POLITECNICO DI MILANO
DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING
DOCTORAL PROGRAMME IN ENVIRONMENTAL AND INFRASTRUCTURE ENGINEERING

Probabilistic Study of Fluid Migration and Allocation of Underground Energy Resources

Doctoral Dissertation of:
Rafael Leonardo Sandoval Pabon

Supervisor:
Prof. Monica Riva

Tutor:
Prof. Alberto Guadagnini

The Chair of the Doctoral Program:
Prof. Monica Riva

35 Cycle – 2023

A Alba, María Fernanda,
Verónica y Eliana.

Acknowledgements

I would like to express my deep gratitude to my professors, Monica and Alberto, who believed in me from the very beginning and tirelessly guided me on this journey. I humbly acknowledge significant personal and scientific growth since the start of my PhD, which would not have been possible without their unwavering support and mentorship. I am profoundly thankful for the opportunity to work alongside them, and their support has been invaluable. Pursuing this objective has been challenging, but I have no doubt that doing it with them has been one of the best decisions of my life. In the words of Alberto, "Si lavora tanto ma a volte anche ci divertiamo" (We work hard, but sometimes we also have fun), a sentence whose verbs change depending on the season.

I would also like to extend my gratitude to the other professors in the research group who have always been available to address my questions, share their perspectives, or simply enjoy a coffee together. Special thanks to Giovanni, who generously shared his knowledge and provided me with valuable advice.

Special acknowledgment goes to Geolog and Geotech for funding my PhD and research. I am indebted to my colleagues at these organizations, particularly Ivo, who served as a wonderful mentor, always offering support

and assistance with a smile. I also appreciate the camaraderie of my office mates, with special mentions to Roberto and Placido, who supported my research efforts, and to Luca, who was unfailingly friendly and inclusive in non-technical activities.

To my family, especially my mother, who has consistently supported and encouraged me to pursue my dreams, I owe an unpayable debt of gratitude. I would also like to express my appreciation to my sister, Maria Fernanda, and her husband, Juan Camilo, who not only supported me from afar but also took care of our family while I was away. My little nephews, who entered my life at the beginning and midpoint of my PhD, brought immense joy. My heart holds a special place for the rest of my family in Colombia, who provided unwavering support and encouragement throughout my journey. I miss them dearly. I know they pray for me and I am sure that their prayers have been heard. I love them and can only wish them all the best.

I extend my thanks to all my friends, old and new. I embarked on this journey with three close friends, each pursuing a PhD from different corners of the world. Laura, Pedro, and Edwin understand the challenges of this path, especially the difficulties of being far from home and family. They have always been a source of kind words and positive energy. To Jorge and Viviana, friends I left behind in Colombia and saw only a few times over four years, I am grateful for their unwavering support, no matter the distance.

I also want to express my appreciation to all the friends and colleagues I met during these years, with a special mention of Verónica, Chiara, Panagiotis, Andrea, and Juan, who have become integral parts of my daily life and have always been available to share a coffee, a meal, or good company. Additionally, I am thankful for all the other friends and

colleagues I met in Italy, who made my time here an enriching and memorable experience.

Abstract

In the context of energy production, hydrocarbons have consistently played a salient role in the global energy matrix. Current reports and policies suggest that their importance will persist in the short and medium term. Studying the processes involved in the fluid migration in underground environments, specifically in hydrocarbon extraction, is key to quantitatively assess the behavior of an oil and gas recovery system and possible interactions between anthropic activities and the environment during hydrocarbon production. This also offers the possibility for transferability of key methodological approaches towards other applications to quantify and predict better the effects of the use of underground resources on the environment.

Studying and modeling underground systems (and in general any natural system) is challenging due to a variety of reasons. Conceptualization of the system behavior is often hindered by the generally incomplete knowledge of domain properties and configuration. Models employed to predict system behavior are typically abstractions of the actual physical phenomena taking place in a domain. The associated model parameters are often estimated on the basis of indirect observations with potential scale and representativity limitations. Moreover, the intrinsic hetero-

geneity of natural systems adds remarkable layers of complexity to the modeling process.

This Doctoral Thesis focuses on the study of three critical aspects of underground energy resources within the framework of uncertainty quantification. The primary goal of this Thesis is to deepen our understanding of geosystems and optimize the efficiency of energy recovery projects.

First, the influence of parametric uncertainty is analyzed for a model representing gas (such as hydrogen, carbon dioxide, and methane) migration across low-permeability materials. The outcomes of this study allow us to identify the most influential model parameters as well as their relationships with the statistical moments of model results. The latter aspect is of critical importance for uncertainty quantification. Along these lines, the contribution of diverse flow mechanisms is quantified, and the probability density functions of effective permeability and diffusion coefficients are attained.

Then, a rigorous probabilistic risk assessment of groundwater contamination after hydraulic fracturing operations for oil production is conducted. The study analyzes two scenarios of contamination risk. Due to the system complexity and scale, it relies on surrogate modeling to quantify the migration of hydrocarbons and fracking fluids across the subsurface. The outcomes of this study enable one to quantify from a probabilistic point of view the risk of groundwater contamination that might arise as well as to identify the parameters and hydraulic fracturing settings that play the most important role in the migration of contaminant fluids after fracturing operations.

Finally, we examine fingerprinting for the allocation of individual oil sources during oil production. The latter is a technique that employs

the outcomes of gas chromatography experiments with the aim of identifying the source formation of hydrocarbon mixtures. Recent advances in experimental methods enable the acquisition of high-resolution gas chromatograms. Here, we review the methods for fingerprinting production allocation and propose a novel algorithm that (*i*) is capable to process high-resolution chromatograms, (*ii*) shows higher accuracy with respect to traditional approaches, and (*iii*) reduces the experimental burden of fingerprinting production allocation.

By addressing these three critical aspects of hydrocarbon production, this Doctoral Thesis contributes to enhancing the efficiency of hydrocarbon production processes and quantifying the potential environmental impacts of their extraction.

Keywords: Gas Migration, Hydraulic Fracturing, Production Allocation, Uncertainty, Sensitivity Analysis, Surrogate Modeling, System Identification, Hydrocarbons, Low Permeable Materials

Sommario

Nel contesto della produzione di energia, gli idrocarburi hanno costantemente svolto un ruolo di rilievo nella matrice energetica globale. Rapporti e politiche attuali suggeriscono che la loro importanza persistirà nel breve e medio termine. Lo studio dei processi coinvolti nella migrazione dei fluidi in ambienti sotterranei, in particolare nell'estrazione di idrocarburi, è fondamentale per valutare quantitativamente il comportamento di un sistema di recupero di petrolio e gas e le possibili interazioni tra le attività antropiche e l'ambiente durante la produzione di idrocarburi. Ciò offre anche la possibilità di trasferire approcci metodologici chiave verso altre applicazioni al fine di quantificare e prevedere meglio gli effetti dell'uso di risorse sotterranee sull'ambiente.

Lo studio e la modellazione dei sistemi sotterranei (e in generale di qualsiasi sistema naturale) sono sfidanti per una varietà di motivi. La concettualizzazione del comportamento del sistema è spesso ostacolata dalla conoscenza generalmente incompleta delle proprietà e della configurazione del dominio. I modelli utilizzati per prevedere il comportamento del sistema sono tipicamente astrazioni dei fenomeni fisici effettivi che si verificano in un dominio. I parametri del modello associati sono spesso stimati sulla base di osservazioni indirette con potenziali limitazioni di scala e rappresentatività. Inoltre, l'eterogeneità intrinseca dei sistemi naturali aggiunge strati notevoli di complessità al processo di modellazione.

Questa Tesi di Dottorato si concentra sullo studio di tre aspetti critici delle risorse energetiche sotterranee nel quadro della quantificazione dell'incertezza. L'obiettivo principale di questa Tesi è approfondire la nostra comprensione dei geosistemi e ottimizzare l'efficienza dei progetti di recupero energetico.

In primo luogo, si analizza l'influenza dell'incertezza parametrica per un modello che rappresenta la migrazione di gas (come idrogeno, biossido di carbonio e metano) attraverso materiali a bassa permeabilità. I risultati di questo studio ci consentono di identificare i parametri del modello più influenti e le loro relazioni con i momenti statistici dei risultati del modello. Quest'ultimo aspetto è di fondamentale importanza per la quantificazione dell'incertezza. In questa prospettiva, viene quantificato il contributo di diversi meccanismi di flusso e vengono ottenute le funzioni di densità di probabilità della permeabilità efficace e dei coefficienti di diffusione.

Successivamente, viene condotta un'attenta valutazione del rischio probabilistico di contaminazione delle acque sotterranee dopo operazioni di fratturazione idraulica per la produzione di petrolio. Lo studio analizza due scenari di rischio di contaminazione. A causa della complessità e della scala del sistema, si basa sulla modellazione sostitutiva per quantificare la migrazione di idrocarburi e fluidi di fratturazione nel sottosuolo. I risultati di questo studio consentono di quantificare da un punto di vista probabilistico il rischio di contaminazione delle acque sotterranee che potrebbe sorgere e di identificare i parametri e le impostazioni di fratturazione idraulica che svolgono il ruolo più importante nella migrazione dei fluidi contaminanti dopo le operazioni di fratturazione.

Infine, esaminiamo la tracciatura per l'allocazione delle singole fonti di petrolio durante la produzione di petrolio. Quest'ultima è una tecnica

che impiega i risultati di esperimenti di cromatografia gassosa allo scopo di identificare la formazione di origine delle miscele di idrocarburi. Gli avanzamenti recenti nei metodi sperimentali consentono l'acquisizione di cromatogrammi a alta risoluzione. Qui, esaminiamo i metodi per l'allocazione della produzione basata sulla tracciatura e proponiamo un nuovo algoritmo che (*i*) è in grado di elaborare cromatogrammi ad alta risoluzione, (*ii*) mostra una maggiore precisione rispetto agli approcci tradizionali e (*iii*) riduce l'onere sperimentale dell'allocazione della produzione basata sulla tracciatura.

Affrontando questi tre aspetti critici della produzione di idrocarburi, questa Tesi di Dottorato contribuisce a migliorare l'efficienza dei processi di produzione di idrocarburi e a quantificare gli impatti ambientali potenziali della loro estrazione.

Parole chiave: Migrazione del Gas, Fratturazione Idraulica, Allocazione della Produzione, Incertezza, Analisi di Sensibilità, Modellazione Surrogata, Identificazione del Sistema, Idrocarburi, Materiali a Bassa Permeabilità.

Resumen

En el contexto de la producción de energía, los hidrocarburos han desempeñado consistentemente un papel destacado en la matriz energética global. Los informes y políticas actuales sugieren que su importancia perdurará a corto y mediano plazo. El estudio de los procesos involucrados en la migración de fluidos en entornos subterráneos, específicamente en la extracción de hidrocarburos, es fundamental para evaluar cuantitativamente el comportamiento de un sistema de recuperación de petróleo y gas, así como las posibles interacciones entre las actividades humanas y el medio ambiente durante la producción de hidrocarburos. Esto también ofrece la posibilidad de transferir enfoques metodológicos clave a otras aplicaciones para cuantificar y predecir de manera más precisa los efectos del uso de recursos subterráneos en el medio ambiente.

Estudiar y modelar sistemas subterráneos (y en general cualquier sistema natural) es un desafío debido a una variedad de razones. La conceptualización del comportamiento del sistema a menudo se ve obstaculizada por el conocimiento generalmente incompleto de las propiedades y configuración del dominio. Los modelos empleados para predecir el comportamiento del sistema suelen ser abstracciones de los fenómenos físicos reales que tienen lugar en un dominio. Los parámetros del modelo asociados a menudo se estiman sobre la base de observaciones indirectas con posibles limitaciones de escala y representatividad. Además, la hetero-

geneidad intrínseca de los sistemas naturales agrega capas notables de complejidad al proceso de modelado.

Esta Tesis Doctoral se centra en el estudio de tres aspectos críticos de los recursos energéticos subterráneos en el marco de la cuantificación de la incertidumbre. El objetivo principal de esta Tesis es profundizar nuestra comprensión de los geosistemas y optimizar la eficiencia de los proyectos de recuperación de energía.

En primer lugar, se analiza la influencia de la incertidumbre paramétrica en un modelo que representa la migración de gases (como hidrógeno, dióxido de carbono y metano) a través de materiales de baja permeabilidad. Los resultados de este estudio nos permiten identificar los parámetros del modelo más influyentes, así como sus relaciones con los momentos estadísticos de los resultados del modelo. Este último aspecto es de importancia crítica para la cuantificación de la incertidumbre. En esta línea, se cuantifica la contribución de diversos mecanismos de flujo y se obtienen las funciones de densidad de probabilidad de la permeabilidad efectiva y los coeficientes de difusión.

Luego, se lleva a cabo una rigurosa evaluación de riesgos probabilísticos de contaminación de aguas subterráneas después de operaciones de fracturación hidráulica para la producción de petróleo. El estudio analiza dos escenarios de riesgo de contaminación. Debido a la complejidad y escala del sistema, se basa en la modelización sustitutiva para cuantificar la migración de hidrocarburos y fluidos de fracturación en el subsuelo. Los resultados de este estudio permiten cuantificar desde un punto de vista probabilístico el riesgo de contaminación de las aguas subterráneas que podría surgir, así como identificar los parámetros y las configuraciones de fracturación hidráulica que desempeñan el papel más importante en la migración de los fluidos contaminantes después de las operaciones de

fracturación.

Finalmente, examinamos la huella digital para la asignación de fuentes individuales de petróleo durante la producción de petróleo. Esta última es una técnica que emplea los resultados de experimentos de cromatografía de gases con el objetivo de identificar la formación de origen de las mezclas de hidrocarburos. Los avances recientes en los métodos experimentales permiten la adquisición de cromatogramas de gases de alta resolución. Aquí, revisamos los métodos para la asignación de producción basada en huella digital y proponemos un nuevo algoritmo que (*i*) es capaz de procesar cromatogramas de alta resolución, (*ii*) muestra una mayor precisión con respecto a enfoques tradicionales y (*iii*) reduce la carga experimental de la asignación de producción basada en huella digital.

Abordando estos tres aspectos críticos de la producción de hidrocarburos, esta Tesis Doctoral contribuye a mejorar la eficiencia de los procesos de producción de hidrocarburos y a cuantificar los posibles impactos ambientales de su extracción.

Palabras clave: Migración de gas, fracturamiento hidráulico, asignación de producción, incertidumbre, análisis de sensibilidad, modelización de sustitución, identificación de sistemas, hidrocarburos, materiales de baja permeabilidad.

Contents

1 | Introduction

1.1. Background and Motivation

For decades, the global energy demand has been dominated by fossil fuels, these sources accounting for around 80% of the overall share. By 2020, natural gas, oil, and coal accounted for 24%, 30%, and 26% of the global energy demand respectively [IEA, 2022].

The preponderance of these fuels can be attributed to several salient factors, as outlined by Vassiliou [2009]. Firstly, fossil fuels are highly energy-dense and represent a highly efficient and cost-effective source of energy. Secondly, they are versatile and can be used across a wide range of applications, encompassing electricity generation, transportation, heating, and industrial processes. Thirdly, fossil fuel reserves are widely distributed across the world, this element contributing to make them a readily accessible energy source. Then, they are highly reliable energy sources that can provide a consistent supply of energy with only moderate variability. Finally, the production and consumption of fossil fuels can yield significant economic benefits and promote energy independence for Countries that produce and export these fuels [Arthur et al., 2009].

Despite their marked importance, there are growing concerns about the

environmental impact of a continuous use of fossil fuels as energy sources. Combustion of these fuels is associated with release of greenhouse gases into the atmosphere, with a potentially significant contribution to climate changes [Ritchie et al., 2020].

During the United Nations Climate Change Conference in Paris in 2015, an international treaty aimed at minimizing climate change effects and reducing greenhouse gas emissions was adopted by 195 countries. The primary objective of the Paris Agreement is to restrict the escalation of global warming to a level lower than 2°C above pre-industrial levels, and to pursue efforts to limit it to 1.5°C. This is considered a critical threshold beyond which the impacts of climate change become much more severe, including more frequent and intense heatwaves, droughts, storms, and sea level rise [Savaresi, 2016]. In light of this accord, several nations have already announced a series of policies aimed at mitigating the effects of climate change. Furthermore, emerging technologies play a key role in the reduction of fossil fuel consumption, which is an indispensable component in minimizing the adverse consequences of climate change.

The 2022 World Energy Outlook report, which has been published by the International Energy Agency, presents an analysis of three distinct scenarios for energy production: the Stated Policies Scenario (STEPS), the Announced Pledges Scenario (APS), and the Net Zero Emission Scenario (NZE). The STEPS scenario envisions the trajectory that is implied by current policy settings, whereas the APS scenario assumes that all ambitious targets announced by governments are achieved on time and in full (including their long-term net-zero and energy access goals). The NZE scenario outlines a pathway aimed at achieving a 1.5°C stabilization in the global temperature rise, while simultaneously ensuring universal access to modern energy by 2030. Remarkably, in this latest edition, the

STEPS scenario foresees a peak (or plateau) in the share of fossil fuels in the global energy demand for the first time in the history of these reports (See Figure 1.1). According to this scenario, the contribution of fossil fuels is expected to decrease to 75% by 2030 and further decline to 60% by the end of 2050 [IEA, 2022].

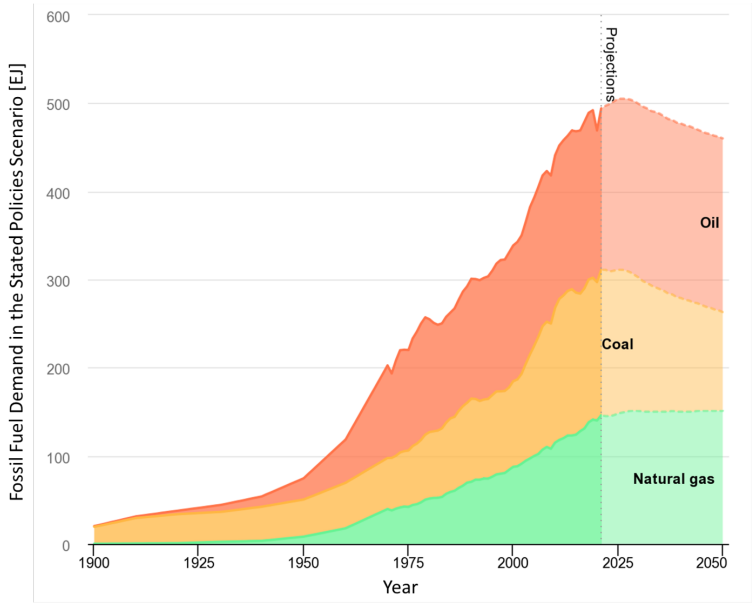

Figure 1.1: General context of hydrocarbon generation and production processes [IEA, 2022].

