Transport in Porous Media
https://doi.org/10.1007/s11242-023-01977-7

The FluidFlower Validation Benchmark Study for the Storage of CO_2

Bernd Flemisch[1] · Jan M. Nordbotten[2,3] · Martin Fernø[3,4] · Ruben Juanes[5] · Jakub W. Both[2] · Holger Class[1] · Mojdeh Delshad[6] · Florian Doster[7] · Jonathan Ennis-King[8] · Jacques Franc[9] · Sebastian Geiger[7,10] · Dennis Gläser[1] · Christopher Green[8] · James Gunning[8] · Hadi Hajibeygi[10] · Samuel J. Jackson[8] · Mohamad Jammoul[6] · Satish Karra[11] · Jiawei Li[9] · Stephan K. Matthäi[12] · Terry Miller[11] · Qi Shao[12] · Catherine Spurin[9] · Philip Stauffer[11] · Hamdi Tchelepi[9] · Xiaoming Tian[10] · Hari Viswanathan[11] · Denis Voskov[10] · Yuhang Wang[10] · Michiel Wapperom[10] · Mary F. Wheeler[6] · Andrew Wilkins[13] · AbdAllah A. Youssef[12] · Ziliang Zhang[10]

Received: 16 January 2023 / Accepted: 17 June 2023

Abstract
Successful deployment of geological carbon storage (GCS) requires an extensive use of reservoir simulators for screening, ranking and optimization of storage sites. However, the time scales of GCS are such that no sufficient long-term data is available yet to validate the simulators against. As a consequence, there is currently no solid basis for assessing the quality with which the dynamics of large-scale GCS operations can be forecasted. To meet this knowledge gap, we have conducted a major GCS validation benchmark study. To achieve reasonable time scales, a laboratory-size geological storage formation was constructed (the "FluidFlower"), forming the basis for both the experimental and computational work. A validation experiment consisting of repeated GCS operations was conducted in the FluidFlower, providing what we define as the true physical dynamics for this system. Nine different research groups from around the world provided forecasts, both individually and collaboratively, based on a detailed physical and petrophysical characterization of the FluidFlower sands. The major contribution of this paper is a report and discussion of the results of the validation benchmark study, complemented by a description of the benchmarking process and the participating computational models. The forecasts from the participating groups are compared to each other and to the experimental data by means of various indicative qualitative and quantitative measures. By this, we provide a detailed assessment of the capabilities of reservoir simulators and their users to capture both the injection and post-injection dynamics of the GCS operations.

✉ Bernd Flemisch
bernd@iws.uni-stuttgart.de

Extended author information available on the last page of the article

Published online: 18 August 2023

Keywords Geological carbon storage · Validation benchmark · Validation experiment · Model intercomparison

1 Introduction

Geological carbon storage (GCS) has the potential to close the gap between CO_2 emissions from legacy carbon-based power sources and the required emission reductions as outlined in the IPCC reports (Bachu et al. 2007; Pacala and Socolow 2004; Halland et al. 2013; Metz et al. 2005). Furthermore, GCS can play a role in negative emissions strategies in combination with biofuels (Johnson et al. 2014), and in the production of so-called "blue hydrogen" (Noussan et al. 2021). In order to realize this potential in a safe and cost-efficient manner, large-scale deployment of GCS relies heavily on modeling and numerical simulation studies to assess the suitability of potential geological formations (predominantly subsurface aquifers). Such modeling studies have been heavily relied upon in existing assessments of storage potential (Juanes et al. 2010; Lindeberg et al. 2009; Kopp et al. 2009a, b; Niemi et al. 2016; Sharma et al. 2011). The generation of simulation-based data and knowledge in application fields like GCS with huge societal impact eventually requires communication to political decision makers. Transparent simulation work flows, reproducibility of data and increased confidence in simulation results, e.g. as a result of comprehensive benchmarking, are key factors for communication or participation of stakeholders in the modeling process (Scheer et al. 2021).

On the other hand, only a few dozen large-scale carbon storage operations are currently active globally (Steyn et al. 2022), and of these, none are in a post-injection phase following a multi-decadal injection period. As such, the modeling and simulation community does not have a robust data set to assess their forecasting skill, and significant uncertainty is associated with our ability to accurately capture the dominant physical processes associated with GCS. Pilot studies provide some measure of information (Sharma et al. 2011; Preston et al. 2005; Hovorka et al. 2006; Lüth et al. 2020; Niemi et al. 2020), yet the fundamental nature of the subsurface means that the data collected will always be relatively sparse, in particular spatially. As a partial remedy to this, several code comparison studies have been conducted (Pruess et al. 2004; Class et al. 2009; Nordbotten et al. 2012). However, none of these studies were conducted in the presence of a physical ground truth.

This study aims to provide a first assessment of the predictive skills of the GCS modeling and simulation community. To achieve this goal, we are exploiting the newly constructed "FluidFlower" experimental facility at the University of Bergen. Within this experimental rig, a geological model with characteristic features from the Norwegian Continental Shelf was constructed. Initial geological and petrophysical characterization was completed, together with a single-phase tracer test. With this basis, we conducted a double-blind study: On one hand, laboratory scale GCS was repeatedly conducted and measured at the University of Bergen, where the corresponding group will be labelled as `ExpUB` in the following. On the other hand, academic research groups active in GCS around the world were invited to participate in a validation benchmark study, coordinated by the University of Stuttgart, in the following indicated by `CoordUS`. Aided by the fact that the pandemic reduced academic travel, we were able to fully ensure that no physical interaction was present between the participating groups, and all digital communication was restricted, moderated, and archived to ensure the integrity of the double-blind study. As detailed in the following, the participants

of the study were both asked to provide independent forecasts, and then subsequently invited to update their forecasts in view of group interactions.

In this contribution, we report the final results of the validation benchmark study, emphasizing (1) The degree of correlation between forecasts from the diverse set of participating groups, and (2) The degree of correlation between the forecasts and the measurements from the laboratory scale GCS conducted in the FluidFlower. Seen together, this provides both a measure of repeatability among forecasts (seen from an operational perspective), and also an indication of forecasting skill.

The paper contains a substantial amount of results, projected onto axes. In particular, the participants provided dense spatial results (sparse in time), time-series of integrated quantities (sparse in space), and certain predefined target quantities (sparse data). These simulated quantities are naturally compared both to each other, as well as to the experiment conducted in parallel. Substantial discussion can therefore be considered throughout the exposition. However, we have endeavored to provide the results in a relatively factual manner in Sects. 3–5, thus reserving the majority of the discussion for Sect. 6. Readers mostly interested in the high-level findings of the study may therefore choose to read Sect. 6 first.

To be precise, we structure the paper as follows. Section 2 introduces some basic required terminology, describes the validation experiment, and illustrates the benchmarking process. The participating groups and corresponding models are introduced in Sect. 3. In Sect. 4, the modeling results are presented and compared by means of qualitative and quantitative assessments. Section 5 provides a comparison of the modeling results with the experimental data. A concluding discussion and outlook are given in Sect. 6.

2 Benchmarking Methodology

We start this section by introducing some fundamental concepts and terminology based on Oberkampf and Roy (2010), American Society of Mechanical Engineers (2006). While the term *verification* describes "the process of determining that a computational model accurately represents the underlying mathematical model and its solution", *validation* refers to "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model". In addition, *calibration* is the process of adjusting parameters in the computational model to improve agreement with data.

A *validation experiment* like the one presented below in Sect. 2.1 is "designed, executed, and analyzed for the purpose of quantitatively determining the ability of a mathematical model expressed in computer software to simulate a well-characterized physical process". As described in further detail below in Sect. 2.2, we perform a *validation benchmark* (Oberkampf and Trucano 2008; Oberkampf and Roy 2010), where the experiment provides measured data against which the simulation results are to be compared. The simulation results are forecasts in the sense that the experimental results are unknown to the modeler.

2.1 The Validation Experiment

In the following, we provide a very brief description of the experiment performed with the FluidFlower rig. For details, we refer to the original benchmark description (Nordbotten et al. 2022) and the experimental paper (Fernø et al. 2023). Figure 1 shows the geometrical setup where the rig has been filled with six different sands to build up several layers of varying permeability, including three fault-like structures.

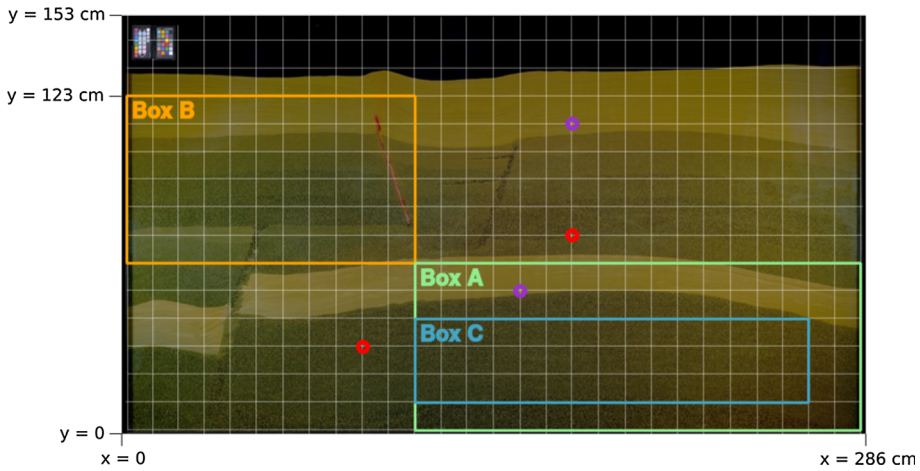

Fig. 1 Photograph image of the validation benchmark geometry with overlaid laser grid (Nordbotten et al. 2022, Figure 8). The brightest facies are the fine-sand barriers. The red circles indicate the injection points, while the purple circles depict the locations of pressure sensors. Boxes A-C correspond to regions for the evaluation of different system response quantities

Initially, the pore space was fully water-saturated and the top of the water table was subject to atmospheric conditions in terms of pressure and temperature. Gaseous CO_2 was injected over five hours at a rate of ten standard milliliters per minute through the lower left injection port, and, beginning 2:15 h after the start of the first injection, over 2:45 h at the same rate through the upper right port. The distribution of CO_2 throughout the rig was monitored over five days after the injection start. In total, five experimental runs were performed between November 2021 and January 2022. The experimental team `ExpUB` tried to establish identical operational conditions during the runs.

The description of the experimental setup in Nordbotten et al. (2022) addressed the external geometry, stratification, facies properties, faults, fluid properties, operational conditions and well test data. In particular, the stratification was described by high-resolution photographs, from which the participating groups had to determine the location of the different sand layers. This was complemented by details on the sedimentation process and pre-injection flushing procedures. Concerning the facies, information was provided on grain size distributions as well as on measurements of absolute permeability, porosity, relative permeability endpoints and capillary entry pressures, see Section A.1 in the appendix for the most important spatial parameters. The purpose of the well test data was to allow for calibration of the numerical models. In particular, the provided pressure[1] and tracer flow data could be employed to estimate the permeability distribution over the different facies.

The experimental setup, while at atmospheric pressure and room temperature, nevertheless shares characteristic dimensionless groups with real geological storage sites, as discussed in detail in Kovscek et al. (2023). The main differences, as relevant from the perspective of a validation benchmark study, are discussed in detail in Sect. 6.

The description also defined the *System Response Quantities* (SRQs) which should be reported by the participants. The individual SRQs will be introduced in detail in Sect. 4.

[1] The injection pressures were reported at a sensor that was separated from the injection point by the length of small diameter tubing. Taking the pressure drop along that tubing into account influences the result of the calibration.

Table 1 Chronology of the FluidFlower validation benchmark process

30.04.2021	Potential participants are invited
15.06.2021	Participation invitation expires
15.07.2021	Preliminary benchmark description supplied to participants
16.07.–19.08.2021	Preparation phase, discussion possible among all participants and ExpUB
20.08.2021	Deadline for feedback on preliminary benchmark description
16.09.2021	Kick-off Zoom meeting, second iteration of benchmark description distributed
17.09.–08.10.2021	Open discussion for finalizing the description
08.10.2021	Final benchmark description circulated to participants
09.10.2021–11.01.2022	Blind phase, no direct communication between different participants or with ExpUB
09.01.2022	Deadline for submitting blind benchmark data
12.01.2022	Virtual workshop and comparison of "fully blind" simulation forecasts
12.01.–25.04.2022	Synchronization phase, communication between all participants enabled, but not with ExpUB
22.04.2022	Deadline for submitting final benchmark data
26.–28.04.2022	Workshop in Norway with presentation of final simulation forecasts, experimental results, model calibration study, and synthesis of results

2.2 Benchmarking Process

Table 1 shows the chronology of the benchmarking process. After a common preparation phase for finalizing the description (Nordbotten et al. 2022), a so-called blind phase of three months started, where there was no direct communication between different participating groups or with ExpUB allowed.

All upcoming issues of the modelers were directed to CoordUS and potentially anonymously forwarded to ExpUB. After agreeing on an answer between CoordUS and ExpUB, that answer was either broadcasted to all participating groups or given to the questioner only. At the end of the blind phase, each participating group provided initial forecasts to CoordUS. This was followed by a first meeting of all participating groups where the results were revealed and discussed, still without any involvement of ExpUB. This meeting initiated a so-called synchronization phase of another three months, allowing the forecasting groups to learn from each other's work and bring this knowledge into their own forecasts. In particular, the synchronization phase included two more common participant meetings. At its end, final forecasts were recorded before an in-person workshop outside of Bergen, Norway, where forecasts and experiments were compared for the first time.

In order to protect the integrity of the results, dedicated communication rules were followed during the different phases of the benchmarking process. To facilitate remote communication between participants, and also to store this communication for evaluating the benchmarking process, a Discord server was set up.[2] Apart from a general channel that was initially open to everyone involved, a private channel was installed for each participating group which could be used for communicating with the benchmark organizers.

All result data was uploaded by the participants to Git repositories within a GitHub organization "FluidFlower".[3] Each participating group got write access to a dedicated repository

[2] https://discord.gg/8Q5fZS3T47.

[3] https://github.com/fluidflower.

named after their institution. During the blind phase, only the participants themselves had access to their respective repositories. For the synchronization phase, read access to all participant repositories was granted for all participants. After the workshop in April, the repositories were opened further to include also the results from the physical experiments. Upon submission of this paper, the relevant repositories have been turned public.

3 Participating Groups and Models

In total, nine groups, each consisting of two to five individuals, participated in the Fluid-Flower validation benchmark study. In the following, they are indicated by the location or name of the corresponding institution as `Austin` (M. Delshad, M. Jammoul, M.F. Wheeler), `CSIRO` (J. Ennis-King, C. Green, J. Gunning, S.J. Jackson, A. Wilkins), `Delft-DARSim` (H. Hajibeygi, Y. Wang, Z. Zhang), `Delft-DARTS` (X. Tian, D. Voskov, M. Wapperom), `Heriot-Watt` (F. Doster, S. Geiger), `LANL` (S. Karra, T. Miller, P. Stauffer, H. Viswanathan), `Melbourne` (S.K. Matthäi, Q. Shao, A.A. Youssef), `Stanford` (J. Franc, J. Li, C. Spurin, H. Tchelepi) and `Stuttgart` (H. Class, D. Gläser). Table 2 lists relevant modeling choices of the participating groups.

In terms of the partial differential equations constituting the main part of the mathematical model, almost all participants employ component mass balances. Apart from two exceptions `Austin` and `Melbourne`, the choice of spatial discretization is uniform with cell-centered finite volumes. All groups except `Melbourne` employ a standard implicit Euler time discretization and solve the resulting discrete equations in a fully-coupled fully-implicit manner. Modeling choices start to differ more when it comes to the constitutive relations. While the majority of the participants uses Brooks–Corey relationships for the capillary pressure and relative permeability, also other approaches such as linear relationships are employed. Moreover, various equations of state for determining the phase compositions as well as the phase densities are considered. Additionally to these principal choices, the participating computational models differ in their employed spatial parameters such as the assumed intrinsic permeabilities, porosities, residual saturations and others. These parameters may depend on the considered sand type, i.e., on the spatial location. They have been collected for each participating group in a file `spatial_parameters.csv` in the top level of the respective GitHub repository and are also provided as tables in Sect. A.2. While the participants mostly followed the parameters provided by `ExpUB`, some groups varied the intrinsic permeability values as the result of a model calibration step. Depending on the type of relationships for capillary pressure and relative permeability, additional parameters such as the Brooks-Corey pore-size distribution index had to be selected.

4 Modeling Results

In the following, we provide and discuss the modeling results which were requested in form of SRQs by the benchmark description and submitted as final forecasts at the end of the synchronization phase. They are grouped into three categories: dense data spatial maps in Sect. 4.1, dense data time series in Sect. 4.2, and sparse data in Sect. 4.3.

Table 2 Modeling choices of the participating groups

Type	Austin	CSIRO	Delft-DARSim	Delft-DARTS	Heriot-Watt	LANL	Melbourne	Stanford	Stuttgart
PDEs	CMB	CMB	CMB	CMB	pseudo black oil	CMB	pseudo black oil + transport of dissolved CO_2	pseudo black oil	CMB
p_c, k_r	BC	BC	BC	const. p_E, power law	power law for rel perms, BC for fine sands, vG for coarse sands	linear	BC	BC	BC
EOS	Peng and Robinson (1976)	Spycher et al. (2003), Spycher and Pruess (2005)	Spycher et al. (2003)	Liquid: Ziabakhsh-Ganji and Kooi (2012), gas: Peng and Robinson (1976)		Duan and Sun (2003)	Span and Wagner (1996), Span and Wagner (2003)	Weiss (1974), Sandve et al. (2021), Duan and Sun (2003)	Spycher and Pruess (2005), Duan and Sun (2003), Spycher et al. (2003)
Density	Peng and Robinson (1976)	Liquid: IAPWS (2007), Garcia (2001)), gas: Span and Wagner (1996)	$\varrho_l = (\varrho_b^{STC} + \varrho_{CO2}^{STC} R_S)/B_b$, Soave (1972)	Liquid: IAPWS (2007), Garcia (2001), gas: Peng and Robinson (1976)	Liquid: exp. with p, linear with CO_2 conc., gas: exp. with p	Span and Wagner (1996)	Derived from miscibility data reported in Carroll et al. (1991)	Fenghour et al. (1998), Span and Wagner (1996)	IAPWS (2007), Span and Wagner (1996)

Table 2 continued

Type	Austin	CSIRO	Delft-DARSim	Delft-DARTS	Heriot-Watt	LANL	Melbourne	Stanford	Stuttgart
Solubility limit [kgm^{-3}]	1.496	1.786	1.649	1.9	2.0	2.0	$f(p)$, 1.752 to 2.0093	1.5	1.845
Domain volume [m^2]	9.1e−2	8.65e−2	8.3e−2	9.2e−2	8.4e−2	8.4e−2	8.18e−2	8.4e−2	8.75e−2
Disc.	MFEM	CC-FV	CC-FV	CC-FV	CC-FV	CC-FV	DFEFVM	CC-FV	CC-FV
# cells	9,100	44,284	43,758	48,274	42,000	42,000	14,822	6,094 / 21,392	26,099
avg. cell diam. [cm]	2.83	1.26	1.40	1.21	1.43	1.29	2.18	3.40 / 1.81	1.64
Software	IPARS	MOOSE (Wilkins et al. 2021)	DARSim (Wang et al. 2022)	DARTS (Lyu et al. 2021)	MRST-2021b (Lie 2019)	FEHM (Zyvoloski et al. 1997), PFLOTRAN (Lichtner et al. 2015)	CSMP++ (Matthäi et al. 2001)	AD-GPRS (Zhou et al. 2013; Garipov et al. 2018; Zhou 2012; Younis et al. 2010)	$DuMu^x$ (Koch et al. 2021)

Component mass balances "CMB", phase mass balances "PMB", Brooks–Corey "BC", van Genuchten "vG", mixed finite elements "MFEM", cell-centered finite volumes "CC-FV", collocated finite-element finite-volume method with embedded discontinuities "DFEFVM"

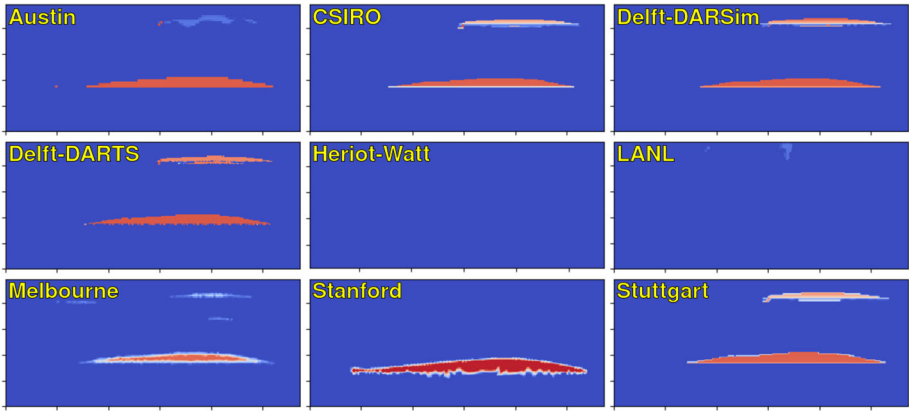

Fig. 2 Spatial distribution of gaseous CO_2 after 24 h. The minimum for the color map is at 0 CO_2 saturation indicated by blue, the maximum at 1 indicated by red

4.1 Dense Data Spatial Maps

The participants were asked to provide snapshots of the spatial phase distribution at 24, 48, 72, 96 and 120 h (hours) after injection start, particularly, the saturation of gaseous CO_2 as well as the concentration of CO_2 in the liquid phase. While each participating group was free to define the computational grid for performing simulations, results should be reported on a uniform grid consisting of 1cm by 1cm cells, extending from (0, 0) to (286cm, 123cm) cf. to Fig. 1.

