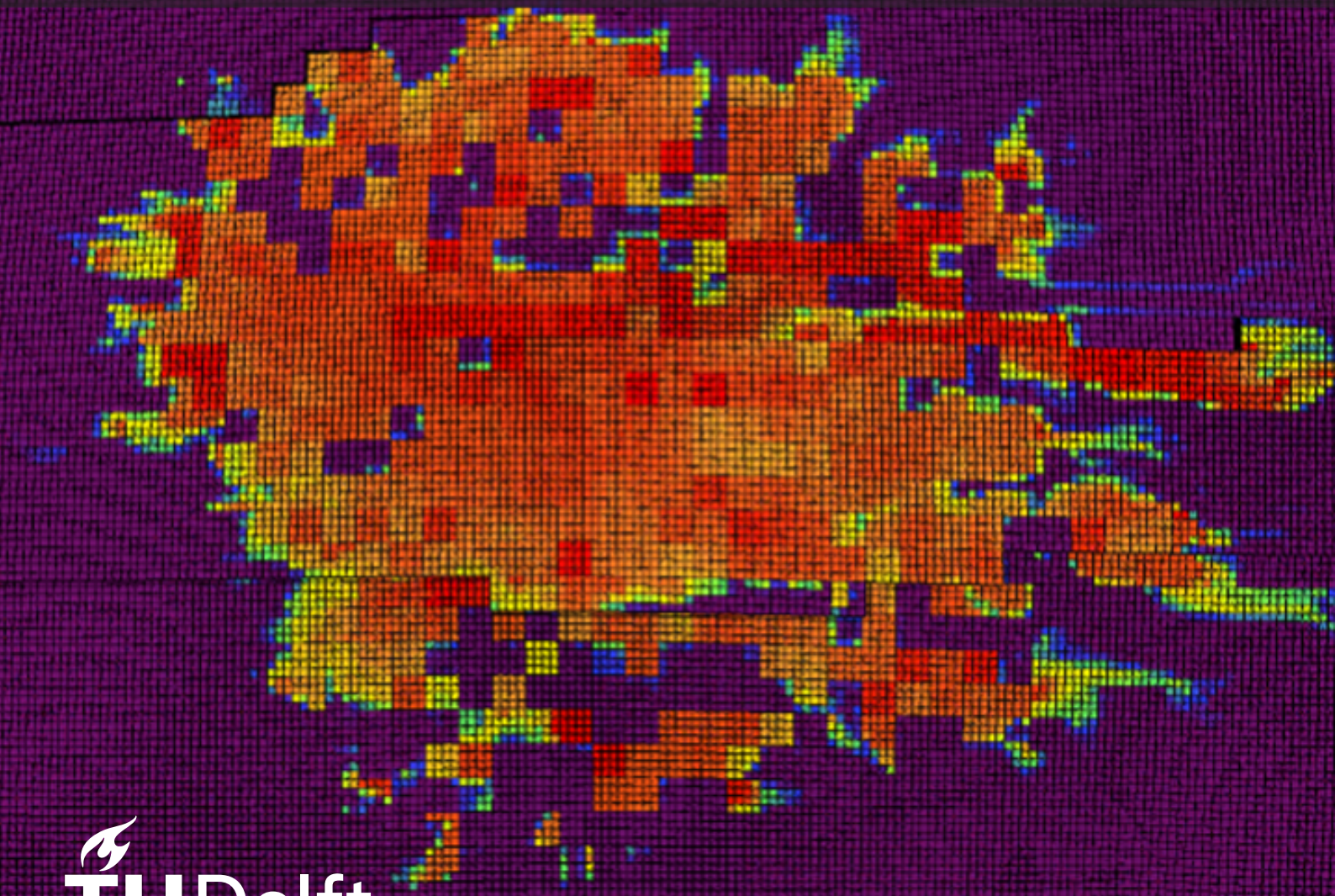
Optimizing Uncertainty Quantification for CO_2 Subsurface Storage:

Model Ranking Using Flow Diagnostics

Maarten de Nooijer



TU Delft

Optimizing Uncertainty Quantification for CO_2 Subsurface Storage:

Model Ranking Using Flow Diagnostics

by

Maarten de Nooijer

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday September 30th, 2025 at 10:45.

Student number:	4942329	
Project duration:	February, 2025 – September, 2025	
Thesis committee:	Prof. dr. S. Geiger	TU Delft, Supervisor
	Dr. A. Daniilidis	TU Delft, Supervisor
	Prof. dr. A.W. Martinius	TU Delft

Cover:	CO_2 saturation field created in StudioSL by Maarten de Nooijer
Style:	TU Delft Report Style, with modifications by Maarten de Nooijer

TU Delft

Preface

Living in a money-driven world, where short-term goals often prevail over long-term outcomes, one can respond in three ways: accept the status quo and move along, fight and blame others for how things are, or take action and try to shape the narrative instead.

A money-driven world has at least one advantage: its objective is clear. If you can offer an alternative that is cheaper and more effective, change can happen remarkably fast. History has shown often enough how adaptable we can be when the right incentives appear.

This thesis[1] is written in that spirit. By helping move CO_2 subsurface storage toward the economically viable side of the spectrum, I hope to make this long-term climate solution competitive in a short-term-focused world.

Specifically, this study focuses on optimizing uncertainty quantification for CO_2 subsurface storage, enabling computationally efficient prediction of reservoir responses under inherent geological uncertainty. The aim is to make uncertainty quantification a practical and cost-effective tool so that decision makers can evaluate and design CO_2 storage projects with greater reliability. In turn, this can lead to more accurate cost forecasts and stronger safety assessments, supporting the responsible and economically viable implementation of the large-scale CO_2 storage so urgently needed to address climate change.

I would like to express my sincere gratitude to Sebastian Geiger for his unconditional support. I have had the pleasure of working with him on several occasions over the past two years, and throughout my master's he has been an incredibly involved and supportive mentor. I deeply appreciate how engaged he has been, both as my thesis supervisor and as a professor who consistently puts his students first. Your guidance, expertise, and insightful feedback were invaluable throughout my studies.

I would also like to thank Alexandros Daniilidis, who was also a member of my bachelor's graduation committee, for his constructive feedback and for the valuable discussions we had during the past months. A special thanks goes to Marco Thiele, Rod Batycky, and Matteo Di Geovanazzo from StreamSim Technologies, who not only provided the streamline-based flow simulator used throughout this study but also offered direct help whenever it was needed. I truly appreciated how quickly you were able to set up meetings and respond to emails, and your input shaped my thinking about the distance-based clustering approach that forms the central thread of this thesis. I would also like to thank Dennis Voskov, who was incredibly supportive whenever I struggled with open-DARTS, and especially Ilshat Saifullin. Without your help, this thesis would never have seen the light of day. I also extend my gratitude to George Hadjisotiriou, Michiel Wapperom, Sajjad Moslehi, Yuan Chen, and all the others from the DARTS team who made time to brainstorm with me and discuss the challenges I faced.

Reaching the end of my thank-you list, I want to thank Gilles Douwes in particular, with whom I have spent almost every day over the past two years, including an unforgettable project in Madrid. You have not only been an incredible colleague but also a true friend, and I will miss working on projects together.

Finally, I am deeply thankful to my family, friends, and especially my girlfriend, who did not always get the attention they deserved during the past few months. Thank you for your patience; I promise to make it up to you.

Or at least try!

Maarten de Nooijer
Delft, September 2025

[1] This thesis made use of ChatGPT to refine language and was used for coding purposes. All scientific content, and conclusions were developed independently by the author.

Summary

Carbon Capture and Storage (CCS) is gaining increasing attention. In 2023, announced 2030 storage capacity rose by about 70%, reaching ~ 615 Mt CO_2 per year [1]. Yet the execution gap remains large: up to 2020, roughly 70% of announced projects did not materialize, owing to high capital costs and uncertain revenue streams [2]. Experience from oil and gas, on which many CCS methods still draw, shows that about 75% of O&G projects fail to deliver planned production levels [3]. Such shortfalls appear particularly problematic for CCS, where narrow margins, long-term liability, contractual obligations to emitters and regulators [4], and public scrutiny demand a more rigorous and transparent treatment of uncertainty. Robust uncertainty quantification (UQ) is therefore essential: to bound plausible storage outcomes, identify the uncertainties that dominate them, and build confidence in the safety and bankability of CCS projects.

Paradoxically, UQ is more expensive in CCS than in oil and gas. It requires simulating large, long-timescale domains with multi-scale heterogeneity, reflecting CCS's shifting force balance (viscous near-well, capillary in the plume wake, gravity in the far field). These physical challenges are compounded by the limited availability of data at prospective CCS sites, such as large aquifers, which are required to deliver the gigatonne-scale volumes needed for meaningful climate impact [5]. As a result, both the number of realizations (needed to span poorly constrained heterogeneities across scales) and the cost per realization (due to longer horizons and larger domains) are substantially higher than in oil and gas, rendering brute-force Monte Carlo impractical. This motivates practical screening workflows that identify the most influential uncertainties and concentrate full-physics runs on representative models, enabling feasible probabilistic assessment of reservoir response.

Building on this motivation, the thesis develops and tests an ensemble-reduction workflow, introduced by Scheidt and Caers [6, 7, 8], that clusters realizations by distances computed on flow diagnostics, targeting preservation of ensemble percentile bounds (P_{90}, P_{50}, P_{10}) for injection rate, maximum plume migration, and plume areal coverage within $\leq 5\ \%$ relative RMSE per percentile. In a 108-member geomodel ensemble incorporating top-surface, fault, fault-transmissibility multiplier, cutoff, and facies-modeling uncertainties, early-time full-physics (FP) injection rates proved highly informative for injection-rate uncertainty quantification. Distances derived from only 10 days of FP simulation (FP-D10, $\approx 1.3\%$ of a 20-year runtime) showed strong Spearman rank correlations with long-term injection behavior ($\rho = 0.93$ and $\rho = 0.85$ for two ensembles) and, when embedded using t-SNE for dimensionality reduction and clustered with k-means, produced stable percentile reconstructions with relative RMSE $\leq 5\%$ for P_{90}, P_{50}, and P_{10} across most tested cluster counts K. A slight conservative bias was observed (underestimation of reconstructed percentiles), consistent with dense clustering of low-performing models under early BHP control and sparser distribution of high-performing models, which is favorable from a risk perspective. In addition, occasional erratic RMSE behavior at specific K values was mitigated by an internal guidance strategy combining the inertia elbow with the Davies–Bouldin index, Silhouette score, and Calinski–Harabasz index. Applied to FP-D10 + t-SNE, this selected compact sets of 9 and 6 representatives for the two ensembles, cutting relative simulation cost to $9.7\ \%$ and $6.9\ \%$ while meeting the $\leq 5\ \%$ criterion (Ensemble 1: $P_{90} = -3\ \%, P_{50} = +5\ \%, P_{10} = -4\ \%$; Ensemble 2: $P_{90} = -1\ \%, P_{50} = -5\ \%, P_{10} = +4\ \%$).

Evidence found in this study indicates that local injectivity dominates injection-rate variability: even 1-day FP rate distances supported credible reconstructions across K, and single-phase streamline diagnostics (injectivity-focused) matched or exceeded two-phase immiscible variants, suggesting that added multiphase complexity in the streamline-based flow simulator can dilute alignment without improving rank fidelity. Beyond injectivity, early FP rates also encoded system-scale effects (e.g., low fault transmissibility), as shown by distance-based generalized sensitivity analysis (dGSA) on cluster assignments, which identified the same influential parameters as variance-based sensitivity on full histories (facies method, cutoff, and fault-transmissibility multiplier in Ensemble 1, and fault transmissibility in Ensemble 2) at only $\sim 1.3\ \%$ of the simulation cost, thus providing both efficiency and a geological

interpretation of cluster structure. A genetic algorithm calibrated on short simulation windows yielded robust results only beyond ~ 100 days and proved less time-efficient than direct early-time distance clustering. A two-stage variant, applying t-SNE and k-means to reduce the ensemble to $\sim 60\%$ before running the GA, improved competitiveness, but the initial ensemble reduction relied on knowledge of true ensemble percentiles that is unavailable in practice, limiting its baseline utility.

For plume metrics, an two-phase immiscible saturation-field diagnostic (11 pressure solves over 20 years) gave the most accurate percentile reconstructions, with a single-phase diagnostic (one pressure solve over 20 years) showing only modest degradation while being far cheaper ($\sim 35\,\mathrm{s}$ vs. $\sim 6\,\mathrm{min}$). Both were strongest for P_{50} and, crucially, the high-end P_{10}, with lower reliability at P_{90}, which is acceptable given the need to treat rare extremes with more care. Notably, cluster representatives chosen for injection-rate UQ via FP-D10 also transferred effectively to plume UQ, reconstructing maximum plume-extent percentiles with $(P_{90}, P_{50}, P_{10}) = (-11\ \%, +4\ \%, -5\ \%)$ for Ensemble 1 and $(-8\ \%, -2\ \%, -5\ \%)$ for Ensemble 2, and plume areal coverage with $(-6\ \%, +9\ \%, -1\ \%)$ and $(+6\ \%, +4\ \%, -6\ \%)$, respectively. This dual use likely reflects a volume signal whereby higher early injection rates correlate with broader plume spread.

Finally, this study confirms that reduced subset CDFs should be weighted by cluster population, a choice that improved accuracy and stability across K. In sum, distance-based clustering enables an order-of-magnitude reduction in ensemble size for interpretational uncertainties while preserving percentile reconstructions within $\sim\leq 5\ \%$ relative RMSE, supporting transparent, defensible, and cost-effective UQ for CCS. The results reinforce the application-agnostic value of the approach and motivate further testing of early-time full-physics simulation diagnostics for long-term injection-rate UQ and pressure-risk assessment under diverse geological and operational settings.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
BHP	Bottom-Hole Pressure
CCS	Carbon Capture and Storage
CDF	Cumulative Distribution Function
CH	Calinski–Harabasz Index (cluster validity metric)
CO	Cut-Off
dGSA	Distance-based Generalized Sensitivity Analysis
D	Days
DB	Davies–Bouldin Index (cluster validity metric)
DR	Dimensionality Reduction
FD	Flow Diagnostic (flow-based distance metric)
FM	Fault Model
FP	Full-Physics
Geomodel	Geological model representing a possible subsurface reservoir realization
IMM	Two-phase Immiscible Streamline
K	Number of clusters or selected models
KPCA	Kernel Principal Component Analysis
MDS	Multidimensional Scaling
Mult	Fault Transmissibility Multiplier
OBJ	Object-based Facies Distribution Modeling
OOIP	Original Oil in Place
OWC	Oil–Water Contact
P_{10}	Optimistic estimate (10% probability of being exceeded, **90th percentile**)
P_{50}	Median estimate (50% probability of being exceeded, **50th percentile**)
P_{90}	Conservative estimate (90% probability of being exceeded, **10th percentile**)
PIX	Pixel-based Facies Distribution Modeling
PRD	Production
PS	Pressure-Solve Index
RMSE	Root Mean Square Error
SAT	Saturation Field
SP	Single-Phase Streamline
t-SNE	t-Distributed Stochastic Neighbor Embedding
TS	Top Surface
UQ	Uncertainty Quantification

1

Introduction

Carbon Capture and Storage (CCS) is gaining increasing attention, with around 45 commercial facilities currently applying carbon capture, utilization, and storage (CCUS) to industrial processes, fuel transformation, and power generation. Momentum has accelerated in recent years, with over 700 projects in various stages of development across the CCUS value chain. In 2023 alone, announced storage capacity for 2030 increased by 70%, bringing the total to around 615 Mt of CO_2 per year. While this progress is encouraging, it still represents only about 60% of the circa 1 Gt per year of CO_2 storage required under the Net Zero Emissions by 2050 (NZE) Scenario [1]. Moreover, meeting this target will require a significant expansion of injection infrastructure, with an estimated 10,000 to 14,000 wells required globally by 2050 [9].

While the recent wave of project announcements is promising, it is also important to acknowledge the historical challenges of CCS deployment. Of the 149 projects proposed to be operational by 2020 (intended to store approximately 130 Mt of CO_2 annually), around 70% failed to materialize. The primary barriers were high capital costs and the absence of reliable revenue streams [2].

In reservoir engineering, and consequently CCS, one of the main challenges during project initiation is addressing the geological heterogeneity of the subsurface, which is inherently uncertain. Heterogeneity can be categorised as structural (for example, faults, folds, fracture intensity), stratigraphic (for example, layering, channels, barriers) [10], and petrophysical (for example, porosity–permeability distributions, anisotropy, relative permeability, capillary entry pressure). Heterogeneity significantly impacts reservoir behavior, including available storage capacity, CO_2–brine displacement, and potential leakage pathways [11], [12]. Therefore, reliable predictions of reservoir behavior require a detailed understanding of the multi-scale heterogeneity of the subsurface. In practice, however, developing such an understanding is hindered not only by the scarcity of high-resolution data (such as seismic data) and the limited spatial coverage of measurements (such as well logs and core samples), but also by modelling and data-interpretation decisions (e.g., property upscaling, facies grouping, and cut-off criteria). These factors collectively contribute to substantial uncertainty in characterising reservoir heterogeneity [13], [14].

With potentially hundreds of uncertain parameters, accurately predicting reservoir response to CO_2 injection using flow simulation models remains a major challenge [6]. Uncertainty quantification (UQ) plays a crucial role in addressing this challenge by mapping how input uncertainty propagates through the model to affect predicted outcomes. This enables better assessment of CCS project feasibility and supports decisions on whether additional data should be acquired to reduce critical uncertainties [15], [16].

Strikingly, in the oil and gas industry (from which CCS currently derives much of its methodological foundation, given its own limited operational track record) it appears to be tolerated that approximately 75% of projects fail to deliver their planned production levels [3]. Such widespread underperformance illustrates the risks of insufficient or overly simplistic treatment of uncertainty. Nevertheless, the industry continues to invest heavily and expand operations, perhaps reflecting the financial resilience and risk

tolerance embedded in traditional hydrocarbon development. In contrast, comparable shortfalls are likely to be far less acceptable in the context of CCS, where narrow profit margins, long-term liability, contractual obligations with emitters and regulators [4], and public perception demand a much more rigorous and transparent treatment of uncertainty. Robust UQ is therefore not optional; it is essential for quantifying the range of possible storage outcomes, identifying the key uncertainties that drive them, and building confidence in the safety and viability of CCS projects.

One way to quantifying this uncertainty is the 'rationalist' approach, which can be viewed as a form of 'traditional' determinism. In this method, a base scenario is defined as the preferred model, with uncertainty incorporated either by applying a percentage factor to the input parameters or model output, or by flanking the base case with separate low and high cases [17, 18]. However, quantifying uncertainty using a single preferred geological scenario with limited variations risks failing to capture the full range of uncertainty in flow response due to the sparse sampling of the parameter space [19], and the best guess is only reliable when the system being described is well ordered and well understood, which in practice is rarely the case [20]. As such, overreliance on the rationalist method may contribute to project underperformance and could expose CCS efforts to the same high failure rates in meeting projected targets as those seen in traditional oil and gas developments.

To better capture and quantify the full range of uncertainty, two more appropriate approaches are the multiple stochastic and multiple deterministic approach. Both involve creating a set of geomodels to span uncertainties in parameters such as porosity, permeability, and structural geology [21]. The multiple stochastic approach generates a large ensemble of equiprobable models using geostatistical simulations. In this method, input parameters are sampled randomly but guided by a starting seed, allowing for the creation of numerous realizations. These realizations can then be combined into a probability distribution, providing a statistical representation of uncertainty. In contrast, the multiple deterministic approach can be viewed as a deterministic, scenario-driven approach, where an number of models is created that are not statistically generated but instead represent discrete deterministic scenarios. This method involves building a smaller number of models compared to the stochastic approach, with each model reflecting a complete real-world outcome based on an explicitly defined reservoir concept. Ultimately, these techniques may converge if multiple stochastic cases are generated from a set of multiple deterministic concepts, resulting in a large ensemble [17].

Consequently, after creating the ensemble of (equiprobable) geological models, flow simulations can be conducted on the entire set. Combining the results of these simulations provides a more comprehensive understanding of the potential reservoir behavior during operation. However, this approach is limited by the need to process each model individually using a full-physics 3D multiphase simulator. Given the multiphysical and multiphase non-linear nature of CO_2 flow, accounting for flow instabilities, phase transitions, petrophysical properties, chemical effects, and other complexities such as geomechanics, these simulations can be computationally expensive [22] and each simulation often requires several CPU hours per realization, depending on the grid size. As a result, brute-force Monte Carlo simulation for the entire ensemble is often impractical, and only a limited number of realizations can typically be analyzed [6].

What makes this problem particularly acute for CCS is the scale and physics of storage systems. Unlike oil and gas, where modelling can often be restricted to the producing reservoir interval, CO_2 storage requires simulating the entire storage complex, including over- and underburden, since pressure perturbations can propagate tens of kilometers from the injection well. Moreover, the dominant physical forces vary across scales: viscous forces near the wellbore, capillary forces in the wake of the passing plume, and gravity in the far field [5]. This shifting force balance makes CCS simulations inherently more physically complex than typical hydrocarbon projects. No single modelling simplification applies everywhere, and heterogeneities ranging from millimeter-scale lamination to kilometer-scale stratigraphic barriers can exert first-order control on CO_2 migration and trapping [23].

Compounding these physical challenges is the limited availability of data at prospective CCS sites. Although most current projects focus on re-using depleted hydrocarbon fields, their global storage capacity is insufficient to deliver the gigatonne-scale volumes required for meaningful climate impact. Consequently, the focus is shifting towards large saline aquifers. Unlike depleted hydrocarbon fields, which typically benefit from extensive seismic surveys, long production histories, and dense well control, aquifer storage projects often begin with only sparse well data and limited high-resolution geophysical

information [5]. This scarcity, combined with the need to forecast plume migration and pressure evolution over much longer timescales (decades to centuries), means that both the number of realizations required (to represent a wider range of poorly constrained heterogeneities across multiple scales) and the cost per realization (due to longer simulation horizons and larger model domains) are substantially greater in CCS than in oil and gas. As a result, robust and effective approaches to uncertainty quantification are essential. Practical workflows are required to screen the most influential uncertainties and focus full-physics simulations on representative models, making probabilistic performance assessment feasible.

A complementary line of work focuses on rapid screening methodologies that prioritize which geological uncertainties deserve detailed simulation. Jackson et al. [24] introduced a workflow that combines experimental design, sketch-based reservoir modelling, and single-phase flow diagnostics to efficiently evaluate the impact of sedimentological heterogeneity on CO_2 migration and trapping. While this approach simplifies physics by neglecting multiphase effects such as capillarity and dissolution, it provides a fast and geologically intuitive way to highlight the most influential heterogeneities. Such screening can guide the selection of scenarios for more detailed full-physics simulation and thus sits between purely conceptual geological modelling and computationally expensive ensemble-based UQ.

In addition, an alternative to full-physics modelling of all possible realizations, proxy models can be used to approximate the results of full-physics multiphase simulations at significantly lower computational cost. Examples include statistical emulators such as response surface models or polynomial chaos expansions [22]. The drawback of such approaches is that they require extensive training, validation, and updating, and they can suffer from overfitting or poor generalisation outside the training domain [25], [26]. Therefore, although they are computationally much cheaper than full-physics simulators, the latter remain superior for reliably estimating reservoir responses of interest.

Assuming one is interested in using the more accurate full-physics simulations to quantify the full range of a reservoir response of interest across an ensemble, but also wishes to reduce computational cost, a practical strategy is to simulate only a selected subset of realizations rather than the complete ensemble. Since random sampling from the ensemble would most likely capture only a limited portion of the uncertainty space, ranking techniques can be employed to identify and select models that better represent the full range of uncertainty in reservoir behavior. This method is used in the oil and gas industry to choose a subset of realizations that estimate the P_{90}, P_{50}, and P_{10} quantiles of key responses [6], [27]. However, the effectiveness of these techniques largely depends on the ranking criteria, which are typically derived from simple static geological properties [7], such as “original oil-in-place” (OOIP), often used as a preliminary estimate of potential oil production performance [28]. The problem is that such static properties reduce complex geological concepts, like depositional setting, facies architecture, or structural configuration, into bulk values. As a result, they generally show weak correlations with dynamic flow responses of interest [6], [29]. For instance, two realizations may have similar OOIP values but differ in structural features such as the presence or absence of faults, which can strongly influence flow behavior. This illustrates why relying solely on static-property–based ranking risks overlooking important geological controls on reservoir performance.

To address these issues, a notable approach to model selection was introduced by Scheidt and Caers [6], who developed a metric-space method for selecting a small yet representative subset of realizations that closely preserves the uncertainty bounds (P_{90}, P_{50}, P_{10}) of the full ensemble. Unlike ranking on static properties, which only provide indirect and often weak indicators for flow behavior, their method relies on dynamic flow responses that inherently reflect connectivity, heterogeneity, and other geological controls on reservoir performance. In their study, these responses were computed efficiently using a fast streamline-based flow simulator, which has been shown to correlate well with full-physics outcomes. The resulting pairwise distance matrix captures behavioral similarity between models and serves as input for dimensionality reduction and clustering techniques. These are then used to identify a reduced subset of realizations. Simulating only this subset with a full-physics model enables accurate reconstruction of key ensemble statistics, thereby capturing the spread of uncertainty with a fraction of the computational cost [8].

Although this method has been applied in the oil and gas industry, where commercial organizations such as Streamsim Technologies make use of it [30], it has, to our knowledge, not yet been explored for CO_2 subsurface storage. However, the dynamic-response–based metric-space approach is fundamentally

application-agnostic and could potentially be adapted to CCS. Therefore, the aim of this study is to investigate whether such methods can be used to select a representative subset from a large ensemble of geological models to capture the full range of uncertainty in terms of storage capacity and plume migration for CO_2 injection.

To this end, the study evaluates dynamic flow responses, hereafter referred to as flow diagnostics, which can serve as dissimilarity metrics while remaining far less computationally demanding than simulating the full ensemble over its lifetime. Specifically, two sources of dynamic information are considered: (i) streamline-based flow diagnostics, and (ii) early-time flow diagnostics obtained from full-physics simulations. Dimensionality reduction and clustering techniques are then applied to identify the most robust workflows for accurately reconstructing key ensemble statistics. In addition, a calibration-informed genetic algorithm (GA) selection strategy is tested as a complementary approach to ensemble reduction, assessing whether representative models chosen from short simulated calibration periods can still accurately reconstruct ensemble percentiles over the full simulation lifetime.

This, in turn, would accelerate uncertainty quantification workflows and enable computationally efficient probabilistic performance forecasts for specific well configurations and control strategies.

1.1 Research Questions

To investigate whether certain metric-space methods can be used to select a subset of models from a large ensemble of geological realizations, while still capturing the full range of uncertainty in key CO_2 storage performance metrics, the following research questions will be addressed throughout this study.

Main Research Question

> To what extent can clustering methods based on dissimilarity metrics be used to reduce simulation costs while accurately reconstructing key uncertainty bounds (P_{90}, P_{50}, P_{10}) of CO_2 storage performance metrics?

Subquestions

1. Which combinations of flow diagnostics, dimensionality reduction, and clustering approaches most accurately reconstruct ensemble percentiles (P_{90}, P_{50}, P_{10}) across different output types (e.g., injection rate, plume migration, coverage)?
2. How can the minimum subset size of geological models be determined to ensure reliable reconstruction of ensemble percentiles, and how does reconstruction accuracy evolve as the number of selected models increases?
3. Which parameters most strongly influence variability in CO_2 storage responses, and how can clustering-based analyses be used to identify them?
4. To what extent can an informed genetic algorithm improve model subset selection and percentile reconstruction, compared with dynamic flow-based clustering methods, in terms of accuracy and simulation cost?

1.2 Structure of the Report

Chapter 2 provides background on the proposed distance-based clustering workflow developed by Scheidt and Caers [6]. It also introduces the geomodel ensemble used in this study to represent interpretational uncertainty in model construction and presents the full-physics simulator (open-DARTS) and the streamline-based flow simulator (3DSL) employed throughout this work. Chapter 3 describes the methodology applied in this study, including the clustering workflows evaluated for the storage metrics of interest and the complementary parameter sensitivity analysis. Chapter 4 presents the results for the various CO_2 storage metrics, while Chapter 5 examines potential reasons for the differences in workflow performance and explores the key mechanisms driving variability in the storage metrics. This chapter also generalizes the most promising workflows and highlights directions for future research. Finally, Chapter 6 summarizes the main findings of this study.

2

Background

This chapter provides the background required to place the methodology of this study into context. Section 2.1 introduces the concepts of dimensionality reduction and clustering in reservoir modelling, highlighting how distance-based methods and dynamic screening have been applied in prior studies. Section 2.2 describes the construction of the geological model ensemble used in this study, including the sources of geological uncertainty and the modelling choices that define the realizations. Section 2.3 then presents the open-DARTS full-physics reservoir simulator used for this study, including its governing equations, assumptions, and numerical formulation. Finally, Section 2.4 introduces the 3DSL streamline simulator used for this study and outlines its underlying principles and formulations.

2.1 Dimensionality Reduction and Clustering

In 2007, Scheidt and Caers [6] introduced a novel workflow for selecting a subset of realizations (N_s) to capture the full range of uncertainty in a reservoir response of interest from the full geomodel ensemble consisting of N_r realizations. Earlier screening approaches had relied on static metrics such as the Dykstra–Parsons coefficient [31] or original oil-in-place to reduce ensembles, but these measures only characterize heterogeneity in petrophysical properties and do not capture differences in dynamic flow behaviour. Scheidt and Caers advanced the concept by introducing dynamic screening, in which realizations were grouped according to similarities in predicted production or flow response. Because this approach targets flow-relevant characteristics, the reduced subsets were able to preserve the uncertainty in dynamic outcomes, ensuring that reconstructed percentiles P_{90} (10th percentile), P_{50} (median), and P_{10} (90th percentile) closely matched those of the full ensemble. More recently, Watson et al. [21] confirmed the advantage of dynamic measures by demonstrating that flow diagnostics offer a computationally efficient way to rank and screen realizations while being more informative than static metrics.

To achieve the close resemblance with the key ensemble statistics, Scheidt and Caers parameterized uncertainty using a realization-based distance metric. This metric, represented by a single parameter (δ), quantifies the dissimilarity between any two realizations and can be tailored to specific applications. In their study, distances were computed based on the field oil production rate at a specific time, obtained using the tracer simulation capability of a commercial streamline-based flow simulator. The advantage of this approach is that streamline simulators are significantly faster (up to two orders of magnitude) than full-physics simulators [32], while providing flow responses that are more strongly correlated with those from full-physics simulators compared to static reservoir properties. By comparing these distances, which reflect similarity in both geological features and dynamic behavior, a distance matrix D was constructed to quantify pairwise relationships across the entire ensemble.

This ($N_r \times N_r$) distance matrix is then used to map the realizations into a Euclidean space R using Multidimensional Scaling (MDS). MDS is employed for its ability to represent the pairwise dissimilarities as a configuration of points in a low-dimensional Euclidean space (typically $n = 2$), where each point corresponds to a realization. In this space, clustering techniques such as Principal Component Analysis

(PCA) or k-means can be applied to group similar realizations and select representative models.

However, Scheidt and Caers observed that the structure of the mapped points in R often exhibited non-linear relationships, whereas traditional statistical methods like PCA and k-means assume linearity. To address this, they applied kernel methods to transform the Euclidean space R into a higher-dimensional feature space F, where the relationships between realizations became more linear. This transformation enabled the effective application of linear clustering techniques such as k-means.

The k-means algorithm was chosen for its simplicity and its compatibility with the assumption that realizations with similar flow responses (as defined by the distance metric) would be located near each other in feature space. Consequently, selecting cluster centroids in space F provided a subset of representative models that preserved the diversity of dynamic behavior across the ensemble [8].

These regression methods facilitated the identification of a small subset of representative realizations from the larger ensemble. By applying the full-physics simulations to this subset, it became possible to quantify uncertainty in the response variable, such as estimating the P_{90}, P_{50}, and P_{10} quantiles, which closely correlated with the values obtained from flow simulations conducted on the complete ensemble. Consequently, they demonstrated that their metric space method significantly reduced the number of simulations required to capture the range of uncertainty of the complete ensemble.

Figure 2.1 shows the workflow applied in their study.

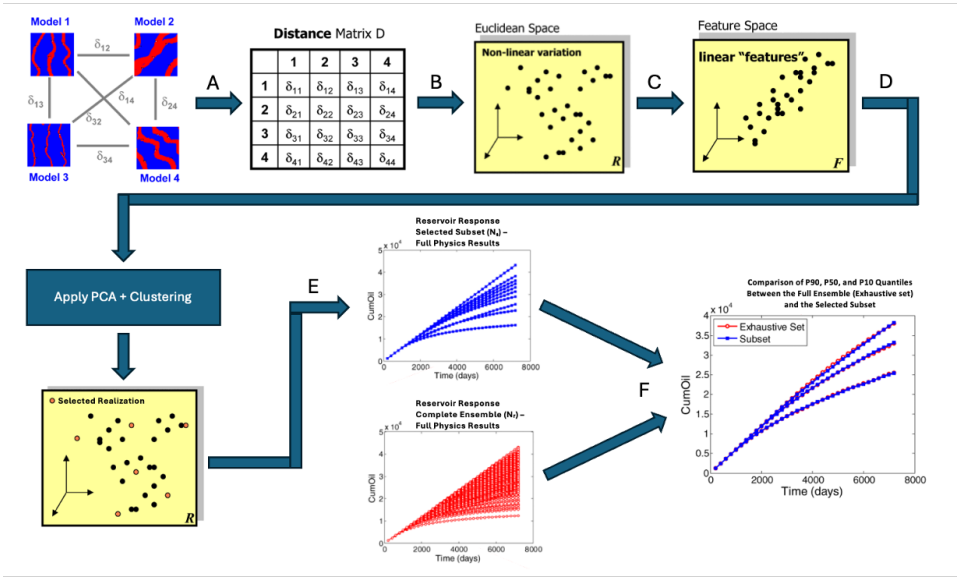


Figure 2.1: Proposed workflow for uncertainty quantification. (A) Compute pairwise distances between realizations and assemble into a distance matrix. (B) Map realizations into Euclidean space. (C) Apply kernel methods to obtain a feature space with improved linearity. (D) Apply Principal Component Analysis and clustering to select representative realizations for flow simulation. (E) Simulate representative realizations with a flow simulator. (F) Assess accuracy of quantile reconstruction (P90, P50, P10) by comparing quantiles of the selected subset with those of the full ensemble. Adapted from Scheidt and Caers [7].

While the workflow of Scheidt and Caers has shown to be effective, the use of two consecutive dimensionality reduction techniques (MDS followed by KPCA) raises methodological questions. Both MDS and KPCA are designed to extract structure from the data by projecting it into a lower-dimensional space, but applying them sequentially may obscure or distort the original distance relationships. One could argue that once MDS has transformed the pairwise distance matrix into Euclidean coordinates, directly applying clustering methods (e.g., k-means or k-medoids) would suffice. Conversely, if nonlinear structures are suspected from the outset, KPCA (or an alternative nonlinear technique) might be applied directly to the distance matrix, bypassing MDS altogether.

In this study, it is therefore decided to explore several alternative metric-space workflows, each starting from a realization-based distance matrix D but diverging in the dimensionality reduction and clustering steps and which will be explained in more detail in Section 3.4

2.2 The Geomodel Ensemble

This section provides an overview of the geomodel ensemble used in this study. The ensemble is a set of alternative geological models (realizations) that honor the available data and geological concepts while varying within plausible uncertainty ranges, derived from the Watt Field case study [33]. It integrates interpretational uncertainty hierarchically across the workflow, including top surfaces, fault models, fault transmissibility multipliers, RPD cutoffs, and facies-distribution modelling techniques. The section first describes the geological setting of the Watt Field, then outlines the principal geological uncertainties and modelling concepts, and finally summarizes the simulation assumptions applied in this study.

2.2.1 Geology

Geologically, the interpreted depositional environment of the studied reservoir is a braided river system which forms the geological conceptual model of the study [33], [34]. A conceptual illustration of this system is shown in Figure 2.2, characterized by three typical facies: fluvial channel sands, overbank fine sands, and shales.

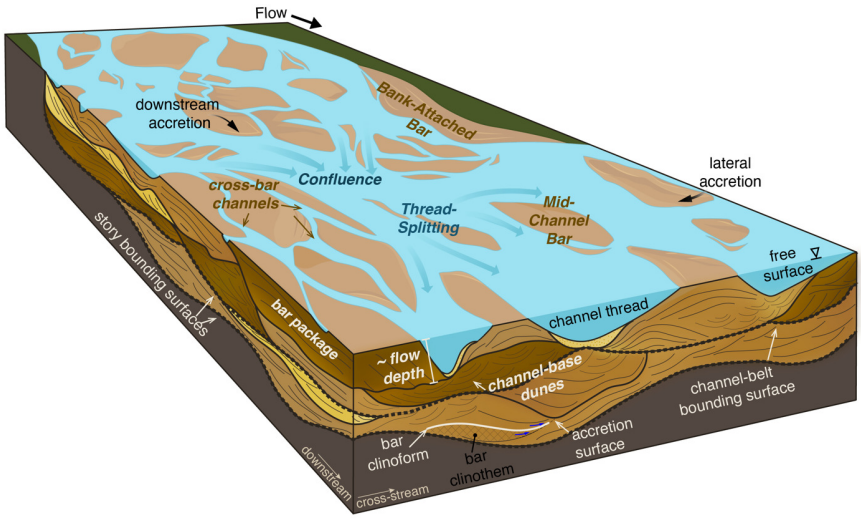


Figure 2.2: Conceptual representation of a braided river system and its stratigraphic deposition [35].

2.2.2 Geological Uncertainties, Interpretations, and Concepts

Although the Watt Field Case Study is semi-synthetic, where fluid properties, relative permeabilities, and capillary data are synthetic, the structural geology (including possible fault networks, facies distribution, and top structures) is based on real field data. Uncertainty arises at multiple levels of the reservoir modelling workflow, following a hierarchical structure: from depositional interpretation through the structural definition of top surfaces and faults, the classification of facies from wireline logs, and the spatial distribution of facies and petrophysical properties. At each stage, uncertainties reflect the indirect nature of the available data (such as seismic and log measurements that are noisy, incomplete, and resolution-limited) and the interpretational judgements required to translate these data into geological models.

The top surface was interpreted from seismic data converted to depth, but seismic resolution and velocity-model assumptions introduce ambiguity in horizon picking. To reflect this uncertainty, three alternative top surface structures were included (TS1, TS2 and TS3). These structural variations affect reservoir thickness and closure geometry, which in turn influence the available storage volume and the migration pathways of injected CO_2.

Similarly, three different fault systems were constructed to represent uncertainty in fault geometry. Faults in the Watt Field are primarily identified from reflector discontinuities in seismic data, yet only displacements larger than the seismic resolution ($\approx 10\,\mathrm{m}$) can be confidently detected. Since all observed faults in the seismic strike east–west and no large-scale faulting was identified in other orientations, two of the fault models (FM1 and FM2) included only the east–west striking faults visible in the seismic. A

third model (FM3), however, incorporated additional north–south striking faults that may exist below the detection threshold. These additional faults cannot be ruled out geologically and, given their potential impact on reservoir connectivity, were included as a modeller's interpretation.

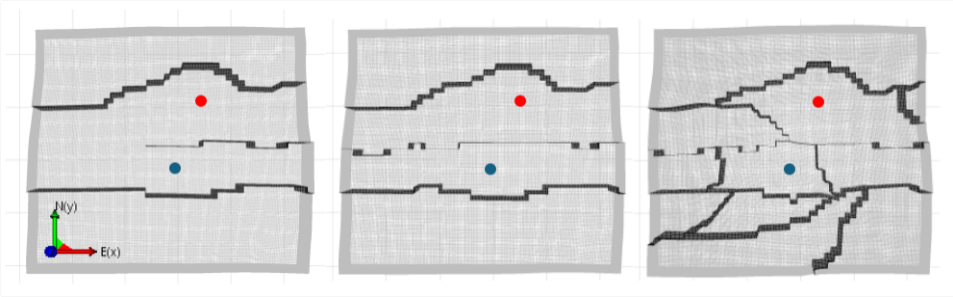


Figure 2.3: Top view of the three fault model scenarios (FM1, FM2, and FM3), illustrating variations in fault geometry. The blue dot represents well location 1 (Ensemble 1), and the red dot represents well location 2 (Ensemble 2).

In addition to geometry, uncertainty also exists in fault transmissibility. Even when fault positions are well defined, their capacity to transmit or impede flow cannot be uniquely determined because it depends on fault-zone processes that are not captured at reservoir-model scale. Factors such as fault gouge development, clay smear continuity, cataclasis, cementation, and facies juxtapositions across the fault plane are highly variable in nature, generally lie below seismic resolution, and are rarely measured directly, making them impossible to constrain with confidence. For this reason, transmissibility multipliers are commonly applied in reservoir simulation as a pragmatic way to span plausible ranges of fault behaviour [36]. In this study, two values were used uniformly across all faults: a high multiplier of 0.9 (MULT1), representing nearly open faults, and a low multiplier of 0.01 (MULT2), representing nearly sealing faults [34].

In terms of facies predictions, the Watt Field Study derived facies interpretations from core plug data and wireline logs. Core data were used to characterize porosity and permeability, while electrofacies identification relied on the Relative Porosity Difference (RPD) calculated from neutron and density porosity logs. Because the RPD cutoff is interpretational, three thresholds (0.6 (CO1), 0.7 (CO2), and 0.8 (CO3)) were tested, producing alternative electrofacies logs. These thresholds directly control the classification of net versus non-net facies and thus influence the reservoir's net-to-gross ratio. Consequently, the cutoff choice affects the spatial distribution and continuity of permeable facies, which in turn govern flow connectivity and predicted CO_2 plume migration. Permeability prediction was then performed using porosity–permeability cross-plots from cored wells, adopting a single predictive relationship (Appendix A.1) to extrapolate permeability values to uncored wells and uncored reservoir intervals [33].

To further reflect conceptual uncertainty, two facies distribution modelling techniques were applied: object-based and pixel-based modelling.

Object-based modelling inserts geological features such as channels or lobes as discrete bodies into the reservoir model, selecting from prescribed input distributions and conditioning to well data. This approach is geologically intuitive and captures realistic geometries but may create unrealistic correlations between wells and requires careful parameter tuning.

Pixel-based modelling, in contrast, assigns facies values cell-by-cell using geostatistical algorithms, guided by well data and variogram models. This method avoids funnelling effects and handles complex conditioning better, but struggles to reproduce well-defined features such as channels and may produce over-smoothed patterns [17, 37].

By including both approaches, the study spans alternative conceptual representations of fluvial architecture, ensuring differences in facies continuity, connectivity, and geometry are propagated into the flow simulations. Combined with the three distinct facies logs derived from RPD cutoffs, this resulted in six unique permeability distribution scenarios. The corresponding horizontal permeability distributions for these six scenarios are presented in Appendix A.3.

Finally, combining these varied parameters resulted in an ensemble of 108 geological models. This ensemble directly reflects the main sources of geological uncertainty, as each modelling parameter (structural surfaces, fault networks and transmissibility, facies distribution, and electrofacies cutoffs) embodies a plausible but uncertain interpretation of the subsurface. By combining all possible parameter combinations, the ensemble captures the compounded impact of multiple uncertainty sources, providing a basis for quantifying their influence on flow simulation outcomes. An overview of the parameters and their associated uncertainty types is provided in Table 2.1.

Table 2.1: Overview of ensemble parameters, specification, and associated uncertainty types.

Ensemble Parameter	Name	Uncertainty Type
Top structure	TS1, TS2, TS3	Interpretational
Fault model	FM1, FM2, FM3	Interpretational
Facies model (cut-offs)	CO1, CO2, CO3	Parametric
Modeling method	OBJ, PIX	Conceptual
Fault transmissibility	MULT1, MULT2	Parametric

2.2.3 Simulation Settings

The original reservoir models of the Watt Field study [33] cover a surface area of $20\ \mathrm{km} \times 5.9\ \mathrm{km}$, elongated in the east–west direction, forming a shallow anticline with a total modelled thickness of approximately $190\ \mathrm{m}$. However, since this study focuses on a single CO_2 injector system, the area of interest was reduced to a cropped region of $4.5\ \mathrm{km} \times 4\ \mathrm{km}$ (see Figure 2.4) to reduce simulation time. Grid cells were defined with dimensions of $25 \times 25 \times 2.5\ \mathrm{m}$, resulting in a grid size of $\sim$2.3 M cells ($180 \times 160 \times 80$).

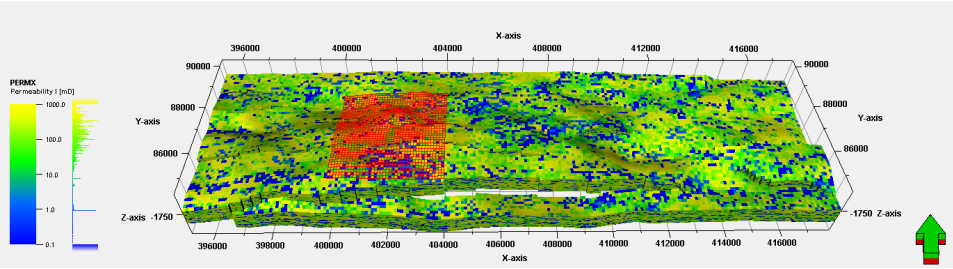


Figure 2.4: Cropped area of interest (shown in red) extracted from the full-field reservoir model of the Watt Field case study.

In this study, two well locations were investigated across the full set of geological realizations (Figure 2.4). The blue dot marks Well Location 1 and the red dot marks Well Location 2. For clarity, the set of 108 realizations with injection at Well Location 1 is referred to as *Ensemble 1*, while the same 108 realizations with injection at Well Location 2 is referred to as *Ensemble 2* in this study.

The reservoir is assumed to be fully saturated with brine prior to CO_2 injection, classifying it as a saline aquifer. The top surface of the reservoir lies at a depth of approximately $1{,}555\ \mathrm{m}$. A hydrostatic pressure gradient of $97.5\ \mathrm{bar/km}$ and a thermal gradient of $30\ ^\circ\mathrm{C/km}$ were assumed, resulting in pressure values ranging from approximately 152 to $170\ \mathrm{bar}$ and temperature values from 66 to $72\ ^\circ\mathrm{C}$ from top to bottom.

Each facies was assigned a distinct relative-permeability curve, using facies-dependent Corey parameters (Table 2.2) provided through personal communication with Prof. Farajzadeh. The resulting curves are presented in Appendix A.2. Capillary-pressure effects were neglected to reduce simulation time. While this simplification can influence CO_2 migration [23], it is considered acceptable for this study because the focus is limited to the injection stage (20 years). Capillary forces become particularly important in the post-injection phase, where they control long-term migration and trapping, which are beyond the scope of this work.

The reservoir was assumed compressible, with porosity modeled as a pressure-dependent property and permeability kept constant. No explicit geomechanical deformation was considered. Key reservoir and rock properties are summarized in Table 2.3.

Table 2.2: Facies-dependent Corey relative permeability parameters

Facies	n_w	n_g	S_{wc}	S_{gc}
Fluvial channel sands	4.0	2.0	0.20	0.30
Overbank fine sands	3.0	1.5	0.20	0.20
Shales	1.5	1.5	0.32	0.10

Table 2.3: Reservoir and rock properties .

Property	Symbol	Value	Unit
Static reservoir properties			
Porosity	ϕ	0 to 0.266	–
Permeability (x, y)	k_h	0 to 9,442	mD
Permeability anisotropy	k_v/k_h	$1/10$	–
Reservoir depth	D	1,505.5 to 1,845.0	m
Pressure gradient	–	97.5	bar/km
Geothermal gradient	–	30	°C/km
Surface temperature	T_{surf}	20	°C
Initial reservoir pressure	p_0	150 to 181	bar
Initial reservoir temperature	T_0	65 to 75	°C
Lateral boundary volume	V_{lat}	1×10^{10}	m^3
Rock properties			
Rock compressibility	c_r	1×10^{-5}	1/bar
Facies 1, 2 (fluvial channel sands; overbank fine sands)			
Thermal conductivity	κ	2.59×10^5	J/$(m \cdot d \cdot K)$
Volumetric heat capacity	C_r	2.45×10^6	J/$(m^3 \cdot K)$
Facies 3 (shale)			
Thermal conductivity	κ	1.90×10^5	J/$(m \cdot d \cdot K)$
Volumetric heat capacity	C_r	2.30×10^6	J/$(m^3 \cdot K)$

2.3 open-DARTS

For this study, the open Delft Advanced Research Terra Simulator (open-DARTS), developed at Delft University of Technology [38], was used as the full-physics reservoir simulator to model CO_2 sequestration. The simulator was configured to represent two-phase, two-component flow (aqueous brine and CO_2-rich phases), with partitioning of CO_2 between phases through dissolution. A non-isothermal formulation was adopted, capturing advective and conductive heat transport, as well as thermal exchange between fluids and the rock matrix. This was necessary because CO_2 was injected at 300 K, while the in-situ reservoir temperature was significantly warmer at around 340 K. Such thermal disequilibrium can arise in practice, for example when injected CO_2 is cooled during surface handling and wellbore transport, and may therefore influence fluid properties and plume dynamics, justifying the inclusion of energy transport in the model. Kinetic dissolution of CO_2 was included; molecular diffusion, mineral reactions, and long-term geochemical transformations were not. To represent an open saline aquifer, open boundary conditions were applied at the model edges.

Open-DARTS combines a high-performance C++ core with a flexible and customizable Python interface. One of the key differentiators between open-DARTS and other state-of-the-art simulators such as TOUGH2 and AD-GPRS is its Operator-Based Linearization (OBL) approach. This method caches evaluation points and computes derivatives through interpolation, resulting in a reduction in CPU time [38] and making it especially interesting when running geomodels of an ensemble in sequence as done with this study.

2.3.1 Governing Equations

The open-DARTS simulator solves a coupled system of partial differential equations that govern the conservation of mass and energy and describe flow and transport in porous media [39]. These equa-

tions form the basis for full-physics modeling of multiphase, multicomponent CO_2 sequestration under non-isothermal conditions. This section outlines the mathematical formulation of the conservation laws as implemented in open-DARTS, including the underlying assumptions and numerical discretization strategies adopted in this study.

The modeled system consists of n_c chemical components (here, CO_2 and H_2O) distributed across n_p fluid phases: an aqueous phase (*Aq*) and a vapor phase (*V*). Conservation equations are solved for each component index $c \in \{1, \ldots, n_c\}$, as well as for thermal energy, represented by an extended index $c = n_c + 1$. The general conservation law for each quantity is expressed in integral form over a control volume Ω with boundary Γ as follows

$$\frac{\partial}{\partial t} \int_{\Omega} M^c \, d\Omega + \int_{\Gamma} \mathbf{F}^c \cdot \mathbf{n} \, d\Gamma = \int_{\Omega} Q^c \, d\Omega, \tag{2.1}$$

where M^c is the accumulation term for the c^{th} component, $\mathbf{F}^c$ the flux term of the c^{th} component, Q^c the source or sink term of the c^{th} component, and $\mathbf{n}$ the unit normal pointing outward to the domain boundary.

The mass accumulation term collects each component distribution over n_p fluid phases in a summation form

$$M^c = \phi \sum_{j=1}^{n_p} x_{cj} \rho_j s_j, \qquad c = 1, \ldots, n_c, \tag{2.2}$$

where ϕ is the porosity, s_j is the saturation of phase j, ρ_j its molar density [$kmol/m^3$], and x_{cj} the mole fraction of component c in phase j. For thermal energy ($c = n_c+1$), accumulation includes contributions from both fluids and the rock matrix

$$M^{n_c+1} = \phi \sum_{j=1}^{n_p} \rho_j s_j U_j + (1 - \phi) U_r, \tag{2.3}$$

where U_j denotes the specific internal energy [kJ] of phase j, and U_r that of the rock.

The rock was assumed compressible, and porosity was modeled as a pressure-dependent property using a linear relation

$$\phi = \phi_0 \left(1 + c_r (p - p_{\text{ref}})\right), \tag{2.4}$$

where ϕ_0 is the initial porosity, p_{ref} the reference pressure [bars], and c_r is the rock compressibility [1/bar].

Component transport is modeled as purely advective. For each component $c \leq n_c$, the mass flux is given by the summation over n_p fluid phases

$$\mathbf{F}^c = \sum_{j=1}^{n_p} x_{cj} \rho_j \mathbf{u}_j, \tag{2.5}$$

where $\mathbf{u}_j$ is the Darcy velocity of phase j, calculated by

$$\mathbf{u}_j = -\mathbf{K} \frac{k_{rj}}{\mu_j} \left(\nabla p_j - \rho_j g \nabla z\right), \tag{2.6}$$

where $\mathbf{K}$ is the permeability tensor [mD], k_{rj} is the relative permeability of phase j, μ_j is the viscosity of phase j [mPas], p_j is the pressure of phase j [bar], $\gamma_j = \rho_j g$ is the specific weight [N/m^3], and z is the depth vector [m].

For energy transport ($c = n_c+1$), the energy flux includes both convective and conductive contributions

$$\mathbf{F}^{n_c+1} = \sum_{j=1}^{n_p} h_j \rho_j \mathbf{u}_j + \kappa \nabla T, \tag{2.7}$$

where h_j is the specific enthalpy of phase j [kJ/kg], κ is the effective thermal conductivity [kJ/m/day/K], and T is the temperature [K].

The source term Q^c is expressed as the contribution from wells

$$Q^c = \sum_{j=1}^{n_p} x_{cj}\, \rho_j\, q_j, \qquad c = 1, \ldots, n_c, \tag{2.8}$$

$$Q^{n_c+1} = \sum_{j=1}^{n_p} \bar{h}_j\, \rho_j\, q_j. \tag{2.9}$$

Here, q_j denotes the volumetric well source or sink term for phase j (positive for injection, negative for production), which is nonzero only in perforated cells. The quantities ρ_j and x_{cj} are the molar density and mole fraction of component c in phase j, while $\bar{h}_j$ represents the molar enthalpy of phase j.

The full system is discretized in space using a finite-volume method on unstructured grids, while time integration is performed using an implicit backward Euler scheme. For each control volume i and component index c, the discretized residual takes the form

$$R_i^c = V_i \left(M_i^c(\omega_i^n) - M_i^c(\omega_i^{n-1}) \right) - \Delta t \left(\sum_l A_l F_l^c(\omega) + V_i Q_i^c(\omega) \right) = 0, \tag{2.10}$$

where V_i is the volume of cell i, A_l denotes the interface area between neighboring cells, Δt the time step, and ω represents the primary variables, including pressure, temperature, saturations, and phase compositions.

Finally, phase behavior is computed via a compositional flash algorithm using a hybrid equation of state (EOS) formulation. Specifically:

- The **Peng–Robinson** (PR) EOS is used for the CO_2-rich vapor phase [40].
- The **Aqueous EOS** (AQ) represents the water phase and dissolved species. Thermophysical properties of water and ions are described following Jäger et al. (2003) [41], while CO_2 solubility in aqueous solutions is modeled using the formulation of Ziabakhsh-Ganji and Kooi (2012) [42].

2.4 3DSL

In addition to the full-physics simulator open-DARTS, this study uses 3DSL, a commercial streamline-based flow simulator developed by Streamsim Technologies [43]. Conventional finite-volume simulators such as open-DARTS solve the full system of coupled conservation equations directly on the grid, transporting fluids from cell to cell at each timestep. In contrast, 3DSL reformulates the transport problem in terms of one-dimensional streamlines: curves that are everywhere tangent to the local velocity field at a given instant. It therefore makes use of a dual-grid approach: a static Eulerian grid is used to compute the pressure field and construct the velocity distribution from Darcy's law, while a dynamic Lagrangian grid of streamlines is used to transport fluids from sources (in this study the injection well) to sinks (in this study the open reservoir boundaries) [44]. This mapping reduces grid-orientation effects and numerical diffusion, and because no global Courant–Friedrichs–Lewy (CFL) condition restricts timestep size, much larger timesteps can be taken. As a result, 3DSL can be significantly faster than full-physics simulators (up to two orders of magnitude [32]) while remaining particularly effective for large, heterogeneous reservoirs where advective displacement dominates. Multiphase gravity effects are included through operator splitting, where convective transport along streamlines is advanced first, followed by a separate vertical gravity step. Capillary and dispersion effects are neglected.

Streamline geometry and velocities are determined by the spatial distribution of petrophysical properties (e.g. permeability, porosity, relative permeability) in combination with the imposed well and boundary conditions. Streamlines are held fixed over a time interval Δt, during which components are transported along them, after which the system state (pressures and compositions) is updated. Because the streamline network reflects the evolving flow field, it must be periodically re-traced whenever mobilities or boundary conditions significantly change. The tracing procedure itself is analytical: within each gridblock, the velocity field is assumed to vary linearly in each coordinate direction, which allows the exact travel time to each cell face to be calculated. The streamline exits through the face with the shortest transit time, and the entry point into the neighboring block is then computed explicitly. Repeating this process from block to block reconstructs complete trajectories connecting injectors to producers. Since entry and exit coordinates are obtained in closed form, numerical integration is avoided, ensuring both accuracy and efficiency even in complex three-dimensional flow fields [32]. An illustration of the streamline workflow is shown in Figure 2.5.