Although the APS Scenario envisions a lower share of fossil fuels, it assumes that all announced targets are accomplished on time and in full. Otherwise, in practice, the implementation of such targets is frequently hindered by various factors. A conservative attitude would then correspond to anticipate that fossil fuels will continue to play a key role in

the global energy production in the near future. Therefore, it is critical to investigate the recovery and transformation processes of fossil fuels during energy production to enhance their efficiency and mitigate their environmental impacts.

1.2. Hydrocarbons Generation and Production

Genesis of hydrocarbons, such as oil and natural gas, is a complex process that takes place across geological time scales spanning millions of years. It begins with the accumulation and burial of organic matter, such as plankton and algae, in sedimentary rocks. Over time, the organic matter undergoes a series of chemical and physical transformations, induced by heat and pressure. Organic matter is initially converted into a waxy substance, known as kerogen. As temperature and pressure continue to increase, the kerogen is then subject to catagenesis, leading to the formation of liquid and gaseous hydrocarbons. Hydrocarbons generated during catagenesis can migrate through the surrounding rock until they are trapped in a reservoir that is situated below a low-permeability rock layer, commonly referred to as caprock. Specific conditions required for hydrocarbon generation and accumulation depend on a variety of factors, including the quality and quantity of organic matter, temperature, pressure, and geological structure [Dembicki-Jr., 2017].

Retrieval of hydrocarbons from a reservoir is typically accomplished through a production process (See Figure 1.2). The latter begins with drilling of a well to penetrate a hydrocarbon-bearing formation. During the drilling operation, the well is encased and cemented to prevent exposure of the surrounding rock layers to the reservoir fluids and collapse of the rock formations into the wellbore [Yan et al., 2020]. Following completion of the drilling, casing, and cementing procedures, the well is equipped with production devices (such as, e.g., a wellhead, tubing, and packers), to make it operational. Hydrocarbon fluids are then extracted from the reservoir and transported to the surface. Once on the surface,

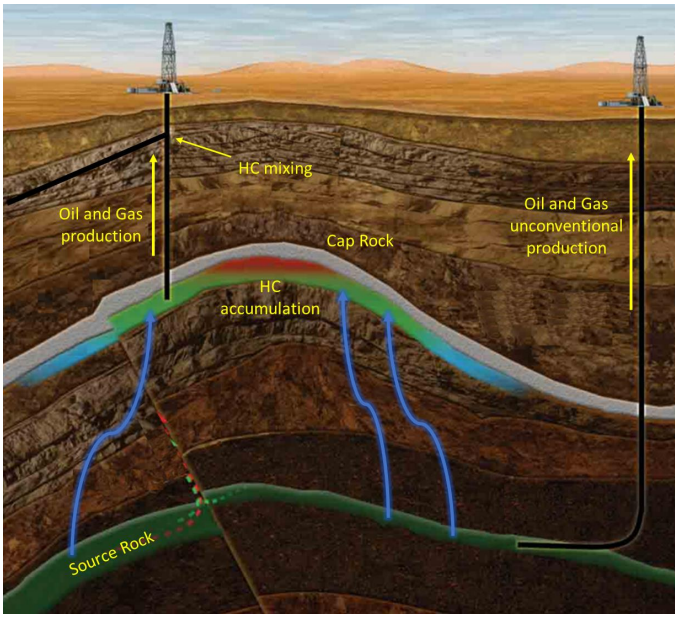

Figure 1.2: General context of the hydrocarbon generation and production processes. Image adapted from SLB (2023).

the hydrocarbon phases are separated using ad-hoc equipment.

Unconventional production methods such as hydraulic fracturing have emerged in recent years to recover hydrocarbons from formations where traditional recovery is technically and/or economically unfeasible. Hydraulic fracturing is employed to extract hydrocarbons directly from the source rock. The process relies on fracturing the host rock, thus leading to increase its permeability and to fast migration of the hydrocarbons [Smith and Montgomery, 2015]. In addition to the associated economic benefits, this technique is particularly valuable in regions in which con-

ventional production is unfeasible or where coal is the only available type of fossil fuel available [Arthur et al., 2009]. As an additional benefit, it is associated with a potential reduction of greenhouse gas emissions since hydrocarbons are associated with lower emissions than coal [Safari et al., 2019].

A common practice in the production process is hydrocarbon mixing. It is associated with the benefit of cost reduction as a result of sharing facilities and production equipment. Crude hydrocarbons originated from various reservoirs, wells, and/or fields are mixed and jointly produced through commingling operations. For technical, fiscal, economic, and management reasons, it might be necessary to assess the individual contribution of a hydrocarbon type to the overall production. This process is known as production allocation and has been subject to renewed interest in recent years [Carati et al., 2020; Murray and Peters, 2021; Patience et al., 2021; van Bergen and Gordon, 2020; Yang et al., 2019].

In this context, a part of the dissertation aims at analyzing processes associated with hydrocarbons generation and production with the aim of (*i*) enhancing our knowledge on the physics governing flow of hydrocarbons across geomaterials, (*ii*) quantifying side effects of unconventional oil production on groundwater bodies, and (*iii*) increasing the efficiency of production-associated processes in oil recovery projects.

1.2.1. Gas Movement in Low-Permeable Materials

Natural gas is mainly composed of methane (CH_4), which is a simple hydrocarbon molecule made up of one carbon atom and four hydrogen atoms. Natural gas can also contain small amounts of other hydrocarbons, including ethane (C_2H_6), propane (C_3H_8), butane (C_4H_10), and pentane (C_5H_12), as well as non-hydrocarbon gases such as carbon dioxide (CO_2), nitrogen (N_2), and hydrogen sulfide (H_2S). Considerable reserves of natural gas are associated with subsurface reservoirs worldwide [U.S. Energy Information Administration, 2015]. Natural gas has the potential to assist the transition to a carbon-free energy landscape as it is associated with lower emissions of greenhouse gases in comparison with coal [Hughes, 2013; Safari et al., 2019]. Gas typically accumulates in reservoir regions subdued to low permeability layers that prevent its upward migration [Dembicki-Jr., 2017]. Due to the partial sealing efficiency of caprocks, some amount of gaseous phase hydrocarbons might cross such barrier. Reservoir gas can then be released into the overburden to (eventually) reach the surface [Schloemer and Krooss, 2004; Schlömer and Krooss, 1997]. In this context, appropriate modeling approaches to quantify gas migration across low-permeability geomaterials can assist in the appraisal of the feasibility of a natural gas recovery project. However, the applications are not limited to energy production but can also be extended to energy storage applications such as the injection of hydrogen or methane in subsurface reservoirs or the capture and storage of greenhouse gasses such as carbon dioxide.

Low permeable caprock formations are frequently associated with mudrocks. Depending on their fissility (i.e., the tendency of a rock to split along flat planes), these can be classified as mudstones (no fissility) or shales (show fissility) [Folk, 1980]. These materials are associated with

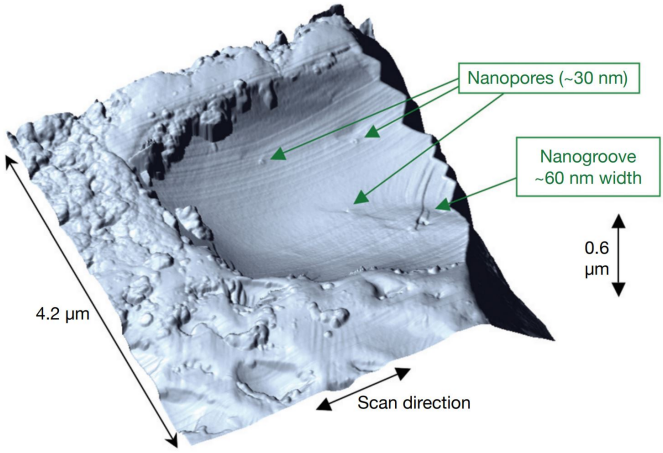

Figure 1.3: Atomic Force Microscope image of nanopores in a low permeability geomaterial. Image taken from Javadpour [2009].

very fine grains that create pores of very small size and typically exhibit a permeability range between 10^{-20} and 10^{-14} m^2. Javadpour [2009] observed the structure of mudrocks employing Atomic Force Microscopy (See figure 1.3) and found that the smallest pores in mudrocks are of the order of nanometers. At such a scale, the continuum hypothesis breaks down and flux of fluids cannot be represented by Darcy's law. Thus, new models that can represent the movement of fluids at such and small scale must be proposed and validated [Javadpour et al., 2021].

A variety of models depicting gas movement in low permeability geomaterials have been recently proposed [Rani et al., 2018; Sun et al., 2017; Wang et al., 2019; Wu et al., 2016]. These models typically estimate the mass flow rate of gas as the result of a combination of various gas transport mechanisms taking place across the porous system. Parameters

associated with these models, describing chemical, mechanical, flow, and transport features governing feedbacks between gas and the host rock matrix are always affected by uncertainty. In addition, these models typically embed numerous parameters which are typically estimated through (direct or indirect) laboratory-scale experiments. Considering the set of complex mechanisms involved, these types of experiments are costly, time demanding, and their results are prone to uncertainty. The latter is also related to the intrinsic difficulties linked to replicating operational field conditions at the laboratory scale as well as to the challenges stemming from the transferability of results to heterogeneous field scale settings [Pan et al., 2010; Tan et al., 2018; Yuan et al., 2014].

Due to our still incomplete knowledge of the critical mechanisms driving gas movement in low permeability media [Javadpour et al., 2021; Singh and Myong, 2018] and the complexities associated with the estimation of model parameters, model outputs should be carefully analyzed considering all possible (aleatoric and epistemic) sources of uncertainty. In this sense, sensitivity analysis approaches are important tools enabling us to (*i*) quantify uncertainty, (*ii*) enhance our understanding of the relationships between model inputs and outputs, and (*iii*) tackle the challenges of model- and data-driven design of experiments [Dell'Oca et al., 2017]. Hence, sensitivity analysis techniques may be effectively used in the context of gas flow modeling efforts to (*i*) quantify and rank the contribution of our lack of knowledge on model parameters to the uncertainty associated with model outputs; (*ii*) identify model input-output relationships; and (*iii*) enhance the quality of parameter estimation workflows, upon focusing efforts on parameters with the highest influence to target model outputs [Dell'Oca et al., 2020; Saltelli et al., 2010]. In cases where parameters associated with a model have already been estimated (e.g., through model calibration), the main purpose of a Global Sensitivity Analysis

(GSA) is to assist quantification of the uncertainty still remaining after model calibration, thus guiding additional efforts for its characterization (e.g., Dell'Oca et al. [2020] and references therein). The probability density function related to each model parameter at this step might differ from the one employed before model calibration and some model parameters might be associated with a reduced uncertainty.

In this context, one objective of this Doctoral Thesis is to answer the following research question:

> Considering the complex physics governing the migration of gases in low-permeable materials and the intrinsic uncertainty associated with the estimation of the parameters employed by the models representing such a process, which parameters control migration of gas in low-permeable materials and in which way do these parameters affect the output of a given gas migration model?

1.2.2. Impacts of unconventional hydrocarbons production on groundwater bodies

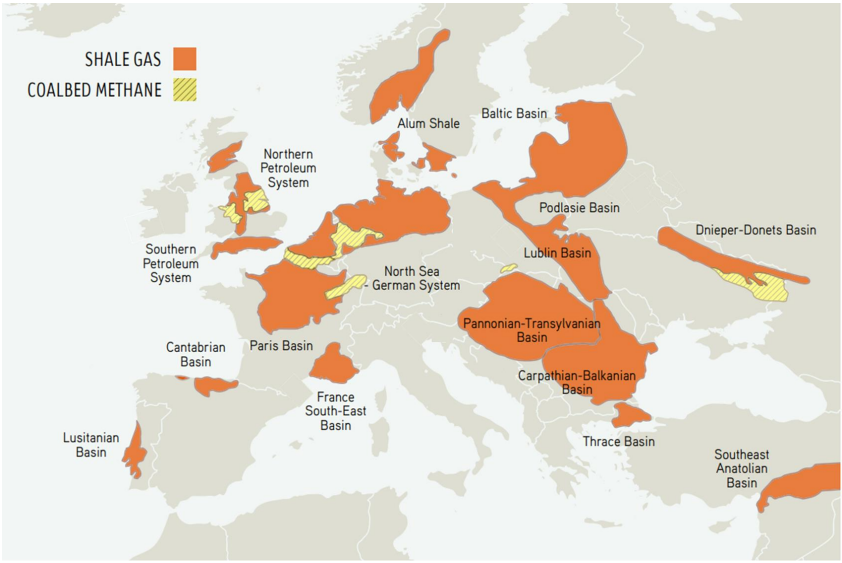

Figure 1.4: Shale plays in Europe. Image adapted from Eberhardt et al. [2013].

Unconventional production of hydrocarbons from shale rocks is a key research topic, particularly for Europe, where conventional gas production is in a declining stage [International Energy Agency, 2014]. The European Commission estimates a considerable potential for shale rocks in Europe, with approximately 16 trillion cubic meters of technically recoverable shale gas [European Comission, 2014]. The combined area of the seven primary shale plays in Europe (i.e., Paris, Northwest German, Baltic, Lublin, Bowland, Carpathian-Balkan, and Pannonian-Transylvanian plays) exceeds 90,000 km^2 (See Figure 1.4). Typically,

the hydrocarbon-bearing formations of these plays are 2,000-4,000 meters deep on average, with an associated thickness and porosity ranging between 150 and 200 meters and 4 and 10%, respectively [Edlmann and McDermott, 2016].

Hydraulic fracturing can be considered as the most popular unconventional recovery method used to extract hydrocarbons from geologic plays with low permeability - particularly shales [Arthur et al., 2009; Britt, 2012; Smith and Montgomery, 2015; Wu, Chen and Li, 2015]. The technique involves drilling a well vertically to reach a hydrocarbon-bearing formation and subsequently altering the orientation of drilling to attain a horizontal configuration within the source rock layer. Upon completion of the well, small perforations are made at discrete positions along the horizontal section of the well. These allow for the injection of a blend of water, proppant agent, and chemicals with the aim of fracturing the rock. The proppant agent, which may consist of sand or other inert solids, prevents fractures from closing after the pressure in the system has been released, whereas the co- injected chemicals serve a multitude of functions in the hydraulic fracturing process. These include, e.g., inhibiting bacterial growth, facilitating pumping of proppant down-hole and into the fractured formation, and minimizing mineral scaling of the well [Stringfellow et al., 2014]. Once fracturing and injection operations have been completed, the mixture of water and chemicals is extracted from the well and treated on the surface [Jabbari et al., 2017]. The entire length to be fractured is typically divided into smaller segments. Fracturing is then conducted in each segment in a different stage. By doing so, fractures can be evenly distributed throughout the domain. Additionally, fracturing in stages allows for the use of different proppants or chemical doses and types depending on the properties of the various zones within the hydrocarbon-bearing formation.

Due to its relatively recent development and lack of firm regulations in many Countries, there is still limited understanding of potential adverse effects of hydraulic fracturing [Howarth et al., 2011]. Of particular concern are possible risks of water contamination [Osborn et al., 2011; Vengosh et al., 2014], air pollution [Moore et al., 2014], and induced seismicity [De Pater and Baisch, 2011; Walker et al., 2014]. Specifically, water pollution may arise as a consequence of (*i*) improper handling or treatment of water used to fracture the source rock once it has been extracted from the well [Glazer et al., 2014] and (*ii*) migration of the injected water or source rock hydrocarbons from the fractured area towards subsurface freshwater bodies [Jabbari et al., 2017]. Thus, conducting rigorous assessments of potential contamination risks is critical to ensure the secure exploitation of shale rock hydrocarbons [Howarth et al., 2011]. This process enables the prevention and mitigation of environmental impacts while enhancing public awareness and comprehension of potential risks.

In this context, one objective of this Doctoral Thesis is to answer the following research question:

> In view of concerns associated with water contamination from hydraulic fracturing operations for hydrocarbon production, how can one provide a robust assessment of the risk of groundwater contamination considering the uncertainties associated with the characterization of the parameters employed in the conceptual models that describe flow of hydrocarbons across the subsurface?

1.2.3. Production allocation to optimize hydrocarbons production

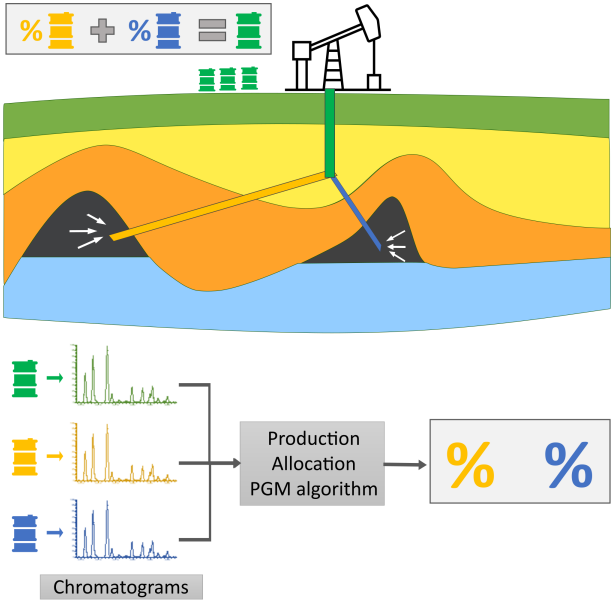

Figure 1.5: Schematic representation of production allocation

Assessment of the spatial and temporal chemical evolution of fluids in oil and gas systems is a key element of modern reservoir engineering and use of underground energy resources. It provides strong support to the planning and implementation of effective and sustainable strategies for reservoir development and production [England, 2007]. In this context, interpretations arising from geochemical analyses can have significant impacts on cost and production scenarios. Advancements in geochemical tools and related analytical methodologies enable further constraining

of uncertainties associated with reservoir characterization [Cubitt et al., 2004]. Within this broad framework, geochemical fingerprinting is nowadays considered a robust technique in reservoir geochemistry applications [Koolen et al., 2018; Permanyer et al., 2002; Rudyk et al., 2013; van Bergen and Gordon, 2020; Yang et al., 2016]. One of the most widely known applications of geochemical fingerprinting is production allocation, which entails determining the relative proportions of distinct components within an oil mixture obtained through the shared utilization of oil production facilities, whether located on the surface or underground.