4.1.1 Saturation

Figures 2, 3, 4, 5 and 6 visualize the reported saturation values for all participating groups at the selected daily time steps. Focusing first on Fig. 2, it can be observed that most participants report a very similar CO_2 plume shape under the lower fine sand barrier after 24 h.

Moreover, no or almost no gaseous CO_2 is reported within Box B in the upper left (cf. Fig. 1) after one day. Considerably less agreement can be seen for the upper barrier in the right part of the domain. This can be explained by the fact that the amount of CO_2 injected in the lower and upper part differs by a factor of more than 2 and, correspondingly, a variation in the dissolution behavior becomes visible earlier in the upper part of the domain.

The two participants `Heriot-Watt` and `LANL` report that no or almost no gaseous CO_2 is present throughout the domain after the first day of simulation. In case of `Heriot-Watt`, this is due to the choice of the van-Genuchten relationship for the capillary pressure, as explained in more detail below in Sect. 4.2. The reported results from `Heriot-Watt` are the ones with the smallest capillary fringe that was possible to resolve within the computing time constraints and an overestimation of dissolution was anticipated. The situation is different for `LANL`, where CO_2 leaves the system because almost no trapping occurs, see also below.

Examining the saturation distributions over the different time steps in Figs. 2, 3, 4, 5 and 6 reveals the effect of differences in modeling CO_2 dissolution in aqueous phase.

In particular, `CSIRO`, `Delft-DARSim`, `Delft-DARTS` and `Melbourne` report a vanishing CO_2 gas plume over time, while the plume shape stays rather constant for `Austin`,

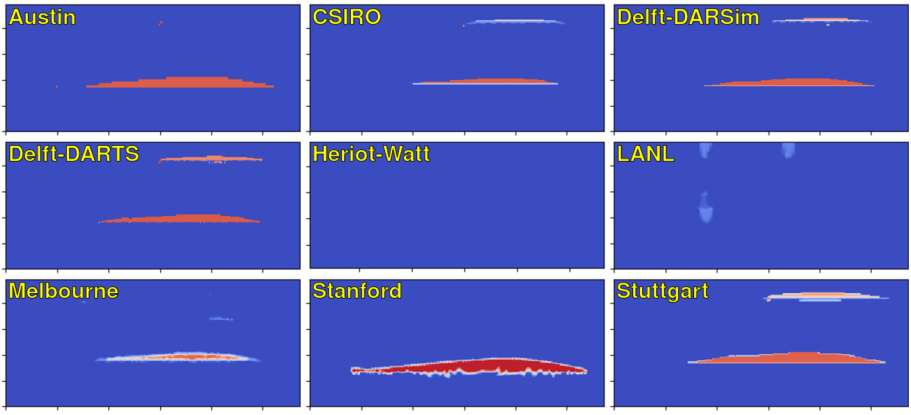

Fig. 3 Spatial distribution of gaseous CO_2 after 48 h. The minimum for the color map is at 0 CO_2 saturation indicated by blue, the maximum at 1 indicated by red

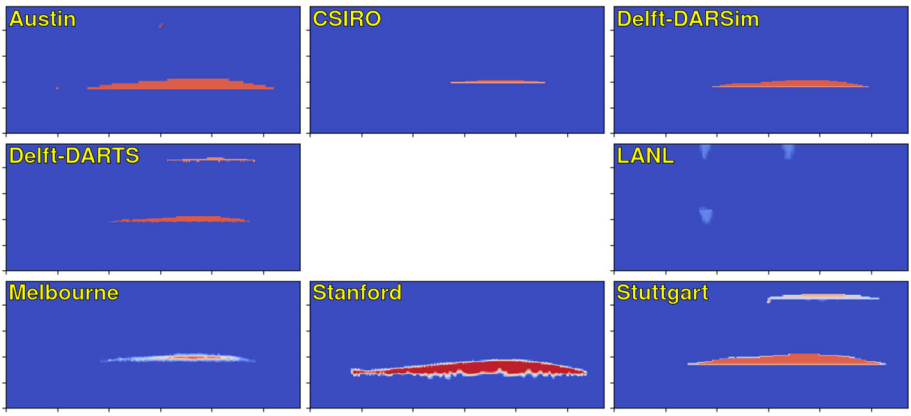

Fig. 4 Spatial distribution of gaseous CO_2 after 72 h. The minimum for the color map is at 0 CO_2 saturation indicated by blue, the maximum at 1 indicated by red

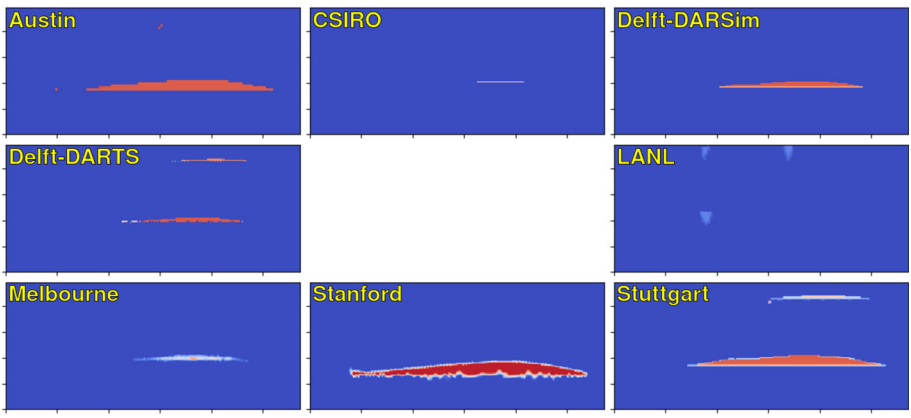

Fig. 5 Spatial distribution of gaseous CO_2 after 96 h. The minimum for the color map is at 0 CO_2 saturation indicated by blue, the maximum at 1 indicated by red

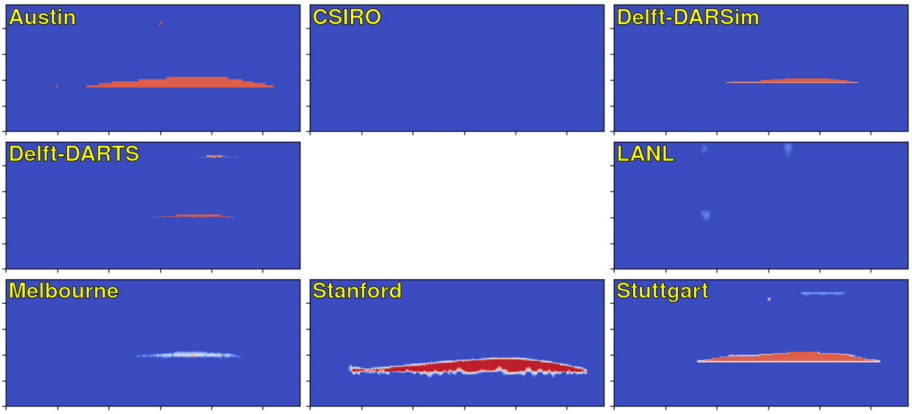

Fig. 6 Spatial distribution of gaseous CO_2 after 120 h. The minimum for the color map is at 0 CO_2 saturation indicated by blue, the maximum at 1 indicated by red

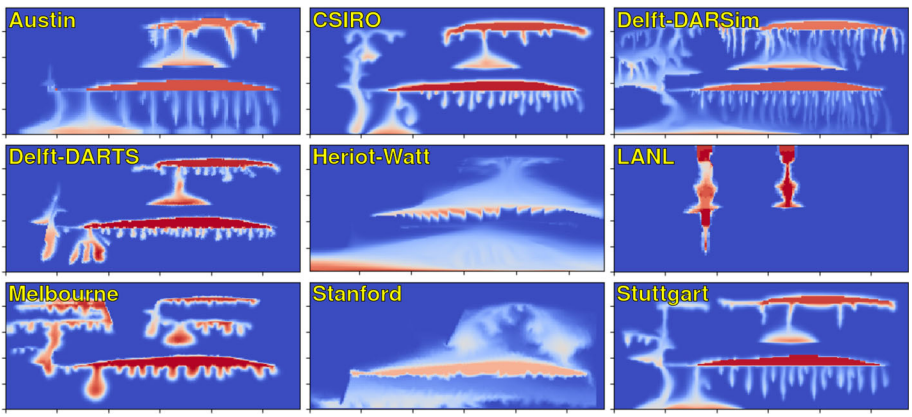

Fig. 7 Spatial distribution of CO_2 concentration in the liquid phase after 24 h. The minimum for the color map is at 0kgm^{-3} indicated by blue, the maximum at 1.8kgm^{-3} indicated by red

Stanford and Stuttgart. Starting with 72 h, Heriot-Watt did not report any spatial map data.

4.1.2 Concentration

Analogous to the saturation, Figs. 7, 8, 9, 10 and 11 visualize the reported concentration values for all participating groups at the selected daily time steps. While at first glance, the variation in the results appears to be larger than for the saturation, the reported qualitative behavior is similar for most groups.

The CO_2 dissolves into the liquid phase and, due to the density difference between pure and CO_2 -enriched water, the latter is moving downwards by developing fingers. This motion is impeded by fine-sand barriers or the bottom of the domain.

A clear outlier to this rather uniform qualitative behavior is given by LANL, whose simulations indicate that gaseous CO_2 has moved relatively straight upward without being hindered substantially by the fine-sand barriers and also not leaving any residual gas. A variety of

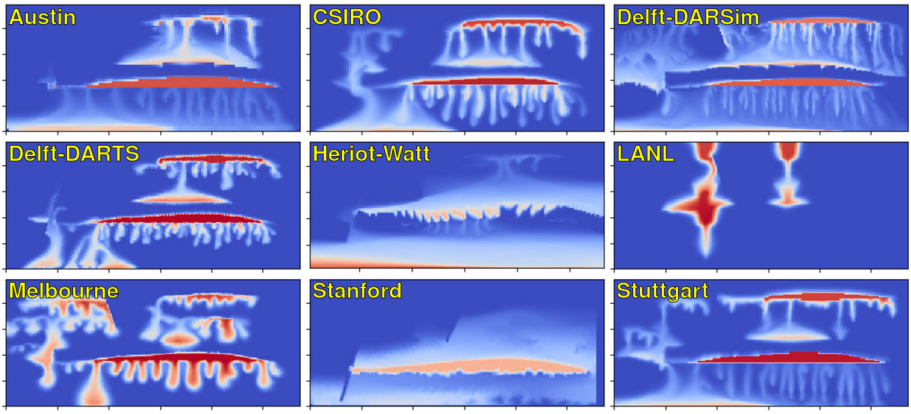

Fig. 8 Spatial distribution of CO_2 concentration in the liquid phase after 48 h. The minimum for the color map is at 0kgm^{-3} indicated by blue, the maximum at 1.8kgm^{-3} indicated by red

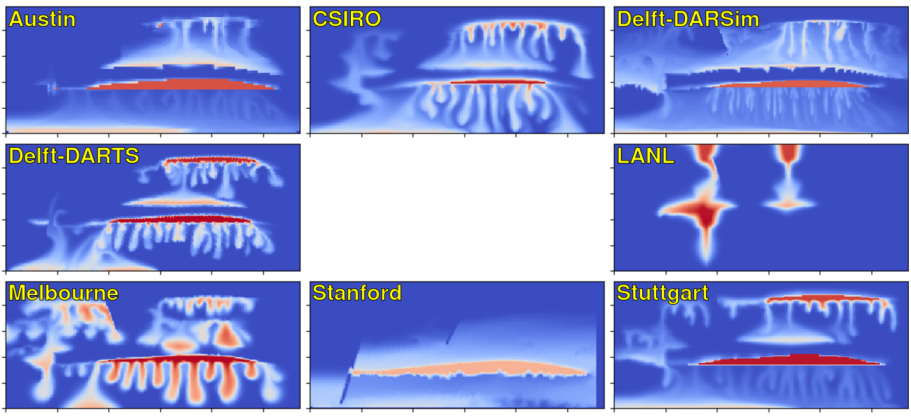

Fig. 9 Spatial distribution of CO_2 concentration in the liquid phase after 72 h. The minimum for the color map is at 0kgm^{-3} indicated by blue, the maximum at 1.8kgm^{-3} indicated by red

possible reasons exist, ranging from differently interpreted facies geometries and realized computational grids over too small variations in spatial parameters up to insufficient constitutive relationships. As running two codes with PFLOTRAN and FEHM yielded similar results, the exact reason could not be determined during the course of the study. Still, the descent over time of CO_2 dissolved in the aqueous phase is captured correctly.

The main quantitative differences which can be observed among the remaining groups arise due to the different speeds at which dissolution is taking place. In particular, dissolution for `Heriot-Watt` and `Stanford` appears to be much faster than for the other participating groups.

Moreover, quite some disagreement can be observed on how much CO_2 is reaching the upper left part of the domain, i.e., Box B, via the corresponding fault zone.

Another interesting measure is the amount and respective thickness in horizontal direction of the evolving fingers. Differences here can be largely attributed to different grid resolutions. For example, the participating groups `CSIRO`, `Delft-DARSim` and `Delft-DARTS` with relatively high resolution and correspondingly small cell diameters (cf. Table 2) show

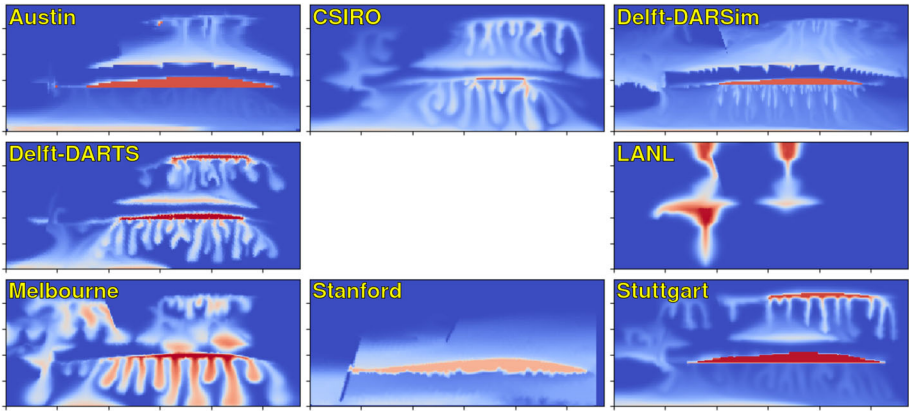

Fig. 10 Spatial distribution of CO_2 concentration in the liquid phase after 96 h. The minimum for the color map is at 0kgm^{-3} indicated by blue, the maximum at 1.8kgm^{-3} indicated by red

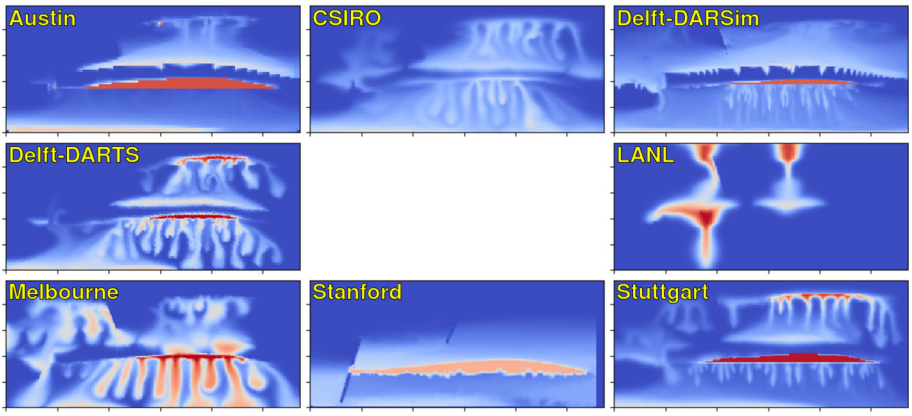

Fig. 11 Spatial distribution of CO_2 concentration in the liquid phase after 120 h. The minimum for the color map is at 0kgm^{-3} indicated by blue, the maximum at 1.8kgm^{-3} indicated by red

substantially more and thinner fingers than `Austin` and `Melbourne` with a relatively low resolution.