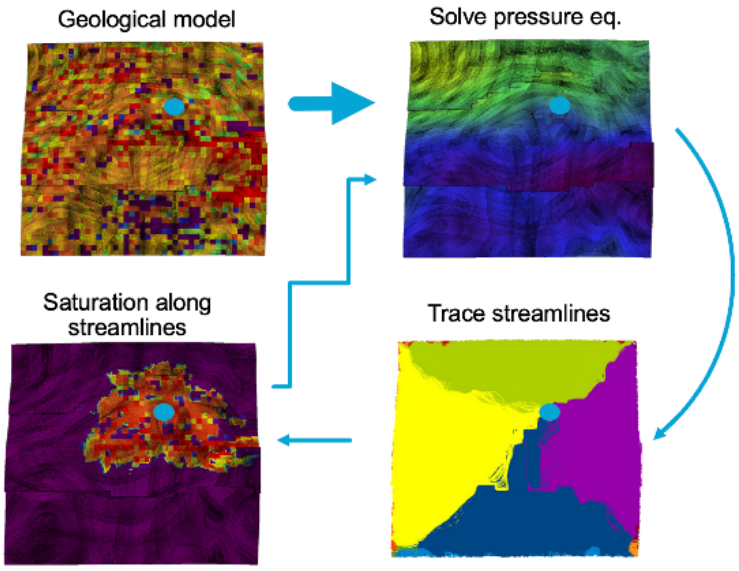


Figure 2.5: Illustrative workflow of streamline simulation for CO_2 injection (workflow structure inspired by Stüben et al. [45]). The blue dot marks the injector, shown consistently across all panels. The geological model (top left) shows the permeability distribution, which provides the basis for solving the pressure field on the static grid (top right). From this velocity field, streamlines are traced (bottom right), defining flow paths to the reservoir boundaries. Fluid transport is then solved along these streamlines, and the resulting saturations are mapped back to the simulation grid for visualization (bottom left).

Beyond its computational efficiency, the streamline framework is particularly valuable for the objectives of this study. Because clustering requires a large number of simulation runs, using a full-physics simulator for this task could be prohibitively expensive. Streamline simulation offers a reduced-order but physics-grounded alternative: it captures the essential advective structure of the flow field while neglecting less dominant effects such as capillarity and dispersion. The resulting streamline geometry and associated time-of-flight values naturally define flow-based distances between injectors and reservoir regions. These distances are physically meaningful, as they reflect reservoir connectivity, pressure gradients, and heterogeneity, in contrast to purely geometric or statistical similarity measures. In this sense, 3DSL is not used here as a black-box proxy, but rather as a physics-based dimensionality reduction tool that enables flow-informed clustering of ensemble members at manageable computational cost. Streamline simulation is especially well-suited for pressure-driven displacements and immiscible multiphase flow, but is less accurate for capillary-dominated processes, strong compressibility effects, or highly transient flow behavior, where frequent re-tracing would be required.

Within this streamline framework, the transport equations can be solved under different flow assump-

tions. 3DSL provides several formulations that balance physical realism against computational efficiency. In this study, two options are considered: the two-phase immiscible formulation, which models multiphase displacement without component exchange between phases, and the simplified single-phase formulation, which represents a limiting case with constant mobility and linear flow behavior. The following subsections outline the governing equations and assumptions for both formulations and discuss their relevance to CO_2 sequestration.

2.4.1 Governing Equations: Two-Phase Immiscible Flow

The two-phase immiscible formulation assumes that there is no solubility between phases; the gas component resides only in the gas phase, and water only in the aqueous phase. Each phase is assigned its own PVT properties (viscosity, compressibility, and density), which are treated independently in the model. This simplification provides potentially a reasonable reduced-physics approximation for CO_2–brine displacement during the injection phase, when advective migration dominates, although it does not capture dissolution processes that become relevant on longer timescales.

The pressure field is obtained from an IMPES formulation of the total velocity equation [32]

$$\nabla \cdot \Big(\mathbf{k} \big(\lambda_t \nabla p + \lambda_g \nabla D \big) \Big) = 0, \tag{2.11}$$

where $\mathbf{k}$ is the absolute permeability tensor, p the pressure, and D the depth below datum.

The total mobility λ_t and gravity mobility λ_g are defined as

$$\lambda_t = \sum_{j=1}^{n_p} \frac{k_{rj}}{\mu_j}, \qquad \lambda_g = \sum_{j=1}^{n_p} \frac{k_{rj}}{\mu_j} \rho_j g, \tag{2.12}$$

with k_{rj} the relative permeability of phase j, μ_j the phase viscosity, and ρ_j the phase density.

To complement the pressure equation, the material-balance equation for each phase j enforces conservation of mass and closes the system

$$\phi \frac{\partial S_j}{\partial t} + \mathbf{u}_t \cdot \nabla f_j + \nabla \cdot \mathbf{G}_j = 0, \tag{2.13}$$

where ϕ is the porosity. Eqs. (1) and (3) form the core of the IMPES formulation, while the following definitions specify the phase contributions.

The fractional flow of phase j is defined as

$$f_j = \frac{k_{rj}/\mu_j}{\sum_{i=1}^{n_p} k_{ri}/\mu_i}. \tag{2.14}$$

The total Darcy velocity is then given by

$$\mathbf{u}_t = -\mathbf{k} \big(\lambda_t \nabla p + \lambda_g \nabla D \big), \tag{2.15}$$

and the gravity segregation flux of phase j is

$$\mathbf{G}_j = \mathbf{k}\, g\, f_j \nabla D \sum_{i=1}^{n_p} \frac{k_{ri}}{\mu_i} (\rho_i - \rho_j). \tag{2.16}$$

2.4.2 Governing Equations: Single-Phase Flow

The single-phase formulation arises as a special case of the immiscible model. In this representation, a two-phase system is assumed in which all phases have identical PVT properties (constant viscosity and density), no phase interactions, and no saturation dependence. In addition, all phases have straight-line relative permeabilities with no residual saturations. Rock and fluids are taken as incompressible, and since all densities are equal, gravity effects vanish and are therefore neglected.

The total mobility $\lambda_t = 1/\mu$ is constant and independent of space and time. An intuitive way to see this is to imagine two phases (oil and water) with identical viscosities, so that

$$\lambda_t = \frac{S_o}{\mu} + \frac{S_w}{\mu} = \frac{1}{\mu}, \tag{2.17}$$

where S_o and S_w denote oil and water saturations. Since $S_o + S_w = 1$, the total mobility remains constant regardless of the phase distribution.

The governing pressure equation therefore reduces to a simple elliptic problem

$$\nabla \cdot \big(k\, \lambda_t \nabla p\big) = q, \tag{2.18}$$

where k is the permeability [m^2], p the pressure [Pa], μ the viscosity [Pa·s], and q denotes source or sink terms [1/s] (with $q > 0$ for injection). Defining a rescaled pressure $P := \lambda_t p$, the system can be expressed in a compact form

$$\nabla \cdot (k \nabla P) = q. \tag{2.19}$$

The associated Darcy flux is then

$$\mathbf{u} = -k \nabla P, \tag{2.20}$$

where $\mathbf{u}$ is the Darcy velocity (volumetric flux per unit cross-sectional area) with units [m/s]. The solution of the pressure equation therefore defines a steady-state velocity field, which forms the basis for streamline tracing. The flux is steady in time once p is solved, provided q and the boundary conditions remain time-independent. Consequently, the pressure field remains unchanged unless the well controls or boundary conditions are modified.

In streamline simulation, particles are advected with the pore velocity

$$\frac{d\mathbf{x}}{dt} = \frac{\mathbf{u}}{\phi}, \tag{2.21}$$

where ϕ is the porosity [-]. Any passive scalar quantity, such as a tracer concentration, is then transported purely by advection along these trajectories. Since there is no coupling to saturation, the streamlines do not change over time, and transport reduces to one-dimensional advection along fixed paths. This makes the single-phase model computationally very efficient and well suited as a flow diagnostic or for rapid geological screening, while acknowledging that multiphase displacement effects are neglected.

3

Methodology

This chapter outlines the methodology used in this study. Section 3.1 introduces the storage metrics that form the basis for evaluating ensemble behavior. Section 3.2 describes the simulation setup and key model assumptions, including both the full-physics and streamline simulations. Section 3.3 explains how flow-based distance metrics are derived, while Section 3.4 presents the dimensionality reduction and clustering workflows applied to identify representative subsets of models. Section 3.4.3 discusses the internal validation criteria used to guide the selection of the number of clusters. Section 3.5 details how reduced ensembles are benchmarked against the full ensemble, including accuracy and simulation-cost metrics. Section 3.6 consolidates complementary sensitivity analyses, including a distance-based global sensitivity analysis (dGSA; Subsection 3.6.1) and a variance-based, time-weighted partial η^2 formulation (Section 3.6.2). Finally, Section 3.7 introduces a frequency-informed genetic-algorithm approach for model selection.

3.1 CO_2 Storage Metrics

The primary objective of this study is to evaluate whether flow diagnostics can be used for distance-based clustering to reduce the number of full-physics simulations required, while still preserving the accuracy of key uncertainty bounds (P_{90}, P_{50}, P_{10}) for CO_2 storage metrics. This section outlines the metrics used to assess the representativeness of reduced ensembles: (i) injection rate, (ii) maximum plume extent, and (iii) plume areal coverage. Together, these metrics capture both injectivity and the spatial evolution of CO_2 in the subsurface, which are critical for capacity estimation and risk assessment.

3.1.1 Injection Rate

Injection rate [tonnes per day] reflects the reservoir's capacity to accommodate CO_2 under a fixed bottomhole pressure and directly constrains the feasible scale of storage operations. Geological variability across the ensemble (e.g., permeability and connectivity) results in distinct injection capacities, and the distribution of these outcomes forms the basis for quantifying uncertainty through ensemble percentiles (P_{90}, P_{50}, P_{10}).

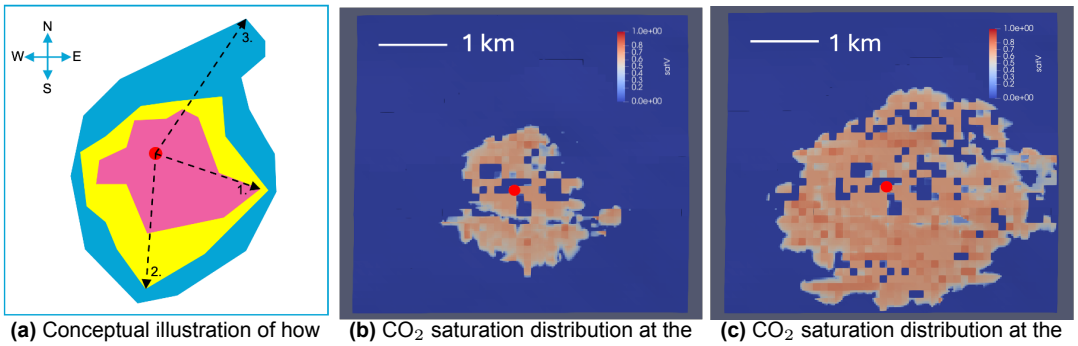
3.1.2 Maximum Plume Extent

In addition to injection performance, the study investigated variability in plume migration behavior across the ensemble. From both a regulatory and safety perspective, understanding the spatial extent and direction of CO_2 migration is critical in CCS operations. Uncontrolled plume propagation may increase the risk of leakage through faults, abandoned wells, or interactions with other subsurface uses such as potable aquifers [46, 47, 48].

To capture plume migration, the maximum plume extent was quantified as the furthest migration distance from the injection well. The plume was defined as the set of grid cells with a CO_2 saturation greater than 0.05. At the end of each simulation year, the furthest boundary cell relative to the injector

was identified in each of the eight cardinal and intercardinal directions (N, NE, E, SE, S, SW, W, NW). This directional breakdown enables targeted analysis where migration risks are spatially dependent, such as toward a known fault. Additionally, the absolute maximum extent, irrespective of direction, was recorded annually, providing a concise measure for comparing plume behavior across models.

Figure 3.1(a) provides a conceptual illustration of this procedure, where the plume areas after 1, 2, and 3 years are shown in pink, yellow, and blue, respectively. The arrows mark the maximum distances from the injector to the plume boundary in each year. Figure 3.1(c) shows an example of the CO2 saturation distribution for one realization after 12 years to give the reader a sense of how these plumes look.



(a) Conceptual illustration of how maximum plume extent is quantified. The colored regions show plume areas after 1, 2, and 3 years; arrows mark the maximum distances from the injector (red dot) to the plume boundary each year.

(b) CO_2 saturation distribution at the top of the reservoir after 6 years of injection for model realization FM2_CO3_TS2_OBJ_Mult1, Ensemble 1. The red dot marks the injector location.

(c) CO_2 saturation distribution at the top of the reservoir after 12 years of injection for model realization FM2_CO3_TS2_OBJ_Mult1, Ensemble 1. The red dot marks the injector location.

Figure 3.1: Comparison of (a) a conceptual representation of maximum plume extent quantification with (b) and (c) showing actual simulated saturation fields from open-DARTS after 6 and 12 years.

3.1.3 Plume Areal Coverage

Plume areal coverage characterizes the overall footprint of the CO_2 plume in the reservoir. It is computed annually by summing the surface areas of all grid cells with a CO_2 saturation greater than 0.05, where each cell area is given by $\Delta x \times \Delta y$ (25×25 m^2). This yields the total plume footprint in square meters, which is then converted to square kilometers. The metric provides insight into reservoir utilization and supports the design of spatially appropriate monitoring strategies during storage operations.

It should be noted that this approach may underestimate the effective coverage in cases where isolated cells fall below the saturation threshold despite being surrounded by plume cells as can be see in figure 3.1(c). However, for the purposes of this study, this simplification is considered acceptable, as it still provides a consistent and comparative measure of plume spread across the ensemble.

3.2 Simulation Setup and Model Assumptions

Two types of flow simulations were used in this study: full-physics flow simulations with open-DARTS (Section 2.3) and streamline-based simulations with 3DSL (Section 2.4). They serve different roles: full-physics runs provide the reference storage metrics for benchmarking reduced ensembles and generate flow diagnostics that serve as flow-based distance metrics for clustering, whereas streamline simulations are used solely to derive those flow diagnostics for clustering.

3.2.1 Full-Physics Simulations (open-DARTS)

Multiphase CO_2 sequestration was simulated with open-DARTS under fixed bottomhole pressure (BHP) control with a single injector, as described in Section 2.2. The injector was represented as a multi-segmented well (MSW), and BHP control was applied at the top perforation (i.e., the shallowest perforated cell / top MSW segment). The injection temperature was set to 300 K, and each realization was run for 20 years.

For each realization, the target BHP was defined relative to the hydrostatic reservoir pressure at the

top perforation. A hydrostatic gradient of 97.5 bar/km with a 1 bar surface reference was used, after which a constant offset of +10 bar was added to maintain a pressure margin

$$p_{\mathrm{BHP}} = 1 + \frac{97.5}{1000}\, z_{\mathrm{top}} + 10 \quad [\mathrm{bar}], \tag{3.1}$$

where z_{top} is the depth (m; positive downward) of the top perforation. The wellbore radius was 0.075 m.

This setup keeps initial injection conditions consistent across realizations with different structural depths. The specific model assumptions are described in Section 2.3, and the numerical settings are summarized in Table 3.1.

Table 3.1: Time-stepping and solver parameters used for open-DARTS runs.

Parameter	Value	Unit
Simulation time	20	years
First timestep	1×10^{-5}	days
Maximum timestep	10	days
Timestep multiplier	1.3	–
Max Newton iterations	20	–
Newton tolerance	1×10^{-3}	–
Max linear iterations	50	–
Linear tolerance	1×10^{-5}	–

3.2.2 Streamline Simulations (3DSL)

The 3DSL simulator was employed to compute flow diagnostics that served as flow-based distance metrics for quantifying dynamic (dis)similarity between geological realizations. As described in Section 2.4, both the two-phase immiscible formulation and the simplified single-phase formulation were considered.

Single-Phase Formulation

The single-phase formulation neglects multiphase interactions and treats fluids as incompressible with identical properties as shown in Section 2.4.2. In this case, the flow field is governed solely by permeability distribution and boundary conditions, with streamlines remaining fixed over time. This isolates the role of geological heterogeneity (e.g. permeability architecture, connectivity, and faults) without the added complexity of multiphase displacement.

Although it does not account for multiphase flow effects, the single-phase model remains relevant for CO_2 sequestration screening because it captures the first-order influence of permeability architecture on both injection rates and plume pathways. It could therefore offer a fast and informative diagnostic for pre-screening realizations.

For the single-phase streamline screening, the assumptions and parameters are summarized in Table 3.2. The bottomhole pressure (BHP) schedule was identical to that used in the full-physics simulations (see Eq. 3.1); a single-well model was used for the injector. With fixed controls, the pressure field is steady and accordingly, the pressure equation was solved once at $t = 0$, after which transport was advanced along fixed streamlines over the 20-year horizon. Although all fluids share identical properties, displacement was tracked by assigning distinct diagnostic labels (tracers) to the resident (oil) and injected (water) fluids. These labels do not affect the single-phase flow solution but enable computation of time-of-flight and partitioned volumes for the flow-based distance metrics. To ensure a clear distinction between resident and injected fluid, the oil–water contact (OWC) was set at a depth of 2000 m, resulting in a fully oil-saturated reservoir into which water was injected.

Immiscible Two-Phase Formulation

The immiscible two-phase formulation represents CO_2 and brine as distinct phases with different mobilities governed by relative permeability curves, while neglecting dissolution and capillary pressure, as described in Section 2.4.1. As saturations evolve, mobility ratios change, and streamlines are retraced, allowing displacement patterns to adapt dynamically. This enables the model to capture the combined effects of heterogeneity, mobility contrast, and buoyancy (through inclusion of gravity) on injection behavior and plume migration. For the immiscible runs, the same BHP schedule as in the full-physics simulations was applied (see Eq.3.1); a single-well model was used for the injector. Facies-dependent relative permeability curves were assigned consistent with the full-physics simulations (Table 2.2, Appendix 2.2). To ensure a clear distinction between resident and injected fluid, the oil–water contact (OWC) was set at a depth of 1000m, resulting in a fully brine-saturated reservoir into which CO_2 was injected. Additional assumptions and parameters are summarized in Table 3.2

To account for evolving saturation fronts, 11 pressure solves were scheduled at 0, 1, 3, 10, 25, 50, and 100 days, and at 1, 5, 10, and 15 years. This allowed streamline patterns to evolve over time while keeping the computational cost far below that of the full-physics simulations. Although simplified, the immiscible formulation provides a more realistic screening framework than the single-phase case, as it reproduces first-order multiphase effects relevant to CO_2 sequestration, particularly rate attenuation under fixed BHP and plume migration influenced by heterogeneity and buoyancy. The resulting flow diagnostics, which incorporate multiphase flow and buoyancy effects, served as distance metrics for clustering.

Single-phase formulation

Parameter	Value	Unit
Datum depth	1555	m
Datum pressure	155.55	bar
Oil–Water Contact (OWC)	2000	m
Gas–Oil Contact (GOC)	500	m
Viscosity	1.0	cP
Density	1019	kg/m^3

Two-phase formulation

Parameter	Value	Unit
Datum depth	1555	m
Datum pressure	155.55	bar
Oil–Water Contact (OWC)	1000	m
Gas–Oil Contact (GOC)	500	m
Viscosity (gas / water)	0.07 / 1.0	cP
Density (gas / water)	810 / 1019	kg/m^3

Table 3.2: Key input parameters for the 3DSL streamline formulations (single-phase and two-phase).

3.3 Flow-Based Distance Metrics

As demonstrated by Scheidt and Caers [6, 7, 8], it is possible to parameterize the spatial variability of a large ensemble of geostatistical realizations, and the effect this variability has on the dynamic reservoir response by defining a distance function that measures the 'dissimilarity' between any two models. Their study showed that using a single flow-based distance metric to quantify the difference between realizations enables the identification of models that exhibit similar behavior in terms of reservoir response. This, in turn, allows for the grouping of similar realizations and the selection of a representative subset that spans the diversity of the ensemble.

By clustering models based on such distances as explained in Section 2.1, one can significantly reduce the number of full-physics simulations required, while still capturing the essential variability in reservoir behavior. This approach supports the reconstruction of the ensemble's probabilistic distribution using only a small, well-chosen subset of distinct realizations, thereby offering substantial computational savings.

The choice of flow-based distance metric is critical to the success of this strategy. Several options for defining distances will be outlined in the following sections, tailored to the different CO_2 storage metrics studied in this work. In this study, such a flow-based distance metric is referred to as a flow diagnostic (FD), and its effectiveness depends on its ability to reflect the model outputs of interest. Specifically, for the distances to meaningfully separate realizations based on dynamic behavior, the chosen FD must be reasonably correlated with the corresponding full-physics output it aims to approximate, while being

computationally inexpensive, ideally requiring only a small fraction of the effort needed for full-physics simulation.

This study refers to diagnostics using shortcodes where SP = single-phase streamline, IMM = two-phase immiscible streamline, FP = full-physics; PS = pressure-solve index; D = days; SAT = saturation field; PRD = production.

3.3.1 Flow Diagnostics for Injection Rate

To approximate the ensemble behavior of injection rates, three FDs were tested. Each was selected for its low computational cost relative to full-ensemble simulation, while being expected to correlate with the full-physics injection response:

1. Injection rate after the first pressure solve of the streamline simulator under the single-phase formulation (SP-PS1).
2. Injection rates extracted from the two-phase immiscible streamline simulations after the first pressure solve (day 1), fourth pressure solve (day 25), and eleventh pressure solve (year 20) (IMM-PS1, IMM-PS4, IMM-PS11).
3. Injection rates extracted from the full-physics simulator at day 1, day 10, and day 100 (FP-1D, FP-10D, FP-100D).

3.3.2 Flow Diagnostics for Maximum Plume Extent

To approximate the ensemble behavior of maximum plume migration, three FDs were tested:

1. CO_2 saturation fields after 20 years (1 pressure solve), obtained from the single-phase simulations (SP-SAT-PS1).
2. CO_2 saturation fields after 20 years (11 pressure solves), obtained from the two-phase immiscible streamline simulations (IMM-SAT-PS11).
3. Cumulative production responses from eight equally spaced production wells, placed on a circle of radius r centered on the injector and evaluated after 20 years (immiscible formulation only) (8-PRD-IMM-PS11).

For the production-well diagnostic, each well was constrained to a fixed low rate of 1 m^3/day. This choice was intended to minimize perturbations of the streamline solution, while still providing a diagnostic signal of plume arrival and migration directionality across realizations. In addition, r is chosen per injector location (Ensembles 1 and 2; Section 2.2) as the largest radius that fits entirely within the model domain so that the eight production wells form a full circle inside the grid. Thus, r differs between the two well locations but is fixed across realizations for a given location.

3.3.3 Flow Diagnostics for Plume Areal Coverage

To approximate the ensemble behavior of plume areal coverage, two FDs were tested:

1. CO_2 saturation fields after 20 years (1 pressure solve), obtained from the single-phase simulations (SP-SAT-PS1).
2. CO_2 saturation fields after 20 years (11 pressure solves), obtained from the two-phase immiscible streamline simulations (IMM-SAT-PS11).

3.3.4 Distance Matrix

To compare geological realizations based on flow diagnostics, pairwise distances between all models were computed. These values form the entries of a distance matrix (Figure 3.2), which is the basis for subsequent dimensionality reduction and clustering as discussed in Section 2.1. Two distance definitions were applied, depending on whether the FD was time-dependent (rates) or spatial (saturation fields).

	A	B	C	D	E
A	D_{AA}	D_{AB}	D_{AC}	D_{AD}	D_{AE}
B	D_{BA}	D_{BB}	D_{BC}	D_{BD}	D_{BE}
C	D_{CA}	D_{CB}	D_{CC}	D_{CD}	D_{CE}
D	D_{DA}	D_{DB}	D_{DC}	D_{DD}	D_{DE}
E	D_{EA}	D_{EB}	D_{EC}	D_{ED}	D_{EE}

Figure 3.2: Schematic example of a distance matrix D_{ij}, where each entry represents the dissimilarity between realizations i and j. Diagonal elements (D_{ii}) are zero, while off-diagonal values quantify differences with respect to the diagnostic.

Time-Series Distance (Rates)

For dynamic responses such as injection or production rates, pairwise distances between realizations A and B were computed using a weighted root-mean-square (RMS) mismatch

$$D_{AB} = \sqrt{\sum_{k} \frac{1}{n_w} \sum_{w=1}^{n_w} \frac{\sum_{i=1}^{n_t} \Delta t_{i,w,k} \left(Y^{A}_{i,w,k} - Y^{B}_{i,w,k}\right)^2}{\sum_{i=1}^{n_t} \Delta t_{i,w,k}}}, \tag{3.2}$$

where k is the variable index (e.g., rate type), w the well index, i the time index, and $\Delta t_{i,w,k}$ the time step length. This formulation ensures temporal differences are properly weighted and that distances are averaged consistently across wells.

Spatial Distance (Saturation Fields)

For static or grid-based diagnostics such as CO_2 saturation fields, the following L_2 norm was used

$$D_{AB} = \sqrt{\sum_{b=1}^{n_b} \left(Y^{A}_{b} - Y^{B}_{b}\right)^2}, \tag{3.3}$$

where $b = 1, \ldots, n_b$ indexes the grid blocks, and Y^{A}_{b}, Y^{B}_{b} denote the saturation values in block b for realizations A and B. This metric quantifies the overall spatial mismatch in plume distribution at a given snapshot in time (e.g., after 20 years of injection).

3.4 Dimensionality Reduction and Clustering

This section explains how flow-based dissimilarities are embedded into low-dimensional spaces and subsequently clustered to select representative realizations. Three dimensionality-reduction methods (MDS, KPCA, t-SNE) are evaluated, and procedures for subset selection using k-means, and, in limited cases, k-medoids, are outlined. Internal clustering metrics used to choose the number of clusters are introduced, and the overall workflow is summarized.

3.4.1 Dimensionality Reduction Techniques

To visualize and cluster the ensemble of models based on pairwise dissimilarities derived from the flow diagnostics, several dimensionality reduction (DR) methods were tested. The goal was to obtain clear separation of realizations in a low-dimensional space (typically two dimensions) while preserving the structure of the original distance matrix. Three DR techniques were selected, namely Multidimensional Scaling (MDS), Kernel Principal Component Analysis (KPCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE).

MDS and KPCA were chosen because they were successfully applied by Scheidt and Caers [6, 7, 8] for flow-based clustering problems. In their work, KPCA was applied after MDS in a sequential fashion to further linearize the embedded space, whereas in this study the two methods are applied independently, as described in Section 2.1. t-SNE was included to investigate whether a method tailored to emphasize local neighborhood structure could reveal additional patterns within the ensembles.

Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) is a classical dimensionality reduction method that aims to preserve the pairwise distances between models in the projected space. It constructs a configuration of points in low dimensions such that the distances between them closely match the dissimilarities in the original space. MDS is useful for identifying global structure but can be sensitive to noise and nonlinearity in the data [49].

Given a set of n objects with dissimilarities d_{ij}, the squared distance matrix is first constructed as

$$D^{(2)} = \left[d_{ij}^2\right]_{i,j=1}^{n}. \tag{3.4}$$

To recover coordinates from dissimilarities, distances are transformed into inner products via *double centering*. Let

$$J = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \tag{3.5}$$

where I_n is the $n \times n$ identity matrix and $\mathbf{1}$ is a vector of ones. The centered Gram matrix is then

$$B = -\frac{1}{2} JD^{(2)}J. \tag{3.6}$$

The matrix B represents inner products between objects in the embedded space. By eigen-decomposition,

$$B = V\Lambda V^\top, \tag{3.7}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ are eigenvalues in descending order and V contains the corresponding eigenvectors. Selecting the top k positive eigenvalues, the low-dimensional embedding is obtained as

$$X_k = V_k \Lambda_k^{1/2}, \tag{3.8}$$

where $V_k \in \mathbb{R}^{n\times k}$ contains the first k eigenvectors and $\Lambda_k^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$. The rows of X_k provide the k-dimensional coordinates of the embedded objects, preserving the geometry implied by the dissimilarities.

The dimensionality k is determined by inspecting the proportion of variance explained by the positive eigenvalues of B. A common criterion is to select the smallest k such that the cumulative explained variance exceeds a specified threshold, here set to 0.95

$$\mathrm{Explained}(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} > 0.95, \tag{3.9}$$

where λ_i are the positive eigenvalues of B. For instance, if the first three eigenvalues account for more than 95% of the total variance, then $k = 3$ is considered sufficient.

Kernel Principal Component Analysis (KPCA)

In this study, KPCA is applied directly to the original distance matrix. KPCA extends the standard PCA technique by using kernel functions to project data into a higher-dimensional feature space before applying PCA. This allows KPCA to capture nonlinear relationships in the input data, making it suitable for detecting curved or clustered manifolds that linear PCA would miss.

Given a dataset $\{x_1, \ldots, x_n\}$, KPCA first computes a kernel matrix [50]

$$K_{ij} = k(x_i, x_j), \tag{3.10}$$

where $k(\cdot, \cdot)$ is a positive semi-definite kernel function. In this study, the radial basis function (RBF) kernel is employed, defined as

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{3.11}$$

where σ is the kernel width parameter.

The kernel matrix K is centered and subjected to the eigenvalue problem

$$Ke_i = \lambda_i e_i, \tag{3.12}$$

where λ_i are the eigenvalues in descending order and e_i the corresponding eigenvectors. The projection of a data point x onto the i-th principal component in the feature space is then given by

$$z_i(x) = \sum_{j=1}^{n} e_{ij}\, K(x_j, x). \tag{3.13}$$

The dimensionality k is determined by the cumulative variance explained by the eigenvalues of K. Consistent with standard practice, the smallest k satisfying

$$\text{Explained}(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} > 0.95 \tag{3.14}$$

is selected, ensuring that at least 95% of the variance in the feature space is preserved.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction method developed by van der Maaten in 2008 [51], designed primarily for visualizing high-dimensional data. It extends Stochastic Neighbor Embedding (SNE) by introducing a symmetrized cost function and a heavy-tailed Student-t distribution in the low-dimensional space, which mitigates the so-called ”crowding problem”.

Given a dataset $\{x_1, \ldots, x_n\}$ in high-dimensional space, t-SNE converts pairwise distances into probabilities that represent similarities. The conditional probability of point x_j being a neighbor of x_i is defined as

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}, \tag{3.15}$$

where σ_i controls the local bandwidth around x_i. To ensure symmetry, the joint probabilities are defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \qquad p_{ii} = 0. \tag{3.16}$$

In the low-dimensional embedding, similarities between points y_i and y_j are modeled using a Student-t distribution with one degree of freedom

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}, \qquad q_{ii} = 0. \tag{3.17}$$

The mismatch between high-dimensional similarities p_{ij} and low-dimensional similarities q_{ij} is minimized by reducing the Kullback–Leibler divergence

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{3.18}$$

The gradient of the cost with respect to y_i is

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}, \tag{3.19}$$

which can be interpreted as attractive and repulsive forces: nearby points (large p_{ij}) are pulled together, while dissimilar points that are too close are pushed apart.

Optimization is performed using gradient descent with momentum, along with heuristics such as early exaggeration, which temporarily increases the p_{ij} values to encourage separation of clusters. The main user-defined parameter is the perplexity, which controls the effective number of neighbors by determining each σ_i and was set at 30 for this study.

t-SNE therefore provides a probabilistic embedding that preserves local neighborhood structure and is particularly effective at revealing cluster separation in high-dimensional datasets. However, the relative distances between well-separated clusters in the low-dimensional map should be interpreted cautiously, as they may not correspond to their true dissimilarities in the original space.

Summary

Table 3.3 summarizes the main strengths and limitations of the dimensionality reduction techniques discussed above.

Table 3.3: Comparison of dimensionality reduction techniques evaluated in this study.

Method	Strengths / Motivation	Limitations
MDS	Preserves global pairwise distances; easy to interpret.	Assumes Euclidean geometry; sensitive to noise and nonlinearity.
KPCA	Captures nonlinear relationships via kernel trick; retains variance explanation like PCA.	Embedding depends on kernel and bandwidth; global distances reflect similarity in the kernel space, not always intuitive in the original space.
t-SNE	Optimized for local neighborhoods; excellent for revealing tight clusters and complex manifolds.	Does not preserve global distances; stochastic results; requires user-defined hyperparameters (e.g., perplexity).

3.4.2 Clustering and Representative Model Selection

K-Means

Following dimensionality reduction, the projected feature space is partitioned using the K-means clustering algorithm. K-means minimizes the within-cluster variance by iteratively assigning points to the nearest centroid and updating centroid positions until convergence [52]. Formally, it solves

$$\min_{\{C_k\}_{k=1}^K} \sum_{k=1}^{K} \sum_{x_i \in C_k} \left\|x_i - \mu_k\right\|^2 \tag{3.20}$$

where C_k denotes the k-th cluster and μ_k its centroid.

To reduce sensitivity to initialization, the K-means++ seeding algorithm [53] was adopted, selecting diverse starting centroids to improve convergence and stability.

Because centroids are abstract points in the reduced space rather than actual models, a representative realization is identified for each cluster by selecting the model closest to its centroid. These representatives form a reduced subset of the ensemble, which is then used to reconstruct percentile curves and benchmark ensemble behavior.

K-means (with K-means++ initialization) is computationally efficient and well-suited for large datasets, providing a simple and effective means of grouping behaviorally similar models. Its main limitations are the assumption of roughly spherical cluster shapes and the lack of a probabilistic interpretation, but for this study its speed and straightforward representative selection outweighed these drawbacks.

K-Medoids

Finally, it should be noted that this study primarily applied dimensionality reduction followed by K-means clustering to identify representative models. As an alternative, the K-medoids algorithm can be applied directly to the pairwise distance matrix without prior dimensionality reduction. In contrast to K-means, which defines abstract centroids, K-medoids selects actual data points (medoids) as cluster centers, ensuring that representatives correspond to realizations in the original ensemble. This property makes K-medoids more robust to outliers and can be more interpretable because the clusters are represented by data examples albeit at the cost of increased computational effort [52]. K-medoids was not adopted as one of the primary workflows in this study due to the higher computational cost of K-medoids compared to the other methods considered. Nevertheless, it was applied in a limited number of cases to evaluate its performance relative to the proposed approaches described previously.

3.4.3 Internal Metrics for Cluster Count Selection

A central challenge in ensemble reduction is determining how many clusters, and thus how many representative models, should be selected. In practical applications, this decision cannot be guided by direct comparison with the full ensemble, since running all computationally expensive full-physics reservoir simulations is precisely what ensemble reduction aims to avoid. It is therefore essential to rely on robust internal clustering metrics that assess cluster quality without requiring external reference data.

Several well-established internal metrics were considered in this work. They provide complementary perspectives on cluster compactness and separation and are widely used for clustering validation.

Inertia (Within–Cluster Sum of Squares)

For clusters C_i with centroids μ_i, inertia is

$$W = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{3.21}$$

Smaller values of W indicate that samples lie close to their cluster centers, i.e., the clusters are tight. Because W decreases monotonically as the number of clusters k grows, an “elbow” in the curve $W(k)$ is typically sought to choose a suitable k.

Davies–Bouldin Index (DB)

For k clusters with centroids c_i and scatter $\sigma_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i\|$, let $d_{ij} = \|c_i - c_j\|$ be the distance between centroids. The index is

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d_{ij}} \tag{3.22}$$

Lower DB values mean that clusters are compact and well separated from each other, while higher values suggest that clusters overlap or lack clear structure.

Silhouette Score

For each sample i, let $a(i)$ be the mean distance to all other points in its own cluster, and $b(i)$ the smallest mean distance to any other cluster. The silhouette coefficient is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}}, \qquad -1 \le s(i) \le 1 \tag{3.23}$$

and the overall silhouette score is the mean of $s(i)$ over all samples. Scores close to $+1$ indicate well-defined, cohesive clusters; values near 0 indicate overlapping clusters, and negative values imply that some samples may be assigned to the wrong cluster.

Calinski–Harabasz Index (CH)

Let B and W denote the between-cluster and within-cluster dispersion matrices

$$B = \sum_{i=1}^{k} n_i(\mu_i - \bar{x})(\mu_i - \bar{x})^\top, \qquad W = \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^\top \tag{3.24a}$$

where $\bar{x}$ is the overall mean. The CH index is

$$CH = \frac{\mathrm{tr}(B)}{\mathrm{tr}(W)} \cdot \frac{n - k}{k - 1} \tag{3.25}$$

Higher CH values indicate that clusters have high between-cluster dispersion and low within-cluster variance, i.e., they are compact and well separated.

These metrics were computed for each candidate number of clusters across the approaches considered (MDS+KMeans, KPCA+KMeans, and t-SNE+KMeans), and their values were examined to determine whether they could support the selection of an appropriate cluster count.

3.4.4 Overview of Proposed Workflows

This section provides a concise overview of the workflows developed for different CO_2 storage performance metrics. Table 3.4 lists the flow diagnostics considered for each metric (see Section 3.3 for details), while Figure 3.3 summarises how these diagnostics are transformed into distance measures, clustered, and used to evaluate reduced ensembles.

Table 3.4: Flow diagnostics considered for each CO_2 storage metric.

Storage Metric	Flow Diagnostics
Injection Rate	(a) SP-PS1 (b) IMM-PS1, IMM-PS4, IMM-PS11 (c) FP-1D, FP-10D, FP-100D
Maximum Plume Extent	(a) SP-SAT-PS1 (b) IMM-SAT-PS11 (c) 8-PRD-IMM-PS11
Plume Areal Coverage	(a) SP-SAT-PS1 (b) IMM-SAT-PS11

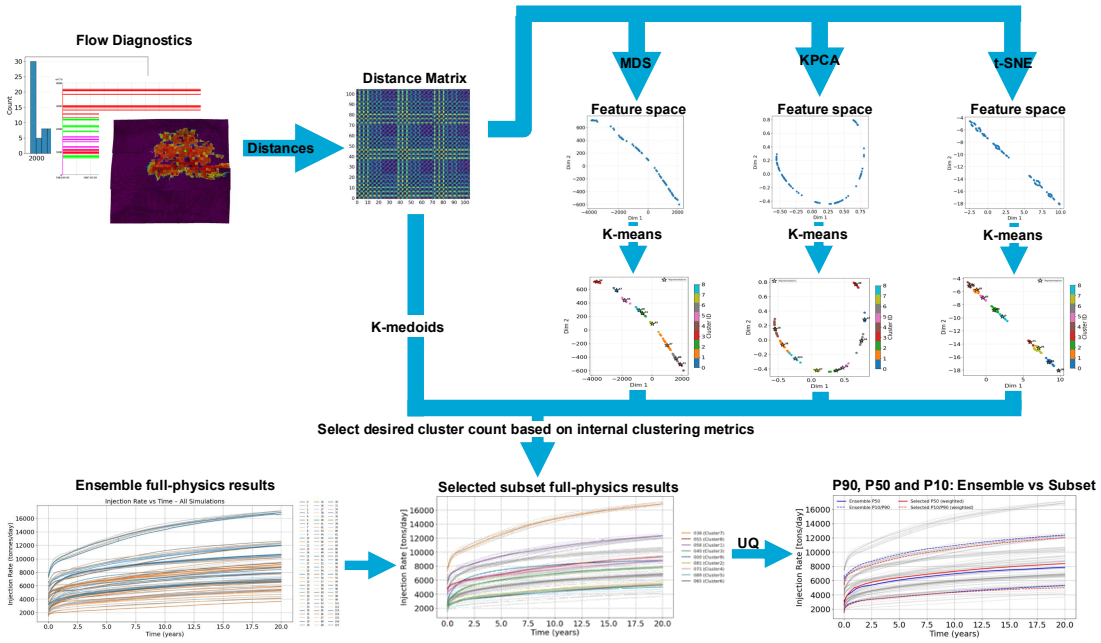


Figure 3.3: Workflow for uncertainty quantification and ensemble reduction. Flow diagnostics are computed for each storage metric and converted into pairwise distances, from which distance matrices are constructed. These distances are either embedded into low-dimensional spaces (MDS, KPCA, t-SNE) and clustered with k-means, or clustered directly with K-medoids. The number of clusters is determined using internal cluster validation metrics. The precomputed full-physics results of the selected representative realizations form the reduced subset; its reservoir responses are used to compute P_{90}, P_{50}, and P_{10} percentiles, which are then compared with the "true" percentiles from the full ensemble.

3.5 Performance Metrics

To evaluate how accurately a selected subset of models can reproduce the behavior of the full ensemble, this section outlines the performance evaluation framework. It first defines how ensemble percentiles (P_{90}, P_{50}, and P_{10}) are computed and reconstructed for both the full ensemble and selected subsets. Then, it introduces the error metric used to quantify the differences between reconstructed and reference percentile curves.

3.5.1 Ensemble and Selected Subset Percentiles

The full ensemble of 108 geological models serves as the reference for evaluating uncertainty in CO_2 storage performance. For each studied storage metric, as described in Section 3.1, the P_{90}, P_{50}, and P_{10} percentile curves are computed across all realizations. These percentiles characterize the statistical spread of model behavior and are treated as the "true" ensemble reference.

To avoid ambiguity, it is important to clarify the percentile terminology used throughout this study. Here, P_{90} refers to the 10th percentile of the ensemble (i.e., the lower bound), and P_{10} refers to the 90th percentile (i.e., the upper bound). While this naming convention may appear inverted, it aligns with standard reservoir engineering practice, where the P_{90} value denotes a conservative estimate, indicating that there is a $90\ \%$ chance the actual performance will exceed this value. Similarly, P_{10} represents an optimistic bound, with only a $10\ \%$ chance that the true outcome will be higher.

For any given model selection method (e.g., distance-based clustering or GA-informed selection which will be explained in Section 3.7), a subset of representative models is chosen from the ensemble. The same percentile curves (P_{90}, P_{50}, and P_{10}) are then reconstructed based on the simulation results of only this reduced subset. These reconstructed percentiles are directly compared to the full ensemble to assess how well the subset captures the overall uncertainty range.

For the purpose of this study, two different percentile reconstruction methods are applied: an *unweighted* and a *weighted* approach.

Unweighted Percentile Reconstruction

In the unweighted reconstruction, each selected cluster representative contributes equally to the percentile calculation. Let $\{y_1(t), \ldots, y_m(t)\}$ denote the time-dependent responses of m selected models at time t. The empirical CDF is then constructed as

$$F_{\text{unweighted}}(y;t) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\, y_i(t) \le y\} \tag{3.26}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Percentiles $P_p(t)$ are obtained by taking the inverse CDF

$$P_p(t) = F^{-1}_{\text{unweighted}}(p), \qquad p \in [0, 100] \tag{3.27}$$

Weighted Percentile Reconstruction

In contrast, the weighted reconstruction accounts for the relative population size of the clusters from which the representative models were selected. Let n_i denote the size of the cluster represented by model i. Normalized cluster weights are then defined as

$$w_i = \frac{n_i}{\sum_{j=1}^{m} n_j}, \qquad \sum_{i=1}^{m} w_i = 1 \tag{3.28}$$

The weighted CDF is given by

$$F_{\text{weighted}}(y;t) = \sum_{i=1}^{m} w_i\, \mathbf{1}\{\, y_i(t) \le y\} \tag{3.29}$$

and weighted percentiles follow as

$$P_p^{(w)}(t) = F^{-1}_{\text{weighted}}(p) \tag{3.30}$$

In this formulation, representatives from larger clusters exert greater influence on the reconstructed percentiles, while those from smaller or anomalous clusters have proportionally less impact. This approach aims to improve the fidelity of reconstructed percentiles by better reflecting the underlying distribution of the full ensemble.

Although the concepts of clustering and the Genetic Algorithm are introduced in Section 3.7, it is important to clarify that the weighted reconstruction can only be applied when using models selected directly through cluster-based methods. Since the GA approach does not necessarily retain information about cluster populations during its selection process, weighted reconstruction is not applicable in that context.

3.5.2 Relative RMSE and Evaluation Strategy

To evaluate how accurately the reconstructed percentiles from a selected subset of cluster representatives reflect the true percentiles of the full ensemble, this study uses the *Relative Root Mean Square Error* ($\mathrm{RMSE}_{\mathrm{rel}}$) as the primary performance metric. In addition to reporting continuous error values, it is also useful to apply a threshold that distinguishes between accurate and inaccurate reconstructions. In this work, a threshold of $\mathrm{RMSE}_{\mathrm{rel}} \le 5\%$ was adopted as the criterion for accuracy. This level was chosen to provide a practical balance between strictness and interpretability: values below 5% are considered sufficiently close to the true ensemble percentiles, while larger deviations are flagged as potentially misleading.

To ensure that differences in simulation timestep spacing (arising from varying convergence behavior) do not bias the results, $\mathrm{RMSE}_{\mathrm{rel}}$ is computed in a time-weighted form, such that timesteps with longer durations contribute proportionally more to the total error.

The metric is calculated independently for each percentile of interest (P_{90}, P_{50}, and P_{10}) and is expressed as a percentage to allow comparison across output types and magnitudes.

The formal definition of the time-weighted relative RMSE is given by

$$\mathrm{RMSE}_{\mathrm{rel}} = 100 \cdot \sqrt{\frac{\sum\limits_{t\in\mathcal{V}} \left(\frac{s(t)-a(t)}{a(t)+\varepsilon}\right)^2 \Delta t(t)}{\sum\limits_{t\in\mathcal{V}} \Delta t(t)}} \tag{3.31}$$

In this expression, t denotes a discrete simulation timestep and $\mathcal{V} \subseteq T$ is the set of timesteps included in the evaluation window T. The functions $a(t)$ and $s(t)$ represent, respectively, the reference ensemble percentile (e.g. the full-ensemble P_{90}) and the percentile reconstructed from the reduced set at time t. $\Delta t(t)$ is the time interval between t and the following timestep, which acts as a weight so that longer intervals have greater influence. A small positive constant ε (here 10^{-6}) is added to $a(t)$ in the denominator to avoid division by zero.

In addition to the $\mathrm{RMSE}_{\mathrm{rel}}$, a Signed $\mathrm{RMSE}_{\mathrm{rel}}$ is computed. This variant retains the magnitude information of the relative RMSE but incorporates the sign of the time-weighted relative mean difference (RMD) (Eq. 3.32). It therefore provides insight not only into the overall error magnitude but also into whether the reconstructed percentiles systematically overestimate or underestimate the ensemble reference.

Relative Mean Difference (RMD)

The time-weighted Relative Mean Difference is defined as

$$\mathrm{RMD}_{\mathrm{rel}} = 100 \cdot \frac{\sum\limits_{t\in\mathcal{V}} \frac{s(t)-a(t)}{a(t)+\varepsilon} \Delta t(t)}{\sum\limits_{t\in\mathcal{V}} \Delta t(t)}. \tag{3.32}$$

3.5.3 Simulation Cost Estimation

The relative simulation costs of clustering-based model selection were calculated differently for streamline flow diagnostics and full-physics flow diagnostics, reflecting how the diagnostic overhead enters the workflow.

For streamline flow diagnostics, the diagnostic simulations are performed outside of the full-physics simulator. Consequently, the overhead of running the diagnostic for all realizations must be added to the cost of running the selected cluster representatives. The total cost is therefore

$$\mathrm{Cost}\% = 100\,\frac{t_{\mathrm{fd}}}{T_{\mathrm{full}}} + 100\,\frac{K}{N} \tag{3.33}$$

where N is the ensemble size, K the number of selected models run to full lifetime, t_{fd} the runtime per realization of the flow diagnostic, and T_{full} the runtime of one full-physics realization to full lifetime. For clarity, this study denotes overhead as

$$\mathrm{Overhead}_{\mathsf{FD}} = 100\,\frac{t_{\mathrm{fd}}}{T_{\mathrm{full}}}.$$

For full-physics flow diagnostics, the situation is reversed: all realizations must already be simulated up to the chosen diagnostic time window (e.g. 1, 10, or 100 days). This overhead is therefore shared across the ensemble and does not need to be repeated for the selected models. The effective cost of running K cluster representatives to full time is computed as

$$\mathrm{Cost}\% = \underbrace{\mathrm{Overhead}_{\mathsf{FD}}}_{\text{all realizations to FD time}} + \underbrace{\frac{K}{N}\left(100 - \mathrm{Overhead}_{\mathsf{FD}}\right)}_{\text{cluster representatives}} \tag{3.34}$$

Tables 3.5a and 3.5b summarize the per-realization overhead costs associated with streamline- and full-physics–based flow diagnostics, expressed in walltime (HH:MM:SS). These values enter the cost formulas defined above and determine whether the diagnostic overhead is added (streamline) or shared and subtracted (full physics).

Table 3.5: Per-realization simulation time and relative overhead, normalized to a full-physics 20-year run (FD-FULL).

(a) Streamline-based flow diagnostics

FD	Time (h:m:s)	Overhead [%]
SP-PS1	00:00:35	0.63
IMM-PS1	00:00:35	0.63
IMM-PS4	00:01:45	1.90
IMM-PS11	00:06:20	6.86

(b) Full-physics flow diagnostics

FD	Time (h:m:s)	Overhead [%]
FD-D1	00:00:36	0.65
FD-D10	00:01:10	1.26
FD-D100	00:03:44	4.04
FD-FULL	01:32:18	100.00

All full-physics simulations were executed on high-performance servers equipped with two AMD Rome 7742 processors (64 physical cores each, 128 in total) and two Nvidia A100 80 GB GPUs. Each run was restricted to eight CPU cores.

The streamline-based flow simulations were performed on a separate workstation equipped with an Intel® Core™ i7-14700K processor (20 cores / 28 threads) and 32 GB of RAM. Each run was restricted to four CPU cores, with no GPU acceleration required.

Although both workflows are compared in terms of relative simulation cost to provide a practical indication of computational efficiency, the results should be interpreted with care because they were obtained on different hardware with different levels of parallelism and GPU support.

3.6 Complementary Sensitivity Analysis of Parameter Effects: dGSA and Time-Weighted Partial η^2

To assess how input parameters shape ensemble behavior for the storage metrics of interest, this study combines a distributional, cluster-aware measure (dGSA) with a variance-based effect-size measure (time-weighted partial η^2). This pairing enables comparison of regime-separating power with uniquely explained variance over the full simulation lifetime and evaluation of whether the low-cost dGSA flags the same influential parameters identified by partial η^2. Both methods are outlined below.

3.6.1 Distance-based Generalized Sensitivity Analysis (dGSA)

To complement the clustering-based model reduction, this study applies distance-based Generalized Sensitivity Analysis (dGSA) to determine which uncertain input parameters most strongly influence how the ensemble separates into distinct behavioral regimes of injection-rate behavior. The method follows the framework of Fenwick et al. [54], which links clustering of model responses with statistical comparisons of the corresponding input-parameter distributions.

In this work, clustering is carried out on a suitably correlated flow diagnostic rather than on the complete injection-rate trajectories, enabling the identification of key drivers without simulating every realization over the full forecast horizon. Realizations are grouped into clusters of similar diagnostic values using a flow-based distance metric.

After clustering, dGSA evaluates, for each input parameter, how its distribution inside each cluster differs from its distribution across the full ensemble. The influence of a parameter is quantified by the maximum vertical gap between the empirical cumulative distribution functions (CDFs) of the parameter within the cluster and in the overall sample. To assess whether an observed gap is larger than could be expected from sampling noise, a bootstrap is applied: many random subsets of the same size as the cluster are drawn, and their CDF distances form a reference distribution. The observed distance divided by the 95$^{\text{th}}$ percentile of this bootstrap reference yields a normalized cumulative distance S; values of $S > 1$ imply that the parameter helps define the cluster beyond random variation.

In practical terms, dGSA provides insight that complements clustering. While clustering reduces the ensemble to a concise set of representative realizations that approximate ensemble percentiles, dGSA explains why different regimes arise by identifying the inputs most responsible for that separation. Here, this means highlighting which geological or conceptual parameters (fault configuration, fault transmissibility multiplier, conceptual model type, top structure, and cut-off values) exert the strongest influence on the system's behavior. These findings help direct uncertainty-reduction efforts toward the parameters that matter most.

Figure 3.4 provides a simplified illustration of how dGSA is interpreted. On the left, realizations are embedded in a low-dimensional feature space (here two-dimensional) obtained by applying one of the dimensionality-reduction techniques described in Section 2.1. Clustering is then applied to group nearby realizations, shown as dashed circles encircling three realizations each.

For this toy example two uncertain parameters are assumed: `FM` and `CO`, each with three possible levels (FM1 to FM3 and CO1 to CO3). Inspection of the clusters shows that the `FM` levels are spread almost evenly across clusters, whereas each cluster tends to contain a single dominant level of `CO`. This pattern suggests that `CO` is more influential in driving cluster separation than `FM`.

The dGSA bar plot on the right quantifies this observation. It shows the normalized cumulative distance score S for each parameter, where $S > 1$ indicates that the parameter's distribution inside clusters deviates more from the overall ensemble than expected by random sampling. Here, `CO` has $S > 1$ and is classified as influential, while `FM` remains below the threshold, indicating limited impact on the ensemble's regime separation.

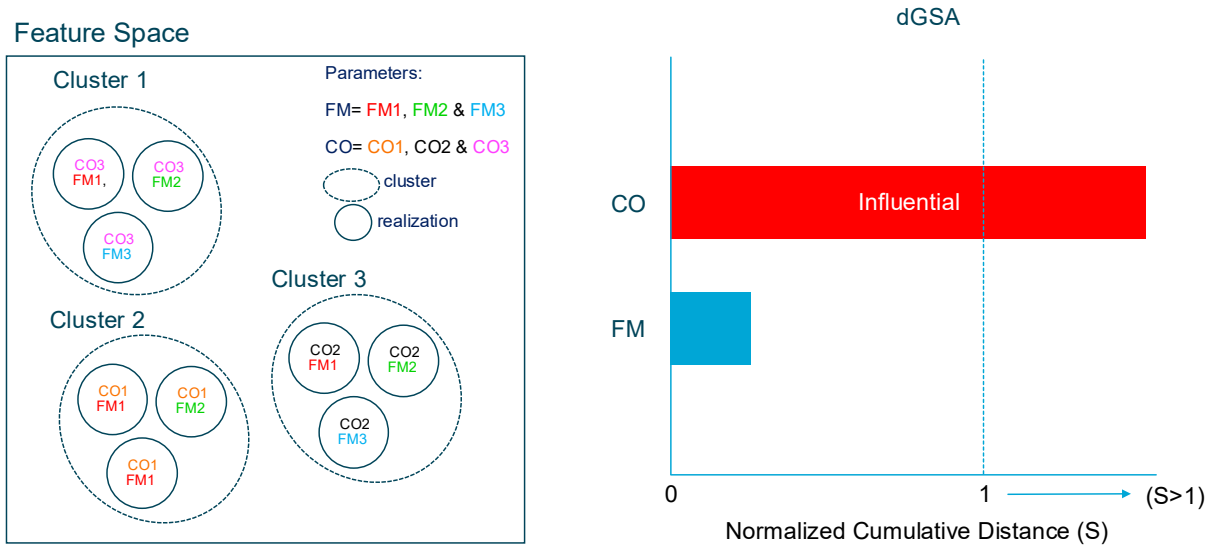


Figure 3.4: Simplified illustration of distance-based Generalized Sensitivity Analysis (dGSA). Left: realizations embedded in a low-dimensional feature space and grouped into clusters (dashed circles). Right: normalized cumulative distance scores (S) for the two example parameters (`CO`, `FM`); $S > 1$ indicates an influential parameter for cluster separation.

3.6.2 Variance–based Sensitivity Analysis via Time–Weighted Partial η^2

While dGSA may effectively highlight parameters that separate the ensemble into distinct regimes, its scores do not indicate how much of the total response variability each parameter uniquely explains. To complement and validate the dGSA findings, a variance–based sensitivity analysis using time–weighted partial η^2 is applied.

Partial η^2 is an effect–size metric that estimates how much of the variability in the injection–rate histories is uniquely attributable to one parameter after accounting for all others [55]. Each input (e.g., fault configuration, transmissibility multiplier, conceptual model type, top structure, cut-off value) is encoded as categorical variables. For every time step of the common simulation grid, two regression models are compared: one including the parameter of interest in addition to the other controls, and one without it. The increase in explained variance when the parameter is included is the per–time partial η^2.

Because the simulation time steps vary in length, the per–time values are averaged using weights proportional to the step durations. This produces a single time–weighted score $\overline{\eta}^2$ for each parameter, representing the fraction of total variance in the full injection–rate history that the parameter uniquely explains.

