International Journal of Greenhouse Gas Control 135 (2024) 104162

Contents lists available at ScienceDirect

International Journal of Greenhouse Gas Control

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

Experimentally assessing the uncertainty of forecasts of geological carbon storage

Jan M. Nordbotten [a,b,*], Martin Fernø [c,b], Bernd Flemisch [d], Ruben Juanes [e], Magne Jørgensen [f]

[a] Center for Modeling of Coupled Subsurface Dynamics, Department of Mathematics, University of Bergen, Bergen, Norway
[b] Center for Sustainable Subsurface Resources, Norwegian Research Center, Postboks 22 Nygårdstangen, 5838 Bergen, Norway
[c] Department of Physics and Technology, University of Bergen, Bergen, Norway
[d] Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Stuttgart, Germany
[e] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Boston, USA
[f] Simula Metropolitan Center for Digital Engineering, Pilestredet 52, 0167 Oslo, Norway

ARTICLE INFO

Keywords:
Forecasting
Carbon storage
Computational science
Overconfidence
Uncertainty quantification

ABSTRACT

Geological storage of carbon dioxide is a cornerstone in almost every realistic emissions reduction scenario outlined by the Intergovernmental Panel on Climate Change. Our ability to accurately forecast storage efficacy is, however, mostly unknown due to the long timescales involved (hundreds to thousands of years). To study perceived forecast accuracy, we designed a double-blind forecasting study. As ground truth, we constructed a laboratory-scale carbon storage operation, retaining the essential physical processes active on the field scale, within a time span of five days. Separately, academic groups with experience in carbon storage research were invited to forecast key carbon storage efficacy metrics. The participating groups submitted forecasts in two stages: First independently without any cross-group interaction, then finally after workshops designed to share and assimilate understanding between the forecast groups. Their confidence in reported forecasts was monitored throughout the forecasting study. Our results show that participating groups provided forecasts that appear bias-free with respect to carbon storage as a technology, yet the forecast intervals are too narrow to capture the ground truth (overconfidence bias). When asked to qualitatively self-assess their forecast uncertainty (and later when asked to provide an external assessment of other forecast groups), the assessment of the participants indicated an understanding that the forecast intervals (both their own and those of others) were too narrow. However, the participants did not display an understanding of how poorly the forecast intervals calibrated to the ground truth. The quantitative uncertainty assessments contrast the qualitative comments supplied by the participants, which indicate an acute awareness of the challenges associated with assessing the uncertainty of forecasts for complex systems such as the geological storage of carbon dioxide.

1. Introduction

Models of large-scale geophysical systems often share characteristics such as a complex interplay of non-linear processes, reliance on constitutive laws of varying degrees of sophistication, and spatio-temporal domains that prohibit fully resolved computer simulation. Arguably, this is the case for both above-surface and sub-surface systems, including seasonal weather models, general circulation models, and, as considered in this study, long-term geological carbon storage.

With the exception of weather, forecasts of such complex, large-scale geophysical systems are usually challenged by the lack of robust datasets for which the reliability of forecasting tools can be established. Uncertainty quantification based on forecasts provided by computer simulation may therefore be limited to exploring the propagation of parameter uncertainty (e.g., by an ensemble of simulations), often without addressing —with the same rigor— other sources of uncertainty, including (but not limited to) model deficiencies, approximations and user errors (Ferson et al., 2004; Smith, 2013; Qian et al., 2018). Crucially, the impact of these frequently ignored uncertainties need not be unbiased, and as such, neither the mean nor the span of the simulation ensemble may be representative of the actual uncertainty one tries to capture in the forecast of the system (Morgan and Henrion, 1990, Cooke, 1991).

The use of probabilistic forecasts is a well-established methodology

* Corresponding author.
E-mail address: jan.nordbotten@uib.no (J.M. Nordbotten).

https://doi.org/10.1016/j.ijggc.2024.104162
Received 2 February 2024; Received in revised form 13 May 2024; Accepted 20 May 2024
Available online 3 June 2024

to communicate uncertainty. As pointed out in Murphy (1998), probabilistic statements on how much confidence one should ascribe to a weather forecast date back to the 18th century. Since then, forecast intervals and probabilistic single-point forecasts have been used in many disciplines. In the context of subsurface energy (petroleum or geothermal) and groundwater extraction, business decisions related to field development are routinely dependent on forecast intervals, and the calibration of these intervals in terms of actual field performance can be assessed in a matter of years or decades. This previous experience with giving probabilistic uncertainty forecasts also means that we now have some general knowledge about our ability to provide well-calibrated forecast intervals and unbiased single-point forecasts at yearly or decadal timescales (see e.g. Floris et al. (2001); Tavassoli (2004); Bickel and Bratvold (2008)). In contrast, in the context of nuclear waste disposal, the existing field data is insignificant compared to the forecast time, and several studies have illustrated the challenges of providing well-calibrated forecasts (see e.g. results from the long-running DECOVALEX project reported in Jing et al. (1995), Birkholzer (2019)).

Forecast intervals in general tend to be too narrow to reflect the stated confidence level, i.e., there is an overconfidence bias. For example, well-calibrated 80 %-confidence forecasting intervals should, in the long run, give an 80 % hit rate (inclusion rate) of observed outcomes, but most reported studies on forecast intervals find that the actual hit rate is much lower, see for example Soll et al. (2004) and Glaser et al. (2013). The root cause has been debated extensively and includes both psychological explanations, such as the desire to be informative and not expose uncertainty (Cesarini, 2006), and model-based explanations, such as shortcomings in including all types of uncertainties in the model simulations (Klas, 2011). Most previous research also points at a tendency for overoptimistic single-point forecasts of, for example, the mean outcome, although this tendency seems to be less robust and more context-dependent (Halkjelsvik og Jørgensen, 2012).

Industrial-scale geological carbon storage (GCS), which is a key carbon mitigation technology (IPCC, 2005, 2022), is an important example of a complex geophysical system where long-term datasets are lacking. Only a handful of industrial-scale storage sites are in operation (Zhang et al., 2022) (these include the Sleipner project in the North Sea (Furre et al., 2017), the Weyburn project in Canada (White, 2009), and the Gorgon project in Western Australia, (Trupp, 2021)), and the geological environment varies greatly from site to site. Hence, there is a danger of overemphasizing learning from existing sites when assessing and planning new sites. Furthermore, commercial-scale carbon storage sites have only been operated for a few decades, and none are in the post-injection phase of operation. As such, there exist no datasets relating to the long-term (centuries) fate of industrial carbon storage for which our forecasting ability can be assessed.

Here, we examine the forecasting ability of a set of modeling groups active in the field of carbon dioxide storage. We focus on their ability to capture the uncertainty of the forecasts, and to give unbiased forecasts, based on the ground truth provided by a room-scale experiment of carbon dioxide injection and trapping in geologically realistic media constructed explicitly for this study (Fernø et al., 2024). The forecast accuracy and model comparisons are discussed elsewhere (Flemisch et al., 2024). Throughout this study, forecasts are given as forecast intervals, i.e., minimum-maximum intervals associated with respectively 10 % and 90 % probability of not exceeding the actual experimental outcome, and as forecasts of the median outcome, i.e., the outcome with a 50 % probability to be exceeded (see Nomenclature for a summary of definitions).

To our knowledge, there have not been any prior studies on overconfidence or overoptimism when forecasting carbon storage operations. Here we address this gap and establish a baseline understanding of our forecasting abilities in the context of carbon storage, together with a rigorous examination of how uncertainty assessment improved with group discussions and requests for alternative framing of uncertainty.

The research questions guiding our study were the following:

1. To what extent does the stated confidence in the forecast intervals correspond to the actual hit rate of the outcome (level of calibration)?
2. To what extent is there a tendency towards optimism, with respect to the safety of carbon storage as a technology, in the forecasts of the median outcomes?
3. What is the impact of community interactions (communication with other teams targeting the same forecasts) on calibration and optimism bias?
4. How accurate are assessments of the hit rate of their own (meta-assessment) and the other groups' forecast intervals?

We address these research questions through a forecasting study defined around a laboratory-scale carbon storage operation. A carbon storage site is constructed and operated in the laboratory, which is our representation of *ground truth*, i.e., the reference values for the evaluation of the forecasts, and multiple academic groups from around the world participated in model building and forecasts. The laboratory storage site is realized as an intermediate-scale visual quasi-2D experiment—an approach that has proven extraordinarily insightful in understanding subsurface fluid flow and transport (Illangasekare et al., 1995; Trevisan et al., 2017). It was constructed to be as realistic as possible with respect to physical processes, geologic complexity, and information availability, and it can be considered comparable to a multi-decadal commercial injection at field conditions, with a multi-century post-injection period (Kovscek et al., 2024). The forecast groups were monitored throughout the study to address their skill in representing the accuracy of their forecasts, both as individuals and as a collective, and quantitative and qualitative information was collected and analyzed. The study was strictly double-blind, in the sense that the experimental group was fully isolated from the forecasting groups. Furthermore, during the first part of the study, the forecasting groups were also blind to each other, while in the second part of the study, the forecasting groups were interacting with each other.