4.1.3 Quantitative Comparison

As a quantitative measure, we apply the Wasserstein metric to analyze the difference between two snapshots, combining a saturation and a concentration field to one mass field. The metric works on distributions of equal mass and measures "the minimal effort required to reconfigure the mass of one distribution in order to recover the other distribution" (Panaretos and Zemel 2019). In order to apply the Wasserstein metric to the reported results, which in general have a slightly different mass (see detailed discussion in Sect. 4.2.1), we first approximate roughly the CO_2 mass density in each cell by combining the reported concentration and saturation values via the formula

$$\widetilde{m} = \varrho_g s + c(1 - s).$$

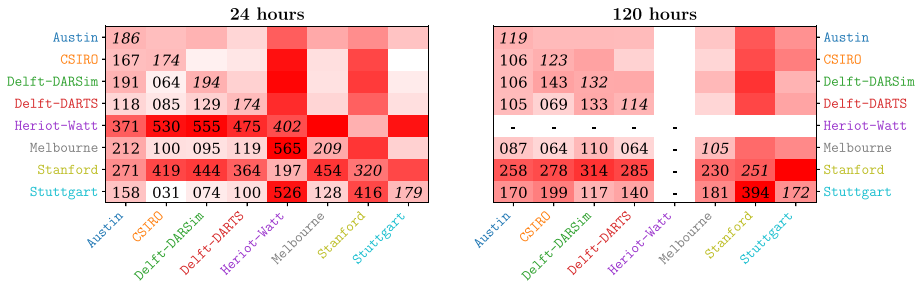

Fig. 12 Wasserstein distances in gram times centimeter for the first and last time step. Colors range from white for low values to red for high values. A value on a diagonal is the mean value of the respective row/column, where the calculation of the mean includes the zero self-distance. Values above the diagonal are not displayed as they are symmetric. As no spatial map has been reported by Heriot-Watt for 120 h, the corresponding fields are left empty

Above, s and c indicate the saturation and concentration value, while the density ϱ_g of gaseous CO_2 is set to 2kgm^{-3} to reflect the experimental conditions. The resulting values can be visualized by corresponding grayscale pictures which have been uploaded to the participants' data repositories. The final step to make these values comparable is their normalization such that they can be treated formally as two-dimensional probability distributions over the experimental domain. Given the normalized values, the Python library POT (Flamary et al. 2021) can be applied to calculate the Wasserstein distances. The values are listed in Appendix B for every requested individual timestep. The full data including distances between results from different timesteps is provided in the FluidFlower general GitHub repository. This approach provides a reasonable estimation for the groups with approximately equal mass in the reported results, however, it is not appropriate for the results from LANL, whose simulations indicate that a significant fraction of the injected mass leaves the domain. Therefore, the results from LANL are excluded from the Wasserstein distance calculations.

We show the calculated Wasserstein distances exemplarily for the first and last time step in Fig. 12.

The values have been dimensionalized by multiplying with the real mass of CO_2 in the system and are provided in units of gram times centimeter. Thus a value of 100 gr.cm corresponds to one gram of mass (e.g. about 20% of the CO_2 in the system) being shifted by one meter (e.g. about one third of the full simulation domain). Values on the order of 100 gr.cm or less thus correspond to what we consider relatively close results, while results in significant excess of 100 gr.cm indicate substantial discrepancies. Figure 12 thus quantifies the qualitative results discussed in the subsections above. In particular, the spatial maps from Heriot-Watt and Stanford show the largest distances to the other groups over all time steps. Their mean distances are between two and three times larger than the ones from the other groups, due to their different dissolution behavior. Overall, the mean distances are mostly decreasing from the first to the last time step, as CO_2 further dissolves in the water and its mass distributes more over the domain. We remark that the calculation of the mean values displayed on the diagonals in Fig. 12 includes the self-distance of zero. This is done for consistency with the calculation of distances between modeling and experimental results in Sect. 5.1.

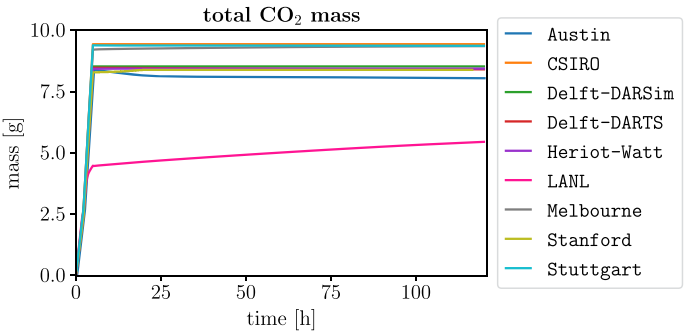

Fig. 13 Temporal evolution of the total CO_2 mass inside the computational domain

4.2 Dense Data Time Series

The participating groups were instructed to report several scalar SRQs in ten-minute intervals over a time span of five days: total mass of CO_2 inside the domain, pressure at two locations, phase composition in Boxes A and B, as well as convection in Box C.

4.2.1 Total Mass of CO_2

Figure 13 depicts the temporal evolution of the total mass of CO_2 inside the computational domain, as reported by the different participating groups.

The benchmark description prescribes the injection rates in terms of Standard Cubic Centimeters per Minute (SCCM) (Nordbotten et al. 2022). While the underlying standard conditions are not explicitly specified, the instrument employed by ExpUB uses the NIST definition of standard conditions, i.e. 293.15K and 1.013 bar. This would yield a final total mass of approximately 8.5g, assuming that no CO_2 leaves the domain. While the majority of the modeling groups employed the corresponding interpretation of standard conditions, three groups report a higher value of approximately 9.4g. The participant LANL reports considerable lower values which is due to the fact that CO_2 leaves the domain, as has been explained in more detail in Sect. 4.1. In most results, the total amount of CO_2 stays constant after injection stops, indicating that no mass leaves the system. Nevertheless, some groups report a further increase or also a further decrease, which can be explained by numerical effects in case of Melbourne (Youssef et al. 2023) or again the circumstance that gaseous CO_2 leaves the computational domain in case of Austin, respectively. The participant LANL reported the CO_2 mass in the box $(0, 0) \times (286\text{cm}, 123\text{cm})$ instead of the whole computational domain, see Fig. 1. When the injection stopped, the dissolved CO_2 in the volume between the top of the actual computational domain and the top of reported bounding box (that coincides with the top of Box B), moved back into the reported bounding box, leading to the increase in the mass in the reported domain with time.

4.2.2 Pressure

The next reported SRQ is the temporal evolution of the pressure, measured at two sensors in the domain. Figure 14 illustrates the reported results.

Most of the results show at most a minor influence of the CO_2 injection on the observed pressure values. The pressure at each sensor stays rather constant at the prescribed initial and

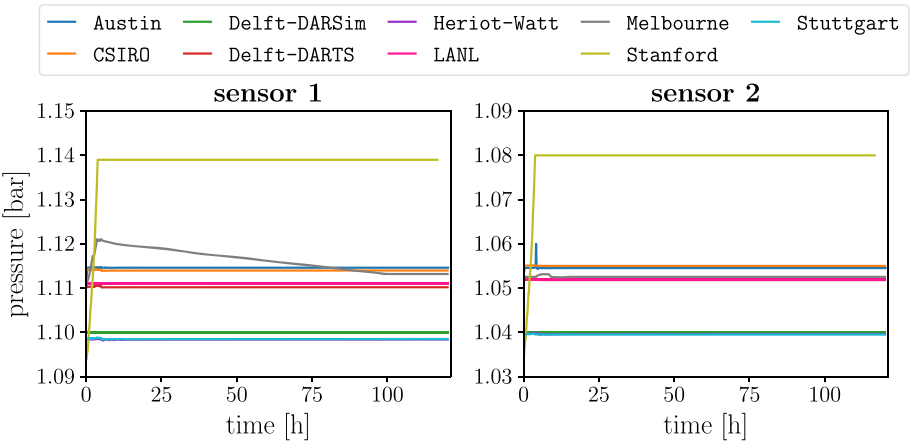

Fig. 14 Temporal evolution of the pressure at two locations inside the computational domain, Sensor 1 (left) and Sensor 2 (right)

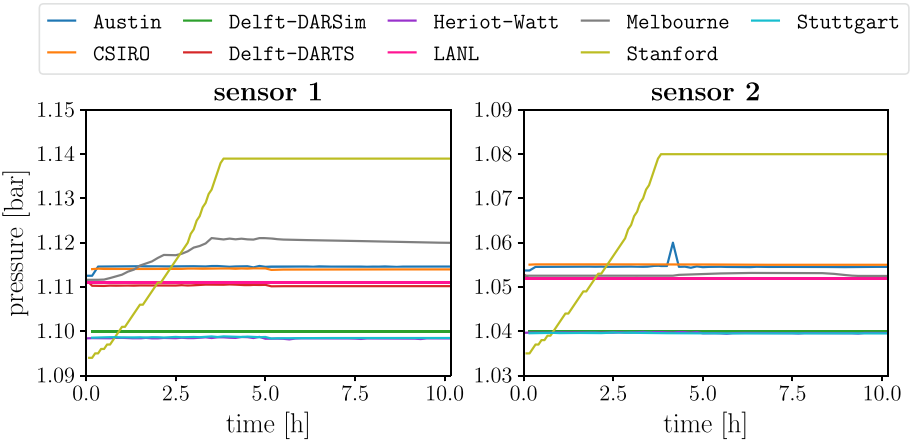

Fig. 15 Temporal evolution of the pressure at two locations inside the computational domain, Sensor 1 (left) and Sensor 2 (right). Zoom into the first ten hours

possibly boundary conditions which correspond to an assumed ambient atmospheric pressure plus the effect of the water table. Nevertheless, two groups, `Stanford` and `Melbourne`, report a considerable influence of the injection processes. In order to examine this in more detail, Fig. 15 depicts a zoom into the first ten hours of simulation.

The results from `Melbourne` show a considerable increase only for the first sensor which decays slowly to a constant level after the stop of injection. Here, the difference in the buildup between the two sensors can be explained by their respective proximity to the injection wells. In contrast to this, `Stanford` reports the same pressure buildup for both sensors. A possible explanation is that the fluids are assumed to be only slightly compressible and that the atmospheric boundary condition on top of the domain is realized by artificial large-volume cells. Moreover, the group detected an even higher buildup followed by a decrease to the officially reported values during the injection phase in the original simulation

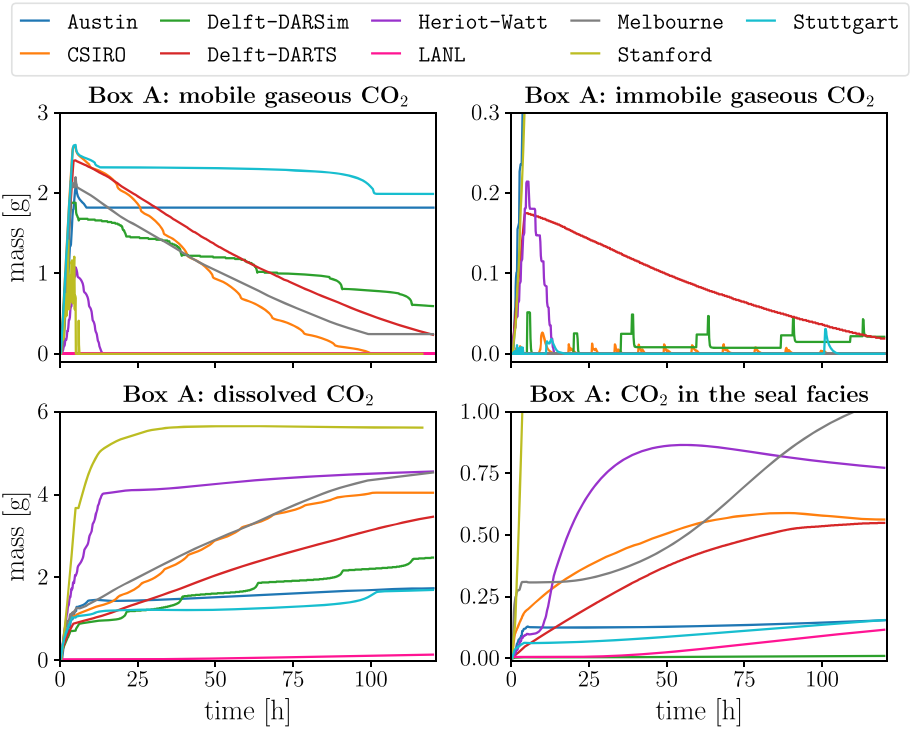

Fig. 16 Temporal evolution of the CO_2 phase distribution in Box A

data which was suppressed due to erroneous post-processing. Notably, both groups report a stop of the pressure buildup at around 3.5 h, before the stop of CO_2 injection at 5 h.

4.2.3 Phase Composition

In the following, we discuss the reported distribution of CO_2 over the two fluid phases in Boxes A and B. In particular, the participants reported the evolution of the amount of mobile and immobile gaseous CO_2, CO_2 dissolved in the liquid phase, as well as CO_2 contained in the seal facies. We first focus on Box A and the respective Fig. 16.

It can be seen immediately that the variation of the results across the participating groups is much larger than for the previous SRQs. All results have in common that mobile gaseous CO_2 reaches a peak value at approximately five hours (coinciding with the injection stop) and then dissolves at different rates. Eight results can be grouped into three clusters showing a similar rate. The largest cluster consists of the participants `CSIRO`, `Delft-DARSim`, `Delft-DARTS` and `Melbourne`. Here, the dissolution takes place over the whole simulation period at an intermediate rate compared to the other two clusters. The two participants `Austin` and `Stuttgart` both show after an initial decay a very slow dissolution behavior. In contrast to this, `Heriot-Watt` and `Stanford` predict the fastest dissolution with zero mobile gaseous CO_2 left after less than one day. Moreover, the fact that `Stanford` reports a very high amount of gaseous CO_2 becoming immobile can be attributed to their non-standard identification of immobile gas. Rather than evaluating the mobility, they declare gaseous CO_2 to be immobile if the change of saturation between two time steps doesn't exceed a certain

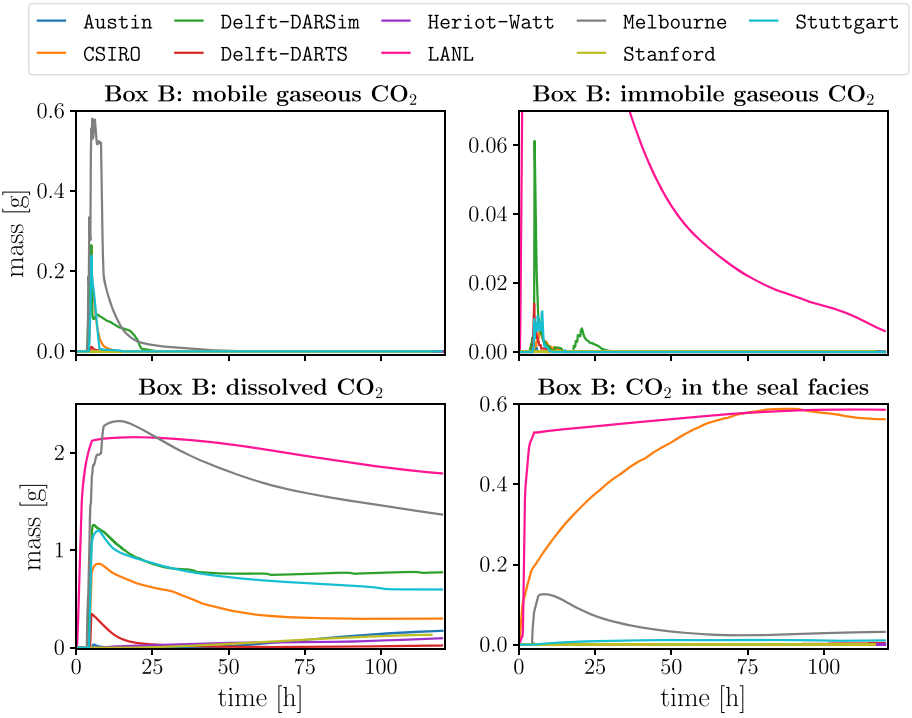

Fig. 17 Temporal evolution of the CO_2 phase distribution in Box B

threshold. An outlier with respect to all reported SRQs can be identified by LANL, where no CO_2 at all reaches Box A. All these observations are consistent with the results and discussion concerning the spatial maps in Sect. 4.1. In addition here, a remarkable characteristic is the step-like progression of several curves, as reported particularly by CSIRO, Delft-DARSim and Stuttgart. This numerical effect is due to grid-dependent bursts in dissolution when the water-gas contact coincides with cell faces. It has also been observed initially by Heriot-Watt, who decided to employ the capillary pressure–saturation relationship by van Genuchten for the coarser sands in order to prevent the effect, see also Table 2.

Turning to Box B and Fig. 17, the results exhibit even more variation. This can be attributed to the location of the box with the challenge of quantifying how much CO_2 reaches the fault zone in the lower left and subsequently the upper left region of the domain.

While all participants predict the disappearance of mobile gaseous CO_2 after at most two days, the peak amount varies strongly between zero and 0.6g. These different peak amounts together with different dissolution rates explain the high variation in dissolved CO_2 as seen in Fig. 17. Nevertheless, almost all models predicting a substantial amount of CO_2 in Box B report very similar times of appearance.

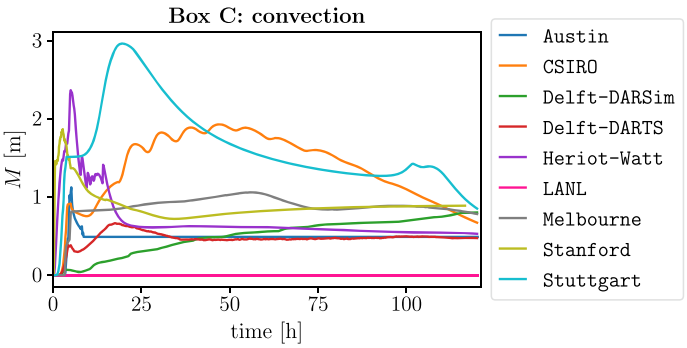

Fig. 18 Temporal evolution of $M(t)$ as a measure for convection in Box C

4.2.4 Convection

As a measure for convection, the participants were asked to report the total variation of concentration within Box C over time, see the definition of $M(t)$ in Nordbotten et al. (2022, Section 2.8.3). The results are depicted in Fig. 18.

A relatively large spread with peak values ranging from 0 to 3 m can be observed. Also the dynamic behavior is very different, ranging from a monotone increase to rather strong oscillations. Nevertheless, most participants report a stabilization over time to a stationary value between 0.5 and 1 m.

4.3 Sparse Data

In this section, we describe the reported so-called sparse data. Each of the sparse data items had to be reported as six numbers, representing the prediction of the mean quantity as obtained by the experiments (stated in terms of P10, P50 and P90 values), as well as the prediction in the standard deviation of the quantity over the ensemble of experiments (again stated as P10, P50, and P90 values). Since most groups did not report any P10 and P90 values for the expected standard deviations, we only consider the P50 values for the following comparisons. As basis for generating the predictions and uncertainties, any preferred methodology could be chosen, ranging from ensemble runs and formal methods of uncertainty quantification to human intuition from experience. We start with the maximum pressure at the two sensors, then focus on the times of maximum mobile gaseous CO_2 in Box A and onset of convective mixing in Box C, before we investigate the predicted phase distributions after three days in Boxes A and B. The numerical values are also recorded in Appendix C.

4.3.1 Maximum Pressure at the Two Sensors

The participants were asked for the expected maximum pressure at Sensors 1 and 2 as a proxy for assessing the risk of mechanical disturbance of the overburden. The reported values are depicted in Fig. 19.