Geochemical analyses of reservoir fluids are routinely adopted to support our conceptual understanding of fluid connectivity. The latter represents a major aspect in reservoir modeling and management and can markedly impact oil and gas production [Ekpo et al., 2018; Milkov et al., 2007]. Additionally, geochemical data can be used for dynamic reservoir performance assessment and evaluation of compositionally graded fluid column depletion [Chuparova et al., 2010]. The adoption of such techniques yields robust results, which are overall consistent with findings obtained with other classical methods (such as, e.g., pressure tests or wireline logs for compartmentalization assessment, as well as production logging tools and flowmeters for production monitoring) and are characterized by marked advantages in terms of cost savings and practical convenience [Elsinger et al., 2010; Kanshio, 2020; McCaffrey et al., 2012].

Conventional production allocation techniques [Hwang et al., 2000; Kaufman et al., 1987; Peters et al., 2008] make use of the molecular differences between individual End Members, EMs, i.e., fluids belonging to a distinct region in the system. In production allocation scenarios, commingled hydrocarbons and EMs are typically analyzed through geochemical techniques (e.g., gas chromatography, GC). The ensuing data (chro-

matograms, also termed GC fingerprints) are then processed upon relying on deconvolution algorithms to evaluate the contribution of each EM to the commingled produced oil. Recently, advancements in gas chromatography, have allowed measuring chromatograms of improved quality. New chromatograms (*i*) provide the ability to identify more discriminating features in case of highly similar hydrocarbons and (*ii*) observe molecules associated with higher boiling points which allows to compensate for poor sampling practices or improper storage conditions. Given the current advancements in experimental methods, it is desirable to improve also the available deconvolution techniques to employ algorithms that use in the most efficient way the information content provided by the experimental chromatograms.

In this context, one objective of this Doctoral Thesis is to answer the following research question:

> Given the recent advancements in the experimental methods for gas chromatography, which deconvolution algorithm can be employed to maximize the accuracy of production allocation estimates?

1.3. Thesis structure

This Thesis is organized into three Chapters. Each of these includes a brief introduction of the main chapter topics, a methodology section in which the details of the techniques employed in each section are explained, a Results and Discussion section in which the results of each study are documented and the most important results are discussed and analyzed, and conclusions, where the most important findings of each study are reported.

2 | The Role of Multiple Transport Mechanisms and Parameters in Gas Flow Models for Low Permeability Systems

2.1. Abstract

Recent models represent gas (methane) migration in low permeability media as a weighted sum of various contributions, each associated with a given flow regime. These models involve numerous chemical and physical parameters that are difficult to assess experimentally. In this context, modern sensitivity analysis techniques help us understand how uncertainties in model inputs affect the model's output. This study employs two global sensitivity analysis methods to evaluate a recent interpretive model that breaks down gas migration into surface diffusion and two weighted bulk flow components. It quantitatively explores the influence of uncertain model parameters and their associated probability distributions on methane flow assessment. Then, the structure of an effective diffusion coefficient embedding all complex mechanisms of the model considered is derived, such formulation enables measuring the contribution of each

flow mechanism to the overall gas flow[1].

2.2. Introduction

In this work, we illustrate the methodological framework and the workflow required for Global Sensitivity Analysis (GSA) of a gas flow model on low permeability geomaterials and provide the elements to perform such an analysis for diverse scenarios. We use an exemplary model, the conceptual model proposed by Wu et al. [2016], which depicts the mass flow rate of a gas across a low permeability medium as the sum of three key processes: (*i*) a surface diffusion, and two weighted bulk diffusion components corresponding to (*ii*) slip flow and (*iii*) Knudsen diffusion. In addition, this model takes into account changes in the porous system caused by mechanical deformation and adsorption/desorption dynamics.

In this thesis chapter, we rely on GSA approaches to study the behavior of the aforementioned gas migration model targeting low permeability media. While previous works focus on only a few selected model parameters [Song et al., 2016; Sun et al., 2017; Wu et al., 2017], a comprehensive diagnosis of the system behavior based on rigorous and modern GSA approaches taking into account the way all model parameters influence model output uncertainty is still missing. Here, we do so by implementing two GSA techniques, respectively based on the evaluation of (*i*) the classical (variance-based) Sobol' indices [Saltelli and Sobol', 1995] and (*ii*) the recent moment-based GSA metrics proposed by Dell'Oca et al. [2017]. We recall that GSA approaches relying on Sobol' indices are widely used to quantify the relative expected reduction of variance of the target model output due to the knowledge of (or conditioning on)

[1] Results of this study were published in a research article which can be consulted at https://doi.org/10.1007/s11242-022-01755-x.

a given parameter. These have been employed in several applications, including diagnosis of models related to, e.g., flood risk assessment [Koks et al., 2015], overpressure risk assessment in sedimentary basins [Colombo et al., 2017], and energy storage [Xiao et al., 2021]. A critical limitation of variance-based GSA methodologies is that the uncertainty of the output is considered to be completely characterized by its variance. Such an assumption can lead to an incomplete characterization of the system behavior. The moment-based GSA approach introduced by Dell'Oca et al. [2017] is designed to enhance our capability to evidence model behavior upon including the quantification of model parameter uncertainty on the (statistical) moments of the pdf of a model output of interest. As such, this comprehensive approach yields information enabling us to characterize various aspects of uncertainty, without being limited solely to the concept of variance. The ensuing indices (termed AMA indices, after the initials of the authors [Dell'Oca et al., 2017]) have been effectively employed in a variety of contexts, including geophysical analyses related to gravimetric responses due to pumping tests [Maina et al., 2021], biochemical degradation of compounds such as gliphosate in soils [la Cecilia et al., 2020], and groundwater flow, including its feedbacks with evapotranspiration [Bianchi Janetti et al., 2019; Maina and Siirila-Woodburn, 2020].

This chapter is organized as follows: Section 2.3.1 briefly illustrates the complete model we consider to describe methane flow in low permeability media. The main theoretical elements of the GSA approaches employed are described in Section 2.3.2. Key results of the GSA are presented in Section 2.4, where we also assess the relative contribution of diverse gas migration mechanisms to the overall flow. In addition, we derive and discuss novel formulations describing an effective diffusive behavior and encapsulating all physical-chemical mechanism included in the full

methane flow model described in Section 2.3.1. Finally, conclusions are drawn in Section 2.5.

2.3. Methodology

2.3.1. Gas Flow in Low Permeability Media

Models adopted to quantify gas migration in low permeability media can be classified according to their complexity, in terms of, e.g., conceptualization and mathematical rendering of the embedded processes, as well as the number of their characteristic parameters. Among existing models associated with a high degree of complexity and including multiple transport processes jointly contributing to the total gas migration across the system [Javadpour et al., 2021; Mehmani et al., 2013; Sun et al., 2017; Wu, Chen and Li, 2015; Wu et al., 2016, 2017; Zhang et al., 2018], here we consider the model of Wu et al. [2016]. The selected model allows considering mechanical deformation as well as relevant features associated with real gases such as variations in the gas viscosity (η), and the effects of gas compressibility (C_g) and deviation (Z) factors caused by pressure and temperature changes.

The model introduced by Wu et al. [2016] rests on a conceptual picture according to which the total mass flow rate of gas per unit of area (J) is rendered through the sum of (i) a surface diffusion (J_s) and two weighted bulk diffusion components, corresponding to (ii) slip flow (J_v), and (iii) Knudsen diffusion (J_k) (See Figure 2.1), i.e.,

$$J = J_s + w_v J_v + w_k J_k. \tag{2.1}$$

The surface diffusion component aims to model the migration of molecules adsorbed to the pore surface, such migration driven by a chemical potential gradient [Krishna and Wesselingh, 1997]. In the case of ideal gases (or quasi-ideal gases), this gradient can be approximated as a linear function

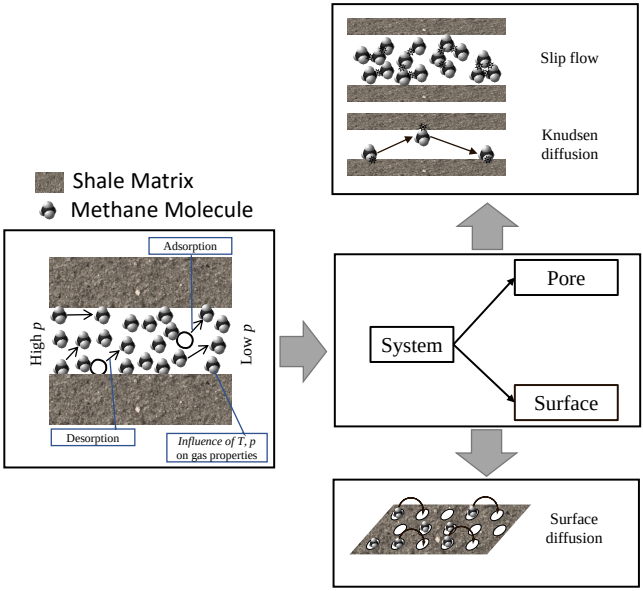

Figure 2.1: Conceptual picture of the model proposed by Wu et al. [2016], including three flow mechanisms. Two mechanisms take place in the pore space of the geomaterial (i.e., Slip flow and Knudsen diffusion) and one at the surface of the grains of the geomaterial (i.e., Surface Diffusion).

of the gas pore pressure gradient [Wu, Li, Wang, Yu and Chen, 2015]. Therefore, surface diffusion can be expressed as

$$J_s = -\zeta_{ms} \frac{D_s C_{sc}}{p} \frac{\partial p}{\partial l}, \tag{2.2}$$

where p is (gas) pore pressure and $\frac{\partial p}{\partial l}$ represents the strength of the driving force through the system. The (dimensionless) coefficient ζ_{ms} is intended to take into account the possibility of applying the model (originally developed for capillary tubes) to a complex pore space and is defined in Equation (A.1) of the Appendix, where it is shown that ζ_{ms} depends on porosity (ϕ), tortuosity (τ), pore radius (r), and gas coverage on the geomaterial (θ) expressed as a fraction of the total coverage area. The term D_s in Equation (2.2) is the surface diffusion coefficient, which is expressed (as shown in Equation (A.9)) in terms of gas temperature (T), isosteric adsorption heat of the geomaterial (ΔH), a parameter (κ) related to the blockage/migration ratio of the adsorbed molecules, and θ. Finally, C_{sc}, defined in Equation (A.12), is the adsorbed concentration, which in turn depends on θ and on the gas molecule diameter (d_m). Note that Equation 2.2 assumes that the rate of surface diffusion is significantly lower than the rate of gas adsorption and desorption, thereby establishing a dynamic equilibrium between the gas in the bulk phase and the adsorption phase within nanopores.

The model proposed by Wu et al. [2016] allows representing the mechanical deformation of the pore space (in terms of variation of permeability and porosity with pressure) through power-law relationships and making use of the classical Kozeny-Carman equation. Here, we rest on their original model formulation, which naturally leads to Equations (A.3) and (A.4), clearly evidencing that both r and ϕ evolve with p as a function of a reference pore radius (r_o) and reference porosity (ϕ_o), respectively.

The weight coefficients of the slip flow (w_v) and Knudsen diffusion (w_k) components in Equation (2.1) are given by [Wu et al., 2016]

$$w_v = \frac{1}{1 + K_n}, \tag{2.3}$$

$$w_k = \frac{1}{1 + 1/K_n}. \tag{2.4}$$

Here, K_n is the (dimensionless) Knudsen number defined as

$$K_n = \frac{\lambda}{2r}, \tag{2.5}$$

with

$$\lambda = \frac{\eta}{p}\sqrt{\frac{\pi ZRT}{2M}}, \tag{2.6}$$

where M and R are the gas molar mass and universal constant, respectively. Note that K_n relates the mean free path of the gas molecules (λ) to a representative length of the system [Civan, 2010], here taken as the pore diameter.

In cases where $K_n < 0.0001$, it is reasonable to assume that, within a system where gas flows through nanopores, the velocity of gas molecules at the pore wall becomes negligible, and Darcy's law holds true. This transport mechanism is commonly referred to as continuum flow. Conversely, when $0.001 < K_n < 10$, the assumption of zero gas molecule velocity at the pore surface is no longer valid. Consequently, the gas flux increases and Darcy's law loses its applicability, leading to what is known as rarefied gas transport [Karniadakis et al., 2005]. Rarefied gas transport can be further divided into slip flow, transition flow, and Knudsen diffusion, with the latter characterized by velocity profiles exhibiting constant values.

The slip flow component, which is dominant in systems where $0.0001 < K_n < 0.1$ [Ziarani and Aguilera, 2012], can be evaluated as [Karniadakis et al., 2005; Wu et al., 2016]

$$J_v = -\zeta_{mb} \frac{r^2 pM}{8\eta ZRT}(1 + \alpha K_n)\left(1 + \frac{4K_n}{1 + K_n}\right)\frac{\partial p}{\partial l}. \tag{2.7}$$

Here, ζ_{mb} is intended to take into account the possibility of applying the slip flow formulation (2.7) to a complex pore space (see Equation. (A.7)) and α is the rarified effect coefficient for a real gas which, according to Karniadakis et al. [2005], is evaluated through Equation (A.8).

The Knudsen flow component is dominant in systems where $K_n > 10$ [Ziarani and Aguilera, 2012] and is evaluated as [Darabi et al., 2012; Liu et al., 2016]

$$J_k = -\frac{2}{3}\zeta_{mb} r \delta^{D_f - 2}\left(\frac{8ZM}{\pi RT}\right)^{1/2}\frac{p}{Z} C_g \frac{\partial p}{\partial l}. \tag{2.8}$$

Here, D_f represents the fractal dimension of the pore surface and δ denotes the ratio between d_m and r.

In this study, we conduct model evaluations by randomly selecting values for all 15 model parameters from their respective probability density functions. These values are then used to calculate Equations 2.7, 2.8, and 2.2. Subsequently, we evaluate Equation 2.1 while taking into consideration the weight coefficients defined in Equations 2.3 and 2.4. It is worth noting that the outcomes of the sensitivity analysis are unaffected by the specific value of the gas pressure gradient, which is set at 0.1 Pa/m for all test cases in this study.

We conclude by noting that the model here described includes a total

of 15 parameters, which are related to the variety of physical processes embedded therein (i.e., Slip flow, Knudsen diffusion, Surface diffusion, mechanical deformation, effects of real gases, adsorption/desorption dynamics, see also Section 2.4.4). All quantities here introduced are listed in Table 1 and in the list of symbols and nomenclature Section.

2.3.2. Global Sensitivity Analysis

We perform a rigorous sensitivity analysis of the model illustrated in Section 2.3.1 to diagnose its behavior with reference to the estimate of methane flow as driven by imperfect knowledge of the associated parameters. Here, we note that uncertainties associated with the selection of the interpretative model is not analyzed. GSA can also be tailored to consider quantification of uncertainty of model outputs in the presence of multiple interpretive models. In this context, uncertainty of a target variable which might result from the use of a collection of interpretive (conceptual and mathematical) models could be assessed upon relying, for example, on the approach illustrated by Dell'Oca et al. [2020]. Our analysis is intended to yield a robust quantification of the relative importance of uncertain model parameters to a model output of interest. As mentioned in the Introduction, we rely on two GSA approaches, corresponding to (*i*) the classical variance-based technique grounded on the evaluation of the well-known Sobol' indices [Saltelli and Sobol', 1995] and (*ii*) the moment-based GSA framework introduced by Dell'Oca et al. [2017].

Model parameters are treated as statistically independent, as the amount of available information does not enable us to clearly identify cross-correlations amongst parameters and to quantify joint distributions. We consider three differing characterizations of pdf describing uncertainty of

model parameters: (*a*) all parameters are represented through uniform pdfs, (*b*) all parameters are represented by truncated Gaussian pdfs, and (*c*) the reference pore radius is characterized by a truncated log-normal pdf, while all remaining parameters are associated with uniform pdfs. Case *a* is representative of an approach where information on the considered parameters is limited so that all parameter values within the identified range of variability are equally weighted in the analysis (other studies relying on the same assumption include, e.g., Bianchi Janetti et al. [2019]; Ciriello et al. [2013]; Dell'Oca et al. [2020]; Laloy et al. [2013]; Sochala and Le Maître [2013]). Case *b* is implemented as an alternative uninformed case, making use of the widely adopted hypothesis that model parameters are normally distributed. Case *c* takes advantage of the findings of Naraghi et al. [2018] who provide some experimental evidence suggesting that the pdf of pore radii in shales can be interpreted through a log-normal model. Our choice of performing sensitivity analyses according to configurations associated with diverse pdfs characterizing uncertain model parameters enables us to analyze the influence of model parameter pdf (which is generally unknown a priori) on the results of the GSA and, ultimately, on gas flow forecasting.

Considering the computational cost associated with multiple model evaluations (requiring 10^{-4} seconds per simulation on an Intel Xeon Gold 6148 CPU @ 2.4 GHz) required for these analyses, along with the corresponding expense for random sampling across the high-dimensionality parameter space of the model, we have conducted our analyses based on 2×10^7 model evaluations. This quantity has been deemed to represent an acceptable balance between the necessity of obtaining stable sensitivity analysis outcomes and the computational burden entailed (details not shown).

Variance-based Sobol' Indices

Sobol' indices [Saltelli and Sobol', 1995] can assist the appraisal and quantification of the relative expected reduction of the variance of a target model output due to knowledge of (or conditioning on) a given model parameter, which would otherwise be subject to uncertainty. In this context, considering a model output y, which depends on N random parameters collected in vector $\mathbf{x} = (x_1, x_2, ..., x_N)$ and defined within the space $\Gamma = \Gamma_1 \times \Gamma_2 \times ... \times \Gamma_N$ ($\Gamma_i = [x_{i,min}, x_{i,max}]$ corresponding to the support of the i-th parameter, x_i), the principal Sobol' index S_{x_i} associated with a given model parameter x_i is evaluated as

$$S_{x_i} = \frac{V\left[E\left[y|x_i\right]\right]}{V\left[y\right]}. \tag{2.9}$$

Here, $E\left[\cdot\right]$ and $V\left[\cdot\right]$ represent expectation and variance operators, respectively; the notation $y|x_i$ denotes conditioning of y on x_i. Note that S_{x_i} describes the relative contribution to $V[y]$ due to variability of only x_i. Joint contributions of x_i with other model parameters included in $\mathbf{x}$ to the variance of y are embedded in the total Sobol' indices (details not shown). We recall that relying on Sobol' indices to diagnose the relative importance of uncertain model parameters to model outputs is tantamount to identifying uncertainty with the concept of variance of a pdf. As such, while Sobol' indices are characterized by a conceptual simplicity and straightforward implementation and use, they provide only limited information about the way variations of model parameters can influence the complete pdf of model outputs.