To assess whether an observed $\overline{\eta}^2$ could arise by chance, a permutation test is performed: the labels of the target parameter are randomly shuffled many times while all other parameters remain fixed, producing a null distribution of $\overline{\eta}^2$ values [56]. The proportion of permutations exceeding the observed value forms a p-value; small p (e.g., < 0.05) indicates a statistically significant unique effect.

As a practical guide, values of $\overline{\eta}^2 \gtrsim 0.10$ (with $p < 0.05$) are considered strongly influential, $[0.05, 0.10)$ moderately influential, and < 0.02 negligible.

3.7 Informed Genetic Algorithm Selection Approach

This section introduces a Genetic Algorithm (GA)–based model selection method, developed by this study, that leverages frequency information from multi-*k* clustering. The aim is to determine whether such a GA can reconstruct the key ensemble percentiles of CO_2 injection-rate behavior (P_{90}, P_{50}, P_{10}) as accurately as, or better than, the distance-based clustering strategy, while requiring similar or lower simulation cost. Because a GA applied directly to the full ensemble faces a very large combinatorial search space and may become trapped in suboptimal regions, this study restricts its search to a curated, frequency-ranked candidate pool.

The motivation arises from two observations. First, this study already uses early-time injection rates as flow diagnostics to enable distance-based clustering; these same early responses might also help guide a search-based selection method. Second, while clustering provides an unsupervised partitioning of the ensemble, a GA offers a way to directly optimize the subset for percentile reconstruction once a suitable candidate pool is defined.

3.7.1 Frequency Pool from Multi-*k* Clustering

To initialize the GA population with behaviorally relevant candidates, a frequency-ranked pool of model indices based on multi-*k* clustering results is created. Clustering is performed on t-SNE embeddings for several values of k (number of clusters). For each k, the representative model of each cluster is recorded. Let f_i denote the *selection frequency* of realization i, i.e., the number of times i is chosen as a cluster representative across all considered k values. Models are ranked by f_i, which is treated as a proxy for behavioral centrality and diversity. In practice, the GA draws exclusively from the top-ranked realizations (e.g., top-N by f_i or those exceeding a frequency threshold), forming a reduced and prioritized search pool.

This frequency pool serves two purposes: (1) it narrows the search space to high-potential candidates, improving GA efficiency and convergence, and (2) it embeds clustering-derived insights into the GA without requiring real-time clustering during optimization.

3.7.2 GA Optimization Procedure

The GA is used to optimize the selection of a fixed number of models from the frequency pool, targeting high performance in percentile reconstruction. A single GA variant is employed that restricts candidate

models to the frequency-ranked pool; no additional clustering is performed during optimization.

The GA iteratively refines model subsets through tournament selection, crossover, and mutation, while maintaining population diversity using controlled mutation probabilities and occasional random restarts. Each subset is evaluated based on its ability to reconstruct the P_{90}, P_{50}, and P_{10} percentiles of the full ensemble, computed only over the calibration period.

Calibration Periods.

GA optimization is repeated using truncated injection-rate histories of 10 days, 100 days, 200 days, 1 year, and 3 years to represent different levels of early information availability. Shorter calibration windows reduce simulation cost but may reduce long-term percentile reconstruction accuracy.

Fitness Function and Selection Strategy

For each generation, the GA evaluates candidate subsets using the average time-weighted relative RMSE between the reconstructed and reference P_{10}, P_{50}, and P_{90} profiles. These benchmarks are computed strictly over the calibration period. The objective is to minimize this RMSE while keeping the number of selected models fixed.

The fitness function used is

$$\text{Fitness} = \tfrac{1}{3}\left(\text{RMSE}_{\text{rel},P_{90}} + \text{RMSE}_{\text{rel},P_{50}} + \text{RMSE}_{\text{rel},P_{10}}\right) \tag{3.35}$$

The relative RMSE is defined in Section 3.5.

Simulation Cost Considerations

Simulation cost for the GA approach is estimated using the same methodology as for full-physics flow diagnostics described in Section 3.5.3, accounting for both the calibration simulations and the selected long-term runs.

4

Results

This chapter presents the results of the ensemble-reduction workflows applied to key CO_2 storage metrics. Section 4.1 introduces the full-ensemble percentiles for injection rate, maximum plume extent, and plume areal coverage, which serve as the reference for evaluating reduced subsets. Section 4.2 examines the impact of weighted versus unweighted percentile reconstruction. Section 4.3 presents injection-rate percentile reconstruction results, based on both streamline diagnostics and early-time full-physics simulations. Section 4.4 discusses injection-rate reconstruction when the number of selected clusters is determined by internal cluster validation metrics. Section 4.5 reports the results of the η and distance-based generalized sensitivity analysis performed on the injection-rate profiles and clusters. Section 4.6 evaluates a calibration-informed genetic algorithm as an alternative approach for reconstructing injection-rate percentiles. Finally, Sections 4.7 and 4.8 present the results for plume migration and plume areal coverage, completing the assessment of accuracy and efficiency across the different CO_2 storage metrics analyzed in this study.

4.1 Ensemble Percentiles

To establish a reference for subsequent ensemble-reduction analyses, the full ensemble of 108 geological realizations was simulated using open-DARTS (Section 2.3). These simulations provide benchmark distributions of key CO_2 storage metrics against which reduced subsets are evaluated. Specifically, injection rate, maximum plume extent, and plume areal coverage are analyzed to characterize injectivity and plume-migration behavior across the ensemble, as outlined in Section 3.1.

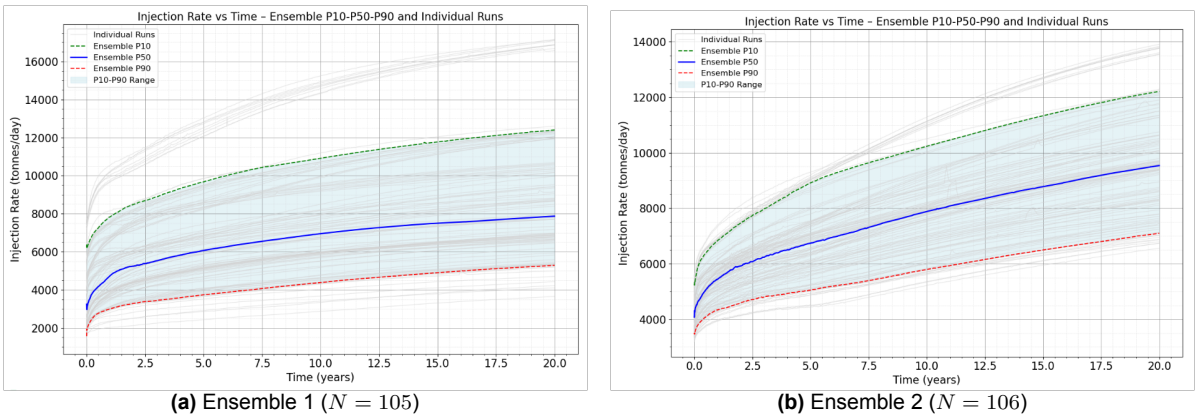
Before assessing the performance of the proposed workflows for the different storage metrics, the overall ensemble results are first presented. The true ensemble percentiles (P_{90}, P_{50}, and P_{10}) of the three CO_2 storage metrics are shown to provide context and establish a baseline for interpreting the subsequent results.

Note on ensemble sizes. Although the original geomodel ensemble comprised 108 geological realizations, convergence issues prevented a few runs from completing: in Ensemble 1, three realizations did not converge, and in Ensemble 2, two did not converge, reducing the effective sample sizes to $N_1 = 105$ and $N_2 = 106$. Here, N_i denotes the number of realizations used for Ensemble i after excluding non-converged runs. “Ensemble 1” and “Ensemble 2” refer to the two different well locations used in this study, as described in Section 2.2; the locations are shown in Figure 2.3. Appendix B, Table B.1 lists the realizations that did not converge for both ensembles.

4.1.1 Injection Rate

Figures 4.1(a) and 4.1(b) present the full-physics injection-rate results over a 20-year period for Ensembles 1 and 2. All individual simulation profiles are shown as light grey curves in the background. From these results, ensemble percentiles were derived: P_{90} (lower dashed red curve), P_{50} (solid blue curve), and P_{10} (upper dashed green curve), computed pointwise in time, as described in Section 3.5.1. This study uses the exceedance convention, where P_x is the value exceeded by $x\%$ of realizations; therefore

(P_{90} < P_{50} < P_{10}) for injection rates, with (P_{90}) serving as a conservative lower bound and (P_{10}) as an optimistic upper bound. These percentile curves provide the benchmark against which reconstructed percentiles for selected subsets will be evaluated later in terms of (signed) relative RMSE.



(a) Ensemble 1 ($N = 105$) **(b)** Ensemble 2 ($N = 106$)

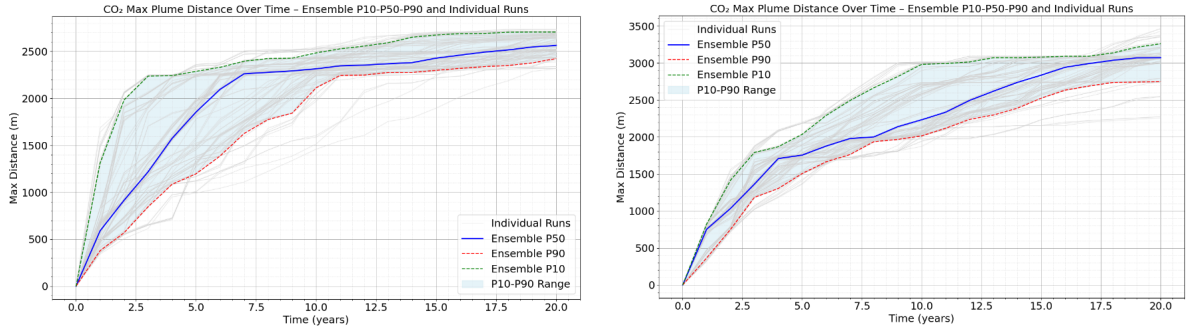
Figure 4.1: Injection-rate percentiles over the 20-year simulation window for both Ensemble 1 and 2. Thin grey curves show full-physics simulation results for individual realisations; bold curves are the pointwise-in-time ensemble percentiles (P_{50} solid blue; P_{10} and P_{90} dashed green and red respectively). Shading indicates the P_{10}–P_{90} band.

4.1.2 Maximum Plume Extent

Figures 4.2(a) and 4.2(b) show the maximum plume extent over 20 years for Ensembles 1 and 2, with ensemble percentiles derived from the full set of simulations. The maximum plume extent is computed following the procedure in Section 3.1.2, which tracks this metric over time for each realization.

In Ensemble 1, the upper bound (P_{10}) rises quickly during the first three years and then levels off, while P_{50} and P_{90} continue to grow. As a result, the percentile band narrows after about ten years, indicating a tighter overall distribution. Ensemble 2 shows a similar pattern, with P_{10} flattening around ten years and the other percentiles gradually approaching it.

This plateau reflects the finite spatial extent of the reservoir in this study: once the CO_2 plume reaches structural boundaries, further outward migration can no longer be tracked, which limits growth of the maximum plume extent over the remaining simulation period.

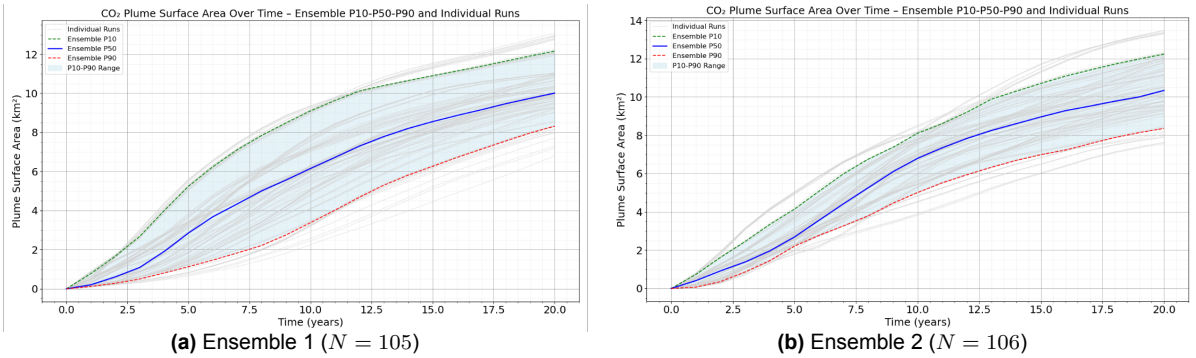


(a) Ensemble 1 ($N = 105$) **(b)** Ensemble 2 ($N = 106$)

Figure 4.2: Maximum plume extent percentiles over the 20-year simulation window for Ensembles 1 and 2. Thin grey curves show full-physics simulation results for individual realisations; bold curves are the pointwise-in-time ensemble percentiles (P_{50} solid blue; P_{10} and P_{90} dashed green and red respectively). Shading indicates the P_{10}–P_{90} band.

4.1.3 Plume Areal Coverage

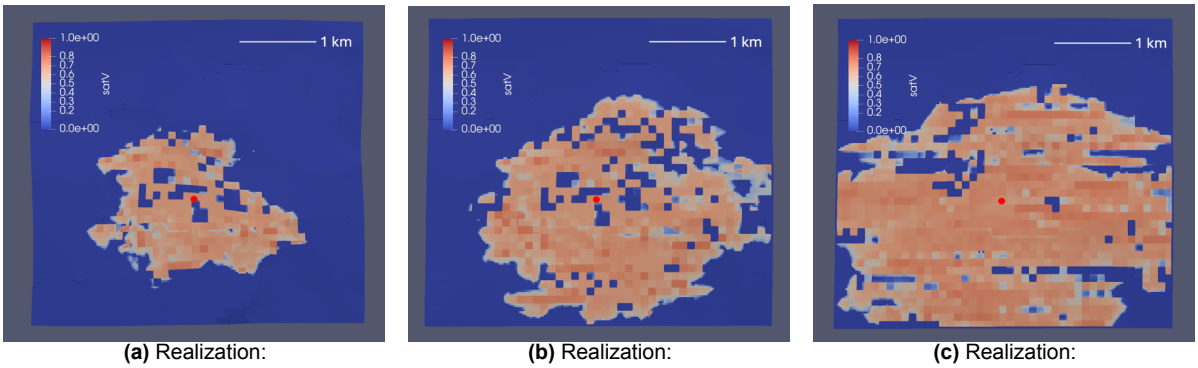
Figures 4.3(a) and 4.3(b) show the plume areal coverage over 20 years for Ensembles 1 and 2, with ensemble percentiles derived from the full set of simulations. Areal coverage is computed pointwise in time following the procedure in Section 3.3.3.



(a) Ensemble 1 ($N = 105$) **(b)** Ensemble 2 ($N = 106$)

Figure 4.3: Plume areal coverage percentiles over the 20-year simulation window for Ensembles 1 and 2. Thin grey curves show full-physics simulation results for individual realisations; bold curves are the pointwise-in-time ensemble percentiles (P_{50} solid blue; P_{10} and P_{90} dashed green and red respectively). Shading indicates the P_{10}–P_{90} band.

Figures 4.4(a), 4.4(b), and 4.4(c) show plan views of CO_2 saturation after 12 years for three different realisations, illustrating spatial variability in plume coverage across parameter configurations.



(a) Realization: FM3–CO3–TS1–OBJ–Mult 2 **(b)** Realization: FM2–CO3–TS2–OBJ–Mult 1 **(c)** Realization: FM2–CO1–TS2–PIX–Mult 1

Figure 4.4: CO_2 saturation maps after 12 years for three realizations, highlighting spatial variability in plume coverage for different parameter settings. The red dot marks the injector (Ensemble 1).

4.2 Weighted vs. Unweighted Percentile Reconstruction

Before presenting the results of clustering-based subset selection, it is important to highlight a consistent finding throughout this study: percentile reconstruction using cluster-population weighting consistently outperforms the unweighted approach, which assumes equal importance for all selected realizations as explained in Section 3.5.1.

Figure 4.5 illustrates the difference in percentile-reconstruction performance between the two methods. The top row shows relative RMSE values, per cluster count (number of selected clusters), for each percentile when reconstructed with cluster-population weighting, while the bottom row shows the same metric for the unweighted reconstruction. The RMSE quantifies the deviation between the percentile curve reconstructed for a given number of selected clusters and the corresponding reference curve from the full ensemble. Lower RMSE values indicate a closer match and thus better reconstruction performance, whereas higher values reflect reduced accuracy as described in Section 3.5.2.

The weighted method (assigning greater influence to representatives of larger clusters) yields smoother and more stable RMSE–versus–number-of-clusters curves for the P_{90}, P_{50}, and P_{10} percentiles. In contrast, the unweighted approach often exhibits erratic behaviour and larger errors. This is most evident for P_{50} and P_{10}: with weighting, after roughly 15 selected clusters all dimensional-reduction workflows tend to converge to stable, accurate behaviour, with RMSE values dropping below the 5 % threshold and remaining low. Without weighting, particularly for MDS + K-means and KPCA + K-means, the RMSE for P_{50} and P_{10} fluctuates strongly with the number of clusters, sometimes reaching values

up to $30\ \%$, as is especially clear for the P_{10} percentile.

Overall, the weighted approach yields consistently lower and more stable errors across all percentiles. These findings underline the value of accounting for cluster population when reconstructing ensemble statistics from clustering-based selections. Section 5.1 discusses possible reasons for the superior performance of the weighted method, and Appendix C provides further examples confirming its robustness across outputs and clustering setups.

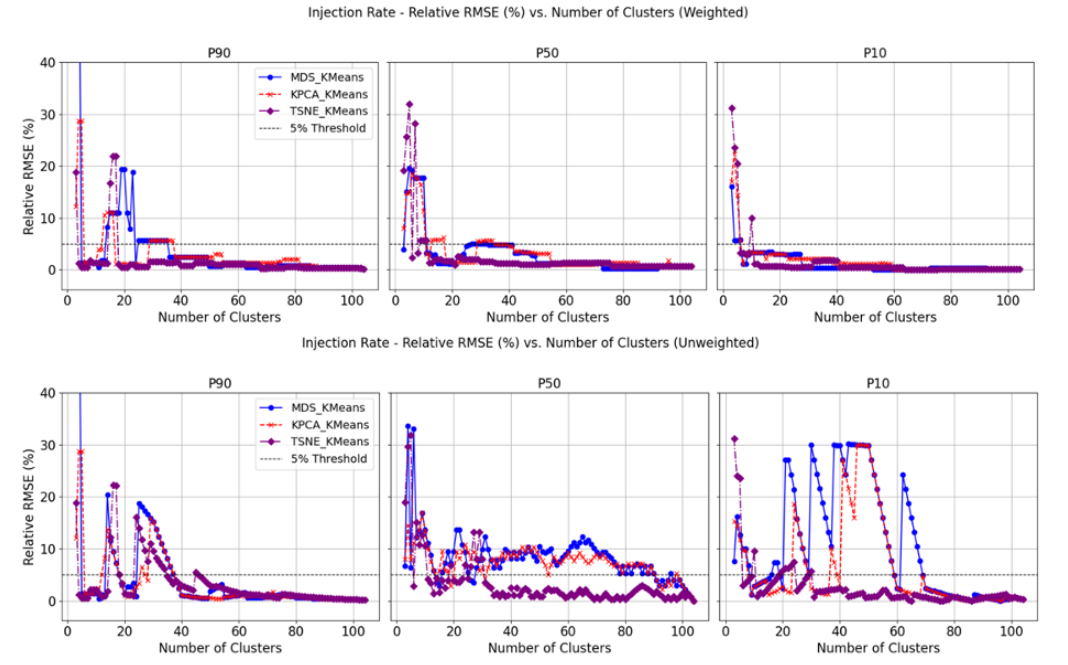


Figure 4.5: Relative RMSE as a function of the number of selected models for percentile reconstruction using two strategies: cluster-population–based weighting (top row) and an unweighted approach (bottom row). The unweighted method exhibits more irregular RMSE trends across varying cluster counts compared with the weighted approach. Results are derived from distance-based clustering of the day-1 full-physics rate diagnostic used to reconstruct the ensemble injection-rate percentiles.

4.3 Injection Rate Results

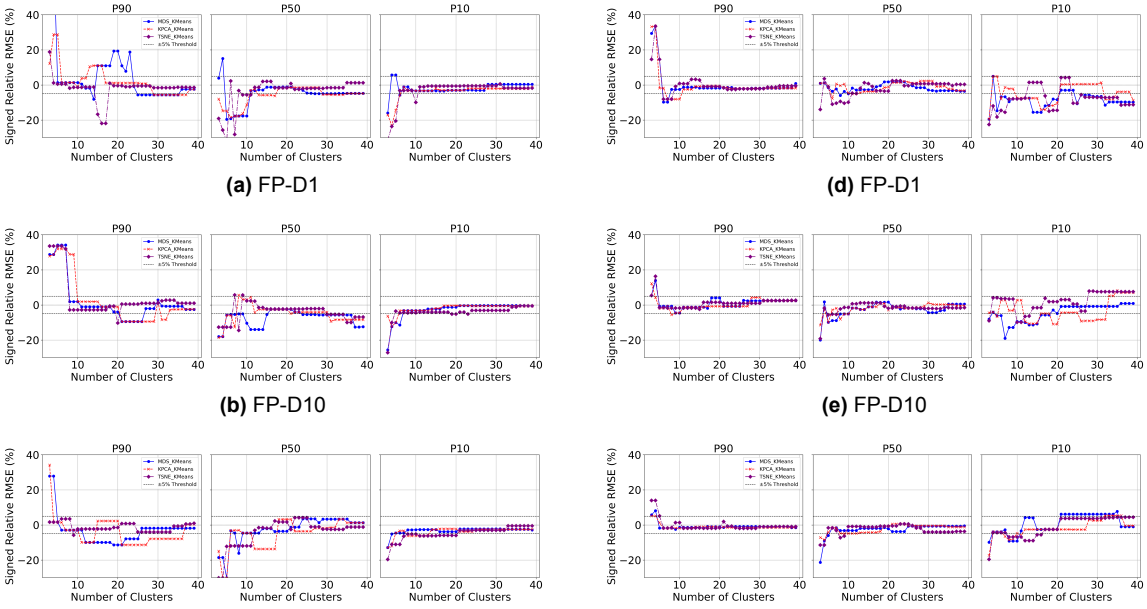
As discussed in Section 3.1, injection rate is one of the key CO_2 storage metrics considered in this study, alongside maximum plume extent and plume areal coverage. In the context of the proposed workflows, injection rate serves as the target metric against which the accuracy of reduced model subsets is evaluated. To construct the flow-based distance metrics that form the input to dimensionality reduction and clustering, flow diagnostics were derived either from the 3DSL streamline-based flow simulator or from early-time injection rates of the full-physics simulator as discussed in Section 3.3.

The representative models identified through clustering were simulated over the full injection period using the full-physics reservoir simulator. From this subset of realizations, the P_{90}, P_{50}, and P_{10} percentiles were computed and compared with those of the full ensemble to evaluate how well the ensemble reductions produced by each workflow preserve the distribution of injection behavior of the full ensemble. Results are presented in three parts: first, using FDs based on early-time full-physics injection rates (Section 4.3.1); second, using 3DSL-derived FDs (Section 4.3.2); third, a comparative evaluation of both diagnostics (Section 4.3.3).

4.3.1 Results for Full-Physics Flow Diagnostics

Figure 4.6 shows signed relative RMSE (%, Section 3.5.2) versus the number of selected clusters for Ensembles 1 and 2, based on clustering early-time injection-rate flow diagnostics (FP-D1, FP-D10, FP-D100) computed with the full-physics Open-DARTS simulator (Section 3.3). Curves are reported for percentiles P_{90}, P_{50}, and P_{10} and for three dimensionality-reduction (DR) methods: t-SNE (purple),

KPCA (red), and MDS (blue) (Section 3.4.1).



(a) FP-D1 **(d)** FP-D1

(b) FP-D10 **(e)** FP-D10

(c) FP-D100 **(f)** FP-D100

Ensemble 1 Ensemble 2

Figure 4.6: Signed relative RMSE (%) as a function of the number of selected clusters for Ensembles 1 and 2. Early-time injection-rate flow diagnostics (FP-D1, FP-D10, FP-D100) are computed with the full-physics Open-DARTS simulator (Section 3.3) and used for distance-based clustering. Results are reported for P_{90}, P_{50}, and P_{10} and for three DR workflows: t-SNE (purple), KPCA (red), and MDS (blue). Left: Ensemble 1 (a–c). Right: Ensemble 2 (d–f).

Best-Performing Dimensionality-Reduction Workflow

While no single DR workflow consistently outperforms the others across all cases, t-SNE shows a modest advantage over MDS and KPCA. Across most flow diagnostics, t-SNE more often reconstructs percentiles within the ± 5 % RMSE threshold, achieves higher accuracy at low cluster counts ($k < 10$), and exhibits fewer large RMSE excursions as k increases. In addition, t-SNE shows an overall tendency to slightly underestimate the reconstructed percentiles (negative signed RMSE), suggesting a conservative bias. In the context of probabilistic injection-rate forecasts, such conservatism is preferable to optimistic bias because it reduces the risk of overpromising capacity and for instance failing to meet contracted storage targets. Section 5.2.3 discusses potential reasons for this downward bias.

Overall, these results indicate that t-SNE provides slightly more accurate and stable reconstructions in this study when early-time injection rates are used as flow diagnostics for injection-rate uncertainty quantification. Consequently, the following analysis focuses primarily on t-SNE.

Comparison of Flow Diagnostic Time Windows

Across both ensembles, all early-time injection rates tend to provide accurate reconstructions for all percentiles at relatively low cluster counts, particularly when combined with t-SNE for dimensionality reduction. In Ensemble 1, the 1-day diagnostic requires only six models to reconstruct all percentiles (P_{90}, P_{50}, and P_{10}) with signed relative RMSE values within the ± 5 % RMSE threshold (Figure 4.6a), compared to nine models with the 10-day diagnostic (Figure 4.6b) and 12 models with the 100-day diagnostic (Figure 4.6c). In Ensemble 2, the same error threshold is reached with 13 models at 1 day (Figure 4.6d), eight models at 10 days (arguably already at five, since the P_{50} signed RMSE is only ≈ -5.1 %; Figure 4.6e), and seven models at 100 days (Figure 4.6f).

The corresponding simulation costs, relative to running the entire ensemble, are summarized in Ta-

ble 4.1. These costs include both the overhead of simulating the flow diagnostic itself and the additional costs of simulating the selected models required to achieve the specified reconstruction accuracy, as outlined in Section 3.5. Based on these results, the 10-day diagnostic is slightly favored overall: it is the second-cheapest option for Ensemble 1 and the cheapest for Ensemble 2. In contrast, the 1-day diagnostic is the cheapest in Ensemble 1 but the most expensive in Ensemble 2, while the 100-day diagnostic is the least cost-effective, particularly given its performance in Ensemble 1.

Table 4.1: Comparison of early-rate full physics flow diagnostics for Ensemble 1 ($N = 105$) and Ensemble 2 ($N = 106$). The reported costs include both the overhead of computing the flow diagnostic for all realisations and the cost of simulating the selected models to full lifetime, expressed as a percentage of the total cost of simulating the complete ensemble to full lifetime.

FD	Ensemble 1			Ensemble 2			Average (1 & 2)		
	K	Cost [%]	Walltime	K	Cost [%]	Walltime	K	Cost [%]	Walltime
FP-D1	6	6.3	10:05:11	13	12.8	20:38:24	9.5	9.6	15:21:48
FP-D10	9	9.7	17:00:41	8	8.7	14:02:15	8.5	9.2	15:31:28
FP-D100	12	15.1	24:11:37	7	10.4	16:46:58	9.5	12.8	20:29:18

Looking more broadly at the use of early-time full-physics injection rates as flow diagnostics, both ensembles indicate that the day-10 and day-100 diagnostics generally yield the most stable reconstruction performance across different numbers of selected clusters. Both exhibit fewer extreme fluctuations, whereas the day-1 diagnostic occasionally produces RMSE values exceeding $20\ \%$, even after more than ten clusters have been selected (see, for example, the P_{90} of Ensemble 1 in Figures 4.6a). In addition, extending the diagnostic window from 10 to 100 days does not provide significant additional benefits, with both ensembles showing little or no improvement. This is particularly important given that longer simulation times entail higher computational costs.

Therefore, based on these results, the day-10 rate diagnostic appears to be the most appropriate choice: it provides stable and accurate percentile reconstructions (P_{10}, P_{50}, and P_{90}) over varying K, while being computationally more efficient than the day-100 diagnostic.

These findings suggest that early-time injection rate signals are generally sufficient for effective clustering in this study, and that extending the diagnostic window to 100 days does not necessarily improve performance. By contrast, the day-1 diagnostic seems least appropriate, as waiting slightly longer yields more stable behavior across P_{90}, P_{50}, and P_{10}. Section 5.2.1 will discuss potential reasons why early-time injection rate signals serve as effective flow diagnostics for clustering, while Section 5.2.2 will further examine the considerations involved in selecting an appropriate diagnostic window.

Fluctuations in Relative RMSE

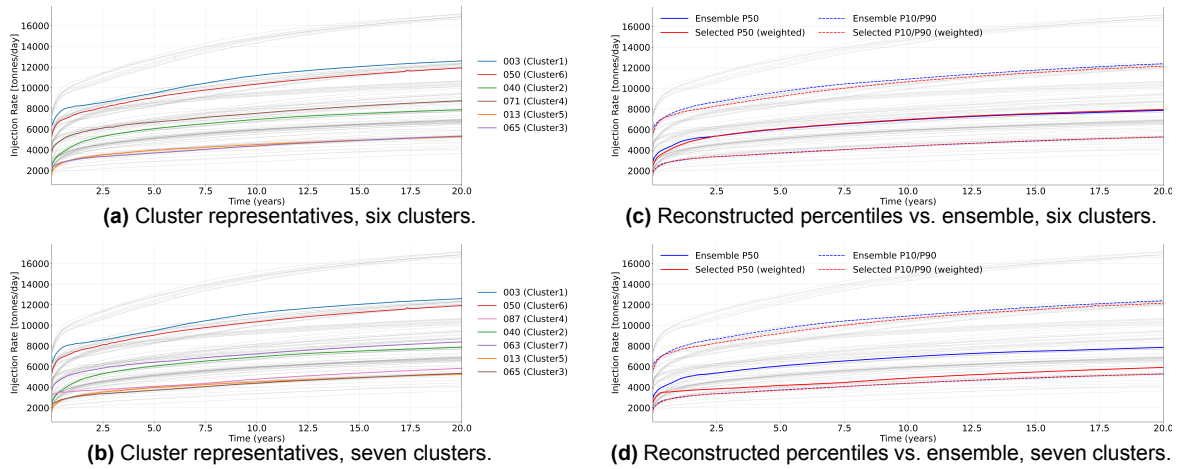
Note that the analysis above is ex post: full-ensemble, full-lifetime simulations are used to compute the true percentiles and to identify the earliest K meeting the $\leq 5\ \%$ relative RMSE threshold. In practice this vantage point is unavailable, because simulating every realization is precisely what clustering seeks to avoid.

This matters because all dimensionality-reduction and clustering workflows exhibit some degree of erratic behavior in percentile reconstruction accuracy, with sudden jumps in RMSE at specific numbers of selected models. While adding models generally improves accuracy and stability, the trend is not monotonic: in some cases, additional realizations disrupt the CDF reconstruction and cause large deviations from the true percentiles.

This behaviour is, for example, evident in Ensemble 1 with t-SNE, where the P_{50} percentile spikes to nearly $-30\ \%$ when the number of selected clusters increases from six to seven using the 1-day diagnostic (Figure 4.6a). The impact of this jump on the reconstructed injection-rate percentiles is shown in Figure 4.7, where the transition from six (Figure 4.7c) to seven (Figure 4.7d) clusters leads to a clear degradation in the reconstructed P_{50} curve (red) relative to the true ensemble P_{50} (blue). For the same workflow, the P_{90} percentile also exceeds $-20\ \%$ between 15 and 17 models (Figure 4.6a)

before stabilizing. Ensemble 2 shows fewer extreme deviations but still exhibits moderate fluctuations in the P_{50} and P_{10} percentiles across different diagnostics.

These results highlight that percentile reconstruction accuracy cannot be guaranteed simply by selecting a larger number of models, and without access to the true ensemble percentiles such fluctuations cannot be managed by visual inspection. Consequently, an appropriate K must be selected using internal clustering metrics; this will be discussed in Section 4.4.



(a) Cluster representatives, six clusters.

(c) Reconstructed percentiles vs. ensemble, six clusters.

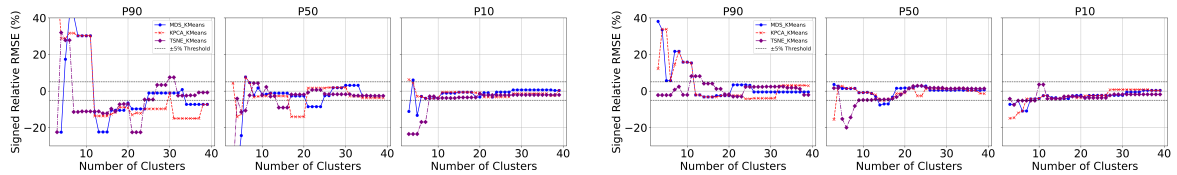
(b) Cluster representatives, seven clusters.

(d) Reconstructed percentiles vs. ensemble, seven clusters.

Figure 4.7: Effect of increasing the number of selected clusters from six to seven on the reconstruction of injection-rate percentiles for Ensemble 1 using the FP-D1 diagnostic in combination with t-SNE. Panels (a)–(b) highlight the selected cluster representatives for six and seven clusters, respectively.Panels (c)–(d) show the reconstructed P_{90}, P_{50}, and P_{10} compared with the true ensemble percentiles.

4.3.2 Results for Streamline-Based Flow Diagnostics

For the injection-rate analysis, two different streamline-based flow models were used: a single-phase formulation and a two-phase immiscible formulation, as described in Section 2.4 and 3.2.2. The results obtained from the different flow diagnostic approaches are presented in this section. Figures 4.8(a) and 4.8(b) show the signed relative RMSE (%) as a function of the number of selected clusters for Ensemble 1 and Ensemble 2, respectively, using the single-phase formulation, across three dimensionality-reduction methods: t-SNE (purple), KPCA (red), and MDS (blue). Figure 4.9 presents the corresponding results for the immiscible diagnostic, evaluated from the injection rate at three simulation times: (a)/(d) after the first pressure solve (day 1, IMM-PS1), (b)/(e) after the fourth pressure solve (25 days, IMM-PS4), and (c)/(f) after 11 pressure solves (20 years, IMM-PS11).



(a) Ensemble 1 - SP-PS1

(b) Ensemble 2 - SP-PS1

Figure 4.8: Signed relative RMSE versus the number of selected clusters for the single-phase flow diagnostic (SP-PS1). Results are reported for P_{90}, P_{50}, and P_{10} and for three DR workflows: t-SNE (purple), KPCA (red), and MDS (blue). Results are shown for Ensemble 1 (left) and Ensemble 2 (right).

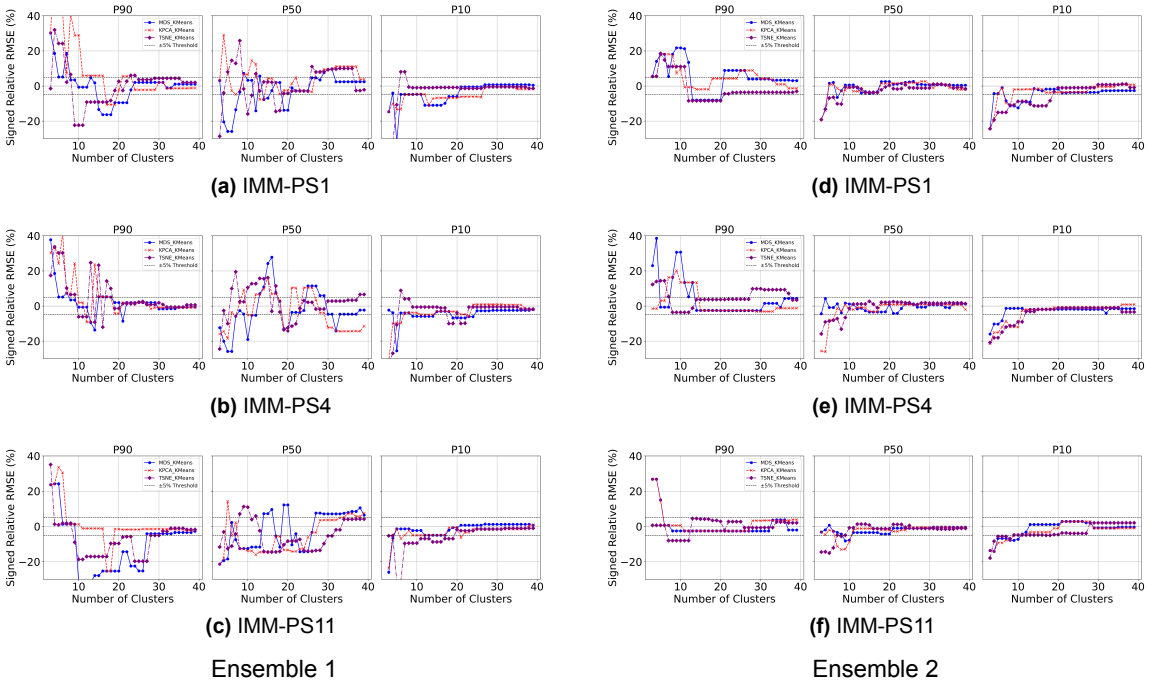


Figure 4.9: Signed relative RMSE versus number of clusters for the immiscible flow diagnostics, computed from the injection rate at three simulation times. Results are shown for P_{90}, P_{50}, and P_{10} across three dimensionality-reduction workflows: t-SNE (purple), KPCA (red), and MDS (blue). Left: Ensemble 1 at (a) first pressure solve (∼1 day), (b) fourth pressure solve (∼25 days), and (c) 11 pressure solves (∼20 years). Right: Ensemble 2 at (d) first pressure solve (∼1 day), (e) fourth pressure solve (∼25 days), and (f) 11 pressure solves (∼20 years).

Best Performing Dimensionality Reduction Workflow

While no single dimensionality-reduction workflow consistently dominates across all cases, t-SNE shows a slight advantage over MDS and KPCA. For most flow diagnostics, t-SNE more frequently reconstructs percentiles within the ±5 % RMSE threshold, demonstrates greater accuracy at low cluster counts ($K < 10$), and exhibits fewer large fluctuations in RMSE as the number of clusters increases. These trends indicate that t-SNE generally provides more accurate and stable reconstructions, particularly at lower cluster counts, in this study. Consequently, the following results primarily focus on t-SNE. Nevertheless, the observed advantage is modest and not sufficient to recommend t-SNE as a universally superior approach when applying streamline-based flow diagnostics for injection-rate uncertainty quantification.

Single-Phase vs. Two-Phase Immiscible Performance

Overall, the results indicate that using the injection rate after the first pressure solve in the single-phase model (SP-PS1) as a flow diagnostic yields more stable and accurate clustering outcomes across the tested range of cluster counts (3 to 40) compared to injection rates from the immiscible model. This is particularly evident for the P_{90}, P_{50}, and P_{10} of Ensemble 2, and for the P_{50} and P_{10} of Ensemble 1 as shown in Figure 4.9. At medium cluster counts (10-20), the advantage of the single-phase diagnostic is especially clear, with t-SNE results for these percentiles more frequently remaining within the ±5 % RMSE threshold compared to the other diagnostics.

By contrast, immiscible-flow diagnostics appear to show a stronger tendency toward instability in percentile-reconstruction accuracy across the DR workflows, with abrupt RMSE jumps as the number of clusters K increases. Since increasing K subdivides existing clusters, the irregular RMSE behavior suggests that the resulting partitions for the immiscible diagnostics could become less aligned with the actual ensemble structure for certain K. This indicates that additional partitions could capture noise rather

than meaningful signal, leading to cluster representatives that reproduce the ensemble percentiles less effectively. Consequently, the reconstructed P_{90}/P_{50}/P_{10} values shift away from the true ensemble percentiles, and RMSE tends to increase. The effect seems most pronounced in Ensemble 1 and somewhat milder in Ensemble 2. In comparison, the single-phase diagnostic tends to reconstruct percentiles more consistently across K, suggesting cleaner, more stable partitions and less dependence on the specific cluster configuration. Section 5.3 examines why the single-phase diagnostic tends to yield overall more stable and accurate percentile reconstructions across the tested range of clusters.

An important exception is the P_{90} of Ensemble 1 (Figure 4.8(a)), where reconstruction with the single-phase diagnostic is comparatively poor: signed RMSE values remain close to -10% between $K = 7$ and $K = 24$, and only fall within the $\pm 5\%$ RMSE bound beyond that point. That said, the P_{90} from the immiscible-flow diagnostics is not significantly more accurate or stable and also tends to show relatively large RMSE values across all DR workflows for varying K. Therefore, because two of the three percentiles (P_{50} and P_{10}) exhibit the most stable and accurate behavior across both ensembles, and the P_{90} of Ensemble 2 is reconstructed reliably even at low cluster counts (<10) (Figure 4.8(b)), while reconstructing the P_{90} of Ensemble 1 appears to be challenging for all flow diagnostics, the single-phase diagnostic can still be regarded as the most reliable option on average in this study. Section 5.3.2 discusses why the P_{90} of Ensemble 1 may be particularly difficult to reconstruct accurately with the streamline-based flow diagnostics.

To further substantiate this conclusion, Table 4.2 reports the total simulation costs of the different flow diagnostics. The table shows the costs of selecting the required number of clusters to reconstruct all percentiles within the $\leq 5\%$ RMSE constraint, expressed relative to the total simulation cost of simulating the complete ensemble. Although it is important to again acknowledge the poor P_{90} reconstruction of the single-phase formulation for Ensemble 1 (which requires 24 models to meet the RMSE threshold) the superior performance for Ensemble 2 results in the lowest average cost overall: 13.5 % on average across both ensembles, compared to 19.6 %, 18.5 %, and 15 % for the immiscible 1-day, 25-day, and 20-year diagnostics, respectively.

Table 4.2: Comparison of streamline flow diagnostics for Ensemble 1 ($N = 105$) and Ensemble 2 ($N = 106$). The reported costs include both the overhead of computing the flow diagnostic for all realisations and the cost of simulating the selected models to full lifetime, expressed as a percentage of the total cost of simulating the complete ensemble to full lifetime.

Flow diagnostic	Ensemble 1			Ensemble 2			Average (1 & 2)		
	K	Cost [%]	Walltime	K	Cost [%]	Walltime	K	Cost [%]	Walltime
SP-PS1	24	23.5	37:24:52	3	3.5	05:37:31	13.5	13.5	21:31:12
IMM-PS1	19	18.7	29:45:49	21	20.5	32:56:46	20.0	19.6	31:21:18
IMM-PS4	23	23.8	37:53:23	12	13.2	21:12:50	17.5	18.5	29:33:06
IMM-PS11	4	10.8	17:12:12	13	19.2	30:51:45	8.5	15.0	24:01:59

After the single-phase formulation, the IMM-PS11 diagnostic appears to be the most appropriate alternative in this study when considering the average cost of meeting the ± 5 % RMSE threshold for all percentiles. It differs only slightly in average cost (about 1.5 %) compared with the single-phase diagnostic while requiring fewer cluster selections to reconstruct all percentiles within the ± 5 % RMSE threshold. However, two aspects of its behaviour merit closer examination.

First, although for Ensemble 1 only $K = 4$ models are required to achieve accurate percentile reconstruction with this immiscible diagnostic, the solution is highly cluster-sensitive. For example, with $K = 3$ clusters the signed RMSE of P_{90} increases to approximately $+35$ %, and the P_{50} exceeds -10 %; with $K = 5$ models, the P_{10} decreases to about -30 %, and the P_{50} again exceeds -10 % (Figure 4.9c). In addition, if exactly $K = 4$ models are not selected, the threshold is not satisfied again until roughly $K = 30$ models are included. A similar sensitivity occurs with the single-phase diagnostic for Ensemble 2 (Figure 4.8(b)), where the constraint is not consistently met until $K = 9$ clusters if $K = 3$ are not chosen. This deviation, however, is much less severe: at $K = 4$ clusters only the P_{10} signed RMSE rises to around -7 %, and between $K = 5$ and $K = 8$ clusters only the P_{50} exceeds the threshold, fluctuating between -8 % and -20 %, before all percentiles again satisfy the constraint at $K = 9$ clus-

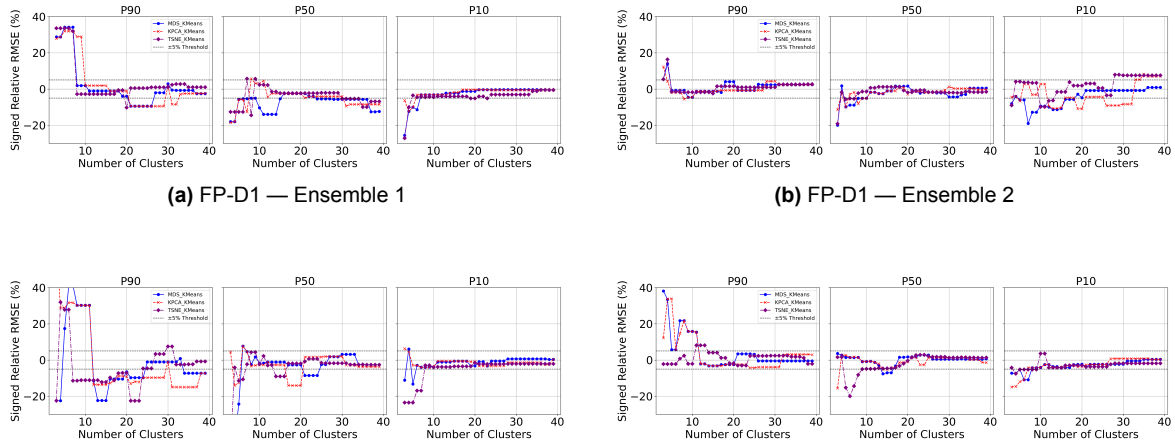
ters. Hence, the gap in the number of models required to re-attain the threshold is much smaller for the single-phase diagnostic, and its RMSE fluctuations are far less pronounced.

Second, the 20-year immiscible diagnostic has substantially higher overhead costs than the single-phase model (> 6 min versus ≈ 35 s). This means that, in this study, even if the longer diagnostic reduces the number of selected models, its additional overhead may limit any real cost advantage. For example, selecting $K = 24$ models with the single-phase diagnostic for Ensemble 1 corresponds to $23.5\ \%$ of the cost of simulating the full ensemble, whereas the immiscible diagnostic requires only $K = 13$ models but still results in $19.2\ \%$ of the total cost. That said, it should be acknowledged that this study relied on GPU-accelerated full-physics simulations, against which the overhead is measured: on average, a 20-year simulation required only 92 minutes, as shown in Section 3.5.3. As a result, even small differences in overhead time can translate into relatively large percentage differences. This effect would have been less pronounced if CPU-based simulations had been used, since in that case the overhead would represent a smaller fraction of the total cost.

As a final observation, diagnostics based on two-phase immiscible flow, using injection rates sampled at different pressure solves, reveal that in Ensemble 1, accuracy across varying cluster counts declines slightly when moving from rates at the first timestep to those at later timesteps, although the difference is minimal; both choices show strong, erratic behavior across cluster counts for all dimensionality-reduction methods. By contrast, in Ensemble 2 the pattern reverses: accuracy and stability improve when rates from later timesteps are used. Section 5.3.3 discusses mechanisms that could cause the immiscible formulation's clustering performance to degrade with simulation time in some ensembles yet improve in others.

4.3.3 3DSL vs Full-Physics Flow Diagnostics

Figures 4.10(a) and 4.10(b) show the signed relative RMSE results obtained for different numbers of clusters when using the FP-D10 flow diagnostic from the full-physics simulations for Ensembles 1 and 2. Figures 4.10(c) and 4.10(d) show the corresponding results for the single-phase flow diagnostic SP-PS1 obtained with 3DSL. Both diagnostics were found to perform best among the various FDs tested within the streamline-based and full-physics-based flow-diagnostic setups, as discussed in Sections 4.3.1 and 4.3.2. They are therefore briefly compared with each other in this section.



(a) FP-D1 — Ensemble 1

(b) FP-D1 — Ensemble 2

(c) SP-PS1 — Ensemble 1

(d) SP-PS1 — Ensemble 2

Figure 4.10: Signed relative RMSE (%) as a function of the number of selected clusters for two flow-diagnostic approaches. Top row: results obtained with the 10-day injection-rate flow diagnostic from full-physics simulations for Ensembles 1 and 2 ((a), (b)). Bottom row: corresponding results for the single-phase injection-rate diagnostic (3DSL) for Ensembles 1 and 2 ((c), (d)). Results are reported for P_{90}, P_{50}, and P_{10} and for three DR workflows: t-SNE (purple), KPCA (red), and MDS (blue).

Overall, the early-time rate diagnostic from the full-physics simulator performs better in terms of RMSE accuracy and stableness over different number of selected clusters, particularly when using the t-SNE dimensionality reduction workflow. For Ensemble 2, the streamline diagnostic (Figure 4.10(d)) some-

times performs slightly better for the P_{10} percentile compared to the full-physics FD (Figure 4.10(b)). However, this advantage is outweighed by its much worse performance for the P_{90} percentile in Ensemble 1 (Figure 4.10(c)).

Looking at the point where all percentiles fall below the ≤ 5 % RMSE threshold, and focusing on the t-SNE DR workflow, Ensemble 1 reaches this level of accuracy with as few as nine clusters when using the full-physics diagnostic, whereas the streamline diagnostic requires around 24 clusters to achieve the same performance. For Ensemble 2, the gap is smaller and reversed: the full-physics diagnostic meets the threshold with eight selected clusters, compared with only three clusters for the streamline diagnostic. However, one could argue that, in the full-physics case, the constraint is effectively satisfied already at five clusters, since the slight exceedances above 5 % are only a few decimal points and the reconstruction remains stable up to ten clusters, at which point a small jump in the P_{10} percentile appears (Figure 4.10(b)). By contrast, although the streamline simulator performs impressively at three clusters, its accuracy deteriorates immediately afterward: the threshold is not met again until nine clusters, owing to a significant spike in the P_{50} percentile (Figure 4.10(d)). This suggests that the full-physics diagnostic is also generally more reliable for Ensemble 2 at low cluster counts.

Section 5.2.1 discusses possible reasons why early injection rates obtained with the full-physics simulator may be more effective than those from the streamline simulator.

4.4 Selecting the Appropriate Number of Clusters

As outlined in Sections 3.4.3 and 4.3.1, selecting an appropriate number of clusters is a central challenge for distance-based ensemble reduction. Internal clustering metrics offer a practical way to guide this choice, as they provide quantitative indicators of cluster compactness and separation without requiring comparison against the full set of computationally expensive full-physics simulations.

To this end, four widely used internal metrics were evaluated: the Silhouette score, the Davies–Bouldin (DB) index, Inertia, and the Calinski–Harabasz (CH) index, as discussed in Section 3.4.3. As Section 4.3.3 showed, the FP-D10 rate diagnostic combined with t-SNE as the dimensionality-reduction technique produced the most stable and accurate percentile reconstructions among all workflows tested. Therefore, this section first focuses on the internal cluster-metric results for this specific configuration. A more general overview is then provided for all workflows when cluster selection is guided solely by the inertia elbow.

4.4.1 Internal Metrics for the Day-10 Full-Physics Rate Diagnostic with t-SNE

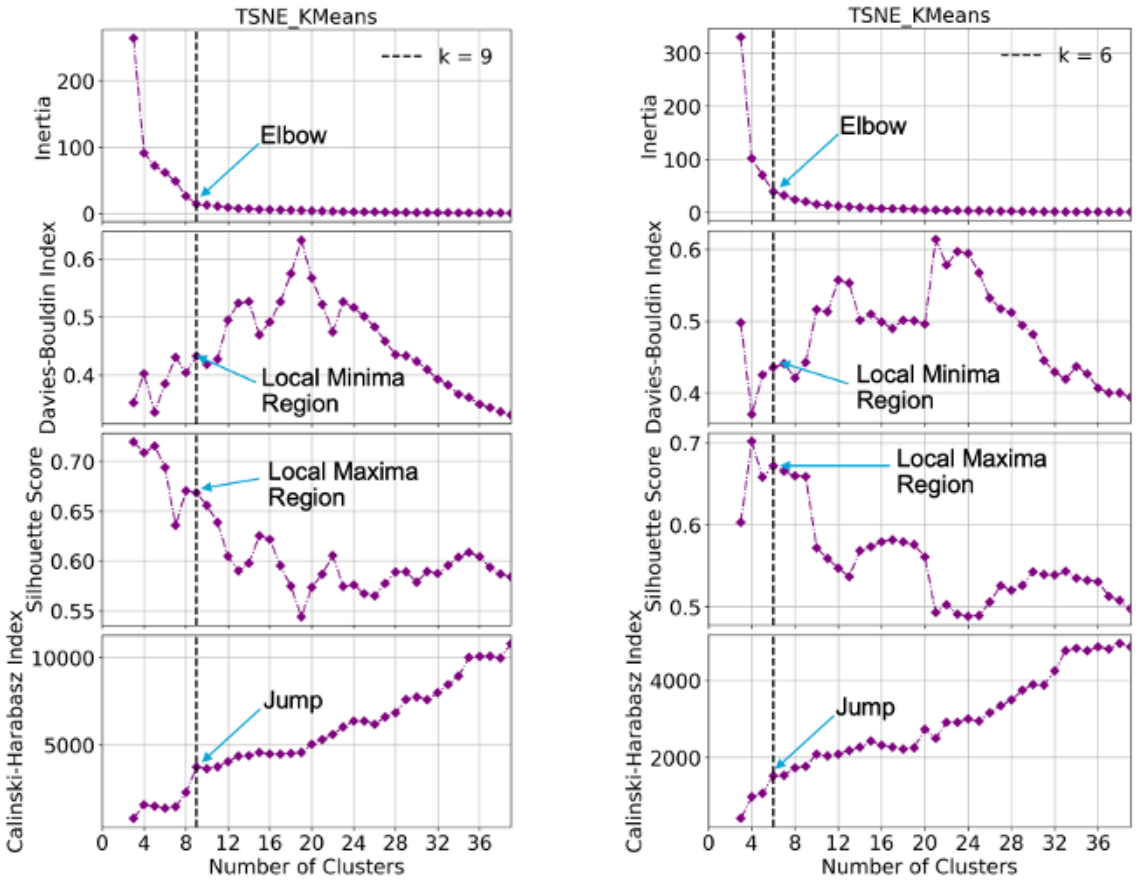
Figure 4.11 presents the internal clustering metrics for Ensembles 1 and 2 for different number of clusters selected. Inspection of these results shows that Inertia exhibits a clear elbow around nine clusters for Ensemble 1 and around six clusters for Ensemble 2, suggesting suitable thresholds for cluster selection. The other internal metrics provide consistent support in the same regions: the DB index is close to a local minimum with only minor fluctuations, the Silhouette score is near a local maximum, and the CH index shows a noticeable increase when moving from eight to nine clusters for Ensemble 1 and from five to six clusters for Ensemble 2. Together, these trends indicate that clustering quality is strongest in these ranges, with nine clusters emerging as an appropriate choice for Ensemble 1 and six clusters for Ensemble 2.

This observation aligns with the reconstruction accuracy at these cluster counts. For both ensembles, all percentiles remain within the ≤ 5 % RMSE threshold, with one exception: the P_{50} exceeds it slightly (by about 0.1–0.2 %) in both ensembles. Such a minor deviation can be considered acceptable, particularly when contrasted with the scenario in which eight clusters would have been selected for Ensemble 1, which would have caused the P_{50} signed relative RMSE to drop sharply below -15 % (Fig. 4.10(a)).

In addition, special attention should be given to cluster selections 6, 7, 8, and 9 for Ensemble 2, as shown in Figure 4.11(b). The “elbow” at $K = 6$ could arguably be placed slightly later, although this study identified it at 6. That said, if it had been placed at 7, 8, or 9, neither the DB index nor the Silhouette index would have opposed this, as these metrics tend to remain relatively stable around local minima. However, choosing 10 clusters would have been clearly discouraged by both indices, given the significant increase in DB and the sharp drop in Silhouette values when moving from 9 to 10 clusters

(see Figure 4.11(b)). This highlights the value of cross-referencing multiple internal metrics when the inertia "elbow" is not entirely distinct. Inertia provides a strong primary guide because it quantifies the compactness that k-means explicitly optimizes and typically exhibits a clear point of diminishing returns as K increases. Other indices (Silhouette, Davies–Bouldin, Calinski–Harabasz) can then be used to verify that the inertia-based elbow also corresponds to well-separated and meaningful clusters.