In the remaining part of this paper, we describe the study design in Section 2 and the results in Section 3, before we discuss the results and conclude in Section 4.

2. Study design

To address the research questions, we designed a study consisting of two experiments, running concurrently. One experiment, denoted the "ground truth" experiment, defines the results to be forecasted, and has been reported on elsewhere (see references below). The second experiment, which we report herein, is denoted the "forecast experiment", and targets the forecasts themselves. The steps and the timeline of the two experiments are depicted in Fig. 1 and explained below.

2.1. The physical ground truth experiment

For the purpose of creating a realistic yet well-defined environment for forecasts, we constructed a long-term carbon storage site at laboratory scale (Fernø et al., 2024). By long-term, we mean that we allow for a full injection phase, and a post-injection phase that is on the order of ten times the length of the injection. Our carbon storage experiment, the FluidFlower experiment, was built to the maximal size allowable within laboratory limitations, and measures almost two meters high by three meters wide. The experiment is quasi two-dimensional, having a depth of approximately two centimeters, and the front panel is transparent for accurate optical-based monitoring of the carbon storage processes. While an exact scaling from field to lab is impossible, the choice of dimensions and sands provides a reasonable consistency of dimensionless groups when compared to ongoing carbon storage operations (Kovscek et al., 2024).

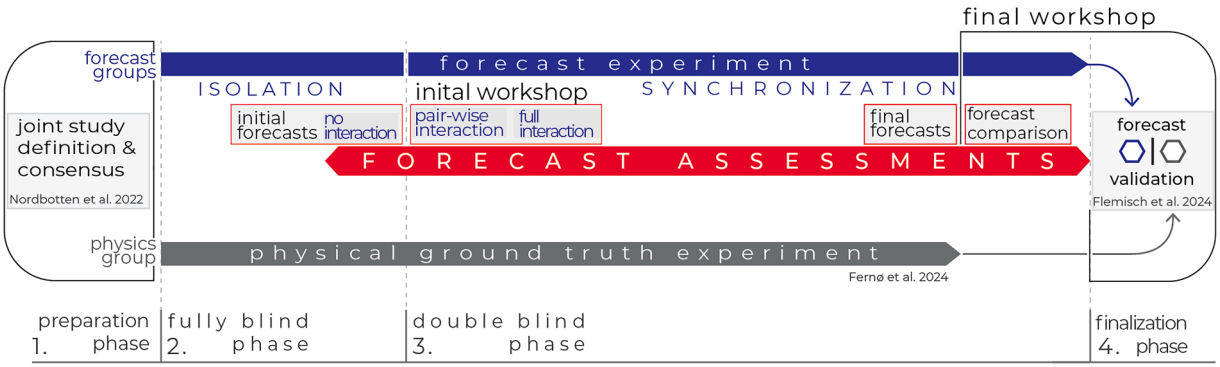

Fig. 1. Overview of the steps and timeline of the study, including both the physical ground truth experiment and the components of the forecast experiment. The overall timescale of the figure spans from Spring 2021 to April 2022.

Experimentally measurable input parameters were documented and released to the participating forecasting groups, as well as a tracer injection run for model calibration. The data made available was chosen to be realistic as compared to the pre-injection assessment phase of a real storage operation. In the preparation phase, comments and clarifications were allowed by all participants, and the final system description was made static early in the blind phase (Nordbotten et al. 2022). The experimental group, led by one of the authors (Fernø), based at the University of Bergen, Norway, conducted five repetitions of the experiment (Fernø et al., 2024), without sharing results or any other material information with the forecasting groups or the co-authors outside of Bergen. The idealized laboratory carbon storage site was constructed using unconsolidated sands (six separate size distributions ranging from an average grain size of 0.2 to 2.5 mm) with absolute permeability ranging from 5 to 10 000 Darcy. A typical carbon storage experiment is shown in Fig. 2 (Panel I). One injection well was located in the lower and homogenous reservoir (to the left of Box A and C), whereas the second injection well was located in the fining upwards sequence above the lower reservoir seal. Gaseous CO_2 was injected at constant rate through the two injection wells over the initial 5 h of the experiment, with the upper injection active only during the last 2.5 h. Difference in capillary entry pressures dictated the observed flow and trapping patterns, and CO_2-saturated water was distinguished from formation water by a color shift of the aqueous pH sensitive solution. The gaseous CO_2 phase was observed by lack of water in the sand. All data were analyzed and quantified using open-source image analysis software (Nordbotten et al., 2024a). Repeated CO_2 injections with identical operational conditions allow physical variability to be addressed using the same geological geometry.

2.2. The forecast experiment

Leading academic and industry groups with active carbon storage research were invited to participate in the forecasting study, of which a total of 10 groups (all academic) participated in the forecast study, of which seven groups provided sufficient data to be included in the results reported herein. Of these seven, four groups provided the forecast quantities requested in this study, while the remaining three groups submitted forecast quantities only at the final workshop. The three groups who did not provide data for this study, were nevertheless part of the knowledge exchange at the initial and final workshops. Typical numerical simulations of the carbon storage experiment used to inform the forecasts are shown in Fig. 2 (Panel II). To ensure that the study was double-blind, the forecasting groups were coordinated by one of the authors (Flemisch), based at the University of Stuttgart, Germany, with moderated and archived communication between forecasting and experimental groups by means of a dedicated online platform.[1]

2.3. The forecasted quantities

The forecasting groups were asked to provide forecasts of six proxy questions associated with storage capacity and security (see Fig. 2, Panel I, and (Nordbotten et al., 2022) for the precise definition of Boxes A-C and injection/pressure port positions). Three of the proxies comprise several numbers, referred to as quantities:

1. *As a proxy for assessing risk of mechanical disturbance of the overburden*: Maximum pressure at sensor number 1 and 2 (referred to as quantity 1a and 1b). Unit: Newtons per square meter.
2. *As a proxy for when leakage risk starts declining*: Time of maximum mobile free phase in Box A. Unit: Seconds.
3. *As a proxy for the ability to accurately forecast near well phase partitioning*: Phase composition in Box A at 72 h after CO_2 injection starts (3a through 3d). Unit: Kilograms. The four quantities are specified as:
   a. Mobile CO_2 in gas phase
   b. Immobile CO_2 in gas phase
   c. Dissolved CO_2.
   d. CO_2 in seal facies.
4. *As a proxy for the ability to handle uncertain geological features*: Phase composition in Box B at 72 h after injection starts (4a through 4d, according to the same convention as in Proxy 3). Unit: Kilograms.
5. *As a proxy for the ability to capture onset of convective mixing*: Time for which the quantity

$$\int_C \left| \nabla \left(\frac{\chi_c^w(t)}{\chi_{c,\max}^w} \right) \right| dx$$

   first exceeds 110 % of the width of Box C, where χ_c^w is the mass fraction of CO_2 in the water phase. Unit: Seconds.
6. *As a proxy for the ability to capture migration into low-permeable seals*: Total mass of CO_2 in the top seal facies (areas of the finest sand layers) at final time within Box A. Unit: Kilograms.

Noting that the answers to Proxies 1, 3 and 4 include multiple numbers, the total response to the six proxies involves providing forecasts of 13 numerical quantities. See Fig. 2, Panel III, for an example of the forecasts of the four quantities comprising Proxy 3. These quantities will in turn not be completely independent, as they survey different