Moment-Based AMA Indices

The recent moment-based GSA approach proposed by Dell'Oca et al. [2017, 2020] rests on the idea that the quantification of the effects of model parameter uncertainty on various statistical moments of the ensuing pdf of model outputs can provide enhanced understanding of model functioning. Dell'Oca et al. [2017] introduce Moment-Based sensitivity metrics (termed AMA indices) according to which one can evaluate the influence of uncertain model parameter on key elements of the model output pdf, as embedded in its associated statistical moments. The AMA indices are defined as follows (Dell'Oca et al. [2017]):

$$\mathrm{AMA}M_{x_i} = \frac{1}{|M[y]|} E\left[|M[y] - M\left[y|x_i\right]|\right]. \tag{2.10}$$

Here, $\mathrm{AMA}M_{x_i}$ represents the indices associated with a model parameter x_i and a given (non-zero) statistical moment M of the pdf of model output y. The AMA indices are intended to quantify the expected change of each statistical moment of y due to our knowledge of x_i. Large values of these indices indicate that variations of the associated parameter strongly affect the statistical moments of y.

2.4. Results and Discussion

2.4.1. GSA of Methane Flow Model

Parameter - (Units)	Range (CV%)	Criteria for the support - Reference
r_o - (nm)	2-100 (55)	Literature [Wu et al., 2016]
ϕ_o - (-)	0.005-0.1 (52)	Literature [Li et al., 2006]
p - (MPa)	0.5-50 (57)	Literature [Wu et al., 2016]
τ - (-)	2.8-5.8 (20)	Literature [Mohd Amin et al., 2014]
T - (K)	337-473 (10)	Literature [Chiquet et al., 2007]
p_c - (MPa)	51-90 (16)	Literature [Chiquet et al., 2007]
q - (-)	0.014-0.056 (35)	Literature [Dong et al., 2010]
t - (-)	0.02-0.04 (19)	Literature [Dong et al., 2010]
κ - (-)	0.1-2 (52)	Literature [Wu, Li, Wang, Yu and Chen, 2015]
D_f - (-)	2.1-2.9 (9)	Theoretical Limits [Coppens, 1999]
ΔH - (J/mol)	12000-16000 (8)	Literature [Wu, Li, Wang, Yu and Chen, 2015]
p_{L_o} - (Pa)	41-128 (30)	CV [Wu, Li, Wang, Yu and Chen, 2015]
α_0 - (-)	1.02-1.36 (8)	Literature [Karniadakis et al., 2005]
α_1 - (-)	2-6 (30)	CV [Karniadakis et al., 2005]
β - (-)	0.2-0.6 (30)	CV [Karniadakis et al., 2005]

Table 2.1: Ranges of variability for the methane migration model uncertain parameters considered in the GSA. Values of the coefficient of variation, criteria for the selection of the range of variability, and references considered for the definition of each range of variability are also listed.

The 15 uncertain model parameters of model (2.1) are considered to vary across the support defined through the ranges of variability listed in Table 2.1. These ranges have been designed upon considering available liter-

ature references (values typically employed for the model parameters in low permeability geomaterials). With reference to three of the model parameters i.e., p_{L_o}, α_1, and β, only very limited information is available from the literature, to the best of our knowledge [Karniadakis et al., 2005]. Thus, we take the values considered by Wu et al. [2016] and Karniadakis et al. [2005] as the centers of corresponding ranges of variability. We then consider their (uniform) distributions to be characterized by a given coefficient of variation (that we set as 30%), thus enabling us to imprint these parameters with a sufficiently broad range of variability, which is also consistent with the degree of variability documented for the remaining uncertain parameters (see Table 2.1). Finally, we allow the fractal dimension D_f to vary within a range of variability close to its theoretical bounds (i.e., $2< D_f <3$) [Coppens, 1999; Coppens and Dammers, 2006]. Methane properties (such as viscosity, compressibility, and deviation factor) are estimated using miniREFPROP [Lemmon et al., 2018], a tool that incorporates equations of state for a variety of gas species. With reference to methane miniREFPROP relies on the equation of state proposed by Setzmann and Wagner [1991].

x_i	AMAE_{x_i}	AMAV_{x_i}	S_{x_i}	AMAγ_{x_i}	AMAk_{x_i}
r_o	**0.728**	**0.798**	**0.417**	**0.562**	**0.757**
ϕ_o	**0.453**	**0.643**	**0.160**	**0.345**	**0.464**
p	**0.335**	**0.484**	**0.091**	**0.208**	**0.476**
τ	**0.181**	**0.356**	0.026	**0.114**	**0.213**
T	0.094	**0.163**	0.007	0.027	0.046
q	0.061	0.119	0.003	0.011	0.022
t	0.057	0.114	0.003	0.01	0.021
p_c	0.028	0.063	0.001	0.008	0.014
κ	0.010	0.005	0	0.004	0.007
ΔH	0.001	0.002	0	0.002	0.005
D_f	0.002	0.003	0	0.002	0.004
p_{L_o}	0.002	0.003	0	0.002	0.004
α_0	0.001	0.002	0	0.002	0.004
α_1	0.001	0.002	0	0.002	0.004
β	0.001	0.002	0	0.002	0.004

Table 2.2: Moment-based GSA indices AMAM_{x_i} and Sobol' principal indices S_{x_i} for all x_i parameters included in Equation (2.1). All model parameters are described by uniform pdfs (Case *a*). Values of each metric identifying the most influential parameters are reported in bold.

Table 2.2 lists the moment-based GSA indices related to mean (AMAE_{x_i}), variance (AMAV_{x_i}), skewness (AMAγ_{x_i}), and kurtosis (AMAk_{x_i}) of J as well as the principal Sobol' indices (S_{x_i}) evaluated for methane flow rate values rendered by Equation (2.1) for the case in which all model parameters are modeled as independent and identically

distributed random variables, each characterized by a uniform pdf (Case *a*).

While the strength of the influence of the reference pore radius (r_o) on the model output is not the same for the (first four) statistical moments, the AMA indices clearly suggest that conditioning on r_o has (overall) the strongest impact on the first four statistical moments of methane flow. This is then followed by reference porosity, pore pressure, tortuosity, and temperature. While the remaining model uncertain parameters still exert some influence on the (first four) statistical moments of J (as evidenced by the non-zero values of AMA indices), the strength of their influence can be considered as marginal when compared to the above-mentioned quantities, which are seen to be key in driving the main features of the pdf of methane flow. In the following we denote as *most influential parameters* for metrics AMAM_{x_i} or S_{x_i} all parameters x_i corresponding to AMA$M_{x_i}/\sum_{x_i}$ AMA$M_{x_i} \geq 5\%$ or $S_{x_i}/\sum_{x_i} S_{x_i} \geq 5\%$, respectively. Parameters identified as most influential by each of these metrics are reported in bold in Table 2.2. Values of the Sobol' principal indices are generally consistent with the results stemming from the moment-based GSA, even as τ and T are not identified as influential to the model output according to the Sobol' principal index. This result is consistent with the observation that conditional variance can be larger or smaller than its unconditional counterpart (see also Figure 2.2b) in a way that its integral over Γ_T vanishes. A similar effect associated with the principal Sobol' indices was identified by Dell'Oca et al. [2017] with reference to the Ishigami function, which is a widely used analytical benchmark in sensitivity studies.

Figure 2.2 depicts the first four statistical moments of J conditioned on values of the five most influential uncertain parameters selected on the

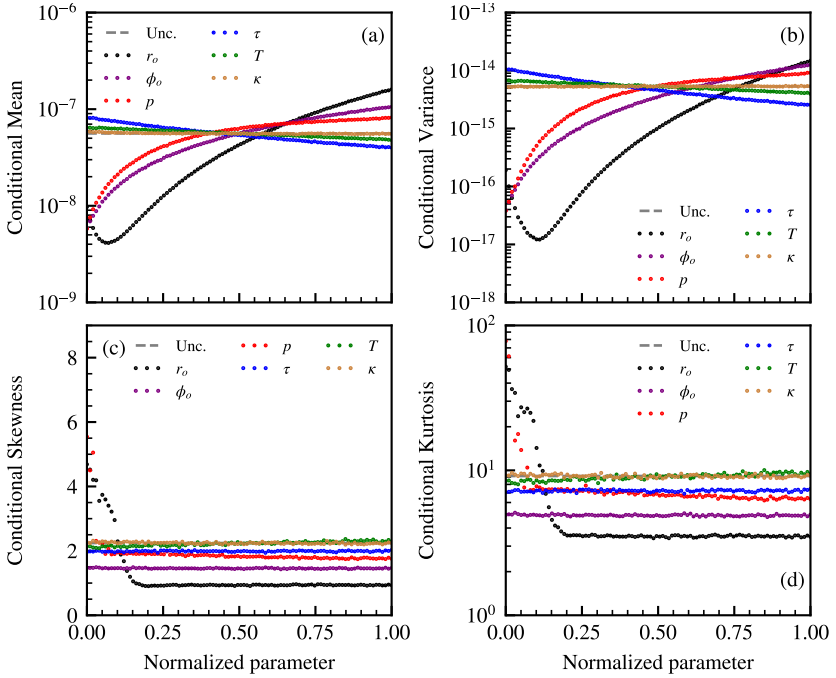

Figure 2.2: First four statistical moments of methane flow J (Ton/m^2 year) conditional on values of the most influential model parameters (see Table 2.2): (a) expected value, (b) variance, (c) skewness, and (d) kurtosis. The corresponding unconditional moments are also depicted (gray bold horizontal lines). Intervals of variation of the uncertain model parameters are rescaled within the unit interval for graphical representation purposes. All model parameters are described by uniform pdfs (Case *a*).

basis of Table 2.2. Uncertain parameters are normalized to span the unit interval, for ease of interpretation. Unconditional moments are also depicted as a reference. We note that, when considering conditioning on the model parameters which have been identified as non-influential according to the metrics employed, the difference between conditional and unconditional moments is negligible (details not shown).

As expected, conditioning on values of the reference pore radius (r_o) yields the most marked effects to all of the statistical moments considered (see black dotted curves in Figure 2.2). Mean and variance of methane flow generally increase with r_o. A minimum mean methane flow value is attained for 2< r_o <15 nm (corresponding to the range of normalized values comprised between 0 and 0.15 in Figure 2.2). The dominant transport mechanism for r_o <15 nm is surface diffusion, the strength of its contribution decreasing with increasing r_o (See Figure 2.6). As r_o increases, the strength of the contribution related to surface diffusion decreases faster than the corresponding increase of the slip flow contribution, thus resulting in a minimum value for the expected methane flow for values of the reference pore radius comprised in the aforementioned range. Otherwise, skewness and kurtosis (*i*) are affected by variations of the reference pore radius when the latter is smaller than 20 nm (corresponding to a normalized value of 0.18); and (*ii*) are generally constant for $r_o > 20$ nm. Nevertheless, we note that these statistical moments are still remarkably different from their unconditional counterparts even for large r_o values, thus evidencing the impact of acquired knowledge on r_o on reducing the asymmetry (as rendered by the skewness) and the peakedness and tailedness (i.e., the probability associated with extreme values, as rendered by the kurtosis) of the methane flow pdf.

Conditioning on pore pressure imprints variations to the statistics of the

model output which are qualitatively similar to those associated with r_o. Larger values of mean and variance of J are linked to larger values of p. This result descends from the linear relationship between pore pressure and slip flow (Equation (2.7)), the latter being the dominant mechanism in systems formed by larger pores (See Figure 2.6). Conditional skewness and kurtosis are constant (albeit different from their unconditional counterpart) across most of the variability range of p, sharp variations of these quantities being associated with conditioning on low values of p (i.e., corresponding to pore pressure values smaller than 10 MPa). Our findings about the influence of p on J are consistent with the results of Sun et al. [2017]. These authors find that increasing values of pore pressure lead to an increase of apparent permeability (which is in turn linearly proportional to gas flow) for $r_o > 10$ nm. Wu et al. [2016] document a similar behavior due to the dominance of the slip flow component (which is proportional to p; see Equation (2.7)) in systems characterized by large pores.

While the impact of reference porosity and tortuosity is not analyzed in any of the available previous studies [Sun et al., 2017; Wu et al., 2016, 2017; Wu, Li, Wang, Yu and Chen, 2015; Zhang et al., 2018], our results rank ϕ_o and τ as the second and fourth most influential parameters in the evaluation of the pdf of J, respectively (see Table 2.2). The correction factors for bulk (Equation (A.7)) and surface (Equation (A.1)) diffusion flow increase linearly with reference porosity. Thus, increased values of ϕ_o yield corresponding increases of the methane flow (and hence of its first two statistical moments) independent of the dominant transport mechanism. Conditional mean and variance of J decrease with increasing values of tortuosity. This is in line with the observation that all gas transport mechanisms are characterized by an inverse proportionality between J and τ through the correction factor which is related to these processes

taking place within a porous domain. These elements are consistent with a physical picture according to which fluid flow rates across a porous geomaterial are expected to increase and decrease with increasing porosity and tortuosity, respectively. Unlike pore pressure and reference pore radius, conditioning on reference porosity and tortuosity yields a reduction of skewness and kurtosis of the pdf of J, whose conditional values remain constant independent of the value of ϕ_o and/or τ.

Conditioning on temperature (T) affects the mean and variance of the methane flow pdf in a way that is qualitatively similar to the effect of tortuosity (albeit quantitatively to a lesser extent) due to the inverse proportionality between J and T. Otherwise, the overall shape of the pdf of J is not significantly influenced by the knowledge of T, values of conditional skewness and kurtosis practically coinciding with their unconditional counterparts.

The results listed in Table 2.2 suggest that statistical moments of methane flow are virtually insensitive to the remaining parameters (i.e., 10 of the 15 model parameters). Therefore, setting any of these parameters at given values within the variability space considered in our analysis yields negligible changes in the prediction of J. In this context, our results suggest that methane flow can be assessed with an acceptable degree of reliability even in the presence of scarce information about several parameters embedded in Equation (2.1) such as, e.g., the overburden pressure (i.e., p_c), the power-law exponents related to porosity (i.e., q) and pore radius (i.e., t), the fractal dimension of the pore surface (i.e.,D_f), or the isosteric adsorption heat of the geomaterial (i.e., ΔH). Further to this, our results suggest the opportunity to prioritize allocation of resources to robust characterization of (in descending order) reference pore radius, reference porosity, pore pressure, tortuosity, and temperature.

2.4.2. Impact of the Model Parameter pdfs on GSA Results

In this section, we analyze the impact of the choice of model parameter distribution on the pdf of J. As described in Section 2.3.2, we compare the GSA outcomes obtained with a uniform pdf for all model parameters (Case a) and illustrated in Section 2.4.1 against those computed when (i) all model parameters are characterized through truncated Gaussian pdfs (Case b) and (ii) r_o is described by a truncated log-normal pdf while the remaining parameters are described as in Case a (Case c). To provide a consistent comparison, Gaussian and log-normal pdfs are defined to honor the same mean and variance of the scenario associated with Case a.

x_i	AMAE_{x_i}	AMAV_{x_i}	S_{x_i}	AMAγ_{x_i}	AMAk_{x_i}
r_o	**0.787**	**0.828**	**0.761**	**0.608**	**0.692**
ϕ_o	**0.452**	**0.674**	**0.242**	**0.306**	**0.402**
p	**0.321**	**0.481**	**0.131**	**0.152**	**0.302**
τ	**0.182**	**0.363**	0.041	**0.088**	**0.157**
T	**0.100**	**0.178**	0.012	0.027	0.042
q	0.063	0.122	0.005	0.010	0.018
t	0.059	0.117	0.004	0.009	0.016
p_c	0.025	0.056	0.001	0.006	0.011
κ	0.007	0.005	0	0.005	0.008
ΔH	0.001	0.002	0	0.003	0.007
D_f	0.001	0.003	0	0.002	0.005
p_{L_o}	0.001	0.002	0	0.003	0.006
α_0	0.001	0.002	0	0.002	0.005
α_1	0.001	0.002	0	0.003	0.006
β	0.001	0.002	0	0.003	0.006

Table 2.3: Moment-based GSA indices AMAM_{x_i} and Sobol' principal indices S_{x_i} for all x_i parameters included in Equation (2.1). All model parameters are described by truncated Gaussian distributions (Case *b*). Values of each metric identifying the most influential parameters are reported in bold.