As can be seen from the scaling of the vertical axis, all participating groups report very similar pressure values. Most groups also report P10, P50 and P90 values for the expected mean which are very close to each other, with the largest difference for one group being around 10 mbar. With `Austin` and `Melbourne`, only two groups expect any substantial

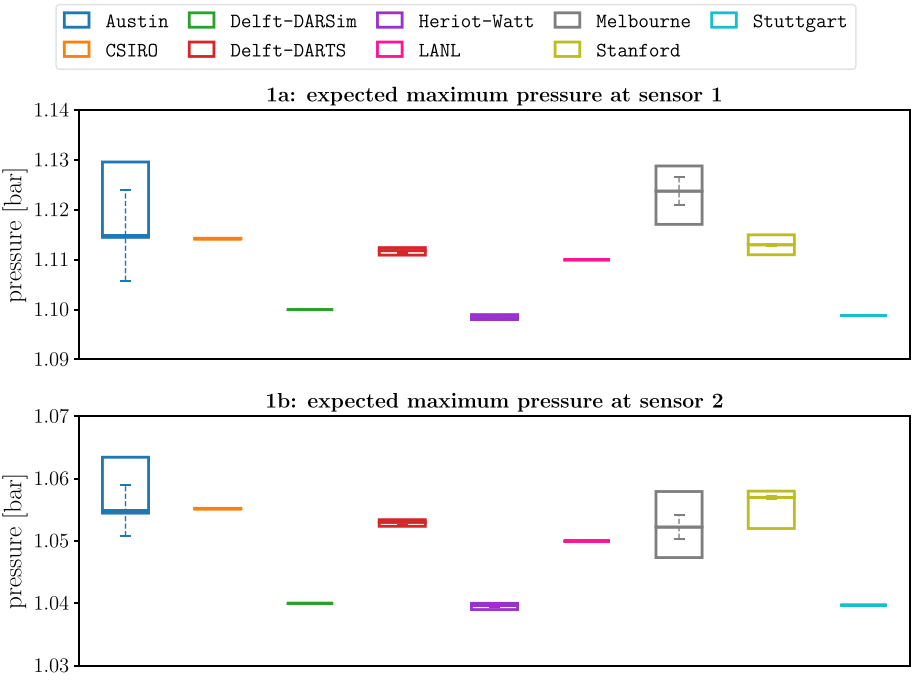

Fig. 19 Reported sparse data for the maximum pressure at sensors 1 and 2. Bottom, middle and top horizontal lines of the boxes indicate the reported P10, P50 and P90 values for the expected mean value, respectively. Dashed vertical lines extend from the mean values by ± the reported P50 of the expected standard deviations

standard deviation over the ensemble of experiments. The difference over all groups between the minimum P10 and maximum P90 reported pressure value is less than 40 mbar for each of the two sensors. This indicates that the typical variation in atmospheric pressure at the location of the experimental rig was not taken into account, exceeding 50 mbar over the winter months. Although the exact days of the experimental runs have not been provided explicitly to the participants, the information on the usual pressure variation is publicly available. [4]

4.3.2 Times of Maximum Mobile Gaseous CO_2 in Box A and Onset of Convective Mixing in Box C

We now focus on the time of maximum mobile gaseous CO_2 in Box A as a proxy for when leakage risk starts declining. The corresponding reported values are visualized in the upper picture of Fig. 20.

The majority of the participating groups now report substantial differences between the P10 and P90 values of both the expected mean and standard deviation. Nevertheless, several groups are very certain on the expected mean value and report narrow ranges. The variation between the groups is considerably larger than for the pressure discussed above. This can be explained by the larger variation in the modeling results as discussed in Sects. 4.2.2 and 4.2.3.

[4] https://weatherspark.com/h/s/148035/2021/3/Historical-Weather-Winter-2021-at-Bergen-Flesland-Norway#Figures-Pressure.

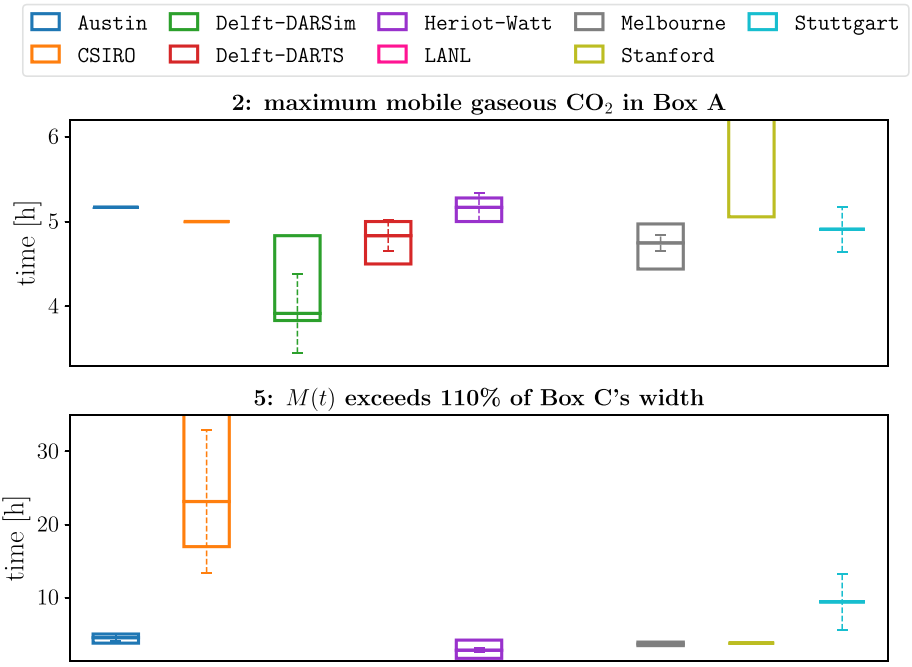

Fig. 20 Reported sparse data for the times of maximum mobile gaseous CO_2 in Box A (top) and for which the integral $M(t)$ first exceeds 110% of the width of Box C (bottom). Bottom, middle and top horizontal lines of the boxes indicate the reported P10, P50 and P90 values for the expected mean value, respectively. Dashed vertical lines extend from the mean values by $\pm$ the reported P50 of the expected standard deviations

As a proxy for the ability to capture the onset of convective mixing, we focus on the time for which the quantity $M(t)$ defined in Nordbotten et al. (2022, Section 2.8.3) first exceeds 110% of the width of Box C, as depicted in the lower picture of Fig. 20. We first note that three groups do not report any value at all. Out of the remaining six, four report very similar values around 4 h and narrow ranges between P10 and P90. With CSIRO, one group reports much larger expected values and also variations between P10 and P90. In order to examine this in more detail, we perform a comparison with the corresponding temporal evolution of $M(t)$ as depicted in Fig. 18. With 110% of the width of Box C being equal to 1.65 m, we can observe that several results do not reach this value at all over the whole simulation period. In turn, this explains that three groups did not report any value for the sparse data. Zooming closer into the first ten hours of simulated time as done in Fig. 21 allows to put the reported time series values in explicit relation to the sparse data.

As can be identified from the vertical lines representing the reported expected mean values, the measured value for $M(t)$ is usually well below the 110%. Therefore, it becomes obvious that several participating groups did not rely only on the reported simulation results for the SRQ considered here.

4.3.3 Phase Distributions After 3 Days in Boxes A and B

We now turn to the reported sparse data for the phase distribution in Box A at 72 h after injection starts as a proxy for the ability to accurately predict near well phase partitioning.

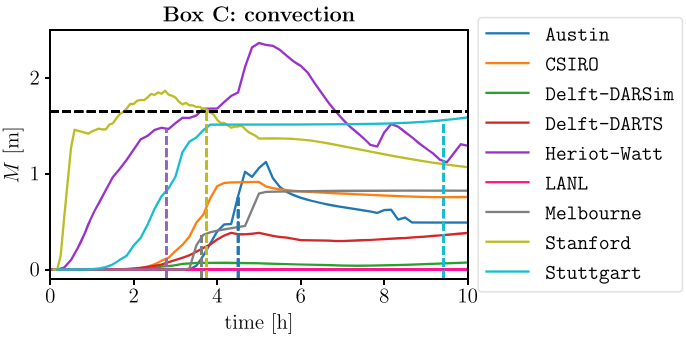

Fig. 21 Zoom into the first ten hours of the temporal evolution of $M(t)$. The black horizontal dashed line depicts 110% of Box C, dashed vertical lines correspond to the reported expected mean values

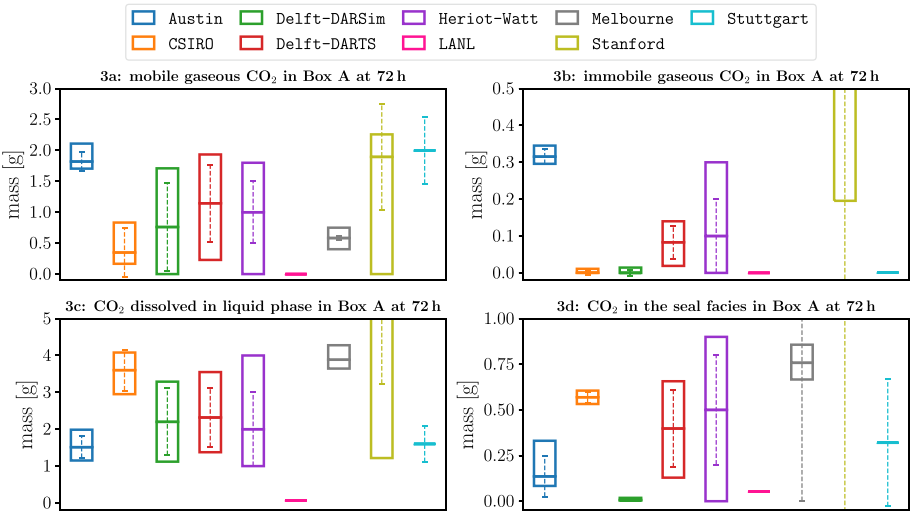

Fig. 22 Reported sparse data for the phase distribution in Box A at 72 h after injection starts. Bottom, middle and top horizontal lines of the boxes indicate the reported P10, P50 and P90 values for the expected mean value, respectively. Dashed vertical lines extend from the mean values by ± the reported P50 of the expected standard deviations

From the corresponding Fig. 22, it can be seen immediately that the reported ranges between the P10 and P90 values of the expected mean values are substantially larger than for the preceding measures, going along with increased expected standard deviations.

Concerning the amount of mobile gaseous CO_2, the expected P50 of the mean value ranges between 0.5 and 2g, while for the amount of dissolved CO_2, values range mostly between 1 and 4g.

The expected phase distribution in Box B at 72 h after injection starts is depicted in Fig. 23, interpretable as a proxy for the ability to handle uncertain geological features.

It can be observed that mostly no mobile gaseous CO_2 is expected, while the associated uncertainty is considered to be quite high. In case of `Stanford`, the large variation comes from the fact that a simulation with immiscible fluid phases was included in the underlying uncertainty quantification as a limit case. Turning to the lower left picture, the amounts of predicted dissolved CO_2 show a strong variation over the participating groups.

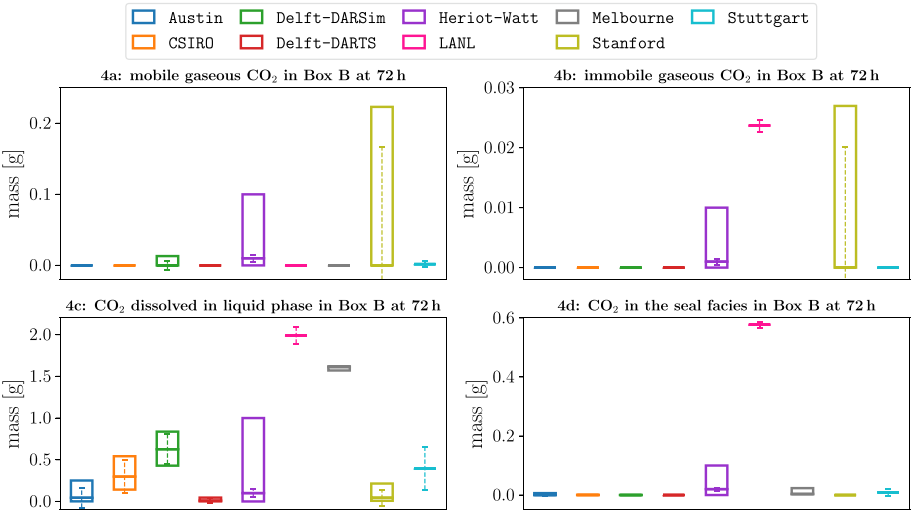

Fig. 23 Reported sparse data for the phase distribution in Box B at 72 h after injection starts. Bottom, middle and top horizontal lines of the boxes indicate the reported P10, P50 and P90 values for the expected mean value, respectively. Dashed vertical lines extend from the mean values by ± the reported P50 of the expected standard deviations

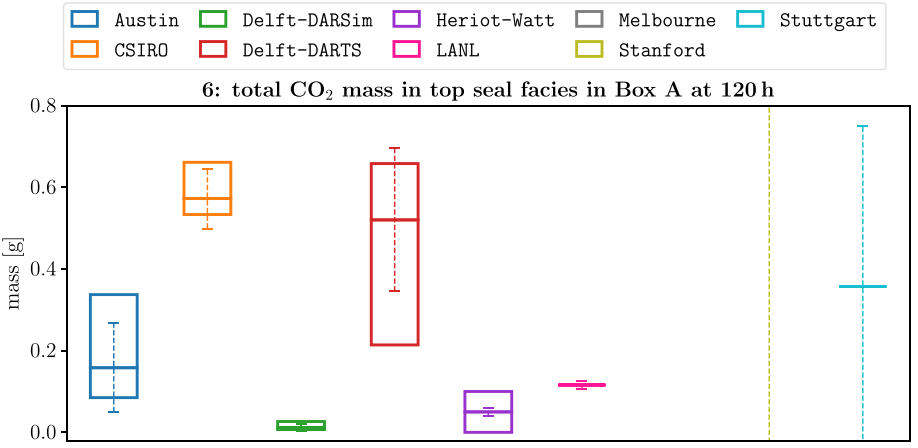

Fig. 24 Reported sparse data for the total mass of CO_2 in the top seal facies at final time within Box A. Bottom, middle and top horizontal lines of the boxes indicate the reported P10, P50 and P90 values for the expected mean value, respectively. Dashed vertical lines extend from the mean values by ± the reported P50 of the expected standard deviations

4.3.4 Total CO_2 Mass in Top Seal Facies Within Box A

As the last SRQ, we examine the expected total mass of CO_2 in the top seal facies at final time within Box A for evaluating the ability to capture migration into low-permeable seals. Figure 24 depicts the corresponding reported results.

Also here, large variations can be observed, not only in the expected mean values, but also in the expected standard deviations.

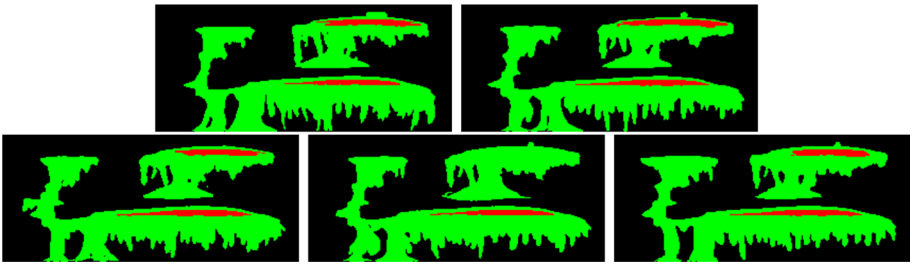

Fig. 25 Segmentation data after 24 h for five experimental runs. Black, green and red indicate pure water, water with dissolved CO_2 and gas, respectively

5 Comparison to Experimental Data

In the following, we will compare the modeling results described in the previous section with the actually observed experimental data. The underlying experimental methodology and original dataset is presented in Fernø et al. (2023), while the image analysis approach is discussed in Nordbotten et al. (2023a). We focus first on the dense data spatial maps and time series and investigate afterwards the sparse data SRQs.

5.1 Dense Data Spatial Maps

We will first perform a visual comparison of segmentation maps and subsequently perform a quantitative comparison by means of the Wasserstein distance.

5.1.1 Segmentation Maps

In the following, we compare daily spatial maps given in form of segmentation data. For the experiments, this data has been generated by analyzing corresponding images using the newly developed toolbox DarSIA (Nordbotten et al. 2023a). In Fig. 25, the snapshots at 24 h are shown for five experimental runs.

Visually, there is a very good agreement over all five runs and differences can only be detected in the details. One slight exception is given by the fourth run, where no gas appears to be present in the upper right part of the domain. However, this is attributable to numerical effects in the image analysis procedure, rather than a different physical truth. We will perform a quantitative analysis further below.

Before that, a visual comparison with the modeling results is carried out. For this, the concentration and saturation maps at 24 h provided by the participants are converted into segmentation data. Thresholds of $1e{-}2$ for saturation and $1e{-}1 kgm^{-3}$ for concentration are used above which a cell is declared to contain gaseous CO_2 and CO_2-rich water, respectively. To allow for a more direct comparison, the modeling results are overlaid by the contour lines corresponding to the experimental data. The result is shown in Fig. 26.

It can be seen that the locations of the two gas plumes are reasonably well captured by several participants, namely, `Austin`, `CSIRO`, `Delft-DARSim`, `Delft-DARTS`, `Melbourne` and `Stuttgart`, while their sizes are overestimated in general. As already suggested by the strong variability of the concentration distributions discussed in Sect. 4.1, considerably less agreement can be observed concerning the region covered by water with dissolved CO_2. This becomes particularly apparent for Box B in the upper left part of the

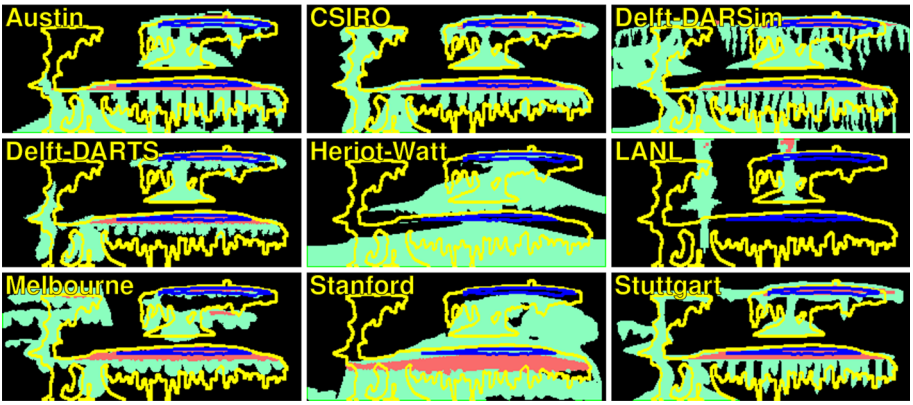

Fig. 26 Comparison of segmentation data after 24 h. Each modeling result is overlaid by the contour lines of experimental run 2. The forecasts are colored by black, pale green and pale red, indicating pure water, water with dissolved CO_2 and gas, respectively. Concerning the experimental data, yellow contour lines indicate the region of water with dissolved CO_2, while blue lines illustrate the gas plume

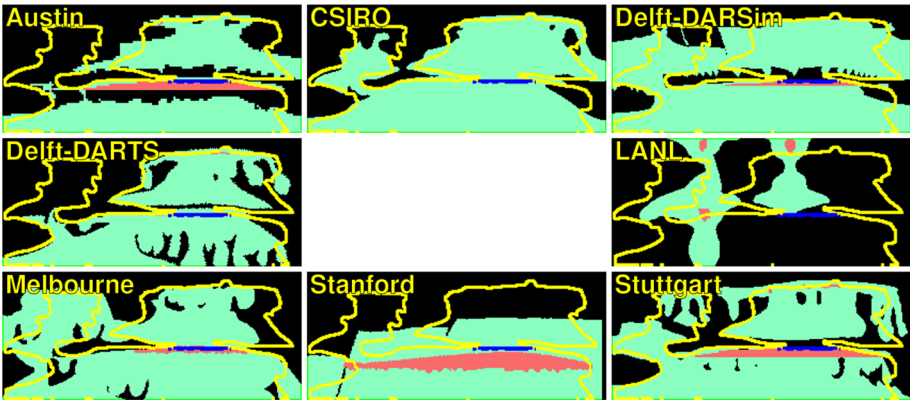

Fig. 27 Comparison of segmentation data after 120 h. See Fig. 26 for more details on the color coding

domain, where only the CSIRO modeling result matches the basic shape and extension in a visually satisfactory way.