Finally, it should be noted that, if instead of prioritizing Inertia, the DB and Silhouette indices had been used as the primary criteria, one would most likely have chosen five clusters for Ensemble 1 and four clusters for Ensemble 2, since at those counts the metrics suggest approximate global optima (lowest DB and highest Silhouette values). However, as shown in Figure 4.10(b), this would have resulted in P_{90} signed relative RMSE reconstructions of $+34$ % and $+17$ %, P_{50} reconstructions of -13 % and -2 %, and P_{10} reconstructions of -4 % and $+5$ % for Ensembles 1 and 2, respectively. This represents particularly poor performance for the P_{90} percentile reconstruction and indicates that the DB and Silhouette indices alone are not highly reliable indicators of cluster quality. That said, when combined with Inertia, identified here as the most robust indicator, they can still provide useful confirmation or help identify options to avoid, as noted earlier.



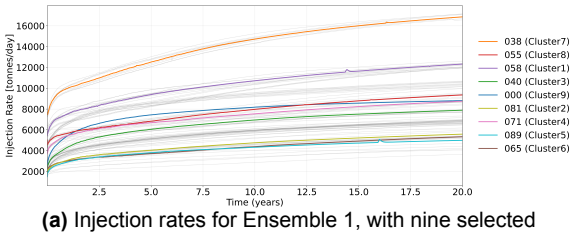
(a) Internal clustering metrics for Ensemble 1 (FD-D10 + t-SNE)

(b) Internal clustering metrics for Ensemble 2 (FD-D10 + t-SNE)

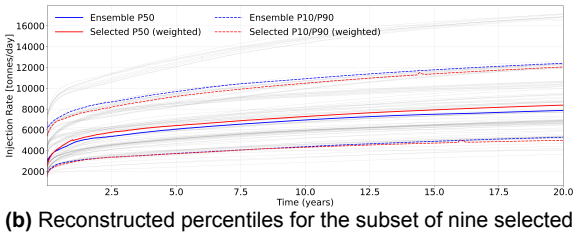
Figure 4.11: Internal clustering metrics for the day-10 full-physics rate diagnostic with t-SNE. Inertia shows a clear elbow (nine clusters for Ensemble 1, six for Ensemble 2), while the Davies–Bouldin, Silhouette, and Calinski–Harabasz indices provide consistent evidence supporting these choices.

To give the reader a clearer sense of what such low RMSE values per percentile imply in practice, Figures 4.12(a) and 4.13(a) show the selected realizations in Ensembles 1 and 2 when the number of clusters is chosen based on the internal metrics (nine clusters for Ensemble 1 and six for Ensemble 2). Figures 4.12(b) and 4.13(b) then present the reconstructed percentile curves for these subsets, consist-

ing of the selected cluster representatives, compared with the full-ensemble percentiles. These plots demonstrate that, even with a (very) limited number of models, the key statistics can be reproduced with high accuracy: the reconstructed percentile curves (red) closely track the corresponding full-ensemble percentiles (blue) over the entire simulation period, confirming the low RMSE values reported.

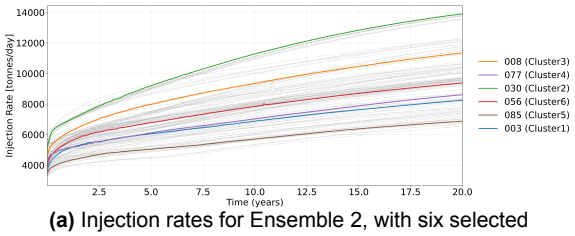


(a) Injection rates for Ensemble 1, with nine selected realisations highlighted as indicated by the internal clustering metrics.

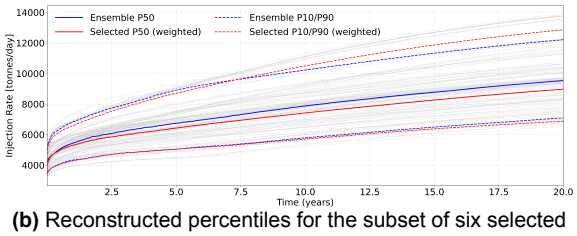


(b) Reconstructed percentiles for the subset of nine selected realisations (red), compared with the full-ensemble percentiles (blue) (Ensemble 1).

Figure 4.12: Impact of selecting nine realisations for Ensemble 1, as suggested by the internal cluster metrics. Left: highlighted realisations in the injection-rate profiles. Right: reconstructed percentiles based on this subset compared with the full ensemble.



(a) Injection rates for Ensemble 2, with six selected realizations highlighted as indicated by the internal clustering metrics.



(b) Reconstructed percentiles for the subset of six selected realizations (red), compared with the full-ensemble percentiles (blue) (Ensemble 2).

Figure 4.13: Impact of selecting six realizations for Ensemble 2, as suggested by the internal cluster metrics. Left: highlighted realizations in the injection-rate profiles. Right: reconstructed percentiles based on this subset compared with the full ensemble.

4.4.2 Comparative Assessment of All Workflows Using Inertia-Based Guidance

Tables 4.3 and 4.4 summarise, for Ensemble 1 and Ensemble 2, the number of clusters (K) selected for injection-rate percentile reconstruction when the elbow in the inertia curve is used as the guidance criterion. This inertia-based selection is used because earlier results indicated that it tends to provide reasonable estimates of appropriate cluster counts while remaining straightforward and consistently applicable across the different flow-diagnostic settings considered. For each flow-diagnostic and dimensionality-reduction workflow tested in this study, the tables report the selected K, the total simulation cost of running the selected models (including the overhead of computing the flow diagnostics) expressed relative to the cost of simulating the complete ensemble for the full duration with the full-physics simulator (Section 3.5.3), and the signed relative RMSE values for each percentile reconstruction. Appendix D shows the inertia curves of all workflows with the indicated elbow points used in this study.

Table 4.3: Summary of the optimal number of clusters (K), relative simulation cost (%), and signed relative RMSE (%) for the P_{90}, P_{50}, and P_{10} injection-rate percentiles for each tested flow-diagnostic and dimensionality-reduction workflow (t-SNE, KPCA, and MDS) for Ensemble 1. The number of clusters was selected by locating the elbow in the inertia curve; costs include both the flow-diagnostic overhead and the simulations of the selected models.

FD	t-SNE					KPCA					MDS				
	K	Cost (%)	P_{90} (%)	P_{50} (%)	P_{10} (%)	K	Cost (%)	P_{90} (%)	P_{50} (%)	P_{10} (%)	K	Cost (%)	P_{90} (%)	P_{50} (%)	P_{10} (%)
FP-D1	6	6.3	1	2	-3	7	7.3	1	-18	-1	7	7.3	1	18	-1
FP-D10	9	9.7	-3	5	-4	10	10.7	2	3	-3	6	6.9	34	-5	-11
FP-D100	9	12.3	-5	-11	-5	10	13.2	-3	-5	-6	6	9.5	-3	-3	-5
SP-PS1	11	11.1	-11	2	-4	7	7.3	32	5	-4	7	7.3	41	5	-4
IMM-PS1	6	6.3	24	14	8	7	7.3	7	-4	-5	8	8.3	4	-4	-5
IMM-PS4	7	8.6	10	20	4	7	8.6	7	-6	-4	7	8.6	7	-5	-4
IMM-PS11	8	14.5	-2	7	-10	8	14.5	1	-13	-2	6	12.6	1	2	-1

Table 4.4: Summary of the optimal number of clusters (K), relative simulation cost (%), and signed relative RMSE (%) for the P_{90}, P_{50}, and P_{10} injection-rate percentiles for each tested flow-diagnostic and dimensionality-reduction workflow (t-SNE, KPCA, and MDS) for Ensemble 2. The number of clusters was selected by locating the elbow in the inertia curve; costs include both the flow-diagnostic overhead and the simulations of the selected models.

FD	t-SNE					KPCA					MDS				
	K	Cost (%)	P_{90} (%)	P_{50} (%)	P_{10} (%)	K	Cost (%)	P_{90} (%)	P_{50} (%)	P_{10} (%)	K	Cost (%)	P_{90} (%)	P_{50} (%)	P_{10} (%)
FP-D1	9	9.1	-1	-10	-8	10	10.0	-8	-5	-8	6	6.3	-10	-4	-7
FP-D10	6	6.9	-1	-5	4	11	11.5	-1	-2	3	11	11.5	-1	-2	-9
FP-D100	9	12.2	1	-6	-7	10	13.1	-2	-5	-5	5	8.6	-2	-6	-4
SP-PS1	5	5.3	-2	-15	-5	9	9.1	16	-1	-4	9	9.1	16	-1	-4
IMM-PS1	6	6.3	15	-6	-15	8	8.2	17	-3	-11	9	9.1	21	1	-11
IMM-PS4	6	7.6	14	-8	-15	7	8.5	16	-1	-9	9	10.4	30	2	-1
IMM-PS11	9	15.4	-8	-5	-5	8	14.4	1	-13	-8	9	15.4	-3	-8	-8

Upon analyzing the results, several points stand out. First, t-SNE combined with the day-10 full-physics rate FD (FP-D10) again appears to be the best-performing workflow. It not only yields the most accurate and stable percentile reconstructions, as reflected by the relatively low variability and low RMSE values across different choices of K, as concluded in Section 4.3.3, but also produces accurate results when Inertia is used as the primary cluster-count heuristic, with RMSEs within $5\ \%$ for all percentile reconstructions. In this respect, both t-SNE and KPCA paired with FP-D10 perform very well, but because t-SNE already achieves this performance at lower K, this workflow stands out in terms of accuracy, computational efficiency, and predictable cluster-indicator behavior.

Second, when the FDs are examined at the inertia-selected cluster counts, FP-D1 also performs best with t-SNE: it is strongest on Ensemble 1, acceptable on Ensemble 2 (with the P_{90} reconstruction standing out, showing a signed relative RMSE of just -1%), and none of its percentile reconstructions exceed 10% RMSE. That said, no workflow using this diagnostic meets the 5% RMSE threshold simultaneously for both ensembles. FP-D100 also does not meet the RMSE threshold for both ensembles simultaneously, but it misses it only slightly (-6% P_{10} on Ensemble 1 with KPCA, and -6% P_{50} on Ensemble 2 with MDS), making it a strong performer when paired with KPCA or MDS. t-SNE performs slightly weaker on Ensemble 1 due to elevated P_{50} RMSE. Overall, the early full-physics injection-rate diagnostics, combined with the various DR workflows, show promising results, indicating that they could be reliable flow-based distance metrics for clustering with a clear potential for guiding cluster selection based on Inertia.

Surprisingly, the single-phase flow diagnostic (SP-PS1), appraised in Sections 4.3.2 and 4.3.3 for its stable and accurate percentile reconstruction once $K \gtrsim 10$, is overall slightly less accurate than the immiscible 20-year formulation (IMM-PS11) when results are aggregated over all dimensionality-reduction techniques and cluster selection is guided by inertia. That said, when combined with t-SNE, it still deliv-

ers reasonable results, with four out of six percentiles across both ensembles reconstructed within the 5% RMSE bound. In addition, the two outliers (the reconstructed P_{90} of Ensemble 1 and P_{50} of Ensemble 2) are at least underestimated (negative RMSE), which is preferable given the potential downside of overestimation. In contrast, using the single-phase diagnostic with KPCA or MDS yields discouraging results due to the large positive RMSE values.

In general, using the IMM-PS1 or IMM-PS4 rate FDs extracted from two-phase immiscible simulations produces weak results for all dimensionality-reduction techniques, particularly for the P_{90} (notably for Ensemble 2, with RMSE $>$ 14%).

Finally, the IMM-PS11 rate FD merits special attention. It provides reasonably accurate cluster guidance, with acceptable percentile reconstructions (absolute RMSE ≤ 10 %) for t-SNE and MDS in both Ensembles 1 and 2, while for KPCA only the P_{50} of Ensembles 1 and 2 shows notable RMSE exceedances (about -13 %), whereas P_{90} and P_{10} remain acceptable. Considering the type of deviation, it is noteworthy that across both ensembles, percentile reconstructions for all workflows generally tend to be underestimated rather than overestimated: in fact, only one percentile exceeds the $+5$ % signed relative RMSE ($+7$ %, P_{50} with t-SNE on Ensemble 1), whereas all other exceedances of the 5 % RMSE bound occur in the negative direction.

That this workflow emerges as the most accurate after the early injection-rate FDs when cluster selection is guided by inertia is encouraging, yet it should be interpreted with caution given the pronounced fluctuations across K observed for a range of cluster counts, particularly in Ensemble 1 (Figure 4.9c). Although the present study identifies acceptable cluster counts, this may not always hold in future applications, and stronger fluctuations could increase the risk of more severe deviations if K is not selected carefully. Therefore, while this FD combined with t-SNE or MDS is promising when guided by inertia, accurate cluster selection remains essential, whereas the day-10 full-physics rate diagnostic (and likewise the single-phase diagnostic) proved far more stable and less sensitive to the exact choice of K. Overall, although these results are promising, this study concludes that using the early-rate and single-phase diagnostics remains more appropriate for clustering purposes in injection-rate percentile reconstruction. Sections 5.2 and 5.3 further outline the reasoning behind this conclusion.

As a final note regarding the IMM-PS11 FD, this study emphasizes the following. Focusing on FP-D10 t-SNE (Ensemble 1, Table 4.3) and IMM-PS11 t-SNE (Ensemble 2, Table 4.4), both select $K = 9$. However, for FP-D10 the total cost is estimated to be ~ 10 % of simulating the complete ensemble, whereas for the immiscible formulation it is approximately 15 %, resulting in an overhead difference of about 6 %. These percentages assume access to a GPU-accelerated full-physics flow simulator, which substantially speeds up the full-physics FD and causes even small differences in FD simulation time to translate into notable differences in relative cost. Without such acceleration, the overhead of the immiscible formulation would likely decrease proportionally and could even become slightly lower than that of the full-physics FD, making it a more competitive alternative in terms of computational efficiency.

4.5 Parameter Sensitivity: time–weighted η^2 and dGSA

As shown in Section 4.4, the internal cluster metrics suggest selecting $K = 9$ representative models for Ensemble 1 and $K = 6$ for Ensemble 2 when using the t-SNE workflow in combination with the FP-D10 injection-rate FD from the full-physics simulations. This choice provides an adequate reconstruction of the true ensemble percentiles for injection rate, with all percentile RMSE values remaining below the 5 % threshold, demonstrating that clustering can substantially reduce ensemble size while preserving the response distribution and thus enabling efficient uncertainty quantification.

Beyond ensemble reduction, clustering also provides a basis for assessing parameter influence through distance-based Generalized Sensitivity Analysis, as introduced in Section 3.6.1. Unlike variance-based methods, dGSA evaluates how strongly the values of an input parameter differ across the response clusters; parameters whose values show clear separation between clusters are considered influential, as they help to characterise the behavioural regimes represented by those clusters.

These insights could be useful for guiding uncertainty reduction. By identifying the parameters most responsible for distinct behavioral regimes, dGSA could highlights where additional data or tighter prior constraints would be most effective in narrowing ensemble variability while remaining cheap in terms

of computing.

To complement this perspective, a variance-based analysis was performed using the time-weighted partial η^2 introduced in Section 3.6.2. Figures 4.14(a) and 4.14(b) present the results for the two ensembles. These statistics quantify each parameter's unique contribution to the variability in injection rates and are only available when all realisations have been simulated over the full simulation duration. Both η^2 and permutation p-values are reported: η^2 is the effect size (the fraction of variance uniquely explained by a parameter), while p measures how surprising that effect would be under the no-effect null.

For Ensemble 1, the facies distribution model (`mod`) explains by far the largest share of variance ($\overline{\eta}^2 \approx 0.51$, $p = 0.005$), followed by the fault transmissibility multiplier (`mult`, ≈ 0.20, $p = 0.005$) and the structural cut (`cut`, ≈ 0.15, $p = 0.005$). Both `fault` and `top` have only minor or negligible effects; `fault` is small ($\overline{\eta}^2 \approx 0.01$, $p = 0.005$), and `top` shows no unique effect ($\overline{\eta}^2 \approx 0.00$, $p = 0.985$).

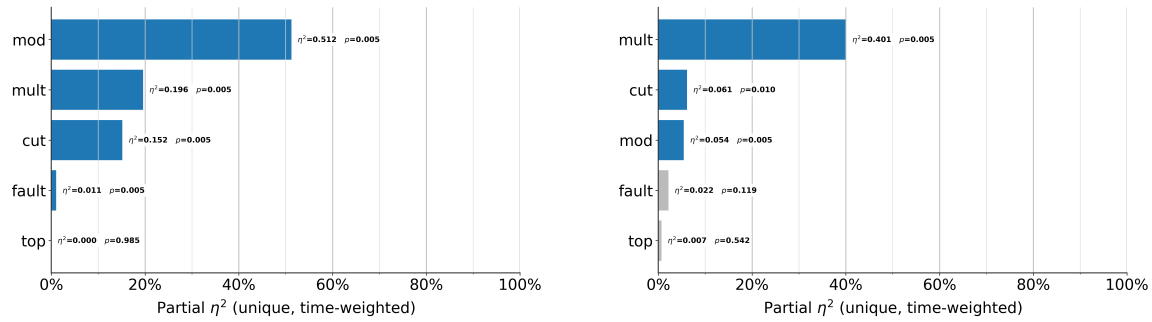
In Ensemble 2, the picture is simpler: behaviour is dominated by `mult` ($\overline{\eta}^2 \approx 0.40$, $p = 0.005$), with smaller contributions from `cut` ($\overline{\eta}^2 \approx 0.06$, $p = 0.01$) and `mod` ($\overline{\eta}^2 \approx 0.05$, $p = 0.005$), while `fault` and `top` again play a negligible role.

Having established these variance-based effects, it is instructive to compare them with the dGSA results shown in Figures 4.15(a) and 4.15(b). These figures present the dGSA results based on the nine and six response clusters, respectively for Ensembles 1 and 2, retrieved using the FD-D10 FD from the full-physics simulations. Bars show the normalised distance S for each parameter, and the dashed vertical line marks $S = 1$. Parameters with $S > 1$ (red bars) are deemed sensitive: they help separate the response clusters beyond random variation, making them influential. Conversely, $S \leq 1$ (blue) indicates little or no evidence of influence, as described in Section 3.6.1. Bar length ranks importance, and parameters are ordered accordingly. Upon comparing the results between η^2 and dGSA, a clear resemblance emerges between the two approaches.

For Ensemble 1, both η^2 and dGSA highlight `mod`, `mult`, and `cut` as the main controls on injection-rate behaviour, with both analyses clearly identifying `mod` as the most important contributor to ensemble variance. For Ensemble 2, the transmissibility multiplier is identified as the dominant driver in both analyses, again demonstrating dGSA's ability to correctly identify and prioritise parameter importance.

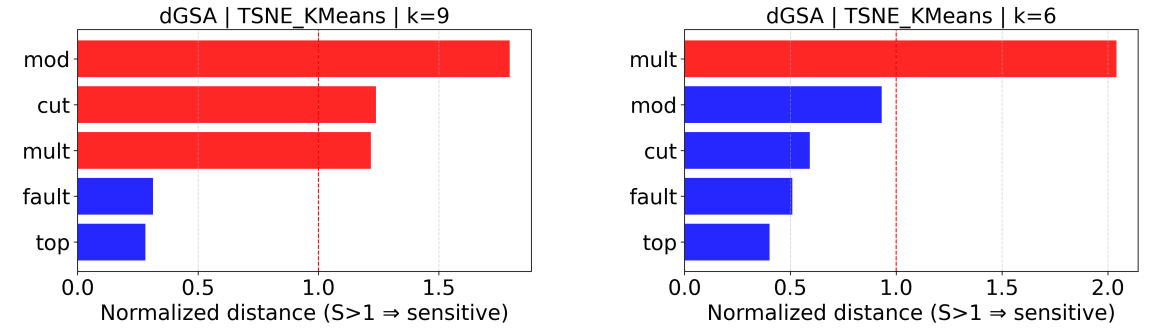
These findings indicate that dGSA, in combination with the clusters obtained using the FD-D10 + t-SNE workflow, is effective at identifying which parameters are most important for explaining the different response regimes. Even though dGSA relies only on the cluster structure at an early simulation time, it is able to reveal the same key drivers that would be identified by a full variance-based assessment over the entire simulation lifetime. This agreement also helps explain why clustering based on the t-SNE workflow combined with the FD-D10 FD from the full-physics simulations achieves such accurate percentile reconstructions: it correctly captures the parameter importance within the feature space.

Finally, the differing parameter influences observed across the two well locations (Ensemble 1 and 2) did not impede the FP-D10 + t-SNE workflow from achieving accurate clustering, indicating robustness to site-specific sensitivities. While this is encouraging, claims of general applicability require confirmation across additional geological settings, well configurations, and operational regimes.



(a) Ensemble 1 **(b)** Ensemble 2

Figure 4.14: Time-weighted partial η^2 for injection rate over the full simulation duration, with permutation p-values. The bars show each parameter's unique contribution to variance (effect size). (a) Ensemble 1: behaviour is mainly controlled by `mod`, followed by `mult` and `cut`, while `fault` has only a minor effect and `top` is negligible. (b) Ensemble 2: `mult` is the dominant driver, with smaller contributions from `cut` and `mod`, and little influence from `fault` or `top`.



(a) Ensemble 1 **(b)** Ensemble 2

Figure 4.15: Distance-based generalized sensitivity analysis for injection rate using response clusters obtained with the FP-D10 feature descriptor and t-SNE. Bars show the normalised distance S per parameter; the dashed line marks the sensitivity threshold $S{=}1$. Parameters with $S{>}1$ (red) are influential and are ordered by importance. (a) Ensemble 1: `mod`, `mult`, and `cut` emerge as the main controls, consistent with the η^2 results. (b) Ensemble 2: `mult` dominates, with `cut` and `mod` as secondary factors, again matching the variance-based analysis.

4.6 Informed Genetic Algorithm - Injection Rate Percentile Reconstruction

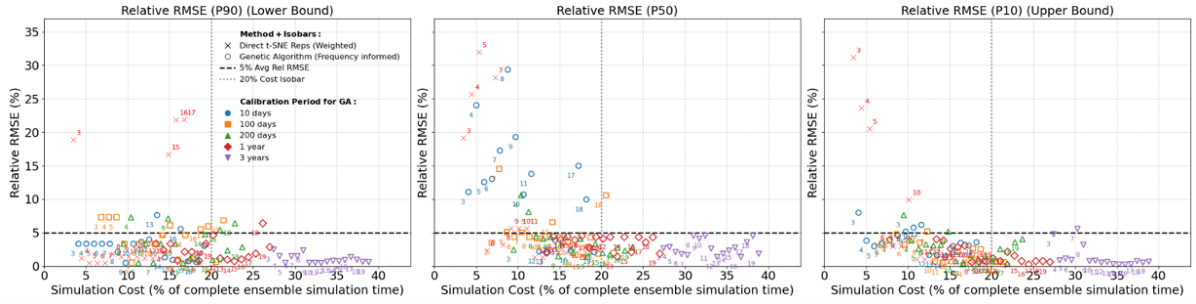
In Section 4.3.1 it was shown that t–SNE selection performs well across both ensembles when using early-time injection rates from the full-physics simulations (1 and 10 days) as flow-diagnostic input to the clustering workflow. Section 4.4 further demonstrated that internal clustering metrics provide a practical way to determine an appropriate number of clusters upfront. However, the results also showed that even a one-cluster difference can sometimes trigger accuracy losses exceeding 10–15 %, as all workflows occasionally exhibited erratic percentile reconstruction with sudden jumps in RMSE at specific subset sizes. These discontinuities are typically caused by the inclusion of a realization that disrupts the reconstructed distribution. While internal metrics proved effective in this study, the risk remains that selecting an inappropriate cluster count can substantially degrade performance.

To assess whether this volatility can be reduced, this study evaluated a calibration-informed Genetic Algorithm. The GA uses calibration RMSE over fixed time windows to guide subset selection as explained in Section 3.7, with the aim of improving percentile reconstruction at comparable or lower cost. Results are compared directly with t–SNE on percentile-wise relative RMSE (P_{90}, P_{50}, P_{10}), average

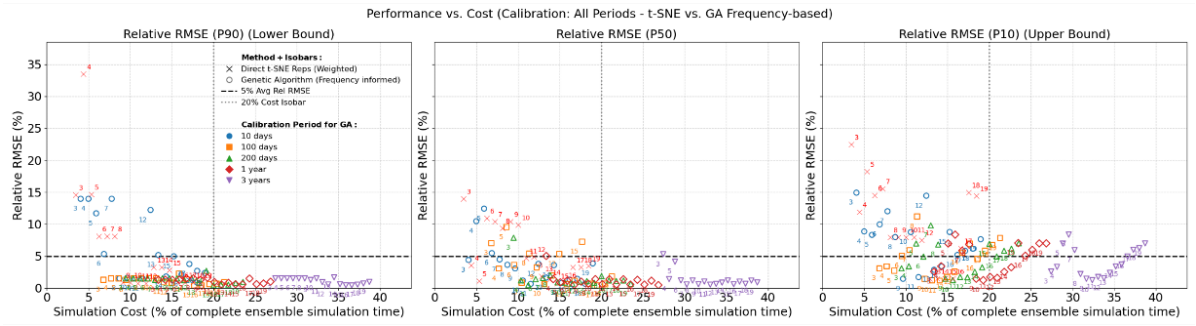
RMSE, and total simulation cost. In addition, calibration RMSE is examined as a potential indicator for selecting an appropriate subset size.

Figures 4.16(a) and 4.16(b) show the performance of the GA frequency-based selection in terms of relative RMSE for each percentile (P_{90}, P_{50}, P_{10}) as a function of simulation cost. The results are compared with the baseline of direct t–SNE selection using the 1–day full-physics flow diagnostic.

To provide an overall view, Figures 4.17 and 4.18 present the average relative RMSE across all three percentiles as a function of simulation cost, again including the 1–day t–SNE benchmark.



(a) Ensemble 1.



(b) Ensemble 2.

Figure 4.16: Relative RMSE for percentiles P_{90}, P_{50}, and P_{10} versus simulation cost for the GA frequency-based selection compared with direct t–SNE selection using the 1–day full-physics diagnostic. Dashed lines indicate the 5% RMSE threshold; dotted lines mark the 20% cost isobar.

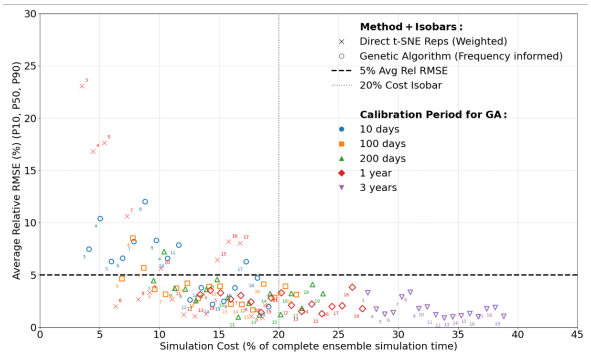


Figure 4.17: Ensemble 1.

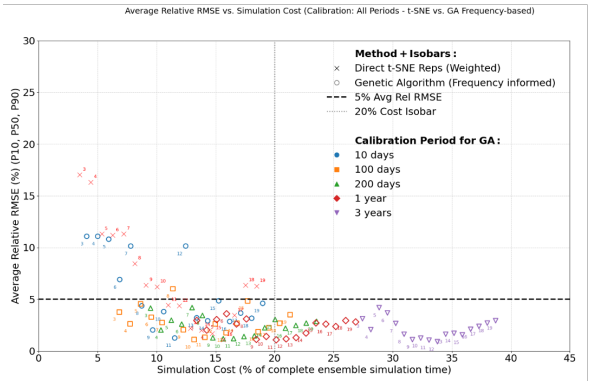


Figure 4.18: Ensemble 2.

Figure 4.19: Average relative RMSE across P_{90}, P_{50}, and P_{10} versus simulation cost for the GA frequency-based selection, compared with direct t–SNE (day–1 rate full-physics diagnostic).

Looking at the results in Figures 4.17 and 4.18, several observations can be made. First, the GA approach almost always avoids the worst-performing selections observed with direct t-SNE: while some t-SNE subsets exceed an average relative RMSE of $15\ \%$, GA-informed selections remain well below these peaks across all calibration periods.

Second, calibration length has a strong impact on accuracy. The 10-day calibration period is the least reliable, with most selections yielding average RMSEs between 5–10 $\%$. At 100-day calibration, three subsets also fall within this 5–10 $\%$ range, and for the 200-day calibration the four-model subset of Ensemble 1 slightly exceeds the 5 $\%$ threshold. By contrast, the 1-year and 3-year calibration periods consistently achieve average RMSEs below 5 $\%$, showing that longer calibration windows produce more robust percentile reconstructions.

This confirms that longer calibration periods allow the GA to match or surpass the accuracy of shorter calibrations with fewer selected realizations, which is a desirable feature that lets the modeller balance accuracy and simulation cost early in the workflow. In practice, choosing a slightly longer calibration window can reduce the total number of full-lifetime simulations required later, improving tractability and providing greater control over computational resources.

Examining the per-percentile performance in Figures 4.16(a) and 4.16(b) shows that, although the GA produces far fewer extreme fluctuations in average RMSE, noticeable deviations remain when percentiles are considered separately. This is most evident for the 10-day calibration, where several subsets show severe errors, for instance, the P_{50} of Ensemble 1 reaches almost 30 $\%$ relative RMSE for the subset with eight selected models. The 10-day calibration also contains many other subsets with smaller (yet still substantial) fluctuations in the 10–20 $\%$ range. The 100- and 200-day calibration periods perform significantly better, with the vast majority of reconstructed percentiles across all tested subset sizes remaining below the 5 $\%$ RMSE threshold. Only twice does a percentile reconstruction for the 100-day period yield a relative RMSE between 10 and 15 $\%$, compared with a single occurrence for the 200-day period. Going further, even the 1-year and 3-year calibrations contain a few subsets above the 5 $\%$ threshold, though these are rare, and none of the percentile reconstructions exceed 10 $\%$.

Overall, most subsets, apart from those derived from the 10-day calibration, remain comfortably below 5 $\%$ RMSE. Across both ensembles, P_{90} and P_{50} are reconstructed with good accuracy for all calibration lengths once the 10-day period is excluded. The main exception is the P_{10} of Ensemble 2, which shows slightly reduced accuracy across calibration periods, although RMSE values remain below 10 $\%$.

Selecting an Appropriate Subset Size

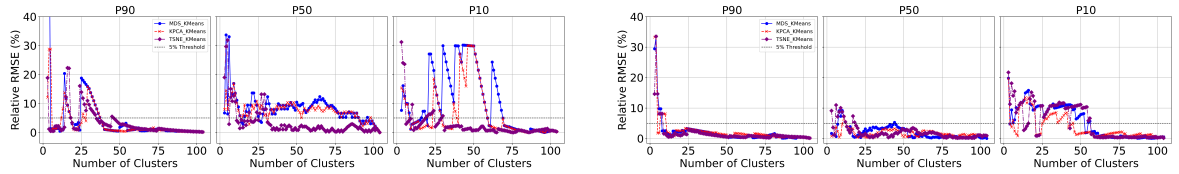
Although the GA does not completely eliminate fluctuations (“jumpiness”) in RMSE as a function of the number of selected models, it does dampen them, leading to more stable and reasonable error levels overall. The average RMSE across percentiles tends to converge more smoothly when guided by calibration-informed GA selection than when using direct cluster-selection methods. Nevertheless, the remaining variability highlights the need for an internal metric to guide the choice of an appropriate number of models.

One candidate for such a metric is the calibration RMSE. In principle, the calibration fit could indicate the point at which adding more models begins to degrade rather than improve accuracy. In this study, however, no strong correlation was found between calibration RMSE and the full-lifetime RMSE of the selected subsets up to the 1-year calibration period. Beyond one year, some correlation became visible, but it remained too weak to provide a reliable basis for determining subset size. Consequently, within the GA framework, no direct feedback mechanism was identified for selecting an appropriate number of models.

GA Using a Reduced Ensemble

The previous section showed that the informed GA method can produce a robust reconstruction of the ensemble percentiles, with reliability improving as the calibration period increases and becoming particularly promising from the 100-day window onwards. Building on this, the present study proposes an additional step that can be applied before calibrating on the full ensemble. Because the primary aim of this work is to evaluate different cluster-based approaches for effective model-selection workflows, one important finding is that t-SNE exhibits increasingly stable and accurate percentile reconstructions as the number of selected selected clusters grows. This finding can be exploited to the advantage of the proposed GA method.

To illustrate this, Figures 4.20(a) and 4.20(b) present the relative RMSE per percentile when reconstructing ensemble percentiles using unweighted selection, which directly reflects the true distribution percentiles of the reduced ensemble. The results indicate that, while KPCA (red) and MDS (blue) often produce unstable or erratic RMSE patterns, t-SNE (purple) begins to yield remarkably stable and accurate reconstructions once approximately 45 cluster representatives are selected, with RMSE values remaining below $5\,\%$. This trend holds across most percentiles, except for the P_{10} curve in Ensemble 2, which stabilizes around 60 selected models. Beyond this threshold, t-SNE-based reconstructions consistently approximate the full-ensemble percentiles with RMSE values below $5\,\%$, indicating that the reduced ensemble closely mirrors the distribution of the complete ensemble. Section 5.7.2 will discuss possible reasons for t-SNE's superior performance within this study.



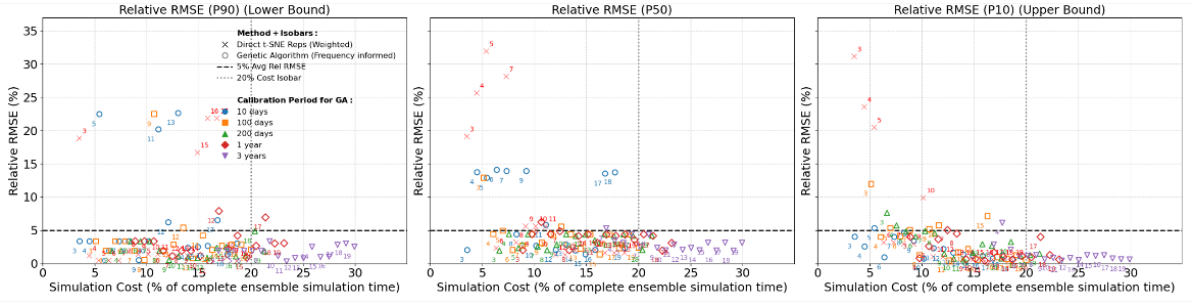
(a) Ensemble 1 **(b)** Ensemble 2

Figure 4.20: Relative RMSE per percentile when reconstructing ensemble percentiles using unweighted selection. While KPCA (red) and MDS (blue) often produce unstable RMSE patterns, t-SNE (purple) yields stable and accurate reconstructions once approximately 45 cluster representatives are selected, with RMSE values below $5\,\%$ for most percentiles. Only the P_{10} curve in Ensemble 2 stabilizes later, around 60 selected models. Beyond these thresholds, t-SNE-based reduced ensembles closely mirror the full-ensemble percentiles.

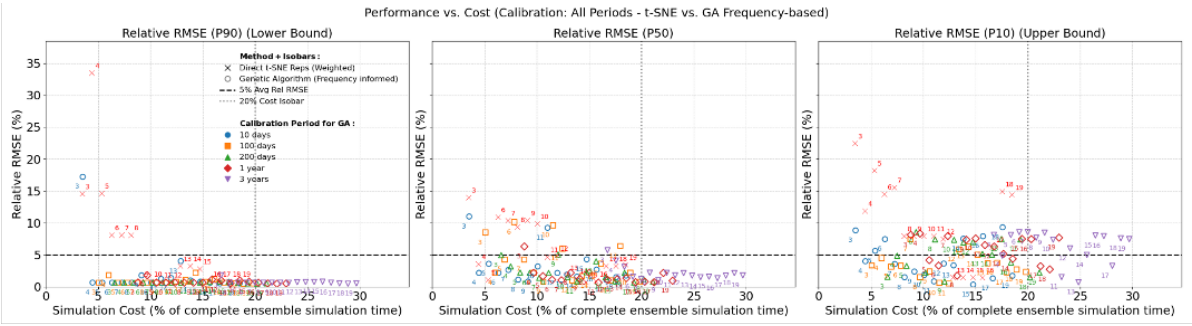
This opens the door to a two-stage strategy: by first applying t-SNE to reduce the ensemble to a more manageable subset (e.g., 60 models), the reduced set can be treated as a surrogate full ensemble for subsequent calibration and GA-based selection. This approach reduces the computational burden of calibration while preserving diversity and representativeness within the reduced ensemble.

Figures 4.21(a) and 4.21(b) present the GA performance for the reduced ensembles in terms of relative RMSE for each percentile (P_{90}, P_{50}, P_{10}) as a function of simulation cost. These results are shown alongside the baseline of direct t–SNE selection using the 1-day full-physics flow diagnostic.

Figures 4.22(a) and 4.22(b) display the average relative RMSE across all three percentiles as a function of simulation cost for the reduced ensembles, again benchmarked against the 1-day t–SNE results.

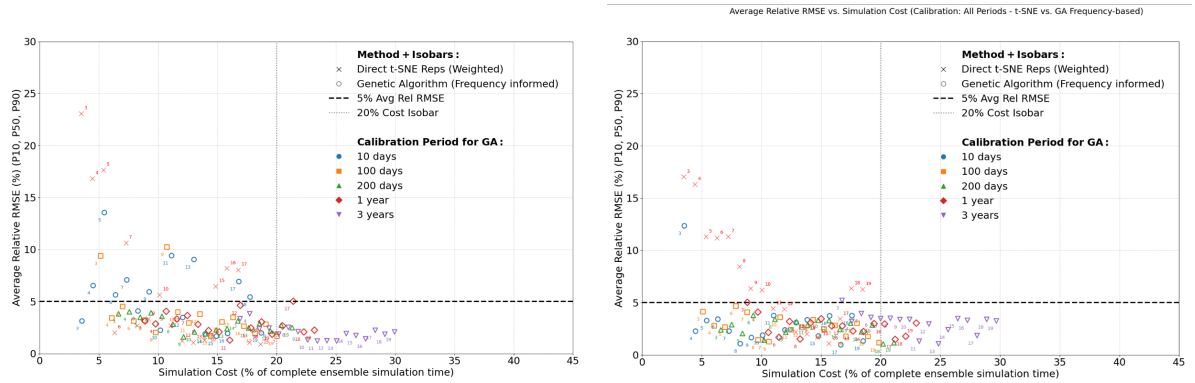


(a) Ensemble 1.



(b) Ensemble 2.

Figure 4.21: Relative RMSE for percentiles P_{90}, P_{50}, and P_{10} versus simulation cost for GA frequency–based selection on the t–SNE–reduced ensembles (e.g., 60 models), compared against direct t–SNE selection using the 1–day full–physics diagnostic. Dashed lines indicate the 5% RMSE threshold; dotted lines mark the 20% cost isobar.



(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.22: Average relative RMSE across P_{90}, P_{50}, and P_{10} versus simulation cost for GA frequency–based selection on the t–SNE–reduced ensembles, benchmarked against the 1–day direct t–SNE baseline. Dashed lines indicate the 5% RMSE target; dotted lines mark the 20% cost isobar.

Upon inspecting the average RMSE versus simulation cost results, it is evident that reducing the ensemble to 60 models before applying GA (Figures 4.22(a) and 4.22(b)) does not significantly affect the percentile reconstructions compared with applying GA directly to the full ensemble (Figures 4.17 and 4.18). The main difference is observed for the P_{10} of Ensemble 2 (Figure 4.21(a)), where more subsets exceed the $5\ \%$ threshold than when GA is applied to the complete ensemble (Figure 4.16(b)), although they still remain within the $10\ \%$ RMSE range.

A second observation is that the two-step approach (first reducing the ensemble with t-SNE and then applying GA-based selection) becomes particularly advantageous when longer calibration periods are used. Because longer calibration windows inherently require more simulation time, reducing the ensemble size upfront leads to a greater reduction in total computational cost. This, in turn, enables

the use of longer calibration periods, which generally yield more accurate reconstructions than shorter windows.

Overall, the accuracy of percentile reconstructions is largely preserved across most calibration periods when applying GA to the reduced ensemble, with the exception of the 10-day window, where reconstruction quality remains more variable and less reliable. However, this behaviour was also observed when using the same calibration window for the full ensemble.

Therefore, even though reducing the ensemble to approximately $60\ \%$ of its original size inevitably alters the true percentiles (due to changes in the underlying statistical distribution), the calibration process on this reduced set still produces reconstructions that remain closely aligned with those of the full ensemble, typically within a $5\ \%$ RMSE margin. This seemingly counterintuitive outcome may be explained by two key factors:

1. **Targeted diversity through clustering**: Dimensionality reduction and clustering (via t-SNE) ensure that the reduced ensemble retains the dominant modes of variability present in the full ensemble. By selecting representative realizations from structurally distinct clusters, the essential features of the system's behavior are preserved, even with fewer models.
2. **Reduced search space improves GA performance**: The Genetic Algorithm is more effective in a constrained and informative search space. Removing redundant or near-identical realizations reduces noise and the risk of convergence to suboptimal solutions, thereby improving both robustness and accuracy in percentile reconstruction.

In summary, reducing the ensemble not only cuts simulation cost but can also enhance the robustness of GA calibration. Crucially, in this study, the reduction does not come at the expense of percentile accuracy with the exception of the P_{10} of Ensemble 2, highlighting the efficiency gains of a targeted, structure-preserving model selection strategy.

GA vs Distance-Based Clustering

Although the GA approach has generally been shown to reconstruct percentiles accurately across a range of subset sizes for all calibration periods except the 10-day window, this study still recommends using the direct distance-based clustering approach to select representative ensemble models.

There are several reasons for this recommendation. Foremost among these is, unlike cluster selection, no feedback mechanism was identified for the GA method that can reliably indicate the optimal subset size. To illustrate this limitation, Section 4.4.2 showed that inertia would have guided t-SNE to select six clusters for Ensemble 1 and nine for Ensemble 2, yielding average RMSE values of approximately $2\ \%$ in Ensemble 1 and $6\ \%$ in Ensemble 2.

Looking at the GA results for Ensemble 1 (Figure 4.22(a)), none of the GA subsets achieved comparable accuracy at similar or lower simulation cost. For Ensemble 2 (Figure 4.22(b)), a few GA-selected subsets performed slightly better, such as with 8–9 models at 10 days, 3–6 models at 100 days, and 3 models at 200 days. However, there is no clear way to know a priori that these particular subsets should be chosen. As shown previously, the 10-day calibration period should generally be avoided because of its large variability across subset sizes, with even selections exceeding 10 models sometimes producing average RMSE values above $10\ \%$, making it difficult to determine a practical number of models. By contrast, from 100 days onwards, the results become more reliable as the number of selected models increases; yet, without clear guidance, choosing only three or four models risks selecting an inadequate subset.

This study would place considerably more confidence in selecting three or four models when using the 1- or 3-year calibration period, as results showed that reliable percentile reconstructions were achieved at such low model counts for both ensembles. However, these cases come with substantially higher computational overhead, undermining the goal of reconstructing percentiles as accurately and efficiently as possible. For instance, selecting three models with the 1-year or 3-year calibration window can be as costly as selecting nine or seventeen models with direct t-SNE.

Applying the GA method to the reduced ensemble does, however, reveal cases in both ensembles, and particularly in Ensemble 2, where GA outperforms t-SNE. Yet guidance for identifying appropriate

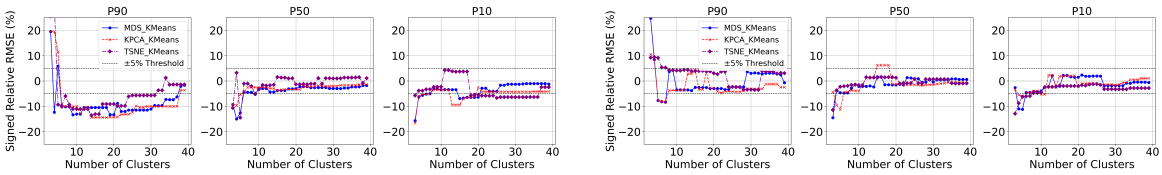
subset sizes remains unclear. Moreover, although reducing the ensemble to 60 models proved effective in this study, that number was chosen primarily to demonstrate the method's potential; in practice, the user would not know where to set this threshold. This introduces the risk of reducing the ensemble too aggressively, producing a set that no longer reflects the true ensemble percentiles and thus undermines subsequent calibration.

To illustrate, if this study had reduced the ensemble to 50 models, the P_{10} percentile of the reduced set would have shown an RMSE of $10\ \%$ relative to the corresponding percentile of the full ensemble (Figure 4.20(b)). Such a deviation would almost certainly have caused problems when calibrating subsequent model selections based on a reconstructed percentile that no longer aligns closely with the true ensemble percentile.

That said, t-SNE's stability compared with other dimensionality-reduction techniques as the number of clusters increases is noteworthy, making it an area that merits further investigation. Overall, while the GA method shows promise, its current inability to identify the correct number of models, combined with the fact that ensemble reduction has so far been tested on only two ensembles, highlights the need for further research to address these open questions.

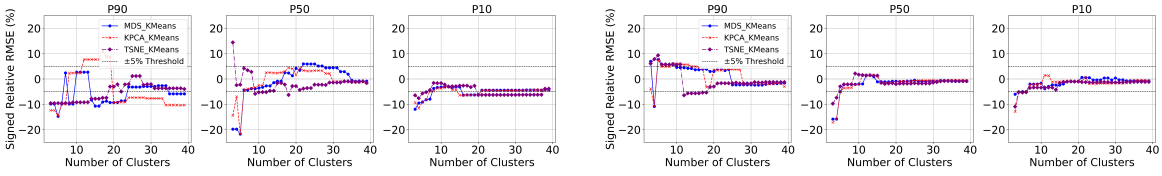
4.7 Results for Maximum Plume Extent

Figures 4.23(a) and 4.23(b) show the signed relative RMSE as a function of the number of selected clusters for the single-phase saturation-field diagnostic, with curves reported for percentiles P_{90}, P_{50}, and P_{10} and for three DR methods: t-SNE (purple), KPCA (red), and MDS (blue). Figures 4.24(a) and 4.24(b) present the corresponding results for the two-phase immiscible saturation-field diagnostic. Figures 4.25(a) and 4.25(b) display the results for the immiscible eight-producer cumulative-rate diagnostic. Finally, Figures 4.26(a) and 4.26(b) show the results for the 10-day injection-rate diagnostic from the full-physics Open-DARTS simulator, combined with t-SNE, which was identified in Sections 4.3.3 and 4.4 as the preferred workflow for injection-rate uncertainty quantification. This analysis is included to examine the impact of using the same clusters for plume-migration uncertainty quantification as for injection-rate uncertainty quantification, thereby achieving both objectives with a single clustering procedure. Additional details on the flow diagnostics are provided in Section 3.3.3.



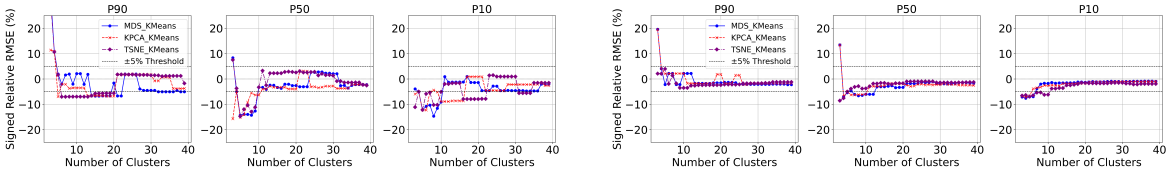
(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.23: Signed relative RMSE for percentiles P_{90}, P_{50}, and P_{10} as a function of the number of selected clusters for the single-phase saturation-field diagnostic (SP-SAT-PS1). Results are shown for t–SNE (purple), KPCA (red), and MDS (blue).



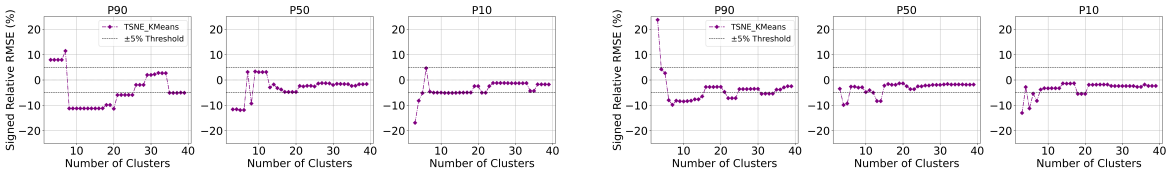
(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.24: Signed relative RMSE for P_{90}, P_{50}, and P_{10} for the two-phase immiscible saturation-field diagnostic (IMM-SAT-PS11) as a function of the number of selected clusters. Curves are shown for t–SNE (purple), KPCA (red), and MDS (blue).



(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.25: Signed relative RMSE for P_{90}, P_{50}, and P_{10} for the two-phase immiscible eight-producer cumulative-rate diagnostic (8-PRD-IMM-PS11). The curves compare t–SNE (purple), KPCA (red), and MDS (blue) for different numbers of selected clusters.



(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.26: Signed relative RMSE for P_{90}, P_{50}, and P_{10} for FP-D10 diagnostic from the full-physics Open-DARTS simulator. The diagnostic is combined with t–SNE clustering, which was identified as the preferred workflow for injection-rate uncertainty quantification.

Best Performing Dimensionality Reduction Workflow

An examination of the results across the tested DR and clustering workflows applied to the 3DSL flow diagnostics shows that no single method consistently dominates across all diagnostics, percentiles, and ensembles. Overall performance is comparable among t-SNE, KPCA, and MDS when combined with K-means. A slight, case-specific edge is observed for MDS with K-means: at the P_{90} of the immiscible eight-producer cumulative-rate diagnostic for Ensemble 1 (Figure 4.25(b)), MDS more frequently remains within the 5 % RMSE threshold, particularly when fewer than ten clusters are selected. This advantage, however, is modest and confined to that setting, and does not justify recommending MDS as a universally superior DR workflow for plume-migration uncertainty quantification based on these flow diagnostics.

Comparison of Flow Diagnostics

When comparing the different flow diagnostics, particularly for reconstructing the P_{50} and P_{10}, it can be observed that all three diagnostics applied within 3DSL are capable of producing reasonably accurate reconstructions of plume migration behavior even at low cluster counts ($K < 10$). Most reconstructed percentiles show RMSE values below $5\ \%$, and none exceed the $\pm 10\ \%$ threshold after selecting more than five clusters, with the exception of the P_{50} of the immiscible eight-producer cumulative-rate diagnostic in Ensemble 1. Notably, even the FP-D10 rate diagnostic from the full-physics simulations, combined with t-SNE, yields acceptably accurate results, again particularly for the P_{50} and P_{90}, with most reconstructed percentiles showing RMSE values below $5\ \%$ and, after more than six clusters, remaining within the $\pm 10\ \%$ RMSE bound for both ensembles.

To determine which of the tested flow diagnostics provide the best performance, the evaluation criteria were chosen based on the fact that, for quantifying the uncertainty of maximum plume migration, the upper percentiles are generally of greater interest because of the potential impact associated with plume extension, which could result in black swan events, as discussed in Section 3.1.2. Therefore, the focus was placed on the P_{50} and P_{10} reconstructions, with an emphasis on lower cluster counts ($K < 10$) to minimize computational costs. Based on these criteria, the single-phase and immiscible saturation-field diagnostics demonstrate the most consistent overall performance. For both Ensemble 1 and Ensemble 2, these diagnostics show that most reconstructed percentiles remain within $5\ \%$ RMSE at lower cluster counts compared to the other diagnostics.

What is particularly noteworthy is that the single-phase diagnostic performs on par with the immiscible formulation. This suggests that, despite the simplifications of the single-phase formulation—such as neglecting multiphase flow and buoyancy effects—it preserves the essential spatial dynamics of plume behavior, making it an effective flow diagnostic for subset selection that captures the variability of the full ensemble. From a modelling perspective, this is encouraging given the significant difference in computational cost: the single-phase diagnostic requires approximately 40 seconds per realization, whereas the immiscible formulation takes more than six minutes, as shown in Section 3.5.3. Section 5.4.1 will discuss potential reasons that could explain the comparable behavior between the two diagnostics.

Finally, it is worth highlighting the again surprisingly strong performance of the day-10 injection rate from the full-physics simulation as a flow diagnostic for maximum plume extent. Although the streamline-based diagnostics show more accurate and stable results in terms of percentile reconstruction acros both ensembles, it is noteworthy that using the cluster counts specified in Section 4.4 for t-SNE, namely nine clusters for Ensemble 1 and six clusters for Ensemble 2, it would have resulted in RMSE values for the reconstructed percentiles of approximately (P_{90}: 11 %, P_{50}: +4 %, P_{10}: −5 %) for Ensemble 1 and (P_{90}: −8 %, P_{50}: −2 %, P_{10}: −5 %) for Ensemble 2. Section 5.4.2 will discuss potential reasons why this approach may prove to be particularly effective.

To give the reader a sense of how this would look in terms of the actual versus reconstructed percentiles, Figures 4.27 and 4.30 illustrate the results for Ensembles 1 and 2. On the left of each row, the highlighted curves show the maximum plume distance over time for the selected realizations compared with all ensemble runs (in grey). On the right, the reconstructed P_{90}, P_{50}, and P_{10} percentiles for the selected subset are plotted against the true ensemble percentiles, illustrating the quality of the reconstruction: the closer the red line follows the blue line, the more accurate the reconstruction. In addition, Appendix E presents the same percentile comparisons based on the ten-day injection-rate flow diagnostic for both ensembles, evaluated separately for each direction (N, NE, E, SE, S, SW, NW). These directional results also show a reasonably close alignment between the reconstructed and true ensemble percentiles, especially for P_{50} and P_{10} over time in each direction.

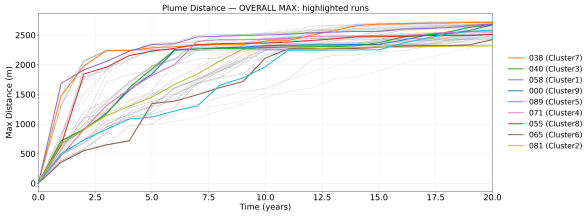


Figure 4.27: Maximum plume migration distance over time for Ensemble 1. Coloured curves represent the realizations selected via distance-based clustering on the ten-day injection-rate diagnostic (t-SNE, nine clusters), while grey curves denote all ensemble members.

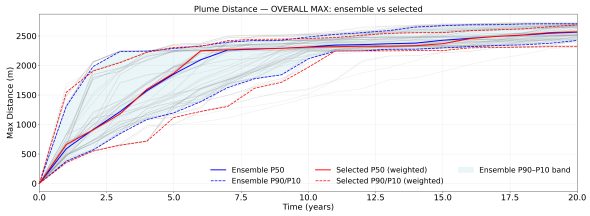


Figure 4.28: Reconstructed P_{90}, P_{50}, and P_{10} of the maximum plume migration distance for Ensemble 1, obtained from the selected realizations (red) and compared with the reference percentiles from the full ensemble (blue).

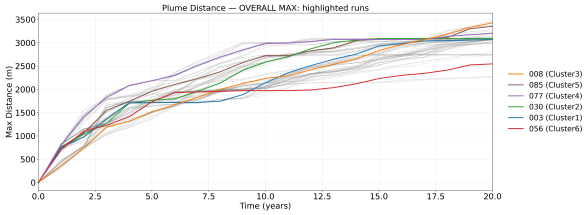


Figure 4.29: Maximum plume migration distance over time for Ensemble 2. Coloured curves represent the realizations selected via distance-based clustering on the ten-day injection-rate diagnostic (t-SNE, six clusters), while grey curves denote all ensemble members.

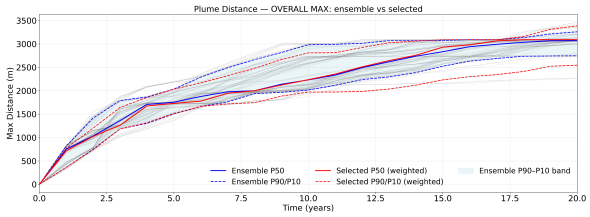
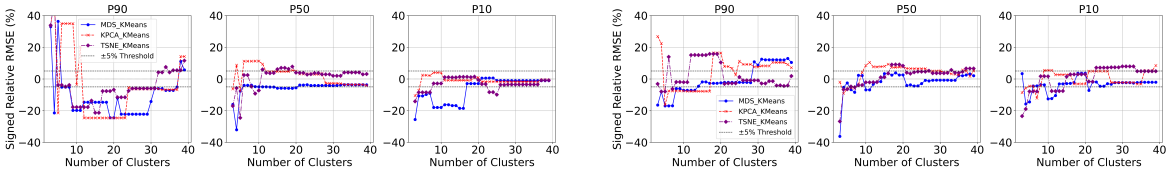


Figure 4.30: Reconstructed P_{90}, P_{50}, and P_{10} of the maximum plume migration distance for Ensemble 2, obtained from the selected realizations (red) and compared with the reference percentiles from the full ensemble (blue).

