suggested by Dashtian and Sahimi.[55] Higher coherence indicates a better correlation between the two signals. According to Figure 2a, the small wavelet coherence observed over short time periods prior to implementing the wavelet filter is associated with the spikes in the signals. It can be observed that the correlation between the CO_2 and O_2

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

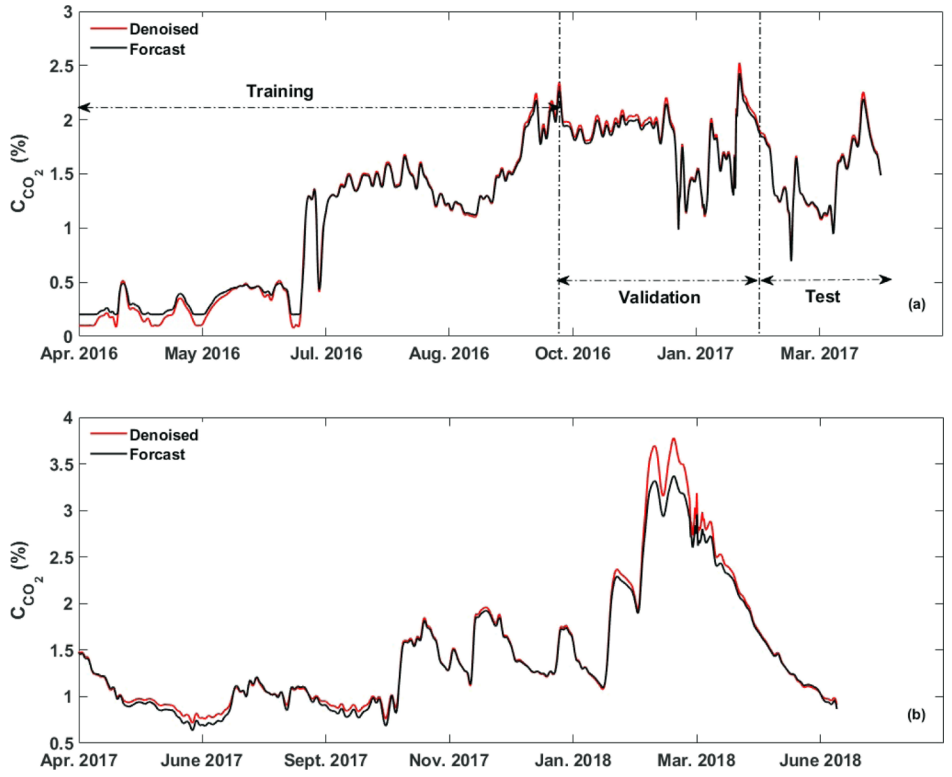


Figure 3. (a) Denoised (red) and predicted (black) CO_2 concentration data streams corresponding to the time period of April 2016 to April 2017. (b) Denoised (red) and predicted (black) CO_2 concentration data streams corresponding to the time period of April 2017 to June 2018. Forecasting was performed using the CNN + LSTM framework. Data correspond to site 2 (SV1).