In Fig. 27, the same comparison is made at 120 h.

CO_2-rich water has spread throughout large parts of the domain in both the experimental data and most of the modeling results. The correspondingly covered regions coincide reasonably well below the original gas plumes. Like at 24 h, the biggest differences can again be observed in the upper left part of the domain. There, the results from CSIRO and also from Stuttgart provide a decent match. Almost all models predict correctly that no gaseous CO_2 is present anymore in the upper part of the domain. Regarding the lower part, some models overestimate and some others underestimate the amount of gaseous CO_2, while Delft-DARSim and Melbourne appear to be closest to the experimental data.

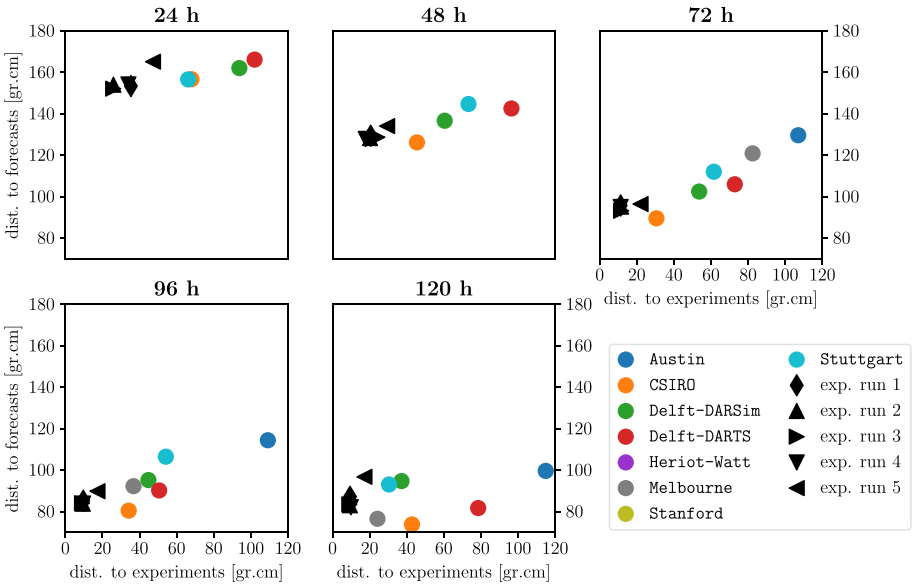

Fig. 28 Wasserstein distances of the segmentation maps to experiments and forecasts. Zoom into the ranges from 0 to 120 gr.cm for the mean distance to the experimental results and from 70 to 180 gr.cm for the mean distance to the modeling forecasts. Some groups with outlying results are therefore not visible in all plots, while `Heriot-Watt` and `Stanford` are consistently outside the range of the plots (confer distances in Fig. 12)

5.1.2 Quantitative Comparison

To develop a more quantitative understanding, a similar analysis as in Sect. 4.1.3 can be performed in terms of the Wasserstein metric. This involves calculating distances for all pairs consisting of two participating groups, two experimental runs, or one participant and one run. For the application of the Wasserstein metric, the segmentation maps discussed above are converted to mass distributions, assigning zero/half/full weights to cells with pure water/CO_2-rich water/gaseous CO_2. Like in Sect. 4.1.3, the calculated distances are multiplied with the total mass of CO_2. Proceeding like this, the mean distances to the other modeling results and now also to the experimental data can be calculated, yielding two values for each segmentation map. Figure 28 plots these values for all segmentation maps at the selected time steps.

We can observe that the experimental data sets are within 50 gr.cm of each other, confirming that the experimental repeatability is strong, and that there is only minor impact of the different experimental conditions (primarily attributed to atmospheric pressure, some chemical alterations within the experimental rig, and very minor amounts of settling sand throughout the experimental period). About half of the modeling results are within about 100 gr.cm of the experimental data for all reporting times, which we consider a relatively good match. At the final time, the closest simulation results are as little as 50 gr.cm away from the experimental mean, which is within twice the experimental variability at that time. This also aligns with the visual impressions for the segmented images shared above. With increasing time, the distances to both the experiments and the forecasts are decreasing for most modeling results; the same holds for the distances of the experimental data sets to the forecasts. This can be explained by the increasing spread of CO_2-rich water over the domain and a corresponding equilibration of CO_2 mass.

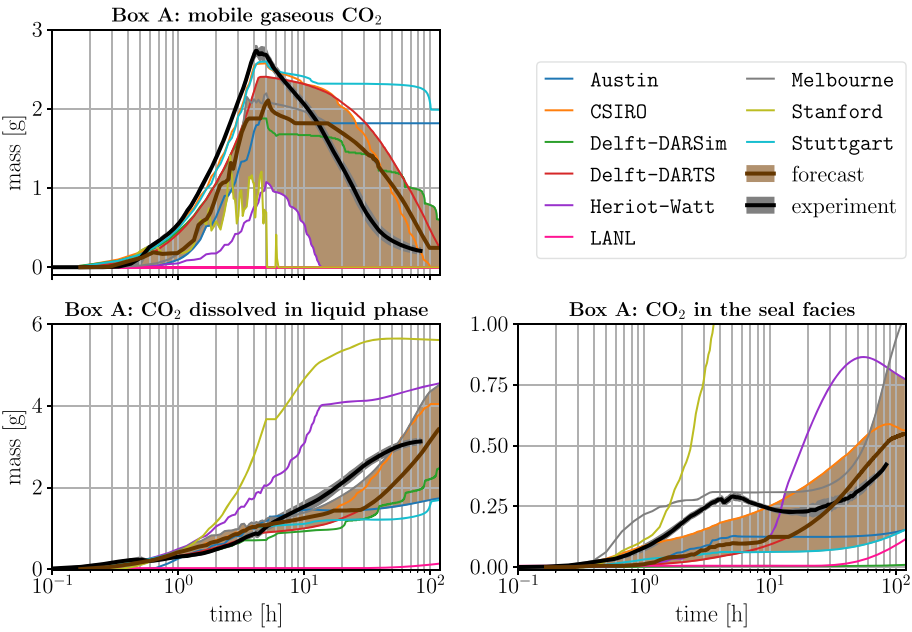

Fig. 29 Comparison between modeling forecasts and experimental observations for the temporal evolution of the CO_2 phase distribution in Box A. A brown line depicts the median of the reported modeling results, while the associated pale brown region illustrates the area between the corresponding first and the third quartile. A black line shows the mean of the experimental data, while the associated grey region depicts the corresponding variation by means of the standard deviation

5.2 Dense Data Time Series

In the following, we compare selected dense data time series as reported by the participating groups with corresponding experimental data. As described in Fernø et al. (2023), Nordbotten et al. (2023a), the derivation of saturation and concentration values from the experimental photographs is a very challenging endeavor based on several assumptions. The correspondingly calculated mass values are subject to significant uncertainties. Therefore, the degree of physical truth behind the comparisons has to be taken with great care.

Figure 29 shows the comparison for the temporal evolution of the phase distribution in Box A.

For being able to observe more details in the beginning of the investigated time frame, the x-axes in the pictures use a logarithmic scaling. Concerning mobile gaseous CO_2, the basic shape of the experimental mean is quite similar to the median of the modeling results. Nevertheless, the peak value for the forecast is considerably lower than the experimental one. The spread of the modeling results during the advection-driven stage of increasing values is substantially less than during the dissolution-driven stage of decreasing values afterwards. This results in a much longer period where the value stays rather constant. While in general the stages of increasing and decreasing values are lagging behind the experimental results, the results from `CSIRO` and `Stuttgart` match the first stage very well. All plots in Fig. 29 and also Fig. 30 reveal that the variation in the experimental data is rather small, illustrating a good repeatability of the experiments.

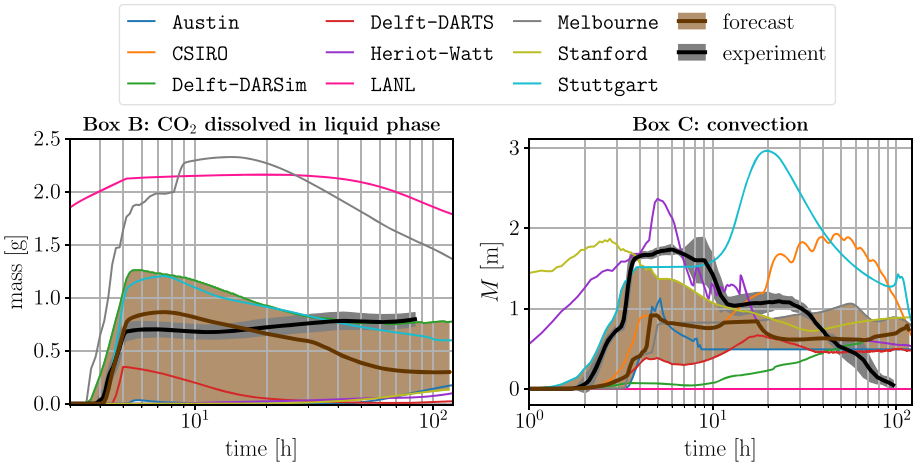

Fig. 30 Comparison between modeling forecasts and experimental observations for the temporal evolution of the dissolved CO_2 mass in Box B (left) and the integral quantity $M(t)$ (right). A brown line depicts the median of the reported modeling results (which coincides with the result reported by CSIRO on the left). The associated pale brown region illustrates the area between the corresponding first and the third quartile. A black line shows the mean of the experimental data, while the associated grey region depicts the corresponding variation by means of the standard deviation

Focusing on the temporal behavior of the dissolved CO_2 mass, it can be seen that most of the modeling results agree well with the experimental data in the beginning. The spread in the forecasts starts to increase after the injection stops and the very different dissolution behaviors discussed earlier become dominant. While most modeling results underestimate the amount of dissolved CO_2 during the majority of the simulated time, the values tend to increase longer than the corresponding experimental data which saturates earlier. Investigating the third picture, the evolution of the CO_2 mass in the seal varies strongly over the participating groups and differs substantially from the experimental data. A reason for the non-monotonic behavior of the experimental mean is discussed in Fernø et al. (2023).

Experimental data has been provided for two other time series and the corresponding comparisons are illustrated in Fig. 30.

Turning first to the amount of dissolved CO_2 in Box B, the large variations in the modeling results are also apparent by the depicted large spread. Like for Box A, the advection-driven increase in the beginning is captured well by two participating groups. Also here, the differences become more pronounced after injection stops. The amount of CO_2 increases further in the experimental data over time due to CO_2-rich water entering Box B from the right. This effect is not captured by most of the models.

We investigate finally the temporal evolution of the convection measure $M(t)$ in the right picture of Fig. 30. However, the differences of the modeling results to the experimental data are too strong to draw any meaningful conclusion here. It is likely that this has to do with the fact that the numerical evaluation of the integral value is not straightforward, strongly discretization-dependent and has been left entirely to the participants.

5.3 Sparse Data

The collection of the sparse data results has been accompanied by questionnaires for monitoring the confidence of each participant in their own prediction as well as in the ones of the

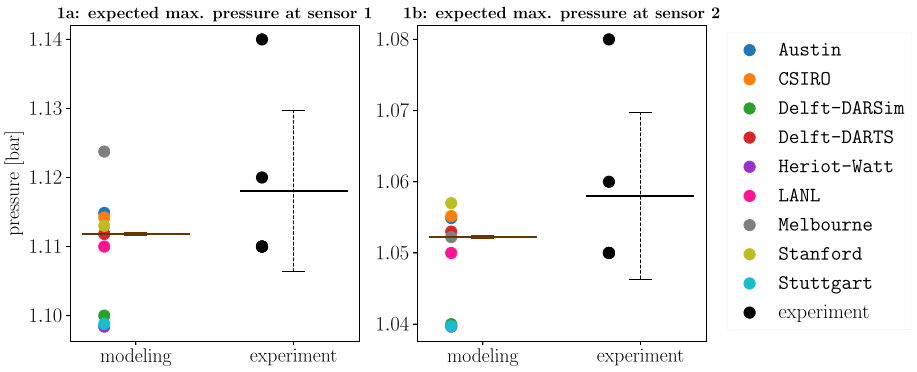

Fig. 31 Comparison of the sparse data reported by the participating groups with the experimental data for SRQ 1. Concerning the modeling results, colored circles correspond to the individual expected means, while the horizontal brown line depicts their median. A dashed vertical brown line extends from this value by ± the median of all reported P50 values for the standard deviation. Regarding the experimental data, black circles depict the results of the individual runs, while the horizontal black line indicates their mean. A dashed vertical black line extends from the mean by ± the standard deviation

respective other working groups. Since the description and analysis of this process and its results would be beyond the scope of this work, a separate paper is devoted to this (Nordbotten 2023b). In the following comparison with the experimental data, we therefore limit ourselves to a rather brief presentation of a few agglomerated measures.

In order to condense the responses by the individual participating groups presented in Sect. 4.3, we only consider the reported P50 values for the expected means and standard deviations. The means will be plotted as individual data points, together with their median and the median of the expected standard deviations. Concerning the experimental data, the results from the individual runs are plotted, together with their mean and standard deviation.

In Fig. 31, we consider first with SRQ 1 the expected and observed maximum pressures in the two sensors.

Like predicted by most of the participating groups, the injection of CO_2 had almost no impact on the pressure observed in the two sensors. The reported measured experimental values correspond to the maximum atmospheric pressure during a respective experimental run plus the hydrostatic contribution by the corresponding overlying water column. The individually reported expected means are within 10 mbar of the experimental mean and the median of the expected means shows a good agreement with the experimental mean. Nevertheless, as already noticed in Sect. 4.3.1, the participants expected almost no variation in the experimental results. Due to the natural fluctuations in atmospheric pressure, the observed variations turn out to be significantly larger than the expected ones.

Figure 32 illustrates the comparison for the SRQs 2 and 5, namely, the time of maximum mobile gas phase in Box A and the time when $M(t)$ exceeds 110% of Box C's width, respectively.

Concerning the former, it can be observed that the experimental mean is overestimated by most participating groups and that the reported and observed ranges are rather disjoint. For the latter, the situation is different as two sets of experimental data are provided which differ in the underlying image analysis parameters and constitute upper and lower bounds for the target quantity. Here, the median of the expected means lies close to the corresponding upper experimental mean.

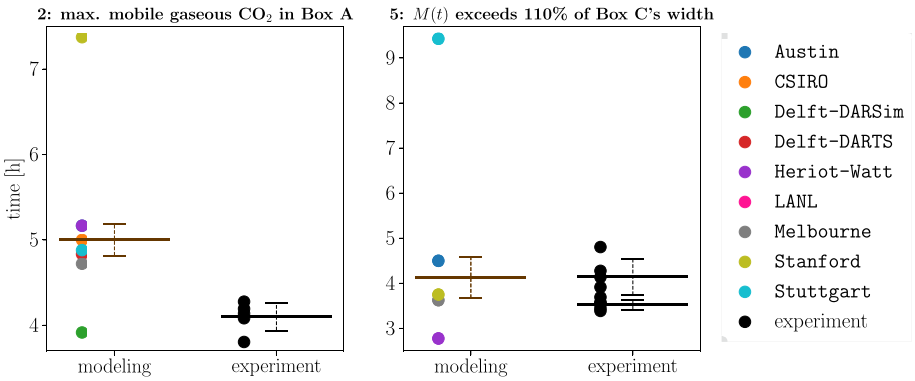

Fig. 32 Comparison of the sparse data reported by the participating groups with the experimental data for SRQs 2 (left) and 5 (right). See Fig. 31 for more details on the plotted quantities. For illustration purposes, the value reported by LANL (2.8e5 h) is not visualized on the left. On the right, this holds for the values from CSIRO (2.3e1 h) and LANL (2.4e3 h)

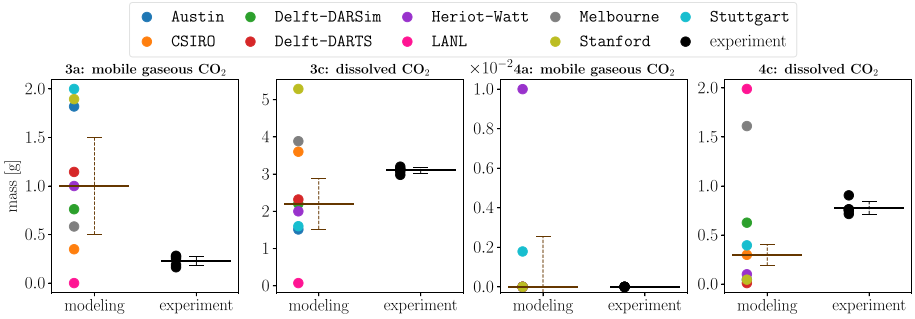

Fig. 33 Comparison of the sparse data reported by the participating groups with the experimental data for SRQs 3a, 3c, 4a and 4c (left to right). See Fig. 31 for more details on the plotted quantities

Next, we perform a comparison for the sparse data SRQs 3a, 3c, 4a and 4c, regarding the phase distribution of CO_2 after 72 h in Box A and B, respectively. Figure 33 depicts the corresponding quantities in terms of CO_2 mass in either gaseous or liquid phase.

Starting with 3a, it can be observed that the mean value of mobile gaseous CO_2 in Box A is overestimated by most participating groups and only some groups report values within the observed experimental range. This is consistent with the visual impressions discussed in Sect. 5.1. Regarding 3c, the mean value of CO_2 dissolved in water in Box A is rather underestimated by the modelers. Moving to Box B, all experimental runs suggest that no gaseous CO_2 is left after 72 h. This has also been expected by most participants, while they nevertheless presumed a slight standard deviation on average. While the reported numbers for the expected mean of dissolved CO_2 are rather widespread, the median value is remarkably close to the observed experimental mean.

With the final SRQ 6, we examine the total CO_2 mass in top seal facies within Box A at final simulation time, as illustrated in Fig. 34.

The median of the expected means is at around 50% of the observed experimental mean. Correspondingly, most participating groups underestimate the amount of CO_2 in the top seal facies. Nevertheless, two groups are very close to the experimental results.

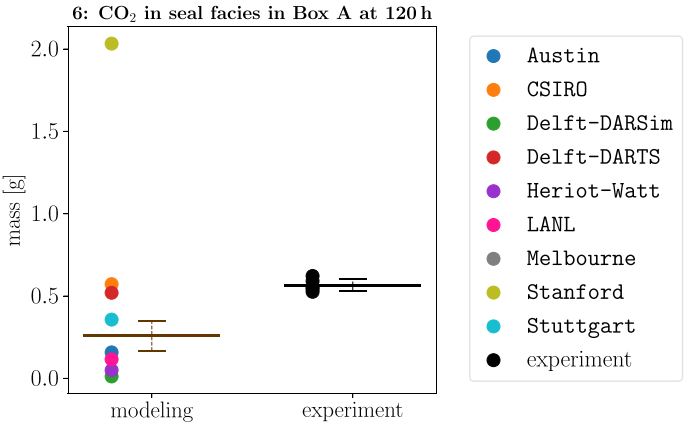

Fig. 34 Comparison of the sparse data reported by the participating groups with the experimental data for SRQ 6. See Fig. 31 for more details on the plotted quantities

6 Concluding Discussion and Outlook

In the following, we will draw several conclusions from this validation benchmark study and present challenges and opportunities for further work.