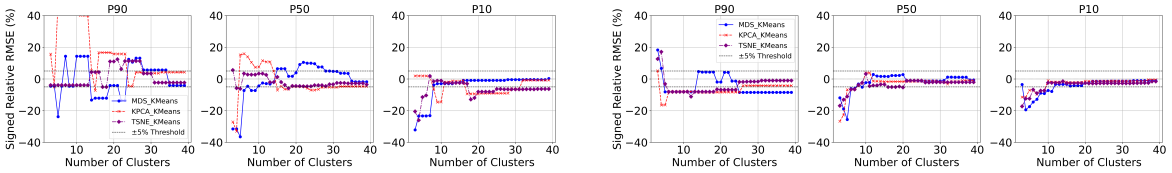
4.8 Results for Plume Areal Coverage

Figures 4.31(a) and 4.31(b) show the signed relative RMSE as a function of the number of selected clusters for the single-phase saturation-field diagnostic, with curves reported for the percentiles P_{90}, P_{50}, and P_{10} and for three dimensionality-reduction methods: t-SNE (purple), KPCA (red), and MDS (blue). Figures 4.32(a) and 4.32(b) present the corresponding results for the two-phase immiscible saturation-field diagnostic. Figures 4.33(a) and 4.33(b) show the results for the ten-day injection-rate diagnostic simulated with the full-physics open-DARTS simulator. This analysis is included to evaluate the impact of using the same clusters for plume-migration uncertainty quantification as for injection-rate uncertainty quantification, thereby enabling both objectives to be addressed through a single clustering procedure. Additional details on the flow diagnostics are provided in Section 3.3.2.



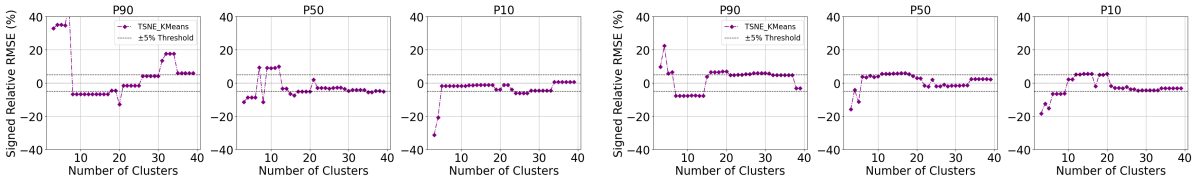
(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.31: Signed relative RMSE for percentile reconstructions of the plume areal coverage metric (P_{90}, P_{50}, P_{10}) as a function of the number of selected clusters for the single-phase saturation-field diagnostic (SP-SAT-PC). Results are shown for t–SNE (purple), KPCA (red), and MDS (blue).



(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.32: Signed relative RMSE for percentile reconstructions of the plume areal coverage metric (P_{90}, P_{50}, P_{10}) for the two-phase immiscible saturation-field diagnostic (IMM-SAT-PC) as a function of the number of selected clusters. Curves are shown for t–SNE (purple), KPCA (red), and MDS (blue).



(a) Ensemble 1. **(b)** Ensemble 2.

Figure 4.33: Signed relative RMSE for percentile reconstructions of the plume areal coverage metric (P_{90}, P_{50}, P_{10}) for the FP-D10 diagnostic from the full-physics Open-DARTS simulator. The diagnostic is combined with t–SNE clustering to evaluate whether the same clusters used for the preferred injection-rate uncertainty quantification can also be used for plume areal coverage.

Best Performing Dimensionality Reduction Workflow

Upon reviewing the results obtained from the various dimensionality-reduction and clustering workflows, no single approach consistently outperforms the others across all cases. If one were to select the most suitable workflow, applying t-SNE in combination with K-means appears to perform slightly better overall, particularly at lower cluster counts (fewer than ten), where it most often remains within the $5\ \%$ RMSE threshold across the reconstructed percentiles. This advantage, however, is modest and does not justify recommending t-SNE as a universally superior dimensionality-reduction workflow for plume areal-coverage uncertainty quantification based on these flow diagnostics.

Comparison of Flow Diagnostics

To determine which of the tested flow diagnostics provide the best performance, the evaluation criteria were chosen based on the fact that, when quantifying the uncertainty of plume areal coverage, the upper percentiles are generally of greater interest because of the potential impact associated with plume extension, as discussed in Section 3.1.2. Consequently, the analysis focused on the P_{50} and P_{10} reconstructions, with an emphasis on lower cluster counts to minimise computational costs. Based on these criteria, the single-phase and two-phase immiscible saturation-field diagnostics demonstrate the most consistent overall performance: for both Ensemble 1 and Ensemble 2, the majority of reconstructed percentiles remain within 5% RMSE at lower cluster counts (fewer than ten) compared with the other diagnostics. Neither method, however, is clearly more consistent than the other for low-K percentile reconstruction across both ensembles.

Although both diagnostics perform on par at low cluster counts ($K < 10$), the single-phase saturation-field diagnostic shows a stronger tendency toward instability in percentile-reconstruction accuracy across dimensionality-reduction workflows as K increases, with more frequent abrupt increases in RMSE. This behaviour is most visible in Ensemble 2 and, to a lesser extent, in Ensemble 1. Because increasing K subdivides existing clusters, the irregular RMSE pattern suggests that, for the single-phase formulation, the resulting partitions can become less aligned with the true ensemble structure at certain values of K. Additional partitions may capture noise rather than meaningful signal, yielding cluster representatives that reproduce the ensemble percentiles less effectively. Consequently, the reconstructed P_{90}, P_{50}, and P_{10} values drift from the true ensemble percentiles and RMSE increases.

By comparison, the two-phase immiscible diagnostic reconstructs percentiles more consistently across K, suggesting cleaner, more stable partitions and reduced dependence on the specific clustering configuration. This observation is consistent with the close relationship between the quantity of interest (CO_2 areal coverage) and the diagnostic based on the streamline-computed saturation field. As described in Sections 2.4.1 and 2.4.2, the two-phase immiscible formulation accounts for multiphase flow, including buoyancy effects, whereas the single-phase formulation neglects these and is effectively advection-dominated. Accordingly, the immiscible diagnostic tends to correlate more strongly with the metric of interest than the single-phase diagnostic.

The similarity in performance at low cluster counts indicates that the first-order driver of ensemble variability in plume areal coverage is advection, which both diagnostics capture. As K increases, the divergence in performance suggests that finer-scale variability becomes more relevant. These second-order effects appear to be governed by multiphase physics, including buoyancy and relative permeability effects, which are represented in the two-phase immiscible formulation but omitted in the single-phase formulation. To make this mechanism more explicit, it is helpful to view the problem through the induced feature space, in which pairwise relationships among realisations are embedded following dimensionality reduction as described in Section 3.4).

From this perspective, the single-phase diagnostic may produce pairwise distances and local neighbourhoods that become less congruent with the true ensemble variability at finer scales, owing to the absence of the governing multiphase physics. As K grows and clustering resolves smaller structures, this misalignment can make the partitions more susceptible to local noise, such that selected representatives explain less of the true ensemble variability and RMSE tends to fluctuate.

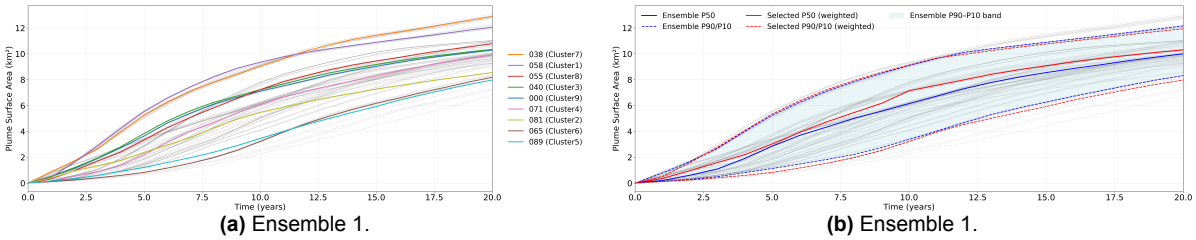
By contrast, the two-phase immiscible diagnostic appears to induce a feature space whose distance ordering and neighbourhood structure align more closely with the underlying ensemble variability. As K increases, the resulting partitions tend to remain coherent, so the selected representatives more consistently explain the ensemble structure rather than overfit small-scale variability, resulting in more stable and accurate percentile reconstructions. While these comparisons rely on dimensionality reduction embeddings and are therefore susceptible to local distortions, the observed instability patterns are consistently stronger for the single-phase diagnostic across all DR methods considered, suggesting that the qualitative conclusions are likely robust to the embedding choice.

At the same time, as noted in Section 4.7, the single-phase diagnostic often performs on par with the immiscible formulation at low cluster counts, particularly for the P_50 and P_10. This suggests that, despite neglecting multiphase flow and buoyancy, the single-phase formulation preserves the essential spatial

dynamics of plume behaviour, making it an effective diagnostic for subset selection that captures the variability of the full ensemble. From a modelling perspective, this is encouraging given the computational costs: the single-phase diagnostic requires approximately 40 seconds per realisation, whereas the immiscible formulation takes more than six minutes. Section 5.4.1 provides further discussion of the differences between the two diagnostics and possible reasons why the immiscible saturation field shows slightly higher overall accuracy and stability, while the single-phase field could still be a nearly comparable yet much cheaper alternative.

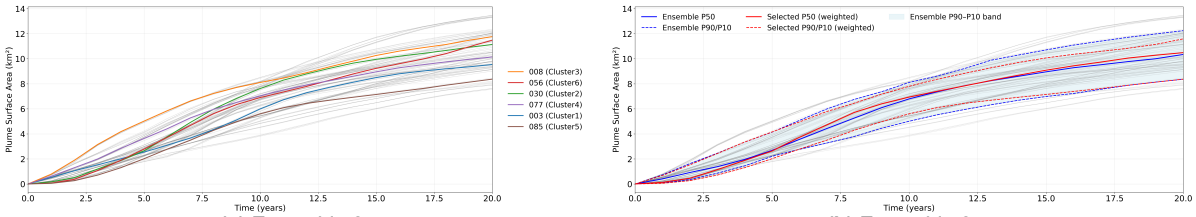
Finally, it is worth highlighting the strong performance of the ten-day injection-rate diagnostic from the full-physics simulations for plume areal coverage. Although the streamline-based diagnostics generally yield more accurate and stable percentile reconstructions, using the cluster counts specified in Section 4.4 for t-SNE (nine clusters for Ensemble 1 and six clusters for Ensemble 2) would produce approximate signed relative RMSE values for the reconstructed plume-coverage percentiles of $P_{90}: -6\%$, $P_{50}: +9\%$, and $P_{10}: -1\%$ for Ensemble 1, and $P_{90}: +6\%$, $P_{50}: +4\%$, and $P_{10}: -6\%$ for Ensemble 2. Section 5.4.2 will discuss potential reasons why this approach may prove to be particularly effective.

To illustrate the correspondence between the reconstructed and reference percentiles, Figures 4.34(b) and 4.35(b) present the results for Ensembles 1 and 2. In each row, the left panel shows the areal plume coverage over time for the selected realisations (highlighted) against all ensemble runs (grey). The right panel compares the reconstructed P_{90}, P_{50}, and P_{10} from the selected subset with the true ensemble percentiles; closer tracking of the reference curves (blue) by the reconstructed curves (red) indicates a more accurate reconstruction.



(a) Ensemble 1. **(b)** Ensemble 1.

Figure 4.34: Reconstruction of plume areal coverage percentiles using the FP-D10 + t-SNE workflow, where cluster selection is guided by internal cluster validation for Ensemble 1. Left: areal plume coverage over time for the selected realizations (highlighted) compared with the full ensemble (grey). Right: reconstructed P_{90}, P_{50}, and P_{10} percentiles from the selected subset (red) compared with the reference ensemble percentiles (blue).



(a) Ensemble 2. **(b)** Ensemble 2.

Figure 4.35: Reconstruction of plume areal coverage percentiles using the FP-D10 + t-SNE workflow, where cluster selection is guided by internal cluster validation for Ensemble 2. Left: areal plume coverage over time for the selected realizations (highlighted) compared with the full ensemble (grey). Right: reconstructed P_{90}, P_{50}, and P_{10} percentiles from the selected subset (red) compared with the reference ensemble percentiles (blue).

5

Discussion and Recommendations

In this chapter, Section 5.1 first explains why this study argues that cluster-population weighted percentile reconstruction performs better than the unweighted approach. Section 5.2 then discusses the performance of full-physics early injection rate diagnostics, their generally superior ability to reconstruct injection-rate percentiles with the distance-based method, and the sensitivity of these results to the chosen diagnostic time window. It also highlights a tendency to underestimate percentiles across varying K and briefly considers the impact of system compressibility on the early-rate diagnostic. Afterward, Section 5.3 evaluates streamline-based flow diagnostics, showing the stronger performance of the single-phase rate diagnostic compared to the immiscible one. Special attention is given to the difficulty in reconstructing the P_{90} of Ensemble 1, the effect of increasing pressure solves on the immiscible diagnostic, and the potential to generalize streamline-based diagnostics for injection-rate uncertainty quantification. Section 5.4 then examines plume migration metrics in more detail, with a focus on the immiscible saturation field diagnostic, which provided the most accurate percentile reconstruction across storage metrics. This section also highlights the surprising effectiveness of early-time full-physics injection rates as a dual-use diagnostic for plume migration UQ, offering potential for more cost-effective workflows. Next, Section 5.5 compares the most effective workflows for injection-rate and plume-coverage metrics when using direct k-medoids clustering on distance matrices instead of applying t-SNE, which was identified as the most effective dimensionality reduction workflow overall. This analysis helps isolate the effect of dimensionality reduction and supports its added value. Section 5.6 explores whether assuming global parameter configurations for minimum and maximum cases can identify extreme scenarios and whether early-rate full-physics and streamline diagnostics can capture these at low computational cost. Section 5.7 then discusses how the most effective UQ workflow found in this study, day-10 full-physics rate diagnostics combined with t-SNE, can be generalized along with the broader use of the distance-based clustering method. Section 5.8 considers the use of open-DARTS as an industry pre-screener for CO_2 storage projects. Finally, Section 5.9 summarizes remaining limitations and recommends future research directions, and Section 5.10 closes with a reflection on the overall purpose and implications of this study for sustainable reservoir engineering.

5.1 Weighted vs Unweighted Percentile Reconstruction

As shown in Section 4.2, the cluster–population–weighted percentile reconstruction (introduced in Section 3.5.1) consistently outperforms the unweighted reconstruction for varying subsets selected with the flow–based distance clustering workflow. The unweighted approach exhibits unstable behaviour, with strongly fluctuating RMSE as the number of clusters K varies.

This study beliefs the mechanism explaining this is relatively straightforward. As K changes, one coherent group of realizations that lie close together in feature space and produce similar reservoir responses can potentially be split into multiple clusters or merged into one. In the unweighted reconstruction, each cluster representative has the same weight, so a split effectively counts that group multiple times, while a merge counts it only once. This distorts the reconstructed percentiles. For example, if an additional partition causes two selected representatives to come from what is essentially the same influence re-

gion (now split across two clusters), both are counted, inflating that region's influence and amplifying RMSE variability.

By contrast, the weighted reconstruction assigns each representative a weight proportional to its cluster population (see Section 3.5.1). When a region is split, the total weight of that region is divided proportionally across its representatives; when regions merge, their weights add. This preserves the relative influence of the underlying regions and stabilizes the percentile curves. Beyond split and merge effects, weighting also improves robustness to outliers and to small or singleton clusters. In the unweighted scheme each representative has leverage $1/C$, so even a singleton outlier can move the percentile reconstruction materially. With cluster-population weights $w_i = n_i/N$, small or singleton clusters receive proportionally less influence, reducing their ability to distort the reconstructed percentiles. When all cluster sizes are equal, the weighted and unweighted reconstructions coincide.

Figure 5.1 illustrates the principle with an example. The same small dataset is partitioned into $K = 3$, $K = 4$, and $K = 5$ clusters, with the symbol counts n indicating cluster populations. For each partition, the median (P_{50}) is reconstructed from the cluster representatives. The unweighted P_{50} jumps markedly with K ($25 \rightarrow 22 \rightarrow 26$), because splits and merges change how often a region is represented, and small or singleton clusters carry the same weight as large clusters. The weighted P_{50} varies much less across K ($24 \rightarrow 26 \rightarrow 25$), because normalised cumulative weights (NCW) derived from the cluster populations are used in the percentile interpolation. In short, weighting by cluster population reduces sensitivity to arbitrary partitioning and to small extreme clusters, yielding a more faithful and stable percentile reconstruction as K changes.

Reconstructing the P_{50}: Unweighted vs Weighted. Ensemble size N = 13; cluster size n; value represents the value of the cluster representative. K is the total number of clusters.

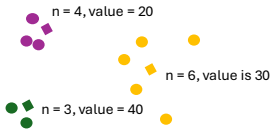
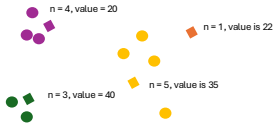
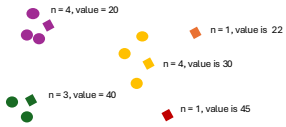
K = 3 , data sorted [20, 30, 40]	K = 4 , data sorted [20, 2, 35, 40]	K = 5 , data sorted [20, 22, 30, 40, 45]
		
Unweighted: Purple (n=4): 20, Orange (n=6): 30, Green (n=3): 40 Cumulative % Ranks: 20 (33.33%), 30 (66.67%), 40 (100%); P50 falls between 20 (at 33.33%) and 30 (at 66.67%). P50 = 20 + ((50 - 33.33) / (66.67 - 33.33)) * (30 - 20) = 20 + 0.5 * 10 = **25**	Unweighted: Purple (n=4): 20, Orange (n=1): 27, Yellow (n=5): 35, Green (n=3): 40 Cumulative % Ranks: 20 (25%), 22 (50%), 35 (75%), 40 (100%); P50 falls exactly on 27 (at 50%). No interpolation between values needed. Result (Unweighted P50): **22**	Unweighted: Purple (n=4): 20, Orange (n=1): 22, Yellow (n=4): 30, Green (n=3): 40, Red (n=1): 45 Cumulative % Ranks: 20 (20%), 22 (40%), 30 (60%), 40 (80%), 45 (100%); P50 falls between 22 (at 40%) and 30 (at 60%). P50 = 22 + ((50 - 40) / (60 - 40)) * (30 - 22) = 22 + (10 / 20) * 8 = 22 + 0.5 * 8 = **26**
Weighted: Designed Values & Weights: (20, n=4), (30, n=6), (40, n=3) Calculation (NCW & Interpolation): NCW for 20: 4/13 ≈ 0.3077 NCW for 30: (4+6)/13 ≈ 0.7692 P50 falls between 20 (0.3077) and 30 (0.7692). P50 = 20 + ((0.50 - 0.3077) / (0.7692 - 0.3077)) * (30 - 20) ≈ 20 + 0.4166 * 10 ≈ **24.2**	Weighted: Designed Values & Weights: (20, n=4), (22, n=1), (35, n=5), (40, n=3) Calculation (NCW & Interpolation): NCW for 20: 4/13 ≈ 0.3077 NCW for 22: (4+1)/13 ≈ 0.3846 NCW for 35: (4+1+5)/13 ≈ 0.7692 P50 falls between 22 (0.3846) and 35 (0.7692). P50 = 22 + ((0.50 - 0.3846) / (0.7692 - 0.3846)) * (35 - 22) ≈ 22 + 0.3 * 13 ≈ **25.9**	Weighted: Designed Values & Weights: (20, n=4), (22, n=1), (30, n=4), (40, n=3), (45, n=1) Calculation (NCW & Interpolation): NCW for 20: 4/13 ≈ 0.3077 NCW for 22: (4+1)/13 ≈ 0.3846 NCW for 30: (4+1+4)/13 ≈ 0.6923 P50 falls between 22 (0.3846) and 30 (0.6923). P50 = 22 + ((0.50 - 0.3846) / (0.6923 - 0.3846)) * (30 - 22) ≈ 22 + 0.375 * 8 = 22 + 3 = **25**

Figure 5.1: Illustrative example of reconstructing the median P_{50} from cluster representatives for the same dataset partitioned into K=3, K=4, and K=5 clusters. Blocks are cluster representatives with their corresponding value shown. Dots with similar color are related cluster members and n is the cluster population. In the unweighted reconstruction every representative has equal weight, so splits/merges across partitions change regional influence and the reconstructed P_{50} jumps (e.g., $25 \rightarrow 22 \rightarrow 26$). In the weighted reconstruction representatives are weighted by cluster population (w_i=n_i/N) and the percentile is obtained via normalized cumulative weights, preserving regional influence and yielding a more stable P_{50} (e.g., $24 \rightarrow 26 \rightarrow 25$).

5.2 Evaluating Early-Time Full-Physics Rate Diagnostics for CO_2 Injection Rates

This section examines why early-time injection rates obtained from full-physics simulations can serve as effective flow diagnostics for uncertainty quantification of CO_2 injection rates. The first part analyzes the physical and statistical reasons that early injection responses provide a stable and informative basis for clustering (Section 5.2.1). The subsequent analysis investigates how the choice of diagnostic time horizon (e.g., day 1, day 10, day 100) affects clustering quality and percentile reconstruction (Section 5.2.2). Systematic biases in reconstructed ensemble percentiles are then explored, focusing on the tendency toward underestimation and its link to cluster density and selection mechanics (Section 5.2.3). Finally, the influence of storage-system compressibility on the usefulness of early-time diagnostics and on the optimal simulation window for different reservoir settings is discussed (5.2.4).

5.2.1 Why Early-Time Injection Rates Are Effective Flow Diagnostics

A key finding of this study is that, for reconstructing the ensemble percentiles of injection-rate behaviour (P_{90}, P_{50}, P_{10}), early-time injection-rate diagnostics from full-physics simulations yield better clustering performance than diagnostics computed with the 3DSL streamline simulator in either single-phase or immiscible formulation. The RMSE of the reconstructed percentiles is lower and more stable across varying K, and performance remains superior when cluster selection is guided by internal clustering metrics; see Sections 4.3.3 and 4.4.

Why early-time injection rates can serve as effective flow diagnostics for injection rate uncertainty quantification becomes clear when comparing streamline-based and full-physics flow simulations. One might initially expect streamline simulations to be better at capturing global reservoir heterogeneity, particularly at early stages. Because streamlines trace particle trajectories along the instantaneous velocity field, which is determined by pressure gradients and permeability (and by phase mobilities in two-phase immiscible formulations; see Section 2.4), they provide an immediate picture of dominant flow patterns and offer a broad view of the system from the outset. In this sense, they can "see further" into the reservoir and quickly identify major flow paths, which may make them more effective for clustering based on large-scale heterogeneity that influences long-term flow behavior.

By contrast, during the early-time phase of a full-physics simulation, the solution is often dominated by the semi-analytical well model, which emphasises near-wellbore effects. These are largely governed by local injectivity, itself controlled by small-scale heterogeneity around the well. At first sight, this suggests that early-time full-physics outputs might be less sensitive to broader reservoir characteristics that influence long-term injection behavior, such as transmissibility across sealing faults.

However, two factors could explain why early-time full-physics simulations still yield meaningful and stable clustering results:

1. **Injectivity-driven flow behavior is stable and hierarchical.**
   Injection-rate behavior and the resulting clustering patterns ("banding") are primarily governed by near-well injectivity. This produces a hierarchical structure in flow paths that remains relatively stable over time, or at least does not significantly disrupt the initial banding. Because these early-time dynamics are dominated by local injectivity contrasts, even the first few simulation days (i.e. within the first 100 days) are sufficient to establish a representative ranking of model behavior. As a result, the clustering outcomes are relatively robust to the specific time window used.

2. **Large-scale reservoir features still influence early full-physics response.**
   While full-physics simulators do not explicitly trace global flow paths like streamlines, they still respond to broader system properties. For instance, low transmissibility across faults increases resistance to flow, leading to measurable pressure build-up or deviation even in the first simulation steps. Conversely, highly transmissible faults allow early pressure equalization. These large-scale features, though not directly visualized, leave a detectable signature in early-time injection rates and pressure responses. Consequently, the full-physics simulation implicitly encodes important system-wide effects, which in this study appear sufficient to enable informative clustering even at the earliest stages.

In contrast, streamline-based flow diagnostics, although designed to efficiently screen large-scale heterogeneity, remain reduced-physics approximations of full-physics behavior. Their diagnostic-implied ranking of realizations is unlikely to map one-to-one onto the ranking derived from full-physics rates. This rank misalignment introduces variability (noise) into the distances used for clustering, reducing accuracy and stability.

To quantify this effect, Spearman's rank correlation (ρ), a non-parametric measure of agreement between two sets of rankings (ranging from -1 for perfect inverse association to $+1$ for perfect agreement) [57], was computed between the diagnostic distance matrices and the pairwise absolute differences in cumulative injected volume at 20 years. This was done to compare the ranking performance of the day-10 full-physics rate FD, which proved to be the most effective FD (Section 4.3.3), with the streamline single-phase and 11-pressure-solve two-phase immiscible rate diagnostics, which were identified as the most appropriate streamline-based FDs for injection-rate uncertainty quantification (Section 4.3.2). Higher ρ values indicate that realizations judged far apart by the diagnostic are also far apart in terms of stored volume, providing a test of alignment between the dissimilarity structure of the diagnostic and the storage metric. For Ensemble 1, ρ was 0.94 for the full-physics FD, 0.73 for the single-phase diagnostic, and 0.55 for the immiscible diagnostic with 11 pressure solves, respectively. For Ensemble 2, the corresponding values were 0.85 (full-physics), 0.53 (single-phase), and 0.54 (immiscible, 11 pressure solves). These findings confirm that early-time injection-rate diagnostics from the full-physics simulator offer the most reliable proxy for long-term storage behaviour in terms of relative model ranking, surpassing the streamline approaches and enabling more effective clustering.

To illustrate how rank alignment affects clustering and percentile reconstruction, Figure 5.2 contrasts the diagnostic-induced ranking with the ranking implicit in the full-physics results using an exaggerated schematic. On the left, the full-physics injection-rate results for all realizations are shown with four clear bands (natural clusters) for illustration; the blue dashed lines denote the true ensemble percentiles. The same realizations $R_1, \ldots, R_n$ are then evaluated by three flow diagnostics: the early-rate full-physics diagnostic (FP FD), the single-phase streamline diagnostic (SP FD), and the immiscible two-phase diagnostic with 11 pressure solves (IM FD). Each diagnostic induces its own ordering (equivalently, distance structure) over realizations, with observed Spearman rank ranges indicated: FP FD $0.85 < \rho < 1$, SP FD $0.5 < \rho < 0.75$, IM FD $\rho < 0.55$. A higher ρ preserves the full-physics ensemble ranking more closely: FP FD largely keeps bands intact, with only minor within-band perturbations, so the reconstructed percentiles (red) closely follow the true percentiles (blue). With SP FD, bands begin to mix; cluster purity drops and reconstructed percentiles deviate more, as illustrated by the larger deviation in the reconstructed P_{50}. With IM FD, the lowest ρ implies the greatest rank misalignment and the noisiest pairwise distance matrix used for clustering, so realizations that appear diverse under the diagnostic may collapse to similar behavior under full physics, biasing the reconstruction and producing the largest departures from the truth. Further analysis of streamline-based flow diagnostics for injection-rate behaviour is provided in Section 5.3.

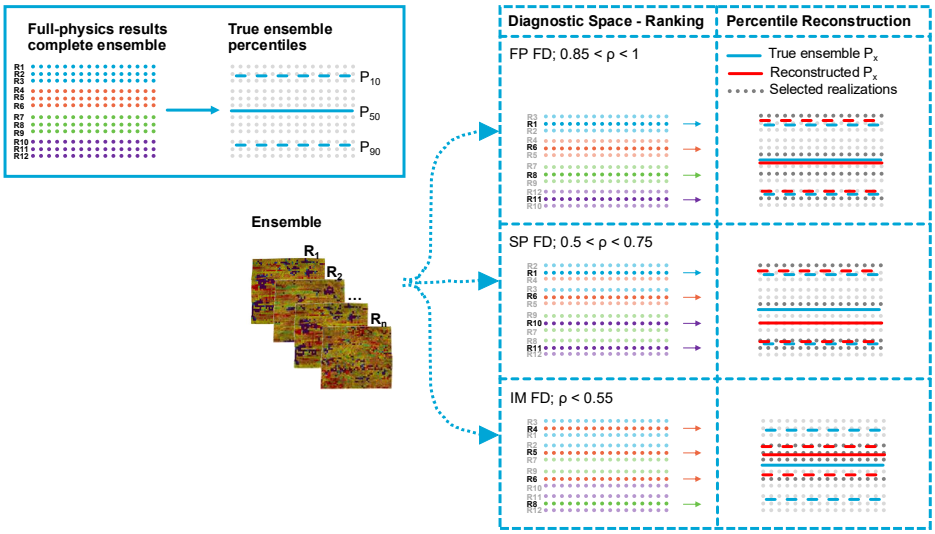
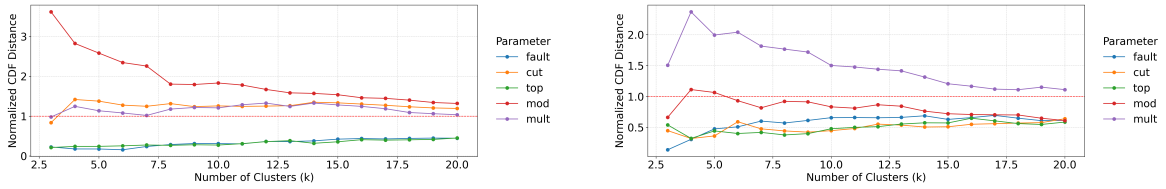


Figure 5.2: Schematic illustrating how rank alignment between flow diagnostics and full-physics results affects flow diagnostic-based clustering and percentile reconstruction. The FP FD (0.85 < ρ < 1) largely preserves the realization ranking of the full-physics ensemble, producing reconstructed P_x close to the truth; the SP FD (0.5 < ρ < 0.75) shows partial band mixing and larger P50 bias; the IM FD (ρ ρ 0.55) exhibits the greatest rank misalignment, the noisiest distance structure for clustering, and the largest departures of reconstructed percentiles from the true ensemble percentiles. The diagnostic space refers to the injection-rate responses predicted by the flow diagnostic

To conclude, the strong spearman rank coefficients obtained with the full-physics flow diagnostic further support the conclusion that, in this study, long-term injection-rate behaviour and the resulting clustering patterns seem to be primarily governed by near-well injectivity, which is effectively captured by the early-time injection-rate full-physics diagnostic.

In addition, to support the claim that the early-rate diagnostic also responds to broader system properties, such as low transmissibility across faults, it is useful to examine the dGSA results over varying cluster counts for the day-10 diagnostic combined with t-SNE (Figure 5.3). As outlined in Section 4.5, η^2 analysis showed that `mod`, `mult`, and `cut` were the key parameters driving ensemble variability in Ensemble 1, whereas in Ensemble 2 the injection-rate behaviour was primarily controlled by `mult`. The dGSA, which flags parameters as influential once they exceed the Normalized CDF threshold ($S > 1$) as explained in Section 3.6.1, revealed that these dominant parameters were consistently captured by the early-injection-rate diagnostic from the first cluster selections onward. This is particularly interesting for Ensemble 2, where `mult` is the dominant variance driver. It was correctly detected by the early-rate diagnostic, emphasizing its ability to encode broader system-wide effects as signal, even under conditions where it's the main driver. Additional results for the day-1 and day-100 diagnostics are provided in Appendix F, Figures F.1 and F.2. These show that the day-1 diagnostic tends to underestimate the importance of `mult`, suggesting that injection rates require some time to develop a meaningful ranking that also captures the system-wide effects established slightly later in the simulation.



(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.3: Distance-based generalized sensitivity analysis over increasing cluster counts for FP-D10 combined with t-SNE. The dGSA curves show how influential each parameter is according to the clusters formed. Parameters with a normalized CDF exceeding $S > 1$ are considered influential, indicating that their values vary strongly across response clusters and thus have a strong impact on cluster formation. Results are shown for Ensemble 1 and Ensemble 2.

5.2.2 Sensitivity of Clustering to the Diagnostic Time Horizon

As shown in Section 4.1.1, clustering performance for both ensembles tends to increase slightly when moving from early-time injection rates day 1 to 100 days, with minimal change between the 10-day and 100-day diagnostics. One reason that even the day-1 diagnostic yields reasonably good results, despite further improvements with longer horizons, is that early injection rates primarily reflect well injectivity, which has been identified as a key driver of ensemble variability (Section 5.2.1).

Under this assumption, the ensemble can be viewed as forming distinct bands in rate space that correspond to initial injectivity levels. In this study, these bands are primarily set by the six *mod–cut* permeability configurations, with `mult` exerting a substantive additional influence on early-time rates and thus on band separation; by contrast, *fault* and *top* play a comparatively minor role, as quantified by η^2 (see Section 4.5). As time progresses, nonlinear CO_2 flow processes (multiphase effects, mobility changes, and saturation-front movement) can change the relative ordering of these bands. Bands may partially overlap or even swap order. If a diagnostic snapshot is taken during such a period (e.g., at day 1 or day 10), two otherwise distinct groups can appear merged. This apparent merging can cause the clustering to collapse distinct clusters, which in turn degrades CDF-based percentile reconstruction and lowers overall clustering performance. With slightly longer simulation times, this overlap diminishes, resulting in clearer distance distinctions between the bands and improved clustering performance.

This trend is supported by Spearman rank correlations computed between the diagnostic distance matrices of the injection rates at 1, 10, and 100 days and the pairwise absolute differences in cumulative injected volume at 20 years (Section 5.2.1). For Ensemble 1, the coefficients increased from 0.84 to 0.94 to 0.98; for Ensemble 2, they increased from 0.60 to 0.85 to 0.97, indicating a stabilizing rank structure with increasing diagnostic horizon.

Finally, Figure 5.4 illustrates this hypothesis: injectivity-driven bands that should be separated in feature space to accurately cluster and capture long-term behavior may, at early times, collapse into fewer apparent groups, reducing clustering performance. This is a transient windowing effect: as bands pass through one another, temporary overlap can obscure their separation at a given snapshot, but the bands separate again at later times. In addition, Appendix F, Figure F.3 shows the injection rates obtained with the full-physics simulator for all realizations in both ensembles over the first 100 days. These plots show that, especially during early simulation times, realizations form bands whose rankings often cross, leading to overlap in the 1 to 10 day window; thereafter, the hierarchy stabilizes.

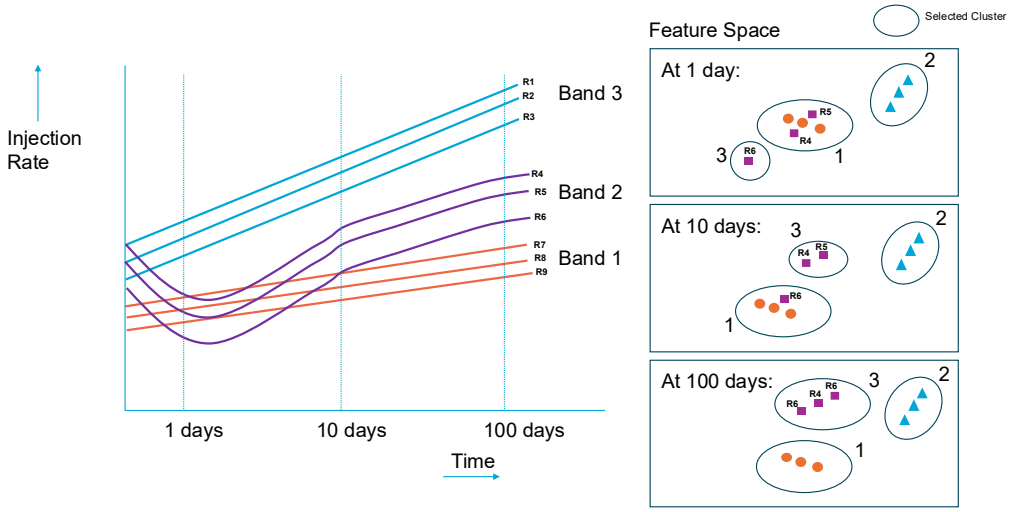
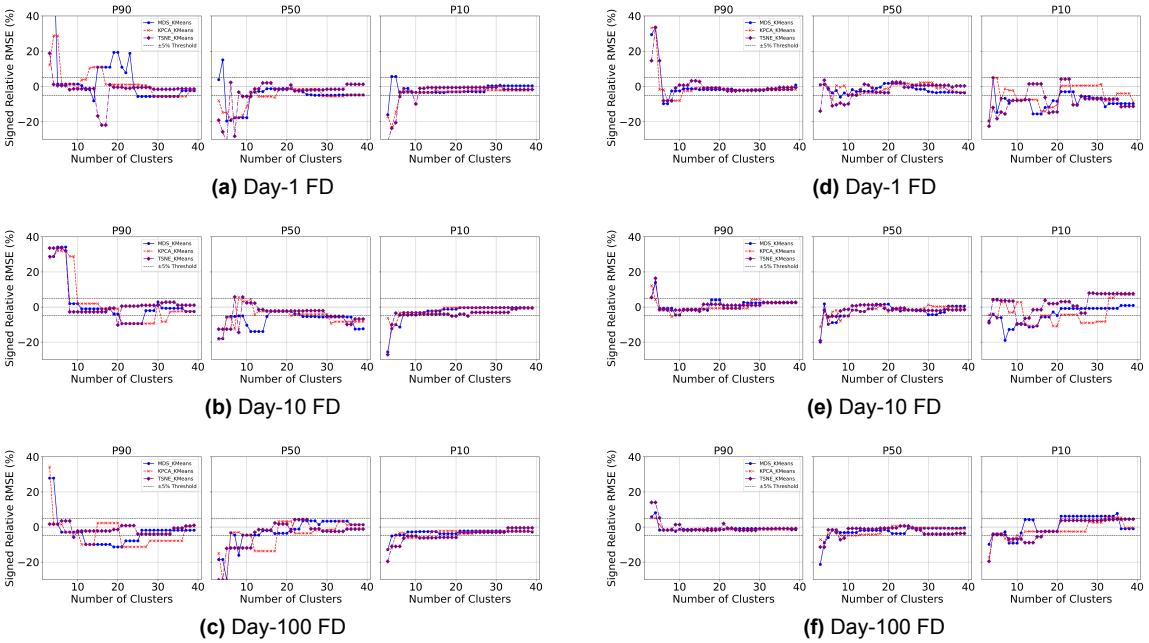


Figure 5.4: Illustration of how early-time injection rates reflect distinct injectivity-driven bands (left), which map to well-separated clusters in feature space (right) as the diagnostic horizon increases. At early stages, multiphase flow effects cause bands to overlap, leading to artificial merging of clusters, particularly at day 1, less so at day 10, and absent by day 100,resulting in increased clustering performance over time.

5.2.3 Underrepresentation of Percentiles

Figure 5.5 show the signed relative RMSE as a function of the number of selected clusters for Ensembles 1 and 2 when using early-time injection-rate flow diagnostics obtained with the full-physics simulator.



(a) Day-1 FD **(d)** Day-1 FD

(b) Day-10 FD **(e)** Day-10 FD

(c) Day-100 FD **(f)** Day-100 FD

Ensemble 1 Ensemble 2

Figure 5.5: Signed relative RMSE (%) as a function of the number of selected clusters for Ensembles 1 and 2. Early-time injection-rate flow diagnostics (Day 1, Day 10, Day 100) are computed with the full-physics Open-DARTS simulator (Section 3.3) and used for distance-based clustering. Results are reported for P_{90}, P_{50}, and P_{10}, and for three DR workflows: t-SNE (purple), KPCA (red), and MDS (blue). Left: Ensemble 1 (a–c). Right: Ensemble 2 (d–f).

Using the signed relative RMSE, defined as relative RMSE multiplied by the sign of the relative mean deviation (Section 3.5.1), allows us to assess not only the magnitude of error but also its direction. The results show that reconstructed percentiles are more often, and more severely, underestimated than overestimated, with negative RMSEs dominating across different cluster counts. Overestimation occurs less frequently and typically remains within the acceptable $+5$ % RMSE threshold. The main exception is the P_{90} of Ensemble 1, where MDS shows significant overestimation ($+10$–20 %) between 15 and 25 selected clusters, and KPCA shows a similar effect ($+10$ %) between 13 and 17 selected clusters.

This general downward asymmetry could be traced to the interaction between dimensionality reduction, k-means clustering, and the underlying performance distribution. In this study, low-performing models often cluster (especially early in the simulations) into relatively large, dense groups. Part of this effect may arise from system behavior under BHP control: the physics impose a practical lower bound on injection capacity, which compresses variation and limits how poorly a case can perform. In other words, when many realizations share ordinary petrophysical combinations, they naturally accumulate near this bound under fixed injection pressure. High-performing models, by contrast, are more heterogeneous: success can be achieved through several distinct combinations of favorable properties, each relatively rare, resulting in smaller, more isolated clusters.

The distribution of injection rates at day 1, day 10, and day 100 of Ensemble 1 (Figure 5.6) illustrates this contrast clearly: low-performing cases accumulate into large, dense groups, while high-performing cases are spread more sparsely across several smaller peaks. This pattern is consistent across both Ensemble 1 and Ensemble 2 (Appendix F, Figure F.4).

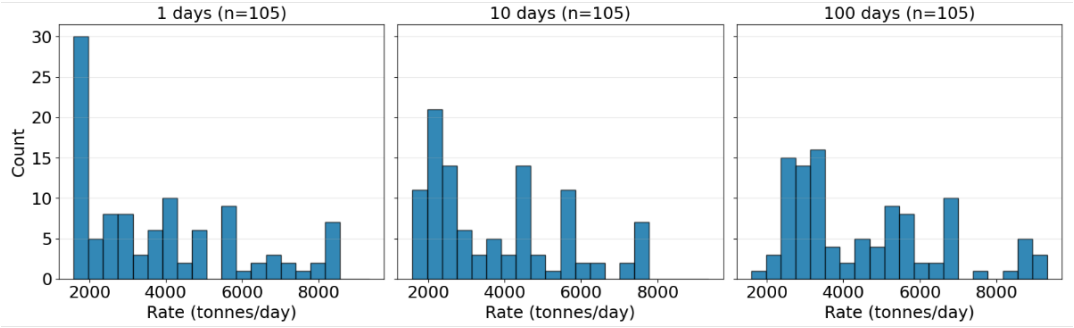


Figure 5.6: Distribution of Injection Rates at Selected Times (1, 10, and 100 days) - Ensemble 1; Full-Physics Results

Because one representative is chosen per cluster, the earliest selections are strongly shaped by the k-means++ initialization. This seeding strategy prioritizes well-separated regions rather than the densest parts of the distribution, increasing the chance of reducing final inertia and avoiding poor local minima. Consequently, at very low cluster counts (fewer than about 7), the compact block of lowest-performing runs is often skipped in favor of more diffuse, mid-performing groups. This induces a temporary bias: P_{90} is overestimated because the lowest performers have not yet been included, whereas P_{10} is underestimated because the highest-performing runs are diffuse and therefore rarely receive an early representative an effect visible in Figure 5.5 for all flow diagnostics. As additional clusters are introduced, the compact low-performing subgroups are eventually captured, pulling P_{90} downward into slight underestimation. The high-performing runs, however, remain scattered and underrepresented until much larger numbers of clusters are used, so P_{10} continues to be slightly underestimated in most cases.

Taken together, these effects could explain the downward asymmetry observed in the signed RMSE results. The P_{90} tends to be underestimated once the dense, low-performing clusters are included, while the P_{10} remains underestimated because the scattered, high-performing clusters are incorporated only at much higher cluster counts. The P_{50} is also affected: its reconstruction depends on which mid-performing clusters are selected, and in practice it too shows a general tendency toward underestimation. Occasional overestimations do occur (e.g., P_{90} with MDS and KPCA in Ensemble 1; Figure 5.5a), but underestimation is both more severe and more systematic. From a risk perspective, this tendency is not necessarily undesirable: conservative reconstructions provide a safeguard

against underestimating worst-case outcomes, which is preferable to being overly optimistic given the economic consequences of contractual defaults on CO_2 storage commitments with contractors.

Future work should examine whether this conservative bias in percentile reconstruction persists under different reservoir conditions, clustering strategies, and control settings, or if it is specific to the early-time, BHP-controlled systems analyzed here.

5.2.4 Role of System Compressibility

The effectiveness of early-time injection rates as flow diagnostics is likely dependent on the compressibility of the storage system. In stiffer systems with low compressibility, such as depleted reservoirs, pressure signals propagate quickly, and early-time injection responses already contain much of the information needed to characterize ensemble variability. By contrast, in more compressible systems such as saline aquifers with open boundaries, pressure diffusion is slower, and longer simulation times may be required before meaningful variability emerges in the injection response.

This suggests that the optimal time window for using injection rates as flow diagnostics is system-dependent: shorter simulations may suffice for depleted reservoirs, while longer time windows may be necessary in aquifers. From a practical perspective, this means that early-time injectivity screening should be calibrated to the compressibility of the target system, rather than assuming a universal diagnostic time frame.

In this study, half of the realizations of the ensembles apply a stringent fault transmissibility multiplier (0.01), which restricts cross-fault communication and limits connected storage at early times. These cases therefore exhibit a stiffer pressure–rate response, whereas the remaining realizations, with a less restrictive multiplier (0.9), are more compliant. This contrast likely explains why the 10-day injection rate already encodes system-scale information in our study (see Section 5.2.1): the presence of distinctly stiffer systems accentuates early differences that mirror the underlying dissimilarity structure of the injection rates over time. Consequently, for ensembles composed of more uniformly compliant reservoirs (i.e., with smaller differences in fault transmissibility), longer calibration periods may be required before injection-rate diagnostics capture system-wide effects, such as fault-controlled connectivity.

Finally, to assess robustness, the analysis was repeated with a reduced contrast in the fault transmissibility multiplier by replacing the 0.01 cases with 0.1 (leaving the others at 0.9). This narrowing decreased the ensemble spread in fault transmissibility but did not materially diminish the effectiveness of the 10-day rate diagnostic compared to the day-100 rate diagnostic, further supporting its robustness. Results are provided in Appendix F, Figure F.5.

5.3 Evaluating Streamline-Based Flow Diagnostics for CO_2 Injection Rates

Section 4.3.2 presented the results for the streamline with respect to the injection-rate storage metric. In the following section, some of these results will be examined in more detail, focusing on the mechanisms that may explain the observed behaviors and differences between the diagnostics.

5.3.1 Single-Phase vs Two-Phase Immiscible

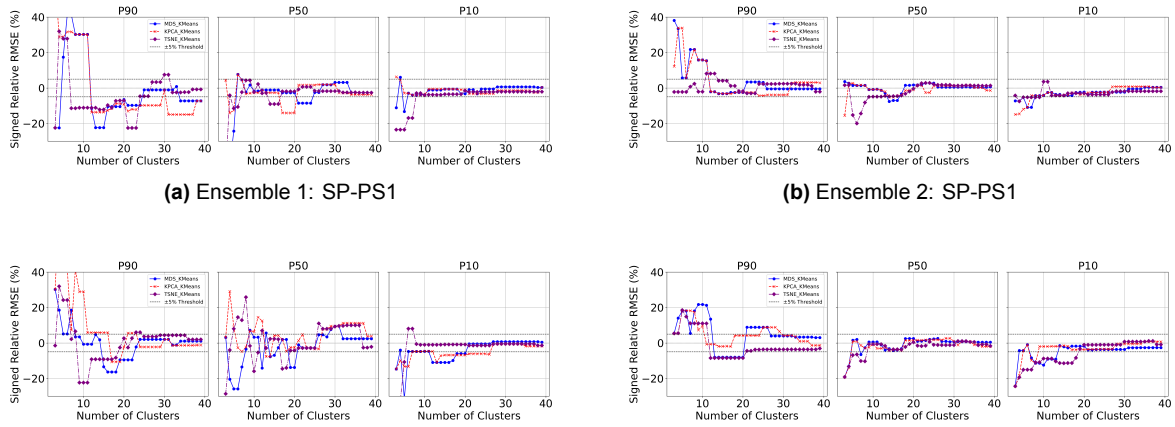
The observed advantage of the single-phase rate diagnostic over the immiscible rate diagnostics, evidenced by its notably more stable and accurate percentile reconstructions across varying cluster counts for Ensemble 1 (Section 4.3.2), raises the question of why a simplified model can outperform a more physically realistic one.

Although the immiscible formulation should, in principle, align more closely with the full-physics CO_2 simulations in open-DARTS by including multiphase effects (Section 2.4.1), ensemble variability appears to be dominated by well injectivity rather than early saturation dynamics. In the single-phase case, the injection response is directly linked to the pressure gradient and local permeability, effectively making it a proxy for well injectivity, defined as the ease with which fluid enters the reservoir. Under the simplifying assumptions of incompressible fluids with equal viscosities and linear relative permeabilities, total mobility remains constant, and the pressure field reflects only static heterogeneities, as it is solved once. As a result, differences in injection rate between runs arise primarily from variations in in-

jectivity. Given the strong clustering performance of this diagnostic, this finding indicates that injectivity is a principal driver of variability in injection-rate behavior, consistent with Section 5.2.1.

By contrast, the immiscible formulation introduces nonlinear, saturation-dependent effects from the first timestep. While these influence long-term behavior in full physics, they appear secondary for distinguishing variability across realizations. As a reduced-physics diagnostic, the two-phase formulation can overweight saturation-driven nonlinearities that contribute little to ensemble variability, shifting realization rankings and weakening alignment with the storage metric (see Section 5.2.1). Consequently, its induced feature space orders distances and neighborhoods less in line with injectivity-driven variability, reducing clustering performance and contributing to more erratic RMSE behavior across K. As K grows, partitions can become less coherent, and representatives may overfit small-scale differences, degrading percentile reconstructions. However, if injectivity dominates variability, the single-phase diagnostic isolates that driver and yields distances and neighborhoods that better match storage-relevant structure. As clustering resolves finer detail, partitions remain coherent and representatives explain more of the ensemble, so reconstructions fluctuate less and accuracy is comparatively stable.

This interpretation is supported by the signed-RMSE comparison between the single-phase diagnostic and the "immiscible–1-pressure-solve" variant: despite both using a single pressure solve, the single-phase diagnostic clearly outperforms the immiscible–1-pressure-solve case across both ensembles in terms of overall stability and accuracy (Figure 5.7).



(a) Ensemble 1: SP-PS1

(b) Ensemble 2: SP-PS1

(c) Ensemble 1: IMM-PS1

(d) Ensemble 2: IMM-PS1

Figure 5.7: Signed relative RMSE versus the number of selected clusters for the single-phase diagnostic (panels **a**, **b**) and the 1-pressure-solve immiscible diagnostic (panels **c**, **d**). Curves are shown for P_{10}, P_{50}, and P_{90} and three DR workflows: t-SNE (purple), KPCA (red), and MDS (blue). Left: Ensemble 1; right: Ensemble 2.

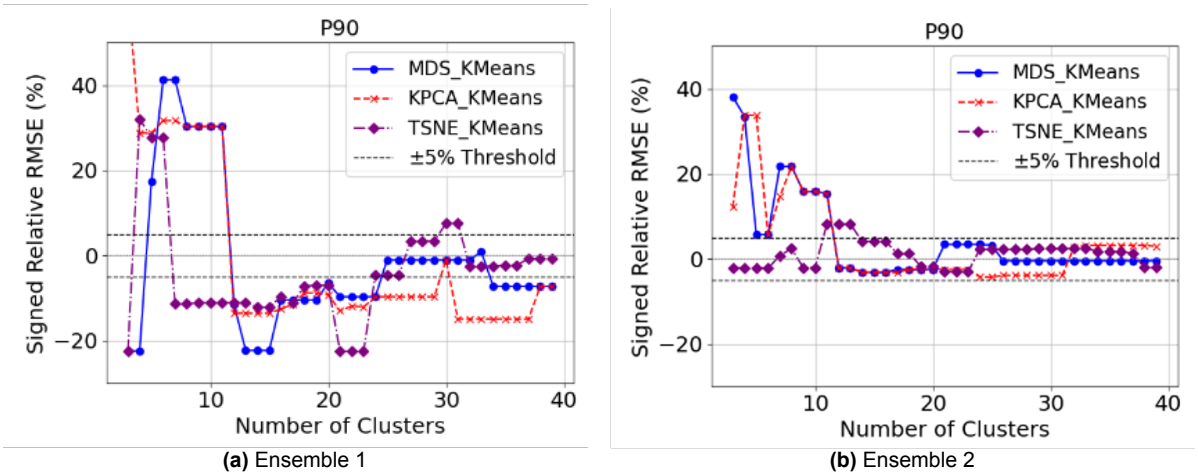
To quantify the alignment between the storage metric of interest and the streamline-based flow diagnostics, this study used Spearman's rank correlation (ρ) between diagnostic distance matrices and pairwise absolute differences in cumulative injected volume at 20 years, as explained in Section 5.2.1. For Ensemble 1, ρ was 0.73 (single-phase) versus 0.54, 0.53, and 0.55 for immiscible with 1, 4, and 11 pressure solves, respectively. For Ensemble 2, ρ was 0.53 (single-phase) versus 0.45, 0.45, and 0.54 (immiscible, 1/4/11 pressure solves). Thus, the single-phase diagnostic preserves storage-relevant ordering especially well in Ensemble 1 and is comparable to the 11-pressure-solve immiscible case in Ensemble 2. Notably, the first immiscible pressure solve yields a lower ρ than the single-phase case, consistent with early saturation effects introducing variation that does not mirror the full-physics ordering and can reduce clustering performance. Overall, these results suggest that, under the tested conditions, injectivity is the key control on ensemble variability in injection-rate behavior. The single-phase signal contains sufficient information to reconstruct the injection-rate distribution over time for at least P_{50} and P_{10} (with RMSE ≤ 5 % across most K), and it shows strong P_{90} performance for Ensemble 2.

That said, this study acknowledges that these conclusions intertwine diagnostic choice with the dimensionality-reduction technique. A sharper comparison includes k-medoids analysis to better isolate the correlation between flow diagnostics and cluster performance; see Section 5.5.1 for supporting results.

Finally, Section 4.3.2 noted a slight decrease in percentile-reconstruction accuracy over time for Ensemble 1 and an improvement for Ensemble 2. For context, first see the discussion of the P_{90} performance of the single-phase diagnostic in Ensemble 1; Section 5.3.3 then builds on this to propose explanations for these observations.

5.3.2 Single-Phase P_{90} Reconstruction in Ensemble 1

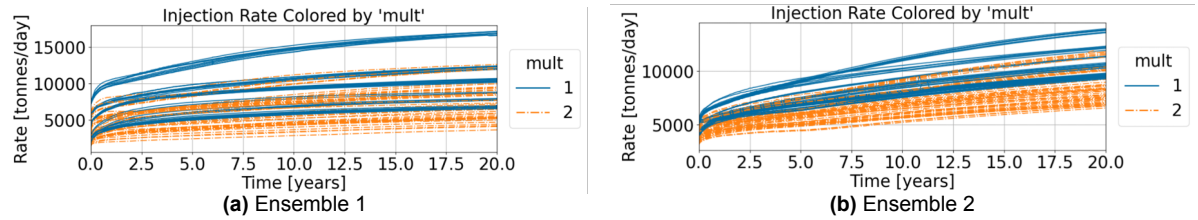
Although the P_{50} and P_{10} percentiles are reasonably well reconstructed by the single-phase diagnostic for both ensembles, as outlined in Section 4.3.2, the clustering performance for the P_{90} percentile is noticeably worse in Ensemble 1. This suggests that the single-phase streamline simulations may not be well suited to capturing the lower-end injection behavior in the geological settings of this ensemble. To illustrate this, Figures 5.8 show the signed relative RMSEs over varying cluster selections for Ensembles 1 and 2 P_{90} using the single-phase diagnostic. For Ensemble 1, all clustering workflows—particularly the t-SNE workflow at low cluster counts ($K < 10$)—generally result in stronger underestimation than overestimation of the true ensemble P_{90}. For Ensemble 2, the opposite tendency is observed, although the deviations are less severe and remain within acceptable limits (≤ 5 % RMSE) for most numbers of selected clusters, particularly for the t-SNE + KMeans workflow. Overall, this pattern suggests that the streamline-based approach may overweight pessimistic, low-rate cases when selecting representatives for Ensemble 1, while it may slightly underweight the lowest-rate cases in Ensemble 2.



(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.8: Signed relative RMSE for the reconstruction of the P_{90} injection-rate percentile using the single-phase streamline diagnostic for Ensembles 1 and 2. Results are shown as a function of the number of selected clusters for each dimensionality-reduction workflow (t–SNE, KPCA, and MDS).

A possible explanation for this overrepresentation of pessimistic, low-rate cases in Ensemble 1 could lie in the influence of fault transmissibility and fault geometry parameters included in the ensembles. Two transmissibility multipliers were applied: 0.9 and 0.01. It seems reasonable to assume that the higher multiplier (0.9) is associated with the high-injectivity cases, while the lower multiplier (0.01) acts as a strong flow barrier, producing restricted cases. This tendency is confirmed in Figures 5.9, which show the full-physics ensemble injection-rate results over time, with realizations highlighted by transmissibility multiplier for both Ensembles 1 and 2. It is clear that the lowest injection-rate realizations are associated with the lower multiplier (Mult 2, highlighted in orange).