Table 2.3 lists values of AMA and principal Sobol' indices for each of the parameters embedded in Equation (2.1) for Case *b*. Results reported in Table 2.3 and Table 2.2 are qualitatively similar, i.e., the GSA yields similar results considering a uniform or a truncated Gaussian pdf for

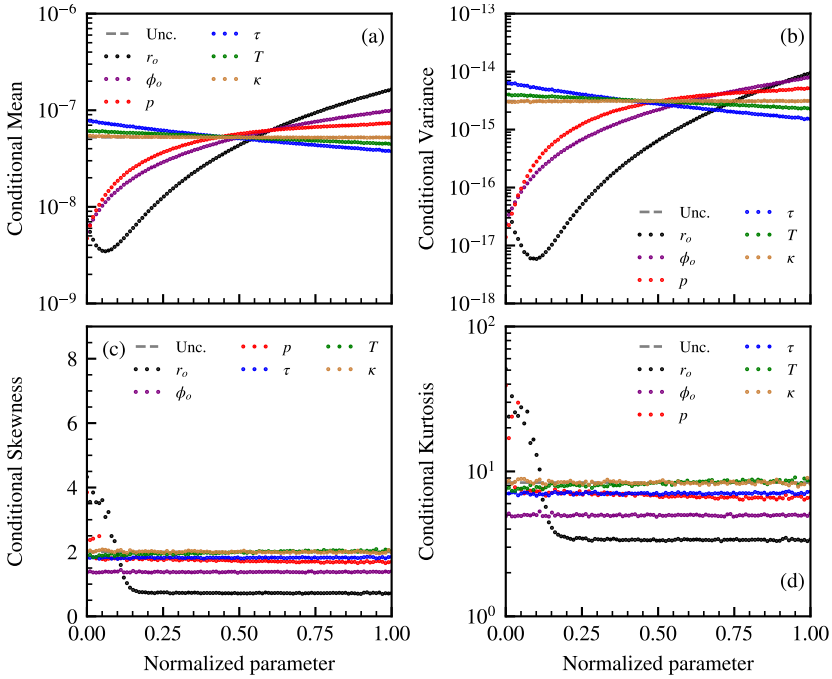

Figure 2.3: First four statistical moments of methane flow J (Ton/m^2 year) conditional on values of the most influential model parameters (see Table 2.3): (a) expected value, (b) variance, (c) skewness, and (d) kurtosis. The corresponding unconditional moments are also depicted (gray bold horizontal lines). Intervals of variation of the uncertain model parameters are rescaled within the unit interval for graphical representation purposes. All model parameters are described by truncated Gaussian pdfs (Case *b*).

all model parameters. Our results imbue us with confidence about the documented ranking of parameter importance, with reference pore radius, reference porosity, pore pressure, tortuosity, and temperature identified as the model parameters being the key drivers to the evaluation of the major features of the pdf of methane flow. Values of statistical moments of J conditioned on model parameters for Case b are very similar to those depicted in Figure 2.2 for Case a (See Figure 2.3).

x_i	AMAE_{x_i}	AMAV_{x_i}	S_{x_i}	AMAγ_{x_i}	AMAk_{x_i}
r_o	**3.342**	**3.585**	**2.843**	**0.791**	**0.884**
ϕ_o	**0.452**	**0.690**	0.065	**0.213**	**0.405**
p	0.183	**0.497**	0.011	**0.152**	**0.261**
τ	0.181	**0.359**	0.011	**0.069**	**0.164**
κ	0.114	0.013	0.005	0.026	0.039
T	0.069	0.134	0.002	0.021	0.045
q	0.062	0.120	0.001	0.012	0.029
t	0.041	0.111	0.001	0.012	0.031
p_{L_o}	0.023	0.016	0.000	0.014	0.030
p_c	0.023	0.062	0.000	0.013	0.029
ΔH	0.019	0.015	0.000	0.013	0.028
α_0	0.004	0.015	0.000	0.010	0.025
β	0.003	0.013	0.000	0.012	0.030
α_1	0.003	0.014	0.000	0.011	0.027
D_f	0.003	0.011	0.000	0.010	0.026

Table 2.4: Moment-based GSA indices AMAM_{x_i} and Sobol' principal indices S_{x_i} for all x_i parameters included in Equation (2.1). Reference pore radius (r_o) is described by a truncated log-normal distribution and the remaining model parameters are described by uniform distributions (Case *c*). Values of each metric identifying the most influential parameters are reported in bold.

Table 2.4 lists the AMA and the principal Sobol' indices associated with J for Case *c*. In this case, it is even more evident that the uncertainty of r_o is strongly dominant on the evaluation of the pdf of methane flow. Additionally, the blockage/migration ratio of the adsorbed molecules (κ)

gains importance with respect to previous cases, quantitatively impacting the pdf of J to an extent which is similar to what exhibited by temperature. This feature is attributed to the abundance of small pores in this scenario, which favors the dominance of the surface diffusion flow mechanism (linked to parameter κ).

Figure 2.4 depicts the first four statistical moments of methane flow conditioned on values of influential uncertain parameters for Case c (see Table 2.4). Unconditional moments are also shown as a reference. Overall, the results are qualitatively similar to those embedded in Figure 2.2 for Case a. The unconditional mean and variance of J in Case c are reduced (to approximately one-fourth and one-sixth, respectively) with respect to the corresponding values for Case a. Otherwise, unconditional skewness and kurtosis increase by about 2.6 and 6 times, respectively. These behaviors are attributed to the larger frequency of small reference pore radius values considered in Case c with respect to Case a (and b). Low values of reference pore radius are associated with large values of surface diffusion (See Figure 2.6) and to small values of mean and variance of methane flow. Conditioning on r_o and ϕ_o imprints variations to the model output mean and variance across the entire range of variability of these parameters (Figure 2.4). We further note that conditioning on r_o strongly reduces skewness and kurtosis of the pdf of J, thus reducing the probability associated with extreme (large) values of J.

Conditioning on p induces variations in the (first four) statistical moments of the model output. Conditioning on larger values of this quantity yields the highest values of mean and variance of the model output. A minimum in the values of conditional variance, skewness, and kurtosis is observed in the interval 1MPa$< p <$15MPa. Finally, the blockage/migration ratio of adsorbed molecules displays (a small but noticeable) influence on the

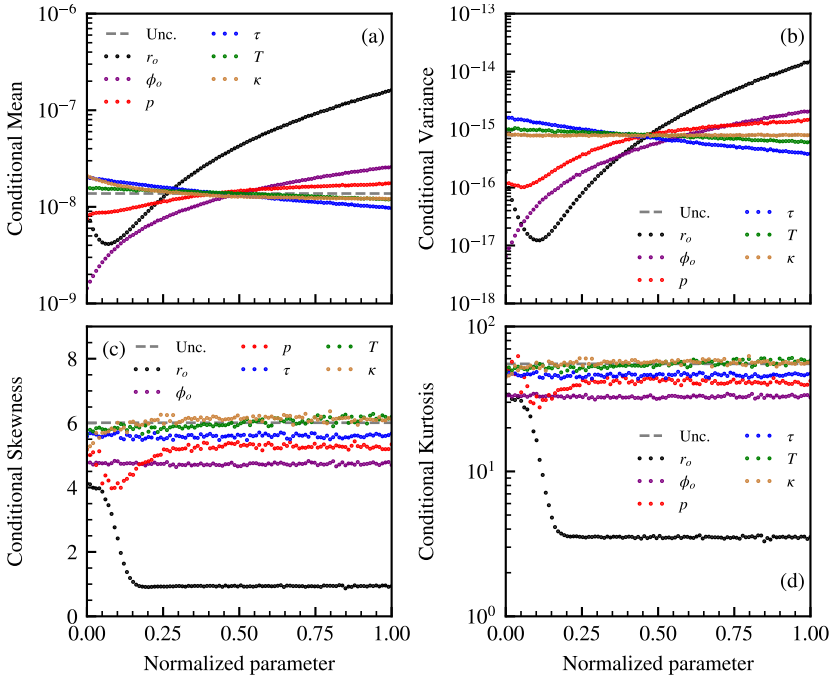

Figure 2.4: First four statistical moments of methane flow J (Ton/m^2 year) conditional on values of the most influential model parameters (see Table 2.4): (a) expected value, (b) variance, (c) skewness, and (d) kurtosis. The corresponding unconditional moments are also depicted (gray bold horizontal lines). Intervals of variation of the uncertain model parameters are rescaled within the unit interval for graphical representation purposes. Note that r_o is described by a truncated log-normal pdf and the remaining model parameters are described by uniform pdfs (Case c).

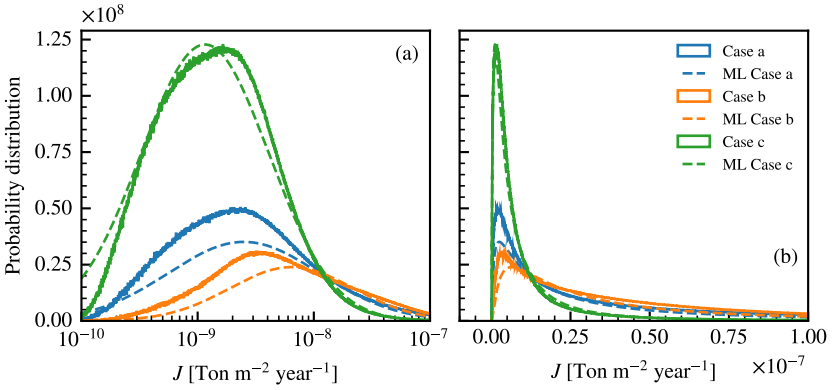

Figure 2.5: Probability density functions (in logarithmic (a) and natural (b) scale) of the overall methane flux rendered by Equation (2.1) for model parameters characterized by: (*i*) uniform distributions (Case *a*), (*ii*) truncated normal distributions (Case *b*), and (*iii*) uniform distributions with the exception of r_o which is represented by a log-normal distribution (Case *c*). Dashed curves represent a ML- based fit with a log-normal model for each case.

model output pdf. Mean and variance of J decrease with increasing values of κ. This behavior is expected, given the nature of κ, high values of this parameter being related to significant blockage of gas molecules on the geomaterial surface.

2.4.3. Overall Methane Flow

We evaluate the pdf of the overall methane flux (J) by making use of Equation (2.1) for all analyzed scenarios. Sample pdfs as well as corresponding Maximum Likelihood (ML) fits of log-normal distributions are depicted in Figure 2.5 in logarithmic and natural scales. Positive

skewness and large kurtosis are evident for all cases, these being larger for Case c, as illustrated in Section 2.4.2. These results reinforce the observation of higher frequencies of low J values in Case c with respect to the other settings investigated. Sample statistical moments (mean, variance, coefficient of variation, skewness, and kurtosis) of the pdf of J are listed in Table 2.5 together with the parameters of the ML-based log-normal models. The overall methane flux can vary across about three orders of magnitude (i.e., between 10^{-10} and 10^{-7} Ton/m^2 year). The largest variance of J is associated with Case c, where all parameters of model (2.1) are characterized by uniform pdfs by exception of r_o which is characterized by a truncated log-normal distribution. Finally, we remark that the results embedded in Figure 2.5 can be of practical assistance, as they allow for fast evaluations of the probability that methane flow in low permeability geomaterials exceeds a given threshold value.

Feature	Case a	Case b	Case c
Mean ($\times 10^{-8}$)	5.61	5.25	1.37
Variance ($\times 10^{-15}$)	5.33	3.10	0.80
CV	1.30	1.06	2.03
Skewness	2.24	2.00	6.01
Kurtosis	9.14	8.40	55.22
μ ($\times 10^{-8}$)	2.32	2.88	5.71
σ	1.49	1.22	1.26

Table 2.5: Sample mean, variance, coefficient of variation, skewness, and kurtosis of the overall methane flux Ton/m^2 year (Equation (2.1)) together with parameters of log-normal models (μ and σ) evaluated through ML fits against sample pdfs.

2.4.4. Identification of Dominant Flow Mechanisms

Figure 2.6 depicts color maps quantifying the relative strength of the contribution of the three flow mechanism (slip flow in red, Knudsen diffusion in green, and surface diffusion in blue) to the overall methane flux considering various combinations of all uncertain parameters embedded in Equation (2.1) for all scenarios investigated. Each sub-plot depicts the average value of the ratio $w_i J_i/J$ (with $i = v, k, s$ and $w_s = 1$) as a function of two parameters (i.e., averaging is performed with respect to uncertain parameters with the exception of the two varying along the (normalized) axes of the subplots), selected amongst those which were classified as most influential to the system (see Sections 2.4.1 and 2.4.2).

Our results indicate that the dominant flow mechanism in defining the methane flow is slip flow (in red in Figure 2.6) in all of the analyzed cases. An exception is observed for small values of the reference pore radius and/or small pore pressure, where surface diffusion is dominant. The contribution of Knudsen diffusion mechanism is always negligible. This suggests that it is possible to simplify Equation (2.1) by neglecting the Knudsen diffusion mechanism in the evaluation of methane flow under caprock conditions. Further simplifications of the methane flux model illustrated in Section 2.3.1 can be considered when the dominance of a given flow mechanism can be clearly established. For example, in cases where the value of reference pore radius (r_o) is known and the value falls in a zone where one can unequivocally distinguish between slip flow and surface diffusion dominance (i.e., far from the transition zone).

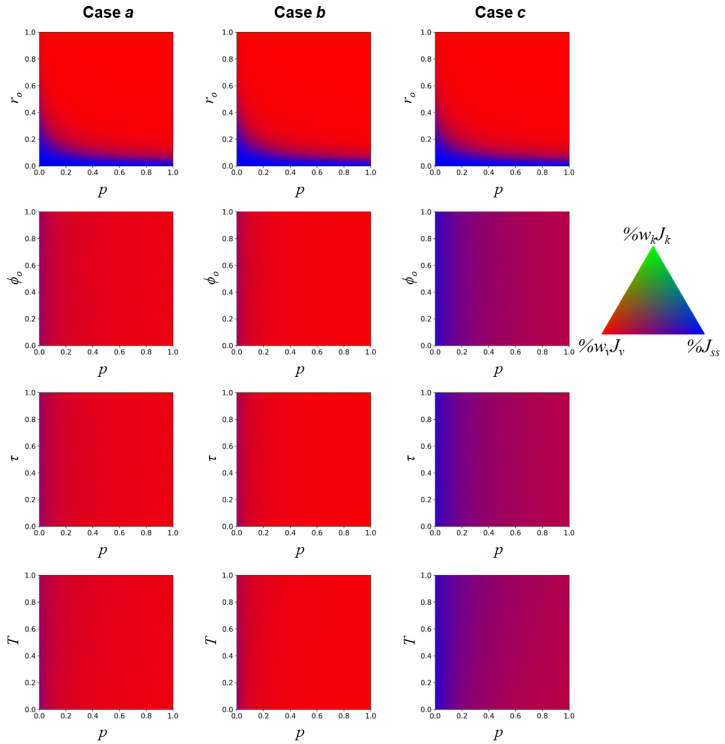

Figure 2.6: Relative contribution of the methane flow mechanisms (w_vJ_v, w_kJ_k, and J_s) to the overall methane flow J rendered by Equation (2.1). Intervals of variation of the uncertain model parameters are rescaled within the unit interval for graphical representation purposes. The results are shown for the three cases evaluated in Sections 2.4.1 and 2.4.2.

2.4.5. Scaling of Gas Flow Model

A pure diffusion modeling approach has been shown to represent with an acceptable degree of accuracy the movement of methane in low permeability media [Lu et al., 2015]. Such a model embeds all physics governing the system dynamics in a unique parameter (i.e., a diffusion coefficient D) and, under steady-state conditions, the mass flow-rate of methane can be expressed as:

$$J_d = -D\frac{\partial C}{\partial l}, \tag{2.11}$$

where $\partial C/\partial l$ represents the spatial gradient of methane concentration (C), i.e., the driving force of the system. Considering an isothermal system, the density of methane, $\rho = pM/RTZ$, and the relationship between concentration and density in a single phase system, $C = \rho$, the Equation (2.11) can be written as:

$$J_d = -\frac{DM}{RTZ}\left(1 - \frac{p}{Z}\frac{dZ}{dp}\right)\frac{\partial p}{\partial l}. \tag{2.12}$$

We complete our set of results and discussion by noting that the model illustrated in Section 2.3.1 coincides with a pure diffusion model (Equation (2.12)) under single-phase conditions, as we illustrate in the following.

Equation (2.1) can be written as:

$$J = -B\frac{\partial p}{\partial l}, \tag{2.13}$$

with $B = B_v + B_k + B_{ss}$, where

$$
\begin{aligned}
B_v &= w_v \zeta_{mb} \frac{r^2 p M}{8 \eta Z R T} (1 + \alpha K_n) \left(1 + \frac{4 K_n}{1 + K_n}\right), \\
B_k &= w_k \frac{2}{3} \zeta_{mb} r \delta^{D_f - 2} \left(\frac{8 Z M}{\pi R T}\right)^{1/2} \frac{p}{Z} C_g, \\
B_{ss} &= \zeta_{ms} \frac{D_s C_{sc}}{p}.
\end{aligned}
\tag{2.14}
$$

Comparing Equations (2.12) and (2.13), it can be seen that the diffusion coefficient D can be decomposed according to each flow mechanism as:

$$
D = D_v + D_k + D_{ss}, \tag{2.15}
$$

with

$$
D_i = \frac{B_i \; R T Z}{M \left(1 - \frac{p}{Z} \frac{dZ}{dp}\right)}, \tag{2.16}
$$

where $i = v, k, ss$. Note that we introduce three effective diffusion coefficients in Equation (2.15). These are associated with the slip flow (D_v), the Knudsen diffusion (D_k), and the surface diffusion (D_{ss}) components of model (2.1), respectively, and, to the best of our knowledge, are new for the flow model considered in this work. The variety of mechanisms included in model (2.1) are fully encapsulated in an overall diffusion coefficient D as illustrated in Equations (2.12), (2.15), and (2.16), where the contribution of each of the processes described in Section 2.3.1 is clearly recognizable.

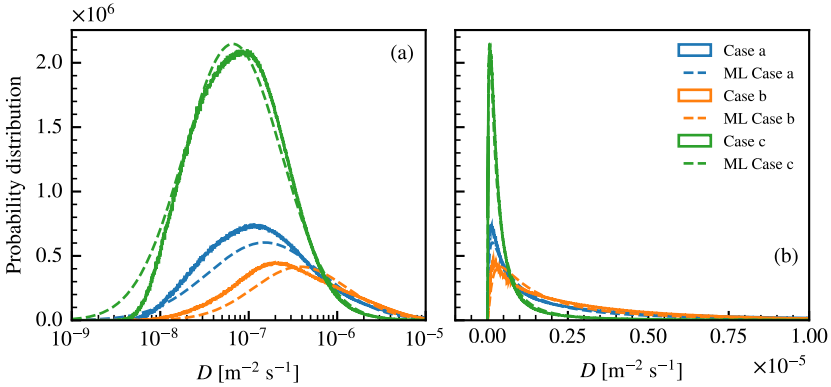

Figure 2.7: Probability density functions (in logarithmic (a) and natural (b) scale) of the overall diffusion coefficient rendered by Equation (2.15) for model parameters characterized by: (*i*) uniform distributions (Case *a*), (*ii*) truncated normal distributions (Case *b*), and (*iii*) uniform distributions with the exception of r_o which is represented by a log-normal distribution (Case *c*). Dashed curves represent a ML- based fit with a log-normal model for each case

Feature	Case a	Case b	Case c
Mean ($\times 10^{-6}$)	3.12	2.94	0.77
Variance ($\times 10^{-12}$)	15.7	9.28	2.40
CV	1.27	1.04	2.00
Skewness	2.17	1.92	5.82
Kurtosis	8.69	7.92	52.02
μ ($\times 10^{-6}$)	1.33	1.65	3.27
σ	1.47	1.21	1.26

Table 2.6: Sample mean, variance, coefficient of variation, skewness, and kurtosis of the overall diffusion coefficient D (m^2/s) (Equation (2.15)) together with parameters of log-normal models (μ and σ) evaluated through ML fits against sample pdfs.