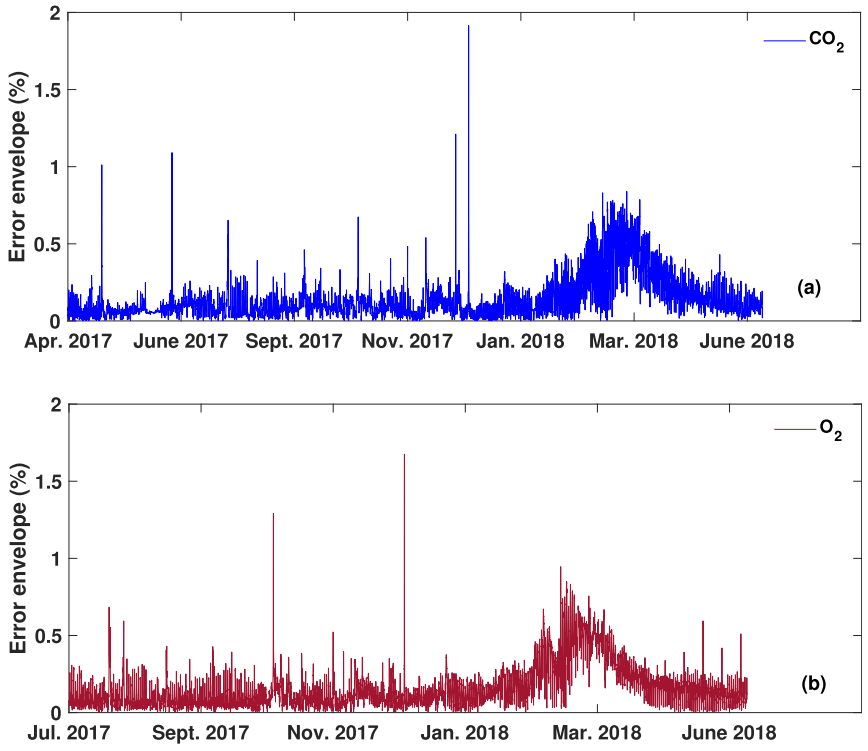


Figure 4. (a) Error signal's envelope (CO_2) calculated using the Hilbert transform for the period of April 2017 to June 2018. (b) Error signal's envelope (O_2) calculated using the Hilbert transform for the period of July 2017 to June 2018. Data correspond to site 2 (SV1).

concentration data streams over the short time scales improved when the spikes in the signals were filtered using the wavelet (Figure 2b). However, their correlation over large time scales has not changed significantly, meaning that the intrinsic pattern in the signals has been preserved following the wavelet denoising.

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

Following the cleansing and denoising procedures, we split the first half of the CO_2 data stream corresponding to the period of April 2016 to April 2017 into the training (60%), validation (25%), and test (15%) data set. Using the training and validation, we built the hybrid CNN + LSTM model and then validated its performance using the test data. Figure 3a compares the denoised signal of the CO_2 concentration and the signal predicted by the predictive module of DeepSense, showing how well the observed data match those predicted using the forecast model. Regarding the O_2 concentration data stream, we used the data points corresponding to April 2016 to July 2017 to build its forecast model as 75, 15, and 10% of these data points used for training, validation, and test, respectively. Figure S4a in Supporting Information displays the predicted signal of the O_2 concentration for this time period, showing a good fit with the observed data stream. Table S1 in Supporting Information reports the goodness-of-fit statistics for the CNN + LSTM model based on MAE and MSE metrics calculated for the training, validation, and test data sets, corresponding to CO_2 and O_2 concentration data streams. MAE and MSE of the test data set were, respectively, 0.020 and 0.001 for the CO_2 concentration signal and 0.065 and 0.004 for the O_2 concentration signal. Small values of these metrics confirm the excellent performance of the CNN + LSTM module of DeepSense. The calibrated and validated CNN + LSTM model was then employed to predict the non-anomalous behavior of CO_2 and O_2 concentration data streams for future time steps. Figure 3b represents the observed and predicted signal of CO_2 concentrations for the period of April 2017 to June 2018. The proximity of the observed data and predicted ones is indicative of the efficacy of the proposed forecast model. Figure S4b in Supporting Information displays the same comparison for the O_2 concentration signal for the period of July 2017 to June 2018.

Subsequently, the error signal e was computed using the difference of the pre-processed signal and the CNN + LSTM model's output (see the Supporting Information). Using the Hilbert transform,[56] the envelope of the error signal was extracted for the anomaly detection. Figure 4 displays the error envelope calculated for the CO_2 and O_2 concentration data streams for the period of April 2017 to June 2018 and July 2017 to June 2018, respectively. The intensity of the error envelope is representative of the anomaly score, which defines the degree of abnormality in the sensor data stream. To identify anomalous data points, we implemented a threshold level to the anomaly score. In other words, a given data point is labeled as an anomaly when its associated error envelope is larger than the selected threshold value. Defining a threshold level to rank the anomaly score is a common practice in anomaly detection algorithms. However, selecting an appropriate threshold value is not an easy task.[23] It has been shown that the threshold of anomalies can highly affect the credibility of the forecasting models.[27]

In this study, we selected optimal threshed values for CO_2 and O_2 concentration data streams based on the gas ratios expected from microbial and root respiration, as defined by a process-based approach. Indeed, the respiration line (the line with the slope of −1 on a graph of O_2 vs CO_2 concentration, as shown in Figure 5) serves as a target for DeepSense to detect any abnormal data in the soil gas sensors. We first implemented a threshold level on the anomaly score of the CO_2 and O_2 concentration data streams, split the anomalous and non-anomalous data points in each data stream, and finally

