Lawrence Berkeley National Laboratory
LBL Publications

Title
The Kimberlina synthetic multiphysics dataset for CO2 monitoring investigations

Permalink
https://escholarship.org/uc/item/300461wh

Authors
Alumbaugh, David
Gasperikova, Erika
Crandall, Dustin
et al.

Publication Date
2023

DOI
10.1002/gdj3.191

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Check for updates

Received: 23 November 2022 | Revised: 15 March 2023 | Accepted: 10 April 2023

DOI: 10.1002/gdj3.191

Geoscience Data Journal | RMetS | WILEY

DATA ARTICLE

The Kimberlina synthetic multiphysics dataset for CO_2 monitoring investigations

David Alumbaugh[1] | **Erika Gasperikova**[1] | **Dustin Crandall**[2] | **Michael Commer**[1] | **Shihang Feng**[3] | **William Harbert**[2] | **Yaoguo Li**[4] | **Youzuo Lin**[3] | **Savini Samarasinghe**[4]

[1]Lawrence Berkeley National Laboratory, Berkeley, California, USA

[2]National Energy Technology Laboratory, Morgantown, West Virginia, USA

[3]Los Alamos National Laboratory, Los Alamos, New Mexico, USA

[4]Colorado School of Mines, Golden, Colorado, USA

Correspondence
David Alumbaugh, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 74R316C, Berkeley, CA 94720, USA.
Email: dlalumbaugh@lbl.gov

Funding information
U.S. Department of Energy, Grant/Award Number: FP00010056

Abstract

We present a synthetic multi-scale, multi-physics dataset constructed from the Kimberlina 1.2 CO_2 reservoir model based on a potential CO_2 storage site in the Southern San Joaquin Basin of California. Among 300 models, one selected reservoir-simulation scenario produces hydrologic-state models at the onset and after 20 years of CO_2 injection. Subsequently, these models were transformed into geophysical properties, including P- and S-wave seismic velocities, saturated density where the saturating fluid can be a combination of brine and supercritical CO_2, and electrical resistivity using established empirical petrophysical relationships. From these 3D distributions of geophysical properties, we have generated synthetic time-lapse seismic, gravity and electromagnetic responses with acquisition geometries that mimic realistic monitoring surveys and are achievable in actual field situations. We have also created a series of synthetic well logs of CO_2 saturation, acoustic velocity, density and induction resistivity in the injection well and three monitoring wells. These were constructed by combining the low-frequency trend of the geophysical models with the high-frequency variations of actual well logs collected at the potential storage site. In addition, to better calibrate our datasets, measurements of permeability and pore connectivity have been made on cores of Vedder Sandstone, which forms the primary reservoir

Dataset The Kimberlina Geophysical Data consists of 54 zip files containing geophysical property models, synthetic geophysical data and well logs, and CT scans of Vedder Sandstone core. The main page (https://edx.netl.doe.gov/dataset/kimberlina-1-2-ccus-geophysical-models-and-synthetic-data-sets) provides a description and links to individual files.

Identifier: DOI: 10.18141/1887287

Creator: Lawrence Berkeley National Laboratory (E. Gasperikova, D. Alumbaugh, and M. Commer), National Energy Technology Laboratory (D. Crandall and W. Harbert), Los Alamos National Laboratory (S. Feng and Y. Lin), and Colorado School of Mines (Y. Li and S. Samarasinghe).

Title: Kimberlina Geophysical Data.

Publisher: National Energy Technology Library's Energy Data Exchange.

Website: https://edx.netl.doe.gov/group/kimberlina-geophysical-data

Publication Year: 2022.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *Geoscience Data Journal* published by Royal Meteorological Society and John Wiley & Sons Ltd.

unit. These measurements provide the range of scales in the otherwise synthetic dataset to be as close to a real-world situation as possible. This dataset consisting of the reservoir models, geophysical models, simulated time-lapse geophysical responses and well logs forms a multi-scale, multi-physics testbed for designing and testing geophysical CO_2 monitoring systems as well as for imaging and characterization algorithms. The suite of numerical models and data have been made publicly available for downloading on the National Energy Technology Laboratory's (NETL) Energy Data Exchange (EDX) website.

KEYWORDS

CO_2 storage, geophysics, monitoring, subsurface

1 | INTRODUCTION

Geologic carbon sequestration (GCS) is a strategy to help mitigate climate change by injecting and storing CO_2 into deep reservoirs rather than letting the greenhouse gas be emitted into the atmosphere. To aid in the deployment of GCS within the United States (US), the US Department of Energy (DOE) recently initiated the Science-informed Machine Learning for Accelerating Real-Time Decisions in Subsurface Applications (SMART) Initiative. The goal of SMART is to develop approaches that facilitate our understanding of the subsurface, and specifically GCS, through near real-time visualization, forecasting and virtual learning. With the focus being on near real-time results, a large emphasis has been placed on testing and applying science-based machine learning and data analytics to transform how people interact with subsurface data and improve the efficiency and effectiveness of field-scale carbon storage operations.

Test data generated from a known model are necessary for the testing of real-time visualization of rock properties at depth, and the development of physics-guided machine learning (ML) algorithms and workflows that can assimilate multiple types of data and measurements made over a wide range of spatial scales. The data types range from high-resolution (μm, micro-meter) laboratory computer tomography (CT) scans of core undergoing CO_2 imbibition at limited collection points, to multi-physics geophysical data that have tens to hundreds of meters of resolution at the depths of interest. Though the geophysical measurements made on the earth's surface have lower spatial resolution compared to the core and well log measurements, they achieve good spatial coverage and provide for the detection of real-time volumetric changes in the reservoir.

One of the questions facing researchers developing the various algorithms under the SMART Initiative is what data to use to prove the efficacy of the imaging and visualization algorithms, and the quality and certainty of the answers these algorithms produce. There is a limited number of GCS sites currently operating in the United States, and existing sites have limited data types available, some of which are publicly available. There have been full-scale tests in other countries or regions (Ringrose, 2020), for example Sleipner in the North Sea (e.g. Torp & Gale, 2004), Aquistore in Canada (e.g. Worth et al., 2014), and Otway in Australia (e.g. Underschultz et al., 2011), and a number of tests where geophysical data have been collected on a smaller scale such as the Cranfield site in Mississippi (e.g. Hovorka et al., 2013) and the Containment and Monitoring Institute's Field Research Station in Canada (e.g. Lawton et al., 2019).

Although the above examples provide case histories of field tests where data have been collected during CO_2 sequestration operations, no matter how well-characterized the sites are with core measurements, well logs, multi-physics imaging using geophysical and other data, none are completely 'known' in terms of the target that is being produced via CO_2 injection. Therefore, to test the ML algorithms and workflows for visualizing CO_2 saturation and estimates of uncertainty in terms of accuracy, a known GCS model must be synthetically created to provide test data.

We employ CO_2 injection simulations using Kimberlina 1.2 model (Birkholzer et al., 2011; Wainwright et al., 2013) to produce geophysical properties and data with the approach similar to that by Yang et al. (2019) and Gasperikova et al. (2020, 2022). The Kimberlina realizations were based on a potential CO_2 storage site in California's Southern San Joaquin Basin. (Note that due to various factors, CO_2 has never been injected at this site). To fully analyse the sensitivity of the system to the injection, over 300 different realizations/perturbations of the porosity and permeability of the Kimberlina base model were stochastically created. We have used 100 of these 3D simulations, each of which has outputs of pressure, temperature and CO_2 saturation at 33-time steps starting from the pre-injection

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

state through 50 years of injection, and from 50 years out to 200 years in a post-injection phase. The 33 output times 100 different realizations yield a total of 3,300 3D models. While two of these models (SIM001 at 0 and 20 years after the start of injection) are being used to provide synthetic test data, the remainder are being employed to provide training datasets for the ML algorithms under development for near real-time visualization.

The rest of this paper is organized as follows. First, we will describe the process of converting the 100 realizations of Kimberlina model hydrologic properties with 33-time steps of pressure and saturation data to 3D models of geophysical properties, including seismic velocity (both pressure and shear wave), bulk rock density and electrical resistivity. Next, we demonstrate the process of generating synthetic 2D and 3D seismic acoustic, gravity and electromagnetic (EM) datasets. To bring in the multi-scale nature of the required data, we present the creation of a series of well logs at simulated monitoring well locations. Next, we outline a series of computer tomography images made on the Vedder Sandstone core samples that were collected during drilling. Last, we describe the data repository on NETL's Energy Data Exchange (EDX) system where the data are publicly available.

We note that the objective of this paper is not to study the specific sensitivity of the different geophysical techniques to the subsurface changes caused by the CO_2 injection in this specific case. Instead, the ultimate goal of this paper is to describe this unique synthetic dataset which can be employed during research and development efforts to develop algorithms and workflows for monitoring and safe geologic storage of CO_2, and to show how other researchers can gain access to the data.

2 | DESCRIPTION OF THE KIMBERLINA 1.2 NUMERICAL MODEL

The Kimberlina 1.2 reservoir model was developed based on a geological study in the Southern San Joaquin Basin, California, using geologic and hydrogeologic data obtained from many oil fields in the region (Birkholzer et al., 2011; Dataset: Kimberlina, n.d.; Wagoner, 2009). The model includes 12 formations, from the crystalline basement to the top shallow aquifer, over an 84 x 112 km area. CO_2 is injected into a saline reservoir of the deep Vedder formation via a single well in the centre of the model.

The Vedder formation is a large permeable sandstone formation that dips upward towards a shallow outcrop area located on the eastern border of the model. The overlying Temblor-Freeman Shale formation is a suitable reservoir seal for the containment of the injected supercritical CO_2 with a porosity of 0.001 and horizontal permeability of 0.002 mDarcy. Based on logs collected in two wells drilled in the area, the Vedder formation contains six laterally continuous layers of alternating sand and shale, with the thickest sand layer located at the top portion of this formation. Birkholzer et al. (2011) state the porosity of the Vedder ranges from 0.27 in the sand units to 0.32 in the shale baffles with horizontal permeabilities of 303 mDarcy and 0.1 mDarcy, respectively. The Vedder formation is about 400 m thick at the injection well, its top elevation is about 2,750 m below the ground surface (Wainwright et al., 2013), and the caprock shale formation is about 200 m thick. Several faults in the area are modelled with a hydraulic conductivity below that of the adjacent sandstone formations (Birkholzer et al., 2011), thus acting as partial barriers. In this model, the lateral permeability of major faults is reduced by a factor of 100 compared to the adjacent formation permeability. The faults are assumed impermeable in shale formations, that is the potential for leakage of CO_2 through permeable faults is not a concern at this site (Wainwright et al., 2013). Additional hydrologic properties for the flow model can be found in Table 1 below which is reproduced from Birkholzer et al. (2011).

All CO_2 injection simulations were conducted using the massively parallel version of the TOUGH2/ECO2N simulator (Pruess, 2005; Zhang et al., 2008). The TOUGH2 3D mesh comprises 64,214 elements, with a fine mesh in the centre and a growing cell size towards the model edges (Wainwright et al., 2013). The simulations employed a constant injection rate of 5 million tons of CO_2 per year for 50 years. This yielded a maximum plume extent of 13 km by 9 km with a maximum reservoir pressure of 23 bars (2.3 MPa) and a residual water saturation away from the injection well of approximately 40%. After injection cessation, the simulations cover a post-injection period of 150 years.

TOUGH2 flow simulations were converted to geophysical property models required for modelling of seismic, gravity and electromagnetic monitoring and data simulations. The resistivity models were constructed using the following empirical relationships. Parameters affecting the pore-fluid electrical conductivity (EC) are salt mass fractions, converted into total dissolved solids (TDS), and temperature (Hayashi, 2004; Walton, 1989):

$$\mathrm{TDS\,(mg/L)} = 7500\ \mathrm{EC\,(S/m)}, \tag{2.1a}$$

$$\mathrm{EC}(t) = \mathrm{EC}(t_0)\left[1 + a\left(t - \mathrm{t}_0\right)\right], \tag{2.1b}$$

where $a = 0.022$, $t_0 = 20°C$, and t is the simulation's time.

The electrical property of interest for geophysical modelling is the bulk formation resistivity (res_b). Using

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Hydrogeologic properties assigned to each formation: k_h is horizontal permeability, k_v is vertical permeability, Φ is porosity, β_p is pore compressibility, α is the van Genuchten parameter for entry capillary pressure, and m is the van Genuchten parameter for pore-size distribution.