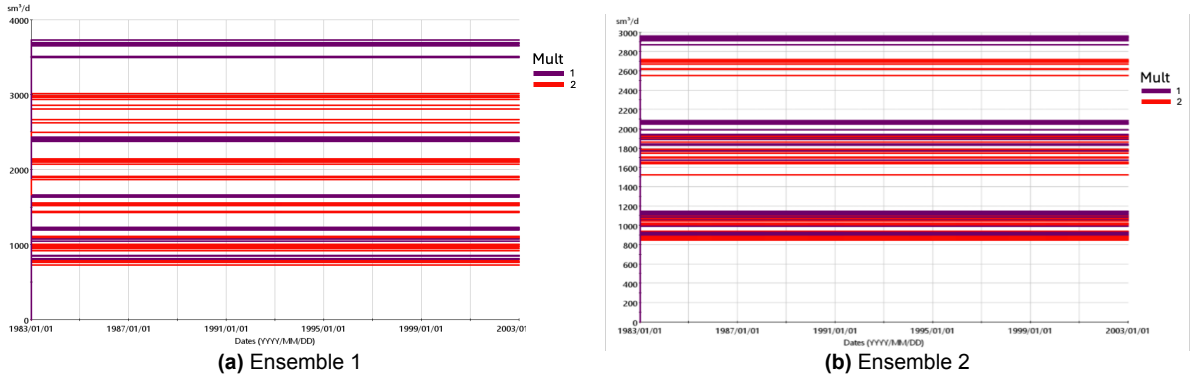


(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.9: Full-physics injection rates for Ensembles 1 and 2, colored by fault transmissibility multiplier.

Building on this observation, a tentative hypothesis is that the single-phase streamline diagnostic may underestimate the effect of the low-transmissibility multiplier, thereby misrepresenting the variability of restricted cases. In the diagnostic space (the injection-rate responses predicted by the flow diagnostic), realizations with transmissibility multipliers of 0.01 could appear more distinct from one another than they actually are, because the diagnostic signal is influenced more strongly by other parameters, such as the cut-off or the choice of facies-distribution model. If this is the case, clustering may then select several of these low-multiplier cases as "representatives," since they seem to span the variability within the diagnostic space. However, when evaluated with the full-physics simulator, their injection responses converge toward a narrower set of low-rate outcomes, leaving the clustered subset biased toward pessimistic behavior and leading to an underestimated P_{90}.

This mismatch in variability between the diagnostic and full-physics spaces (the injection-rate responses predicted by the full-physics simulator) is confirmed by examining the ensemble injection rates obtained with the single-phase flow diagnostic, shown in Figures 5.10(a) and 5.10(b). Comparing these results, where variability is highlighted by the fault-multiplier parameter, with the ensemble full-physics results in Figures 5.9(a) and 5.9(b), it becomes evident that the multiplier variability is represented differently in the diagnostic space than in the actual full-physics results. In particular, when focusing on the lower injection-rate region of the flow diagnostic (around 1000 sm^3/d), variation between Multipliers 1 and 2 is apparent, whereas this variation is clearly not visible in the lowest injection region of the full-physics results in either example. Moreover, the low-multiplier (Mult 2) cases tend to play a more dominant role in the higher-rate regions of the diagnostic space than is reflected in the full-physics results. Consequently, clustering in the diagnostic space can result in multiple low-multiplier cases being selected as representatives because they appear to span a wide range, even though they actually belong to the lower-bound region. This leads to a pessimistic bias in percentile reconstruction due to overrepresentation of restricted cases.

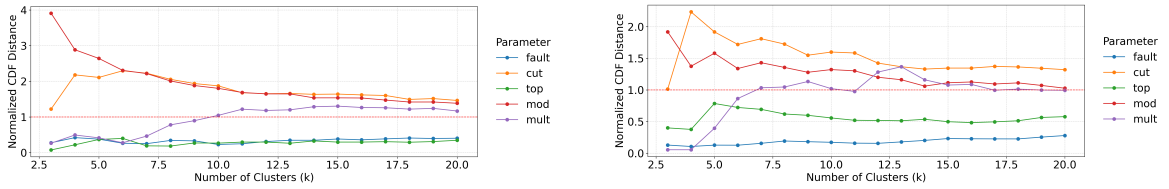


(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.10: Single-phase injection rates for Ensembles 1 and 2, colored by fault transmissibility multiplier.

That the variability in the diagnostic space does not reliably reflect the impact of the `mult` parameter on ensemble variability is further confirmed by examining the dGSA results over varying cluster counts for the single-phase flow diagnostic combined with t-SNE. As outlined in Section 4.5, η^2 showed that `mod`, `mult`, and `cut` were the key parameters driving the underlying ensemble variability in Ensemble 1, while in Ensemble 2 the injection-rate behavior was primarily controlled by `mult`.

Looking at the dGSA results of the single-phase diagnostic, it can be seen that only after selecting 10 clusters is the `mult` parameter flagged as influential in Ensemble 1, as it then surpasses the Normalized CDF threshold ($S > 1$), indicating that at lower cluster counts the parameter is assumed to have less impact on ensemble variability than it actually has according to its η^2 importance. For Ensemble 2, $S > 1$ is reached at six clusters; however, hierarchically the analysis incorrectly assigns greater influence to `cut` and `mod`, which are flagged as significantly influential ($S > 1$) from three clusters onward. This does not align with the η^2 results, which clearly identify `mult` as the key driver. Since dGSA evaluates how strongly parameter values separate across response clusters, this mismatch indicates that clusters in the diagnostic space fail to capture the expected separation for `mult`; its influence is underrepresented and the behavioural regimes are mischaracterised, leading to a less reliable representation of ensemble variability and, ultimately, a potentially pessimistic bias in P_{90} reconstruction, as explained previously.



(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.11: dGSA results over cluster counts for the single-phase flow diagnostic using t-SNE, showing the influence of key parameters (`mod`, `mult`, `cut`) in Ensembles 1 and 2.

That this effect, pessimistic P_{90} reconstruction, is more severe in Ensemble 1 than in Ensemble 2 may seem counterintuitive, especially since the dGSA results for Ensemble 2 showed the strongest misalignment with η^2, with a relatively late flagging of `mult` as an influential parameter while assuming `mod` and `cut` play a more significant role, which η^2 does not support. That said, the explanation may be more straightforward than expected. Overall, this study believes it can be stated with confidence that the impact of `mult` is not correctly reflected with the streamline diagnostic, which can lead to pessimistic P_{90} reconstructions due to the wrong selection of multiple low injection rate realizations that incorrectly span the variability range in the diagnostic space. The fact that this effect appears more amplified in Ensemble 1 could be due to its well placement relative to the fault network (Figure 5.12), where in Ensemble 1 (particularly fault model FM3), the injector is located in a structurally compartmentalized zone encircled by faults.

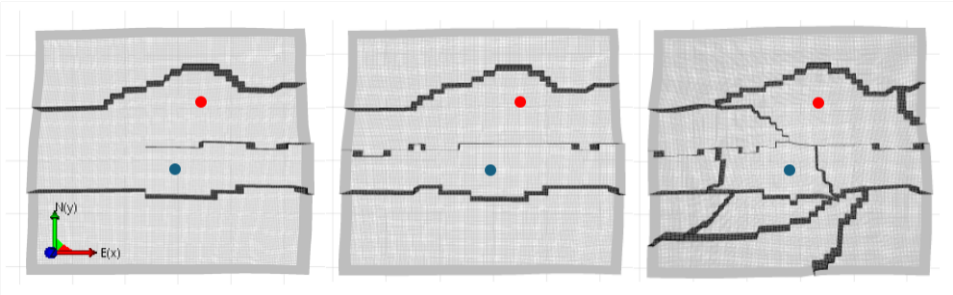
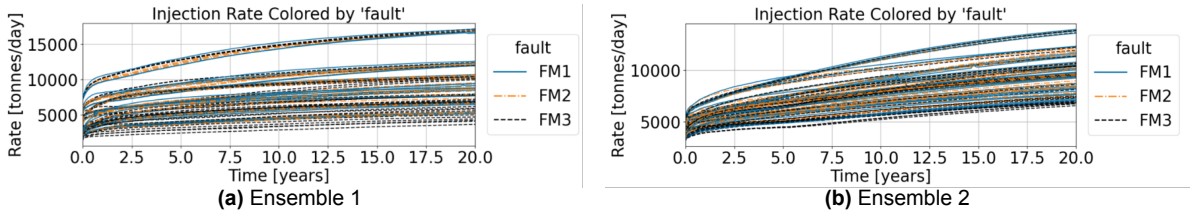


Figure 5.12: Top view of the three fault model scenarios (FM1–FM3). Blue: injector Ensemble 1. Red: injector Ensemble 2.

In such a setting (FM3), injectivity in the full-physics simulations is likely highly sensitive to the transmissibility multiplier: 0.9 allows partial connectivity, whereas 0.01 strongly restricts it. Therefore, the lowest-injectivity cases are likely to result from the combination of FM3 and Mult 2. This is confirmed by Figure 5.13, which shows the full-physics ensemble injection-rate results over time, with realizations highlighted by their corresponding fault model for both Ensembles 1 and 2. It can be seen that the lowest injection rates are associated with FM3 for both ensembles, while Figures 5.9 already showed their association with the lower multiplier (Mult 2).



(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.13: Full-physics injection rates for Ensembles 1 and 2, colored by fault model.

Simultaneously, it should be recognized that, upon inspecting the lowest-rate cases of Ensembles 1 and 2 in Figure 5.13, at least six of the lowest cases are clearly the result of the FM3 and Mult 2 combination in Ensemble 1. By contrast, in Ensemble 2 only the two lowest cases arise from this combination.

Moreover, the six lowest FM3 and Mult 2 cases in Ensemble 1 display significantly lower and more distinct performance compared with Ensemble 2, where the two lowest cases lie relatively close to a much denser lower-rate region composed of different combinations of fault models. This observation is most likely due to the FM3 and Mult 2 combination in Ensemble 2 producing less severe pessimistic cases, since in Ensemble 2 the injector in fault model 3 is located in a less compartmentalized system and is close to an open eastern boundary. Here, pressure dissipation could reduce the dominance of transmissibility, together with the fault-model setting, in controlling injectivity, bringing the injection rates of the realizations closer to one another regardless of the fault-model configuration. By contrast, the highly compartmentalized setting in Ensemble 1 results in a more severe impact of the FM3 and Mult 2 configuration compared with, for instance, the FM2 and Mult 2 or FM1 and Mult 2 configurations, producing distinctly more pessimistic injection-rate behavior.

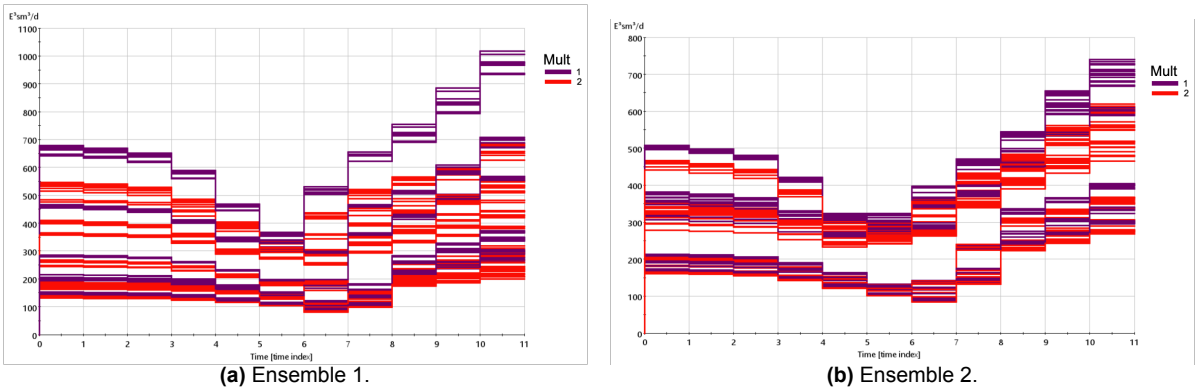
Consequently, if the streamline diagnostic smooths out these mult contrasts and underestimates their importance, thereby spanning a wider apparent variability in the diagnostic space, it may overselect these truly low-injection cases as representatives. This would have a much more severe effect for Ensemble 1 than for Ensemble 2, because it would pull the reconstructed P_{90} more strongly downward relative to the true ensemble P_{90} (Figures 4.1(a) and 4.1(b)).

In summary, the weaker P_{90} performance of the single-phase diagnostic in Ensemble 1 may be explained by a combination of (i) the diagnostic exaggerating the apparent diversity of low-transmissibility cases and (ii) the geological setting amplifying this effect due to stronger compartmentalization around the injector. Taken together, these factors could bias clustering toward over-weighting restrictive cases and thereby underestimating the P_{90}. While consistent with the observed results, this explanation remains a hypothesis that requires further testing to confirm. In addition, it should be acknowledged that the results also depend on the dimensionality-reduction technique applied, which makes it difficult to make claims that depend solely on the flow diagnostic itself. That said, in Ensembles 1 and 2 the different DR workflows behave overall quite similarly, with the most significant differences at lower cluster counts, where both MDS and KPCA show severe overestimation. This may be explained by a mechanism similar to the overpessimistic cases, since the argument cuts both ways: if the impact of multipliers is not appropriately reflected in the diagnostic space, clustering could also result in selecting apparent low cases as representatives that are actually higher-rate cases in the full-physics space.

5.3.3 Effect of Pressure Solve Count on the Two-Phase Immiscible Diagnostic

As outlined in Section 4.3.2, clustering performance for reconstructing ensemble percentiles appears to vary with the number of pressure solves for the two-phase immiscible rate diagnostic. For Ensemble 1, accuracy tends to decrease slightly as more pressure solves are considered, whereas for Ensemble 2 it improves. To understand why performance can improve with additional pressure solves, note that representing the effect of the fault transmissibility multiplier (`mult`) on ensemble variability appears to be challenging not only for the single-phase formulation, as outlined in the previous section, but also for the two-phase immiscible formulation.

Figures 5.14(a) and 5.14(b) show injection rates for Ensembles 1 and 2 across different numbers of pressure solves, coloured by `mult`. The distribution of `mult` in the diagnostic space is only loosely related to its distribution in the full-physics space (Figures 5.9(a) and 5.9(b)). In particular, within the lower-rate region of the flow diagnostic, variation between Multiplier 1 and 2 is apparent, whereas it is absent in the lowest-rate region of the full-physics results. Conversely, low-`mult` (Mult 2) cases appear to exert a stronger influence in the higher-rate part of the diagnostic space than they do in the full-physics space.

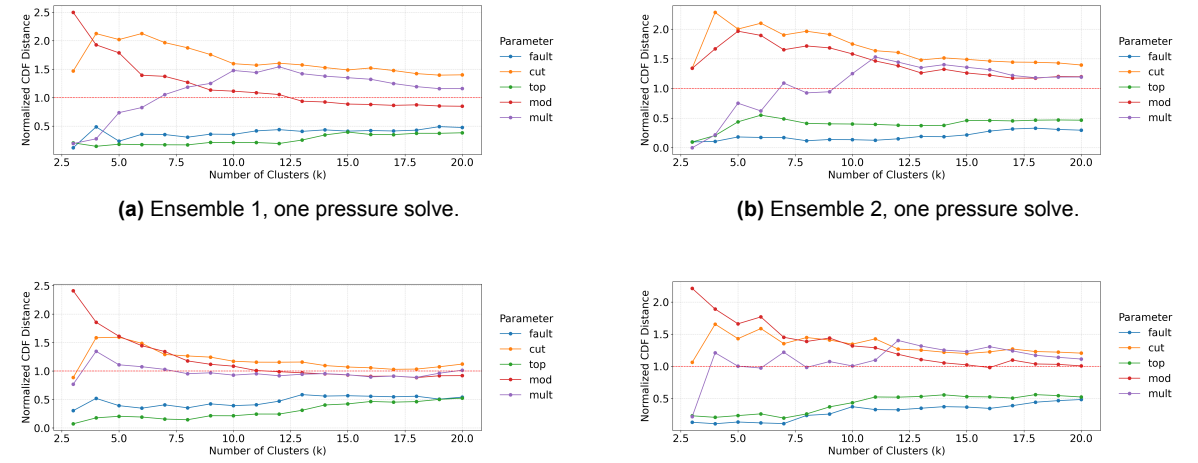


(a) Ensemble 1. **(b)** Ensemble 2.

Figure 5.14: Two-phase immiscible injection rates for Ensembles 1 and 2, coloured by fault transmissibility multiplier. Index 0 corresponds to day 1 (after the initial pressure solve); index 3 corresponds to day 25; index 10 corresponds to the eleventh pressure solve.

After the first pressure update (indices 0 to 1), variation among multiplier values is less distinct, and dense bands of realisations emerge, indicating that the mapping from `mult` to the diagnostic space is initially weak. With increasing pressure solves, particularly by the final one (indices 10 to 11, representing the eleventh pressure solve), `mult`-driven rate patterns become more pronounced. Although they still do not replicate the full-physics parameter-driven variation, the separability of `mult`-correlated realisations improves in the diagnostic space.

This trend aligns with the dGSA behaviour across cluster counts. For the immiscible formulation, `mult` shows relatively low influence after the first pressure solve (Figures 5.15(a) and 5.15(b)) and is flagged as influential only once at least six clusters are considered for both Ensembles 1 and 2. By the eleventh solve (Figures 5.15(c) and 5.15(d)), `mult` becomes more influential: the normalised CDF distance exceeds 1 once at least four clusters are considered.



(a) Ensemble 1, one pressure solve. **(b)** Ensemble 2, one pressure solve.

(c) Ensemble 1, eleven pressure solves. **(d)** Ensemble 2, eleven pressure solves.

Figure 5.15: dGSA over cluster counts for the immiscible diagnostic, showing the influence of `mult` after the first and eleventh pressure solves for Ensembles 1 and 2.

A plausible mechanism is that, as simulation time progresses, streamlines are retraced in response to advancing saturation fronts. The evolving flow field samples heterogeneity along paths increasingly aligned with fluid movement, which enhances sensitivity to transmissibility contrasts and therefore to `mult`. This can improve clustering performance for Ensemble 2 by aligning the diagnostic more closely with the full-physics response, consistent with the increasing Spearman rank correlations reported in Section 5.2.1.

The deterioration for Ensemble 1 is harder to reconcile. Note that, according to dGSA, the importance

of `mult` increases from the first to the eleventh pressure solve and, in terms of statistical significance, begins to align more closely with the true parameter impact η for both ensembles (Section 4.5); for the eleventh-solve diagnostic, significance is reached at earlier cluster counts than for the first-solve diagnostic.

A potential explanation is that the immiscible formulation introduces nonlinearities from the first time step through saturation-dependent properties. These nonlinearities can generate sensitivities that contribute little to the storage metric and thus weaken the correlation between the diagnostic and full-physics responses. If Ensemble 1, particularly the combination `FM3 Mult2` that yields strong compartmentalisation, occupies a geological setting that disfavours consistent alignment between full-physics and diagnostic responses, small conditioning or parameterisation mismatches may amplify as pressure solves accumulate, degrading percentile reconstruction at later times. By contrast, for Ensemble 2 the geological setting may be less unfavourable (for example, an open eastern boundary across all realisations), resulting in less error accumulation over time.

Regarding the hypothesis that poor cluster performance relates to a misalignment in how the diagnostic maps `mult`'s effects, the error pattern is consistent: in Ensemble 1, P90 and P50 tend to be underestimated (RMSE $\sim$20%), while P10 remains comparatively stable and accurate, suggesting that high-`mult` realisations are identified more reliably (Figure 4.9). This is intuitive: with relatively open boundaries (`mult` ≈ 0.9) there is little resistance to flow, leaving fewer mechanisms for divergence between the diagnostic and the full–physics simulator. In contrast, near-sealing barriers (`mult` ≈ 0.01) impose strong restrictions that can magnify differences in how the two approaches represent barrier impacts.

This interpretation should be viewed as a hypothesis. The qualitative patterns in Figures 5.14–5.15(d) support it, but conclusive evidence requires targeted experiments that (i) isolate `mult` from confounding parameters and (ii) verify robustness across alternative diagnostic features, clustering settings, and time horizons.

5.3.4 Generalizability of Streamline-Based Screening for CO_2 Injection Rates

As shown in the previous sections, in general the single-phase formulation performs better than the two-phase immiscible formulation. Performance for the immiscible formulation does improve with additional pressure solves, with both formulations performing approximately on par after 11 pressure solves for Ensemble 1 (Section 4.3.2), but the single-phase case remains superior when accounting for simulation cost (approximately 35 s vs. >6 min per realization). That said, both remain clearly less effective than using early-time injection rates simulated with the full-physics simulator. On the basis of these results, it is difficult to recommend these streamline-based formulations as effective screening tools for CO_2 subsurface injection-rate uncertainty quantification.

It seems clear that, according to the findings of this study, streamline-based simulators are not effective flow-rate screeners in heavily compartmentalized systems, such as Ensemble 1. By contrast, performance is more favorable in Ensemble 2, likely because all models place the well in a region with partial pressure dissipation due to a connection to an open boundary (see Figure 5.12). This indicates that streamline-based simulators may still serve as effective rate screeners in less fault-dominated systems with low fault-transmissibility values. Future studies should evaluate whether this holds under such conditions.

5.4 Evaluation of Flow Diagnostics for Plume Migration Uncertainty Quantification

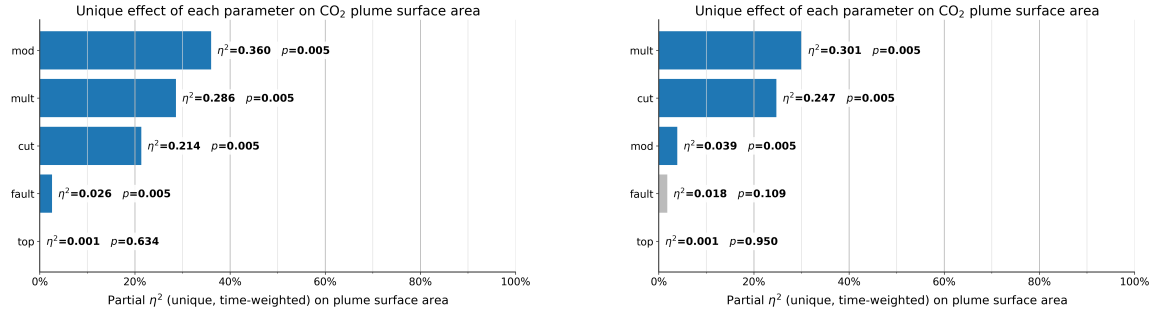
This section discusses several findings related to the uncertainty quantification of both the maximum plume extent and the plume areal coverage workflows.

5.4.1 Streamline-Based Saturation Field Flow Diagnostics

As outlined in Sections 4.7 and 4.8, the two-phase immiscible saturation-field diagnostic after 20 years was identified as the most appropriate metric for reconstructing the key percentiles P_{90}, P_{50}, and P_{10} based on distance-based clustering for both maximum plume extent and areal coverage uncertainty quantification. At the same time, the single-phase diagnostic was found to perform only slightly less

accurately, particularly at low cluster count. To explain this difference, the focus here is placed on the plume coverage metric, as it shows the closest correlation with the applied flow diagnostic while remaining relevant for assessing maximum plume migration.

Figure 5.16 presents the η^2 results for plume coverage in both ensembles, indicating the contribution of individual parameters to the ensemble variability of the full-physics plume coverage results, as described in Section 3.6.1. For Ensemble 1, the parameters `mod`, `mult`, and `cut` are most influential, whereas for Ensemble 2, `mult` and `cut` dominate. The corresponding η^2 results for plume migration are provided in Appendix G, Figure G.1. These results suggest a similar ranking of importance for Ensemble 1, while in Ensemble 2 the `cut` parameter emerges as the strongest driver, followed by `fault`, and then `mod` and `mult`.

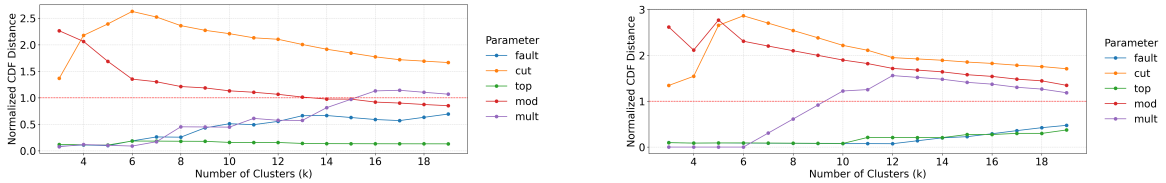


(a) Ensemble 1 coverage

(b) Ensemble 2 coverage

Figure 5.16: η^2 sensitivity indices for plume areal coverage obtained from the full-physics simulations for Ensembles 1 and 2. The indices quantify the contribution of each geological parameter to the variability of plume coverage across the ensemble over the complete simulation duration. For Ensemble 1, `mod`, `mult`, and `cut` are most influential, whereas for Ensemble 2 the variability is dominated by `mult` and `cut`.

To evaluate whether the two-phase immiscible saturation diagnostic could reproduce these parameter rankings, a dGSA analysis across varying cluster counts was performed using the t-SNE dimensionality-reduction workflow. The results are presented in Figure 5.17. The most notable discrepancy compared to the η^2 results is the absence of `mult` as a significant parameter. As shown in Section 5.3.3, increasing the number of pressure solves to eleven revealed `mult` as a significant parameter for cluster response separation at relatively low cluster counts ($K > 4$), indicating its importance for explaining full-physics ensemble variance, which aligned with the actual underlying η^2. Strikingly, the saturation field flow diagnostic, computed using the same two-phase immiscible flow simulation as the 11-pressure-solve rate diagnostic, does not identify `mult` as influential until 18 clusters are selected for Ensemble 1 and 11 clusters for Ensemble 2.



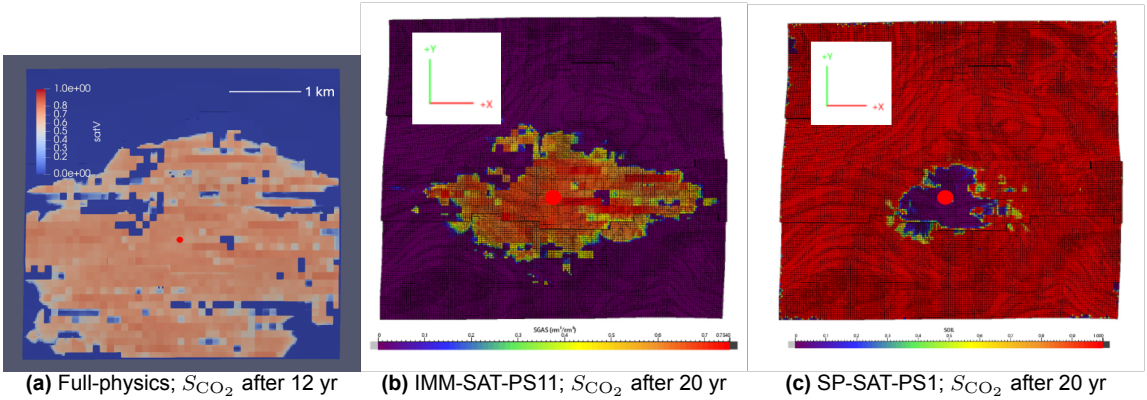
(a) t-SNE+KMeans, Ensemble 1 coverage

(b) t-SNE+KMeans, Ensemble 2 coverage

Figure 5.17: Distance-based generalized sensitivity analysis results for plume areal coverage in Ensembles 1 and 2, obtained using the IMM-SAT-PS11 diagnostic combined with the t–SNE+KMeans clustering workflow. Each plot shows how the η^2 sensitivity indices of the main geological parameters evolve as the number of selected clusters (K) increases.

The absence of a strong `mult` effect in the streamline-based diagnostics can be explained by examining the predicted saturation fields. Figure 4.4 shows realization 42 (configuration FM2-CO1-TS2-PIX-Mult1), which yields one of the largest plume coverage areas according to the full-physics simulations. The figure compares the saturation field after 12 years from the full-physics simulator with the satu-

ration fields after 20 years from the two-phase immiscible and single-phase streamline formulations. Strikingly, both streamline formulations predict plumes that migrate significantly less far after 20 years than the full-physics plume does after only 12 years. As a result, the plumes often do not reach the fault boundaries, or do so only in a limited manner, which most likely explains why the `mult` parameter exerts little apparent influence on plume migration in the streamline diagnostics. The parameter has not yet, or only barely, had the opportunity to counteract flow restriction and thereby limit plume coverage, meaning that its effect cannot yet be captured effectively when evaluating response clusters based on the field diagnostic.



(a) Full-physics; S_{CO_2} after 12 yr **(b)** IMM-SAT-PS11; S_{CO_2} after 20 yr **(c)** SP-SAT-PS1; S_{CO_2} after 20 yr

Figure 5.18: CO_2 saturation maps for realization 42 (FM2_CO1_TS2_PIX_Mult1). (a) Full-physics result after 12 years, compared to (b) two-phase immiscible and (c) single-phase streamline diagnostics after 20 years. Both streamline formulations severely underestimate plume migration. The two-phase immiscible formulation produces a somewhat larger plume than the single-phase due to gravity-driven buoyancy effects. The red dot marks the injector (Ensemble 1).

Due to time constraints, the underlying causes of this discrepancy could not be examined in further detail. Nevertheless, given the reasonably accurate clustering results presented in Sections 4.7 and 4.8, the diagnostics can still be considered to provide useful insights. In particular, local connectivity effects near the well appear sufficient to distinguish between high- and low-performing realizations in diagnostic space, leading to a relative ranking that aligns with the full-physics outcomes. In other words, although streamline simulators underestimate far-field plume migration, they are still able to capture near-wellbore connectivity signals that govern system-scale migration behavior. This is particularly evident for the P_{10} percentile, which across both ensembles shows the most accurate percentile reconstruction (Figures 4.31 and 4.32; Section 4.8), in contrast to the more significantly deviating reconstructions of the P_{90} percentile in both ensembles. This is favorable from an uncertainty-quantification perspective: for plume migration, P_{10} may be the most important percentile to estimate, given the risks associated with underestimating the severity of the worst-case scenario.

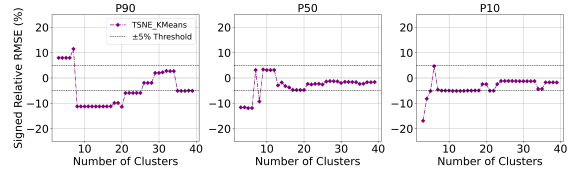
This discrepancy in percentile accuracy reconstruction could be linked to the same reasoning discussed in Section 5.3.3. With relatively open boundaries (`mult` ≈ 0.9), resistance to flow imposed by the fault transmissibility multiplier is limited, leaving fewer mechanisms for divergence between the diagnostic and the full-physics simulator. By contrast, near-sealing barriers (`mult` ≈ 0.01) impose strong restrictions that could magnify differences in how barrier effects are represented. Consequently, lower-bound cases (P_{90}) may be represented less accurately by the diagnostic, whereas upper-bound cases (P_{10}), for which system-scale (global) connectivity tends to dominate the response, potentially show better agreement and more consistent ranking between the two approaches. This would not only hold for the immiscible formulation, but also for the single-phase formulation, as this one will also most likely show the largest plume migration for the cases with the best global connectivity. At the same time, Figure 5.18 also illustrates why the immiscible saturation diagnostic proves to be the most effective: its plume shape more closely resembles the actual plume shape, as it includes two-phase flow with gravitational effects. This results in buoyancy-driven upward flow and a cone-shaped plume that spreads further at the top, as shown in the figure.

Therefore, given the importance of reliably quantifying plume-spread uncertainty for safe CO_2 stor-

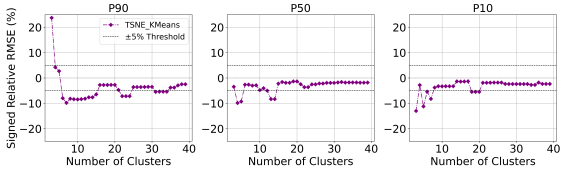
age, the two-phase immiscible saturation-field diagnostic appears to be the most suitable option for distance-based clustering, as it more faithfully represents the physics controlling far-field plume migration. Nevertheless, the single-phase diagnostic demonstrates that even a simplified, advection-dominated formulation can capture the near-wellbore connectivity patterns that appear to largely govern ensemble variability and drive the ranking of realizations. Future work should investigate how robust this connectivity-based signal remains under different geological settings and boundary conditions. Such studies could help define when the single-phase approach provides a sufficiently reliable and cost-effective surrogate, and when the additional physics of the immiscible formulation are required to maintain accuracy for plume migration uncertainty quantification.

5.4.2 Using Early-Time Full-Physics Injection Rates for Plume Migration Analysis

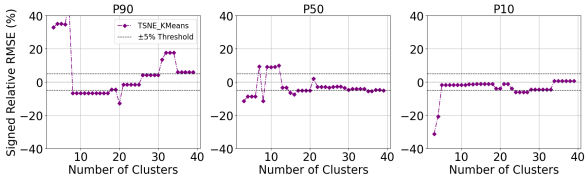
It is worth highlighting the surprisingly strong performance of the FP-D10 diagnostic from the full-physics simulations for both maximum plume extent and plume areal coverage (Sections 4.7 and 4.8). Although slightly less accurate than the saturation-field-based diagnostics, it still performs exceptionally well: even the worst percentile reconstructions remain almost consistently within a 10% RMSE bound, and both P_{50} and P_{10} are accurately reconstructed for both ensembles at low cluster counts ($K > 6$), as illustrated in Figure 5.19.



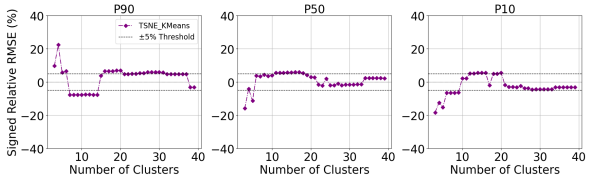
(a) Maximum plume migration, Ensemble 1.



(b) Maximum plume migration, Ensemble 2.



(c) Plume areal coverage, Ensemble 1.



(d) Plume areal coverage, Ensemble 2.

Figure 5.19: Signed relative RMSE for P_{90}, P_{50}, and P_{10} percentile reconstructions using the FP-D10 diagnostic from the Open-DARTS simulator. Top row: maximum plume migration. Bottom row: plume areal coverage. The FP-D10 diagnostic is combined with t–SNE clustering to assess whether the same clusters selected for injection-rate uncertainty quantification can also approximate plume migration and areal coverage.

This is a particularly interesting finding, as it suggests that the same set of realisations used to reconstruct injection-rate distributions can also approximate plume-migration behaviour with reasonable accuracy. Such a dual-purpose capability offers practical advantages: it reduces the need to run an additional subset of full-physics simulations for plume analysis and allows the entire process to be integrated within a single modelling platform, thereby reducing computational effort and workflow complexity.

A possible explanation for this favorable behavior is that the early-time injectivity response already encodes much of the information that controls plume migration. The ability of the reservoir to accept CO_2 during the first day of injection reflects the effective permeability, connectivity, and presence of barriers in the near-well region—factors that also govern the subsequent plume extent. In this sense, the day-10 injection rate acts as a global volumetric signal rather than a spatial diagnostic, yet still captures the reservoir properties most relevant to long-term plume migration.

That said, this study does not attempt to draw definitive conclusions in this regard. Instead, the results highlight a potentially valuable observation that warrants further investigation, particularly under a broader range of reservoir conditions and model configurations.

5.4.3 Spatial Limitation

Although RMSE accounts for percentile reconstruction accuracy over time, the reliability of this metric is somewhat limited in this context. Due to the finite spatial extent of the reservoir, most models gradually approach the maximum plume extent after about ten years. Once this upper limit is reached, the ensemble distribution collapses into a narrow range, artificially reducing variability. As a result, differences between models become less meaningful, and low RMSE values may reflect this spatial constraint rather than true clustering performance. This effect is particularly pronounced in Ensemble 1, but is also visible in Ensemble 2 for the plume migration metric, as shown in Section 4.1.2, while also being relevant for the plume areal coverage metric results (Section 4.1.3).

That said, when visually inspecting the reconstructed percentiles over time based on the clusters selected by the day-10 full-physics rate diagnostic (Figures 4.28, 4.30; Section 4.7), it can be observed that although the distribution tends to collapse over time, the percentiles still show overall acceptable alignment with the true percentiles at early times, when divergence is significantly more pronounced across the different realizations.

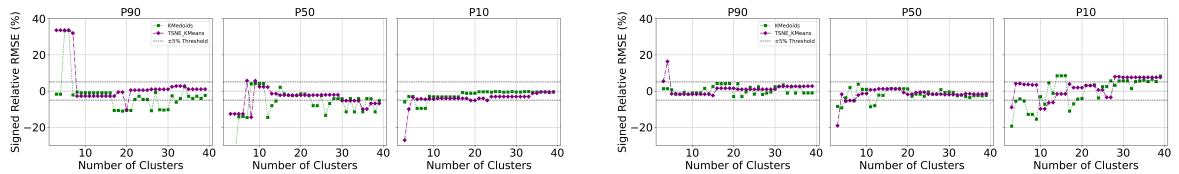
5.5 Comparing Most Effective Workflows with Direct K-medoids Clustering

This section compares the best-performing workflows for the injection-rate and plume areal-coverage storage metrics with an approach that applies K-medoids clustering directly to the distance matrices, without dimensionality reduction. Although the core focus of this study is on dimensionality reduction combined with K-means, primarily for computational efficiency across many cluster counts and diagnostics, K-medoids was included for comparison where time permitted. This is especially helpful for separating the effect of the flow diagnostics from that of the dimensionality-reduction techniques, making statements about their effectiveness more robust. It also enables assessment of whether the dimensionality-reduction step further improves cluster performance compared with applying K-medoids directly to the distance matrices.

The comparison will be shown for both the injection-rate diagnostic and the plume areal coverage obtained with the t-SNE workflow, which was found to be the most accurate and stable across varying cluster counts. A few observations are noteworthy.

5.5.1 Injection-Rate Uncertainty Quantification

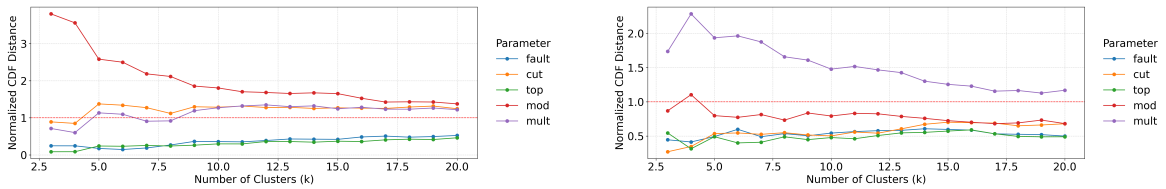
In this study, the FP-D10 diagnostic in combination with t-SNE dimensionality reduction produced the most stable and accurate percentile reconstructions overall. Figure 5.20 shows the signed relative RMSE values of this workflow for both ensembles; it also includes results obtained when omitting dimensionality reduction and applying K-medoids directly to the flow-diagnostic distance matrix. Inspecting the results over varying cluster counts, t-SNE is generally slightly more accurate and more stable across cluster counts. That said, K-medoids still performs reasonably well: most percentile reconstructions remain within the $\leq 5\ \%$ RMSE bound, with only a few exceeding $10\ \%$ RMSE, again highlighting the effectiveness of the day-10 flow diagnostic for ranking the realizations' long-term flow behavior. In addition, also K-medoids shows a tendency to underestimate the percentiles, which further supports the hypothesis that early injection-rate diagnostics tend to produce conservative reconstructions (Section 5.2.3).



(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.20: Signed relative RMSE over increasing cluster counts for the day-10 injection-rate diagnostic using K-medoids. Results are shown for Ensemble 1 and Ensemble 2.

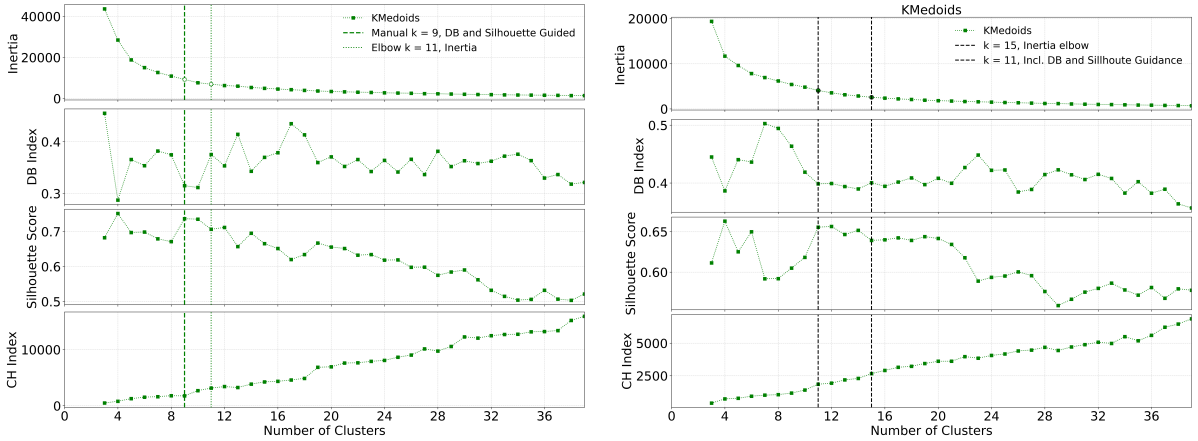
Second, inspecting the dGSA results for K-medoids shows that the day-10 diagnostic identifies the parameters driving long-term ensemble variability, typically for $K \gtrsim 5$, including `mult`, `mod`, and `cut` for Ensemble 1, while `mult` is flagged as key in Ensemble 2. This suggests that the diagnostic's ability to separate responses and reveal parameter importance is not strictly dependent on the choice of dimensionality reduction, but is inherent to the diagnostic itself and the dissimilarity structure it induces.



(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.21: Distance-based global sensitivity analysis (dGSA) over increasing cluster counts for the day-10 injection-rate diagnostic using K-medoids. Parameters with normalized CDF $S > 1$ are considered influential.

Third, while K-medoids tends to perform overall similarly to t-SNE in reconstruction accuracy, showing only slight degradation, its inertia curve tends to be smoother over varying cluster counts, resulting in later elbows, e.g., around $K \approx 11$ for Ensemble 1 and $K \approx 15$ for Ensemble 2 (Figure 5.22). The associated RMSEs (e.g., $\{0, 5, -5\}$ % for Ensemble 1 and $\{+3, 0, +8\}$ % for Ensemble 2) are reasonable, but t-SNE indicated earlier choices while remaining consistently within the ≤ 5 % RMSE bound for all percentiles (e.g., $K = 9$ and $K = 6$; see Section 3.4.3). That said, following DB and silhouette, one might still pick either $K = 9$ or $K = 10$ for Ensemble 1 with little loss in accuracy, and $K \approx 11$ for Ensemble 2 with only minor RMSE deterioration, again showing the effectiveness of guiding cluster-count selection by cross-referencing internal metrics. A likely part of the explanation for sharper inertia elbows with t-SNE is its tendency to exaggerate separation of local neighborhoods, which makes clusters appear more compact and well separated, thereby sharpening the inertia elbow.



(a) Ensemble 1 **(b)** Ensemble 2

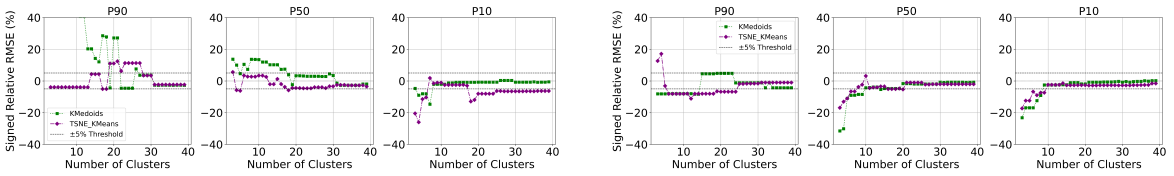
Figure 5.22: Internal clustering metrics (e.g., inertia, Davies–Bouldin, silhouette, Calinski–Harabasz) guiding cluster-count selection for the day-10 injection-rate diagnostic. K-medoids compared with t-SNE-guided choices.

For completeness, Appendix H.1 reports the signed relative RMSE per percentile over varying cluster counts for the single-phase and the first- and eleventh-pressure-solve immiscible rate diagnostics used for injection-rate uncertainty quantification; it also presents dGSA results for K-medoids clustering among the streamline-based rate diagnostics.

5.5.2 Plume Areal Coverage Uncertainty Quantification

Figure 5.23 shows the signed RMSE versus cluster count when comparing K-medoids and t-SNE for the plume-coverage metric. A notable result is the overall superior performance of t-SNE in Ensemble 1, whereas Ensemble 2 shows broadly similar performance for both methods. This difference in

performance likely stems from the distance definition used by the flow diagnostic. Because the metric is based on saturation-field differences aggregated over all cells (see Section 3.3), it can compress spatial structure: two realizations with different plume morphologies may appear similar once differences are summed and normalized. Direct partitioning with K-medoids operates on this compressed distance matrix and does not learn an intermediate representation. By contrast, non-linear embeddings such as t-SNE can recover neighborhood structure when the underlying geometry is non-linear, which can improve cluster separability and, in turn, reconstruction accuracy. This suggests that when distance metrics are expected to encode non-linear structure, methods like t-SNE are preferable.



(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.23: Signed RMSE over increasing cluster counts for the immiscible plume-coverage diagnostic using K-medoids.

Appendix H, Figure H.4 shows the dGSA results across varying cluster counts for K-medoids clustering using the streamline-based saturation-field diagnostic, again showing that `mult` is not flagged as an influential parameter driving response-cluster segregation as outlined in Section 5.4.1.

5.6 Selecting Minimum and Maximum Cases for Uncertainty Assessment

This section evaluates the practice of selecting minimum and maximum cases to represent uncertainty. It first examines whether parameter-based extreme selections capture the true response range and then tests whether early-time flow diagnostics, obtained with both 3DSL and open-DARTS, can better identify extreme cases.

5.6.1 Limitations of Parameter-Driven Extremes

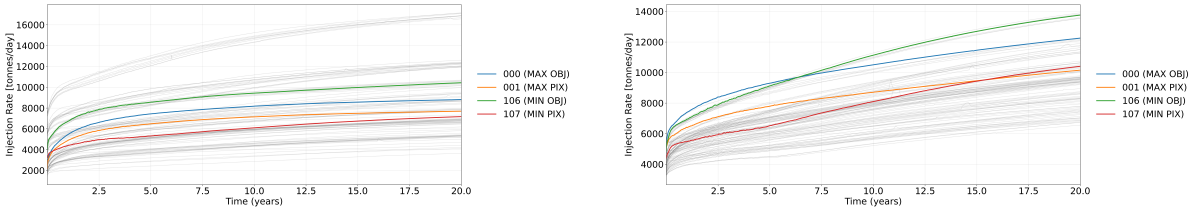
As outlined in the introduction of this thesis, one way to quantify uncertainty is the so-called rationalist approach, which can be regarded as a form of traditional determinism. In this method, a single base scenario is defined as the preferred model, and uncertainty is incorporated either by applying percentage factors to input parameters or model outputs, or by flanking the base case with separate low and high cases.

However, quantifying uncertainty using a single preferred geological scenario with only limited variations risks failing to capture the full range of flow responses, because the parameter space is only sparsely sampled. To illustrate this limitation, the injection-rate results were examined for the realisations that, based on their parameter configuration, would be expected to produce the lowest and highest injection rates according to this study.

Two “low” and two “high” cases were selected, one for the objective-based facies-distribution workflow (OBJ) and one for the pixel-based workflow (PIX). The low cases use the parameter combination `Cut=C03` (lowest net-to-gross), `FM=FM3` (most compartmentalised system), `Mult=2` (most restrictive fault-transmissibility multiplier), and `Top=T3` (lowest top surface). The high cases use `Cut=C01` (highest net-to-gross), `FM=FM1` (least compartmentalised), `Mult=1` (least restrictive multiplier), and `Top=T1` (highest top surface). Details of these parameters are provided in Section 2.2.2. Figure 5.24 shows the resulting injection-rate profiles: realisation 000 corresponds to the highest OBJ case (blue), 001 to the highest PIX case (orange), and 106 (green) and 107 (red) represent the lowest OBJ and PIX cases, respectively.

The results show that these selections do not align with the actual highest and lowest injection rates: even the case assumed to have the lowest OBJ performance ranks among the highest in Ensemble 2. This finding again suggests that local injectivity may exert a dominant influence on flow behaviour.

A geological realisation that appears unfavourable from its global parameter configuration may still achieve high injection rates if the well happens to intersect a highly permeable, well-connected region, such as a channel within an object-based facies model. Although this study cannot confirm that this specific mechanism caused the counterintuitive results, the evidence underscores the importance of uncertainty-quantification workflows that preserve ensemble variability, such as flow-based distance clustering, which has proved effective in this work, rather than relying solely on rankings based on parameter assumptions.



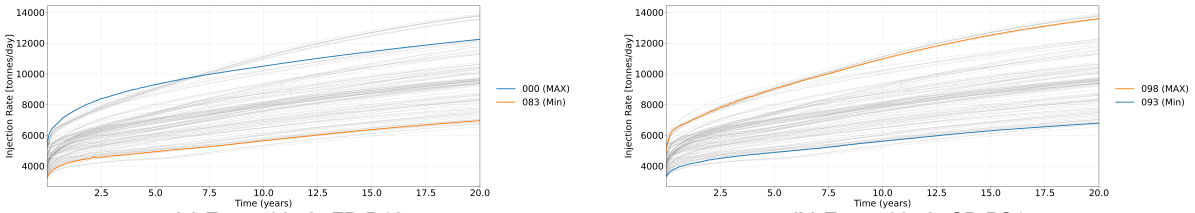
(a) Ensemble 1 **(b)** Ensemble 2

Figure 5.24: Injection-rate profiles for the realisations considered to represent the lowest and highest injection rates in each ensemble, based on their parameter configuration. Colours correspond to objective-based (OBJ) and pixel-based (PIX) selections for both high and low cases.

Appendix I.1 presents a similar analysis for plume-coverage results for the same four cases, again indicating that parameter-based assumptions risk misidentifying the true highest and lowest outcomes.

5.6.2 Early-Rate and Streamline Diagnostics as Screening Tools

Beyond assessing the impact of selecting the assumed highest and lowest injection-rate cases based on model-parameter configuration assumptions, this study also tested whether 3DSL and open-DARTS, using their respective SP-PS1 and FP-D10 diagnostic, can indicate which realizations tend to exhibit the highest injection rates. For Ensemble 1, both methods performed on par. For Ensemble 2, open-DARTS performed reasonably, although the 3DSL single-phase diagnostic more reliably captured the upper extreme, as shown in Figure 5.25. Although observed only for Ensemble 2, this suggests superior suitability of the single-phase diagnostic for predicting long-term extremes, likely due to its early sensitivity to global heterogeneity via streamline analysis as discussed in Sections 2.4 and 5.2.1.



(a) Ensemble 2: FP-D10 **(b)** Ensemble 2: SP-PS1

Figure 5.25: Comparison of screening diagnostics from 3DSL (single-phase) and open-DARTS (10-day early-rate) for identifying extreme injection-rate realizations for Ensemble 2

Appendix I, Figure I.2 shows that both the 10-day full-physics early-rate diagnostic and the single-phase diagnostic can recover the overall lowest and highest cases of Ensemble 1 with similar accuracies.

Unfortunately, it was not possible within this study to extract the smallest and largest plumes predicted by the streamline simulator and verify whether they align with the full-physics results. A practical recommendation is therefore to extend 3DSL with plume-centric diagnostics, for example automated reporting of plume footprint or area and maximum migration distance, under both the immiscible and single-phase formulations. Such functionality could enable rapid screening of candidate “min” and “max” plume realizations prior to full-physics runs. Given that plume extent during the injection period appears to be dominated by advective controls, these streamline-based indicators are likely to be informative for injection-stage extremes, but their skill post-injection may be lower as buoyancy, capillarity, and diffusive processes gain importance, and should therefore be validated against full-physics benchmarks at end-of-injection and at later monitoring times.

5.7 Generalizing the Most Effective UQ Workflow: Full-Physics Early-Rate Flow Diagnostics with t-SNE

Evidence from this study indicates that clustering on early-time injection–rate distances derived from full-physics simulations, combined with t-SNE for dimensionality reduction and subsequent k-means selection of cluster representatives, tends to yield accurate and stable percentile reconstructions across a broad range of cluster counts. When guided by internal clustering metrics (see Section 4.4), this workflow also exhibits clear potential for principled model-count selection. The remainder of this section contextualizes these findings and outlines avenues for future research that are required to establish, with confidence, the method's effectiveness for long-term injection-rate–behavior uncertainty quantification.

5.7.1 Full-Physics Early-Rate Flow Diagnostics

High Spearman rank coefficients (Sections 5.2.1 and 5.2.2) indicate that ranking realizations by early-time injection rates is strongly concordant with the corresponding long-term ranking. Complementarily, distance-based global sensitivity analysis (dGSA) applied to partitions obtained via t-SNE (Section 5.2.1) and k-medoids (Section 5.5.1) identifies parameter influences on cluster separation that are consistent with η^2-based sensitivities (Section 4.5).

Taken together, these findings support the conclusion that, for geomodel ensembles of the type studied here, characterized by categorical structural parameter variations (e.g., fault model structure, fault transmissibility multipliers, top-surface realizations) and limited variability in permeability distributions, early-rate flow diagnostics are effective for (i) identifying and ranking the parameters that drive ensemble variability and (ii) selecting representative realizations using internal clustering guidance. In this study, accurate P_{90}–P_{50}–P_{10} reconstructions were obtained from substantially reduced subsets: selecting 9 representatives for an ensemble of 105 realizations and 6 for an ensemble of 106 realizations corresponds to less than 10% of the full ensemble, while maintaining reconstruction fidelity.

Generalization beyond this setting warrants caution. The effectiveness of early-rate diagnostics may depend on system compressibility (Section 5.2.4) and on the extent to which variability is dominated by categorical rather than continuous heterogeneity; both factors can influence the required diagnostic time window and, consequently, overall effectiveness. A key contextual factor in the present ensembles is their effectively categorical structure: permeability fields vary through six discrete configurations (pixel- vs. object-based facies modeling and facies cutoffs controlling proportions and derived permeability). Moreover, among the five parameters considered, *mod*, *mult*, and *cut* were influential for injection-rate variability, whereas *fault* and *top* were not (Section 4.5). This concentrates the diversity of effective drivers from 108 nominal combinations to roughly a dozen impactful combinations. Consistent with Section 5.2.2, such concentration can produce banding, that is, compact groups that dominate rate variability, while non-influential factors contribute primarily noise rather than meaningful separation.

To delineate the domain of validity, it would be informative to evaluate performance in large ensembles where models differ solely in spatial permeability distributions. In such cases, the banding observed here may diminish, nonlinear flow effects over time may become more important for representing variability, and longer simulation horizons may be required to obtain rate diagnostics with comparable predictive power. In addition, applicability should be assessed in multi-well settings and across diverse reservoir environments.

5.7.2 t-SNE

This study suggests that the t-SNE dimensionality-reduction workflow delivered the most accurate clustering performance across multiple flow diagnostics and storage metrics, with especially strong results when combined with the early-rate flow diagnostic of the full-physics simulation for the rate uncertainty quantification. This study cautions, however, against assuming these findings generalize broadly.

A key contextual factor is the (effectively) categorical structure of this ensemble as stated previously. t-SNE is potentially well-suited to this situation because it strongly preserves *local* neighborhoods in the embedded space (Section 2.1), often forming compact, well-isolated clusters when the underlying structure is banded or quasi-categorical. The trade-off is that t-SNE pays less attention to *global* geometry.

If the ensemble were instead driven by a more continuous spectrum of variations (e.g., 108 realizations differing only in permeability distribution), this study would be less confident that t-SNE would maintain clean cluster boundaries. In such cases, methods that respect global structure, e.g., distance-based K-medoids on the original distance matrix, may yield more coherent partitions and more stable cluster performance. Therefore, to assess generalizability and robustness, future studies should replicate this workflow on ensembles with more continuous variability and benchmark t-SNE against global-structure baselines (e.g., distance-based K-medoids).

In addition, it should be noted that t-SNE requires a predefined perplexity. Perplexity controls the effective neighborhood size that each point trusts during embedding, with lower values emphasizing very local structure (risking over-fragmentation) and higher values smoothing over larger scales (risking band smearing and merged clusters). This study used a value of 30, as stated in Section 2.1, without exploring the effect of varying this number or tuning it for optimal results. Because the chosen value can influence the embedding and, by extension, cluster assignments and performance metrics, a targeted robustness check, sweeping perplexity (e.g., 10–40) and re-evaluating the same downstream metrics, should be considered in future work. The present findings should be interpreted with this caveat in mind.