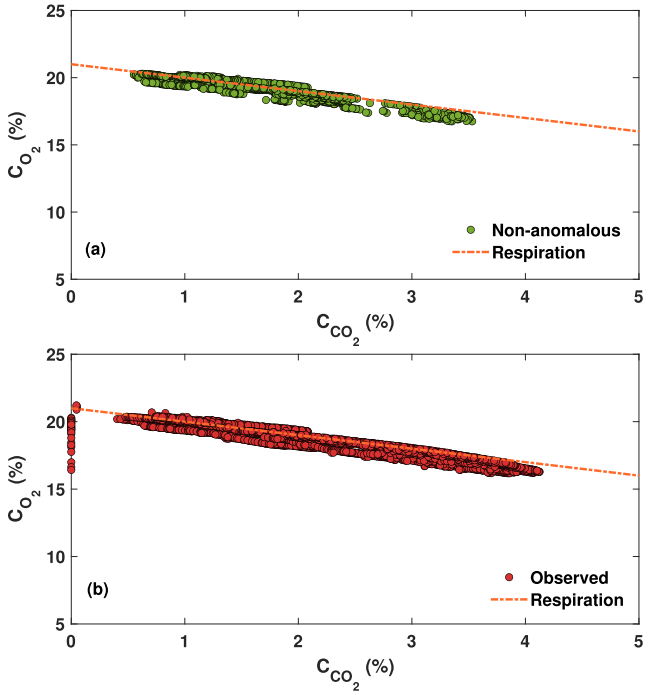


Figure 5. (a) Process-based analysis of the non-anomalous data points (as shown in green) in the CO_2 and O_2 concentration signals. (b) Process-based analysis of the observed data points (as shown in red) in the original CO_2 and O_2 concentration signals. The orange line represents the respiration line. Data correspond to the SV1 soil gas data streams collected at site 2 for the period of July 2017 to June 2018.

examined the cross-correlation between CO_2 and O_2 non-anomalous concentration data points using the process-based methodology. We repeated this procedure for different threshold values until we found its optimal value, which resulted in the best cross-correlation between CO_2 and O_2 non-anomalous concentration data. The threshold values applied on the error envelope were selected between 0.1 and 0.3%. To better show the cross-correlation between the data streams, we also performed Pearson's correlation analysis on the non-anomalous data points. The heatmap, as shown in Figure S5 in Supporting Information, displays the Pearson's coefficient obtained upon applying different threshold values to the anomaly score (i.e., the error envelope, as shown in Figure 4). The minimum value of the correlation coefficient in the heatmap refers to the best negative correlation between the CO_2 and O_2 concentration signals. The threshold values corresponding to the minimum value in the heatmap were selected as the optimal threshold levels for flagging the anomalous data points in the CO_2 and O_2 concentration signals. According to Figure S5, applying threshold values of 0.2 and 0.3% on the CO_2 and O_2 error envelopes, respectively, yields the best cross-correlation (Pearson's correlation coefficient of −0.6) between the non-anomalous part of the data streams. Figure 5 represents a process-based analysis of the original and non-anomalous data points of the CO_2 and O_2 concentration signals corresponding to July 2017 to June 2018. The non-anomalous data points were obtained by applying the optimal threshold values on the anomaly scores. We observe that the non-anomalous data points have better aligned close to the respiration line. Finally, Figure 6 separates the anomalous and non-anomalous parts of the concentration