We evaluate the pdf of the overall diffusion coefficient (D) using Equations (2.15) and (2.16) for all analyzed scenarios. Sample pdfs, as well as the corresponding ML fits of log-normal distributions, are depicted in Figure 2.7 in logarithmic and natural scales. Sample statistical moments, including the mean, variance, coefficient of variation, skewness, and kurtosis, of the pdf of D are listed in Table 2.6, along with the parameters of the ML-based log-normal models. The overall diffusion coefficient can vary across approximately four orders of magnitude, ranging between 10^{-9} and 10^{-5} m^2/s.

Following the approach proposed by Javadpour [2009], the Equation 2.13 can be written in terms of volumetric flow per unit area, q, as

$$q = k_{eff} \frac{\rho}{\eta} \frac{\partial p}{\partial l}, \tag{2.17}$$

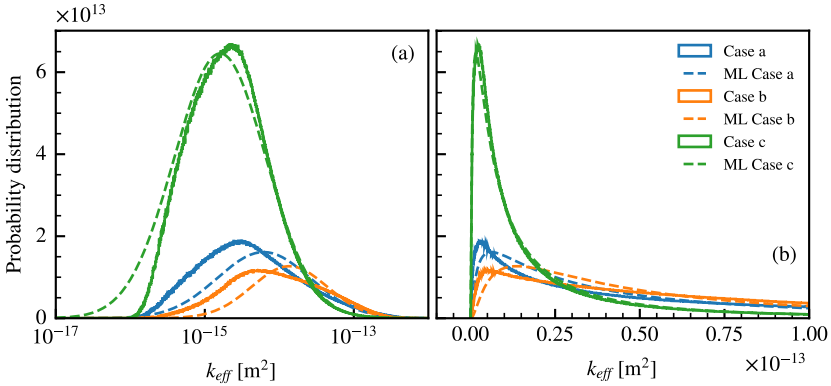

Figure 2.8: Probability density functions (in logarithmic (a) and natural (b) scale) of the effective permeability for model parameters characterized by: (*i*) uniform distributions (Case *a*), (*ii*) truncated normal distributions (Case *b*), and (*iii*) uniform distributions with the exception of r_o which is represented by a log-normal distribution (Case *c*). Dashed curves represent a ML- based fit with a log-normal model for each case

where k_{eff} is an effective permeability value, which is a function of the topology of the pore space of the host porous medium, and of properties of the gas (e.g., viscosity and density), which in turn depend on the pressure and temperature of the system.

Feature	Case a	Case b	Case c
Mean ($\times 10^{-14}$)	10.32	8.99	3.32
Variance ($\times 10^{-27}$)	13.74	7.42	6.00
CV	1.13	0.96	2.33
Skewness	2.39	1.76	13.79
Kurtosis	29.48	9.19	522.68
μ ($\times 10^{-14}$)	4.82	5.32	1.19
σ	1.43	1.17	1.41

Table 2.7: Sample mean, variance, coefficient of variation, skewness, and kurtosis of the effective permeability k_{eff} (m^2) of methane in low permeable materials, together with parameters of log-normal models (μ and σ) evaluated through ML fits against sample pdfs.

For the sake of completeness, we also assess the pdf of the effective permeability (k_{eff}). Sample pdfs, along with the corresponding ML fits of log-normal distributions, are portrayed in Figure 2.8, both in logarithmic and natural scales. The statistical moments of the pdf of k_{eff}, including the mean, variance, coefficient of variation, skewness, and kurtosis, are listed in Table 2.7, along with the parameters of the ML-based log-normal models. The range of variability of the effective permeability spans approximately five orders of magnitude, ranging between 10^{-17} and 10^{-12} m^2, values which are in line with the previous findings of Wang et al. [2020] and Liehui et al. [2019].

2.5. Conclusions

We perform a rigorous Global Sensitivity Analysis (GSA) to assess the impact of uncertain model parameters on the evaluation of methane flow (J) in low permeability media, such as caprocks. We study three scenarios that consider differing characterizations of the probability density function (pdf) describing model uncertain parameters to assess the impact of this choice on the results of the analysis. In details, we consider settings according to which: (*i*) all model parameters are represented through uniform pdfs, (*ii*) all model parameters are represented through truncated Gaussian pdfs, and (*iii*) the reference pore radius is characterized by a truncated log-normal pdf, all remaining parameters being associated with uniform pdfs.

Our work leads to the following main conclusions:

1. The uncertainty of methane flow is governed by uncertainty in the reference pore radius, followed (in decreasing order of importance) by reference porosity, pore pressure, tortuosity, temperature, and to a lesser extent, blockage migration ratio of adsorbed molecules. The remaining parameters of the investigated model (Section 2.3.1) being practically uninfluential. These results can assist future efforts to allocate resources during experimental activities aimed at characterizing methane flow in caprocks.

2. The shape of the pdf employed to characterize uncertain model parameters affects the results of our GSA. Additionally, it has a marked effect in the definition of the dominating transport mechanisms of the model. With reference to the model parameter variability considered in this study, as evaluated on the basis of available information, our results suggest that the dominant transport

mechanism is slip flow. Surface diffusion plays also an important role, especially for low values of reference pore radius and pore pressure, while Knudsen diffusion is negligible in all of the analyzed test cases.

3. The gas flow model introduced by Wu et al. [2016] (Section 2.3.1) can be related to a simple pure diffusion model by introducing an overall diffusion coefficient (D). The latter represented by the contribution of three effective diffusion coefficients, each associated with a well-defined flow mechanism. The ensuing mathematical structure of D allows distinguishing the relative contribution of all flow mechanisms to the overall methane flow. The relationship we derive also enables one to estimate the pdf of D when the model parameters are uncertain.

Symbols and Nomenclature - Chapter 1

Symbol	Refers to	Units	Evaluated
C_g	Gas compressibility	1/MPa	MiniREFPROP
C_{sc}	Adsorbed concentration	kg/m^3	Equation A.12
d_m	Gas molecule diameter	nm	0.38
$\partial p/\partial l$	Gradient of gas pore pressure	MPa/m	0.1
D	Overall diffusion coefficient	m^2/s	Equation 2.15
D_s	Surface diffusion coefficient	m^2/s	Equation A.9
D_k	Knudsen effective diffusion coefficient	m^2/s	Equation 2.16
D_{ss}	Surface effective diffusion coefficient	m^2/s	Equation 2.16
D_v	Slip flow effective diffusion coefficient	m^2/s	Equation 2.16
J	Mass flux of gas per unit of area	kg/(m^{2}s)	Equation 2.1
J_k	Knudsen diffusion	kg/(m^{2}s)	Equation 2.8
J_s	Surface diffusion	kg/(m^{2}s)	Equation 2.2
J_v	Slip flow	kg/(m^{2}s)	Equation 2.7
Kn	Knudsen number	-	Equation 2.5
M	Gas molar mass	kg/mol	1.6×10^{-2}
p_L	Langmuir pressure	MPa	Equation A.6
p_o	Atmospheric pressure	MPa	0.1
r	Pore size	nm	Equation A.3
R	Universal gas constant	J/(mol K)	8,3144
r_{ad}	Thickness of adsorbed gas layer	nm	Equation A.2
w_k	Knudsen diffusion flux weight factor	-	Equation 2.4
w_v	Slip mass flux weight factor	-	Equation 2.3
Z	Gas deviation factor	-	MiniREFPROP
α	Rarified effect coefficient for gas	-	Equation A.8
ζ_{ms}	Correction factor of surface diffusion	-	Equation A.1
ζ_{mb}	Correction factor bulk flow	-	Equation A.7
η	Gas viscosity	Pa s	MiniREFPROP
θ	Gas coverage of the geomaterial	-	Equation A.5
λ	Mean free path of gas molecules	m	Equation 2.6
ϕ	Porosity	-	Equation A.4

3 | Probabilistic Assessment of Groundwater Contamination Following Hydraulic Fracturing Operations

3.1. Abstract

Hydraulic fracturing is a technology with the potential to enhance oil and gas production from low-permeability reservoirs. However, its environmental impacts remain poorly understood. Therefore, comprehensive studies are imperative to minimize the risk of groundwater contamination resulting from hydraulic fracturing operations. In this study, we illustrate a methodology to conduct a probabilistic risk assessment of groundwater contamination following hydraulic fracturing operations, accounting for parametric uncertainty stemming from our limited knowledge of hydrogeological parameters and the precise operating conditions of the fracturing procedure. Our analysis is based on numerical simulations of multiphase flow in the area affected by hydraulic fracturing activities, considering two potential contamination scenarios. Given the probabilistic framework employed in this study, we must simulate numerous scenarios that

encompass various combinations of uncertain model parameters, leading to a substantial computational challenge. To address this issue, we employ polynomial chaos expansion, a cutting-edge technique for reducing model complexity[1].

3.2. Introduction

In this chapter, we present select outcomes stemming from our direct participation in an international research project (i.e., SORBACO - Shale Oil Risk Based Corrective Actions). Research activities were developed in collaboration with the University of Stuttgart (Germany) and a company from the private sector (EWRE, Israel). The project aims at estimating the potential risk of groundwater contamination arising from the migration of hydrocarbons and fracturing fluids following hydraulic fracturing operations performed in the shale rock layer that underlies a sedimentary basin. The operations are implemented for the purpose of retrieving liquid hydrocarbons stored within the shale rock. In light of confidentiality agreements, specific details concerning the study area and fracturing schedule must remain undisclosed. Nevertheless, the methodology and results disclosed herein are transferable to other scenarios and are scientifically relevant, as they contribute to bridging the knowledge gap regarding the environmental impacts of groundwater contamination following hydraulic fracturing operations [Howarth et al., 2011; Jabbari et al., 2017].

Here, our aim is to quantify groundwater contamination associated with two scenarios associated with hydraulic fracturing operations. The first scenario involves estimating the amount of fluids that migrate into the

[1] Data employed in this study is confidential and therefore a publication of these results in a research article is not possible.

source rock overburden through fractures created during the hydraulic fracturing operations. These fluids are of two types: (i) preexisting fluids in the hydrocarbon-bearing formation, and (ii) fluids injected to enhance the permeability of the source rock formation (i.e., fracking fluids). Then, in the second scenario, we ascertain the quantity of hydrocarbons and fracking fluids that flow through a preferential pathway and reach the bottom of a shallow groundwater system.

The estimation of the quantities of interest in each scenario is conducted by relying on numerical simulations, which are designed to model the physical processes occurring in the subsurface after hydraulic fracturing activities. Such numerical simulations are fraught with several complexities, including, but not limited to, (i) the absence of a complete characterization of geological systems, (ii) the lack of knowledge regarding the values of the parameters linked to the numerical models representing the processes in the system, (iii) uncertainties surrounding the features of hydraulic fracturing operations (e.g., injection flowrate), and (iv) the limited ability of numerical models to accurately represent all processes taking place within the system. Note that in the context of the SORBACO project, an international collaboration enabled us to receive and analyze the outcomes of the numerical simulations, which were in turn conducted in collaboration with the University of Stuttgart.

Given the numerous sources of uncertainty present within this study, including domain properties and geometry, mathematical formulation parameters, and system driving force features, we adopt a probabilistic framework [Riva et al., 2015, 2006; Tartakovsky, 2013]. Such a framework enables us to (i) estimate the probability of observing a given mass of contaminating fluids in specified locations of the system, (ii) determine the probability of surpassing a prescribed mass value that could poten-

tially contaminate the water of a groundwater body in a manner that poses a health hazard [Jabbari et al., 2017], and (iii) rigorously evaluate the individual contributions that uncertain model parameters have on the overall model output uncertainty through the application of Global Sensitivity Analysis (GSA) techniques [Dell'Oca et al., 2017; Saltelli and Sobol', 1995].

The successful application of a probabilistic framework often involves multiple evaluations of the model being analyzed. However, when such evaluations require significant computational resources, a probabilistic framing may become impractical. In such cases, utilizing a reduced complexity model can help to minimize the computational burden of model evaluations while still retaining the input-output relationships of the original model [Dell'Oca et al., 2017; Sudret, 2008]. In this study, we employ polynomial chaos expansion (PCE) to develop models of reduced complexity capable to mimic the surface of response, as well as the interaction of the model parameters, of the full numerical model (i.e., surrogate models).

The present chapter is structured as follows: Section 3.3.1 presents the mutiphase flow equations that are considered to represent the migration of contaminating fluids in the system. In Section 3.3.2, the two risk scenarios analyzed in this study are introduced alongside an exposition of the conceptual model adopted to represent the system behavior in each scenario and the uncertain model parameters that influence this process. Key concepts on polynomial chaos expansion are explained in Section 3.3.3. In Section 3.4, a detailed report of the key findings of the probabilistic quantification associated with the quantities of interest in each of the scenarios is presented. Finally, Section 3.5 provides a comprehensive summary of the study's conclusions.

3.3. Methodology

3.3.1. Numerical Assessment of Fluid Migration

In our study, we consider the two-phase (immiscible) mass conservation equations to represent the migration of both fracking fluids and hydrocarbons within the domains of both scenarios. For each phase γ the equation is expressed as follows

$$\phi \frac{\partial \rho_\gamma S_\gamma}{\partial t} - \nabla \left\{ \rho_\gamma \frac{k_{r\gamma}}{\mu_\gamma} \mathbf{k} \left[\nabla \left(p_\gamma - \rho_\gamma \mathbf{g} \right) \right] \right\} - q_\gamma = 0. \tag{3.1}$$

Here, $k_{r\gamma}$ [-], ρ_γ [ML^{-3}], S_γ [-], μ_γ [ML^{-1}T-1], p_γ [ML^{-1}T^{-1}], and q_γ [MT-1] respectively denote relative permeability, density, relative saturation, viscosity, pressure, and source/sink terms of phase γ; $\mathbf{k}$ [L^2] refers to the absolute permeability tensor, ϕ [-] represents porosity and $\mathbf{g}$ [LT^{-2}] represents gravity. Our simulations comprise two phases, a non-wetting phase ($\gamma = n$) and a wetting phase ($\gamma = w$). It should be noted that the saturation of the two phases constitutes a compositional quantity (i.e., $S_w + S_n = 1$) and that capillary pressure is defined as $p_c = p_n - p_w$. The Mualem - van Genuchten model [Van Genuchten, 1980] is employed to estimate the relative permeability of the wetting phase. The relationship between k_{rw}, S_w, and p_c can be expressed as:

$$\begin{aligned} k_{rw} &= \sqrt{S_w} \left[1 - \left(1 - S_w^N \right)^m \right]^2 \\ S_w &= \left[1 + \left(\alpha p_c \right)^N \right]^{-m} \end{aligned} \tag{3.2}$$

where N is a van Genuchten model parameter that depends on the properties of the porous material, α is a characteristic entry pressure of the

porous medium, and $m = 1 - 1/N$. The effects of multiphase flow in the fractures of the domain are considered by employing the approach described in Gläser et al. [2017].

These equations are solved by employing the numerical model DuMu^{x} [Koch et al., 2020], using an implicit Euler method for the time discretization, and a cell-centered finite volume scheme for the spatial discretization. It is worth noting that the numerical model employed in this study does not account for transport diffusive nor dispersive effects, and is thus only valid for Peclet numbers larger than 1, where advection is considered to be the dominant mechanism driving the movement of contaminant fluids in the system.

3.3.2. Scenarios of Analysis

Here we consider two distinct contamination scenarios that are associated with hydraulic fracturing operations. The joined analysis of these scenarios enables a comprehensive estimation of the likelihood of contaminating a shallow groundwater body. In the first scenario, we quantify the mass of fluids that migrate from a shale source rock formation into its associated overburden. These masses can be related to two types of fluids: (i) hydrocarbons that are naturally present in the shale formation, denoted by $M_{h,1}$, and (ii) fluids that are injected during hydraulic fracturing operations (i.e., fracking fluids), denoted by $M_{f,1}$.

Scenario 1

Figure 3.1 presents a scheme of the domain associated with Scenario 1. This scenario is characterized by several sedimentary layers, whose respective thicknesses are determined by the geological configuration of the study site. The range of permeability values associated with each layer

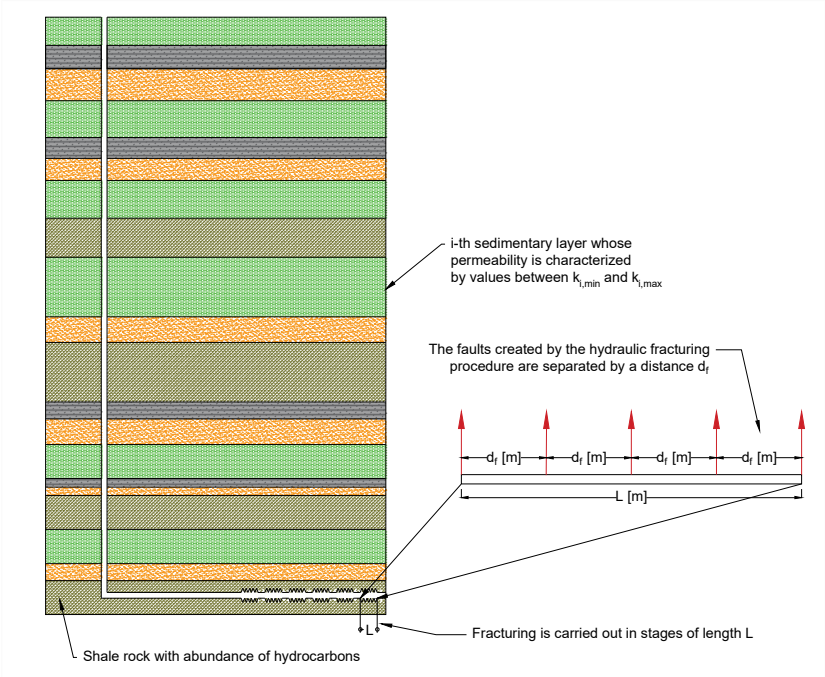

Figure 3.1: Analysis domain for Scenario 1, consisting of multiple sedimentary layers. The horizontal fracturing well is positioned in the middle of the shale rock layer, with the fracturing process carried out in stages in segments of length L. Within each segment, water is injected at five discrete positions spaced apart by a distance of d_f.

corresponds to typical permeability values associated with the material of the layer. The domain is deemed to be fully saturated with water, except for the source rock, whose porous space is considered to store both water and hydrocarbon fluids at a ratio of 0.4 and 0.6 in terms of volume, respectively. During each stage, the fracking fluid is injected at high pressure in segments of length L through the fracturing pipe that is placed in the middle vertical point of the source rock. Within each segment, water is injected at five discrete positions, which correspond to the perforation points of the stage, these injection points are spaced apart from each other by a distance of d_f, as indicated by the red arrows in Figure 3.1.