Formations	k_h [mDarcy]	k_v [mDarcy]	Φ [−]	β_p [10^{-10} Pa^{-1}]	α [10^{-5} Pa^{-1}]	m [−]
Non-fault zones						
Pre-Etchegoin	3,000	3,000	0.35	15.5	5.0	0.457
Etchegoin	1,200	1,200	0.32	15.5	5.0	0.457
Macoma-Chanac	1,900	1,900	0.31	10.5	5.0	0.457
Santa Margarita-McLure	2,000	2,000	0.275	10.5	5.0	0.457
Stevens Sand	240	48	0.22	10.5	5.0	0.457
Fruitvale-Round Mountain	0.002	0.001	0.338	14.5	0.42	0.457
Olcese Sand	170	34	0.336	4.9	5.0	0.457
Temblor-Freeman	0.002	0.001	0.338	14.5	0.42	0.457
Vedder Sand (sand layers)	303	60.6	0.264	4.9	13.0	0.457
Vedder Sand (shale layers)	0.1	0.05	0.32	14.5	0.42	0.457
Tumey-Eocence	0.07	0.07	0.07	14.5	0.42	0.457
Baserock	0.0001	0.0001	0.01	22.7	0.5	0.457

Equation (2.1b) and Archie's equation (Archie, 1942), we obtain

$$\mathrm{res}_b = \frac{1}{\mathrm{EC}\,\phi^2 S_f^2}, \tag{2.2}$$

where ϕ is formation porosity and S_f is fluid saturation, the latter is related to CO_2 saturation through S_f=(1-S_{CO2}). The relevant subsurface hydrologic properties (e.g. fluid salinity, fluid saturation and porosity) were extracted from the TOUGH2 simulation output for the calculation of res_b throughout the 3D volume.

Seismic velocities (V_p and V_s) and density models were created using relationships presented by Wang et al. (2018) and Yang et al. (2019). Both V_p and V_s velocities are related to saturated bulk modulus (K_{sat}), saturated shear modulus (μ_{sat}) and saturated density (ρ_{sat}). Saturated density can be calculated by knowing the porosity (ϕ), densities of the fluid (ρ_{fl}) and framework density (ρ_{frame}). With knowledge of K_{sat}, μ_{sat}, and ρ_{sat} (Sheriff & Geldart, 1995), V_P, V_S and ρ_{sat} can be calculated (Avseth et al., 2007; McKenna et al., 2003) by

$$V_P = \sqrt{\frac{K_{sat} + \frac{4}{3}\mu_{sat}}{\rho_{sat}}}, \tag{2.3a}$$

$$V_S = \sqrt{\frac{\mu_{sat}}{\rho_{sat}}}, \tag{2.3b}$$

$$\rho_{sat} = \varphi * \rho_{fl} + (1-\varphi) * \rho_{frame}, \tag{2.3c}$$

where K_{sat} is calculated as a combination of the bulk modulus of the mineral, framework and the pore filling fluids. The bulk modulus of the pore filling fluid (K_{fluid}), minerals ($K_{mineral}$), framework mineralogy (K_{frame}) and porosity (ϕ) was used to estimate the saturated bulk modulus (K_{sat}) and saturated density (ρ_{sat}):

$$K_{sat} = K_{frame} + \frac{\left(1 - \frac{K_{frame}}{K_{mineral}}\right)^2}{\frac{\varphi}{K_{fl}} + \frac{1-\varphi}{K_{mineral}} - \frac{K_{frame}}{K_{mineral}^2}}. \tag{2.3d}$$

We assumed all layers were 70% quartz and 30% clay by volume in our simplified rock framework model and a Poisson's ratio of 0.2. Note that the shear modulus is not changed by fluid saturation assuming the low-frequency Gassmann-Biot model (Gassmann, 1951):

$$\mu_{sat} = \mu_{dry} \tag{2.3e}$$

As described in Wang et al. (2018), the fluid bulk modulus and density (K_{fl} and ρ_{fl}) were estimated using averaging of the separate pore fluid phases (brine and CO_2 phases) (Kumar, 2006):

$$\frac{1}{K_{fl}} = \frac{S_w}{K_{brine}} + \frac{S_g}{K_{co2}} \text{ and } \rho_{fl} = S_w \rho_{brine} + S_g \rho_{co2}, \tag{2.3f}$$

where S_g is the CO_2 saturation and S_w is the brine saturation. The bulk moduli and densities of pure brine and CO_2 (K_{brine}, K_{co2} and ρ_{brine}, ρ_{co2}) were calculated as functions of temperature, pressure, and salinity (Batzle & Wang, 1992).

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

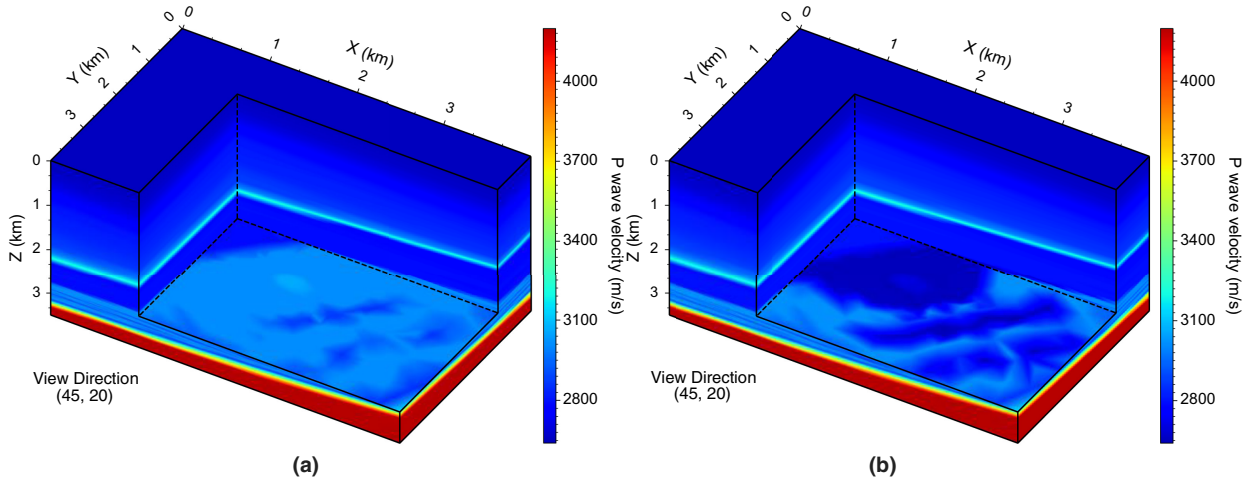

FIGURE 1 Example of the 3D 4 km by 4 km by 3.5 km velocity model at (a) Year-0 and (b) Year-20.

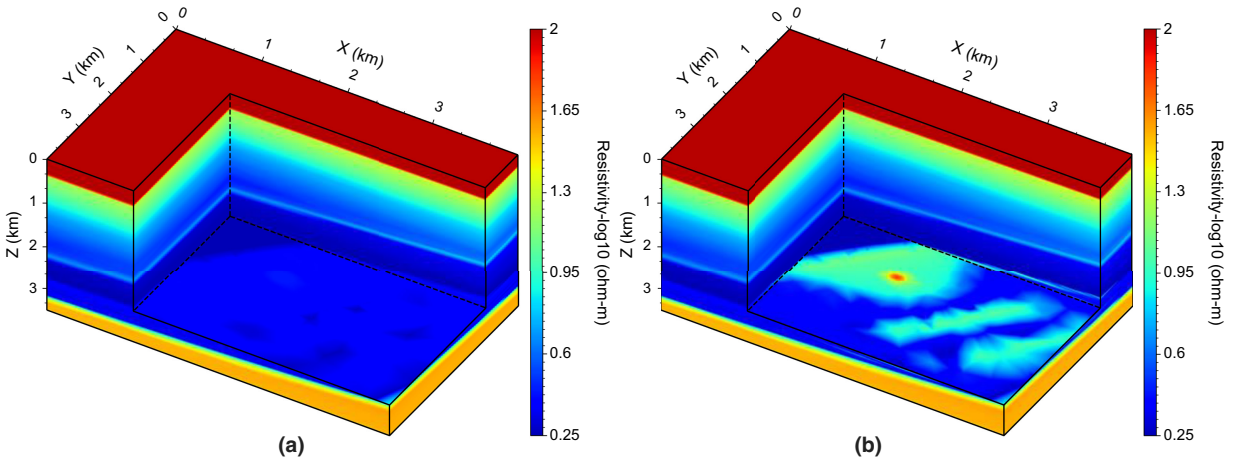

FIGURE 2 Example of the 3D 4 km by 4 km by 3.5 km resistivity model at (a) Year-0 and (b) Year-20.

All geophysical properties were calculated on the original unstructured TOUGH2 grid and then linearly interpolated onto a regular Cartesian $10 \times 10 \times 10$ m grid. The latter spans from −2,000 to 4,000 m in the x-direction, from −2,000 to 4,000 m in the y-direction, and from 0 m to 3,500 m in the z-direction. The interpolation leads to a model size of 600 x 600 x 350 (in grid nodes along x, y, z).

The geophysical property files contain two columns which are Node-ID, Property. Geophysical property files exist for CO_2 saturation, V_p, V_s, density and resistivity. The corresponding (regular) Cartesian grid is identical for all property files and is specified by a separate 'mesh' file which contains the four columns Node-ID, x-coordinate, y-coordinate, z- coordinate. The node-ordering in each property file is identical with the node-ordering in the mesh file. The Node-ID is useful if one wants to select model sub-volumes through specification of subsets of Node-IDs.

3 | SYNTHETIC GEOPHYSICAL DATA CREATION

As mentioned previously, the studies of Yang et al. (2019) and Gasperikova et al. (2020, 2022) used variations of the Kimberlina CO_2 injection model developed under the DOE's National Risk Assessment Partnership (NRAP) program in their studies analysing the sensitivity of various geophysical techniques to shallow acclamations of CO_2. Given that the purpose of this paper is to describe synthetic multi-physics geophysical datasets for testing various ML and imaging algorithms and workflows for monitoring plume evolution and estimating CO_2 saturation at depth, we have not exhaustively sampled different sensor configurations. Rather for seismic monitoring, we only use surface seismic arrays as these have become the industry standard for monitoring. However, to provide a cost-effective solution, we have adopted the approach of showing good

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

sensitivity with a limited number of surface sources (e.g. Correa et al., 2021; Pevzner et al., 2021). Gasperikova et al. (2020, 2022) show that the borehole-to-surface EM technique with a vertical dipole source located at proximity to the storage region exhibits better sensitivity to injections at depth than surface methods. We have thus chosen to simulate the configuration with the source at the bottom of the monitoring well and surface receivers. Note that we chose not to simulate the magnetotelluric (MT) response as it is well known that MT is insensitive to thin resistors at depth (Constable & Weiss, 2006; Um & Alumbaugh, 2007). Lastly, for gravity, we used both borehole and surface three-component measurements of the gravitational acceleration.

Below, we discuss the simulated data for each geophysical technique and provide examples of 2D and 3D datasets. These datasets include some acquisition configurations for the SIM001 realization of the Kimberlina model used for testing and training ML approaches. If other researchers are interested in testing new configurations, these can be simulated and their sensitivities studied using publicly available geophysical property files.

Cut-away views through the 3D velocity model at years 0 and 20 are shown in Figure 1a,b, respectively. The same cut-away views through the 3D resistivity and density models at years 0 and 20 are shown in Figures 2 and 3, respectively. The cut-away view through the 3D CO_2 saturation model for both years is shown in Figure 4.

3.1 | Acoustic seismic data generation

The synthetic seismic data were generated using both 2D and 3D finite-difference codes that simulate the acoustic wave equation (Moczo et al., 2007). To compute the 2D data, we extracted 6-km long longitudinal sections of the P-Wave velocity model along the *Y*-axis from each of the 33 time-steps of the SIM001. These 2D slices of P-wave velocity were extracted at 100 m intervals from $X=-2$ km to $X=4$ km, as shown in Figure 5. Six point-pressure sources were positioned along each line from $Y=-2$ km to 4 km at 1.2-km intervals, and receivers were spaced at 10 m intervals along each line. Figure 6 shows the P-wave velocity slices of the models at Year-0 and Year-20, respectively, used to generate the test data along with the profile at $X=0$ km. Figure 7 provides examples of the generated data for the 2D velocity models shown in Figure 6 for sources at $X=-2$ km, $X=1.6$ km and $X=4$ km. Figure 7a shows the Year-0 data, Figure 7b shows the Year-20 data, and Figure 7c shows the time-lapse difference between Year-20 and Year-0. There are subtle velocity changes in the background over the years so that small direct wave residuals exist in the time-lapse difference. We note the strong response generated by the introduction of the CO_2 plume. As mentioned previously, these data served as the test data for the ML algorithms described in Wu and Lin (2019) and Um et al. (2022).