First, we can state with strong confidence that Darcy-scale balance equations together with standard constitutive relationships for the capillary pressure and relative permeability describe adequately the relevant observed physical processes on the considered spatial and temporal scale. This is revealed clearly from the comparison of the modeled saturation and concentration distributions with the corresponding experimental segmentation maps. In particular, stratigraphic and residual trapping mechanisms are captured well by most participating groups. Moreover, the process of convective mixing due to density differences is considered adequately in a qualitative manner.

Quantitatively, large variations in the modeling results can be observed particularly for the dissolution behavior and the resulting fingering. This can be attributed to different modeling choices for the solubility limit of CO_2 in water as well as for constitutive relations such as capillary pressure - saturation relationships, equations of state for determining phase compositions or phase density calculations. It can also be observed that differences in grid resolution clearly influence the convective mixing behavior. Nevertheless, several participating groups are in close proximity to the experimental results, as quantified by the Wasserstein metric. The corresponding distances decrease with increasing time as more CO_2 is dissolved and its mass equilibrates over the domain.

The study included reporting of pre-defined "sparse data", which were quantities that we can consider as proxies for various aspects of storage capacity and storage security. These quantities were reported with both a most likely exceedance value (P50), as well as P10-P90 intervals. While the P50 values mostly reproduce the reported dense data, the P10-P90 values add an additional dimension to the results. Notably, for the majority of requested quantities, the reported P10-P90 quantities do not overlap between the groups. Logically speaking, if two P10-P90 intervals do not overlap, then one group believes that there is at most a 10% chance that the other group will find the experimental results to be within their reported interval (and conversely). This implies that despite the significant group interaction through the study, the groups did not take the quantitative response of other groups into serious consideration, and

placed high or full confidence in their own results. This observation is complemented by the fact that the interaction helped almost all groups to establish a common understanding regarding the expected qualitative behavior such as the effect of capillary barriers.

A particular critical physical process that is evidenced in this study (both in sparse and dense data) is the role of convective mixing in accelerating dissolution of gaseous CO_2. This is quantified both through the actual phase compositions in Box A and B, as well as in the metric M, which is a proxy for the time of fully developed fingers (for a detailed discussion of various onset times in numerical simulation of density driven fingers, see Elenius and Johannsen 2012). The onset and evolution of convective fingers is particularly challenging for this system, since the low-order numerical methods used in this study (suitable to capture heterogeneity and stable discretization of multi-phase flow) tend to be too diffusive in their representation of the gas-water interface. The result is significantly over-estimating mass transfer from the gas to the water phase, necessitating a fine grid in the vertical direction. Moreover, the characteristic wave-length of density driven fingers for this system is on the order of 5 cm (as seen experimentally), further necessitating a sub-centimeter grid resolution horizontally. Seen together, this may be the cause for large variability in the reported structure and importance of density-driven fingering among the participants, and motivates further study on how to reliably and accurately capture this process within reservoir simulation tools.

While this study is at the laboratory scale, the fundamental physical processes of multi-phase, multi-component flows in heterogeneous porous media are the same as at reservoir conditions. As such, we argue that the findings and observations in this study are indicative of field-scale simulation (for a detailed scaling analysis, see Kovscek et al. 2023). That said, actual field-scale simulation will deviate from this study in several important aspects, of which we highlight:

- Heterogeneity. This study was conducted with homogeneous facies (to the extent possible in laboratory conditions), emphasizing larger-scale structural heterogeneities. On the field scale, it is expected that there will be significant subscale heterogeneity also within each geological structure.
- Quality of geological characterization. This study was conducted in a quasi-2D geometry, which was fairly well characterized (high-resolution photography as well as thickness measurements at the beginning of the experiment). At the field scale, the geological characterization is based on seismic surveys, which are not able to provide the same level of accuracy.
- Dimensionality. Reality is 3D, which will impact simulation time, and thus indirectly the level of grid refinement that can be sought.
- Convective mixing. In field-scale simulations, the spatial and temporal resolutions required for capturing correctly convective mixing are not practically feasible.
- Pressure and temperature conditions. At laboratory conditions CO_2 exists in a gas phase, while at field scale typically reservoirs with pressure and temperature compatible with supercritical CO_2 is sought. This has a minor impact on viscosity, but leads to a denser and less compressible CO_2 phase.

What actually is very different from reservoir conditions at depth is the importance of pressure measurements. In the experiment, pressure signals are rather uninformative and might introduce differences in permeability interpretation, whereas they are valuable in a reservoir context. Another major consideration is that the subsurface is much harder to characterize than the experimental rig, and so the uncertainties in predictions are going to be dominated

by uncertainties in geological characterisation. This validation benchmark study illustrates the range of predictions that are possible in a relatively well-characterised system.

From a reservoir simulation perspective, all participants reported that they struggled to achieve acceptable run times, and were forced to use relatively coarse grids for this study. We speculate that this is due to the low density of the gas phase, which has the consequence that when CO_2 dissolves into water, the resulting mixture has significantly lower volume than before mixing. This study thus provides impetus for further development of efficient non-linear solvers for soluble gas-water systems.

Acknowledgements B. Flemisch thanks the German Research Foundation (DFG) for supporting this work by funding SFB 1313, Project Number 327154368. S. Geiger acknowledges partial funding from Energi Simulation. H. Hajibeygi was sponsored by the Dutch National Science Foundation (NWO) under Vidi Talent Program Project "ADMIRE" (Project Number 17509).

Author Contributions BF, JMN, MF and RJ conceptualized, designed and implemented the validation benchmark study. JWB was involved in the image analysis of the experimental results and contributed the Python script for the calculation of the Wasserstein metric. All other authors constitute the participating groups and correspondingly set up, executed and evaluated the simulations and provided the requested results, together with descriptions of the underlying models. BF and JMN wrote the initial draft of the manuscript, all other authors were involved in the internal review and editing process. BF wrote and executed the scripts for generating all figures in the manuscript. All authors read and approved the final manuscript.

Funding S. Geiger acknowledges partial funding from Energi Simulation. H. Hajibeygi was sponsored by the Dutch National Science Foundation (NWO) under Vidi Talent Program Project "ADMIRE" (Project Number 17509).

Data Availability All data which has been used for generating the figures in this paper is collected in respective repositories of the GitHub "FluidFlower" organization, which is accessible at github.com/fluidflower. In particular, the results provided by the participating groups are collected in repositories github.com/fluidflower/*groupname* .git, where *groupname* is out of austin, csiro, delft, heriot-watt, lanl, melbourne, stanford and stuttgart. The experimental data used for comparison with the modeling results is assembled in github.com/fluidflower/experiment. The scripts for the generation of all figures are contained in github.com/fluidflower/general.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Appendix A Spatial Parameters

A.1 Experimentally Measured Parameters

The following table depicts the experimentally measured spatial parameters for the six facies, namely, the absolute permeability K, the porosity ϕ, the irreducible liquid/gas saturation

$S_{l,i}/S_{g,i}$ with corresponding endpoint gas/liquid relative permeability $k_{r,g}/k_{r,l}$, and the capillary entry pressure $p_{c,e}$. For details on the measurements and more spatial parameters, we refer to Section 2.3 of the description (Nordbotten et al. 2022).

Facies	K [m^2]	ϕ	$S_{l,i}$	$k_{r,g}$	$S_{g,i}$	$k_{r,l}$	$p_{c,e}$ [PA]
G	9.45e−9	0.45	0.10	0.16	0.06	0.75	0.00e0
F	4.20e−9	0.44	0.12	0.11	0.13	0.72	0.00e0
E	1.98e−9	0.45	0.12	0.10	0.06	0.93	0.00e0
D	1.10e−9	0.44	0.12	0.02	0.08	0.95	9.81e1
C	4.67e−10	0.44	0.14	0.05	0.10	0.93	2.94e2
ESF	4.34e−11	0.44	0.32	0.09	0.14	0.71	1.47e3

A.2 Model Parameters

In the following, the spatial parameters as chosen by each participating group are listed. Regarding the constitutive relations, additional parameters than the experimentally measured ones had to be selected, see also Table 2. For most participants, this concerns the Brooks-Corey pore-size distribution index $p_{c,\lambda}$, while `Delft-DARSim` and `Heriot-Watt` had to choose individual exponents n_l and n_g for the power law for each fluid phase.

A.2.1 Austin

Facies	K [m^2]	ϕ	$S_{l,i}$	$k_{r,g}$	$S_{g,i}$	$k_{r,l}$	$p_{c,e}$ [PA]	p_c, λ
G	9.58e−9	0.46	0.10	0.16	0.06	0.75	0.00e0	2
F	4.26e−9	0.43	0.12	0.11	0.13	0.72	0.00e0	2
E	2.01e−9	0.45	0.12	0.1	0.06	0.93	0.00e0	2
D	1.11e−9	0.44	0.12	0.02	0.08	0.95	9.81e1	2
C	4.73e−10	0.43	0.14	0.05	0.1	0.93	2.94e2	2
ESF	4.40e−11	0.44	0.32	0.09	0.14	0.71	1.47e3	2
barrier	1.00e−18	0.01	0	1	0	1	0.00e0	2

A.2.2 CSIRO

Facies	K [m^2]	ϕ	$S_{l,i}$	$k_{r,g}$	$S_{g,i}$	$k_{r,l}$	$p_{c,e}$ [PA]	p_c, λ
G	4.10e−9	0.44	0.1	0.16	0.06	0.75	1.00e1	2
F	4.12e−9	0.45	0.12	0.11	0.13	0.72	1.00e1	2
E	2.52e−9	0.45	0.12	0.10	0.06	0.93	1.00e1	2
D	1.08e−9	0.44	0.12	0.02	0.08	0.95	9.81e1	2
C	4.68e−10	0.44	0.14	0.05	0.1	0.93	2.94e2	2
ESF	5.70e−11	0.43	0.32	0.09	0.14	0.71	1.47e3	2
barrier	0.00e0	0.1	0.05	1.0	0.0	1.0	0.00e0	2

A.2.3 Delft-DARSim

Facies	K_x[m^2]	K_z [m^2]	ϕ	$S_{l,i}$	$k_{r,g}$	$S_{g,i}$	$k_{r,l}$	$p_{c,e}$ [PA]	p_c, λ
G	4.31e−9	4.79e−9	0.46	0.1	0.16	0.06	0.75	0.00e0	2
F	1.92e−9	2.13e−9	0.43	0.12	0.11	0.13	0.72	0.00e0	2
E	9.02e−10	1.00e−9	0.45	0.12	0.1	0.06	0.93	0.00e0	2
D	5.00e−10	5.55e−10	0.44	0.12	0.02	0.08	0.95	9.81e1	2
C	2.13e−10	2.37e−10	0.43	0.14	0.05	0.1	0.93	2.94e2	2
ESF	1.98e−11	2.20e−11	0.44	0.32	0.09	0.14	0.71	1.47e3	2
barrier	4.50e−19	5.00e−19	0.001	0.32	0.09	0.14	0.71	0.00e0	2

The participant `Delft-DARSim` used anisotropic permeabilities with values K_x and K_z in x- and z-direction, respectively.

A.2.4 Delft-DARTS

Facies	K_x [m^2]	a_z	ϕ	$S_{l,i}$	$k_{r,g}$	$S_{g,i}$	$k_{r,l}$	$p_{c,e}$ [PA]	n_l	n_g
G	2.25e−9	1.0	0.44	0.11	0.85	0.06	0.80	0.00e0	2.0	1.5
F	3.20e−9	1.0	0.45	0.11	0.85	0.06	0.80	0.00e0	2.0	1.5
E	1.42e−9	0.9	0.45	0.11	0.85	0.06	0.80	0.00e0	2.0	1.5
D	1.85e−9	0.8	0.44	0.12	0.95	0.08	0.93	1.00e2	2.0	1.5
C	3.81e−10	0.7	0.44	0.14	0.95	0.10	0.93	3.00e2	2.5	2.0
ESF	4.34e−11	0.75	0.43	0.32	0.75	0.14	0.71	1.50e3	2.5	2.0
Fault-1	6.36e−9	1.0	0.44	0.11	0.85	0.06	0.80	0.00e0	2.0	1.5
Fault-2	2.82e−9	1.0	0.44	0.11	0.85	0.06	0.80	0.00e0	2.0	1.5

The participant `Delft-DARTS` used anisotropic permeabilities with values K_x in x-direction and proportionality factors a_z for the z-direction.

A.2.5 Heriot-Watt

Facies	K_x [m^2]	ϕ	$S_{l,i}$	$k_{r,g}$	$S_{g,i}$	$k_{r,l}$	$p_{c,e}$ [PA]	p_c, λ	n_l	n_g
G	9.45e−9	0.45	0.1	.16	0.06	.75	4.00e1	n/a	3	1.5
F	4.20e−9	0.44	0.12	0.11	0.13	.72	4.50e1	n/a	3	1.5
E	1.98e−9	0.45	0.12	0.1	0.06	.93	5.00e1	n/a	3	1.5
D	1.10e−9	0.44	0.12	0.02	0.08	.95	9.80e1	1	3	1.5
C	4.67e−10	0.44	0.14	0.05	0.1	.93	2.94e2	1	3	1.5
ESF	4.34e−11	0.44	0.32	0.09	0.14	0.71	1.47e3	1	3	1.5
barrier	1.00e−16	0.01	0.0	1.0	0.0	1.0	0.00e0	1	3	1.5

A.2.6 LANL

Facies	K [m^2]	ϕ	$S_{\mathrm{l,i}}$	$k_{\mathrm{r,g}}$	$S_{\mathrm{g,i}}$	$k_{\mathrm{r,l}}$	$p_{\mathrm{c,e}}$ [PA]	p_{c}, λ
G	9.58e−9	0.46	0.1	0.16	0.06	0.75	0.00e0	1
F	4.26e−9	0.43	0.12	0.11	0.13	0.72	0.00e0	1
E	2.01e−9	0.45	0.12	0.1	0.06	0.93	0.00e0	1
D	1.11e−9	0.44	0.12	0.02	0.08	0.95	9.81e1	1
C	4.73e−10	0.43	0.14	0.05	0.1	0.93	2.94e2	1
ESF	4.40e−11	0.44	0.32	0.09	0.14	0.71	1.47e3	1
barrier	1.00e−15	0.1	0	1	0	1	0.00e0	1

A.2.7 Melbourne

Facies	K [m^2]	ϕ	$S_{\mathrm{l,i,pc}}$	$k_{\mathrm{r,g,kr}}$	$p_{\mathrm{c,e}}$ [PA]	p_{c}, λ
G	9.45e−9	0.44	0.0999	0.1	1.00e1	3.50
F	4.20e−9	0.45	0.1199	0.12	3.23e1	3.62
E	1.98e−9	0.45	0.1196	0.12	3.63e2	3.70
D	1.10e−9	0.44	0.1167	0.12	6.57e2	3.70
C	4.67e−10	0.44	0.1356	0.14	1.44e3	3.70
ESF (Top Seal)	2.66e−11	0.43	0.31968	0.32	6.06e3	3.00
ESF (Bottom Seal)	4.34e−11	0.43	0.31968	0.32	6.06e3	3.00
Sealing Fault	9.87e−15	0.25	–	0	5.00e5	0.00
Free Flow	1.18e−8	1	–	0	0.00e0	0.00

The participant `Melbourne` used two values for the residual saturation of the liquid phase, $S_{\mathrm{l,i,pc}}$ for evaluating the capillary pressure and $S_{\mathrm{l,i,kr}}$ for the relative permeability.

A.2.8 Stanford

Facies	K [m^2]	ϕ	$S_{\mathrm{l,i}}$	$k_{\mathrm{r,g}}$	$S_{\mathrm{g,i}}$	$k_{\mathrm{r,l}}$	$p_{\mathrm{c,e}}$ [PA]	p_{c}, λ
G	9.58e−9	0.46	0.1	0.16	0.06	0.75	0.00e0	2
F	4.26e−9	0.43	0.12	0.11	0.13	0.72	0.00e0	2
E	2.01e−9	0.45	0.12	0.1	0.06	0.93	0.00e0	2
D	1.11e−9	0.44	0.12	0.02	0.08	0.95	9.81e1	2
C	4.73e−10	0.43	0.14	0.05	0.1	0.93	2.94e2	2
ESF	4.40e−11	0.44	0.32	0.09	0.14	0.71	1.47e3	2
barrier	1.00e−15	0.1	0	1	0	1	0.00e0	2

A.2.9 Stuttgart

Facies	K [m^2]	ϕ	$S_{\mathrm{l,i}}$	$k_{\mathrm{r,g}}$	$S_{\mathrm{g,i}}$	$k_{\mathrm{r,l}}$	$p_{\mathrm{c,e}}$ [PA]	p_{c}, λ
G	4.88e−10	0.46	0.14	0.16	0.1	0.72	0.00e0	2
F	2.58e−10	0.43	0.03	0.11	0.0	1.0	0.00e0	2
E	7.08e−10	0.45	0.21	0.10	0.0	1.0	0.00e0	2
D	4.24e−10	0.44	0.17	0.20	0.0	1.0	9.81e1	2
C	2.39e−10	0.43	0.15	0.20	0.0	1.0	2.94e2	2
ESF	3.90e−11	0.44	0.05	1.0	0.0	1.0	1.47e3	2
barrier	1.00e−16	0.1	0.05	1.0	0.0	1.0	0.00e0	2

B Wasserstein Distances

The following tables list the Wasserstein distances between the spatial maps as provided by the participating groups for each requested timestep. For the calculation, the Python library POT (Flamary et al. 2021) has been used. The full data including distances between results from different timesteps is provided in the FluidFlower general GitHub repository. For obtaining the numbers depicted in Fig. 12, the normalized table values of dimension meter were multiplied by 850 g cm m^{-1} to arrive at the desired dimension of gram times centimeter. The value 8.5 refers to the mass of injected CO_2 in gram.