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

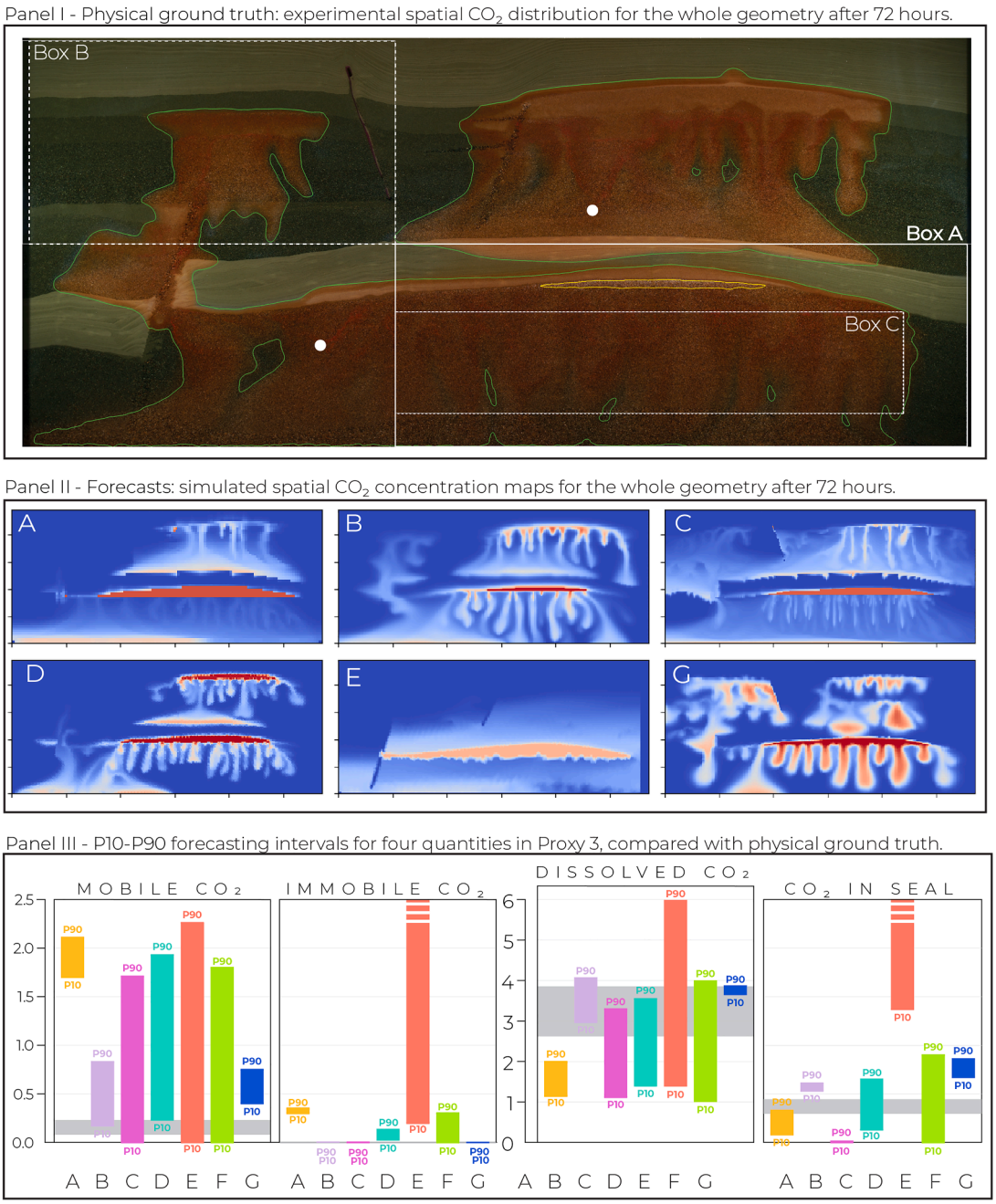

Fig. 2. Panel I: Photo of the third repetition of the ground truth experiment, with analysis boxes A-C referred to in the text overlain. Grayish-beige colors correspond to water-filled sand layers with no CO_2, reddish darker colors suggest a single fluid phase consisting of liquid water with dissolved CO_2, while reddish light colors indicate the additional appearance of free-phase gaseous CO_2. Automatic segmentation indicated by yellow and green curves. White dots mark the position of the injection points. Panel II: Numerical simulations as submitted by six of the forecasting groups. Here, the coloration indicates concentration of CO_2 in the water phase, with gaseous CO_2 appearing where the dissolution limit is exceeded. The seventh group did not trust their simulations, and therefore submitted forecasts based on expert knowledge. Panel III: Submitted P10-P90 forecasts for the four quantities in Proxy 3 from the seven groups (each group has unique color), compared with range of five repeated experiments (gray).

aspects of the forecasts of the same physical system. We emphasize that the quantitative results of this study (in terms of numerical values given in Table 1 and Fig. 4, see Section 3) depend on the choice of proxy questions. However, we believe it is reasonable to expect that the qualitative results (summarized at the beginning of Section 4) should be true also for other proxies.

For each of the quantities, the mean value of the five measurements from the physical ground truth experiments was forecasted by the research groups. The inherent measurement and analysis uncertainty in the outcome of the experiment implies that the mean value of the five measurements is represented by an interval. We term the mean, minimum and maximum of the five empirical measurements, respectively, EI_{Mean}, EI_{Min}, and EI_{max}. The interval $[EI_{Min}, EI_{Max}]$ of the five empirical measurements is denoted EI. We use the intervals of empirically measured values as a reference when evaluating forecast interval calibration, and EI_{Mean} when evaluating forecast optimism bias and forecast accuracy. More on the quantities, size of this uncertainty interval of the measured mean values, and how measurements were made, can be found in (Fernø et al. 2024).

Table 1
Hit rates of P10-P90 intervals and tendency towards optimism bias per question. "Initial" refers to forecasts without interaction with other groups (four groups submitted intervals), "Final" refers to forecasts with the benefit of group interactions and additional time (seven groups submitted intervals). Well-calibrated P10-P90 intervals would give an 80% hit rate. Unbiased P50 forecasts would give a 50% optimism proportion, while values higher than 50% would suggest a tendency towards over-optimistic forecasts.

Quantity	Initial hit rate *HitRateP*	Initial optimism proportion *P50OptP*	Final hit rate P10-P90 *HitRateP*	Final optimism proportion *P50OptP*
1a	25%	25%	0%	29%
1b	25%	25%	14%	29%
2	25%	25%	14%	14%
3a	25%	50%	49%	0%
3b	75%	63%	57%	86%
3c	34%	25%	60%	41%
3d	25%	25%	31%	34%
4a	25%	13%	100%	29%
4b	25%	100%	100%	57%
4c	0%	100%	24%	14%
4d	0%	0%	43%	14%
5	25%	50%	32%	63%
6	29%	63%	24%	70%
Overall mean	**26%**	**43%**	**42%**	**37%**

2.4. The forecasts

Forecasts are classically understood in the context of probability distributions. However, in practice, it is typically considered unrealistic to provide the precise shape of the probability distribution, and the distribution is replaced by a finite number of characteristics. The most common choices are either mean and variance or, as considered in this study, exceedance values (percentiles) of the cumulative probability distribution. In the forecast experiment, we requested forecasts of the following three percentiles of the outcome probability distribution of a quantity:

P10: The value forecasted to be 10 % likely *not* to be exceeded by the empirically measured value.
P50: The value forecasted to be 50 % likely *not* to be exceeded by the empirically measured value.
P90: The value forecasted to be 90 % likely *not* to be exceeded by the empirically measured value.

The interval from P10 to P90, denoted *PI80*, is then the (central) 80 % forecast interval, i.e., the interval with an 80 % probability of including the empirically measured value. Furthermore, we consider the P50 exceedance value as the "most representative" (the median) forecast.

Thus, forecasting groups provided P10, P50, and P90, as exemplified for Proxy 3 in Fig. 2, Panel III.[2] The methods for achieving the answers could be chosen freely by each group, allowing for a broad spectrum ranging from pure intuition and expert knowledge to sophisticated statistical uncertainty quantification approaches.

2.5. Evaluation measures of the forecasts

For this study, ground truth values are defined as the experimental results. However, the ubiquitous presence of measurement error and data analysis uncertainty must be acknowledged. We choose the following approach to score the forecasts in light of the uncertainty associated with the ground truth data: Based on our best estimate of experimental uncertainty (detailed in Fernø et al., 2024), we take a probabilistic perspective where the ground truth is represented by a uniform distribution within *EI*, i.e., the interval between the lowest (EI_{Min}) and the highest (EI_{Max}) value of the relevant quantity as measured in the five ground truth experiments.

We score an 80 % forecast interval (*PI80*) according to its overlap with the experimental uncertainty (*EI*). This approach provides a probabilistically consistent interpretation when aggregating over all participating groups. The hit rates of the *PI80* intervals (*HitRateP*) can then, according to the above argumentation, be expressed as follows (See Fig. 3 for an illustration)[3]:

$$HitRateP = \frac{HitFrac80}{EI_{Max} - EI_{Min}}$$

Here, *HitFrac80* is the overlap between the forecast and the measurement, which formally can be expressed as:

$$HitFrac80 = \min(EI_{Max}, P90) - \max(EI_{Min}, P10)$$

A special case arises for two quantities within Proxy 3 and 4, where the reference value (empirically measured values) is zero with high confidence, and thus $EI_{min} = EI_{max} = 0$. In this case, we define the formula as the traditional measure of hit rate for single point empirical values, i.e. we score a $HitRateP = 1$ if the participants submitted $P10 = 0$.

We use a variant of the measure *HitRateP* for the hit rate of the proxies, denoted *HitRateProxy*. This is defined as the *product* of the hit rates of the quantities within a proxy, which have up to four quantities. To get a hit rate of 1 for a proxy, one consequently needs to have a hit rate of 1 for all sub-quantities within that proxy.[4]

For the purpose of defining a measure of *optimism bias* of the *P50* forecast we first identified whether the positive direction for each quantity, i.e., better outcome regarding CO_2 storage, was towards higher values ("Higher") or lower values ("Lower"). We encode the optimism direction by ξ, taking the value $\xi = 0$ if optimism direction is positive and $\xi = 1$ if optimism direction is negative. The direction of optimism is described and justified for each of the quantities below:

- Pressure values (Quantities 1a and 1b): Higher induces the risk of damage to overburden, lower is positive: $\xi = 1$.
- Both time measures (Quantities 2 and 5): Values describe transitions of the system to safer states, lower is therefore positive: $\xi = 1$.
- CO_2 in mobile free phase (Quantities 3a and 4a): Mobile CO_2 can leak, thus lower is positive: $\xi = 1$.
- CO_2 in immobile free phase (Quantities 3b and 4b): Immobile CO_2 cannot move, thus higher is positive: $\xi = 0$.
- CO_2 dissolved in water (Quantities 3c and 4c): Water with elevated CO_2 levels tends to sink, and this prohibits leakage. Higher is positive: $\xi = 0$.