The 3D synthetic seismic data were generated using a 3D finite-difference acoustic code (Moczo et al., 2007). 3D velocity models are 6 x 6 x 3.5 km volumes. To extract a velocity model with a smaller volume, we used a model decomposition method where a 4 x 4 x 3.5 km block is moved within the original model at 200 m increments in both *X* and *Y* directions. This sub-domain model extraction process was completed for nine block positions in the *X* direction and seven positions in the *Y* direction, yielding 63 (9 x 7) 3D models per SIM001 output time. Repeating this process for each of the 33 time steps in the SIM001 output yields a total of 2,079 velocity models. For each of these 3D models, 3D acoustic simulations were completed for 25 surface pressure sources using a source separation of 1 km in both the *x* and *y* directions. Of these, the test data at times 0 and 20 years with the block centred at $X=3.3$ km and $Y=2$ km served

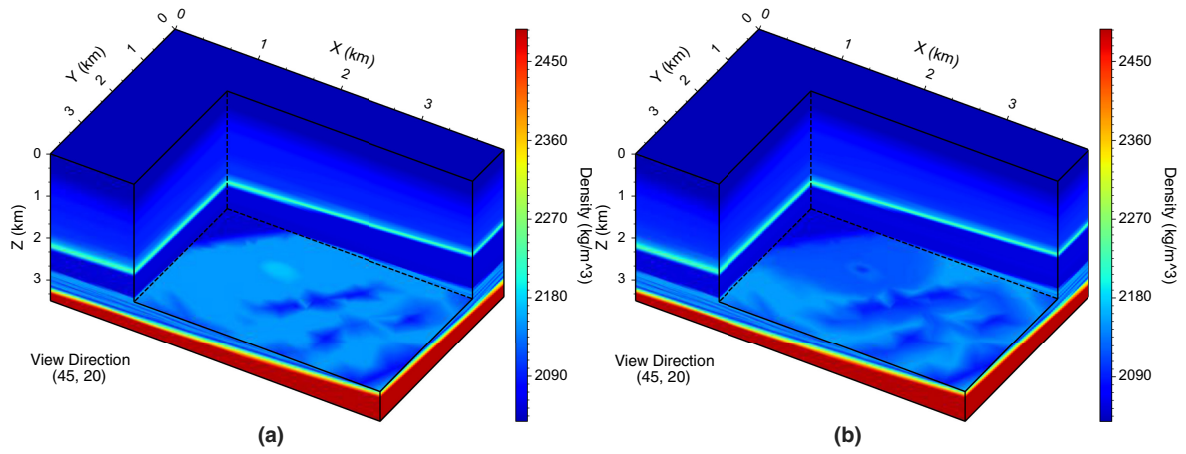

FIGURE 3 Example of the 3D 4 km by 4 km by 3.5 km density model at (a) Year-0 and (b) Year-20.

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

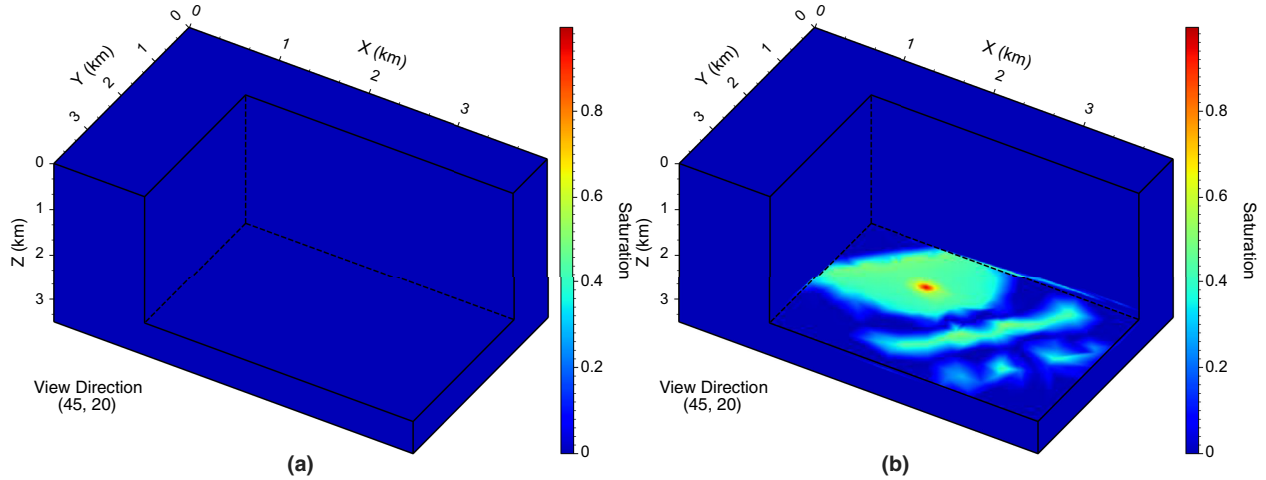

FIGURE 4 Example of the 3D 4 km by 4 km by 3.5 km CO_2 saturation model at (a) Year-0 and (b) Year-20.

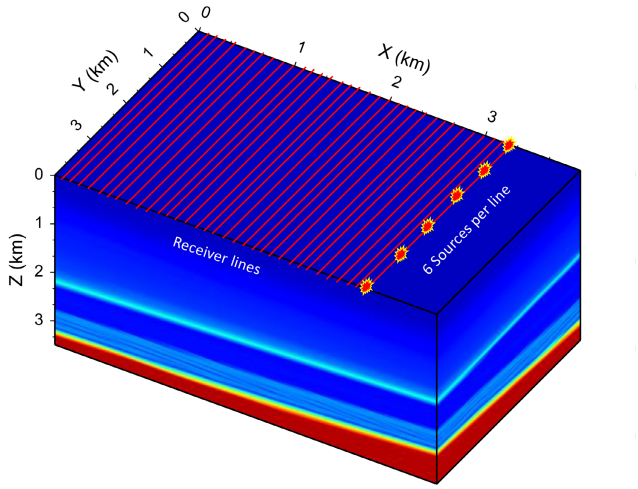

FIGURE 5 Survey lines and configuration for 2D seismic modelling.

as the test data, while the rest served to train the neural net algorithm described in Zeng et al. (2022). In Figure 8a, we provide snapshots of the 3D acoustic response for a shot point at $X = 2$ km $Y = 2$ km for the Year-0 test dataset, Figure 8b, for the Year-20 test data, and Figure 8c, the time-lapse difference.

2D seismic data and models are in the standard binary file format. The shape of the seismic data is $6 \times 10{,}001 \times 600$ (Source number × Time × X-Direction), where the time spacing is 0.0005 s. The size of the model is 350×600 (Depth × X-Direction), where the grid has a uniform 10 m interval in all directions.

3D seismic data and models are also in the standard binary file format. The shape of the seismic data is $25 \times 5{,}001 \times 40 \times 40$ (Source number × Time × X-Direction × Y-Direction), where the time spacing is 0.001 s. The size of the model is $350 \times 400 \times 400$ (Depth × X-Direction × Y-Direction), where the grid has a uniform 10 m interval in all directions.

3.2 | Electromagnetic data simulations

The petrophysical property transformation of Equation (2.2) gives rise to three-dimensional (3D) models of electrical resistivity that provide the modelling input for controlled-source electromagnetic (CSEM) data simulations. Our CSEM simulator, referenced by its code name EMGeo, employs a parallel finite-difference (FD) scheme for approximating Maxwell's equations on a staggered grid (Commer & Newman, 2008). For details on the computational aspects of this code, we refer readers to Commer and Newman (2008) and related references therein.

Figure 9 outlines EM survey configurations covering a surface area of approximately 36 km^2. In order to account for potential effects due to resistivity variations across reservoir or fault structures, our resistivity models preserve the fine-scale discretization of the underlying rock-physics models. The spatial grid node distance of 10 m along each axis leads to a total model size of $N_x \times N_y \times N_z = 601 \times 601 \times 352$ FD mesh cells. In large-scale modelling contexts of this kind, each source excitation typically has a spatially reduced footprint; that is, it only covers a certain model subdomain. Source-centred FD grids with spatially adapted Dirichlet boundary conditions allow for smaller equation systems and more economic solutions. This FD grid-separation scheme and corresponding grid-design considerations are described in detail by Commer and Newman (2008).

A numerical verification of each separate computational grid involves a stepwise grid refinement until field differences, Δ^E, between 3D and 1D-reference fields fall

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

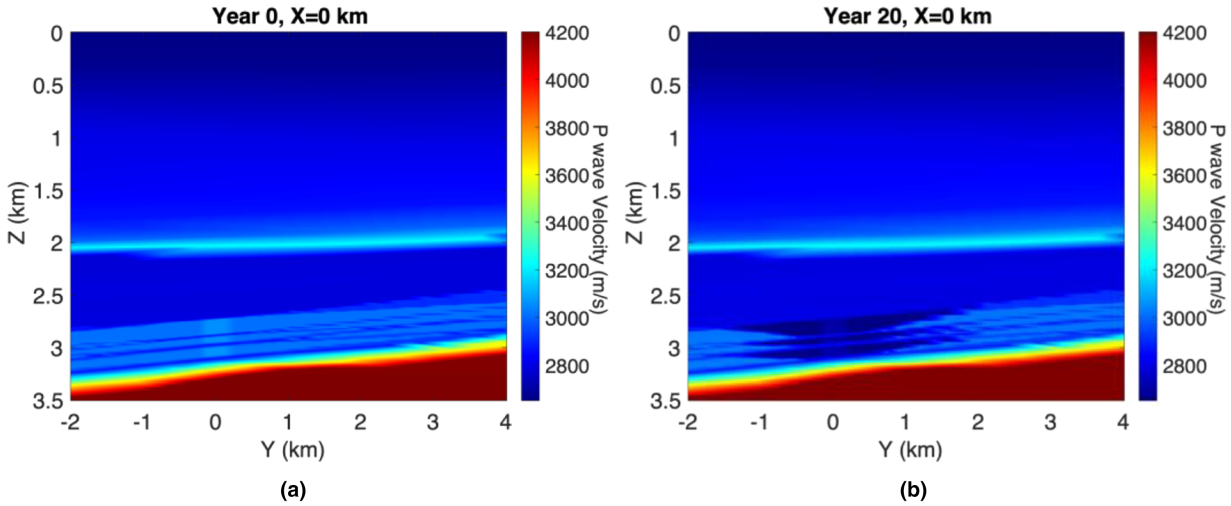

FIGURE 6 2D P-wave velocity cross-sections at $X=0$ for (a) Year-0 and (b) Year-20.

below predefined thresholds. We enforce a threshold of $t=3\%$, specifically

$$\Delta^E = \frac{\left|E^{3D} - E^{1D}\right|}{\left|E^{1D}\right|} \bullet 100 \leq t, \tag{3.1}$$

where E^{3D} and E^{1D} are the real components of the complex E-field. The reference field E^{1D} is obtained via semi-analytical solutions for a horizontally layered half-space model. This 1D reference model comprises 88 layers covering the depth range from $z=0$ to $z=3{,}480$ m. Each layer contains the horizontal variation of the baseline (pre-injection) resistivity of the SIM001 simulation averaged over a fixed thickness of 40 m.

Electromagnetic simulations involve two sets of borehole-to-surface survey configurations, where borehole CSEM sources are located near the reservoir level and electric fields are measured over surface profiles (Figure 9). The first is referred to as pseudo-2D data, because a 2D inline receiver configuration is simulated over 3D resistivity models. The second is referred to as 3D data. Here, E-fields generated by borehole sources at monitoring well locations are measured over a surface receiver grid. Example field calculations are presented for the rock-physics model state realized by SIM001. All FD models include an additional highly resistive air layer at the top of the mesh due to the surface receivers.