5.7.3 Distance-Based Clustering on Flow Diagnostics

This study wants to emphasize the effectiveness of distance-based clustering on flow diagnostics as introduced by Scheidt and Caers [6, 7, 8]. As noted in the introduction, Scheidt and Caers argued that the method is largely application-agnostic, provided that the chosen flow diagnostics correlate with the metric of interest. Based on the results of this study, that claim can be confirmed with greater confidence: the approach not only proves effective for model ranking and selection in oil production context, but also appears to be a very effective workflow for uncertainty quantification in CO_2 subsurface storage. As shown in this study, it can significantly reduce geomodel ensembles arising from interpretational uncertainty while still capturing the ensemble flow distributions for both injection rates and plume migration. Therefore, this study wants to encourage future work to identify additional distance-based clustering workflows that could enhance uncertainty quantification, particularly for sustainable subsurface aims such as geological CO_2 storage and geothermal energy applications.

5.8 Using open-DARTS as an Industry Pre-Screener for CO_2 Subsurface Storage

Based on the findings of this study, open-DARTS can be recommended as an efficient pre-screening tool for CO_2 storage uncertainty quantification focused on long-term injection-rate behaviour. Its open-source availability, GPU support, and scriptable workflow make it well suited to large-ensemble studies where rapid turnaround is essential. In practice, open-DARTS can be used to quickly identify representative and extreme models, after which selected cases can be rerun with commercial full-physics simulators to refine the analysis where needed.

Computational performance

Table 5.1 summarises wall-clock times observed on a shared server. GPU runs were markedly faster than CPU runs and provided substantial speedups for short horizons.

Table 5.1: Observed wall-clock times for open-DARTS on a shared server. Times vary with hardware and load.

Horizon	GPU time	CPU time	Approx. speedup
1 day	∼0.5 min	∼15 min	∼30×
10 days	∼1.15 min	∼22 min	∼19×
100 days	∼3.7 min	∼40 min	∼11×

These results highlight the value of GPU-accelerated workflows for ensemble-based CCS studies. Combined with early-time flow diagnostics, GPU-enabled open-DARTS provides a fast and cost-effective way to explore model variability and to screen scenarios before committing to more computationally ex-

pensive simulations.

Note on 3DSL

This study also evaluated 3DSL. Open-DARTS produced rate diagnostics that yielded more stable and accurate percentile reconstructions across varying numbers of clusters and was more faithfully guided by internal cluster metrics (Section 4.4). In general, open-DARTS is therefore preferred for pre-screening. If GPU hardware is not available, however, CPU-only open-DARTS runs become significantly slower than 3DSL (for example, $\sim$30 s for a single-phase diagnostic in 3DSL versus $\sim$22 min for a day-10 diagnostic in open-DARTS). In such cases, 3DSL with a conservative selection of runs can be a practical alternative. This is particularly appealing if future studies confirm that 3DSL is an effective screener in less compartmentalised systems, as suggested in Section 5.3.4. A full cost versus accuracy comparison under dedicated hardware was beyond the scope of this study and should be examined more systematically in future work.

5.9 Remaining Limitations and Future Work

5.9.1 Neglecting post-injection stage

The present analysis is strongly focused on the injection stage. Post-injection behavior is critical for CCS applications, particularly for assessing plume migration and long-term storage security, but was not evaluated here. This is an important limitation and a priority for future work in terms of uncertainty quantification.

5.9.2 Injection Rates Full-Physics Results

The test cases analyzed in this study are largely advection dominated. With a bottomhole-pressure control set to a ΔP of 10 bar relative to the reservoir pressure (Section 3.2) along a $\sim$200 m well screen, simulated injection rates ranged from about 1 to 5 MTPA per well. These rates are considerably higher than typical single-well CCS deployments and may exaggerate pressure responses that could result in effective distance-based clustering based on early-rates already achieved after simulating 10 days. Future studies should therefore test the early-rate diagnostic under lower, more site-realistic rates.

Preliminary attempts to lower rates in this study led to unstable behavior, including instances where the reservoir pressure exceeded the imposed injection pressure, which is inconsistent with the intended injection-control setup. To avoid such artifacts, the 10 bar ΔP was retained, which produced stable injections without pressure reversal. For screening purposes, the emphasis is on ranking and representativeness rather than exact rate fidelity, so high absolute values can still be informative if applied consistently across realizations. Nevertheless, absolute rates should be interpreted with caution.

5.9.3 Interpretability vs. Performance in Clustering Techniques

While advanced dimensionality-reduction techniques such as t-SNE demonstrate strong performance in reconstructing ensemble behavior in this study, their interpretability remains limited. These methods often function as black boxes, projecting high-dimensional behavior into low-dimensional embeddings that lack a clear physical interpretation.

This limited transparency may be acceptable in purely technical contexts, but it becomes a concern when model selection informs high-stakes decisions. In CO_2 subsurface storage, for instance, stakeholders such as regulators, project developers, and financial institutions may require clear justifications for why particular realizations are selected for further simulation or monitoring. In such cases, more interpretable clustering approaches, such as direct clustering on physically meaningful features using K-medoids, may offer a better balance between performance and explainability. Results in Section 5.5 show promising performance for the early-time injection-rate diagnostic, with little loss of accuracy compared with t-SNE.

That said, dGSA can add valuable clarity regarding cluster behavior by identifying which input parameters dominate the variability captured by a given flow diagnostic within the feature space produced by dimensionality reduction, thereby improving the transparency of methods such as t-SNE.

Finally, as a general workflow recommendation, initial analyses should apply K-medoids directly to

distances derived from the chosen flow diagnostic. This isolates the effectiveness of the diagnostic itself without confounding effects from dimensionality reduction. If the data exhibit non-linear structure that challenges linear embeddings (see Section 5.5.2), exploratory visualization with MDS can help assess curvature; where warranted, non-linear dimensionality-reduction methods such as t-SNE or KPCA may then be introduced.

5.9.4 Parameter Sensitivity Analysis with dGSA

This study has demonstrated that distance-based generalized sensitivity analysis, when combined with the early-rate full-physics diagnostic (FP-D10), is highly effective for identifying parameter importance. Both t-SNE and k-medoids clustering produced sensitivity rankings that closely aligned with the time-weighted partial η^2 obtained from full-lifetime simulations, even under varying parameter importance settings.

Future studies could further validate the robustness of this approach by confirming the effectiveness of the FP-D10 diagnostic for long-term injection-rate screening across different geological settings and modelling scenarios. If this diagnostic proves reliable, dGSA offers a computationally efficient pathway to identify which parameters are most influential early in the workflow. For example, a modeller could parameterise channel width and run an ensemble using only the FP-D10 flow diagnostic, which in this study added just $1.3\ \%$ overhead compared to running the complete ensemble for the full simulation time to derive time-weighted partial η^2. Such an approach would enable early, low-cost screening of parameter importance and help prioritise the most impactful uncertainties for further investigation.

5.9.5 Early-Time Prediction of Critical Pressure Threshold Exceedance

Given that early-time injection rates in DARTS have proven effective in identifying distinct flow regimes, it is worth exploring whether early pressure responses can similarly be used to identify critical pressure risks. Since pressure signals tend to propagate faster than saturation or plume fronts, it may be possible to detect problematic cases (those with excessive pressure buildup) within the first few days of simulation.

A possible workflow could involve simulating the full ensemble for a short period (e.g., 10 days), identifying the models with the highest pressure increases, and then extending only these high-risk cases to full-lifetime simulations. If a critical pressure threshold is exceeded in one of these extended cases, the injection strategy could be iteratively adjusted (e.g., via reduced injection pressure) until the critical limit is no longer violated. Although this would require multiple runs of the complete ensemble to test this theory, it could significantly reduce overall simulation time while adding significant certainty in terms of derisking the critical pressure exceedance in the future if proven to be effective. It would provide a robust means of identifying worst-case scenarios. In this context, the goal is not to capture the full distribution of outcomes, but to detect rare but high-impact “black swan” events.

5.9.6 Clustering for Pressure Plume Migration Reconstruction

In contrast to threshold exceedance, pressure plume migration is a spatially distributed process where the entire response distribution is of interest. Here, early-time pressure profiles could still provide valuable information. By clustering the ensemble based on early pressure responses, it may be possible to reconstruct the full distribution of pressure migration with fewer representative models. This approach would mirror the workflow used for plume extent and injection rate clustering, potentially enabling significant reductions in computational cost without sacrificing the accuracy of pressure plume characterization.

Both directions, namely the identification of pressure threshold exceedance and the reconstruction of pressure migration, are relevant to geomechanical uncertainty quantification and represent promising directions for further investigation.

5.9.7 Injectivity Optimization

The results of this study suggest that near-well injectivity is a key driver of ensemble variability in injection behavior. This raises an important distinction between two sources of uncertainty: (i) global uncertainty, represented by variability across geomodels, and (ii) local uncertainty, arising from the sensitivity of injection performance to small shifts in well location within a single geomodel. While the

present study has focused on global uncertainty by keeping well locations fixed across models, local injectivity variations could introduce significant additional variability.

This observation highlights that local injectivity variability can be interpreted in two ways. On the one hand, it represents an additional source of uncertainty, since small changes in well placement within the same geomodel may lead to significantly different injection responses. On the other hand, because a specific well location must ultimately be chosen in practice, this variability can also be framed as an optimization problem: identifying well placements that maximize injectivity while minimizing variability. In this sense, local injectivity is both an uncertainty quantification challenge and a target for optimization.

If this interpretation holds, it opens up optimization opportunities. For instance, one could simulate a limited set of geomodels while systematically testing multiple well locations within each model, using early-time injection behavior as a rapid diagnostic. Such an approach could enable efficient screening of potential well placements, identifying those that combine high injectivity with reduced variability, thereby guiding field development decisions at relatively low computational cost.

5.10 Serving the Purpose of Sustainable Reservoir Engineering Applications

This chapter concludes with an important remark. The findings and methodologies presented in this study are designed to support sustainable subsurface engineering applications, such as CO_2 storage and geothermal energy production. The overarching goal is to advance efficient and cost-effective uncertainty quantification workflows that can reduce the computational burden of full-physics reservoir simulation. By lowering these barriers, the proposed approaches aim to facilitate the wider adoption of sustainable technologies, particularly in settings where conventional high-fidelity modeling is prohibitively expensive.

This work explicitly does not aim to advance hydrocarbon exploration or enhance oil and gas production strategies. Instead, the goal is to provide robust alternatives that make sustainability-focused modeling more accessible and cost-effective. In doing so, the study hopes to support the broader transition toward cleaner energy systems and responsible subsurface resource management.

6

Conclusion

This thesis evaluated distance-based clustering of flow diagnostics as a principled route to rank and select realizations from a 108-member geomodel ensemble spanning interpretational uncertainty in top surfaces (3), fault models (3), fault transmissibility multipliers (2), cutoffs (3), and facies modeling approaches (pixel- vs. object-based). The central objective was to achieve a substantial reduction in model count while preserving uncertainty-quantification fidelity, defined here as maintaining ensemble percentile bounds (P_{90}, P_{50}, P_{10}) for key CO_2 storage metrics (injection rate, maximum plume migration, plume areal coverage) within ≤ 5 % relative RMSE per percentile. The results demonstrate that flow-based distances provide a robust, computationally efficient basis for ensemble reduction without materially degrading UQ quality.

For injection-rate UQ, clustering performance was compared using rate diagnostics derived from streamline simulations and from full-physics simulations. Distances computed from full-physics injection rates after only 10 days (about 1.3 % of a 20-year simulation's runtime) exhibited high Spearman rank correlations with long-term injection-rate behavior (0.93 and 0.85 across two ensembles). When embedded with t-SNE and clustered, these distances yielded the most accurate and stable reconstructions of full-ensemble percentiles, with relative RMSE ≤ 5 % for P_{90}, P_{50}, and P_{10} across most tested cluster counts. In addition, a slight but consistent negative bias was observed, with reconstructed percentiles tending toward conservative underestimation. This bias likely arises because low-performing models form dense, easily captured clusters under early-time BHP control, while high-performing models are more scattered and therefore underrepresented when only a limited number of clusters is selected. From a risk-management perspective, such conservative reconstructions may be preferable, as they help avoid overly optimistic predictions of injection performance and storage capacity. Future studies should investigate whether this tendency holds under different reservoir and operational conditions.

Although reconstruction accuracy generally improves with increasing K, this study also highlighted occasionally erratic RMSE behavior at certain K values, indicating the inclusion of a cluster representative that can disrupt cumulative distribution function (CDF) construction. To regularize the choice of K, an internal guidance strategy combined an inertia elbow with feedback from the Davies–Bouldin index, Silhouette score, and Calinski–Harabasz index. Applied to the FP-D10 + t-SNE workflow, this approach consistently identified the number of clusters required for accurate selection while remaining computationally inexpensive. Relative to the full ensemble, the recommended selections reduced the model set by approximately an order of magnitude (Ensemble 1: 9 representatives; Ensemble 2: 6 representatives), corresponding to 9.7 % and 6.9 % of the relative simulation cost, while maintaining ≤ 5 % relative RMSE per percentile (Ensemble 1: $P_{90} = -3$ %, $P_{50} = +5$ %, $P_{10} = -4$ %; Ensemble 2: $P_{90} = -1$ %, $P_{50} = -5$ %, $P_{10} = +4$ %). In addition to this detailed investigation of the day-10 rate diagnostic, all tested flow-diagnostic and DR+clustering workflows were evaluated when guided solely by the inertia elbow. This analysis again showed superior performance for early-injection-rate FP diagnostics over streamline-based flow diagnostics, achieving the most accurate percentile reconstructions across both ensembles and workflows at low cluster counts when guided by inertia.

Because early-time FP injection rates yielded such accurate clustering and correlated strongly with long-term behavior, this study concludes that local injectivity is the dominant driver of injection-rate variability. Two observations reinforce this view: (i) even 1-day FP rate distances were sufficiently informative to support credible percentile reconstructions across varying K; and (ii) in streamline diagnostics, the single-phase formulation (isolating injectivity) often matched or exceeded the two-phase immiscible formulation in stability and accuracy, suggesting that multiphase effects introduce alignment challenges that can dilute diagnostic fidelity compared with rankings driven purely by injectivity, as in the single-phase simulator.

Beyond injectivity-driven variability, this study also posits that early-time FP rates can already signal system-scale effects (e.g., low fault transmissibility), encoding these effects into the rate distances and thereby enhancing cluster performance. This was supported by applying a distance-based generalized sensitivity analysis (dGSA) to the clusters produced by the workflow. Combining dGSA with the early-rate diagnostic indicated parameter influence consistent with variance-based sensitivity analysis conducted on the complete ensemble rate histories. For Ensemble 1, dGSA correctly identified the modeling facies-distribution method, cutoff values, and the fault-transmissibility multiplier as drivers of cluster separation. For Ensemble 2, it correctly indicated the fault-transmissibility multiplier as the main driver, supporting the finding that such system-scale effects are already signaled by the early rates. This successful dGSA application demonstrates potential to enhance UQ by highlighting parameters that warrant further attention to reduce uncertainty, and thus variability in ensemble behavior, while the simulation cost represents only $\sim 1.3\ \%$ of that required for running the full ensemble and performing variance-based sensitivity analysis. In addition, it provides a geologically grounded interpretation of cluster formation, thereby enhancing the transparency of cluster-based methods.

A frequency-informed genetic algorithm (GA) that selects subsets via short simulated calibration windows was also evaluated by matching reduced-subset percentiles to those of the full ensemble. While performance became robust for calibration periods $\gtrsim 100$ days, the time savings were limited relative to direct clustering. A two-stage workflow, with first reducing to $\sim 60\ \%$ with t-SNE + k-means, and then applying the GA, improved competitiveness but warrants caution, because in practice one would not know a priori whether such reductions would preserve the true ensemble percentiles used for calibration. More generally, GA's weaker feedback and longer calibration requirements make it less efficient when applied to the complete ensemble, reducing its appeal as a baseline compared with direct early-time distance clustering.

For both maximum plume migration and plume areal coverage, the immiscible saturation-field diagnostic (11 pressure solves over 20 years) produced the most accurate percentile reconstructions for distance-based clustering. The single-phase diagnostic, using only one pressure solve for the same 20-year period, showed only modest degradation while being far cheaper computationally (about $35\,\mathrm{s}$ versus $6\,\mathrm{min}$). Both diagnostics were particularly effective for reconstructing the P_{50} and P_{10} percentiles, with lower reliability for P_{90}. This is acceptable from a risk perspective, since rare extremes are more safety-critical and merit targeted analysis. A plausible explanation for the comparable performance is that local well connectivity strongly influences far-field plume spread, which both diagnostics capture effectively, even if the simulated plume area does not exactly match that of the full-physics saturation field.

In addition, using the cluster representatives selected for injection-rate UQ based on early-time rate diagnostics (FP-D10) resulted in only a small deterioration in performance when applied to plume-behavior UQ relative to the streamline-based saturation-field diagnostic. When cluster-count selection was guided by the internal cluster validation outlined previously (Ensemble 1: 9 clusters; Ensemble 2: 6 clusters), the maximum plume-extent percentiles were reconstructed with relative RMSE values of $P_{90} = -11\ \%$, $P_{50} = +4\ \%$, $P_{10} = -5\ \%$ for Ensemble 1 and $P_{90} = -8\ \%$, $P_{50} = -2\ \%$, $P_{10} = -5\ \%$ for Ensemble 2. For plume areal coverage, reconstructed percentiles of $P_{90} = -6\ \%$, $P_{50} = +9\ \%$, $P_{10} = -1\ \%$ were achieved for Ensemble 1, and $P_{90} = +6\ \%$, $P_{50} = +4\ \%$, $P_{10} = -6\ \%$ for Ensemble 2. This dual use highlights potential additional efficiency gains, because a single cluster selection can serve both injection-rate and plume metrics. A likely explanation is that the diagnostic carries a volume signal: higher early injection rates increase the chance of wider plume spread and vice versa, thereby reasonably reconstructing plume-migration behavior during the injection stage as well. Future work should test the robustness of this result under varying conditions.

Finally, percentile reconstruction based on selected cluster representatives should account for cluster population when constructing cumulative distribution functions. Weighting by cluster population produced superior performance in both accuracy and stability across varying cluster selections.

Overall, this study demonstrates that distance-based clustering can reduce ensembles driven by interpretational uncertainty by an order of magnitude while preserving ensemble-percentile reconstructions with relative RMSE values of $\leq 5\,\%$. Taken together, these results reinforce the claim by Scheidt and Caers that the approach is application-agnostic.

Future work should further explore the generalizability of early-time injection-rate diagnostics for long-term injection-rate uncertainty quantification under diverse geological and operational conditions. Given that early full-physics injection rates have proven effective in distinguishing distinct flow regimes, it would also be valuable to investigate whether early pressure responses could similarly help identify critical pressure risks. In particular, the ability to predict threshold exceedance at an early stage could provide a practical tool for de-risking injection strategies. Moreover, clustering based on early pressure responses may offer an efficient way to reconstruct pressure plume migration.

Finally, this study advocates the broader application of these methods in sustainable subsurface projects, highlighting their application-agnostic nature and their potential to reduce simulation costs while preserving uncertainty quantification accuracy.

References

[1] International Energy Agency [IEA]. *Carbon Capture, Utilisation and Storage*. https://www.iea.org/energy-system/carbon-capture-utilisation-and-storage. Accessed: 2025-07-20. 2024.

[2] Yuting Zhang, Christopher Jackson, and Samuel Krevor. "The feasibility of reaching gigatonne scale CO2 storage by mid-century". In: *Nature Communications* 15.1 (Aug. 2024). DOI: 10.1038/s41467-024-51226-8. URL: https://doi.org/10.1038/s41467-024-51226-8.

[3] Neeraj Nandurdikar and Luke Wallace. "Failure to produce: An investigation of deficiencies in production attainment". In: *Independent Project Analysis Newsletter* 4.2 (2012). URL: https://www.ipaglobal.com.

[4] IEAGHG. "Financial liability for CCS projects". In: *International Journal of Greenhouse Gas Control* 27 (2014), pp. 89–102. DOI: 10.1016/j.ijggc.2014.05.011.

[5] Mark Bentley and Ed Stephens. "Storage v. production: Challenges for reservoir modelling and simulation practitioners". In: *Geological Society, London, Energy Geoscience Conference Series* 1 (2025). Open Access under CC BY 4.0. DOI: 10.1144/egc1-2024-67. URL: https://doi.org/10.1144/egc1-2024-67.

[6] C. Scheidt and J. Caers. *A workflow for spatial uncertainty quantification using distances and kernels*. Tech. rep. 20. SCRF Report, 2007.

[7] Céline Scheidt and Jef Caers. "Representing Spatial Uncertainty Using Distances and Kernels". In: *Mathematical Geosciences* 41.4 (Sept. 2008), pp. 397–419. DOI: 10.1007/s11004-008-9186-0. URL: https://doi.org/10.1007/s11004-008-9186-0.

[8] Céline Scheidt and Jef Caers. "Uncertainty Quantification in Reservoir Performance Using Distances and Kernel Methods—Application to a West Africa Deepwater Turbidite Reservoir". In: *SPE Journal* 14.04 (Aug. 2009), pp. 680–692. DOI: 10.2118/118740-pa. URL: https://doi.org/10.2118/118740-pa.

[9] P. S. Ringrose and T. A. Meckel. "Maturing global CO2 storage resources on offshore continental margins to achieve 2DS emissions reductions". In: *Scientific Reports* 9.1 (Nov. 2019). DOI: 10.1038/s41598-019-54363-z. URL: https://doi.org/10.1038/s41598-019-54363-z.

[10] W. A. Ambrose et al. "Geologic factors controlling CO2 storage capacity and permanence: case studies based on experience with heterogeneity in oil and gas reservoirs applied to CO2 storage". In: *Environmental Geology* 54.8 (July 2007), pp. 1619–1633. DOI: 10.1007/s00254-007-0940-2. URL: https://doi.org/10.1007/s00254-007-0940-2.

[11] S. Krevor et al. "Capillary trapping for geologic carbon dioxide storage–From pore scale physics to field scale implications". In: *International Journal of Greenhouse Gas Control* 40 (2015), pp. 221–237. DOI: 10.1016/j.ijggc.2015.04.006.

[12] Philip Ringrose. *How to Store CO_2 Underground: Insights from Early-Mover CCS Projects*. Cham: Springer, 2020. DOI: 10.1007/978-3-030-43663-9.

[13] D. Arnold et al. "Hierarchical benchmark case study for history matching, uncertainty quantification and reservoir characterisation". In: *Computers & Geosciences* 50 (2013), pp. 4–15.

[14] D. Zhang. "Quantification of uncertainty for fluid flow in heterogeneous petroleum reservoirs". In: *Physica D: Nonlinear Phenomena* 133.1-4 (1999), pp. 488–497.

[15] M. G. Shirangi and L. J. Durlofsky. "A general method to select representative models for decision making and optimization under uncertainty". In: *Computers & Geosciences* 96 (2016), pp. 109–123.

[16] R. S. Middleton and S. Yaw. "The cost of getting CCS wrong: Uncertainty, infrastructure design, and stranded CO_2". In: *International Journal of Greenhouse Gas Control* 70 (2018), pp. 1–11.

[17] Philip Ringrose and Mark Bentley. *Reservoir Model design*. Springer, Jan. 2021. DOI: 10.1007/978-3-030-70163-5. URL: https://doi.org/10.1007/978-3-030-70163-5.

[18] Mark Bentley and Simon Smith. “Scenario-based reservoir modelling: the need for more determinism and less anchoring”. In: *Geological Society London Special Publications* 309.1 (Jan. 2008), pp. 145–159. DOI: 10.1144/sp309.11. URL: https://doi.org/10.1144/sp309.11.

[19] E. Fetel and G. Caumon. “Reservoir flow uncertainty assessment using response surface constrained by secondary information”. In: *Journal of Petroleum Science and Engineering* 60.3-4 (2008), pp. 170–182.

[20] M. Bentley and S. Smith. “Scenario-based reservoir modelling: the need for more determinism and less anchoring”. In: *Geological Society, London, Special Publications* 309.1 (2008), pp. 145–159. DOI: 10.1144/SP309.10.

[21] F. Watson, S. Krogstad, and K. A. Lie. “The use of flow diagnostics to rank model ensembles”. In: *Computational Geosciences* 26.4 (2022), pp. 803–822.

[22] S. Oladyshkin et al. “An integrative approach to robust design and probabilistic risk assessment for CO_2 storage in geological formations”. In: *Computational Geosciences* 15 (2011), pp. 565–577. DOI: 10.1007/s10596-011-9246-5.

[23] Hailun Ni et al. “The impact of capillary heterogeneity on CO_2 flow and trapping across scales”. In: *Earth-Science Reviews* 270 (2025), p. 105257. DOI: 10.1016/j.earscirev.2025.105257. URL: https://doi.org/10.1016/j.earscirev.2025.105257.

[24] S. J. Jackson et al. “A spatial analysis methodology for the carbon capture clusters definition and carbon utilization and storage hubs”. In: *International Journal of Greenhouse Gas Control* 118 (2022), p. 103688. DOI: 10.1016/j.ijggc.2022.103688.

[25] Fei Lu et al. “Limitations of polynomial chaos expansions in the Bayesian solution of inverse problems”. In: *Journal of Computational Physics* 282 (Nov. 2014), pp. 138–147. DOI: 10.1016/j.jcp.2014.11.010. URL: https://doi.org/10.1016/j.jcp.2014.11.010.

[26] Parviz Bahrami, Farzaneh Sahari Moghaddam, and L. Andrew James. “A Review of Proxy Modeling Highlighting Applications for Reservoir Engineering”. In: *Energies* 15.14 (2022), p. 5247. DOI: 10.3390/en15145247. URL: https://doi.org/10.3390/en15145247.

[27] M. Ani et al. “Ranking of geostatistical models and uncertainty quantification using Signal Detection Principle (SDP)”. In: *Journal of Petroleum Science and Engineering* 174 (2019), pp. 833–843.

[28] S. Li, C. V. Deutsch, and J. Si. “Ranking geostatistical reservoir models with modified connected hydrocarbon volume”. In: *Ninth International Geostatistics Congress*. June 2012, pp. 11–15.

[29] Hong Tang and Ning Liu. “Reservoir Static Connectivity and Heterogeneity Analysis (RCHA) and the Impact on Flow Behavior”. In: *International Petroleum Technology Conference* (Dec. 2008). DOI: 10.2523/iptc-12877-ms. URL: https://doi.org/10.2523/iptc-12877-ms.

[30] *Metric Space Methods | Streamsim Connect*. https://www.streamsim.com/technology/uncertainty-quantification/metric-space-methods-key-concepts. Accessed: 2025-07-20. n.d.

[31] Larry Lake and Jerry Jensen. “A Review of Heterogeneity Measures Used in Reservoir Characterization”. In: *In Situ* 15 (Jan. 1991), pp. 409–439.

[32] R. P. Batycky, M. J. Blunt, and M. R. Thiele. “A 3D Field-Scale Streamline-Based reservoir simulator”. In: *SPE Reservoir Engineering* 12.04 (Nov. 1997), pp. 246–254. DOI: 10.2118/36726-pa. URL: https://doi.org/10.2118/36726-pa.

[33] D. Arnold et al. “Hierarchical benchmark case study for history matching, uncertainty quantification and reservoir characterisation”. In: *Computers Geosciences* 50 (Sept. 2012), pp. 4–15. DOI: 10.1016/j.cageo.2012.09.011. URL: https://doi.org/10.1016/j.cageo.2012.09.011.

[34] Daniel O. Schulte et al. “Multi-objective optimization under uncertainty of geothermal reservoirs using experimental design-based proxy models”. In: *Geothermics* 86 (Jan. 2020), p. 101792. DOI: 10.1016/j.geothermics.2019.101792. URL: https://doi.org/10.1016/j.geothermics.2019.101792.

[35] Safiya Alpheus and Elizabeth Hajek. “The Fate of Bars in Braided Rivers”. In: *The Sedimentary Record* 22.1 (June 2024). DOI: 10.2110/001c.117787. URL: https://thesedimentaryrecord.scholasticahq.com/article/117787-the-fate-of-bars-in-braided-rivers.

[36] T. Manzocchi et al. “Fault transmissibility multipliers for flow simulation models”. In: *Petroleum Geoscience* 5.1 (Feb. 1999), pp. 53–63. DOI: 10.1144/petgeo.5.1.53. URL: https://doi.org/10.1144/petgeo.5.1.53.

[37] S.Y. Zheng, V. M. Legrand, and P.W.M. Corbett. “GEOLOGICAL MODEL EVALUATION THROUGH WELL TEST SIMULATION: A CASE STUDY FROM THE WYTCH FARM OILFIELD, SOUTH-

ERN ENGLAND". In: *Journal of Petroleum Geology* 30.1 (Jan. 2007), pp. 41–58. DOI: 10.1111/j.1747-5457.2007.00041.x. URL: https://doi.org/10.1111/j.1747-5457.2007.00041.x.

[38] Denis Voskov et al. "open Delft Advanced Research Terra Simulator (open-DARTS)". In: *The Journal of Open Source Software* 9.99 (July 2024), p. 6737. DOI: 10.21105/joss.06737. URL: https://doi.org/10.21105/joss.06737.

[39] I. Saifullin et al. "DARTS open-source reservoir simulation framework". In: *European Conference on the Mathematics of Geological Reservoirs (ECMOR 2024)*. Vol. 2. EAGE, 2024, pp. 942–960. DOI: 10.3997/2214-4609.202437086. URL: https://pure.tudelft.nl/ws/portalfiles/portal/242933970/86.pdf.

[40] Ding-Yu Peng and Donald B. Robinson. "A New Two-Constant Equation of State". In: *Industrial Engineering Chemistry Fundamentals* 15.1 (Feb. 1976), pp. 59–64. DOI: 10.1021/i160057a011. URL: https://doi.org/10.1021/i160057a011.

[41] M. D. Jager, A. L. Ballard, and E. D. Sloan. "The next generation of hydrate prediction: II. Dedicated aqueous phase fugacity model for hydrate prediction". In: *Fluid Phase Equilibria* 211.1 (2003), pp. 85–107. ISSN: 0378-3812. DOI: 10.1016/S0378-3812(03)00155-9. URL: https://www.sciencedirect.com/science/article/pii/S0378381203001559.

[42] Zaman Ziabakhsh-Ganji and Henk Kooi. "An Equation of State for thermodynamic equilibrium of gas mixtures and brines to allow simulation of the effects of impurities in subsurface CO2 storage". In: *International journal of greenhouse gas control* 11 (Aug. 2012), S21–S34. DOI: 10.1016/j.ijggc.2012.07.025. URL: https://doi.org/10.1016/j.ijggc.2012.07.025.

[43] Inc. Streamsim Technologies. *3DSL® Simulator*. https://www.streamsim.com/technology/3dslr-simulator. 2025.

[44] M. R. Thiele, R. P. Batycky, and D. H. Fenwick. "Streamline Simulation for Modern Reservoir-Engineering Workflows". In: *Journal of Petroleum Technology* 62.01 (2010), pp. 64–70. DOI: 10.2118/118608-JPT.

[45] Klaus Stüben, Patrick Delaney, and Serguei Chmakov. "Algebraic Multigrid (AMG) for Ground Water Flow and Oil Reservoir Simulation". In: *MODFLOW and MORE: Integrated Ground□Water Modeling*. Colorado School of Mines. Golden, CO, Jan. 2003.

[46] S. Taku Ide, S. Julio Friedmann, and Howard J. Herzog. "CO_2 Leakage through Existing Wells: Current Technology and Regulations". In: *Proceedings of the 8th International Conference on Greenhouse Gas Control Technologies (GHGT-8)*. Poster Session II, Trondheim, Norway. 2006. URL: https://sequestration.mit.edu/pdf/GHGT8_Ide.pdf.

[47] David Segura et al. "Physicochemical behavior and impact of CO_2 and CH_4 plumes during gas□rich water leakage in a shallow carbonate freshwater aquifer". In: *Applied Geochemistry* 149 (2024), p. 106122. DOI: 10.1016/j.apgeochem.2024.106122.

[48] Mark G. Little and Robert B. Jackson. "Potential impacts of leakage from deep CO_2 geosequestration on overlying freshwater aquifers". In: *Environmental Science & Technology* 44.23 (2010), pp. 9225–9232. DOI: 10.1021/es102235w.

[49] Florian Wickelmaier. *An Introduction to MDS*. Technical Report. Sound Quality Research Unit, Aalborg University, May 2003.

[50] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. *Nonlinear Component Analysis as a Kernel Eigenvalue Problem*. Tech. rep. Technical Report No. 44. Max Planck Institute for Biological Cybernetics, Dec. 1998. URL: https://www.face-rec.org/algorithms/Kernel/kernelPCA_scholkopf.pdf.

[51] Laurens Van Der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (Jan. 2008), pp. 2579–2605. URL: http://isplab.tudelft.nl/sites/default/files/vandermaaten08a.pdf.

[52] Ryan P. Adams. *K-means clustering and related algorithms*. Lecture notes, Princeton University. n.d. URL: https://www.cs.princeton.edu/courses/archive/fall18/cos324/files/kmeans.pdf.

[53] David Arthur and Sergei Vassilvitskii. "k-means++: The Advantages of Careful Seeding". In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[54] Darryl Fenwick, Céline Scheidt, and Jef Caers. "Quantifying Asymmetric Parameter Interactions in Sensitivity Analysis: Application to Reservoir Modeling". In: *Mathematical Geosciences* 46.4 (2014), pp. 493–511. DOI: 10.1007/s11004-014-9530-5.

[55] John T. E. Richardson. “Eta squared and partial eta squared as measures of effect size in educational research”. In: *Educational Research Review* 6.2 (2011), pp. 135–147. DOI: 10.1016/j.edurev.2010.12.001. URL: https://doi.org/10.1016/j.edurev.2010.12.001.

[56] Jonathan Frossard and Olivier Renaud. “Permutation Tests for Regression, ANOVA, and Comparison of Signals: The permuco Package”. In: *Journal of Statistical Software* 99.15 (2021), pp. 1–32. DOI: 10.18637/jss.v099.i15. URL: https://doi.org/10.18637/jss.v099.i15.

[57] Charles Spearman. “The proof and measurement of association between two things”. In: *American Journal of Psychology* 15.1 (1904), pp. 72–101. DOI: 10.2307/1412159.

A

Additional Details on the Geomodel Ensemble

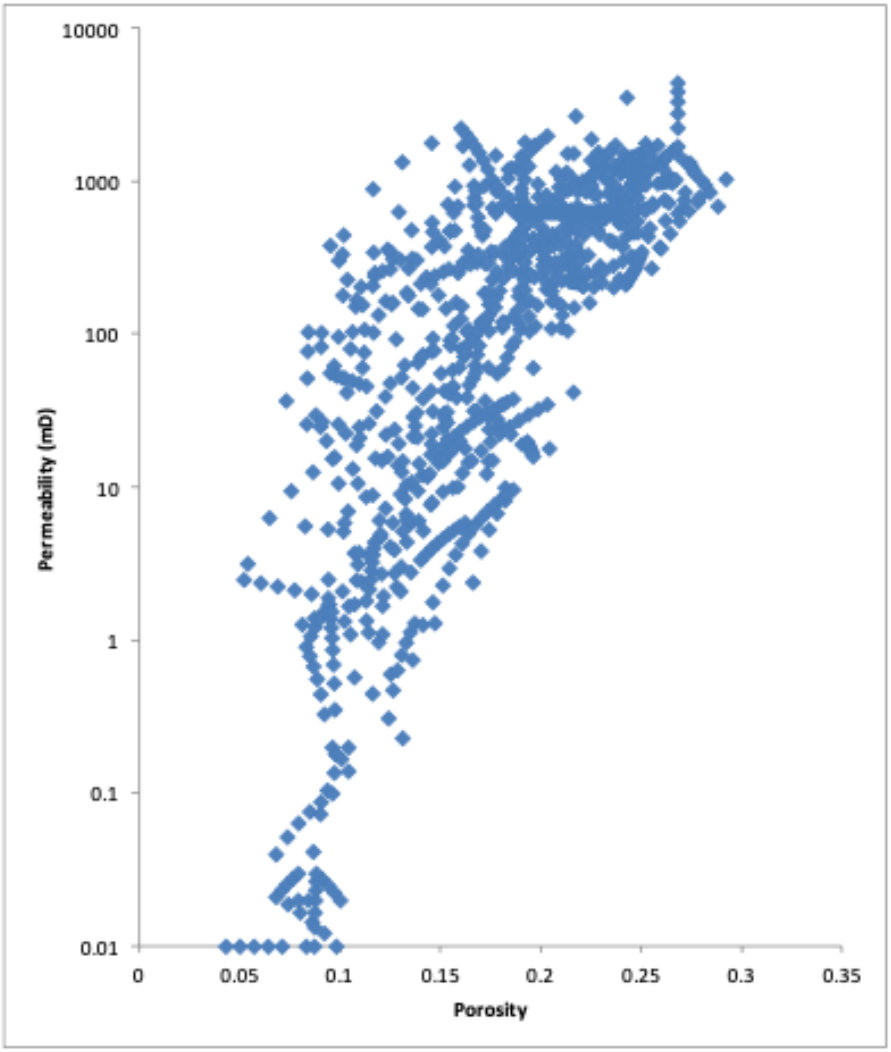


Figure A.1: Porosity–permeability cross-plot derived from core plug data, used as the basis for permeability prediction in the geomodel ensemble [33].

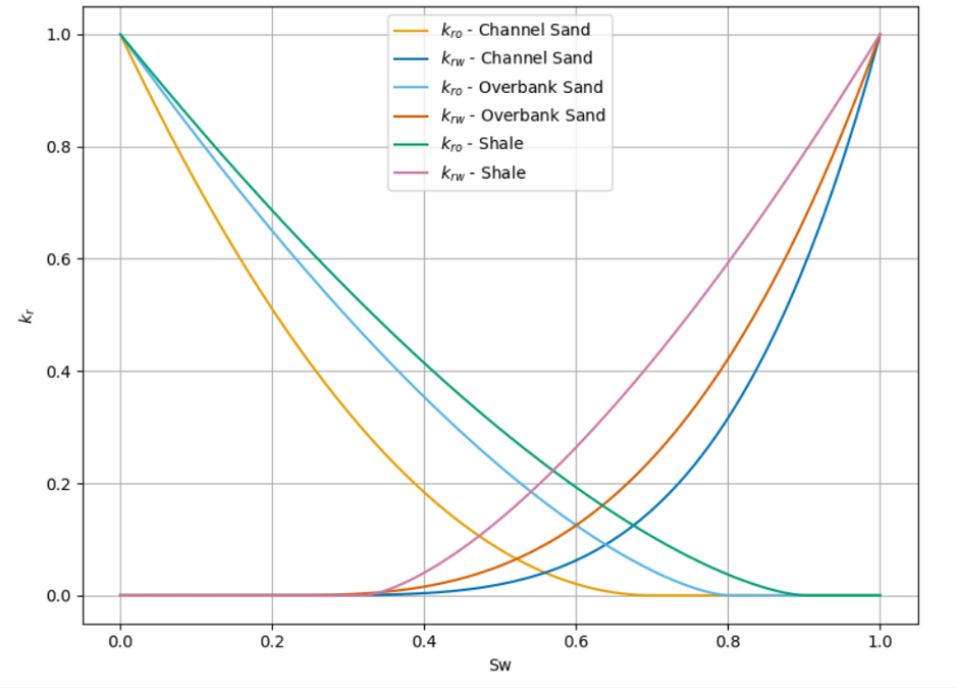


Figure A.2: Relative permeability curves for different facies.

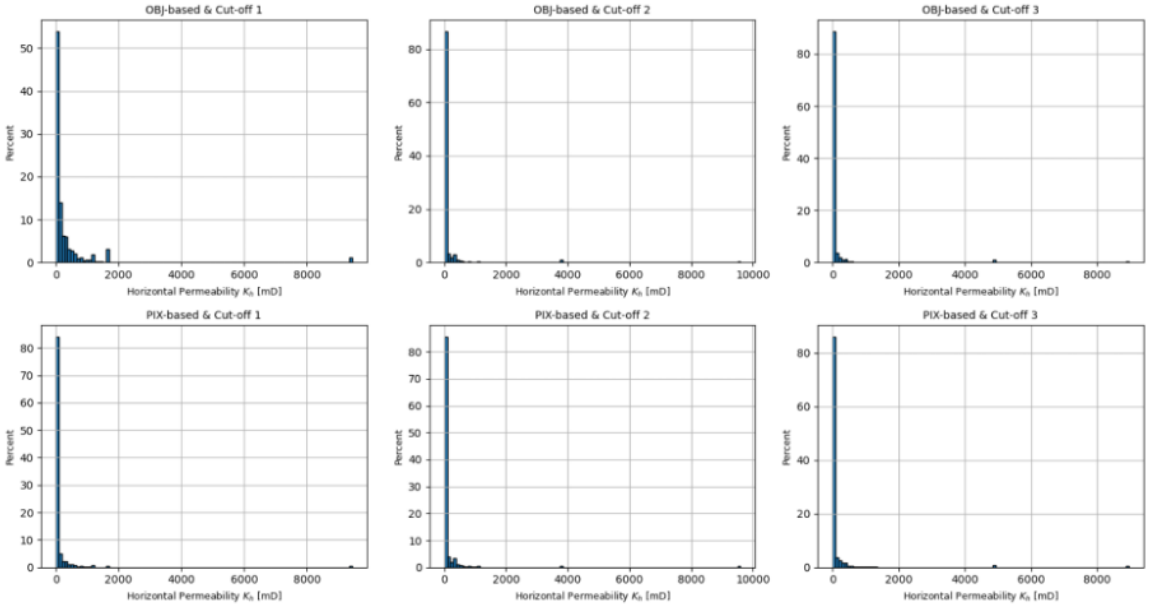


Figure A.3: Permeability distribution for different permeability models.

B

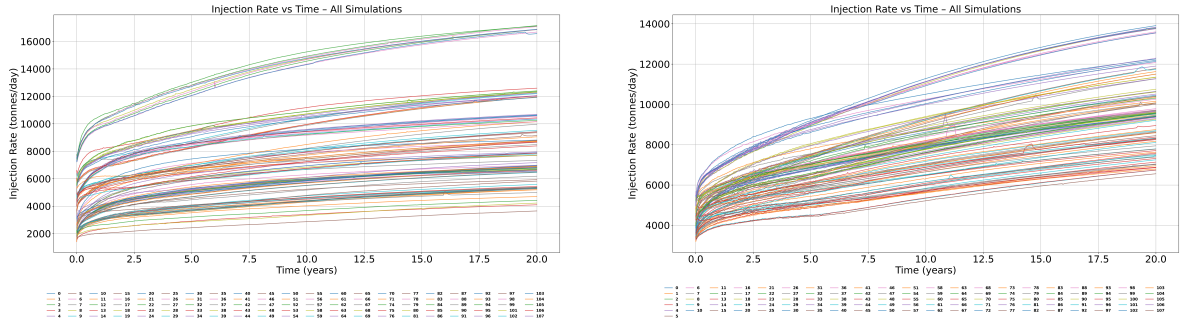
Additional Results open-DARTS

B.1 Non-Converged Realizations

Table B.1: Non-converged open-DARTS (full-physics) runs and corresponding parameter settings by ensemble.

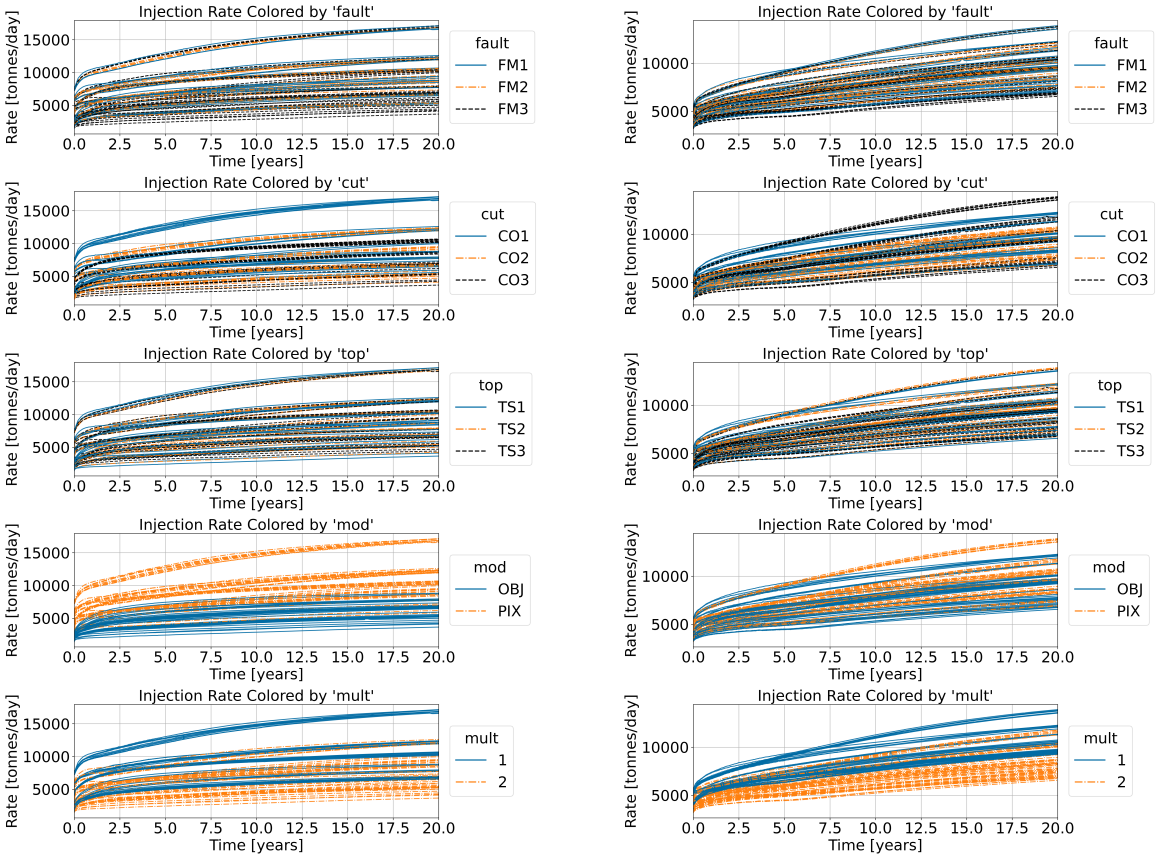
Run	Fault	Cut	Top	Mod	Mult	Failed in Ensemble
52	FM2	CO2	TS2	OBJ	1	Ensemble 2
53	FM2	CO2	TS2	OBJ	2	Ensemble 2
72	FM3	CO1	TS1	OBJ	1	Ensemble 1
73	FM3	CO1	TS1	OBJ	2	Ensemble 1
100	FM3	CO3	TS2	OBJ	1	Ensemble 1

B.2 Injection Rate Results



(a) Ensemble 1. **(b)** Ensemble 2.

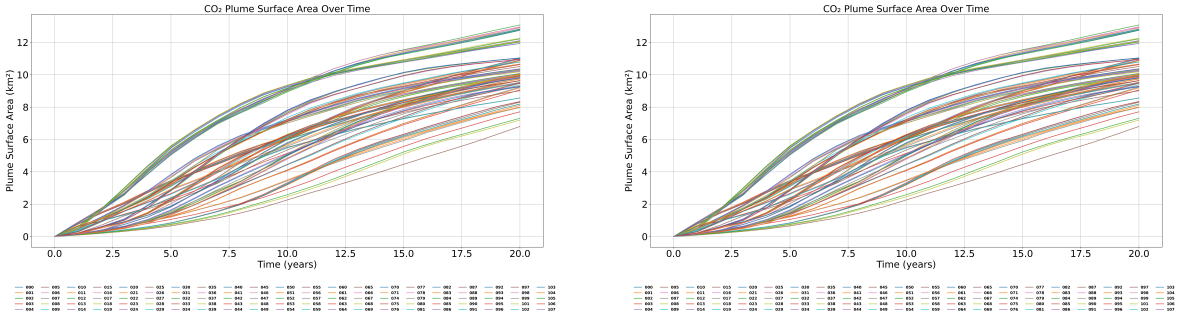
Figure B.1: Injection-rate results from the full-physics simulator, colored per run, for Ensembles 1 and 2.



(a) Ensemble 1. **(b)** Ensemble 2.

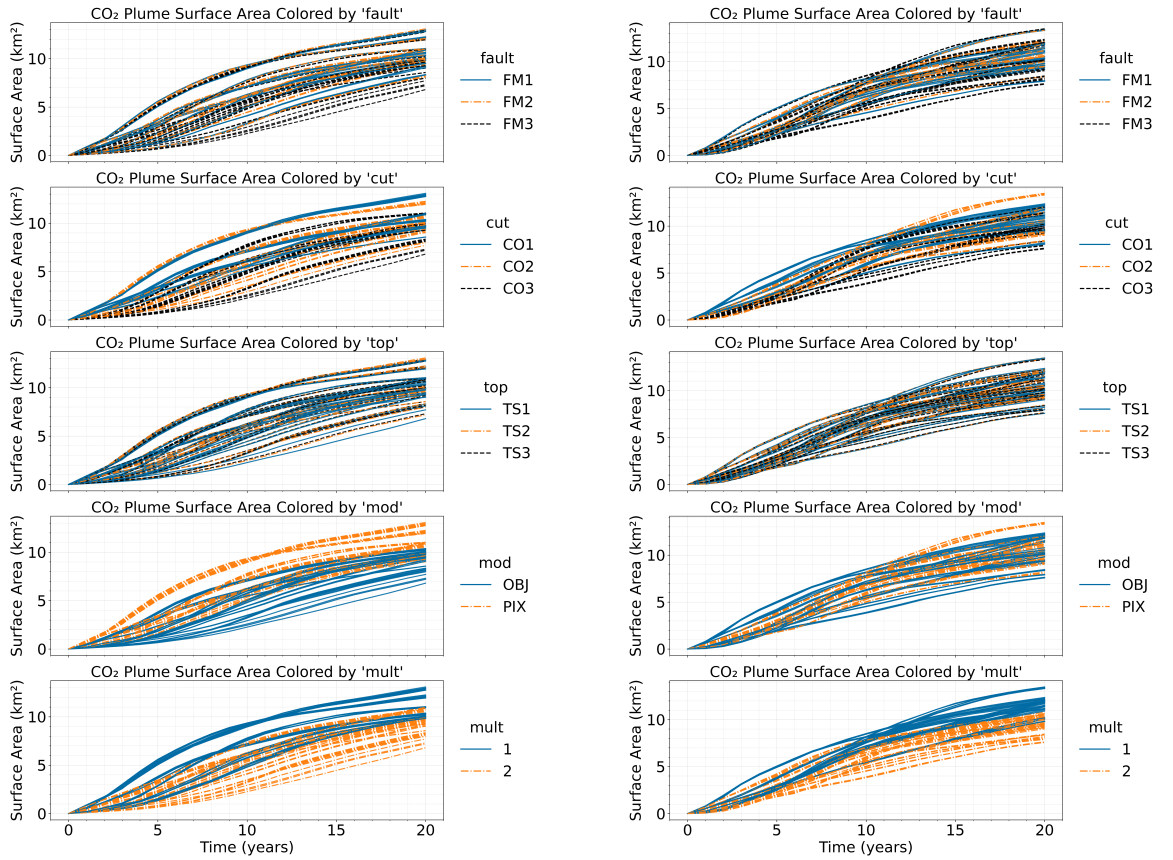
Figure B.2: Injection-rate results from the full-physics simulator, colored by varied parameters, for Ensembles 1 and 2.

B.3 Plume Areal Coverage Results



(a) Ensemble 1. **(b)** Ensemble 2.

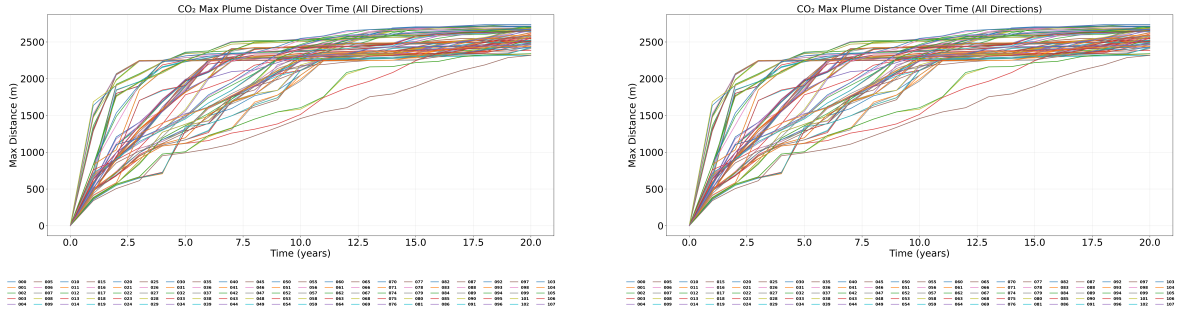
Figure B.3: Plume areal coverage results from the full-physics simulator, colored per run, for Ensembles 1 and 2.



(a) Ensemble 1. **(b)** Ensemble 2.

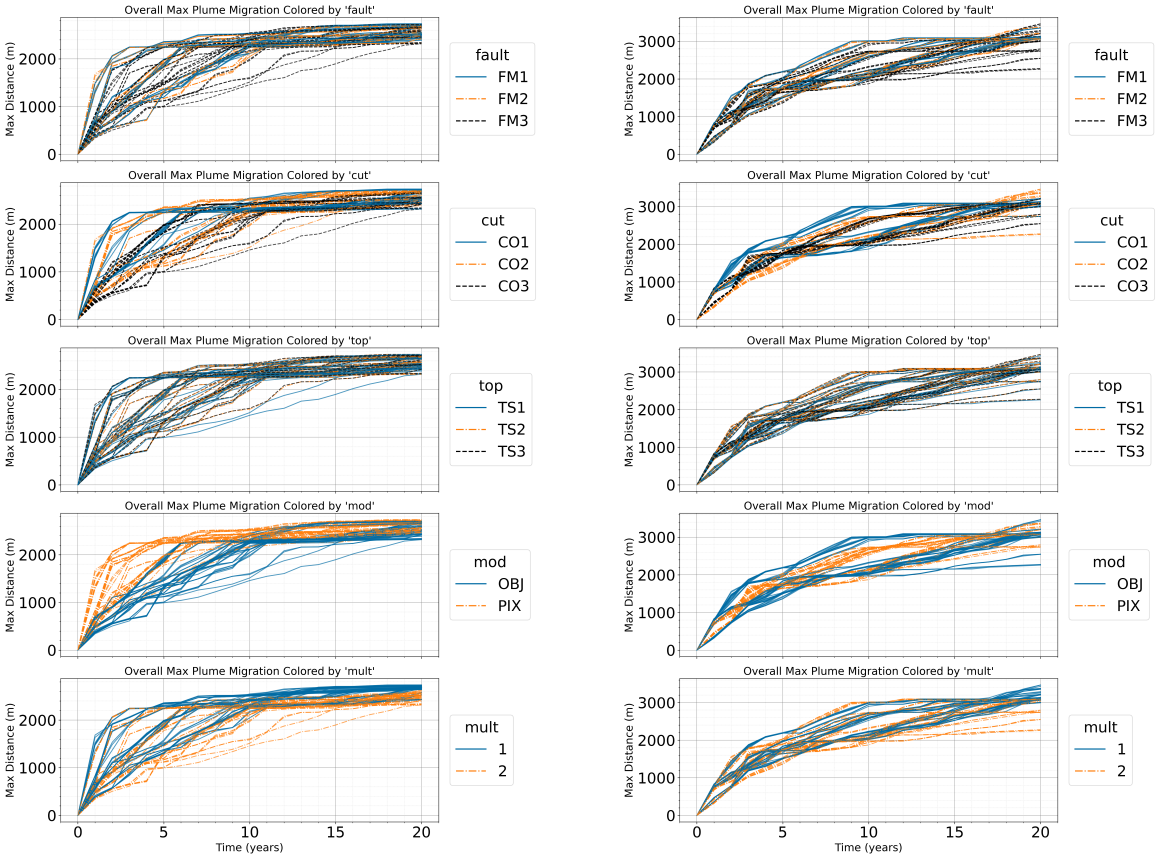
Figure B.4: Plume areal coverage results from the full-physics simulator, colored by varied parameters, for Ensembles 1 and 2.

B.4 Maximum Plume Extent Results



(a) Ensemble 1. **(b)** Ensemble 2.

Figure B.5: Maximum plume extent results from the full-physics simulator, colored per run, for Ensembles 1 and 2.