B.1 24 h

dist [m]	Austin	CSIRO	DARSim	DARTS	LANL	Melbourne	Stanford	Stuttgart
Austin	0.00e0	1.97e−1	2.25e−1	1.39e−1	7.12e−1	2.50e−1	3.19e−1	1.87e−1
CSIRO	1.97e−1	0.00e0	7.62e−2	1.01e−1	5.27e−1	1.18e−1	4.93e−1	3.71e−2
DARSim	2.25e−1	7.62e−2	0.00e0	1.52e−1	5.00e−1	1.13e−1	5.23e−1	8.81e−2
DARTS	1.39e−1	1.01e−1	1.52e−1	0.00e0	5.91e−1	1.40e−1	4.28e−1	1.18e−1
LANL	7.12e−1	5.27e−1	5.00e−1	5.91e−1	0.00e0	4.85e−1	1.02e0	5.33e−1
Melbourne	2.50e−1	1.18e−1	1.13e−1	1.40e−1	4.85e−1	0.00e0	5.35e−1	1.51e−1
Stanford	3.19e−1	4.93e−1	5.23e−1	4.28e−1	1.02e0	5.35e−1	0.00e0	4.90e−1
Stuttgart	1.87e−1	3.71e−2	8.81e−2	1.18e−1	5.33e−1	1.51e−1	4.90e−1	0.00e0

B.2 48 h

dist [m]	Austin	CSIRO	DARSim	DARTS	LANL	Melbourne	Stanford	Stuttgart
Austin	0.00e0	1.58e−1	1.94e−1	1.46e−1	6.47e−1	2.02e−1	3.39e−1	2.32e−1
CSIRO	1.58e−1	0.00e0	8.94e−2	6.62e−2	4.99e−1	9.40e−2	4.52e−1	9.45e−2
DARSim	1.94e−1	8.94e−2	0.00e0	1.17e−1	4.60e−1	9.46e−2	4.92e−1	1.13e−1
DARTS	1.46e−1	6.62e−2	1.17e−1	0.00e0	5.08e−1	1.13e−1	4.43e−1	1.12e−1
LANL	6.47e−1	4.99e−1	4.60e−1	5.08e−1	0.00e0	4.58e−1	9.49e−1	4.20e−1
Melbourne	2.02e−1	9.40e−2	9.46e−2	1.13e−1	4.58e−1	0.00e0	4.92e−1	1.36e−1
Stanford	3.39e−1	4.52e−1	4.92e−1	4.43e−1	9.49e−1	4.92e−1	0.00e0	5.35e−1
Stuttgart	2.32e−1	9.45e−2	1.13e−1	1.12e−1	4.20e−1	1.36e−1	5.35e−1	0.00e0

B.3 72 h

dist [m]	Austin	CSIRO	DARSim	DARTS	LANL	Melbourne	Stanford	Stuttgart
Austin	0.00e0	1.15e−1	1.59e−1	1.43e−1	6.33e−1	1.41e−1	3.22e−1	2.40e−1
CSIRO	1.15e−1	0.00e0	1.32e−1	8.68e−2	5.64e−1	7.48e−2	3.60e−1	1.70e−1
DARSim	1.59e−1	1.32e−1	0.00e0	1.32e−1	4.83e−1	9.89e−2	4.37e−1	1.46e−1
DARTS	1.43e−1	8.68e−2	1.32e−1	0.00e0	5.06e−1	8.39e−2	4.14e−1	1.21e−1
LANL	6.33e−1	5.64e−1	4.83e−1	5.06e−1	0.00e0	5.28e−1	9.19e−1	3.97e−1
Melbourne	1.41e−1	7.48e−2	9.89e−2	8.39e−2	5.28e−1	0.00e0	3.92e−1	1.64e−1
Stanford	3.22e−1	3.60e−1	4.37e−1	4.14e−1	9.19e−1	3.92e−1	0.00e0	5.26e−1
Stuttgart	2.40e−1	1.70e−1	1.46e−1	1.21e−1	3.97e−1	1.64e−1	5.26e−1	0.00e0

B.4 96 h

dist [m]	Austin	CSIRO	DARSim	DARTS	LANL	Melbourne	Stanford	Stuttgart
Austin	0.00e0	1.08e−1	1.37e−1	1.33e−1	6.14e−1	1.10e−1	3.12e−1	2.28e−1
CSIRO	1.08e−1	0.00e0	1.59e−1	9.81e−2	6.10e−1	6.81e−2	3.33e−1	2.25e−1
DARSim	1.37e−1	1.59e−1	0.00e0	1.45e−1	4.86e−1	1.15e−1	4.00e−1	1.50e−1
DARTS	1.33e−1	9.81e−2	1.45e−1	0.00e0	5.27e−1	7.38e−2	3.61e−1	1.43e−1
LANL	6.14e−1	6.10e−1	4.86e−1	5.27e−1	0.00e0	5.74e−1	8.86e−1	3.87e−1
Melbourne	1.10e−1	6.81e−2	1.15e−1	7.38e−2	5.74e−1	0.00e0	3.14e−1	1.96e−1
Stanford	3.12e−1	3.33e−1	4.00e−1	3.61e−1	8.86e−1	3.14e−1	0.00e0	5.01e−1
Stuttgart	2.28e−1	2.25e−1	1.50e−1	1.43e−1	3.87e−1	1.96e−1	5.01e−1	0.00e0

B.5 120 h

dist [m]	Austin	CSIRO	DARSim	DARTS	LANL	Melbourne	Stanford	Stuttgart
Austin	0.00e0	1.25e−1	1.25e−1	1.24e−1	5.91e−1	1.03e−1	3.05e−1	2.00e−1
CSIRO	1.25e−1	0.00e0	1.69e−1	8.21e−2	6.27e−1	7.59e−2	3.27e−1	2.35e−1
DARSim	1.25e−1	1.69e−1	0.00e0	1.57e−1	4.85e−1	1.30e−1	3.69e−1	1.39e−1
DARTS	1.24e−1	8.21e−2	1.57e−1	0.00e0	5.57e−1	7.54e−2	3.36e−1	1.66e−1
LANL	5.91e−1	6.27e−1	4.85e−1	5.57e−1	0.00e0	6.03e−1	8.54e−1	3.92e−1
Melbourne	1.03e−1	7.59e−2	1.30e−1	7.54e−2	6.03e−1	0.00e0	2.71e−1	2.14e−1
Stanford	3.05e−1	3.27e−1	3.69e−1	3.36e−1	8.54e−1	2.71e−1	0.00e0	4.64e−1
Stuttgart	2.00e−1	2.35e−1	1.39e−1	1.66e−1	3.92e−1	2.14e−1	4.64e−1	0.00e0

C Sparse Data Provided by the Participants

The following tables present the sparse data as provided by the participants. The values $P_{10}(\bar{x})$, $P_{50}(\bar{x})$ and $P_{90}(\bar{x})$ indicate the P10, P50 and P90 values of the expected mean of the respective quantity, whereas $P_{10}(\sigma)$, $P_{50}(\sigma)$ and $P_{90}(\sigma)$ refer to the correspondingly expected standard deviation. The values are also contained in the respective participant repositories.

C.1 Austin

SRQ	$P_{10}\,(\bar{x})$	$P_{50}\,(\bar{x})$	$P_{90}\,(\bar{x})$	$P_{10}\,(\sigma)$	$P_{50}\,(\sigma)$	$P_{90}\,(\sigma)$
1a [N/m^2]	1.11e5	1.11e5	1.13e5	9.21e2	9.08e2	1.47e3
1b [N/m^2]	1.05e5	1.05e5	1.06e5	4.34e2	4.14e2	7.75e2
2 [s]	1.86e4	1.86e4	1.86e4	0.00e0	0.00e0	0.00e0
3a [kg]	1.70e−3	1.82e−3	2.11e−3	2.08e−4	1.49e−4	2.93e−4
3b [kg]	2.96e−4	3.15e−4	3.45e−4	2.96e−5	2.04e−5	3.45e−5
3c [kg]	1.15e−3	1.51e−3	1.99e−3	4.72e−4	2.96e−4	5.53e−4
3d [kg]	8.37e−5	1.36e−4	3.31e−4	1.35e−4	1.12e−4	1.99e−4
4a [kg]	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0
4b [kg]	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0
4c [kg]	1.20e−6	4.44e−5	2.51e−4	1.39e−4	1.21e−4	2.10e−4
4d [kg]	1.56e−12	7.85e−8	6.50e−6	3.32e−6	3.29e−6	5.98e−6
5 [s]	1.35e4	1.62e4	1.80e4	2.83e3	1.63e3	2.68e3
6 [kg]	8.49e−5	1.58e−4	3.37e−4	1.42e−4	1.09e−4	1.92e−4

C.2 CSIRO

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	1.11e5	1.11e5	1.11e5	6.10e0	8.32e0	1.05e1
1b [N/m^2]	1.06e5	1.06e5	1.06e5	3.53e0	5.75e0	7.98e0
2 [s]	1.80e4	1.80e4	1.80e4	0.00e0	0.00e0	0.00e0
3a [kg]	1.67e−4	3.49e−4	8.33e−4	2.89e−4	3.90e−4	4.91e−4
3b [kg]	0.00e0	2.80e−7	1.11e−5	3.97e−6	5.35e−6	6.74e−6
3c [kg]	2.95e−3	3.60e−3	4.08e−3	4.11e−4	5.54e−4	6.97e−4
3d [kg]	5.32e−4	5.69e−4	6.06e−4	2.23e−5	3.04e−5	3.85e−5
4a [kg]	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0
4b [kg]	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0
4c [kg]	1.42e−4	2.98e−4	5.42e−4	1.46e−4	1.97e−4	2.48e−4
4d [kg]	2.82e−7	1.03e−7	3.43e−7	9.78e−7	1.34e−6	1.69e−6
5 [s]	6.11e4	8.33e4	1.28e5	2.61e4	3.52e4	4.42e4
6 [kg]	5.33e−4	5.72e−4	6.61e−4	5.38e−5	7.34e−5	9.31e−5

C.3 Delft-DARSim

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	1.10e5	1.10e5	1.10e5	1.92e0	1.92e0	1.92e0
1b [N/m^2]	1.04e5	1.04e5	1.04e5	3.90e0	3.90e0	3.90e0
2 [s]	1.38e4	1.41e4	1.74e4	1.68e3	1.68e3	1.68e3
3a [kg]	0.00e0	7.61e−4	1.71e−3	7.12e−4	7.12e−4	7.12e−4
3b [kg]	0.00e0	0.00e0	1.44e−5	7.85e−6	7.85e−6	7.85e−6
3c [kg]	1.12e−3	2.20e−3	3.29e−3	9.06e−4	9.06e−4	9.06e−4
3d [kg]	3.88e−6	6.68e−6	1.84e−5	6.59e−6	6.59e−6	6.59e−6
4a [kg]	8.05e−10	2.43e−9	1.31e−5	6.67e−6	6.67e−6	6.67e−6
4b [kg]	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0
4c [kg]	4.29e−4	6.26e−4	8.37e−4	1.80e−4	1.80e−4	1.80e−4
4d [kg]	3.38e−9	1.33e−8	1.44e−7	7.46e−8	7.46e−8	7.46e−8
5 [s]	n/a	n/a	n/a	n/a	n/a	n/a
6 [kg]	6.68e−6	1.20e−5	2.65e−5	8.92e−6	8.92e−6	8.92e−6

C.4 Delft-DARTS

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	1.11e5	1.11e5	1.11e5	5.45e1	5.45e1	5.45e1
1b [N/m^2]	1.05e5	1.05e5	1.05e5	3.91e1	3.91e1	3.91e1
2 [s]	1.62e4	1.74e4	1.80e4	6.64e2	6.64e2	6.64e2
3a [kg]	2.29e−4	1.15e−3	1.93e−3	6.27e−4	6.27e−4	6.27e−4
3b [kg]	1.90e−5	8.30e−5	1.40e−4	4.40e−5	4.40e−5	4.40e−5
3c [kg]	1.38e−3	2.32e−3	3.55e−3	8.02e−4	8.02e−4	8.02e−4
3d [kg]	1.29e−4	3.98e−4	6.57e−4	2.13e−4	2.13e−4	2.13e−4
4a [kg]	0.00e0	0.00e0	0.00e0	1.00e−6	1.00e−6	1.00e−6
4b [kg]	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0	0.00e0
4c [kg]	1.00e−6	1.00e−5	4.60e−5	2.90e−5	2.90e−5	2.90e−5
4d [kg]	0.00e0	0.00e0	0.00e0	1.00e−6	1.00e−6	1.00e−6
5 [s]	n/a	n/a	n/a	n/a	n/a	n/a
6 [kg]	2.14e−4	5.20e−4	6.58e−4	1.75e−4	1.75e−4	1.75e−4

C.5 Heriot-Watt

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	1.10e5	1.10e5	1.10e5	2.00e1	2.00e1	2.00e1
1b [N/m^2]	1.04e5	1.04e5	1.04e5	2.00e1	2.00e1	2.00e1
2 [s]	1.80e4	1.86e4	1.90e4	6.00e2	6.00e2	6.00e2
3a [kg]	0.00e0	1.00e−3	1.80e−3	5.00e−4	5.00e−4	5.00e−4
3b [kg]	0.00e0	1.00e−4	3.00e−4	1.00e−4	1.00e−4	1.00e−4
3c [kg]	1.00e−3	2.00e−3	4.00e−3	1.00e−3	1.00e−3	1.00e−3
3d [kg]	0.00e0	5.00e−4	9.00e−4	3.00e−4	3.00e−4	3.00e−4
4a [kg]	0.00e0	1.00e−5	1.00e−4	5.00e−7	5.00e−6	5.00e−5
4b [kg]	0.00e0	1.00e−6	1.00e−5	5.00e−8	5.00e−7	5.00e−6
4c [kg]	0.00e0	1.00e−4	1.00e−3	5.00e−5	5.00e−5	5.00e−4
4d [kg]	0.00e0	2.00e−5	1.00e−4	5.00e−6	5.00e−6	5.00e−6
5 [s]	6.00e3	1.00e4	1.50e4	1.00e3	1.00e3	1.00e3
6 [kg]	0.00e0	5.00e−5	1.00e−4	0.00e0	1.00e−5	1.00e−4

C.6 LANL

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	1.11e5	1.11e5	1.11e5	1.00e0	1.00e0	1.00e0
1b [N/m^2]	1.05e5	1.05e5	1.05e5	1.00e0	1.00e0	1.00e0
2 [s]	1.00e11	1.00e9	4.32e5	4.32e5	4.32e5	4.32e5
3a [kg]	0.00e0	0.00e0	0.00e0	1.00e−6	5.00e−7	1.00e−7
3b [kg]	0.00e0	0.00e0	0.00e0	1.00e−6	5.00e−7	1.00e−7
3c [kg]	7.24e−5	7.24e−5	7.24e−5	1.00e−6	1.00e−6	1.00e−6
3d [kg]	5.24e−5	5.24e−5	5.24e−5	1.00e−6	1.00e−6	1.00e−6
4a [kg]	0.00e0	0.00e0	0.00e0	1.00e−6	1.00e−6	1.00e−6
4b [kg]	2.37e−5	2.37e−5	2.37e−5	1.00e−6	1.00e−6	1.00e−6
4c [kg]	1.99e−3	1.99e−3	1.99e−3	1.00e−4	1.00e−4	1.00e−4
4d [kg]	5.77e−4	5.77e−4	5.77e−4	1.00e−5	1.00e−5	1.00e−5
5 [s]	8.64e6	8.64e6	8.64e6	3.60e3	3.60e3	3.60e3
6 [kg]	1.16e−4	1.16e−4	1.16e−4	1.00e−5	1.00e−5	1.00e−5

C.7 Melbourne

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	1.12e5	1.12e5	1.13e5	3.55e2	2.75e2	3.06e2
1b [N/m^2]	1.05e5	1.05e5	1.06e5	2.36e2	1.88e2	2.79e2
2 [s]	1.60e4	1.71e4	1.79e4	7.41e1	3.23e2	1.20e2
3a [kg]	4.02e−4	5.83e−4	7.50e−4	1.83e−5	3.84e−5	6.04e−5
3b [kg]	n/a	n/a	n/a	n/a	n/a	n/a
3c [kg]	3.64e−3	3.88e−3	4.28e−3	n/a	n/a	n/a
3d [kg]	6.67e−4	7.58e−4	8.57e−4	6.67e−4	7.58e−4	8.54e−4
4a [kg]	3.16e−17	3.39e−17	3.81e−17	n/a	n/a	n/a
4b [kg]	n/a	n/a	n/a	n/a	n/a	n/a
4c [kg]	1.57e−3	1.61e−3	1.62e−3	n/a	n/a	n/a
4d [kg]	2.58e−6	3.81e−6	2.34e−5	n/a	n/a	n/a
5 [s]	1.23e4	1.31e4	1.40e4	1.26e2	1.53e2	2.68e2
6 [kg]	n/a	n/a	n/a	n/a	n/a	n/a

C.8 Stanford

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	1.11e5	1.11e5	1.12e5	n/a	2.37e1	n/a
1b [N/m^2]	1.05e5	1.06e5	1.06e5	n/a	2.38e1	n/a
2 [s]	1.82e4	2.66e4	2.31e4	n/a	1.83e3	n/a
3a [kg]	0.00e0	1.90e−3	2.26e−3	n/a	8.57e−4	n/a
3b [kg]	1.96e−4	2.03e−3	5.89e−3	n/a	2.93e−3	n/a
3c [kg]	1.22e−3	5.29e−3	5.64e−3	n/a	2.06e−3	n/a
3d [kg]	1.37e−3	2.04e−3	5.99e−3	n/a	2.70e−3	n/a
4a [kg]	0.00e0	0.00e0	2.23e−4	n/a	1.66e−4	n/a
4b [kg]	0.00e0	0.00e0	2.70e−5	n/a	2.01e−5	n/a
4c [kg]	7.80e−6	4.50e−5	2.15e−4	n/a	9.52e−5	n/a
4d [kg]	0.00e0	0.00e0	0.00e0	n/a	0.00e0	n/a
5 [s]	n/a	1.35e4	n/a	n/a	n/a	n/a
6 [kg]	1.36e−3	2.03e−3	5.98e−3	n/a	2.70e−3	n/a

C.9 Stuttgart

SRQ	$P_{10}(\bar{x})$	$P_{50}(\bar{x})$	$P_{90}(\bar{x})$	$P_{10}(\sigma)$	$P_{50}(\sigma)$	$P_{90}(\sigma)$
1a [N/m^2]	n/a	1.10e5	n/a	n/a	8.10e0	n/a
1b [N/m^2]	n/a	1.04e5	n/a	n/a	6.21e0	n/a
2 [s]	n/a	1.77e4	n/a	n/a	9.43e2	n/a
3a [kg]	n/a	2.00e−3	n/a	n/a	5.42e−4	n/a
3b [kg]	n/a	1.01e−6	n/a	n/a	1.43e−6	n/a
3c [kg]	n/a	1.60e−3	n/a	n/a	4.96e−4	n/a
3d [kg]	n/a	3.20e−4	n/a	n/a	3.49e−4	n/a
4a [kg]	n/a	1.79e−6	n/a	n/a	4.06e−6	n/a
4b [kg]	n/a	0.00e0	n/a	n/a	0.00e0	n/a
4c [kg]	n/a	3.95e−4	n/a	n/a	2.55e−4	n/a
4d [kg]	n/a	9.05e−6	n/a	n/a	1.23e−5	n/a
5 [s]	n/a	3.39e4	n/a	n/a	1.38e4	n/a
6 [kg]	n/a	3.57e−4	n/a	n/a	3.93e−4	n/a

References

American Society of Mechanical Engineers: Guide for Verification and Validation in Computational Solid Mechanics: an American National Standard. ASME Press, New York (2006). https://www.asme.org/products/codes-standards/v-v-10-2006-guide-verification-validation

Bachu, S., Bonijoly, D., Bradshaw, J., Burruss, R., Holloway, S., Christensen, N.P., Mathiassen, O.M.: CO_2 storage capacity estimation: methodology and gaps. Int. J. Greenh. Gas Control **1**(4), 430–443 (2007). https://doi.org/10.1016/S1750-5836(07)00086-2

Carroll, J.J., Slupsky, J.D., Mather, A.E.: The solubility of carbon dioxide in water at low pressure. J. Phys. Chem. Ref. Data **20**(6), 1201–1209 (1991). https://doi.org/10.1063/1.555900

Class, H., Ebigbo, A., Helmig, R., Dahle, H.K., Nordbotten, J.M., Celia, M.A., Audigane, P., Darcis, M., Ennis-King, J., Fan, Y., Flemisch, B., Gasda, S.E., Jin, M., Krug, S., Labregere, D., Naderi Beni, A., Pawar, R.J., Sbai, A., Thomas, S.G., Trenty, L., Wei, L.: A benchmark study on problems related to CO_2 storage in geologic formations. Comput. Geosci. **13**(4), 409–434 (2009). https://doi.org/10.1007/s10596-009-9146-x