[2] We asked for the P10, P50 and P90 for both the mean and the standard deviation of the experimental results, which in total gives six predictions for each quantity. As only two of the groups gave meaningful P10, P50 and P90 for the standard deviation of the experiments, the analysis in this paper is only based on the P10, P50 and P90 forecasts of the mean of the empirically measured value.

[3] The presence of measurement uncertainty and the use of an interval forecast evaluation measure does not substantially influence the results reported herein. On the level of quantities, the measurement uncertainty was relatively small, in the sense that 86% of the forecast intervals either fully overlap or do not overlap at all with the interval of experimental values. Thus, the precise method chosen for calculating fractional hit-rates for the remaining 14% of the forecasts has only a minor effect in aggregate.

[4] The interpretation of *HitRateProxy* is not straightforward. While a perfectly calibrated 80% confidence forecast interval should give an 80% hit rate, the normatively correct hit rate on the proxy level is more complicated. A calculation of the normatively correct hit rate would require knowledge about both the confidence levels and the dependencies between the quantities that form one proxy.

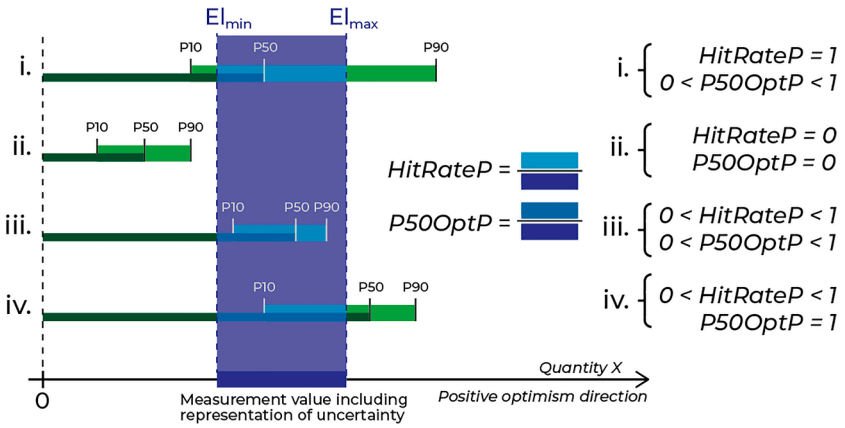

Fig. 3. Examples of calculations of HitRateP and *P50OptP* for some quantity (horizontal axes). Simply speaking, the *HitRateP* is calculated as the overlap interval between the forecast and measurement (*HitFrac80*), divided by the width of the measurement interval ($EI_{Max} - EI_{Min}$). Similarly, *P50OptP* is calculated as the fraction of the measurement interval that is above (or below, for negative optimism direction), the submitted P50 values.

- CO_2 in seals (Quantities 3d, 4d, and 6): CO_2 in seals indicates some migration out of the primary storage formation, thus lower is considered positive: $\xi = 1$.

With the same motivation as for *HitRateP*, i.e., that we do not have a single reference value but instead the empirical interval *EI* of reference values, we define our measure of optimism bias of *P50* forecasts, denoted *P50OptP*, as follows (see again Fig. 3 for an illustration):

$$P50OptP = \xi + (-1)^{\xi}\frac{HitFrac50}{EI_{Max} - EI_{Min}}$$

Here, *HitFrac50* is the overlap of values below *P50* and the measurement, which formally can be expressed as:

$$HitFrac50 = \max(0, \min(EI_{Max}, P50) - EI_{Min})$$

A set of forecasts with a mean *P50OptP* higher than 0.5 indicates a tendency toward over-optimistic forecasts. For the special case where *EI* is zero, we give the value 0.5 if the *P50* forecast also equals zero, and otherwise 0 or 1 depending on the direction of optimism.

The presence of a large spread in the results on scales that are not naturally considered either linear nor logarithmic, render the use of mean values (arithmetic or otherwise) somewhat questionable. We therefore consider the median value as representative of the data, and as a measure of the error of the forecasts, we use the *median absolute error (AE)*, where the absolute error is defined as:

$$AE = |P50 - EI_{Mean}|$$

The median absolute error is a proper evaluation measure of *P50* and gives an expectation of zero error for perfect estimates of *P50*.

2.6. The protocol of the forecasting experiment

The general structure of the forecasting experiment is illustrated in Fig. 1. More precisely, after a common preparation phase for finalizing the benchmark description (Nordbotten et al. 2022), a fully blind phase of three months started, with no direct communication between the various forecasting groups nor the experimental group allowed. All interaction between the forecasting and experimental groups (relating to technical issues, e.g. clarifications, missing data, or measurements) was arbitrated by one of the authors (Flemisch). Information deemed of general interest was broadcasted to all forecasting teams, otherwise bilaterally. The fully blind phase ended when forecasting teams submitted their initial forecasts to the arbitrator. At this point, four groups submitted responses to the quantities forming the basis of this study. An online initial workshop was organized for all forecasting groups, where confidence in the forecasts was measured using questionnaires at three times: First, before any cross-group interaction was allowed, then following pair-wise interaction, and finally after full disclosure of results between all four groups. Thereafter a three-month double-blind synchronization phase followed, where knowledge was shared between the forecasting groups, while the experimental group and their results were still kept double-blind. Community interaction in the synchronization phase allowed the forecasting groups to assimilate the experiences, and also model the results of other groups into their own forecast. In particular, the synchronization phase included two additional, self-assembled workshops, culminating in the submission of final forecasts before an in-person, collective workshop where forecasts and experiments were finally compared. During the synchronization phase the participants identified key differences in their modeling approaches, including the treatment of capillary entry pressure effects and choice of simulation grids. Ultimately, seven groups submitted responses to the quantities forming the basis of this study ahead of the final workshop. Before and during this final workshop, the confidence levels of the participants were again measured using questionnaires. Anonymized responses, both quantitative forecasts and questionnaire responses, are archived as supplementary information.

2.7. Hypotheses

For each of the four research questions, see Section 1, we formulated testable hypotheses. The directions of the hypotheses are based on what typically has been observed in other studies when asking for forecasting intervals and single-point forecasts (Halkjelsvik og Jørgensen, 2012).

- *Hypothesis 1*: The groups' final *PI80*s (80 % confidence forecast intervals) are, on average, too narrow to reflect an 80 % probability of inclusion, i.e., the hit rates tend to be statistically significantly lower than 0.8.
- *Hypothesis 2*: The *P50*-forecasts are, on average, over-optimistic, i.e., there are more than 50 % of the forecasts on the optimistic (positive for CO_2 storage) side of the empirical values.
- *Hypothesis 3*: The calibration of the *PI80s*, defined as the correspondence between hit rate and confidence level, improves with group interaction and improvement-oriented work on the models.
- *Hypothesis 4a*: The groups' forecasts of proxy level hit rates (*HitRateProxys*) are too high to reflect the actual hit rates.
- *Hypothesis 4b*: The research groups are more realistic about the other groups' proxy level hit rates (*HitRateProxys*) than their own.

3. Results

To review the results, we first present the baseline hit rates for the

study. These are purely based on the submitted forecasts, as reported in depth in Flemisch (2024). From this context, we discuss each of the hypotheses posed in the Study Design section.

3.1. Baseline hit rates

Table 1 presents the baseline hit rates per quantity. The *initial hit rate* and *bias* are the hit rate and optimism bias of the initial forecasts in the fully blind phase *before* the interaction with other groups, i.e., before the first workshop (see Fig. 1). At this stage, we had forecast intervals from four groups only. The *final hit rate* and *bias* refer to the hit rate and bias of the final forecasts after the interaction between the groups and more time spent on improving the forecast models, i.e., just before the final workshop. Here, seven groups submitted forecast intervals. As described in the previous section, well-calibrated *PI80* intervals would give an 80 % hit rate (*HitRateP*). Similarly, unbiased P50 forecasts would give a 50 % optimism proportion (*P50OptP*), while values higher than 50 % would suggest a tendency towards over-optimistic forecasts.

The achieved hit rates in Table 1 demonstrate that the forecasted intervals, the *PI80*s, were not well-calibrated. Instead, the hit rates were, for nearly all quantities, lower than the normatively correct 80 %. The tendency to provide too narrow intervals was there initially and prevailed in the final forecasts, although an improvement can be observed with a mean hit rate increase from 26 % to 42 %.

When analyzing the bias of the *P50* forecasts, considering the technology-optimistic direction described earlier, we observe that the forecasts were *not* systematically biased toward optimistic estimates. If anything, they were slightly more likely to provide technology-pessimistic than technology-optimistic forecasts of the stated quantities.