In order to estimate the quantities of interest of the Scenario 1, we conducted numerical simulations whose boundary conditions and conceptual scheme are depicted in Figure 3.2. In these simulations, the multiple layers that form the overburden of the shale rock are replaced by a single layer that is characterized by a height H, a homogeneous porosity ϕ_o, and an anisotropic and homogeneous permeability k_o. The latter is considered as uncertain in our study. The lower and upper limits of the range of variability of the horizontal permeability are respectively defined by the weighted harmonic mean of the minimum and maximum horizontal permeability values associated with each of the rock formations. The shale rock is characterized by a total height $2h$, a porosity ϕ_{sh}, and a homogeneous and anisotropic permeability k_{sh}, whose values were defined to resemble the study site conditions. The anisotropy of the hydrocarbon-bearing formation and its overburden is such that the value of the horizontal permeability is ten times greater than that of the vertical permeability. Within the length associated with each stage, at the five positions where water is injected (red arrows in Figure 3.2) we consider an equal number of vertical fractures that extend from the fracturing pipe

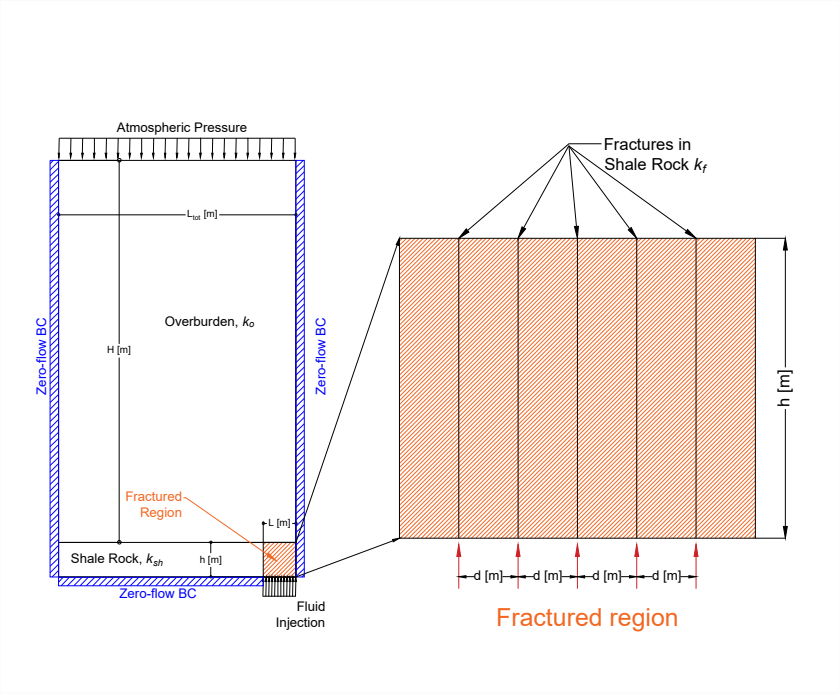

Figure 3.2: Sketch of the two-dimensional domain used in the simulations for Scenario 1. The overburden is represented by a single material. Only the upper half of the source rock, which has a thickness of h, is simulated. The domain has a total width L_{tot}, and height $H + h$. The numerical simulations consider the presence of five fractures within the fractured region, spaced apart by a distance of d_f. Such fractures connect the fracturing well to the bottom of the shale rock overburden. Pressure boundary conditions are imposed at the bottom of the fractured region and at the top of the domain, while zero flow boundary conditions are imposed at the remaining boundaries.

to the bottom of the overburden. The vertical fractures are characterized by a width of 1cm and permeability k_f, which we treat as an uncertain parameter. In accordance with the hydraulic fracturing operation plan, a fixed volume of water is pumped into the system while maintaining a constant wellhead injection pressure. As a consequence, the injection time (t_{inj}) is considered an uncertain parameter in our analysis.

The selection of the geometrical configuration of the fractures is based on the premise that it maximizes the amount of fluids that migrate into the overburden. This is due to the fact that fractures serve as preferential pathways within the source rock. Fractures are assumed to be fully developed before the injection takes place. Although this assumption may not reflect the actual conditions since (i) fracture developement requires time, (ii) fracture orientation may not necessarily be vertical, and (iii) the damage zone may not extend to the bottom of the overburden; it is taken given that the numerical model cannot incorporate the poroelastic processes that would allow for the representation of fracture development. Furthermore, preliminary simulations have shown that the mass of fluids estimated by considering all the stages in a single simulation is nearly identical to the product of the number of stages and the mass estimated by considering only one stage in the simulation. This is because the influence of neighboring stages on the numerical simulations is minimal. Thus, to save computational time and provided that not any error will be induced in the results, the simulations that we conduct here are only associated with one stage. As a consequence, the values that are calculated from the numerical simulations must be multiplied by the number of stages and by a length that accounts for the extension of the fracturing operations in the third dimension (here considered to be h) to obtain the total mass of fluids that migrate into the overburden in Scenario 1.

The numerical simulations of Scenario 1 involve a two-dimensional domain that encompasses the overburden and the upper half of the hydrocarbon-bearing formation (Figure 3.2). The initial condition is established by considering a hydrostatic pressure profile. A constant pressure that is equivalent to atmospheric pressure is imposed at the top of the domain, while a Dirichlet-type boundary condition is specified at the bottom of the domain in the region that is in contact with the stage length (L) to simulate the fracking fluids injection from the well. After the injection time has elapsed, the pressure at the bottom of the stage returns to the hydrostatic pressure. The remaining portion of the bottom boundary, as well as the right and left boundaries of the domain, are specified as zero-flow boundaries. The numerical simulations are carried out for a duration of one year since the mass transfer results indicate that only negligible amounts of fluid mass are transferred into the overburden beyond that period (See Section 3.3.1).

Table 3.1 lists the range of variability of the uncertain parameters that are employed in the numerical simulations of Scenario 1, along with the values of the deterministic parameters that are utilized. In addition to the range of variability, the coefficient of variation of each uncertain parameter is also provided. Moreover, the source that is utilized for the definition of the range of variability is specified for each parameter.

Parameter	Units	CV [%]	Range/Value	Source
k_o	m^2	35	5×10^{-17} - 2×10^{-16}	Study site conditions
k_{sh}	m^2	-	1×10^{-17}	Study site conditions
k_f	m^2	47	1×10^{-14} - 1×10^{-13}	3-4 orders of magnitude larger than k_{sh}
ϕ_o	-	-	0.15	Study site conditions
ϕ_{sh}	-	-	0.136	Study site conditions
ϕ_f	-	-	0.8	[Gläser et al., 2017]
α_o	1/m	-	1×10^{-3}	[Gläser et al., 2017]
α_{sh}	1/m	-	1×10^{-4}	[Gläser et al., 2017]
t_{inj}	min	35	100 - 400	Fracturing schedule
ρ_w	kg/m^3	-	1000	Water density
ρ_n	kg/m^3	-	897	Oil density
μ_w	Pa s	-	0.001	Water viscosity
μ_n	Pa s	-	0.018	Oil viscosity

Table 3.1: Ranges of variability for the model uncertain parameters considered in Scenario 1 and values of deterministic model parameters considered in this study. Values of the coefficient of variation of the uncertain model parameters are also listed.

Scenario 2

In the second scenario, we quantify the mass of hydrocarbons, denoted as $M_{h,2}$; and fracking fluids, denoted as $M_{f,2}$; that travel through a preferential pathway and ultimately reach the bottom of a shallow groundwater body. Figure 3.3 depicts a scheme of the domain associated with Scenario 2. This scheme is oriented perpendicularly to the scheme that is presented in Figure 3.1, resulting in a perpendicular alignment of the fracturing well with the domain. As in Scenario 1, several water-saturated sedimentary layers constitute the overburden of the hydrocarbon-bearing formation, and the permeability of each of these layers is defined based on the typical permeability values that are associated with the material of the layer. A preferential pathway (e.g., a fault) connecting the bottom of the over-

burden with the bottom of a shallow groundwater body is placed at a distance B from the fracturing well. Such a pathway is characterized by a permeability k_p and a thickness h_p. The mass of hydrocarbons and fracking fluids estimated in Scenario 1 are injected at the bottom-left portion of the domain.

In Scenario 2, the mass of hydrocarbons and fracking fluids ($M = M_{h,1} + M_{f,1}$) that is quantified via Scenario 1 is injected into the overburden, which is considered to be constituted of one geomaterial that possesses a height H, a homogeneous porosity ϕ_o, and homogeneous and anisotropic permeability k_o. The anisotropy of the source rock overburden is such that the horizontal permeability value is ten times greater than that of the vertical permeability. The mass M and the horizontal permeability k_o of the overburden are both considered uncertain parameters in our study. The range of variability of the overburden horizontal permeability is computed as for Scenario 1 whereas the range of variability of M is taken from the results of Scenario 1. The product of the permeability and thickness of the preferential pathway ($T_p = k_p \times h_p$), as well as the distance B between the fracturing well and the bottom of the preferential pathway, are considered uncertain quantities.

The numerical simulations conducted in Scenario 2 consider a two-dimensional domain that encompasses the overburden of the shale rock (Figure 3.4), this domain is oriented perpendicularly to the domain of Scenario 1, and the results are reported per unit length of the fracturing well. Consequently, the total mass of fluids reported must be multiplied by the length of the fracturing well in order to estimate the total mass of fluids that reach the bottom of the target formation after the hydraulic fracturing operations are carried out. The initial condition of the simulation is a hydrostatic pressure profile. A constant pressure that is

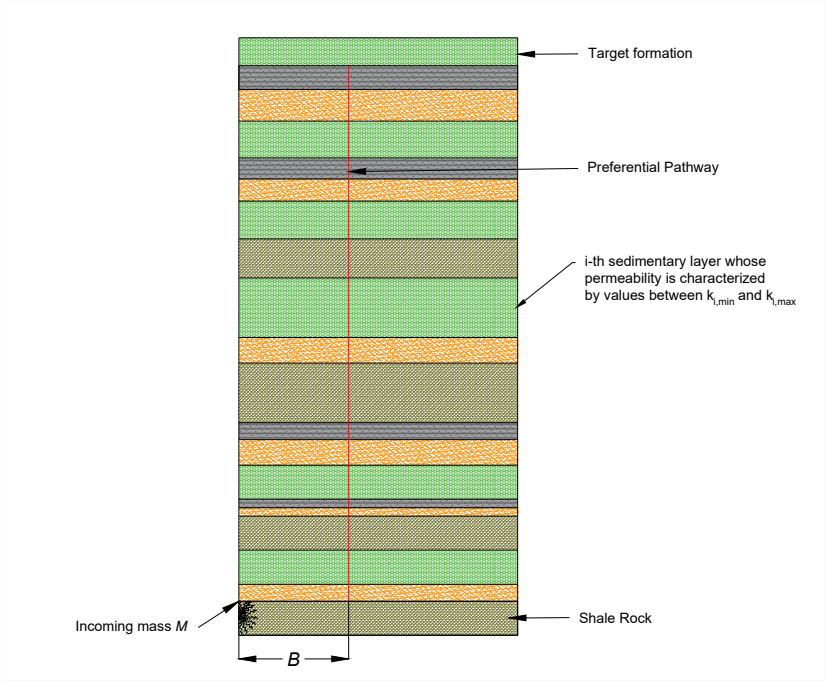

Figure 3.3: Analysis domain for Scenario 2, consisting of multiple layers that mimic the study area. The permeability of each layer is uncertain and defined by a range of minimum and maximum values associated with the layer material. This domain is perpendicular to the Scenario 1 analysis domain, resulting in a perpendicular alignment of the fracturing well with the domain. A preferential pathway, located at a distance of B from the fracturing well, connects the bottom of the overburden with the bottom of a target formation.

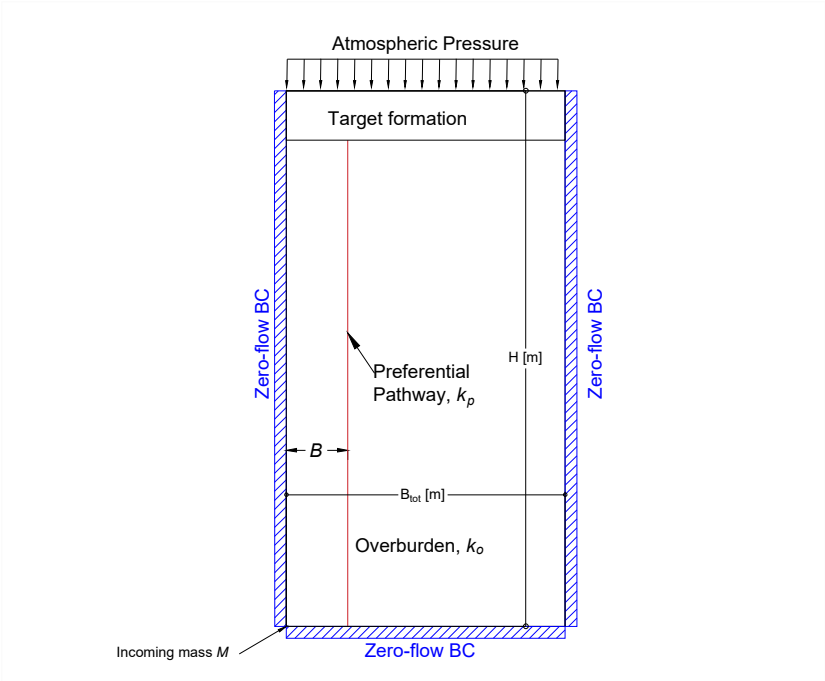

Figure 3.4: Sketch of the two-dimensional domain used in the simulations for Scenario 2. The overburden materials are replaced by a single material with equivalent hydrogeologic properties. The domain has a total width of B_{tot} and a height of H, with a vertical preferential pathway located at a distance of B from the fracturing well, connecting the bottom of the overburden with the bottom of a target formation. At the bottom left of the domain, a Neumann-type boundary condition is imposed to inject the incoming mass M into the domain over a period of one year. Pressure boundary conditions are imposed at the top of the domain, while zero-flow boundary conditions are imposed at the remaining boundaries.

equivalent to atmospheric pressure is imposed at the top of the domain, while the bottom boundary, as well as the right and left domain boundaries, are specified as zero-flow boundaries. To represent the migration of fluids from the source rock to the overburden, a source term is imposed at the bottom left portion of the domain. This source term enables the entire mass of fluids that is estimated in Scenario 1 to enter into the domain during the first year of the simulation. After that period, the source term is set to zero. The variations of relative permeability of hydrocarbons and fracking fluids in the overburden, are represented by employing a Mualem Van-Genuchten constitutive model. The parameter α_o of the Mualem - van Genuchten model is considered uncertain, given that this parameter controls the entry pressure of a fluid phase into a pore of a given size and can significantly impact the results of the analysis. The numerical simulations are carried out for a duration of 100 years since the mass transfer results indicate that only negligible amounts of fluid mass are transferred to the target formation beyond that period (See Section 3.3.1).

It is worth noting that the risk of groundwater contamination associated with hydraulic fracturing operations can not be limited to the study of these two scenarios alone, the risks associated with hydraulic fracturing of deep unconventional reservoirs are manifold. Our analysis, which is motivated by the risk scenarios that are defined in the FrackRisk project [Gläser et al., 2016], is limited to these particular scenarios. However, other possible sources of contamination must be analyzed and taken into consideration to minimize the risk of groundwater contamination after hydraulic fracturing operations. As an example, the work of Jabbari et al. [2017] provides a probabilistic assessment of the risk of groundwater contamination after well casing failure, the work of Kumar et al. [2018] experimentally monitors the concentration of chemicals that are

associated with hydraulic fracturing operations in groundwater bodies that are located in the vicinity of stimulated shale rocks, and in the deterministic study of Lange et al. [2013] and Kissinger et al. [2013], a set of flow scenarios is numerically analyzed to qualitatively estimate critical conditions favoring contamination of groundwater bodies. These studies illustrate the importance of a comprehensive analysis that accounts for various sources of contamination and the need for proactive measures to mitigate the potential risks associated with hydraulic fracturing operations.

Table 3.2 lists the range of variability of the uncertain parameters that are employed in the numerical simulations of Scenario 2, along with the values of the deterministic parameters that are utilized. In addition to the range of variability, the coefficient of variation of each uncertain parameter is also provided. Moreover, the source that is utilized for the definition of the range of variability is specified for each parameter.

Parameter	Units	CV [%]	Range/Value	Source
k_o	m^2	35	5×10^{-17} - 2×10^{-16}	Study site conditions
ϕ_o	-	-	0.15	Study site conditions
ϕ_p	-	-	0.8	[Gläser et al., 2017]
T_p	m^3	58	9.8×10^{-15} - 9.8×10^{-11}	Combination of fracture thickness and permeability uncertainty
α_o	1/m	57	1×10^{-5} - 1×10^{-3}	[Gläser et al., 2017]
M	Kg	35	22 - 90	Results Scenario 1
B	m	47	500 - 1500	Study site conditions
ρ_w	kg/m^3	-	1000	Water density
ρ_n	kg/m^3	-	897	Oil density
μ_w	Pa s	-	0.001	Water viscosity
μ_n	Pa s	-	0.018	Oil viscosity

Table 3.2: Ranges of variability for the model uncertain parameters considered in Scenario 2 and values of deterministic model parameters considered in this study. Values of the coefficient of variation of the uncertain model parameters are also listed.

3.3.3. Surrogate Modeling

Under a probabilistic framework, several (typically thousands) evaluations of the quantities of interest under diverse combinations of the model uncertain parameters of each scenario are required. However, due to the heavy computational burden associated with the multi-phase flow simulations of Scenarios 1 and 2, such a procedure is impractical in our study. To address this challenge, we employ models of reduced complexity that allow for a significant reduction in the computational time required for the execution of the numerical simulations.