The 2D dataset comprises 31 Y-receiver profiles between $X=-2{,}000$ m and $X=+4{,}000$ m, with profile distance $\Delta x=200$ m (Figure 9). Each profile extends from $Y=-2{,}000$ m to $Y=+4{,}000$ m and includes 31 receiver stations spaced at $\Delta x=200$ m. For each profile, horizontal electric dipole (HED) field components with profile-parallel (E_y) and perpendicular (E_x) orientation are calculated for two inline vertical electric dipole (VED) source locations, referred to as T1 and T2. These VED sources are located at $z=3{,}025$ m, have a VED length of 50 m, and operate at a frequency range from 0.1 to 8 Hz. The HED receiver dipole length is 100 m.

Figure 10 exemplifies E_y-field responses for Profile #11 (at $X=0$ m) for the four source frequencies 0.1, 0.6, 1.0, and 6.0 Hz. Responses are compared for Year-0 (pre-injection) and Year-20 where we display the quantities in amplitude and phase relative to a current of 1 Amp in the source dipole. Owing to its proximity to injection-induced reservoir resistivity changes, the responses are more significant for the T1 transmitter compared to the T2 transmitter. In addition, the amplitude differences are above the assumed noise threshold of 10^{-12} V/m at the lower frequencies, but dip below this value at the highest frequencies due to the increased attenuation with increasing frequency. This noise threshold is a factor of 10 times less than that used by Wirianto et al. (2010). We note that this is a best-case scenario and that we have chosen to assume this lower noise value by assuming that a monitoring survey we can use larger dipole moments both on the source and receiver side, and stack data longer to provide lower noise thresholds.

The 3D data calculations employ the whole surface array, as shown by the 31 profiles in Figure 9, totalling 961 surface receiver stations. Figure 11 compares maps of field amplitudes between the Year-0 and the Year-20 for the borehole source located in monitoring well MW2. The two exemplified frequencies (i.e. 0.6 Hz vs. 6.0 Hz) demonstrate that for this kind of scale, where transmitter-receiver distances are in the km range, lower source frequencies benefit the detection of injection-induced reservoir changes owing to a lower degree of spatial field attenuation. Moreover, the higher frequency results in a smaller areal surface footprint of the injection-induced

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

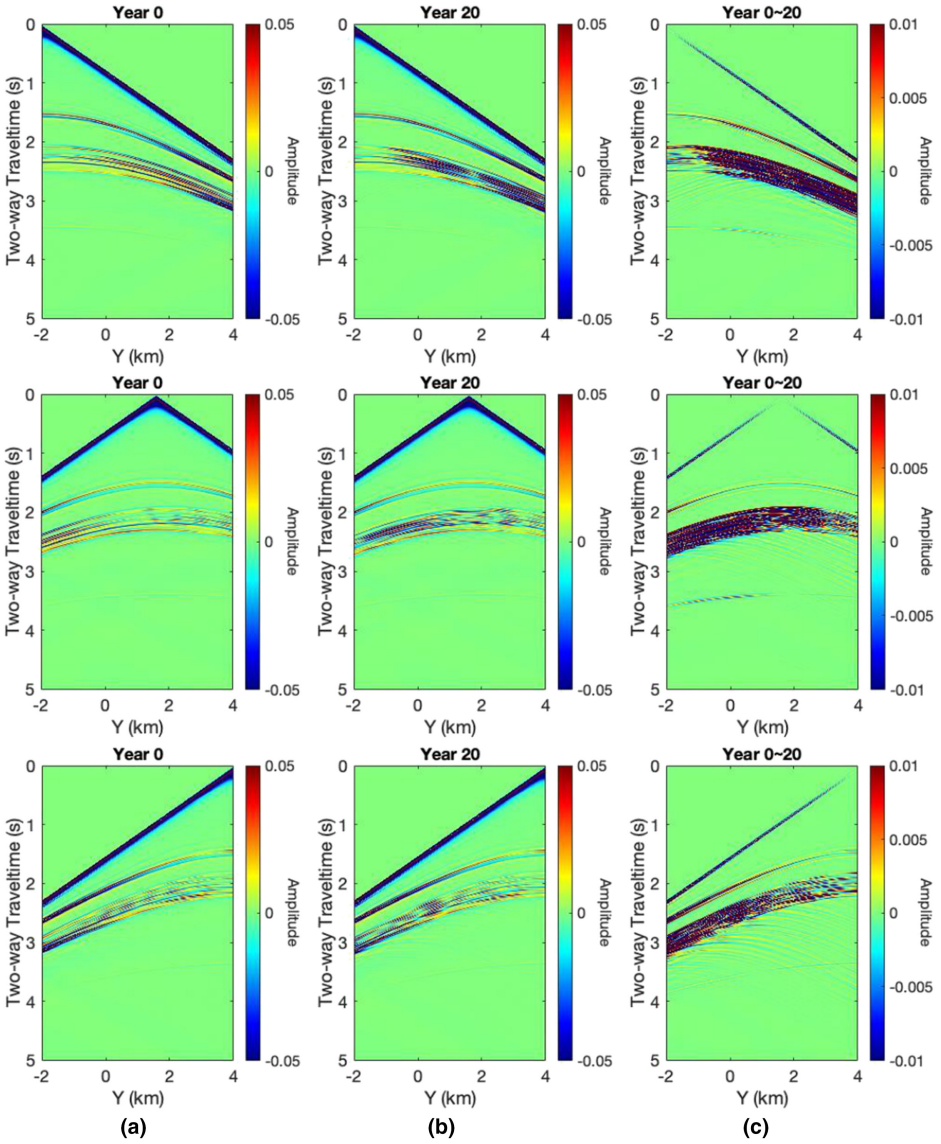

FIGURE 7 The seismic data generated at (a) Year-0, (b) Year-20, and (c) the time-lapse difference with sources located at $X = -2$ km, $X = 1.6$ km, and $X = 4$ km.

differences. Amplitude and difference levels are shown for both normalized (to unit dipole moment and unit source current) and non-normalized *E*-fields. Here, we assume a source current of 20 A and its VED length of 50 m.

The CSEM data output is in form of column-formatted text files with 11 numerical columns per data line. The column entries are as follows: (1) transmitter ID; (2) frequency in Hz; (3, 4, 5) transmitter midpoint (*x*,*y*,*z*) coordinate in *m*; (6) field-component ID; (7, 8, 9) receiver midpoint (*x*,*y*,*z*) coordinate in m; (10, 11) complex datum (real and imaginary field components).

The data output reflects the order of the data input in terms of the transmitter order, where the transmitter ID (column 1) is an integer number specifying the input order. Each dataset associated with a given transmitter ID, *i*, (*i* x number of frequencies), hence comprises of Ni data lines, where Ni is the number of transmitters × number of receivers × number of frequencies × number of calculated field components (e.g. 3D data consist of 46,128 lines = 3 transmitters × 8 frequencies × 2 field components × 961 receivers [31 × 31]). This format is specific for dipole configurations, where transmitter and receiver coordinate output is reduced to midpoints given in meters. Accordingly, field responses (columns 10, 11) are normalized to unit dipole moment and unit source current with units of V/m and A/m for electric and magnetic fields, respectively. The field-component ID (column 6) specifies the receiver field type in form of an integer number ranging from 1 to 6: 1 = Ex; 2 = Ey; 3 = Ez; 4 = Hx; 5 = Hy; 6 = Hz.

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

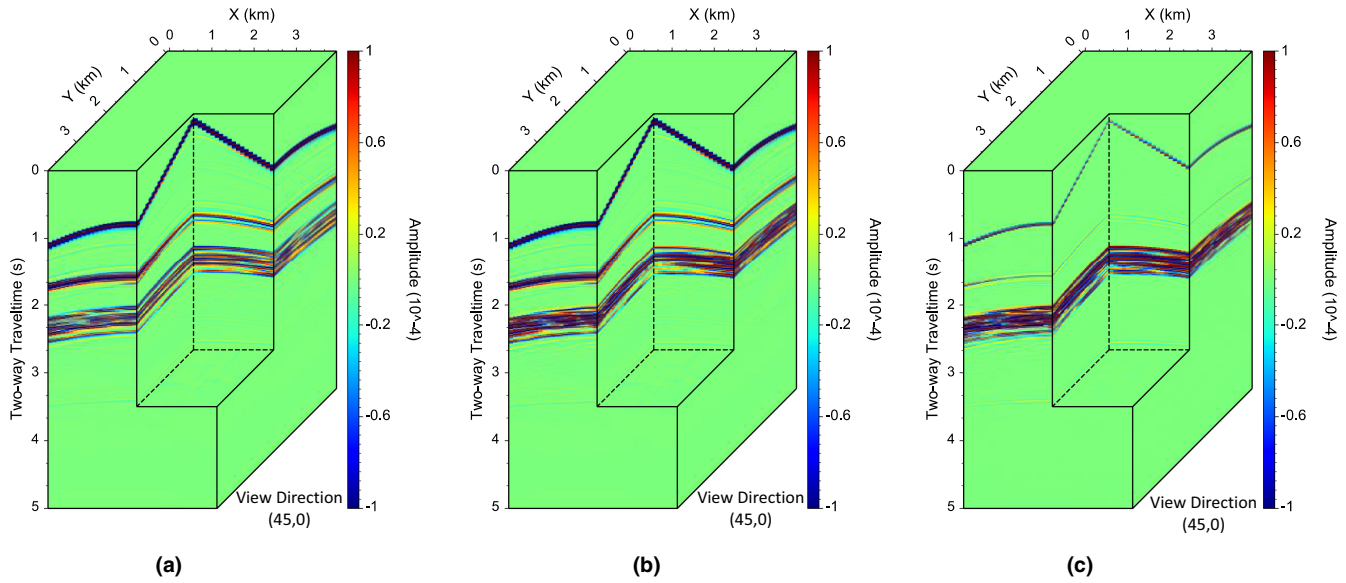

FIGURE 8 The 3D seismic data generated with a source located at $X = 2$ km and $Y = 2$ km at (a) Year-0, (b) Year-20, and (c) the time-lapse difference.

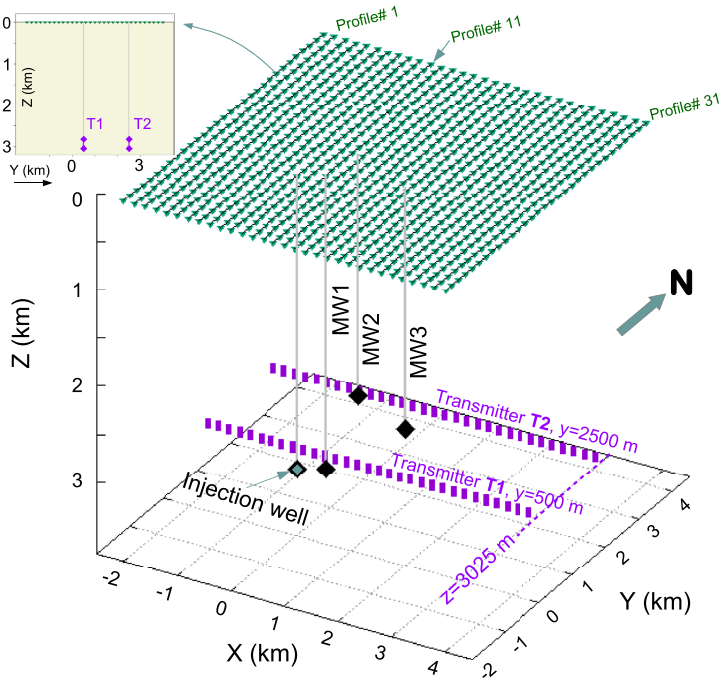

FIGURE 9 Survey layout for pseudo-2D and 3D EM field calculations.