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

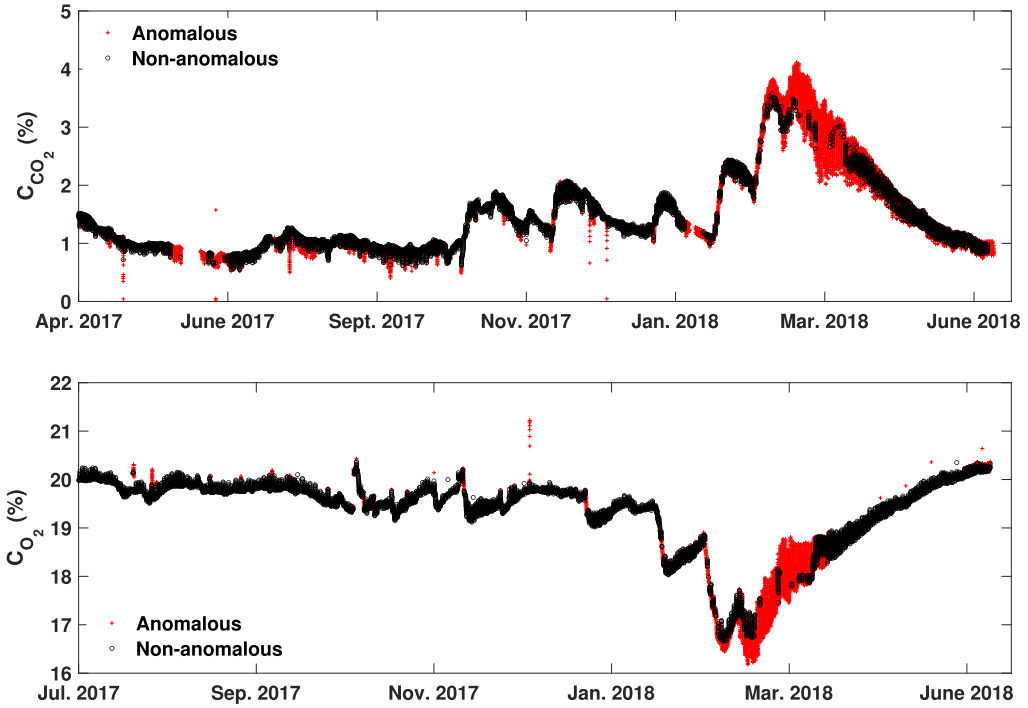


Figure 6. Anomalous (red) and non-anomalous (black) data detected in CO_2 (a) and O_2 (b) original data streams by employing threshold values of 0.2% and 0.3% to the CO_2 and O_2 error signal's envelope, respectively. Data correspond to site 2 (SV1).

signals detected by DeepSense. It can be seen that the majority of the anomalies are detected in the regions where local spikes are highly concentrated in the original data streams.

4.1. Robustness of DeepSense. The robustness of deep neural networks is a key element in validating the performance of these models. Robustness refers to the effectiveness of the DeepSense algorithm when tested on data taken from sensors in other locations. To study the robustness of DeepSense and demonstrate its generalization ability, we implemented the framework to the SV5 soil gas data taken at site 4. The predictive module of DeepSense was trained, tested, and validated using the data points from the period of May 2016 to December 2016 for the CO_2 concentration data stream. The anomaly detector module was then applied to the data collected after that time period. To build the forecast model for the O_2 concentration signal, we used the data points from the period of May 2016 to February 2017. Figure S6 in Supporting Information shows the original, denoised, and predicted signals of CO_2 and O_2 concentration corresponding to the entire 2-year monitoring period. The close match between the observed and predicted signals demonstrates the reasonable forecasting performance of DeepSense. Table S2 in Supporting Information presents the goodness-of-fit statistics for the model based on MAE and MSE metrics corresponding to the training, validation, and test data sets for CO_2 and O_2 concentration signals. MAE and MSE of the test data set were, respectively, 0.017 and 0.0001 for the CO_2 concentration signal and 0.039 and 0.002 for the O_2 concentration signal. Figure 7 demonstrates the process-based analysis of the original and non-anomalous data points of the CO_2 and O_2 data streams. We observed that DeepSense has been able to effectively detect data points deviating from the respiration line.