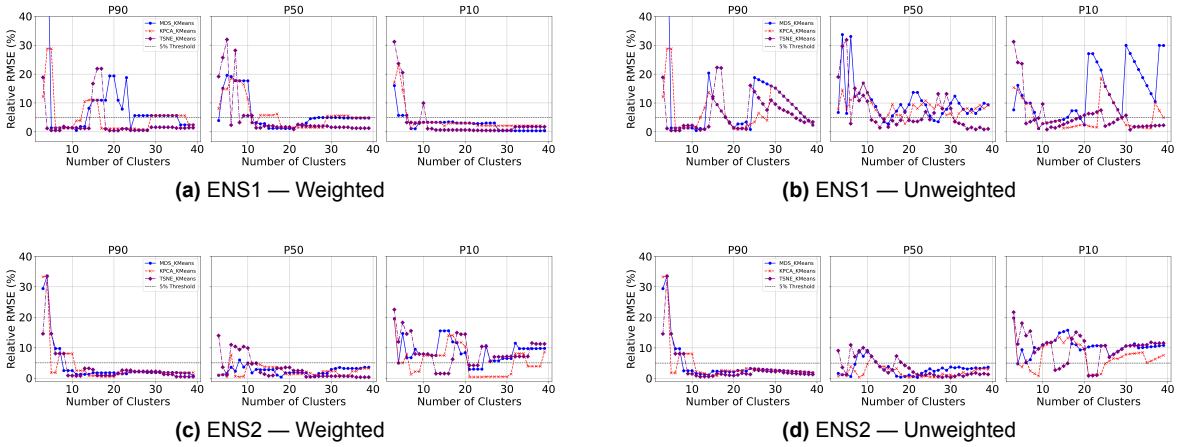


(a) Ensemble 1. **(b)** Ensemble 2.

Figure B.6: Maximum plume extent results from the full-physics simulator, colored by varied parameters, for Ensembles 1 and 2.

C

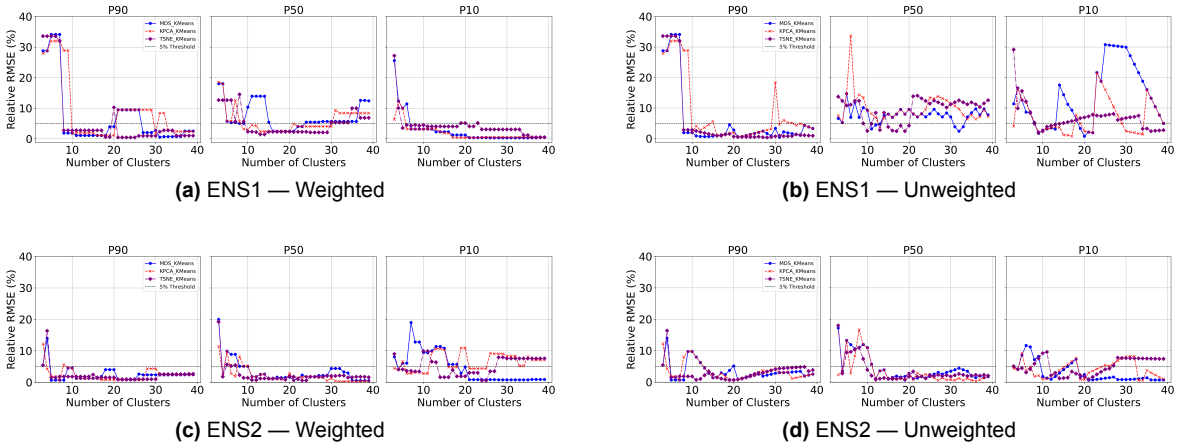
Weighted vs Unweighted Percentile Reconstruction



(a) ENS1 — Weighted **(b)** ENS1 — Unweighted

(c) ENS2 — Weighted **(d)** ENS2 — Unweighted

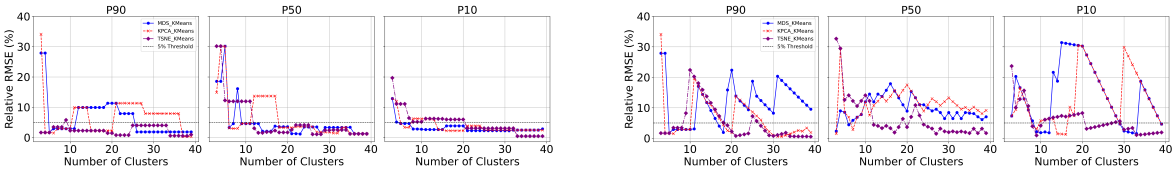
Figure C.1: Weighted vs unweighted percentile reconstruction (relative RMSE, %) for the FP-D1 diagnostic. Rows: ENS1/ENS2. Columns: weighted/unweighted. Curves: P_{90}, P_{50}, P_{10} for t-SNE, KPCA, MDS.



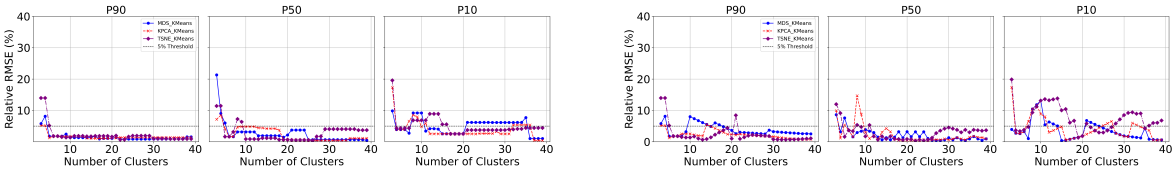
(a) ENS1 — Weighted **(b)** ENS1 — Unweighted

(c) ENS2 — Weighted **(d)** ENS2 — Unweighted

Figure C.2: Weighted vs unweighted percentile reconstruction (relative RMSE, %) for the FP-D10 diagnostic. Rows: ENS1/ENS2. Columns: weighted/unweighted.

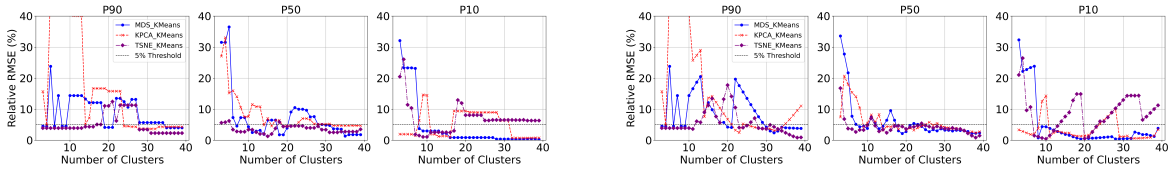


(a) ENS1 — Weighted **(b)** ENS1 — Unweighted

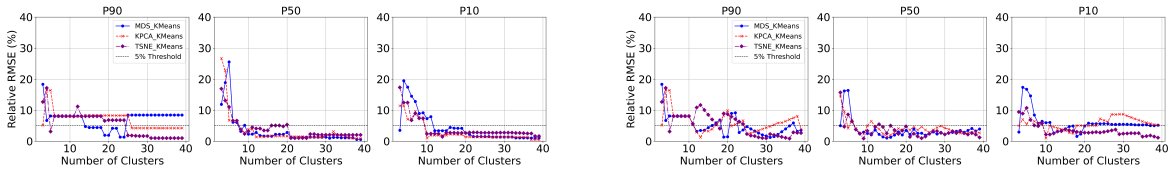


(c) ENS2 — Weighted **(d)** ENS2 — Unweighted

Figure C.3: Weighted vs unweighted percentile reconstruction (relative RMSE, %) for the FP-D100 diagnostic. Rows: ENS1/ENS2. Columns: weighted/unweighted.

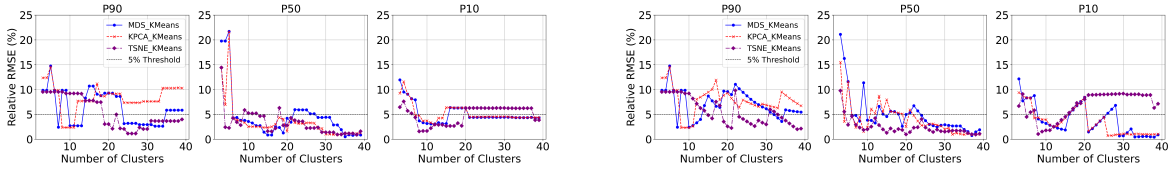


(a) ENS1 — Weighted **(b)** ENS1 — Unweighted

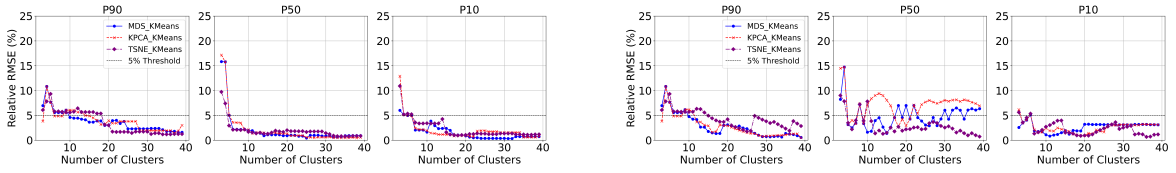


(c) ENS2 — Weighted **(d)** ENS2 — Unweighted

Figure C.4: Weighted vs unweighted percentile reconstruction (relative RMSE, %) for plume areal coverage (IMM-SAT-PS11 diagnostic). Rows: ENS1/ENS2. Columns: weighted/unweighted.



(a) ENS1 — Weighted **(b)** ENS1 — Unweighted

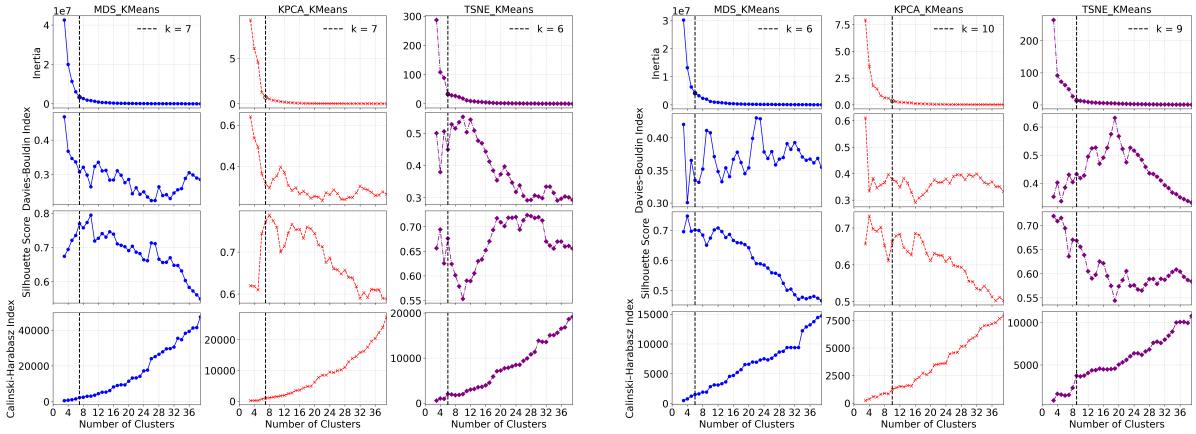


(c) ENS2 — Weighted **(d)** ENS2 — Unweighted

Figure C.5: Weighted vs unweighted percentile reconstruction (relative RMSE, %) for maximum plume extent (IMM-SAT-PS11 diagnostic). Rows: ENS1/ENS2. Columns: weighted/unweighted.

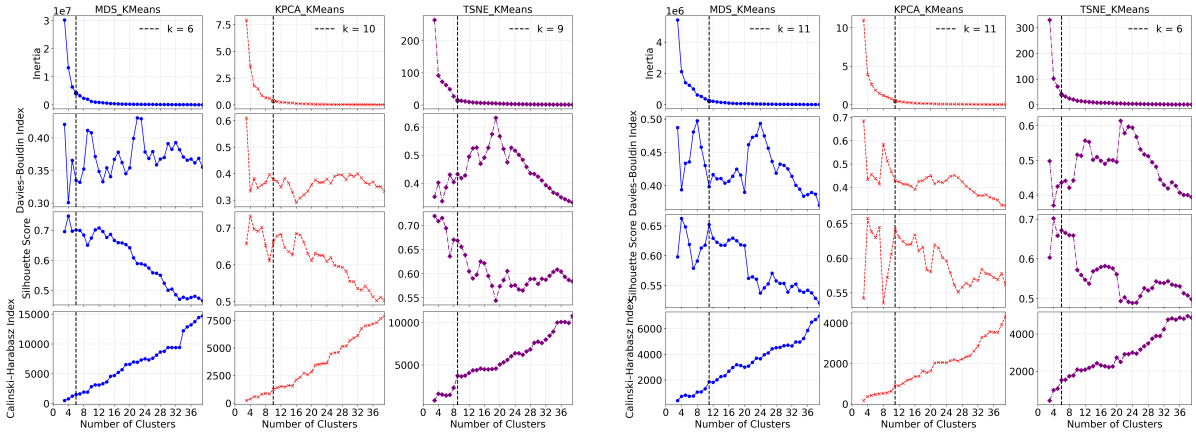
D

Inertia Cluster Selection Guidance



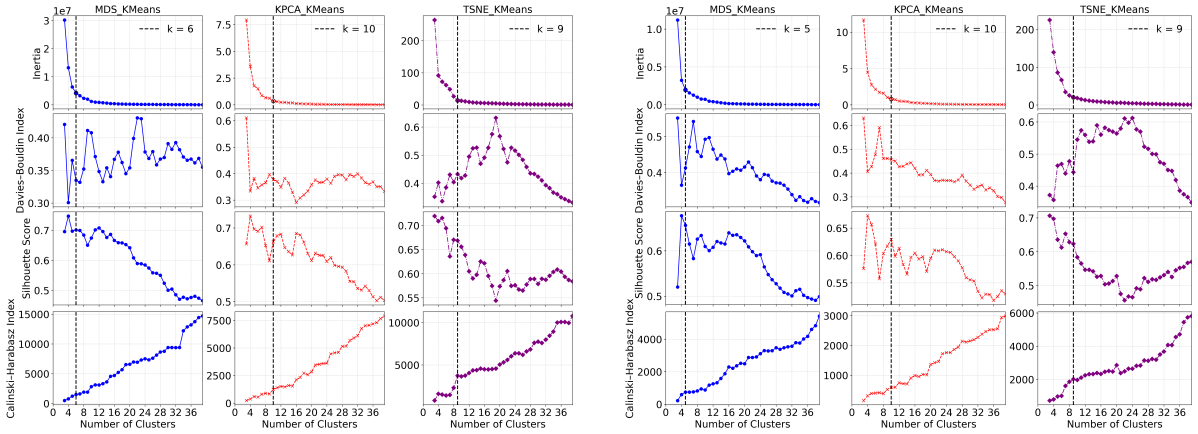
(a) Ensemble 1 **(b)** Ensemble 2

Figure D.1: Internal clustering metrics (inertia, Davies–Bouldin, silhouette, Calinski–Harabasz) versus cluster count for FP-D1. Vertical dashed line indicates identified inertia elbow used in this study.



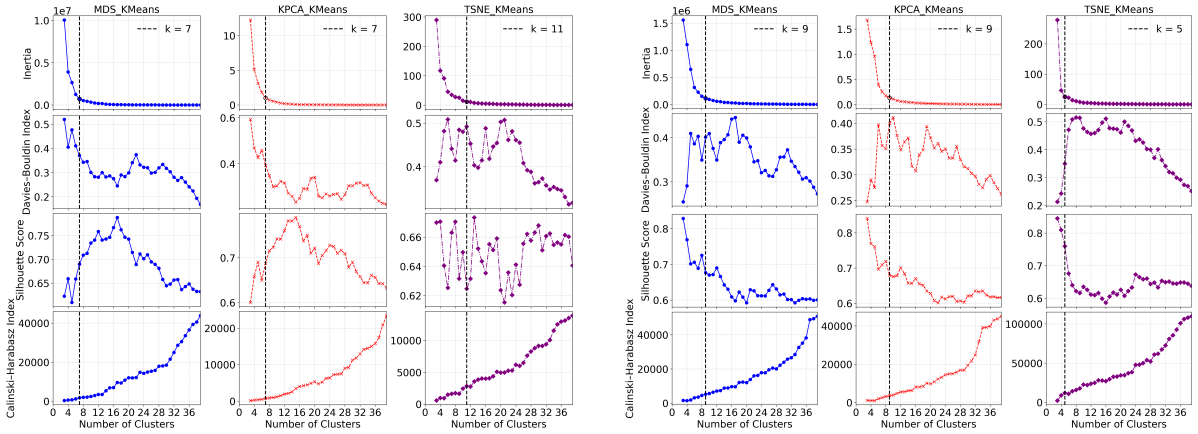
(a) Ensemble 1 **(b)** Ensemble 2

Figure D.2: Internal clustering metrics versus cluster count for FP-D10. Vertical dashed line indicates identified inertia elbow used in this study.



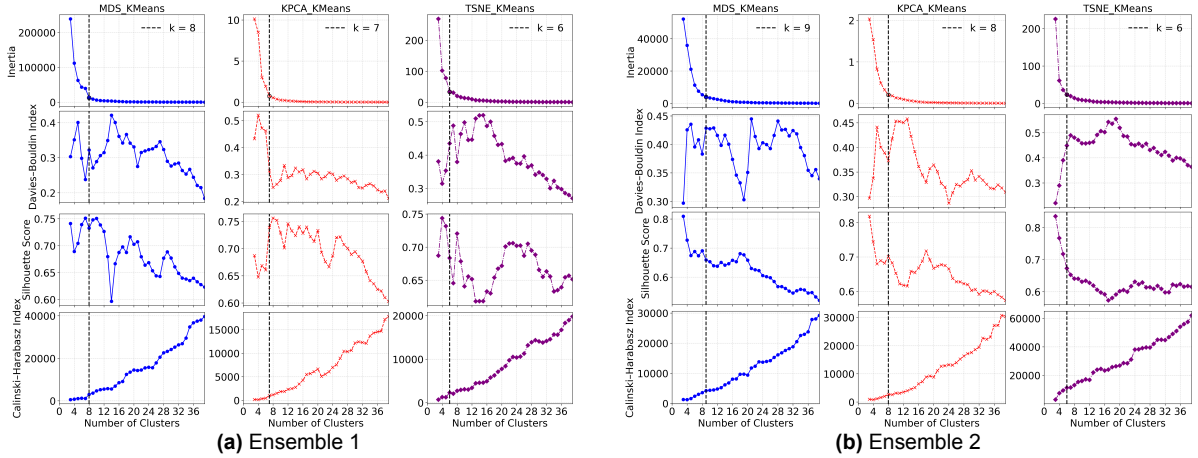
(a) Ensemble 1 **(b)** Ensemble 2

Figure D.3: Internal clustering metrics versus cluster count for FP-D100. Vertical dashed line indicates identified inertia elbow used in this study.



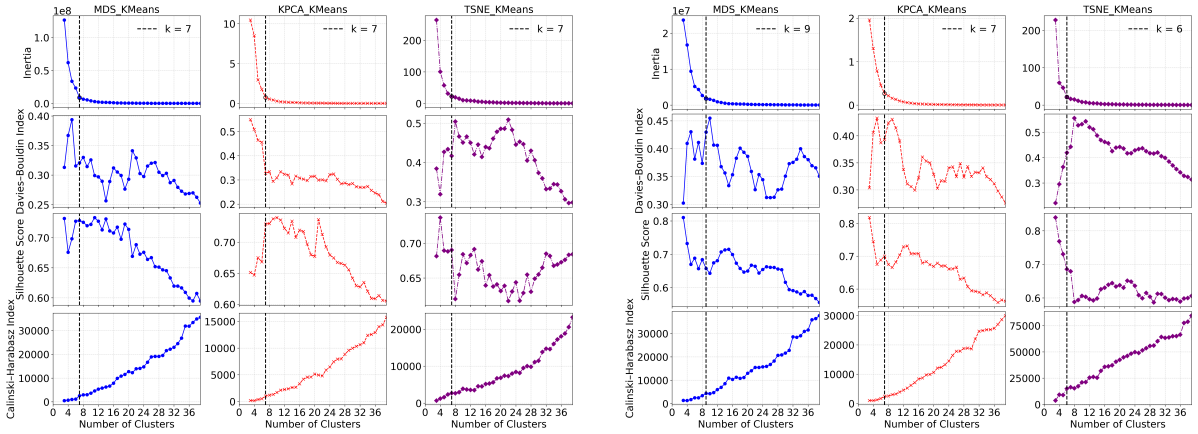
(a) Ensemble 1 **(b)** Ensemble 2

Figure D.4: Internal clustering metrics versus cluster count for SP-PS1. Vertical dashed line indicates identified inertia elbow used in this study.



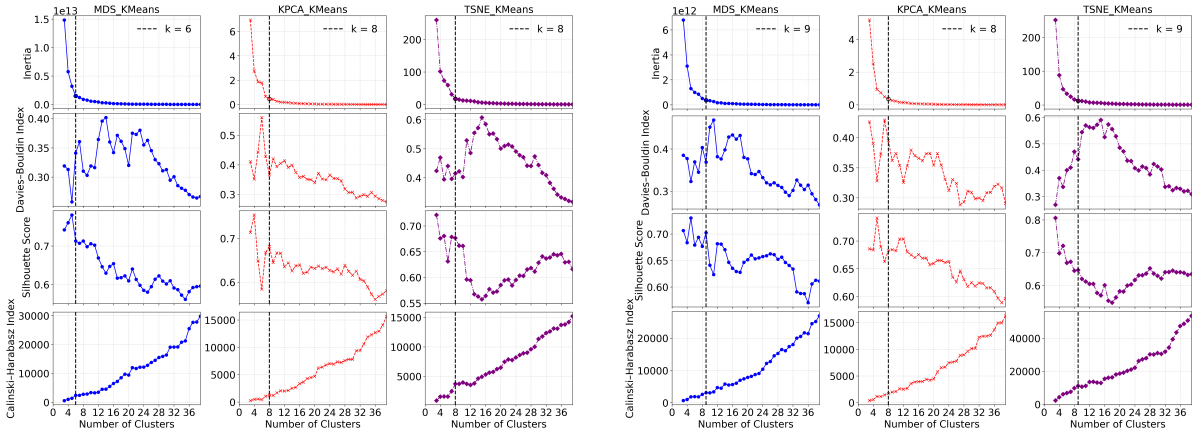
(a) Ensemble 1 **(b)** Ensemble 2

Figure D.5: Internal clustering metrics versus cluster count for IMM-PS1. Vertical dashed line indicates identified inertia elbow used in this study.



(a) Ensemble 1 **(b)** Ensemble 2

Figure D.6: Internal clustering metrics versus cluster count for IMM-PS4. Vertical dashed line indicates identified inertia elbow used in this study.

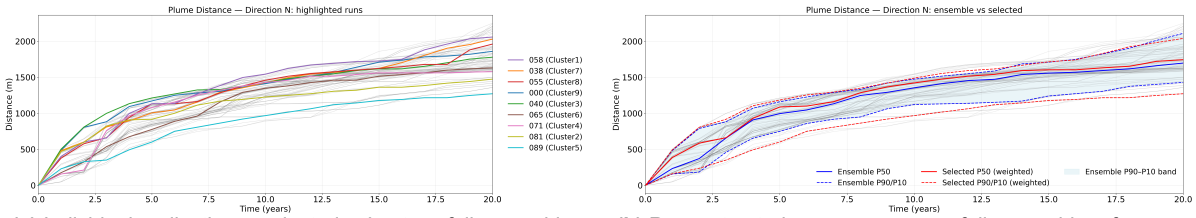


(a) Ensemble 1 **(b)** Ensemble 2

Figure D.7: Internal clustering metrics versus cluster count for IMM-PS11. Vertical dashed line indicates identified inertia elbow used in this study.

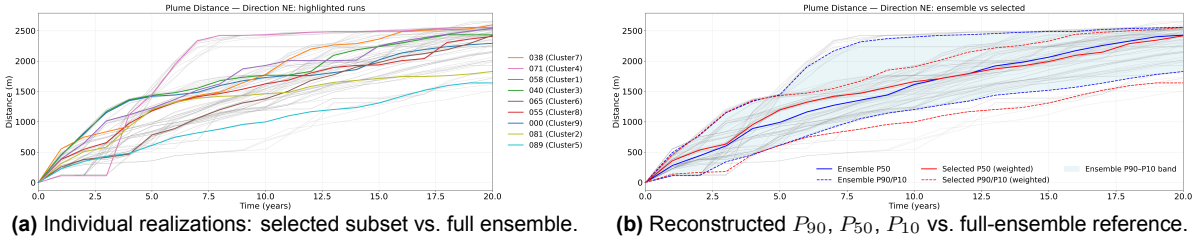
E

Directional Percentile Reconstruction of Maximum Plume Extent (FP-D10 + t-SNE)



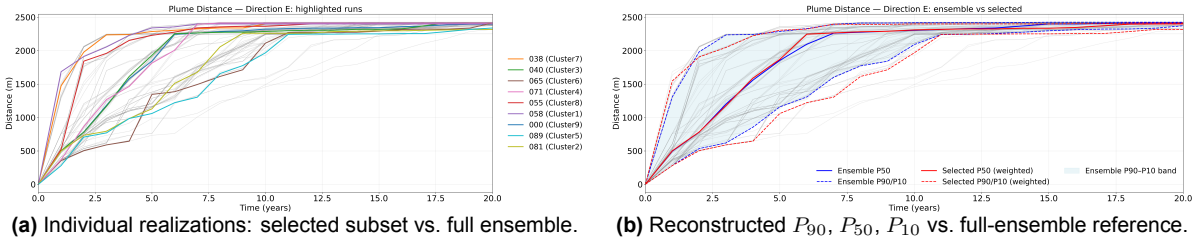
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.1: Ensemble 1, direction N: Maximum plume distance over time (left) for the t-SNE–selected subset (coloured) compared with all runs (grey), and reconstructed percentiles (right) from the subset (red) against the full-ensemble reference (blue).



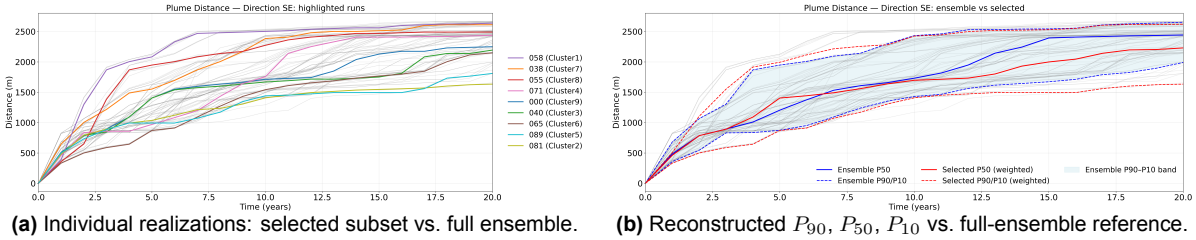
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.2: Ensemble 1, direction NE: As in Fig. N for NE.



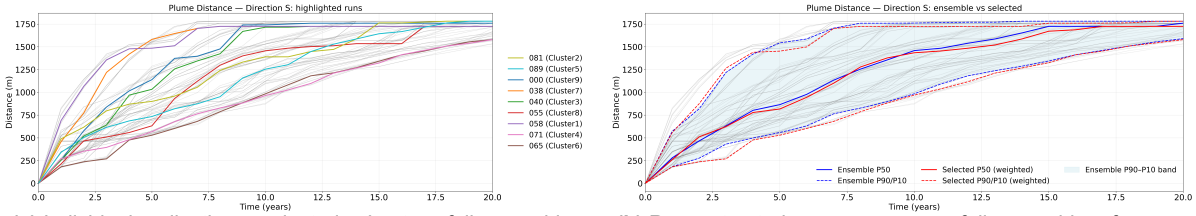
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.3: Ensemble 1, direction E: As in Fig. N for E.



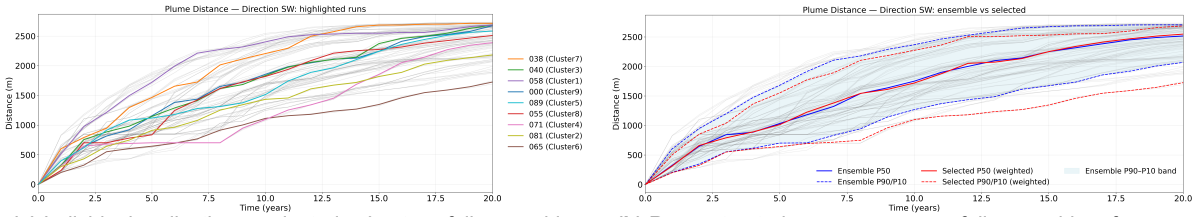
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.4: Ensemble 1, direction SE: As in Fig. N for SE.



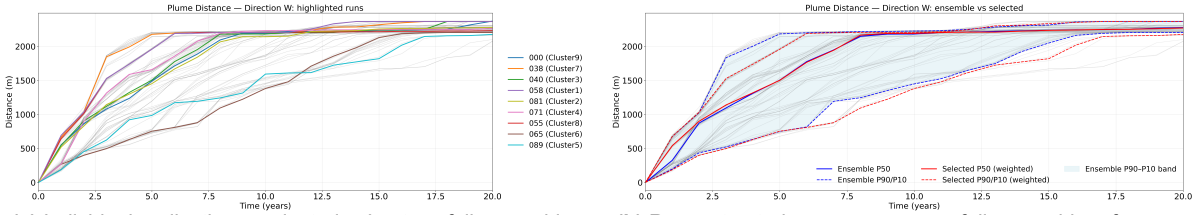
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.5: Ensemble 1, direction S: As in Fig. N for S.



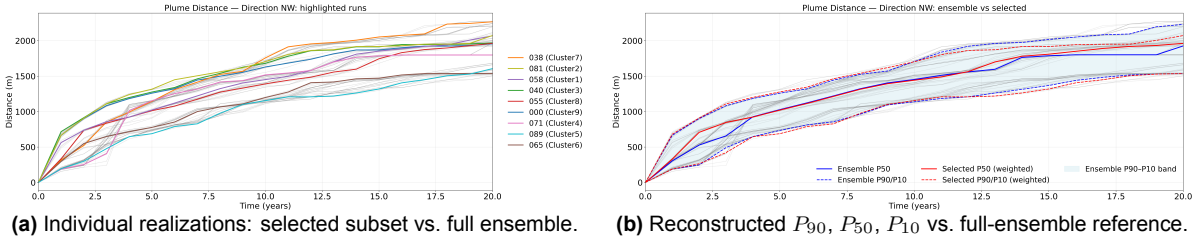
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.6: Ensemble 1, direction SW: As in Fig. N for SW.



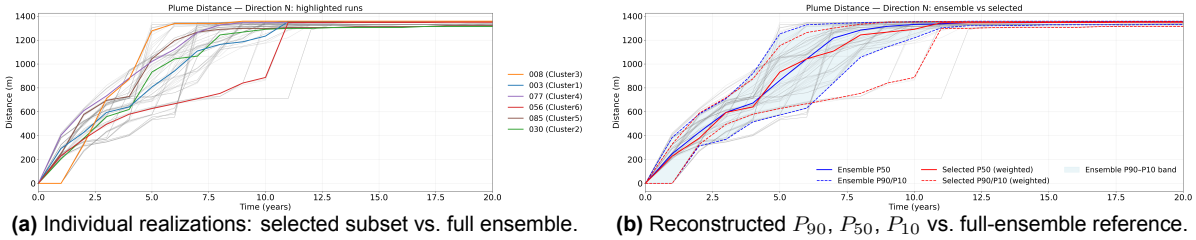
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.7: Ensemble 1, direction W: As in Fig. N for W.



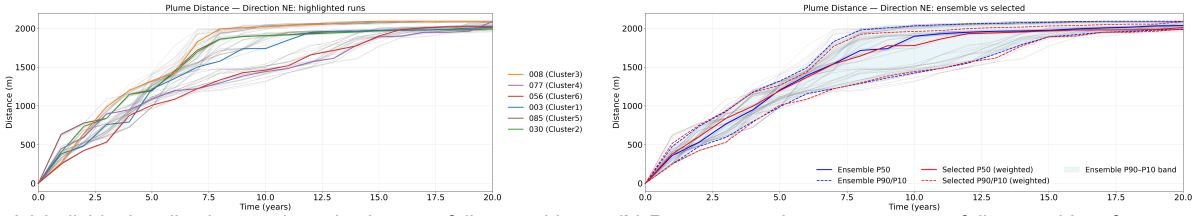
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.8: Ensemble 1, direction NW: As in Fig. N for NW.



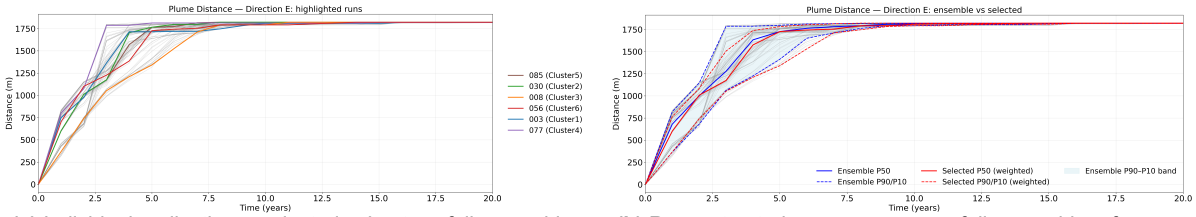
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.9: Ensemble 2, direction N: As in Ensemble 1, direction N.



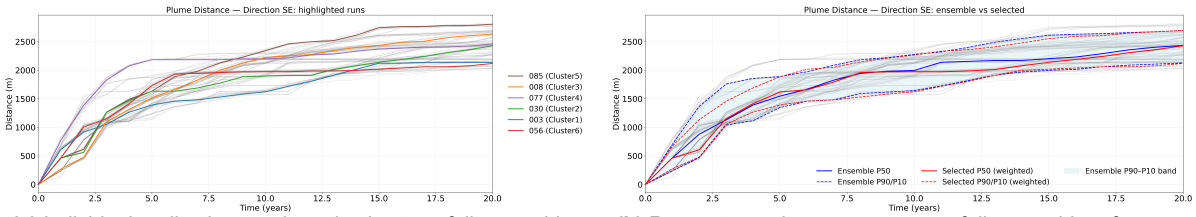
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.10: Ensemble 2, direction NE: As in Ensemble 1, direction NE.



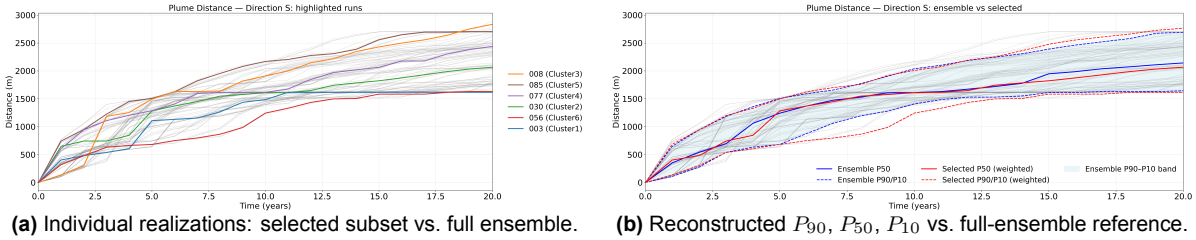
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.11: Ensemble 2, direction E: As in Ensemble 1, direction E.



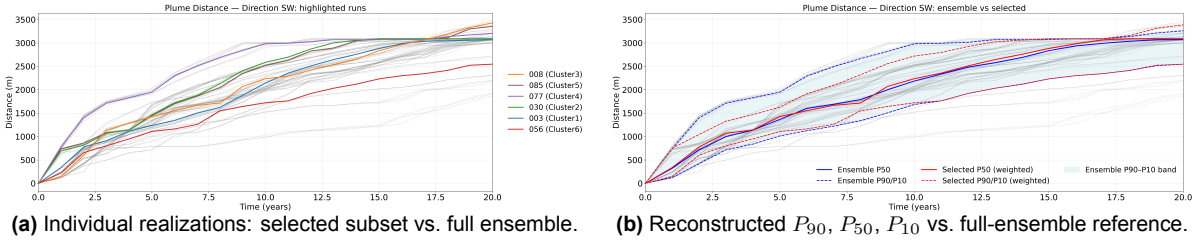
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.12: Ensemble 2, direction SE: As in Ensemble 1, direction SE.



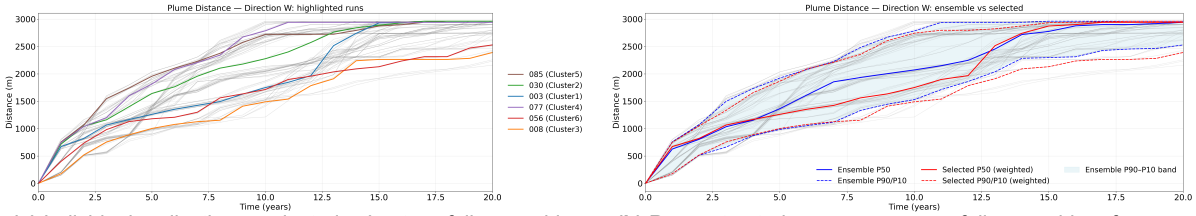
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.13: Ensemble 2, direction S: As in Ensemble 1, direction S.



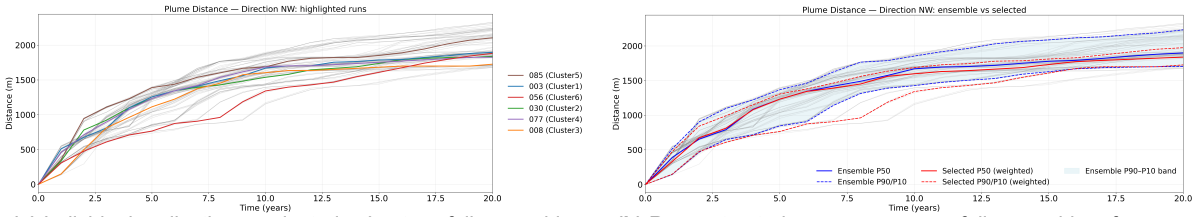
(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.14: Ensemble 2, direction SW: As in Ensemble 1, direction SW.



(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.15: Ensemble 2, direction W: As in Ensemble 1, direction W.

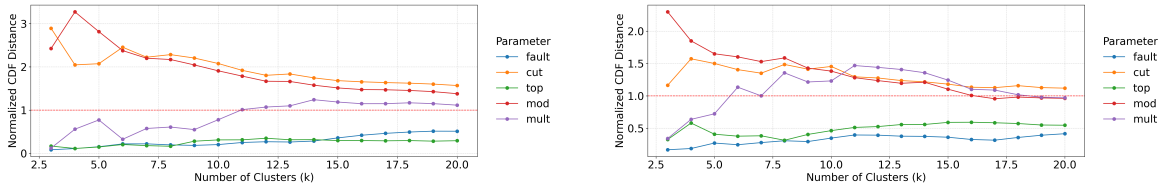


(a) Individual realizations: selected subset vs. full ensemble. **(b)** Reconstructed P_{90}, P_{50}, P_{10} vs. full-ensemble reference.

Figure E.16: Ensemble 2, direction NW: As in Ensemble 1, direction NW.

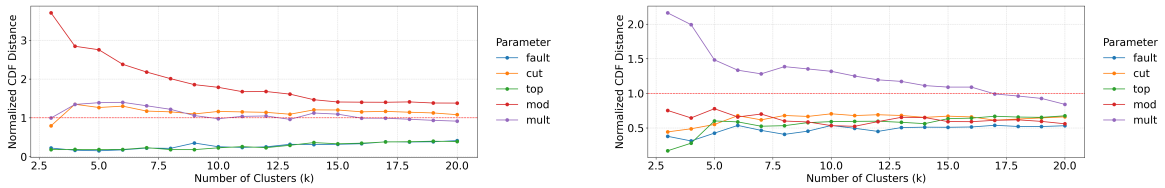
Additional Results Injection Rate Analysis

F.1 dGSA Analysis: FP-D1 and FP-D100 + t-SNE



(a) Ensemble 1 **(b)** Ensemble 2

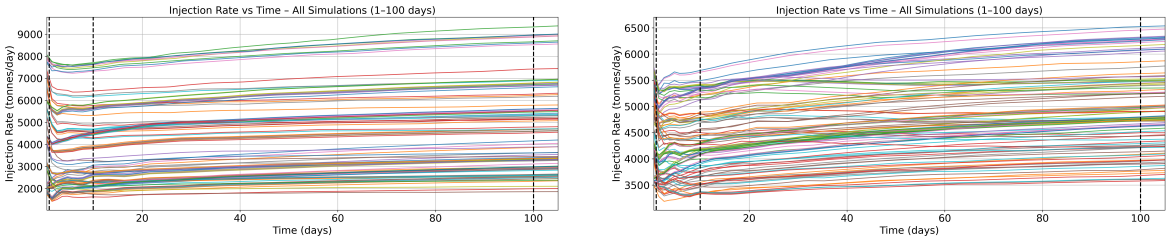
Figure F.1: Distance-based generalized sensitivity analysis over increasing cluster counts for FP-D1 combined with t-SNE. The dGSA curves show how influential each parameter is according to the clusters formed. Parameters with a normalized CDF exceeding $S > 1$ are considered influential, indicating that their values vary strongly across response clusters and thus have a strong impact on cluster formation. Results are shown for Ensemble 1 and Ensemble 2.



(a) Ensemble 1 **(b)** Ensemble 2

Figure F.2: Distance-based generalized sensitivity analysis over increasing cluster counts for FP-D100 combined with t-SNE. The dGSA curves show how influential each parameter is according to the clusters formed. Parameters with a normalized CDF exceeding $S > 1$ are considered influential, indicating that their values vary strongly across response clusters and thus have a strong impact on cluster formation. Results are shown for Ensemble 1 and Ensemble 2.

F.2 Open-DARTS Injection Rate Profiles During the First 100 Days



(a) Ensemble 1 **(b)** Ensemble 2

Figure F.3: Injection rates for all realizations over the first 100 days. Vertical dashed lines mark day 1, day 10, and day 100. Note the more frequent rank crossings among realizations at early simulation times ($\lesssim 10$ days) and the subsequent stabilization of the relative ranking as the simulation horizon increases.

F.3 Distribution of Injection Rates at 1, 10, and 100 Days (Ensemble 2)

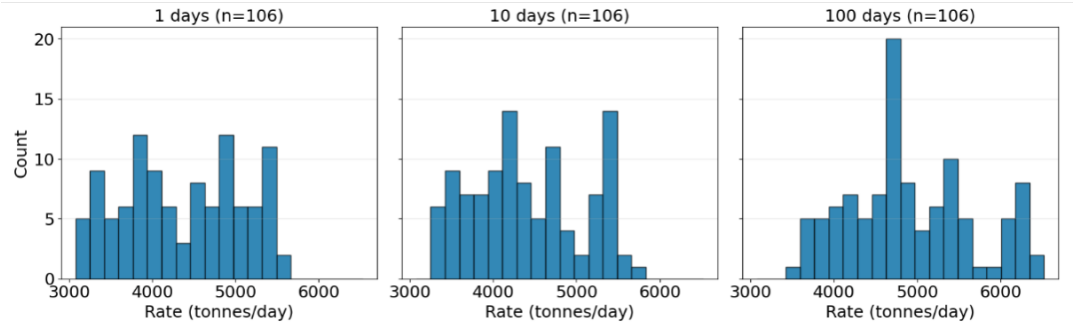
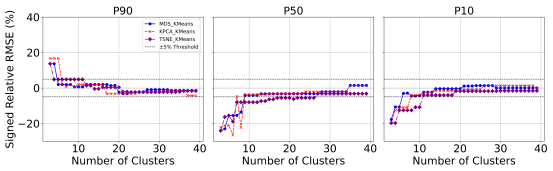
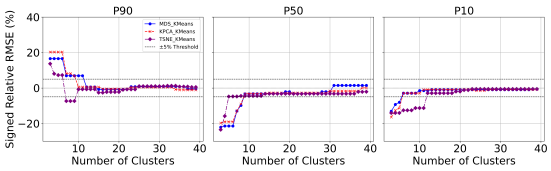


Figure F.4: Distribution of Injection Rates at Selected Times (1, 10, and 100 days) - Ensemble 2

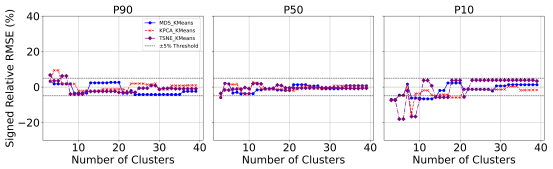
F.4 Signed Relative RMSE for Early Injection-Rate Diagnostics with Reduced Fault-Transmissibility Contrast



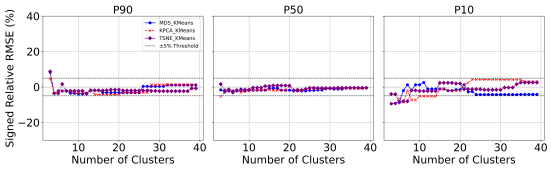
(a) Ensemble 1 — DP-D10



(b) Ensemble 1 — DP-D100



(c) Ensemble 2 — DP-D10



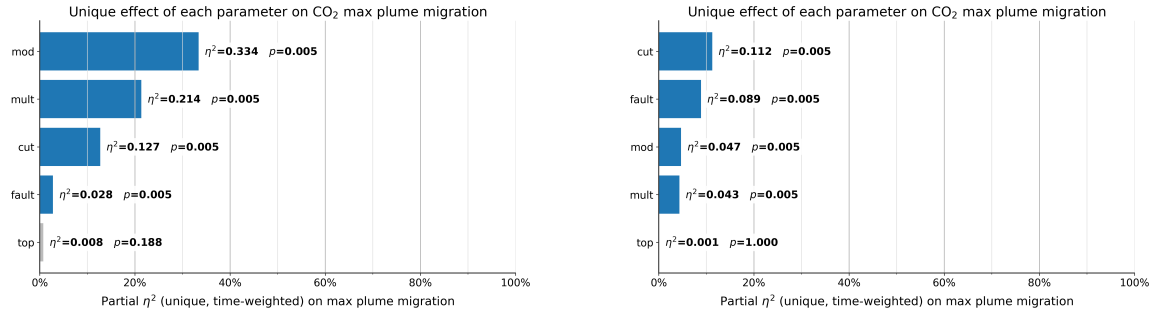
(d) Ensemble 2 — DP-D100

Figure F.5: Signed relative RMSE (weighted) for full-physics injection-rate diagnostics across ensembles. Results show percentile reconstruction RMSE over varying cluster counts for the injection-rate storage metric. Both Ensemble 1 and 2 consist of equal numbers of realizations with a fault transmissibility multiplier of 0.9 and 0.1, in contrast to the originally studied ensembles which used a wider contrast (0.9 vs. 0.01). Notably, the day-10 rate diagnostic still performs accurately, indicating that even with less restrictive flow barriers, early injection rates remain sensitive to system-wide effects such as fault transmissibility and can differentiate realizations on this basis.

G

Additional Results Plume Behavior Analysis

η^2 Sensitivity Analysis for Maximum Plume Extent



(a) Ensemble 1 migration

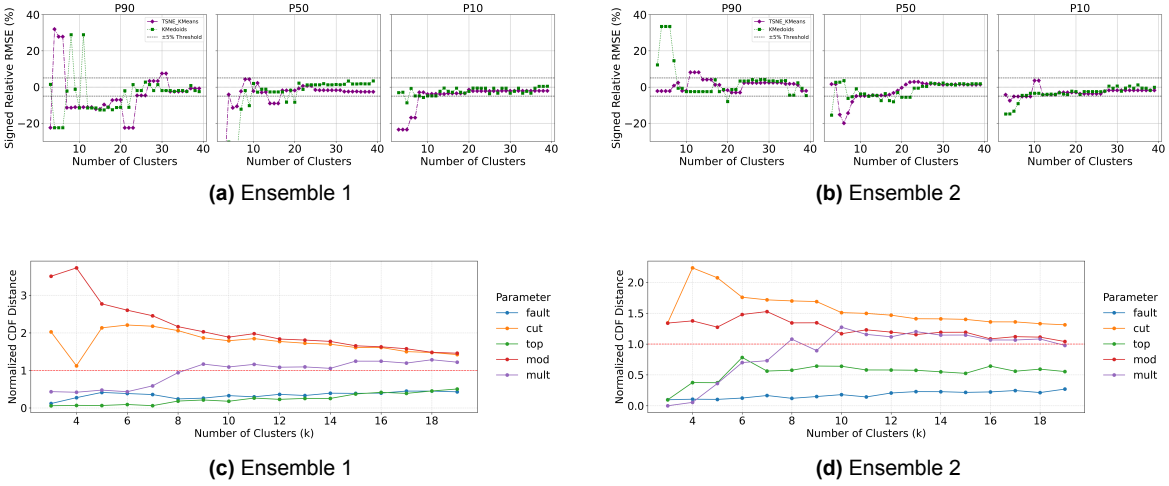
(b) Ensemble 2 migration

Figure G.1: η^2 sensitivity indices for maximum plume migration distance obtained from the full-physics simulations for Ensembles 1 and 2. The indices quantify the contribution of each geological parameter to the variability of plume migration across the ensemble over the complete simulation duration.

H

Additional K-medoids Inclusion Analysis

H.1 K-medoids Results for Streamline-Based Rate Diagnostics



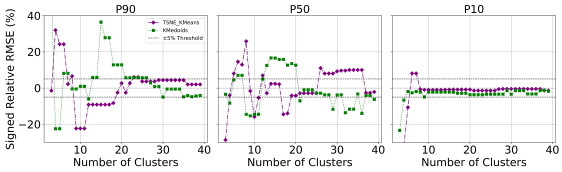
(a) Ensemble 1

(b) Ensemble 2

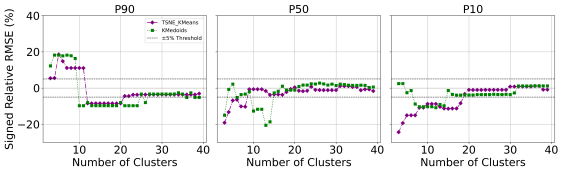
(c) Ensemble 1

(d) Ensemble 2

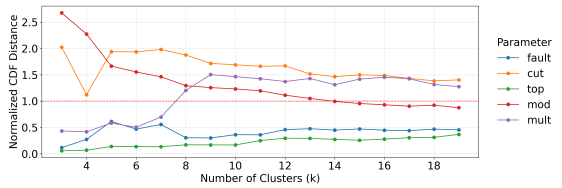
Figure H.1: Single-phase injection-rate (SP-PS1) diagnostic using K-medoids. Top row (a–b): signed relative RMSE over increasing cluster counts. Bottom row (c–d): distance-based generalized sensitivity analysis over increasing cluster counts; parameters with normalized CDF $S > 1$ are considered influential.



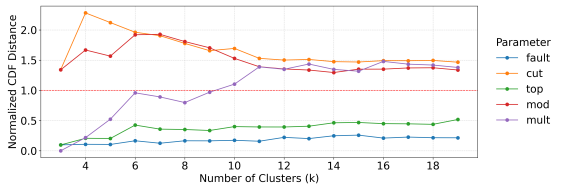
(a) Ensemble 1



(b) Ensemble 2

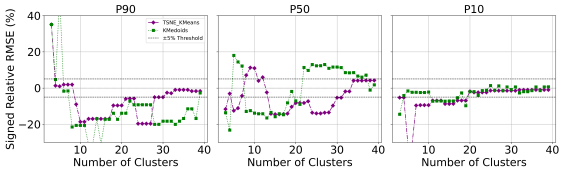


(c) Ensemble 1

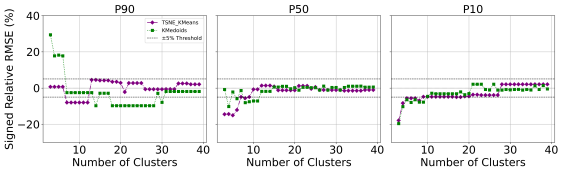


(d) Ensemble 2

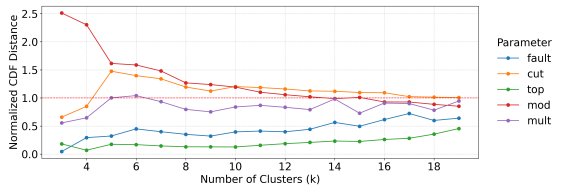
Figure H.2: Immiscible injection-rate diagnostic at the first pressure solve (IMM-PS1) using K-medoids. Top row (a–b): signed relative RMSE over increasing cluster counts. Bottom row (c–d): distance-based generalized sensitivity analysis over increasing cluster counts.



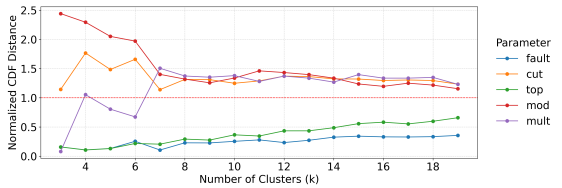
(a) Ensemble 1



(b) Ensemble 2



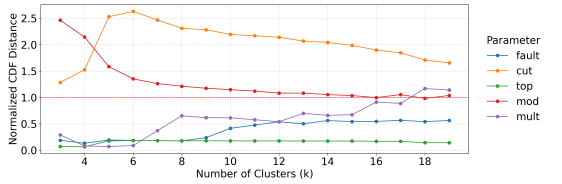
(c) Ensemble 1



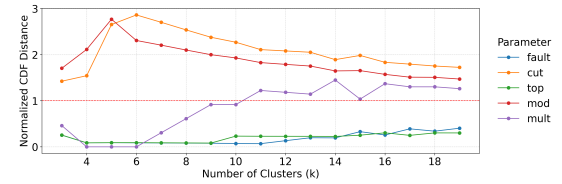
(d) Ensemble 2

Figure H.3: Immiscible injection-rate diagnostics after eleven pressure solves (IMM-PS11) using K-medoids. Top row (a–b): signed relative RMSE over 20 years. Bottom row (c–d): distance-based generalized sensitivity analysis; parameters with normalized CDF $S > 1$ are considered influential.

H.2 K-medoids Results for Immiscible Saturation-Field Diagnostic



(a) KMedoids ENS1 coverage



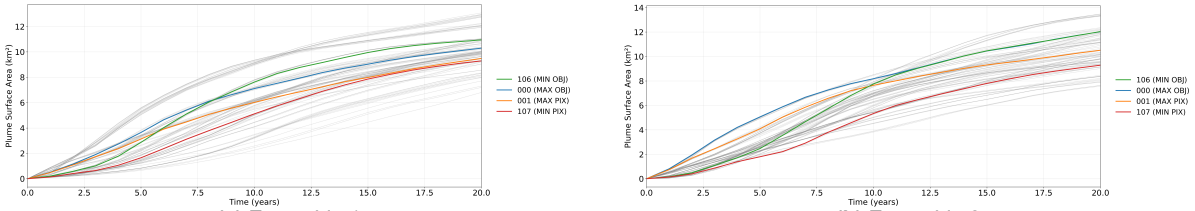
(b) KMedoids ENS2 coverage

Figure H.4: Distance-based generalized sensitivity analysis over increasing cluster counts for plume areal coverage using K-medoids on the immiscible saturation-field diagnostic (IMM-SAT-PS11), for Ensembles 1 and 2. Parameters with normalized CDF $S > 1$ are considered influential.

I

Additional Results: Selecting Minimum and Maximum Cases

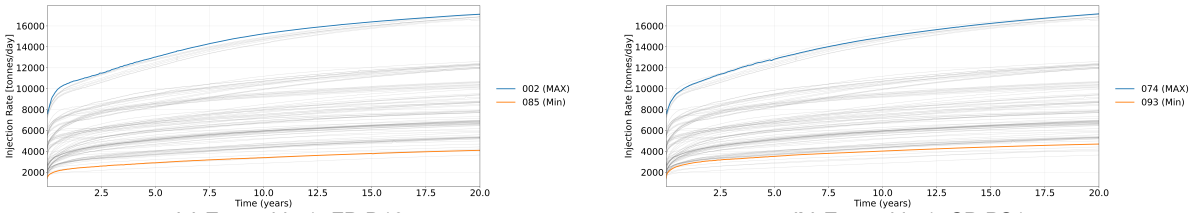
I.1 Plume Areal Coverage: Parameter-Based Extreme Selections



(a) Ensemble 1 **(b)** Ensemble 2

Figure I.1: Plume areal-coverage profiles for the realizations assumed (from parameter configuration) to represent the lowest and highest areal coverage in each ensemble. Colors denote objective-based (OBJ) and pixel-based (PIX) selections for both high and low cases.

I.2 Injection Rate: open-DARTS vs 3DSL



(a) Ensemble 1: FP-D10 **(b)** Ensemble 1: SP-PS1

Figure I.2: Comparison of screening diagnostics from 3DSL (single-phase) and open-DARTS (10-day early-rate) for identifying extreme injection-rate realizations for Ensemble 1