3.3 | Gravity data generation

Another geophysical method that can contribute to monitoring subsurface distribution of CO_2 during sequestration is time-lapse gravity. Both simulation studies (e.g. Gasperikova et al., 2022; Krahenbuhl et al., 2015; Yang et al., 2019) and field trials (e.g. Alnes et al., 2011) have shown the efficacy of the method. A major advantage of time-lapse gravity monitoring stems from the fact that the time-lapse density difference is directly and uniquely dependent upon the CO_2 saturation change, provided that the reservoir porosity change is negligible. Furthermore, similar to the EM response, the gravity response is sensitive to the entire range of CO_2 saturation. The effectiveness of the method may be significantly improved if time-lapse gravity responses are measured downhole by deploying gravimeters in monitoring wells (e.g. Bonneville et al., 2021). A study by Rim and Li (2015) also shows that vector gravity measurements can enhance the information in gravity from sparsely located wells through the inherent direction information contained in the vector gravity data. Therefore, we computed synthetic vector

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

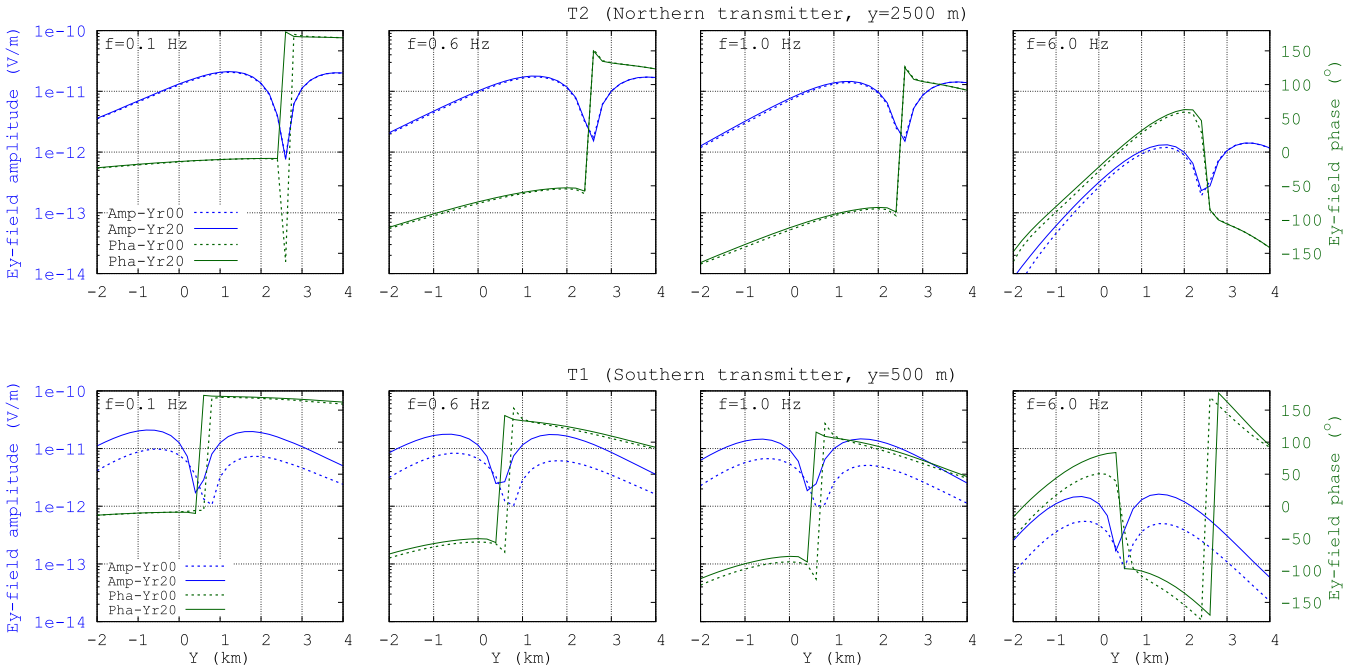

FIGURE 10 Comparison of *E*-fields simulated for the Year-0 (pre-injection, dashed lines) and the Year-20 (solid lines). Field responses are normalized to unit dipole moments and unit source current. Complex fields are displayed as amplitudes and phases.

gravity data both on the surface and down in the injector and monitoring wells. The acquisition scenarios were parallel to those used for the EM:

- Pseudo-2D data were calculated along the same lines and within the same boreholes as shown in Figure 9;
- Full 3D acquisition geometries were completed using the same three monitoring wells shown in Figure 9.

Measured gravity data, in reality, contain a significant component of the common-mode signal that does not vary with time. The sources of the common-mode component include the background rock density and the terrain variation. For this reason, the data in time-lapse gravity monitoring are typically the difference obtained by subtracting the gravity measurements at a reference time from those measured at a later time, provided that the locations of the measurements are repeated with sufficient accuracy. Therefore, the acquisition and processing of time-lapse gravity data in practice seek to extract the time-lapse difference gravity as the final data. For this reason, we only calculated the time-lapse differences in the gravitational acceleration by using a modelling code named *vgfor3d* (Rim & Li, 2015) developed at the Colorado School of Mines and simulations of time-lapse density changes. The use of the algorithm in this manner assumes that data are collected at Year-0 and Year-20 and that the only changes in the subsurface density occurring during that time interval are due to lower-density CO_2 replacing higher-density brine within the storage reservoir. Figure 12 shows the three components of the anomalous gravitational acceleration (henceforth referred to as gravity anomalies) that the CO_2 plume would generate at Year-20 along a *Y*-directed line at $X = 0$ km for the pseudo-2D (i.e. line) acquisition scenario. Figure 13 displays maps of three components of the time-lapse gravity anomaly as measured across an area on the surface directly over the CO_2 plume. Figure 14 displays three components of the time-lapse gravity anomaly that would be measured in monitoring well MW1. We note that the accuracy of current gravity instruments is from 1 to 5 μGal. Therefore, both the surface and borehole anomalies will be measurable.

Both the pseudo-2D and full 3D data have the following format:

- All data files are in ASCII format
- The first line of the file indicates the number of records, which is the number of simulated data locations
- Subsequent lines have six fields (i.e. columns): the first three are *X*, *Y*, and elevation (referenced to the surface that is assumed to be 0 elevation); and columns 4 to 6 are the gravity anomaly components in *Y*, *X*, and vertical (*Z*) directions, respectively.
- All coordinates are in meters, and all gravity values are in milliGals (mGal).

4 | CREATION OF SYNTHETIC WELL LOGS

Some ML algorithms and workflows require well logs as part of the training data. As a part of converting the

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

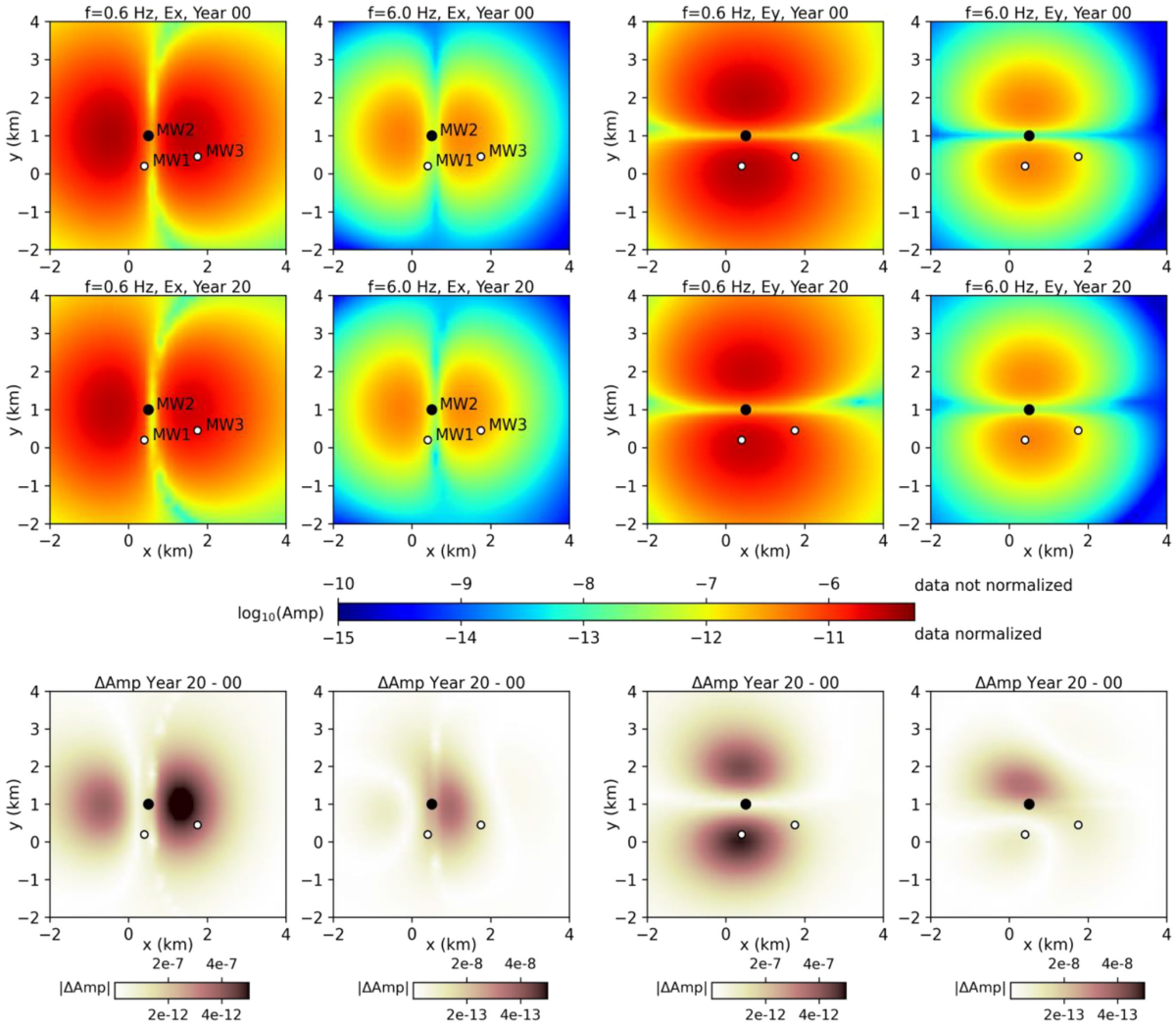

FIGURE 11 *E*-field amplitudes are plotted over horizontal (*x*-*y*) surface sections for the Year-0 (upper row) and Year-20 (middle row) of the SIM001 model. The examples use the source frequencies of 0.6 and 6 Hz. E_x (plot columns 1 and 2) and E_y (plot columns 3 and 4) field responses stem from the borehole source at $z = 3{,}050$ m in monitoring well MW2. Amplitudes and absolute differences (Year-20 – Year-0, bottom row) are shown for both normalized and non-normalized fields, where the latter include a combined HED and VED dipole moment of 10^5 Am.

Kimberlina hydrologic models to geophysical properties, we also created a series of synthetic well logs that both obey the geophysical property models at the coarser scale and have realistic finer-scale variations with depth. Figure 15 demonstrates some of the issues that arise when using the well logs used to create the Kimberlina geologic model and the geophysical property models themselves. Figure 15a is the density log recovered from the Kimberlina-1 well. This was converted from the density porosity log by using a value of 1 g/cm^3 for the water filling the pore space and 2.67 g/cm^3 for the rock matrix which was assumed to consist primarily of quartz.

Figure 15b is the density versus depth from the Year-0 geophysical model. We note a major difference between the model and actual density log (Figure 15a). Whereas the actual log exhibits strong high-frequency variations with depth, due to the coarse discretization used when creating the geophysical model, the model log is smooth with depth. Given the models that have been created using the methodology outlined in Section 2, we needed to develop a methodology that would provide realistic looking well logs that capture the low-frequency trends with depth from the model. To provide for this, we developed the following workflow.

1. Low-pass filter the Kimberlina 1 well logs (sonic velocity, converted density, deep-induction resistivity and density porosity) with a 101 data point averaging window, which corresponds approximately to a depth interval of 15 m.