Duan, Z., Sun, R.: An improved model calculating CO_2 solubility in pure water and aqueous NaCl solutions from 273 to 533 K and from 0 to 2000 bar. Chem. Geol. **193**(3), 257–271 (2003). https://doi.org/10.1016/S0009-2541(02)00263-2

Elenius, M.T., Johannsen, K.: On the time scales of nonlinear instability in miscible displacement porous media flow. Comput. Geosci. **16**(4), 901–911 (2012). https://doi.org/10.1007/s10596-012-9294-2

Fenghour, A., Wakeham, W.A., Vesovic, V.: The viscosity of carbon dioxide. J. Phys. Chem. Ref. Data **27**(1), 31–44 (1998). https://doi.org/10.1063/1.556013

Fernø, M.A., Haugen, M., Eikehaug, K., Folkvord, O., Benali, B., Both, J.W., Storvik, E., Nixon, C.W., Gawthrope, R.L., Nordbotten, J.M.: Room-scale CO_2 injections in a physical reservoir model with faults. Transp. Porous Media (2023)

Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N.T.H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D.J., Tavenard, R., Tong, A., Vayer, T.: POT: python optimal transport. J. Mach. Learn. Res. **22**(78), 1–8 (2021)

Garcia, J.E.: Density of aqueous solutions of CO_2. Technical Report LBNL-49023, LBNL (2001)

Garipov, T.T., Tomin, P., Rin, R., Voskov, D.V., Tchelepi, H.A.: Unified thermo-compositional-mechanical framework for reservoir simulation. Comput. Geosci. **22**(4), 1039–1057 (2018). https://doi.org/10.1007/s10596-018-9737-5

Halland, E.K., Riis, F., Magnus, C., Johansen, W.T., Tappel, I.M., Gjeldvik, I.T., Solbakk, T., Pham, V.T.H.: CO_2 storage atlas of the Norwegian part of the north sea. Energy Proc. **37**, 4919–4926 (2013). https://doi.org/10.1016/j.egypro.2013.06.403

Hovorka, S.D., Benson, S.M., Doughty, C., Freifeld, B.M., Sakurai, S., Daley, T.M., Kharaka, Y.K., Holtz, M.H., Trautz, R.C., Nance, H.S., Myer, L.R., Knauss, K.G.: Measuring permanence of CO_2 storage in saline formations: the Frio experiment. Environ. Geosci. **13**(2), 105–121 (2006). https://doi.org/10.1306/eg.11210505011

IAPWS: Revised Release on the IAPWS Industrial Formulation 1997 for the Thermodynamic Properties of Water and Steam. Technical report, IAPWS (2007). www.iapws.org/relguide/IF97-Rev.pdf

Johnson, N., Parker, N., Ogden, J.: How negative can biofuels with CCS take us and at what cost? Refining the economic potential of biofuel production with CCS using spatially-explicit modeling. Energy Proc. **63**, 6770–6791 (2014). https://doi.org/10.1016/j.egypro.2014.11.712

Juanes, R., MacMinn, C.W., Szulczewski, M.L.: The footprint of the CO_2 plume during carbon dioxide storage in saline aquifers: storage efficiency for capillary trapping at the basin scale. Transp. Porous Media **82**(1), 19–30 (2010). https://doi.org/10.1007/s11242-009-9420-3

Koch, T., Glaeser, D., Weishaupt, K., Ackermann, S., Beck, M., Becker, B., Burbulla, S., Class, H., Coltman, E., Emmert, S., Fetzer, T., Grueninger, C., Heck, K., Hommel, J., Kurz, T., Lipp, M., Mohammadi, F., Scherrer, S., Schneider, M., Seitz, G., Stadler, L., Utz, M., Weinhardt, F., Flemisch, B.: DuMux 3–an open-source simulator for solving flow and transport problems in porous media with a focus on model coupling. Comput. Math. Appl. **81**, 423–443 (2021). https://doi.org/10.1016/j.camwa.2020.02.012

Kopp, A., Class, H., Helmig, R.: Investigations on CO_2 storage capacity in saline aquifers: part: 1. Dimensional analysis of flow processes and reservoir characteristics. Int. J. Greenh. Gas Control **3**(3), 263–276 (2009a). https://doi.org/10.1016/j.ijggc.2008.10.002

Kopp, A., Class, H., Helmig, R.: Investigations on CO_2 storage capacity in saline aquifers–part 2: estimation of storage capacity coefficients. Int. J. Greenh. Gas Control **3**(3), 277–287 (2009). https://doi.org/10.1016/j.ijggc.2008.10.001

Kovscek, A.R., Nordbotten, J.M., Ferno, M.A.: Scaling up FluidFlower results for carbon dioxide storage in geological media (2023). https://doi.org/10.48550/ARXIV.2301.09853

Lichtner, P.C., Hammond, G.E., Lu, C., Karra, S., Bisht, G., Andre, B., Mills, R., Kumar, J.: PFLOTRAN user manual: a massively parallel reactive flow and transport model for describing surface and subsurface processes. OSTI (2015). https://doi.org/10.2172/1168703

Lie, K.-A.: An Introduction to Reservoir Simulation Using MATLAB/GNU Octave: User Guide for the MATLAB Reservoir Simulation Toolbox (MRST). Cambridge University Press, Cambridge (2019). https://doi.org/10.1017/9781108591416

Lindeberg, E., Vuillaume, J.-F., Ghaderi, A.: Determination of the CO_2 storage capacity of the Utsira formation. Energy Proc. **1**(1), 2777–2784 (2009). https://doi.org/10.1016/j.egypro.2009.02.049

Lüth, S., Henninges, J., Ivandic, M., Juhlin, C., Kempka, T., Norden, B., Rippe, D., Schmidt-Hattenberger, C.: Chapter 6.2—Geophysical monitoring of the injection and postclosure phases at the Ketzin pilot site. In: Kasahara, J., Zhdanov, M.S., Mikada, H. (eds.) Active Geophysical Monitoring, 2nd edn., pp. 523–561. Elsevier, Amsterdam (2020). https://doi.org/10.1016/B978-0-08-102684-7.00025-X

Lyu, X., Khait, M., Voskov, D.: Operator-based linearization approach for modeling of multiphase flow with buoyancy and capillarity. SPE J. **26**(4), 1858–1878 (2021). https://doi.org/10.2118/205378-PA

Matthäi, S., Geiger, S., Roberts, S.: Complex systems platform: Csp3d3. 0: user's guide. Technical report, ETH Zurich (2001)

Metz, B., Davidson, O., De Coninck, H., Loos, M., Meyer, L.: IPCC Special Report on Carbon Dioxide Capture and Storage. Cambridge University Press, Cambridge (2005)

Niemi, A., Bensabat, J., Shtivelman, V., Edlmann, K., Gouze, P., Luquot, L., Hingerl, F., Benson, S.M., Pezard, P.A., Rasmusson, K., Liang, T., Fagerlund, F., Gendler, M., Goldberg, I., Tatomir, A., Lange, T., Sauter, M., Freifeld, B.: Heletz experimental site overview, characterization and data analysis for CO_2 injection and geological storage. Int. J. Greenh. Gas Control **48**, 3–23 (2016). https://doi.org/10.1016/j.ijggc.2015.12.030

Niemi, A., Bensabat, J., Joodaki, S., Basirat, F., Hedayati, M., Yang, Z., Perez, L., Levchenko, S., Shklarnik, A., Ronen, R., Goren, Y., Fagerlund, F., Rasmusson, K., Moghadasi, R., Shoqeir, J.A.H., Sauter, M., Ghergut, I., Gouze, P., Freifeld, B.: Characterizing CO_2 residual trapping in-situ by means of single-well push-pull experiments at Heletz, Israel, pilot injection site-experimental procedures and results of the experiments. Int. J. Greenh. Gas Control **101**, 103129 (2020). https://doi.org/10.1016/j.ijggc.2020.103129

Nordbotten, J.M., Flemisch, B., Gasda, S.E., Nilsen, H.M., Fan, Y., Pickup, G.E., Wiese, B., Celia, M.A., Dahle, H.K., Eigestad, G.T., Pruess, K.: Uncertainties in practical simulation of CO_2 storage. Int. J. Greenh. Gas Control **9**, 234–242 (2012). https://doi.org/10.1016/j.ijggc.2012.03.007

Nordbotten, J.M., Fernø, M., Flemisch, B., Juanes, R., Jørgensen, M.: Final Benchmark Description: FluidFlower International Benchmark Study (2022). https://doi.org/10.5281/zenodo.6807102

Nordbotten, J.M., Benali, B., Both, J.W., Brattekås, B., Storvik, E., Fernø, M.: DarSIA: An open-source Python toolbox for two-scale image processing of dynamics in porous media. Transp. Porous Media (2023a)

Nordbotten, J.M., Jørgensen, M., Fernø, M., Flemisch, B., Juanes, R.: Experimentally assessing the uncertainty of forecasts of geological carbon storage (Submitted) (2023b)

Noussan, M., Raimondi, P.P., Scita, R., Hafner, M.: The role of green and blue hydrogen in the energy transition–a technological and geopolitical perspective. Sustainability (2021). https://doi.org/10.3390/su13010298

Oberkampf, W.L., Trucano, T.G.: Verification and validation benchmarks. Nucl. Eng. Des. **238**(3), 716–743 (2008). https://doi.org/10.1016/j.nucengdes.2007.02.032

Oberkampf, W.L., Roy, C.J.: Verification and Validation in Scientific Computing. Cambridge University Press, Cambridge (2010)

Pacala, S., Socolow, R.: Stabilization wedges: solving the climate problem for the next 50 years with current technologies. Science **305**(5686), 968–972 (2004). https://doi.org/10.1126/science.1100103

Panaretos, V.M., Zemel, Y.: Statistical aspects of Wasserstein distances. Ann. Rev. Stat. Appl. **6**(1), 405–431 (2019). https://doi.org/10.1146/annurev-statistics-030718-104938

Peng, D.-Y., Robinson, D.B.: A new two-constant equation of state. Ind. Eng. Chem. Fundam. **15**(1), 59–64 (1976). https://doi.org/10.1021/i160057a011

Preston, C., Monea, M., Jazrawi, W., Brown, K., Whittaker, S., White, D., Law, D., Chalaturnyk, R., Rostron, B.: IEA GHG Weyburn CO_2 monitoring and storage project. Fuel Process. Technol. **86**(14), 1547–1568 (2005). https://doi.org/10.1016/j.fuproc.2005.01.019

Pruess, K., Garciia, J., Kovscek, T., Oldenburg, C., Rutqvist, J., Steefel, C., Xu, T.: Code intercomparison builds confidence in numerical simulation models for geologic disposal of CO_2. Energy **29**(9–10), 1431–1444 (2004). https://doi.org/10.1016/j.energy.2004.03.077

Sandve, T.H., Gasda, S.E., Rasmussen, A., Rustad, A.B.: Convective dissolution in field scale CO_2 storage simulations using the OPM flow simulator. In: TCCS–11. CO_2 Capture, Transport and Storage. Trondheim 22nd–23rd June 2021 Short Papers from the 11th International Trondheim CCS Conference. SINTEF Academic Press (2021)

Scheer, D., Class, H., Flemisch, B.: Subsurface Environmental Modelling Between Science and Policy. Springer, Cham (2021)

Sharma, S., Cook, P., Jenkins, C., Steeper, T., Lees, M., Ranasinghe, N.: The CO2CRC Otway project: leveraging experience and exploiting new opportunities at Australia's first CCS project site. Energy Proc. **4**, 5447–5454 (2011). https://doi.org/10.1016/j.egypro.2011.02.530

Soave, G.: Equilibrium constants from a modified Redlich-Kwong equation of state. Chem. Eng. Sci. **27**(6), 1197–1203 (1972). https://doi.org/10.1016/0009-2509(72)80096-4

Span, R., Wagner, W.: A new equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 K at pressures up to 800 MPa. J. Phys. Chem. Ref. Data **25**, 1509–1596 (1996). https://doi.org/10.1063/1.555991

Span, R., Wagner, W.: Equations of state for technical applications. I. Simultaneously optimized functional forms for nonpolar and polar fluids. Int. J. Thermophys. **24**(1), 1–39 (2003). https://doi.org/10.1023/A:1022390430888

Spycher, N., Pruess, K.: CO_2–H_2O mixtures in the geological sequestration of CO_2. II, partitioning in chloride brines at 12–100 °C and up to 600 bar. Geochim. Cosmochim. Acta **69**(13), 3309–3320 (2005). https://doi.org/10.1016/j.gca.2005.01.015

Spycher, N., Pruess, K., Ennis-King, J.: CO_2-H_2O mixtures in the geological sequestration of CO_2. I. Assessment and calculation of mutual solubilities from 12 to 100 °C and up to 600 bar. Geochimica et Cosmochimica Acta **67**(16), 3015–3031 (2003). https://doi.org/10.1016/S0016-7037(03)00273-4

Steyn, M., Oglesby, J., Turan, G., Zapantis, A., Gebremedhin, R., Zapantis, A., Amer, N.A., Havercroft, I., Ivory-Moore, R., Steyn, M., Yang, X., Gebremedhin, R., Zahra, M.A., Pinto, E., Rassool, D., Williams, E., Consoli, C., Minervini, J.: Global Status of CCS (2022). https://status22.globalccsinstitute.com/wp-content/uploads/2022/11/Global-Status-of-CCS-2022_Download.pdf

Wang, Y., Vuik, C., Hajibeygi, H.: Analysis of hydrodynamic trapping interactions during full-cycle injection and migration of CO_2 in deep saline aquifers. Adv. Water Resour. **159**, 104073 (2022). https://doi.org/10.1016/j.advwatres.2021.104073

Weiss, R.F.: Carbon dioxide in water and seawater: the solubility of a non-ideal gas. Mar. Chem. **2**(3), 203–215 (1974). https://doi.org/10.1016/0304-4203(74)90015-2

Wilkins, A., Green, C.P., Ennis-King, J.: An open-source multiphysics simulation code for coupled problems in porous media. Comput. Geosci. **154**, 104820 (2021). https://doi.org/10.1016/j.cageo.2021.104820

Younis, R., Tchelepi, H.A., Aziz, K.: Adaptively localized continuation-Newton method-nonlinear solvers that converge all the time. SPE J. **15**(02), 526–544 (2010). https://doi.org/10.2118/119147-PA

Youssef, A.A., Shao, Q., Matthäi, S.K.: Simplified numeric simulation approach for CO_2, g-water flow and trapping at near-surface conditions. Transp, Porous Media (2023)

Zhou, Y.: Parallel general-purpose reservoir simulation with coupled reservoir models and multisegment wells. PhD thesis, Stanford University (2012)

Zhou, Y., Jiang, Y., Tchelepi, H.A.: A scalable multistage linear solver for reservoir models with multisegment wells. Comput. Geosci. **17**(2), 197–216 (2013). https://doi.org/10.1007/s10596-012-9324-0

Ziabakhsh-Ganji, Z., Kooi, H.: An equation of state for thermodynamic equilibrium of gas mixtures and brines to allow simulation of the effects of impurities in subsurface CO_2 storage. Int. J. Greenh. Gas Control **11**, 21–34 (2012). https://doi.org/10.1016/j.ijggc.2012.07.025

Zyvoloski, G.A., Robinson, B.A., Dash, Z.V., Trease, L.L.: Summary of the models and methods for the FEHM application–a finite-element heat- and mass-transfer code. OSTI (1997). https://doi.org/10.2172/14903

Authors and Affiliations

Bernd Flemisch[1] · Jan M. Nordbotten[2,3] · Martin Fernø[3,4] · Ruben Juanes[5] · Jakub W. Both[2] · Holger Class[1] · Mojdeh Delshad[6] · Florian Doster[7] · Jonathan Ennis-King[8] · Jacques Franc[9] · Sebastian Geiger[7,10] · Dennis Gläser[1] · Christopher Green[8] · James Gunning[8] · Hadi Hajibeygi[10] · Samuel J. Jackson[8] · Mohamad Jammoul[6] · Satish Karra[11] · Jiawei Li[9] · Stephan K. Matthäi[12] · Terry Miller[11] · Qi Shao[12] · Catherine Spurin[9] · Philip Stauffer[11] · Hamdi Tchelepi[9] · Xiaoming Tian[10] · Hari Viswanathan[11] · Denis Voskov[10] · Yuhang Wang[10] · Michiel Wapperom[10] · Mary F. Wheeler[6] · Andrew Wilkins[13] · AbdAllah A. Youssef[12] · Ziliang Zhang[10]

Jan M. Nordbotten
jan.nordbotten@uib.no

Martin Fernø
martin.ferno@uib.no

Ruben Juanes
juanes@mit.edu

Jakub W. Both
jakub.both@uib.no

Holger Class
holger.class@iws.uni-stuttgart.de

Mojdeh Delshad
delshad@mail.utexas.edu

Florian Doster
f.doster@hw.ac.uk

Jonathan Ennis-King
jonathan.ennis-king@csiro.au

Jacques Franc
jfranc@stanford.edu

Sebastian Geiger
s.geiger@tudelft.nl

Dennis Gläser
dennis.glaeser@iws.uni-stuttgart.de

Christopher Green
chris.green@csiro.au

James Gunning
james.gunning@csiro.au

Hadi Hajibeygi
h.hajibeygi@tudelft.nl

Samuel J. Jackson
samuel.jackson@csiro.au

Mohamad Jammoul
jammoul@utexas.edu

Satish Karra
satkarra@lanl.gov

Jiawei Li
jiaweili@stanford.edu

Stephan K. Matthäi
stephan.matthai@unimelb.edu.au

Terry Miller
tamiller@lanl.gov

Qi Shao
shao.q@unimelb.edu.au

Catherine Spurin
cspurin@stanford.edu

Philip Stauffer
stauffer@lanl.gov

Hamdi Tchelepi
tchelepi@stanford.edu

Xiaoming Tian
x.tian-1@tudelft.nl

Hari Viswanathan
viswana@lanl.gov

Denis Voskov
d.v.voskov@tudelft.nl

Yuhang Wang
wangyuhang17@cug.edu.cn

Michiel Wapperom
m.b.wapperom@tudelft.nl

Mary F. Wheeler
mfw@ices.utexas.edu

Andrew Wilkins
andrew.wilkins@csiro.au

AbdAllah A. Youssef
abdallahy@student.unimelb.edu.au

Ziliang Zhang
z.zhang-15@tudelft.nl

[1] Department of Hydromechanics and Modelling of Hydrosystems, University of Stuttgart, Stuttgart, Germany

[2] Department of Mathematics, Center for Modeling of Coupled Subsurface Dynamics, University of Bergen, Bergen, Norway

[3] Center of Sustainable Subsurface Resources, Norwegian Research Center, Bergen, Norway

[4] Department of Physics and Technology, University of Bergen, Bergen, Norway

[5] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[6] Center for Subsurface Modeling, Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA

[7] Institute of Geoenergy Engineering, Heriot Watt University, Edinburgh, UK

[8] CSIRO Energy, Clayton North, Australia

[9] Department of Energy Science and Engineering, Stanford University, Stanford, CA, USA

[10] Department of Geoscience and Engineering, Delft University of Technology, Delft, The Netherlands

[11] Computational Earth Science, Los Alamos National Laboratory, Los Alamos, NM, USA

[12] Department of Infrastructure Engineering, Peter Cook Center for CCS Research, The University of Melbourne, Parkville, Australia

[13] CSIRO Mineral Resources, Queensland Centre for Advanced Technologies, Kenmore, Australia