We now explore the four hypotheses stated in the conclusion of Section 2 in more detail.

3.2. Overconfidence (Hypothesis 1)

To test our *Hypothesis 1*, i.e., that the groups will tend to have too narrow forecast intervals, we repeated the analysis of the hit rates, now on a group level for the final forecast intervals. The results are shown in Fig. 4. As before, of the seven groups (denoted A-G) that gave the final forecasts, only four had valid responses in the first submission.

As can be seen, the actual hit rates are much lower than the normatively correct 80 %, with means of the initial and final *PI80*s of 26 % and 42 %, respectively. A one-sided *t*-test[5] on observing a mean hit rate of 21 %, given that the actual mean hit rate is 80 %, gives a p-value < 0.001, i.e., the observed mean value is statistically significantly lower than 80 %. We consequently find support for our Hypothesis 1, i.e., support for a bias towards too narrow forecast intervals (overconfidence).

The proportions of *PI80* partially overlapping with the empirical intervals (*EI*s) were low, with 12 % overlapping intervals in the initial and 14 % in the final forecasts, i.e., the majority of empirical intervals were either fully inside or fully outside the forecasted intervals. Thus, despite the presence of measurement uncertainty in the ground truth, this does not significantly impact the results presented in Fig. 4.

To get an impression of to *what degree* the groups' final *PI80s* were too narrow, we added an analysis with the following two steps, where we:

1. Increased the width of the *PI80* intervals so that the *HitRateP* reached 80 %, for the quantities that did not already have a *HitRateP* of 80 % or higher. The altered *HitRateP* was achieved by the following procedure: For all the original P10 and P90 forecasts within the same quantity, adding a value *c* to the original P90 forecasts and subtracting the value min(*c,P10)* from the P10, to avoid negative adjusted P10 values. The value c was chosen so that the new HitRateP for the quantity was approximately 80 %. Each of the *PI80* intervals were consequently transformed into the intervals [min(0, *P10 – c), P90 + c)]*.
2. Calculated the relative increase in the width of the forecast interval as *((P90 + c) – min(0,P10 - c))/(P90 – P10)*. The forecast intervals of the (two) quantities where the PI80 intervals already had a *HitRateP* of 80 % were given a relative increase of 0.

This analysis shows that the median increase in the width of the *PI80*s, across the quantities needed to achieve a *HitRateP* of 80 %, is 61 %, i.e., that the intervals needed to be 1.6 times wider to achieve the normatively correct hit rate. However, we recognize that there is significant impact of algorithmic choices associated with "widening" forecast intervals, due to both non-linear scales and physical constraints (as an example, mass of CO_2 for any quantity must both be positive and no more than the total injected mass of CO_2 in the system). As such, we ascribe low confidence to the quantitative assessment ("1.6 times wider"), but nevertheless believe it is appropriate to qualitatively state that for the forecast interval to be well-calibrated, a substantial increase in width of forecast intervals would be needed.

3.3. Optimism bias (Hypothesis 2)

For the purpose of testing *Hypothesis 2*, i.e., that the groups will tend to have over-optimistic forecasts, we did a group-wise analysis of forecast bias, see Table 2.

A one-tailed *t*-test of data in Table 2, assuming that the mean final P50OptP is 50 % or higher, gives a p-value of 0.99, i.e., no support for an optimism bias on the final forecasts. Indeed, of the 13 forecasted quantities, only four had forecasts that tended to be on the optimistic side. On the contrary, the data suggests that there was a tendency towards pessimism (p-value of 0.01) for the final *P50* forecasts, and this might be an interesting hypothesis to consider in future studies.

3.4. Effect of group interaction and model improvement (Hypothesis 3)

To test *Hypothesis 3*, i.e., that the group interaction and model improvement-oriented work between the two workshops will improve the hit rates of the forecast intervals, we did an analysis per quantity, where we compared the initial and the final forecast intervals (*PI80*). The results are displayed in Fig. 5, both for comparable groups (the same four groups providing both the initial and the final *PI80*s) and for all groups with final *PI80*s.

As can be seen, the calibration improved, as evaluated by that the hit rates are closer to the normatively correct 80 % for the final forecasts. This improvement is close to being statistically significant. A paired *t*-test of the difference gives a p-value of 0.06.

An improvement in calibration from the first to the last submission can be caused either by wider *PI80*s (better awareness of the forecasting uncertainty) and/or by *PI80*s that are better centered around the ground truth values, i.e., improvement of the model to give less forecast error of the *P50* forecasts.

We found that the improvement in the calibration of the forecast intervals over time was a consequence of the improved accuracy of the models, with a median absolute error of the *P50* forecasts reduced by 72 % from the initial to the final submission. The forecast intervals were, on the other hand, not becoming more realistic. Instead, there was a tendency towards decreasing the intervals' width, with the median *PI80* width *reduced* by 61 % from the initial to the final forecasts. The analyses of the change in forecast error and forecast interval width are provided in Appendix A.

These results give support to Hypothesis 3, but does not support that improvement with group interaction is caused by group interaction leading to awareness of the tendency towards too narrow forecast

[5] Due to only four groups with valid interval forecasts in the initial stage, we do the statistical analysis only on the final forecasts.

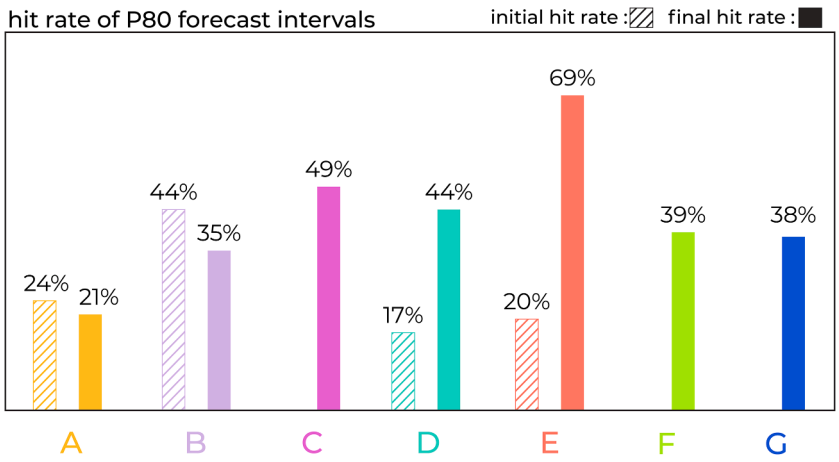

Fig. 4. Hit rates of the first and the final P80 forecast intervals.

Table 2
Evaluation of the bias of the *P50*-values, per group.

Groups	Initial P50OptP	Final P50OptP
A	22%	31%
B	46%	26%
C	–	54%
D	39%	35%
E	65%	46%
F	–	32%
G	–	32%
Mean	**43%**	**37%**

intervals. It is instead likely to be a result of the improved accuracy of the models in response to feedback from the group interaction.

3.5. The assessment of own and others' forecasting ability (Hypotheses 4a and 4b)

To test *Hypothesis 4a*, i.e., overconfidence in the hit rates of own forecasting intervals, and *Hypothesis 4b*, i.e., more realism in the assessment of the forecast intervals of other research groups than their own, we used the responses from the proxy-level self- and other-assessment questionnaires. The self-assessments were, as described earlier, recorded at various stages of the first group workshop and at the final workshop. Group members were also asked to individually assess the forecasts of *other* groups during the final workshop, which we refer to as *external assessment*.

The results are shown in Fig. 6, in terms of the mean of the groups' assessments of hit rate (on the proxy level) relative to their own and the other groups' PI80s. The pairwise disclosing was a process where the groups discussed their results in pairs, while the disclosing of all forecasts, which happened after the pairwise interactions, was a presentation of all groups' forecasts.