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

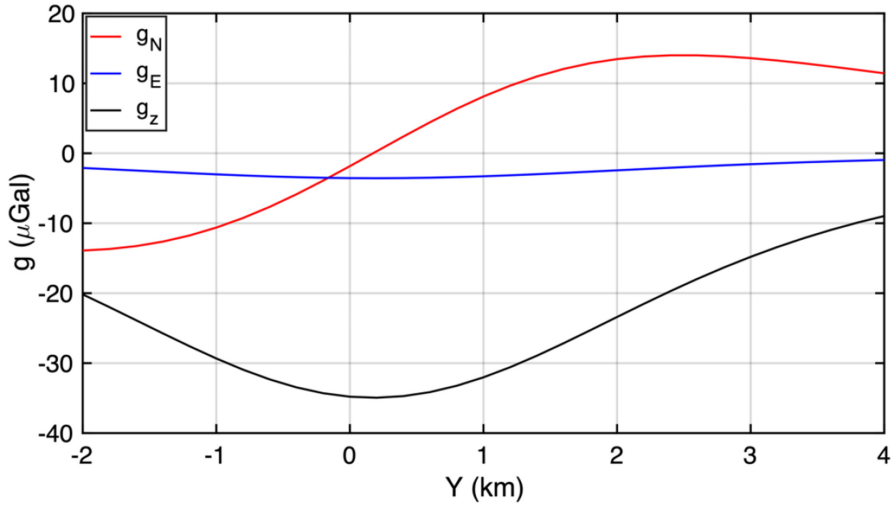

FIGURE 12 Three components of the time-lapse gravity anomaly along a *Y*-profile at $X = 0$ m. Although the anomaly is smooth due to the large depth of the storage reservoir, the northing- and z-directed components are well above the current instrument sensitivity. The small easting directed component on this profile is small because it is located directly near the centre of mass of the CO_2 plume.

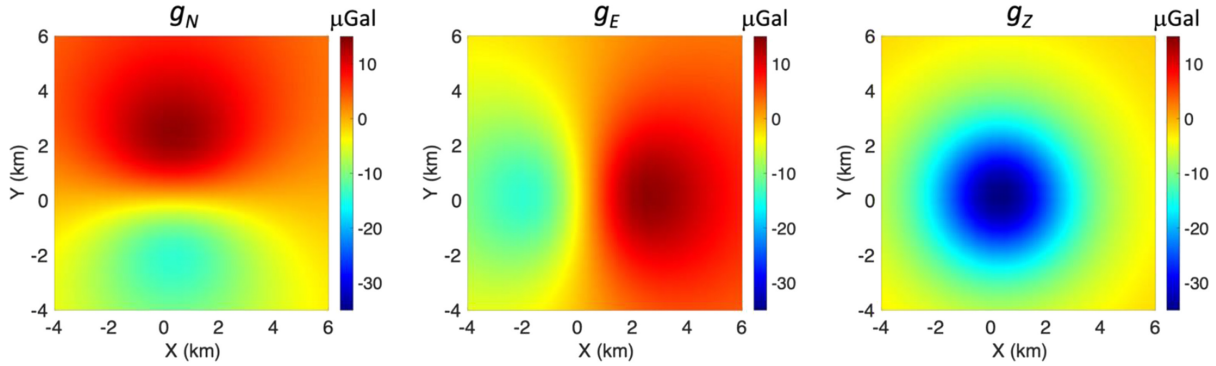

FIGURE 13 Three components of the time-lapse gravity anomaly at Year-20. The anomaly on the surface is smooth due to the depth of the CO_2.

2. Subtract the averaged logs from the actual logs to produce fine-scale 'perturbation' logs for each data/property type.
3. Extract 'logs' from the geophysical property models at each of the well locations shown in Figure 9 to form the long-wavelength component of the synthetic logs.
4. Combine the perturbation logs with the geophysical model property logs to produce the synthetic logs that have both the high-frequency variations of the actual logs as well the general depth trends of the geophysical models.

For step 4, combining the geophysical model logs with the perturbations depends on the type of well log we are synthesizing. Due to the linear nature of the range of sonic velocities and densities within rocks, the perturbations were simply added to the geophysical model logs for these two types of logs. However, because the electrical resistivity of rocks is better represented on a logarithmic scale, the synthetic resistivity logs were constructed by taking the calculated resistivity perturbations and scaling them to have maxima and minima between 1.2 and 0.8, respectively, and then multiplying the resistivity logs extracted from the models by these scaled values. We admit that this process as applied to the resistivity log generation is somewhat ad hoc. However, given the goal of this process is to produce synthetic logs that have the same high-frequency characteristics as the real logs along with the low-frequency characteristics of the geophysical model, we believe that this process resulted in synthetic resistivity logs that have realistic logarithmic scaling.

To generate synthetic CO_2 saturation logs that can be used for converting geophysical property values to estimates of CO_2 saturation, we scaled the density porosity log such that within the reservoir where all the injected

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

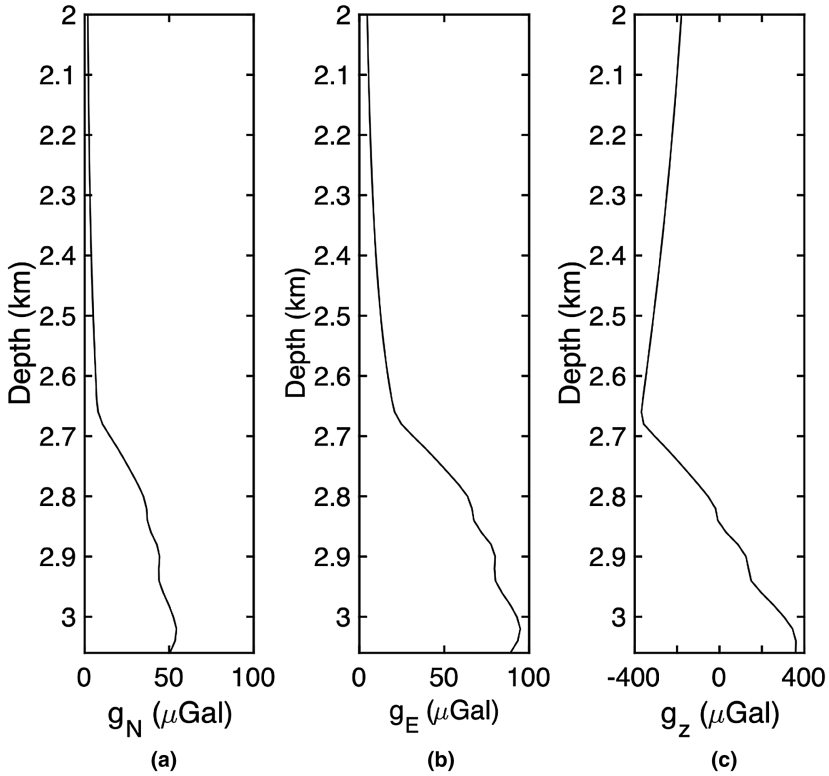

FIGURE 14 Time-lapse gravity anomaly at Year-20 in the MW1. The conventional z-directed (vertical) component (c) has significant magnitudes well above the current instrument accuracy of 5 μGal along the entire length of the well, while the two horizontal components (a,b) show significant magnitudes below the depth of 2,000 m.

CO_2 was contained, the rescaled values ranged from approximately 0.65 to 1.4. These scaled values then multiplied the CO_2 saturations extracted from the hydrologic simulations to produce realistic-looking saturation logs. Note that there is no theoretical justification for this rescaling and using these particular values. Rather we found that these values produced synthetic CO_2 saturation logs with reasonable maximum values and variations within the reservoir.

In addition to adding and multiplying the log perturbations to the extracted models at the well locations, 5% time-varying random noise was added to each of the simulated logs to simulate time-lapse and spatially varying errors and noise in the log data acquisition. As a last note, the log depths were modified between the injection wells, MW1, MW2 and MW3, to account for the moderate dip apparent in the Kimberlina model stratigraphy.

The final suite of well logs was created to correspond to all geophysical models created from zero to 20 years. Thus, synthetic well log data exist at the injection well (INJ) and the three monitoring wells (MW1, MW2, MW3) for years 0, 1, 2, 5, 10, 15 and 20. Note that we do not provide CO_2 saturation logs for Year-0 as there is no injection at that time to warrant this. The injection well (INJ) will likely be steel-cased, which will not allow a resistivity log to be successfully acquired. Hence, there are no synthetic time-lapse resistivity logs for this well. However, we provide time-lapse synthetic density, sonic and saturation logs for INJ, as these logs can be acquired through steel casing.

The synthetic well logs examples for the Year-0 and Year-20 results for MW1 are shown in Figure 16. Figure 16a represents the pre-injection state, while the logs in Figure 16b are the synthetic logs after 20 years of injection. As expected, changes in the density log due to injection are fairly small, while the changes in sonic velocity are more substantial and apparent to the naked eye. The changes in the resistivity logs, on the other hand, are easy to see. Note that the CO_2 saturation log not only appears realistic with saturations confined to the reservoir, but also clearly show the three separate units of the reservoir.

The files containing the converted well logs are in two Excel formats *.csv and *.xlsx. They were created for four wells (INJ, MW1, MW2 and MW3) and seven times (0, 1, 2, 5, 10, 15 and 20 years after start of injection). The data are arranged in columns of depth, CO_2 saturation, density, sonic velocity and resistivity. The *.xlsx files also contain plots of the logs embedded in the spreadsheet.

5 | CT CORE IMAGES DURING CO_2 FLOOD EXPERIMENTS

While the Kimberlina site has proven to be extremely useful for numerical models, it is not an active carbon storage location. As such, no core from this site is directly applicable to upscaling and application to models of the site. Core samples from the Vedder sandstone in the general region of

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

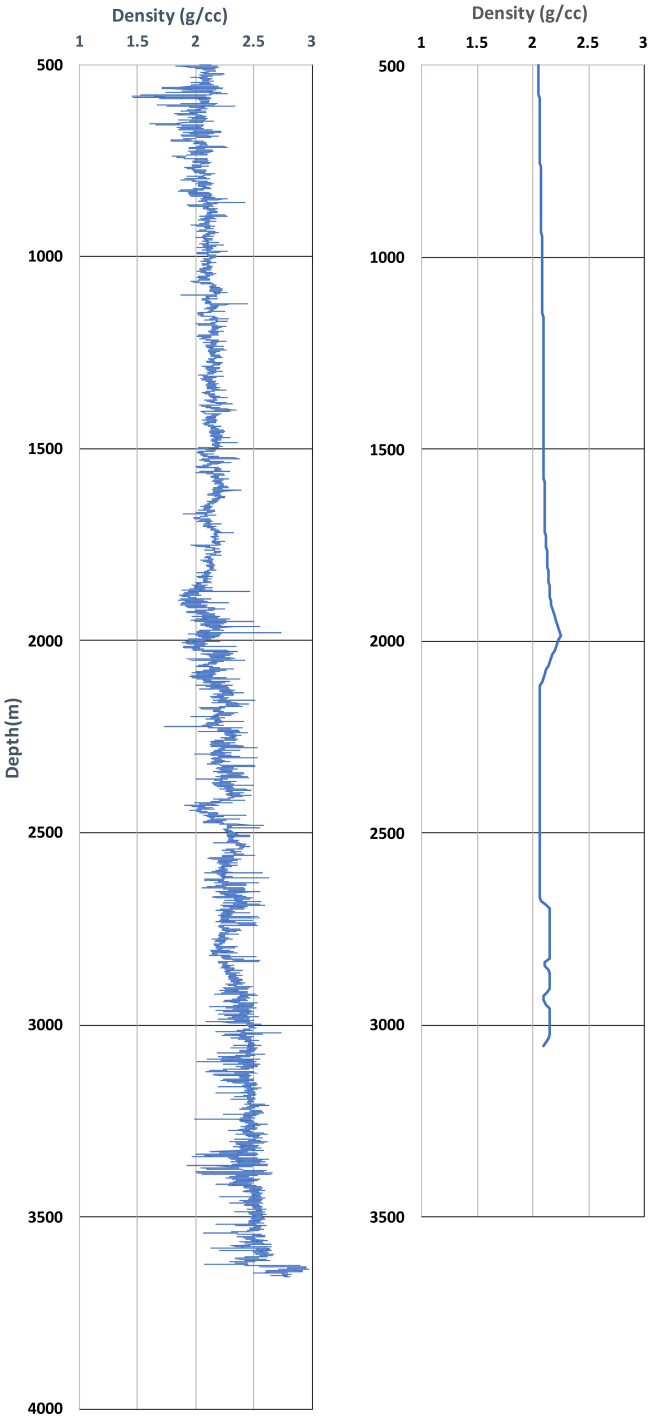

FIGURE 15 Density logs – (a) actual, (b) from the geophysical model.

the originally proposed Kimberlina injection site (Downey & Clinkenbeard, 2005) were obtained from Liaosha Song at the California State University at Bakersfield. The core was originally collected from the Round Mountain Well #1 (API: 04–029-83701) in Kern County, California cored from a depth of 792 m to 1,203 m, fully capturing the Vedder sandstone at this location (from 878 m to 1,181 m). Properties of the four samples are recorded in Table 2.