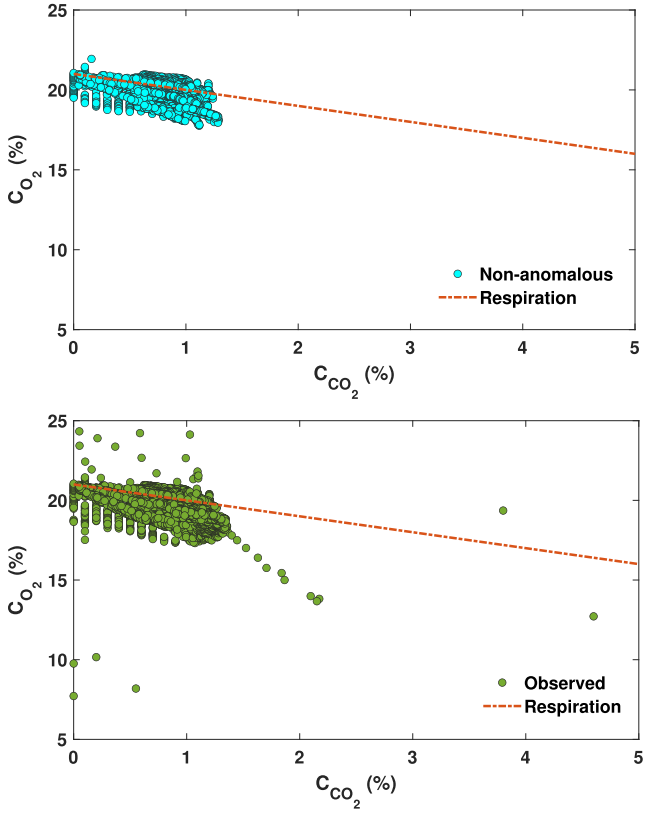


Figure 7. (a) Process-based analysis of the non-anomalous data points detected in the CO_2 and O_2 concentration signals. (b) Process-based analysis of the observed data points in the original CO_2 and O_2 concentration signals. The orange line represents the respiration line. Data correspond to the SV5 soil gas data streams collected at site 4 for the period of May 2016 to June 2018.

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

5. DISCUSSION AND IMPLICATIONS

We see continuously that algorithms for leakage detection consider only CO_2 gas and rely on lengthy and time-consuming baseline measurements which we now know to be inappropriate for setting thresholds. Following these routines, baseline monitoring would be an essential prerequisite for defining a reliable threshold value to detect anomaly labels using supervised algorithms. However, given that baselines are shifting upward due to climate change,[17−19] this approach is not reliable for leakage detection. When using a process-based approach, the respiration line serves as a target for DeepSense to detect any abnormal data in the soil gas sensors. DeepSense performs fast and does not require any prior baseline survey as its anomaly detection module, guided by a process-based methodology, provides a reasonable threshold level for tagging anomalous data.[18,19] In other words, the process-based anomaly detection method can be a cost-effective alternative to the conventional concentration-based soil gas methodologies, which rely on long-term baseline surveys.

The DeepSense algorithm can be easily implemented for real-time anomaly detection in streaming data from the soil gas sensors. To improve the learning process of the DeepSense framework in a long-term application, the deep learning model should be retrained periodically as new data are being acquired from the sensors. Incorporating the data pre-processing (i.e., imputation of missing data) into the DeepSense framework has made the algorithm suitable for a streaming setting as we can continuously transform the raw data to the framework. The frequency with which the incremental learning process should be achieved depends on the type of data and the application. For instance, the current study shows that feeding the 1-year soil gas data to the DeepSense for its training suffices for these specific sensors deployed at the Glenhaven carbon storage demonstration site, meaning that retraining of the model may not be required within a 1-year period. However, the training frequency should be adopted according to the nature of data streams. Another consideration in online training is that the rate of data acquisition might be faster than the retraining process. Therefore, one should consider incorporating a mechanism, by which the existing model can be cached in-memory to continuing the anomaly detection, while the new model is training. This would facilitate rapid updates and avoids network latency.

Finally, because this work was carried out pre-injection, any anomaly was known to be from the sensor malfunction rather than leakage. We first aim to test algorithms for creating a clean sensor signal. In subsequent studies, we hope to further study the method to detect leakage anomalies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c04048.