3.6. The self-assessment

The initial self-assessed (forecasted) hit rate (55 %) is statistically significantly higher than the mean actual hit rate at that stage (17 %). A one-sample *t*-test comparing the forecasted with the actual hit rate gives a p-value of 0.02. Although the mean forecasted hit rate is too high for

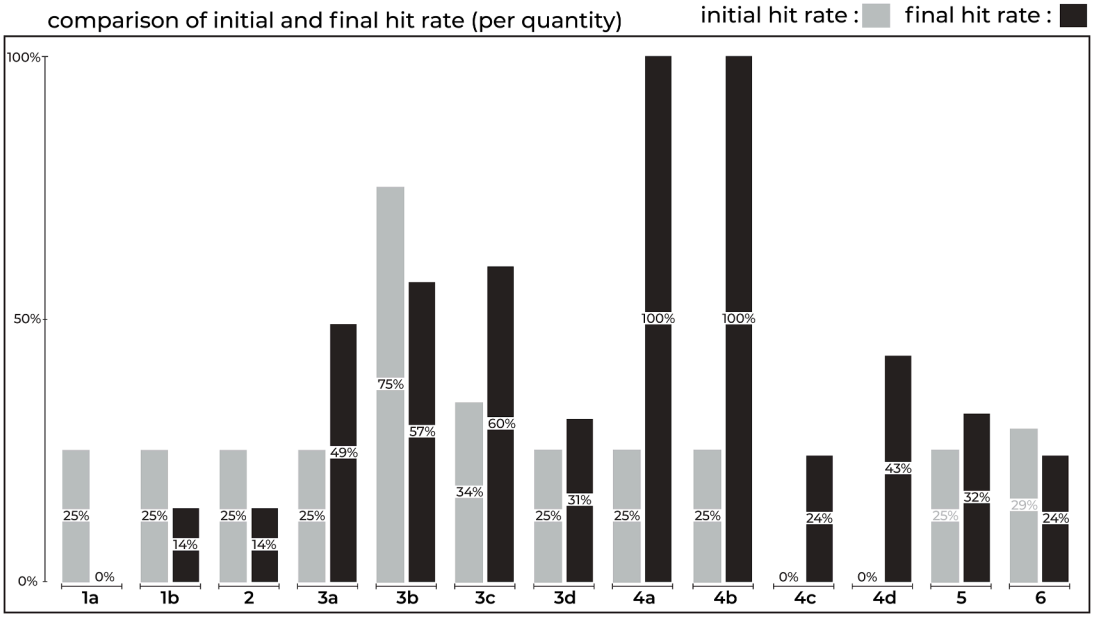

Fig. 5. Change in hit rates from first to final forecasts. The mean initial hit rate across all quantities was 26%, while the mean final hit rate was 42%. Limiting the analysis to only the groups that submitted both initial and final forecasts, the mean final hit rate was 43%.

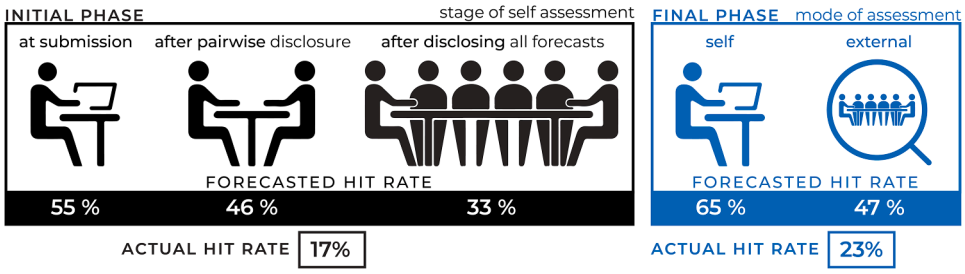

Fig. 6. The mean of the groups' assessments of proxy level hit rate of their own and the other groups' P10-P90 prediction interval (*HitRateProxy)*.

the second and third assessments, these were not statistically significantly different from 17 % ($p = 0.52$ and $p = 0.16$).[6] This may partly be due to fewer responses regarding these two assessments compared to the initial self-assessment, as the difference (effect size) is substantial even after the pairwise and full disclosure of the other groups' forecasts. The final assessments of hit rates were also statistically significantly too high, with a p-value of 0.002, to reflect the actual hit rate of 23 %.

As reported earlier, the groups improved the forecast accuracy of their models between the first and the final workshop. A corresponding belief in improved forecasting accuracy may have contributed to the strong increase in the mean self-assessed hit rate of the proxies from 33 %, provided as the last self-assessment at the initail workshop, to 65 %, provided as the final self-assessment before the final workshop. The mean actual hit rate on the proxy level, in spite of improved models, was, however, just 23 %.

In total, the self-assessed (forecasted) proxy level hit rates show a substantial overconfidence in own ability to give forecast intervals. The results give support for our Hypothesis 4a, that their assessed hit rates, on the proxy level, are too high to reflect the actual hit rates.

We remark that for the data collected in this study, no groups ever scored a partial hit rate (values other than 0 or 1) at the proxy level, thus the presence of measurement uncertainty does not impact the results of the calculated actual hit rates on the proxy level.

As can be seen in Fig. 6, the repeated self-assessments, following the pairwise and total disclosure of the groups' results, during the initial workshop reduced the confidence in one's own forecasts. This may be attributed to the realization by some groups that their own forecasts contained critical flaws, an assertion supported by free-text responses in the questionnaires, or just by seeing that their own forecasts diverged from other groups' forecasts.

The final self-assessment shows that the groups, on average, would have a hit rate of 65 %. This value is not far from the initial self-assessment of 55 %. As argued above, it is likely that the groups' improvement of the models and their corresponding belief in improved accuracy of the forecasts led to the observed decrease in forecast interval widths. This decrease in interval width, combined with an expectation of increased forecasting accuracy seems to, on a self-assessment level, balance such that the participants had an essentially similar expectation of how well their proxy level forecast intervals would match the ground truth.

3.7. The assessment of the other groups

The group members assessed, on average, that the other groups would perform worse than themselves, i.e., they forecasted a hit rate of 65 % of their own and a hit rate of 47 % for the other groups. Both are higher than the actual mean hit rate of 23 %, and all groups had higher self-assessed hit rates than actual hit rates. The difference between the self-assessment and the other assessment is close to being statistically significant, with a p-value of 0.06. The data consequently provides some support for our Hypothesis 4b, i.e., the research groups were more realistic about the calibration of the forecasts of other research groups than their own.

4. Discussion

Our results present answers to our research questions stated in the introduction, and can be summarized as follows:

1) The actual hit rate of the *PI80*s (80 % confidence forecast intervals) of the participating groups was substantially lower than the stated confidence level of 80 %, i.e., the forecast intervals were too narrow, suggesting overconfidence in own forecasts.
2) We found no evidence of systematic bias toward technology optimism in the *P50* forecasts, but rather towards technology pessimistic forecasts.
3) When given a chance to provide updated final forecasts after initial forecasts and a period of community interaction, the forecasts became more accurate, but the width of the forecast intervals did not increase as would be needed for the *PI80*s to correctly capture the physical ground truth. Instead, the median width of the forecasting intervals decreased. Overall, the overconfidence in their forecasts was not significantly mitigated by group interaction.
4) When holistically assessing their own forecast intervals, in terms of expected hit rate on the proxies, the participants provided less confident (and probably more realistic) assessments but still had too high confidence in the forecast intervals. A similar, but weaker, tendency toward too high confidence in the forecast intervals was found when the groups assessed the other groups' forecast intervals after all groups had submitted their final forecasts.

Seen as a whole, the results show a clear tendency of overconfidence among the participants, both with respect to their own ability to provide realistic forecasts, as well as on behalf of the community. While we limit our discussion to the CO_2 storage community as represented in this study, we emphasize that these trends are consistent with expert forecasts in a broad range of disciplines (see Savage et al., 2021, for a recent discussion).

Our survey questions did not ask directly for a justification of how the various groups arrived at their forecast intervals. In an attempt to better understand why the forecast intervals were too narrow, we therefore conceptualize the most common approach to developing a forecast for a complex system in terms of the following four steps (Smith, 2013): First, expert judgment is used to define the scope of the study, primarily in terms of physical processes to consider, computer code to apply, and what physical and numerical parameters to emphasize in a sensitivity study. Second, an uncertainty quantification (UQ) study is conducted using an ensemble of computer simulations following the

[6] The analysis of the forecasted and the actual hit rates are based on the four submitting groups for the first workshop and all seven groups for the final workshop. The results do, however, not change much if we only include the initial four groups in the final workshop. This gives a final self-assessment of a mean forecasted hit rate of 63% instead of 65% and a final mean actual hit rate of 25% instead of 23%.

scope identified in the first step. Third, the results of the computer simulations are assessed statistically to provide a set of forecasts. In this final step, expert knowledge can again be applied to compensate for known deficiencies introduced in the first three steps. Common examples of such deficiencies are: a too limited scope of study; poorly defined prior UQ distributions; limited accuracy of computer simulation tools; and/or the possibility of user errors.

Based on the responses from the participating groups, only a single group deemed their computer simulations completely inadequate, and systematically augmented the UQ with expert knowledge in the third step above (identified as Group F in Fig. 2, see Section 2). The remaining groups followed for most proxies[7] an approach where expert knowledge was applied only in the first step, and the UQ results in the second step were reported using elementary statistics (P10, P50 and P90 values of the simulation ensemble). We will refer to this approach, where no expert knowledge is used in the third step, as an "algorithmic uncertainty quantification" (AUQ). The groups persisted in using an AUQ, even when confronted with substantially deviating forecasts from the other groups during the workshops. Indeed, despite the participants having full access to initial forecasts from all other groups during the second half of the study, groups did not indicate that they had integrated deviating forecasts into their own final forecast. We therefore observe that, within the CO_2 storage community, there is a tendency to provide too much confidence in AUQ based on an ensemble of computer simulations. As a corollary to this, there is hesitation to use external information to modify the forecasts provided by an AUQ.