Dry industrial CT scans were conducted to capture the millimetre to centimetre scale structural features of the samples at NETL using a North Star Imaging M-5000 industrial scanner with a Feinfocus FX variable voltage source and Perkin Elmer detector. Variation in bedding plane porosity and mineral content was apparent from these laminated samples at this resolution. Each digital volume was obtained with source settings of 185 kV and 200 μA. Samples were rotated 360 degrees with 1,440 projections captured, each averaged from 10 radiographs.

Higher resolution images of subsamples from these core sections were obtained at NETL using a Zeiss Xradia micro CT scanner. Each digital volume was obtained with source settings of 150 kV and 66 μA, and using the optical enhancement lens of this system scans of 4.797 and 1.768 micron/voxel resolution were obtained of samples from the following depths: 1,079 m, 1,145 m and 1,155 m.

The Trainable WEKA Segmentation plugin for ImageJ (Arganda-Carreras et al., 2017) was utilized to isolate the porosity within these higher resolution scans and ranged from 9.33% at 1,155 m to 2.41% at 1,079 m. Samples were attempted to be saturated with brine, followed by $scCO_2$ (supercritical) injection to interrogate how CO_2 transmitted and resided in the pore space. The methodology detailed by Dalton et al. (2018) describes the injection of $scCO_2$ through a brine-filled core, followed by brine imbibition to residual conditions. Sample #1 did not permit the transmission of $scCO_2$ under a differential pressure across the core of over 100 psi and had to be abandoned as not permeable enough for laboratory multiphase examination in this fashion. Sample #4 was able to be examined in detail, as both $scCO_2$ and brine were able to be injected and $scCO_2$ was trapped in pore spaces following a brine imbibition step (Figure 17). Full datasets of these images are available for additional analyses.

Data of multi-scale CT images of the Vedder sandstone from the Round Mountain Well #1 (API: 04–029-83701) are in 16-bit tif stacks.

6 | DATA SHARING: EDX DIRECTORY AND FILE STRUCTURE

Simulation models and geophysical data reside on the National Energy Technology Laboratory Energy Data Exchange (EDX) website (https://edx.netl.doe.gov/group/kimberlina-geophysical-data; DOI: 10.18141/1887287). The main page (https://edx.netl.doe.gov/dataset/kimberlina-1-2-ccus-geophysical-models-and-synthetic-datasets) provides a description and links to individual files. Models of CO2_saturation, vp_velocity, vs_velocity, density and resistivity are divided into three part submission (part 1–3) to make file download faster. Each of these

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

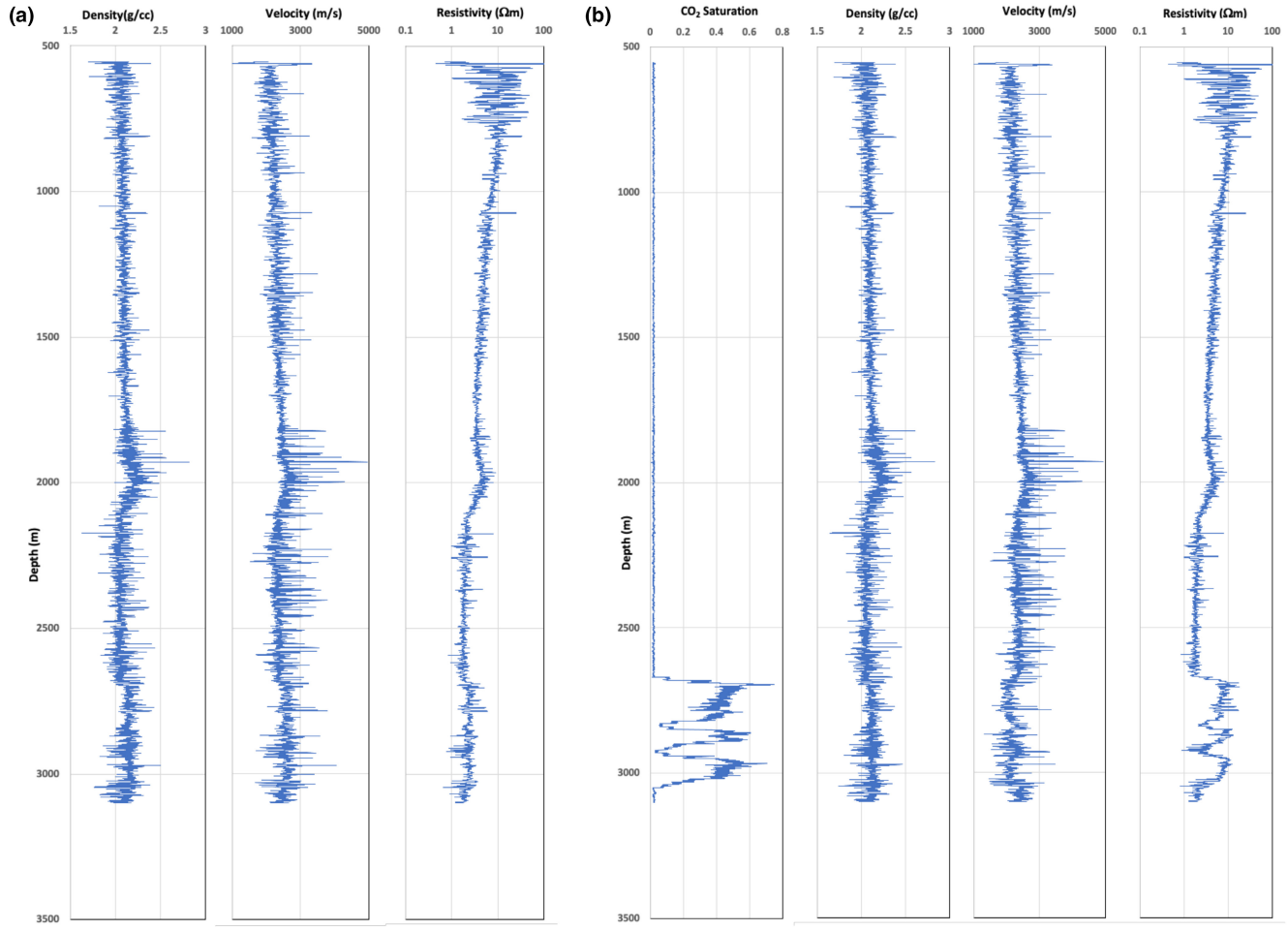

FIGURE 16 (a) Synthetic well logs of density, velocity and resistivity before the injection. (b) Synthetic well logs of CO2 saturation, density, velocity and resistivity after 20 years of CO2 injection.

TABLE 2 Properties of Vedder Sandstone core samples.

Sample #	Depth (m)	Resolution (micron/voxel)
1	1,079	12.4
2	1,145	30.9
3	1,146	33.5
4	1,155	9.1

submissions contains 33 or 34 simulations (100 simulations in total) called simDDD, where DDD is the simulation number (e.g., SIM001 corresponding to simulation 1). The simDDD zip-compressed file contains 33 zip-files containing attributes, CO_2 saturation, saturated density, seismic velocities (V_P and V_S) and resistivity, respectively, for each simulation time step (33-time steps). These files are ASCII csv-files using the naming convention *simmDDD_attributename_timestep*.csv.*zip*, where *attributename* is the calculated attribute for a specific *timestep* of the simulation. This naming format leads to the v_p values for year 80 of the simulation 1 found in the file sim001_vp_080y.csv.zip or resistivity values for year 100 of simulation 50 found in the file sim050_resisitivity_100y.csv.zip.

The main page contains also the links to seismic, EM, gravity, well_logs, and Vedder_CT_images data.

The seismic data submission includes both 2D and 3D data and models used to simulate those data. For the 2D case, there are 33 time-steps of the SIM001, and each time-step has 53 2D slices. For each model, seismic data were calculated for six sources. The naming of files follows the format of csg_year*{n}*_slide2D_*{m}*_*{s}*.bin for seismic data, and vp_year*{n}*_slide*{m}*.bin for velocity models. *{n}* denotes the simulation year (0–200), *{m}* denotes the slice index (1–53), and *{s}* denotes the source index (1–6). For example, csg_year10_slide2D_22_3.bin is a data file for year 10, 22nd 2D slice, and third source, and vp_year100_slide33.bin is a v_p velocity model for year 100 and 33rd 2D slice.

For the 3D case, the velocity models are stored as vp_year*{m}*.zip (e.g., vp_year200.zip), each of which contains 63 binary files (cut1-cut63) (e.g., year200_cut60.

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

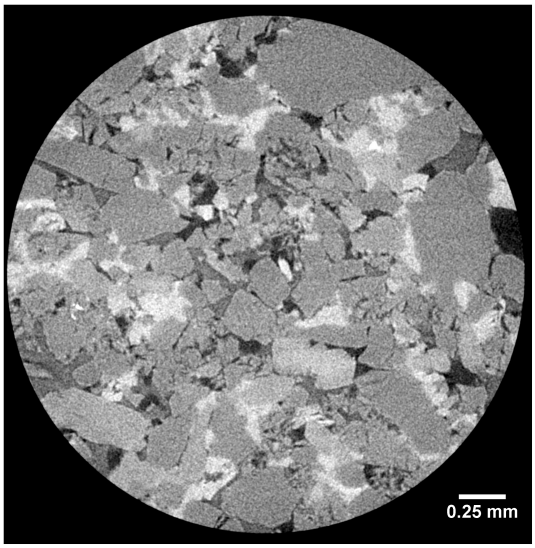

FIGURE 17 2D cross-sectional image of Vedder sandstone core with trapped/residual $scCO_2$ in pore spaces after brine injection to steady state.

bin), where *{m}* represents the simulation year. In order to save the large 3D seismic data, they are split into 25 files, one for each source (1–25). The data and models are in standard binary file format. The shape of the seismic data is 25 × 5,001 × 40 × 40 (Source number × Time × *X*-Direction × *Y*-Direction), where the time spacing is 0.001 s. The size of the model is 350 × 400 × 400 (Depth × *X*-Direction × *Y*-Direction), where the grid has a uniform 10 m interval in all directions.

2D gravity data contain responses of SIM001 along 2D profiles for 33 time-steps. 3D gravity data of the same model include surface responses and responses in four boreholes. EM data comprise of two 2D responses (for two borehole sources), and one set of 3D responses for 33 time-steps. The readme files describe the data file format. Well_logs resource contains the converted well logs in two Excel formats *.csv and *.xlsx. They were created for four wells (INJ, MW1, MW2, and MW3) and seven years (0, 1, 2, 5, 10, 15 and 20). The data are arranged in columns of depth, CO_2 saturation, density, sonic velocity and resistivity. The *.xlsx files also contain plots of the logs embedded in the spreadsheet. The Vedder_CT_images of the Vedder sandstone from the Round Mountain Well #1 (API: 04–029-83,701) are stored as 16-bit tif stacks in two resources: industrial CT images and Micro CT scans.

7 | CONCLUSIONS AND DISCUSSION

We use the Kimberlina 1.2 CO_2 reservoir simulations, based on a potential CO_2 storage site in California's Southern San Joaquin Basin, to produce geophysical models of P and S seismic velocities, saturated density, and electrical resistivity using established petrophysical relationships. We demonstrate the process using the baseline model at year 0 and the model after 20 years of CO_2 injection. These models and acquisition geometries that mimic actual monitoring surveys were used to generate synthetic time-lapse seismic, gravity and electromagnetic responses. We also created a series of synthetic well logs of CO_2 saturation, acoustic velocity, density and induction resistivity in the injection and three monitoring wells. The logs were constructed by combining the low-frequency trend of the geophysical models with the high-frequency variations of actual well logs collected at a potential storage site. In addition, to better calibrate our datasets, measurements of permeability and pore connectivity were made on cores of Vedder Sandstone, the primary reservoir unit. The combined dataset of the reservoir and geophysical models, simulated time-lapse geophysical responses, well logs and core scans, forms a multi-scale testbed for designing and evaluating geophysical CO_2 monitoring systems or imaging and characterization algorithms.

AUTHOR CONTRIBUTIONS

David Alumbaugh: Conceptualization (equal); project administration (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal). **Erika Gasperikova:** Data curation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Dustin Crandall:** Conceptualization (equal); data curation (equal); funding acquisition (equal); project administration (equal); writing – original draft (equal); writing – review and editing (equal). **Michael Commer:** Data curation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Shihang Feng:** Data curation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **William Harbert:** Data curation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Yaoguo Li:** Data curation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Youzuo Lin:** Data curation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Savini Samarasinghe:** Methodology (equal).