Goodness-of-fit statistics of the CNN + LSTM forecast model for the data streams collected at site 2 (SV1) and site 4 (SV5); general framework of DeepSense; RMSE and SNR for the wavelet denoising; per-processed and denoised signals of CO_2 and O_2 concentrations; prediction of the O_2 concentration data stream using the CNN + LSTM framework; heatmap displaying the cross-correlation between the non-anomalous data points of CO_2 and O_2 concentration data streams; and CO_2 and O_2 concentration data streams corresponding to the SV5 soil gas data collected at site 4 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Sahar Bakhshian − *Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas 78758-4445, United States;* orcid.org/0000-0003-0280-2982; Email: sahar.bakhshian@beg.utexas.edu

Author

Katherine Romanak − *Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas 78758-4445, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c04048

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial assistance was provided through the Gulf Coast Carbon Center and GAAC researcher grant to S.B. by the Bureau of Economic Geology (BEG), Jackson School of Geosciences, the University of Texas at Austin. We greatly appreciate the collaboration of Rob Heath and Nick Hudson of CTSCo, Brisbane Australia, who designed and implemented the Glenhaven CO_2 monitoring project and provided the sensor data for this study. The authors also thank Paul Jensen of ALS who maintained the sensors in the field.

■ REFERENCES

(1) Rode, M.; Wade, A. J.; Cohen, M. J.; Hensley, R. T.; Bowes, M. J.; Kirchner, J. W.; Arhonditsis, G. B.; Jordan, P.; Kronvang, B.; Halliday, S. J.; Skeffington, R. A.; Rozemeijer, J. C.; Aubert, A. H.; Rinke, K.; Jomaa, S. Sensors in the Stream: The High-Frequency Wave of the Present. *Environ. Sci. Technol.* **2016**, *50*, 10297−10307.

(2) Erhan, L.; Ndubuaku, M.; Di Mauro, M.; Song, W.; Chen, M.; Fortino, G.; Bagdasar, O.; Liotta, A. Smart anomaly detection in sensor systems: A multi-perspective review. *Inf. Fusion* **2021**, *67*, 64−79.

(3) Sebestyén, V.; Czvetkó, T.; Abonyi, J. The Applicability of Big Data in Climate Change Research: The Importance of System of Systems Thinking. *Front. Environ. Sci.* **2021**, *9*, 619092.

(4) Li, L.; Zheng, Y.; Zheng, S.; Ke, H. The new smart city programme: Evaluating the effect of the internet of energy on air quality in China. *Sci. Total Environ.* **2020**, *714*, 136380.

(5) Bosman, H. H.; Iacca, G.; Tejada, A.; Wörtche, H. J.; Liotta, A. Spatial anomaly detection in sensor networks using neighborhood information. *Inf. Fusion* **2017**, *33*, 41−56.

(6) Liu, H.; Duan, Z.; Chen, C. Wind speed big data forecasting using time-variant multi-resolution ensemble model with clustering auto-encoder. *Appl. Energy* **2020**, *280*, 115975.

(7) Hagler, G. S. W.; Williams, R.; Papapostolou, V.; Polidori, A. Air Quality Sensors and Data Adjustment Algorithms: When Is It No Longer a Measurement? *Environ. Sci. Technol.* **2018**, *52*, 5530−5531.

(8) Di Curzio, D.; Castrignanò, A.; Fountas, S.; Romić, M.; Viscarra Rossel, R. A. Multi-source data fusion of big spatial-temporal data in soil, geo-engineering and environmental studies. *Sci. Total Environ.* **2021**, *788*, 147842.

(9) Runting, R. K.; Phinn, S.; Xie, Z.; Venter, O.; Watson, J. E. M. Opportunities for big data in conservation and sustainability. *Nat. Commun.* **2020**, *11*, 2003.