Interestingly, several of the participations were aware that their AUQ approach was too limited to provide an accurate P10-P90 forecast interval. This is evidenced by free-text responses to the questionnaire used to collect the holistic self-evaluation, where one participant states that "*There are large deviations in the numerical results for most quantities, suggesting that they are very sensitive to details in both geometry and how the physics are represented. I assume that the experimental results may therefore be highly influenced by small heterogeneities/leakage paths that we do not capture in the simulations. So, I stay rather unconfident except for the pressure results maybe.*" Indeed, one participant questions whether such systems may be almost impossible to forecast: "*It is quite eye-opening how different the results are from different groups. Just slightly different model parameters and how interfaces are handled lead to very different results. This brings into question when this can really be modeled at the field scale. The sparse data and uncertainty quantification become very important since these predictions are very uncertain even with this well-characterized experiment.*"

Our interpretation is that in this context of a complex geophysical system, coupling between multiple nonlinear processes may lead to large variations in model output within only small variations in model input – indeed the hallmark of ill-posed problems. This interpretation is being probed in a separate ongoing comparison project (Nordbotten et al., 2024b). Our study has shown that while the carbon storage community is acutely aware of this challenge, there nevertheless remains a strong tendency to understate the uncertainty of storage forecasts. To address this issue, we make the following observations and recommendations:

1) Most groups were averse to augmenting the AUQ with human expert knowledge. In view of the free-text questionnaire responses, we expect that explicitly requesting a post-AUQ human assessment, in addition to the AUQ, may provide broader forecast intervals.
2) While this work contains community interaction, which clearly improved the accuracy of the provided P50 forecasts, the participants did not assimilate the model results of other groups into their own P10-P90 forecast intervals. An additional group stage to develop a "consensus community forecast", including that of forecasting intervals, may be a mechanism to increase cross-team engagement and learning.
3) The substantial time and effort required to construct an efficient computational model may create a bias toward overconfidence in the model outcomes which, in turn, may undermine a full appreciation of the limitations and weaknesses of computer simulations as representations of complex physical systems. A discussion of this risk at the start of the forecasting study may be beneficial.

While our study has focused exclusively on forecasts of geologic carbon storage, the lessons learned are likely applicable to other subsurface technologies such as nuclear waste disposal (Jing et al., 1995; Birkholzer et al., 2019; Cvetkovic et al., 2004), underground gas storage (Conley et al., 2016) and geothermal energy extraction (Deichmann and Giardini, 2009). Like carbon storage, these technologies are faced with risks associated with fluid migration via geologic features like caprocks, faults and fractures or man-made features like existing wells (Kang et al., 2014). Also, like carbon storage, they incur a risk of triggering earthquakes—a risk that has received increased societal scrutiny (National Research Council, 2013; Grigoli et al., 2017). We suggest that increasing the reliability of UQ across subsurface technologies likely will require the participation of multiple groups, cross-examination of the forecast intervals from disparate models, and formally incorporating expert knowledge into the algebraic, model-based UQ outcomes.

Nomenclature

Some words used in this text have different connotations in different research fields. We therefore clarify our use of key terms below:

- Forecast (as opposed to prediction) emphasizes the reliance on numerical simulation tools.
- Forecast Accuracy: Measurement of the deviation between the P50 forecast and the ground truth.
- Forecast bias: Indicates any systematic bias in P50 forecast relative to the ground truth.
- Forecast interval: The numerical values bounded below by P10 (forecasted 10 % probability of not being exceeded) and above by P90 (forecasted 90 % probability of not being exceeded).
- Confidence level: The stated or implied expectation (probability) that the forecast interval includes the ground truth value.
- Hit rate: The actual rate of inclusion of the ground truth within the forecast interval.
- Calibration: The extent to which the hit rate matches the stated confidence level of the forecast interval. Forecast intervals as used in this study (P10 to P90) are well-calibrated if the hit rate is 80 %.
- Overconfidence: The situation where the hit rate is lower than the confidence level (too narrow forecast intervals).
- Proxies: Aspects of the ground truth experiment that are expected to correlate to measures of success for a carbon storage operation (six in total)
- Quantities: Actual numerical quantities involved in forecasting the proxies (thirteen in total)

Classification

Physical Sciences (Engineering) and Social Sciences (Psychological and Cognitive Sciences)

CRediT authorship contribution statement

Jan M. Nordbotten: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Martin Fernø:** Writing – review & editing, Writing – original draft,

[7] As presented in Flemisch et al., (2023), the reported quantities for most proxies can be attributed directly to respective simulation results. Proxy 5 (which is not part of standard numerical simulations), is a notable exception which is discussed in detail there.

Visualization, Methodology, Data curation, Conceptualization. **Bernd Flemisch:** Writing – review & editing, Writing – original draft, Software, Formal analysis. **Ruben Juanes:** Writing – review & editing, Writing – original draft, Conceptualization. **Magne Jørgensen:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The current study has not received any industry funding and is conducted to the best of the author's ability in the absence of conflicts of interest. This said, all authors except for MJ have conducted extensive research related to carbon storage in both the past and present, which at times has been partially funded by industry.

Data availability

Data will be made available on request.

Acknowledgments

B. Flemisch thanks the German Research Foundation (DFG) for supporting this work by funding SFB 1313, Project Number 327154368.

Appendix A. Analysis of the reason for improved calibration over time

The calibration of the forecasts improved (hit rates closer to the normatively correct 80 %) from the first to the final forecasts. This analysis examines whether this improvement is likely to have been caused by a wider *PI80*s (better awareness of the forecasting uncertainty) or *PI80*s better centered around the empirical values, i.e., models with less forecast error. The latter would be indicated by less error of the P50 values.

Table A1 shows the interval width of the PI80 per quantity, with percentage change. A positive (negative) value suggests an interval increase (decrease) from the first to the final forecast. As reported in the main part of this article, the initial interval forecasts were much too narrow to reflect the true uncertainty. To give support for the first explanation, i.e., better awareness of the forecasting uncertainty from the initial to the final forecasts, there should be a substantial increase in interval widths.

Table A1 shows that seven of the thirteen interval widths decreased over time, and the middle observation (median) is negative (61 % decrease in interval width). This means that the increase in hit rate can consequently hardly be explained by wider forecast intervals, as most of the forecast intervals width decreased substantially. The relative change is not statistically significantly different from zero with a p-value 0.31.

The alternative explanation is that the error of the P50 forecast decreased over time and, for this reason, improved the hit rates. The P50 forecast error, measured as the median absolute error, is displayed in Table A2.

With the exception of 3a and 3b, all P50-values improved, some of them very much, with respect to forecast error from the initial to the final submission. A paired *t*-test of the relative change of MdAE gives that the values are statistically significantly lower than 0 (p-value < 0.001 and t-value of -4.56).

This suggests that the improvement in the calibration of the forecasting intervals over time was a result of improved forecast accuracy rather than awareness of what would be a well-calibrated 80 % forecast interval and the need to increase the interval width to reflect an 80 % inclusion rate.

In total, the data gives support to an improvement in the calibration of the forecast intervals over time, but only as a consequence of improved accuracy.

Table A1
Mean forecast interval width of PI80 per quantity.

Quantity	Median interval width of initial forecast	Median interval width of final forecast	Relative change of median interval width* from first to final forecast
1a	2.32E+03	1.26E+02	−95%
1b	2.13E+03	1.03E+02	−95%
2	1.16E+04	5.00E+02	−96%
3a	3.57E-04	1.19E-03	233%
3b	3.72E-06	8.52E-05	2187%
3c	1.29E-03	1.65E-03	28%
3d	2.61E-04	3.88E-04	48%
4a	7.81E-04	0.00E+00	−100%
4b	3.11E-04	0.00E+00	−100%
4c	8.26E-04	3.25E-04	−61%
4d	7.86E-04	3.37E-06	−100%
5	4.32E+03	6.75E+03	56%
6	3.33E-05	1.90E-04	470%
Median			**−61%**

* The relative interval width change is calculated as (Median interval width of final submission – Median interval width of first submission)/Median interval width of first submission.

Table A2
Median absolute forecast error (MdAE) over time per quantity.

Quantity	MdAE of first Submission	MdAE of final submission	Relative change of MdAE* from first to final submission
1a	5.96E+03	4.54E+02	−92%
1b	3.07E+03	5.04E+02	−84%
2	3.13E+04	3.50E+03	−89%
3a	6.07E-04	9.19E-04	52%
3b	0.00E+00	9.15E-05	–
3c	2.35E-03	9.41E-04	−60%
3d	2.96E-04	1.53E-04	−48%

(*continued on next page*)

Table A2 (*continued*)

Quantity	MdAE of first Submission	MdAE of final submission	Relative change of MdAE* from first to final submission
4a	6.04E-04	0.00E+00	−100%
4b	2.65E-04	0.00E+00	−100%
4c	1.76E-03	7.06E-04	−60%
4d	5.67E-04	1.80E-07	−100%
5	1.38E+04	8.75E+03	−36%
6	2.47E-04	2.28E-04	−8%
Median			**−72%**

* The relative change of MdAE is calculated as (MdAE of final submission – MdAE of first submission)/MdAE of first submission. Quantity 3b had MdAE of zero and the relative change could not be calculated.