ACKNOWLEDGEMENTS

This work was completed as part of the Science-informed Machine learning to Accelerate Real Time decision making for Carbon Storage (SMART-CS) Initiative (edx.netl.doe.gov/SMART). Support for this initiative was provided by the US Department of Energy's (DOE) Office

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of Fossil Energy's Carbon Storage Research program through the National Energy Technology Laboratory (NETL). The authors wish to acknowledge Mark McKoy (NETL, Carbon Storage Technology Manager), Darin Damiani (DOE Office of Fossil Energy, Carbon Storage Program Manager), and Mark Ackiewicz (DOE Office of Fossil Energy, Director, Division of Carbon Capture and Storage Research and Development), for programmatic guidance, direction, and support. Technical effort for William Harbert was supported in part by appointment to the NETL Research Participation Program, sponsored by the US DOE and administered by the Oak Ridge Institute for Science and Education. We thank Dr. Zan Wang, for her generous support and sharing of Python code for this work, and Dr. Xianjin Yang for his valuable discussions, sharing of code and constructive suggestions.

FUNDING INFORMATION

The research in this paper was funded as part of the Science-informed Machine learning to Accelerate Real Time decision making for Carbon Storage (SMART-CS) Initiative (edx.netl.doe.gov/SMART). Support for this initiative was provided by the US Department of Energy's (DOE) Office of Fossil Energy's Carbon Storage Research program through the National Energy Technology Laboratory (NETL).

CONFLICT OF INTEREST STATEMENT

None of the authors has any conflict of interest regarding the publication of these results. No human or animal subjects were involved in this research.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally shareable data necessary to reproduce the reported results. The data is available at https://hdl.handle.net/2022/27098. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

ORCID

David Alumbaugh https://orcid.org/0000-0002-6975-7197
William Harbert https://orcid.org/0000-0002-0205-5772

REFERENCES

Alnes, H., Eiken, O., Nooner, S., Sasagawa, G., Stenvold, T. & Zumberge, M. (2011) Results from Sleipner gravity monitoring: updated density and temperature distribution of the CO 2 plume. *Energy Procedia*, 4, 5504–5511.

Archie, G.E. (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. *Transactions of AIME*, 146, 54–61.

Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K.W., Schindelin, J., Cardona, A. et al. (2017) Trainable Weka segmentation: a machine learning tool for microscopy pixel classification. *Bioinformatics*, 33(15), 2424–2426. Available from: https://doi.org/10.1093/bioinformatics/btx180

Avseth, P., Mukerji, T. & Mavko, G. (2007) *Quantitative Seismic Interpretation: Applying Rock Physics Tools to Reduce Interpretation Risk*. Cambridge: Cambridge University Press, p. 359.

Batzle, M. & Wang, Z. (1992) Seismic properties of pore fluids. *Geophysics*, 57(11), 1396–1408.

Birkholzer, J.T., Zhou, Q., Cortis, A. & Finsterle, S. (2011) A sensitivity study on regional pressure buildup from large-scale CO_2 storage projects. *Energy Procedia*, 4, 4371–4378.

Bonneville, A., Black, A.J., Hare, J., Kelley, M.E., Place, M. & Gupta, N. (2021) Time-lapse borehole gravity imaging of CO_2 injection and withdrawal in a closed carbonate reef. *Geophysics*, 86(6), 1–20.

Commer, M. & Newman, G.A. (2008) New advances in three-dimensional controlled-source electromagnetic inversion. *Geophysical Journal International*, 172, 513–535.

Constable, S. & Weiss, C.J. (2006) Mapping thin resistors and hydrocarbons with marine EM methods: insights from 1D modeling. *Geophysics*, 71(2), G43–G51.

Correa, J., Isaenkov, R., Yavuz, S., Yurikov, A., Tertyshnikov, K., Wood, T. et al. (2021) DAS/SOV: Rotary seismic sources with fiber-optic sensing facilitates autonomous permanent reservoir monitoring. *Geophysics*, 86(6), 1–42. Available from: https://doi.org/10.1190/Geo2020-0612.1

Dalton, L.E., Klise, K.A., Fuchs, S., Crandall, D. & Goodman, A. (2018) Methods to measure contact angles in $scCO_2$-brine sandstone systems. *Advances in Water Resources*, 122, 278–290. Available from: https://doi.org/10.1016/j.advwatres.2018.10.020

Dataset: Kimberlina. (n.d.) Regional static 3D geologic model of the southern San Joaquin Basin. https://doi.org/10.18141/1603335

Downey, C. & Clinkenbeard, J. (2005) *An overview of geologic carbon sequestration potential in California*. CGS Special Report 183, WESTCARB Topical Report under DOE-Contract No.: DE-FC26-03NT41984, https://www.osti.gov/servlets/purl/903323

Gasperikova, E., Appriou, D., Bonneville, A., Feng, Z., Huang, L., Gao, K. et al. (2022) Sensitivity of geophysical techniques for monitoring secondary CO_2 storage plumes. *International Journal of Greenhouse Gas Control*, 114, 103585, ISSN 1750-5836. Available from: https://doi.org/10.1016/j.ijggc.2022.103585

Gasperikova, E., Daley, T., Appriou, D., Bonneville, A., Feng, Z., Huang, L. et al. (2020) *Detection thresholds and sensitivities of geophysical techniques for CO2 plume monitoring*. NRAP-TRS-I-001–2020; DOE.NETL-2021.2638. NRAP Technical Report Series; US Department of Energy, National Energy Technology Laboratory: Pittsburgh, PA, 2020; p 64 https://doi.org/10.2172/1735331

Gassmann, F. (1951) Über die elastizität poröser medien. *Vierteljahrss-Chrift der Naturforschenden Gesellschaft in Zurich*, 96, 1–23 English translation at http://sepwww.stanford.edu/sep/berryman/PS/gassmann.pdf

Hayashi, M. (2004) Temperature-electrical conductivity relation of water for environmental monitoring and geophysical data

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

inversion. *Environmental Monitoring and Assessment*, 96, 119–128.

Hovorka, S.D., Meckel, T.A. & Trevino, R.H. (2013) Monitoring a large-volume injection at Cranfield, Mississippi – project design and recommendations. *International Journal of Greenhouse Gas Control*, 18, 345–360.

Krahenbuhl, R.K., Martinez, C., Li, Y. & Flanagan, G. (2015) Time-lapse monitoring of CO_2 sequestration: an integrated site investigation based on reservoir properties and seismic with borehole and surface gravity data. *Geophysics*, 80, WA15–WA24.

Kumar, D. (2006) A tutorial on Gassmann fluid substitution: Formulation, algorithm and Matlab code. *Matrix*, 2(1).

Lawton, D.C., Dongas, J., Osadetz, K., Saeedfar, A. & Macquet, M. (2019) Development and analysis of a geostatic model for shallow CO2 injection at the field Research Station. In: Davis, T.L., Landro, M. & Wilson, M. (Eds.) *Geophysics and geosequestration*. Southern Alberta, AB: Cambridge University Press.

McKenna, J.J., Gurevich, B., Urosevic, M. & Evans, B.J. (2003) Rock physics-application to geologic storage of CO2. *Appea Journal*, 43, 567–576.

Moczo, P., Robertsson, J.O. & Eisner, L. (2007) The finite-difference time-domain method for modeling of seismic wave propagation. *Advances in Geophysics*, 48, 421–516.

Pevzner, R., Isaenkov, R., Yavuz, S., Yurikov, A., Tertyshnikov, K., Shashkin, P. et al. (2021) Seismic monitoring of a small CO2 injection using a multi-well DAS array: operations and initial results of stage 3 of the CO2CRC Otway project. *International Journal of Greenhouse Gas Control*, 110, 103437.

Pruess, K. (2005) *ECO2N: a TOUGH2 fluid property module for mixtures of water, NaCl, and CO_2*. LBNL-57952. Berkeley, CA: Lawrence Berkeley National Laboratory.

Rim, H. & Li, Y. (2015) Advantages of borehole vector gravity in density imaging. *Geophysics*, 80, G1–G13.

Ringrose, P. (2020). *How to store CO2 underground: insights from early-mover CCS projects*, Berlin/Heidelberg, Germany: Springer International Publishing, pp. 129.

Sheriff, R.E. & Geldart, L.P. (1995) *Exploration seismology*. Cambridge, England: Cambridge University Press.

Torp, T.A. & Gale, J. (2004) Demonstrating storage of CO2 in geological reservoirs: the Sleipner and SACS projects. *Energy*, 29, 1361–1369.

Um, E. & Alumbaugh, D. (2007) On the physics of the marine controlled source electromagnetic method for subsurface hydrocarbon detection. *Geophysics*, 72, WA13–WA26.

Um, E., Alumbaugh, D., Lin, Y. & Feng, S. (2022) Real time deep learning inversion of seismic full waveform data for CO_2 saturation and uncertainty in geological carbon storage monitoring. *Geophysical Prospecting*. https://doi.org/10.1111/1365-2478.13197

Underschultz, J., Boreha, C., Dance, T., Stalker, L., Friefeld, B., Kirste, D. et al. (2011) CO_2 storage in a depleted gas field: an overview of the CO2CRC Otway project and initial results. *International Journal of Greenhouse Gas Control*, 5, 922–932.

Wagoner, J. (2009) *3D geologic modeling of the southern San Joaquin Basin for the Westcarb Kimberlina demonstration project - a status report*. Lawrence Livermore National LaboratoryLLNL-TR-410813, 27 https://doi.org/10.2172/957164

Wainwright, H.M., Finsterle, S., Zhou, Q. & Birkholzer, J.T. (2013) Modeling the performance of large-scale CO_2 storage systems: a comparison of different sensitivity analysis methods. *International Journal of Greenhouse Gas Control*, 17, 189205. Available from: https://doi.org/10.1016/j.ijggc.2013.05.007

Walton, N. (1989) Electrical conductivity and total dissolved solids - what is their precise relationship? *Desalination*, 72, 275–292.

Wang, Z., Huang, L., Dilmore, R. & Harbert, W. (2018) Modeling of time-lapse seismic monitoring using CO_2 leakage simulations for a model CO_2 storage site with realistic geology: application in assessment of early leak-detection capabilities. *International Journal of Greenhouse Gas Control*, 76, 39–52.

Wirianto, M., Mulder, W.A. & Slob, E.C. (2010) A feasibility study of land CSEM reservoir monitoring in a complex 3-D model. *Geophysical Journal International*, 181, 741–755. Available from: https://doi.org/10.1111/j.1365-246X.2010.04544.x

Worth, K., White, D., Chalatrunyk, R., Sorensen, J., Hawkes, C., Rostron, B. et al. (2014) Aquistore project measurement, monitoring, and verification: from concept to CO_2 injection. *Energy Procedia*, 63, 3202–3208.

Wu, Y. & Lin, Y. (2019) InversionNet: An Efficient and Accurate Data-driven Full Waveform Inversion. *IEEE Transactions on Computational Imaging*, 6(1), 419–433.

Yang, X., Buscheck, T.A., Mansoor, K., Wang, Z., Gao, K., Huang, L. et al. (2019) Assessment of geophysical monitoring methods for detection of brine and CO_2 leakage in drinking water aquifers. *International Journal of Greenhouse Gas Control*, 90, 102803. Available from: https://doi.org/10.1016/j.ijggc.2019.102803

Zeng, Q., Feng, S., Wohlberg, B. & Lin, Y. (2022) Inversionnet3D: efficient and scalable learning for 3d full waveform inversion. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–16.

Zhang, K., Wu, Y.S. & Pruess, K. (2008) *User's guide for TOUGH2-MP – a massively parallel version of the TOUGH2 code*. Report LBNL-315E. Berkeley, CA: Lawrence Berkeley National Laboratory.

How to cite this article: Alumbaugh, D., Gasperikova, E., Crandall, D., Commer, M., Feng, S., Harbert, W. et al. (2023) The Kimberlina synthetic multiphysics dataset for CO_2 monitoring investigations. *Geoscience Data Journal*, 00, 1–19. Available from: https://doi.org/10.1002/gdj3.191

20496060, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/gdj3.191 by Univ of California Lawrence Berkeley National Lab, Wiley Online Library on [15/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License