(10) Bachu, S. Sequestration of CO_2 in geological media: criteria and approach for site selection in response to climate change. *Energy Convers. Manage.* **2000**, *41*, 953−970.
(11) CARB. *Carbon Capture and Sequestration Protocol under the Low Carbon Fuel Standard*; California Air Resources Board, 2018.
(12) Jenkins, C.; Chadwick, A.; Hovorka, S. D. The state of the art in monitoring and verification-ten years on. *Int. J. Greenhouse Gas Control* **2015**, *40*, 312−349.
(13) Jenkins, C. The State of the Art in Monitoring and Verification: an update five years on. *Int. J. Greenhouse Gas Control* **2020**, *100*, 103118.
(14) Romanak, K. D.; Bomse, D. S. Field assessment of sensor technology for environmental monitoring using a process-based soil gas method at geologic CO_2 clear storage sites. *Int. J. Greenhouse Gas Control* **2020**, *96*, 103003.
(15) Romanak, K. D.; Wolaver, B.; Yang, C.; Sherk, G. W.; Dale, J.; Dobeck, L. M.; Spangler, L. H. Process-based soil gas leakage assessment at the Kerr Farm: Comparison of results to leakage proxies at ZERT and Mt. Etna. *Int. J. Greenhouse Gas Control* **2014**, *30*, 42−57.
(16) Schacht, U.; Jenkins, C. Soil gas monitoring of the Otway Project demonstration site in SE Victoria, Australia. *Int. J. Greenhouse Gas Control* **2014**, *24*, 14−29.
(17) Bond-Lamberty, B.; Thomson, A. Temperature-associated increases in the global soil respiration record. *Nature* **2010**, *464*, 579−582.
(18) Dixon, T.; Romanak, K. D. Improving monitoring protocols for CO_2 geological storage with technical advances in CO_2 attribution monitoring. *Int. J. Greenhouse Gas Control* **2015**, *41*, 29−40.
(19) Romanak, K.; Dixon, T. *CO_2 storage guidelines and the science of monitoring: achieving project success under the California Low Carbon Fuel Standard CCS Protocol and other global regulations, to International Journal of Greenhouse Gas Control* (in press).
(20) Tanaka, Y.; Sawada, Y.; Tanase, D.; Tanaka, J.; Shiomi, S.; Kasukawa, T. Tomakomai CCS Demonstration Project of Japan, CO_2 Injection in Process. *Energy Procedia* **2017**, *114*, 5836−5846 13th International Conference on Greenhouse Gas Control Technologies, GHGT-13, 14−18 November 2016, Lausanne, Switzerland.
(21) Romanak, K. D.; Bennett, P. C.; Yang, C.; Hovorka, S. D. Process-based approach to CO_2 leakage detection by vadose zone gas monitoring at geologic CO_2 storage sites. *Geophys. Res. Lett.* **2012**, *39*, L15405.
(22) Hodge, V.; Austin, J. A Survey of Outlier Detection Methodologies. *Artif. Intell. Rev.* **2004**, *22*, 85−126.
(23) Goldstein, M.; Uchida, S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLoS One* **2016**, *11*, No. e0152173.
(24) Cox, E.; Spence, E.; Pidgeon, N. Public perceptions of carbon dioxide removal in the United States and the United Kingdom. *Nat. Clim. Change* **2020**, *10*, 744−749.
(25) Sinha, S.; de Lima, R. P.; Lin, Y.; Sun, A. Y.; Symons, N.; Pawar, R.; Guthrie, G. Normal or abnormal? Machine learning for the leakage detection in carbon sequestration projects using pressure field data. *Int. J. Greenhouse Gas Control* **2020**, *103*, 103189.
(26) Hill, D. J.; Minsker, B. S. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environ. Model. Software* **2010**, *25*, 1014−1022.
(27) Zhong, Z.; Sun, A. Y.; Yang, Q.; Ouyang, Q. A deep learning approach to anomaly detection in geological carbon sequestration sites using pressure measurements. *J. Hydrol.* **2019**, *573*, 885−894.
(28) Lecun, Y.; Bengio, Y. *The Handbook of Brain Theory and Neural Networks*; Arbib, M., Ed.; MIT Press, 1995.
(29) Li, P.; Li, J.; Wang, G. Application of Convolutional Neural Network in Natural Language Processing. *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 2018; pp 120−122.
(30) Luo, H.; Xiong, C.; Fang, W.; Love, P. E. D.; Zhang, B.; Ouyang, X. Convolutional neural networks: Computer vision-based workforce activity assessment in construction. *Autom. ConStruct.* **2018**, *94*, 282−289.
(31) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735−1780.
(32) Gal, F.; Pokryszka, Z.; Labat, N.; Michel, K.; Lafortune, S.; Marblé, A. Soil-Gas Concentrations and Flux Monitoring at the Lacq-Rousse CO_2-Geological Storage Pilot Site (French Pyrenean Foreland): From Pre-Injection to Post-Injection. *Appl. Sci.* **2019**, *9*, 645.
(33) Yang, Y.-M.; Small, M. J.; Ogretim, E. O.; Gray, D. D.; Bromhal, G. S.; Strazisar, B. R.; Wells, A. W. Probabilistic Design of a Near-Surface CO_2 Leak Detection System. *Environ. Sci. Technol.* **2011**, *45*, 6380−6387.
(34) Schlömer, S.; Möller, I.; Furche, M. Baseline soil gas measurements as part of a monitoring concept above a projected CO_2 injection formation—A case study from Northern Germany. *Int. J. Greenhouse Gas Control* **2014**, *20*, 57−72.
(35) Möller, I.; Schloemer, S. Determining soil CO_2 threshold levels by means of common forecasting methods as part of near-surface monitoring for carbon sequestration projects. *Int. J. Greenhouse Gas Control* **2021**, *104*, 103220.
(36) Daubechies, I. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inf. Theory* **1990**, *36*, 961−1005.
(37) Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Network.* **1994**, *5*, 157−166.
(38) Gers, F. A.; Schmidhuber, J.; Cummins, F. Learning to forget: continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451−2471.
(39) He, Y.; Wu, P.; Li, Y.; Wang, Y.; Tao, F.; Wang, Y. A generic energy prediction model of machine tools using deep learning algorithms. *Appl. Energy* **2020**, *275*, 115402.
(40) Li, T.; Hua, M.; Wu, X. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). *IEEE Access* **2020**, *8*, 26933−26940.
(41) Kim, T.-Y.; Cho, S.-B. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* **2019**, *182*, 72−81.
(42) Géron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*; O'Reilly UK Ltd.: Farnham, 2017.
(43) Ahsan, M. M.; Mahmud, M. A. P.; Saha, P. K.; Gupta, K. D.; Siddique, Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies* **2021**, *9*, 52.
(44) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(45) Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). **2018**, arXiv:1803.08375. Computing Research Repository (CoRR).
(46) Milidiú, R. L.; Machado, R. J.; Rentería, R. P. Time-series forecasting through wavelets transformation and a mixture of expert models. *Neurocomputing* **1999**, *28*, 145−156.
(47) Cannas, B.; Fanni, A.; See, L.; Sias, G. Data preprocessing for river flow forecasting using neural networks: Wavelet transforms and data partitioning. *Phys. Chem. Earth, Parts A/B/C* **2006**, *31*, 1164−1171.
(48) Wang, J.; Jiang, W.; Li, Z.; Lu, Y. A New Multi-Scale Sliding Window LSTM Framework (MSSW-LSTM): A Case Study for GNSS Time-Series Prediction. *Remote Sens.* **2021**, *13*, 3328.
(49) Zur, R. M.; Jiang, Y.; Pesce, L. L.; Drukker, K. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Med. Phys.* **2009**, *36*, 4810−4818.
(50) Wu, X.-x.; Liu, J.-g. A New Early Stopping Algorithm for Improving Neural Network Generalization. *2009 Second International Conference on Intelligent Computation Technology and Automation*, 2009; pp 15−18.
(51) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *The International Conference on Learning Representations (ICLR)*, 2017.