References

Bickel, J.E., Bratvold, R.B., 2008. From uncertainty quantification to decision making in the oil and gas industry. Energy Explor. Exploit. 26 (5), 311–325. https://doi.org/10.1260/014459808787945344.

Birkholzer, J.T., Tsang, C.F., Bond, A.E., Hudson, J.A., Jing, L., Stephansson, O., 2019. 25 years of DECOVALEX-Scientific advances and lessons learned from an international research collaboration in coupled subsurface processes. Int. J. Rock Mech. Mining Sci. 122, 103995.

Cesarini, D., Sandewall, Ö., Johannesson, M., 2006. Confidence interval estimation tasks and the economics of overconfidence. J. Econ. Behav. Organ. 61 (3), 453–470. https://doi.org/10.1016/j.jebo.2005.02.015.

Conley, S., Franco, G., Faloona, I., Blake, D.R., Peischl, J., Ryerson, T.B., 2016. Methane emissions from the 2015 Aliso Canyon blowout in Los Angeles, CA. Science 351 (6279), 1317–1320.

Cooke, R.M., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press.

Cvetkovic, V., Painter, S., Outters, N., Selroos, J.O., 2004. Stochastic simulation of radionuclide migration in discretely fractured rock near the Äspö Hard Rock Laboratory. Water Resour. Res. 40 (2), W02404.

Deichmann, N., Giardini, D., 2009. Earthquakes induced by the stimulation of an enhanced geothermal system below Basel (Switzerland). Seismol. Res. Lett. 80 (5), 784–798.

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., 2024. R.oom-scale CO_2 injections in a physical reservoir model with faults. Transp. Porous Med. https://doi.org/10.1007/s11242-023-02013-4.

Ferson, S., Joslyn, C.A., Helton, J.C., Oberkampf, W.L., Sentz, K., 2004. Summary from the epistemic uncertainty workshop: consensus amid diversity. Reliab. Eng. Syst. Saf. 85 (1–3), 355–369.

Flemisch, B., Nordbotten, J.M., Fernø, M.A., Juanes, R., Both, J.W., Class, H., Delshad, M., Doster, F., Ennis-King, J., Franc, J., Geiger, S., Gläser, D., Green, C., Gunning, J., Hajibeygi, H., Jackson, S.J., Jammoul, M., Karra, S., Li, J., Matthäi, S.K., Miller, T., Shao, Q., Spurin, C., Stauffer, P., Tchelepi, H., Tian, X., Viswanathan, H., Voskov, D., Wang, Y., Wapperom, M., Wheeler, M.F., Wilkins, A., Youssef, A.A., Zhang, Z., 2024. The FluidFlower Validation Benchmark Study for the Storage of CO_2. Transp. Porous Med. https://doi.org/10.1007/s11242-023-01977-7.

Floris, F.J.T., Bush, M.D., Cuypers, M., Roggero, F., Syversveen, A.-R., 2001. Methods for quantifying the uncertainty of production forecasts: a comparative study. Petroleum Geosci. 7 (S), S87–S96. https://doi.org/10.1144/petgeo.7.S.S87.

Furre, A.K., Eiken, O., Alnes, H., Vevatne, J.N., Kiær, A.F., 2017. 20 years of monitoring CO_2-injection at Sleipner. Energy Procedia 114, 3916–3926. https://doi.org/10.1016/j.egypro.2017.03.1552.

Glaser, M., Langer, T., Weber, M., 2013. True overconfidence in interval estimates: evidence based on a new measure of miscalibration. J. Behav. Decis. Mak. 26 (5), 405–417. https://doi.org/10.1002/bdm.1784.

Grigoli, F., Cesca, S., Priolo, E., Rinaldi, A.P., Clinton, J.F., Stabile, T.A., Dost, B., Fernandez, M.G., Wiemer, S., Dahm, T., 2017. Current challenges in monitoring, discrimination, and management of induced seismicity related to underground industrial activities: a European perspective. Rev. Geophys. 55 (2), 310–340.

Halkjelsvik, T., Jørgensen, M., 2012. From origami to software development: a review of studies on judgment-based predictions of performance time. Psychol. Bull. 138 (2), 238–271. https://doi.org/10.1037/a0027039.

Illangasekare, T.H., Ramsey Jr, J.L., Jensen, K.H., Butts, M.B., 1995. Experimental study of movement and distribution of dense organic contaminants in heterogeneous aquifers. J. Contam. Hydrol. 20 (1–2), 1–25.

IPCC, 2005. In: Metz, B., et al. (Eds.), Special Report on Carbon Dioxide Capture and Storage. Cambridge University Press.

IPCC, 2022. In: Shukla, P.R., Skea, J., Slade, R., Al Khourdajie, A., van Diemen, R., McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., Belkacemi, M., Hasija, A., Lisboa, G., Luz, S., Malley, J. (Eds.), Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel On Climate Change. Cambridge University Press. https://doi.org/10.1017/9781009157926.

Jing, L., Tsang, C.F., Stephansson, O., 1995. DECOVALEX—an international co-operative research project on mathematical models of coupled THM processes for safety analysis of radioactive waste repositories. Int. J. Rock Mech. Mining Sci. Geomechanics Abstracts 32 (5), 389–398.

Kang, M., Kanno, C.M., Reid, M.C., Zhang, X., Mauzerall, D.L., Celia, M.A., Chen, Y., Onstott, T.C., 2014. Direct measurements of methane emissions from abandoned oil and gas wells in Pennsylvania. Proc. Natl. Acad. Sci. 111 (51), 18173–18177.

Klas, M., Trendowicz, A., Ishigai, Y., Nakao, H., 2011. Handling estimation uncertainty with bootstrapping: empirical evaluation in the context of hybrid prediction methods. In: 2011 IEEE International Symposium on Empirical Software Engineering and Measurement. IEEE, pp. 245–254. https://doi.org/10.1109/ESEM.2011.40.

Kovscek, A.R., Nordbotten, J.M., Fernø, M.A., 2024. Scaling up FluidFlower results for carbon dioxide storage in geological media. Transp. Porous Med. https://doi.org/10.1007/s11242-023-02046-9.

Morgan, M.G., Henrion, M., 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, Cambridge, p. 332.

Murphy, A.H., 1998. The early history of probability forecasts: some extensions and clarifications. Weather. Forecast. 13 (1), 5–15.

National Research Council, 2013. Induced Seismicity Potential in Energy Technologies. National Academies Press.

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

Nordbotten, J.M., Benali, B.., Both, J.W., Brattekås, B., Storvik, E., Fernø, M.A., 2024a. DarSIA: an open-source Python toolbox for two-scale image processing of dynamics in porous media. Transp. Porous Media. https://doi.org/10.1007/s11242-023-02000-9.

Nordbotten, J.M., Ferno, M.A., Flemisch, B., Kovscek, A.R., Lie, K.A., 2024b. The 11th society of petroleum engineers comparative solution project: problem definition. SPE J. 1–18.

Qian, E., Peherstorfer, B., O'Malley, D., Vesselinov, V.V., Willcox, K., 2018. Multifidelity Monte Carlo estimation of variance and sensitivity indices. SIAM/ASa J. Uncertain. Quantif. 6 (2), 683–706.

Savage, T., Davis, A., Fischhoff, B., Morgan, M.G., 2021. A strategy to improve expert technology forecasts. Proc. Natl. Acad. Sci. 118 (21), e2021558118.

Smith, R.C., 2013. Uncertainty quantification: Theory, implementation, and Applications, 12. SIAM.

Soll, J.B., Klayman, J., 2004. Overconfidence in interval estimates. J. Exp. Psychol. 30 (2), 299. https://doi.org/10.1037/0278-7393.30.2.299.

Tavassoli, Z., Carter, J.N., King, P.R., 2004. Errors in history matching. SPE J. 9 (4), 352–361. https://doi.org/10.2118/86883-PA.

Trevisan, L., Pini, R., Cihan, A., Birkholzer, J.T., Zhou, Q., Gonzalez-Nicolas, A., Illangasekare, T.H., 2017. Imaging and quantification of spreading and trapping of carbon dioxide in saline aquifers using meter-scale laboratory experiments. Water Resour. Res. 53 (1), 485–502. https://doi.org/10.1002/2016WR019749.

Trupp, M., Ryan, S., Barranco Mendoza, I., Leon, D., Scoby-Smith, L., 2021. Developing the world's largest CO2 injection system–a history of the gorgon carbon dioxide injection system. In: Proceedings of the 15th Greenhouse Gas Control Technologies Conference, pp. 15–18.

White, D., 2009. Monitoring CO_2 storage during EOR at the Weyburn-Midale field. Leading Edge 28 (7), 838–842. https://doi.org/10.1190/1.3177022.

Zhang, Y., Jackson, C., Krevor, S., 2022. An estimate of the amount of geological CO_2 storage over the period of 1996–2020. Environ. Sci. Technol. Lett. 9 (8), 693–698. https://doi.org/10.1021/acs.estlett.2c00481.