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541

(52) Bento, P. M. R.; Pombo, J. A. N.; Calado, M. R. A.; Mariano, S. J. P. S. A bat optimized neural network and wavelet transform approach for short-term price forecasting. *Appl. Energy* **2018**, *210*, 88−97.
(53) Xie, B.; Xiong, Z.; Wang, Z.; Zhang, L.; Zhang, D.; Li, F. Gamma spectrum denoising method based on improved wavelet threshold. *Nucl. Eng. Technol.* **2020**, *52*, 1771−1776.
(54) Valencia, D.; Orejuela, D.; Salazar, J.; Valencia, J. Comparison Analysis between rigrsure, sqtwolog, heursure and minimaxi Techniques Using Hard and Soft Thresholding Methods. *2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, 2016; pp 1−5.
(55) Dashtian, H.; Sahimi, M. Analysis of pressure fluctuations in fluidized beds. III. The significance of the cross correlations. *Chem. Eng. Sci.* **2013**, *101*, 390−400.
(56) Zhang, C.; Li, X.; Zhang, M. A Novel ECG Signal Denoising Method Based on Hilbert-Huang Transform. *2010 International Conference on Computer and Communication Technologies in Agriculture Engineering*, 2010; pp 284−287.

https://doi.org/10.1021/acs.est.1c04048
Environ. Sci. Technol. 2021, 55, 15531−15541