In this study, we utilize the generalized Polynomial Chaos Expansion (PCE) surrogate modeling technique [Sudret, 2008]. This technique has been widely employed in the literature to decrease the complexity of nu-

merical models in various applications, such as large-scale stratigraphic simulations [Mahmudova et al., 2023], permeability damage control [Mahmudova et al., 2022], and solute transport in porous media [Dell'Oca et al., 2017]. In addition to reducing the computational burden associated with the probabilistic estimation of groundwater contamination, surrogate models allow to employ model analysis techniques, such as Global Sensitivity Analysis. Thus one can enhance the understanding of the functioning of the numerical model and ultimately the physical system. These techniques also aid in identifying the parameters whose uncertainty contributes the most to the uncertainty of the system behavior.

Within the context of PCE, a complex process $y(\boldsymbol{\theta})$ can be expressed as a linear combination of orthogonal multivariate polynomials, $\psi_i(\boldsymbol{\theta})$, as shown in Equation 3.3.

$$y(\boldsymbol{\theta}) \approx \sum_{i \in \Lambda^{P,D}} \beta_i \psi_i(\boldsymbol{\theta}) \text{ with } \psi_i(\boldsymbol{\theta}) = \prod_{p=1}^{P} \psi_p^d(\theta_p). \tag{3.3}$$

Here, β_i is the coefficient of the i-th term of the PCE surrogate, $\psi_p^d(\theta_p)$ is a univariate polynomial of order d of the parameter θ_p, and $\Lambda^{P,D}$ is a multi-index containing the indices of all the multivariate polynomials ($\psi_i(\boldsymbol{\theta})$) with degree equal to or smaller than the surrogate degree, D (i.e., multivariate polynomials where $\sum_{p=1}^{P} d \leq D$).

Note that in the context of PCE the univariate polynomials must satisfy the orthonormality condition, i.e., $E[\psi_p^j \psi_p^k] = \delta_{jk}$, where δ_{jk} is the Kronecker-delta function, $\delta_{jk} = 1$ if $j = k$ and zero otherwise. Multiple families of polynomials satisfy this condition; however, the selection of the suitable family of polynomials is made based on the shape of uncertain model parameters pdf. In this study, we consider the uncertain model

parameters to be uniformly distributed. Thus, the Legendre polynomial family is employed to construct the surrogates [Yang and Karniadakis, 2013].

The construction of a PCE-based surrogate requires the estimation of the surrogate model coefficients, $\beta = \{\beta_i, \forall\ i \in \Lambda^{P,D}\}$, and the selection of the surrogate model degree, D [Sudret, 2008]. Regarding the evaluation of β, we rely on least-square minimization (also termed as regression approach). According to this technique, the surrogate coefficients β are those that minimize the mean square error between the quantity of interest of each scenario computed with the full numerical model, $y(\boldsymbol{\theta})$, and the corresponding outputs of the surrogate model [Blatman and Sudret, 2011]. Thus, full numerical simulations need to be performed in order to estimate the coefficients β. The number of model evaluations is proportional to the number of model uncertain parameters and the degree of the considered polynomial (i.e., proportional to the number of elements in $\Lambda^{P,D}$). Typically, complex model surface responses are easier to represent with surrogates of higher degrees. Thus, the selection of the degree of the polynomial comes out of a trade-off analysis between the computational cost associated with the number of full model evaluations that one can afford and the acceptable error of the surrogate model.

In this study, we defined a maximum admissible computational burden for which we run a set of N_{FM} full model evaluations encompassing an equal number of randomly selected sets of parameters, such sets of parameters are randomly sampled employing a Quasi-Monte Carlo approach which guarantees that the parameter space is sampled uniformly [Niederreiter, 1992]. The evaluation of β is then performed by minimizing

$$\frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} \left[y(\boldsymbol{\theta}_s) - \sum_{i \in \Lambda^{P,D}} \beta_i \psi_i(\boldsymbol{\theta}_s) \right]^2 , \tag{3.4}$$

where $\boldsymbol{\theta}_s$ is the s-th randomly selected set of the uncertain model parameters.

In our analyses, the selection of D is performed on the basis of an accuracy test of surrogates with degrees varying between 2 and 5 for Scenario 1 and between 2 and 3 for Scenario 2. In such a test, the quantity of interest of 50 randomly selected sets of parameters (termed testing data), that are different from the sets of parameters employed for the estimation of β (termed training data), is evaluated with the numerical model and the surrogate. Then, the mean absolute error (MAE) between these two quantities is evaluated, and the surrogate degree D associated with the smallest MAE is selected and employed for the probabilistic estimation of the quantities of interest and for the GSA.

3.4. Results and Discussion

3.4.1. Numerical Simulations Outputs

Scenario 1

The numerical simulations of Scenario 1 allow for estimating the saturation, pressure, and relative permeability of both phases in the system. We employ Darcy's law to estimate the mass flux of both simulated phases (wetting and non-wetting) at the boundary between the shale rock and its overburden following a two-point flux approximation (Equation 3.5). Such a procedure is repeated for all the pairs of finite volumes adjacent to the interface between the shale and its overburden. Thus, the total mass flux of a phase γ between the shale rock and its overburden at a given time t, can be formulated as

$$Q_\gamma(t) = \sum_{b_{vol}=1}^{N_{b_{vol}}} A_{b_{vol}} \mathbf{k} \frac{k_{r\gamma}\rho_\gamma^2 g}{\mu_\gamma} \frac{\partial h_\gamma}{\partial z}, \tag{3.5}$$

where $A_{b_{vol}}$ is the contact area between two finite volumes, each one of them placed at one side of the boundary. $N_{b_{vol}}$ is the number of pairs of volumes in the boundary between the source rock and its overburden and h_γ is the hydraulic head of phase γ (i.e., $h_\gamma = p_\gamma/\rho_\gamma g + z$) with z denoting the elevation of the point with respect to a reference elevation.

Figure 3.5 presents the mass flux of the non-wetting (Figure 3.5a) and the wetting (Figure 3.5c) phases at the interface between source rock and its overburden, as a function of time. The mass flux is displayed for 351 combinations of model uncertain parameters (i.e., k_o, k_f, and t_{inj}), each combination represented by a gray line. It is observed that, for all the analyzed combinations, the numerical model results indicate a sharp in-

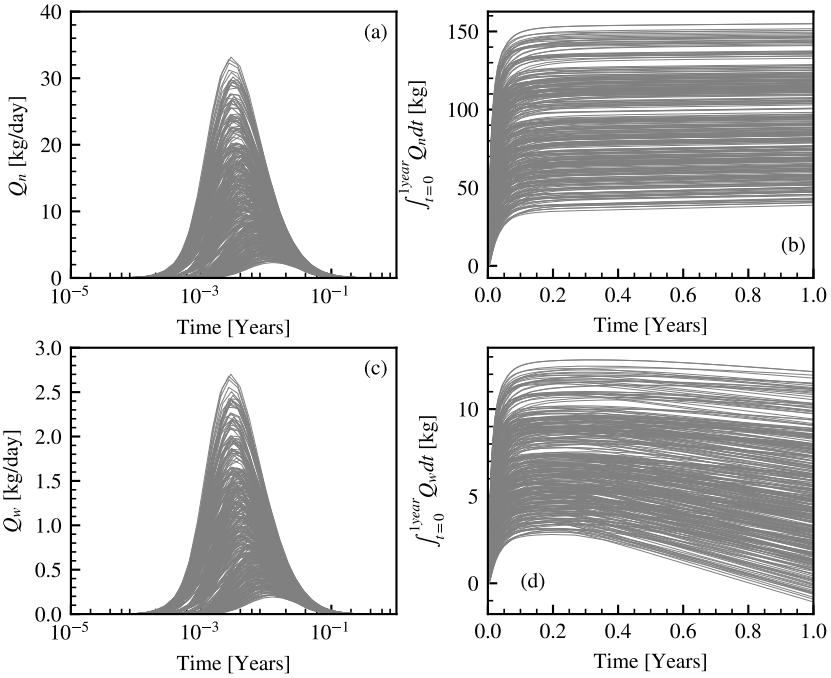

Figure 3.5: Flux [kg/day] and mass transfer [kg] of non-wetting [(a) and (b)] and wetting [(c) and (d)] phases expressed as a function of simulated time. The flux [(a) and (c)] at each time is estimated using Darcy's law (Equation 3.5). The mass [(b) and (d)] is estimated by integrating the fluxes over time. Each line represents the results of a numerical simulation associated with a set of parameter values randomly selected using a Quasi-Monte Carlo approach.

crease in the flux of both phases during and after the injection period, which varies from 100 to 400 minutes (i.e., 2×10^{-4} to 8×10^{-4} years), and reaches a maximum near 1×10^{-2} years (less than one week). Following the cessation of injection, the flux gradually decreases, eventually reaching negligible values after one year of simulation.

The mass of non-wetting (Figure 3.5b) and wetting (Figure 3.5d) phases that reach the overburden of the source rock is presented as a function of time in Figure 3.5. These quantities are computed by numerically integrating the mass fluxes of both phases over time. As expected, the transferred mass of both phases increases rapidly at the beginning of the injection due to the increment of pressure in the fracturing well, which induces an upward fluid movement. Then, once the injection of fracking fluids stops, the effects of a gravity-driven flow are felt in the system, causing the magnitudes of non-wetting and wetting phase masses to increase and decrease with time, respectively. Since our primary interest lies in the effects of injection (i.e., pressure-driven flow), and the bulk of mass transfer resulting from the fracking fluid injection occurs within a short time, we take $M_{h,1}$ and $M_{f,1}$ as the value of the mass of non-wetting and wetting phases transferred after one year, respectively.

The numerical analysis conducted reveals that $M_{h,1}$ varies between 40 and 150 kg, while $M_{f,1}$ ranges between 0 and 12 kg. As these quantities are taken as the mass M injected in Scenario 2, which is oriented perpendicularly to the domain of Scenario 1, the outputs of Scenario 1 must be adjusted by a factor of h/L. This adjustment factor is necessary because the fracturing operation is expected to damage the source rock in the horizontal direction perpendicular to the domain of Scenario 1, and such damage extension is considered to reach a length h. The results of Scenario 1 are then divided by the length of one fracturing stage to

provide the results per unit length of the fracturing pipe. In the present system, h/L is approximately 0.56. Applying this adjustment factor to the results of Scenario 1, the mass of the non-wetting phase transferred to the overburden ranges between 22 and 83.3 kg, while the mass of the wetting phase transferred ranges from 0 to 6.6 kg. Therefore, the total mass M transferred from Scenario 1 and injected in Scenario 2 varies between 22 and 90 kg.

Note that our method for quantifying the mass of fracking fluids transferred to the overburden relies on the underlying assumption that all fluxes of the wetting phase at the interfacial region between the source rock and overburden pertain exclusively to the movement of fracking fluids. However, given that the source rock is considered to be partially saturated with water (also termed resident water), a portion of the wetting phase fluxes may correspond to the displacement of resident water. Such water, while not contaminated with the chemicals utilized in the fracturing process, may contain traces of hydrocarbons. Therefore, as a conservative assumption, and given that the physical equations employed (multiphase-flow equations) are incapable of differentiating between resident water and fracking fluids, we consider that $M_{f,1}$ is solely associated with fracking fluids.

Scenario 2

As for Scenario 1, the numerical simulations carried out in Scenario 2 allow for estimating the saturation, the relative permeability, and the pressure of both phases in the system at each time and at each position of the domain. Then, the mass flux of both phases at the boundary between the preferential pathway and the target formation is estimated. Note that according to the simulations, the non-wetting phase flux is

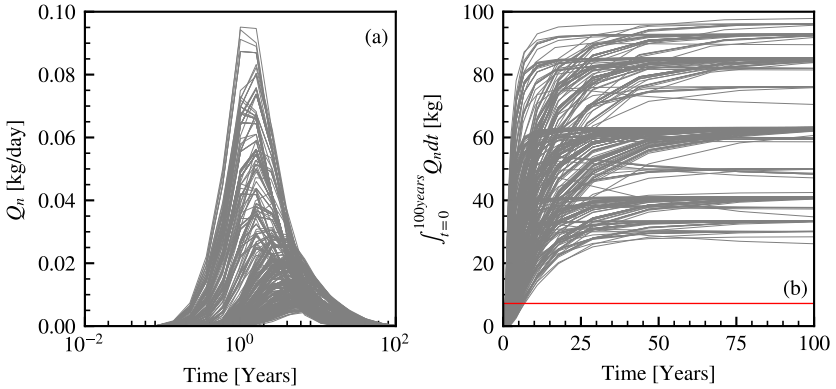

Figure 3.6: (a) Flux and (b) mass transfer of the wetting phase expressed as a function of simulated time. The flux (kg/day) is estimated via Equation 3.5. The mass (kg) is estimated by integrating the fluxes over time. Each line represents the results of a numerical simulation associated with a set of parameter values randomly selected using a Quasi-Monte Carlo approach. In (b) a red line representing the maximum mass of fracking fluids that reach the overburden per unit length of the fracturing well is also depicted.

zero for all times and all parameter combinations. This is explained because the fluids injected in the bottom left part of the domain move only during the first year of the numerical simulation due to the driven force caused by the source term. Once the injection of M finishes, the mass of hydrocarbons stops migrating and the non-wetting phase fluids remain trapped in the overburden. This allows us to safely consider that in this scenario, the probability of contaminating a shallow groundwater body with hydrocarbons is virtually zero. Thus, creating and analyzing a surrogate is not necessary for $M_{h,2}$ in this study case.

Figure 3.6 presents the mass flux of the wetting phase (Figure 3.6a) for 541 combinations of model uncertain parameters (i.e., k_o, M, α_o, B, and T_p), each gray line is associated with one parameter combination. The results of the numerical simulations indicate that the wetting phase flux increases during and after the source term M estimated in Scenario 1 is injected in the domain (during one year) and reaches a maximum between 1 and 10 years, then, the flux decreases and tends to negligible values after 100 years of simulation. The mass of wetting phase fluids that reach the overburden of the source rock is also presented in Figure 3.6b. This quantity is computed by numerically integrating the mass flux over time. As expected, the mass of the wetting phase increases rapidly at the beginning of injection due to the increment of pressure, which induces an upward fluid movement through the preferential pathway, then once the effects of the incoming mass injection finish, the total mass reaches a constant value ranging between 25 and 100 kg.

As the numerical model utilized in this study lacks the capability to differentiate between fracking fluids and resident water, it is not possible to estimate with the results of these simulations the precise quantity of mass reported in Figure 3.6b that can be unequivocally attributed to fracking

fluids. Nonetheless, taking into account that (i) the effect of the driving force is the same for the non-wetting and the wetting phases; (ii) the mass of wetting phase fluids reaching the target formation varies between 25 to 100 kilograms per meter of fracturing pipe; and (iii) considering the findings of Scenario 1, which suggest that the maximal quantity of fracking fluids per unit length reaching the overburden is 6.6 kilograms (indicated by the red line in Figure 3.6b), it is reasonable to conclude that the majority of the water transferred to the target formation is resident water that has not been contaminated with the chemicals used in the injected blend. Therefore, without the need for further analysis, it can be inferred that $M_{h,2}$ is equivalent to zero, and the maximum value of $M_{f,2}$ per meter of fracturing well is approximately equal to $M_{f,1} \times h/L$, albeit likely to be significantly less. For the sake of the completeness of the methodology here proposed, we construct a surrogate model and a sensitivity analysis of the volume of water displaced to the target formation M_w. Such a procedure enables us to assert the contribution of the uncertainty of individual model parameters to the uncertainty of M_w.

3.4.2. Construction of Surrogate Models

In Figures 3.7 (a) and (c), scatter plots are presented, which compare the results of the full numerical model with those of the most accurate surrogate model for $M_{h,1}$ and $M_{f,1}$. The corresponding MAE of the surrogate estimates is reported for both the training and testing data, alongside the order of the polynomial utilized for developing the surrogate model. The similarity of the training and testing error metrics suggests that the PCE surrogates do not suffer from overfitting. Note that the surrogate model for fracking fluids ($M_{f,1}$) is more accurate than the surrogate model for hydrocarbons ($M_{h,1}$). Additionally, the pdfs of $M_{h,1}$ and $M_{f,1}$ obtained

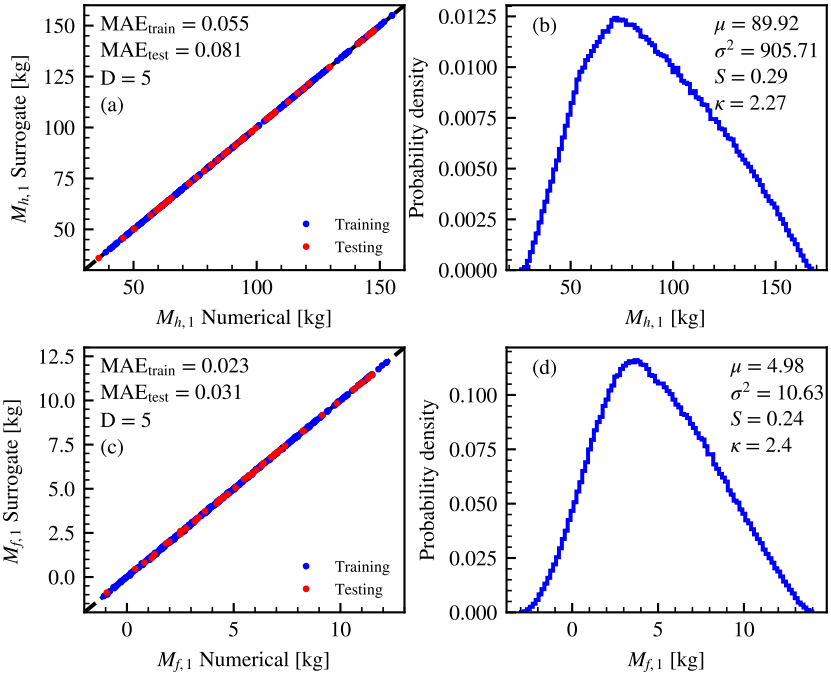

Figure 3.7: Scatter plots comparing the outputs of the full numerical model and the outputs of the most accurate surrogate model for (a) $M_{h,1}$ and (c) $M_{f,1}$, MAE for training and testing data, as well as the degree of the surrogate model are also reported. pdfs of (c) $M_{h,1}$ and (d) $M_{f,1}$ rendered by several evaluations of the surrogate model, the value of the first four statistical moments is also reported.

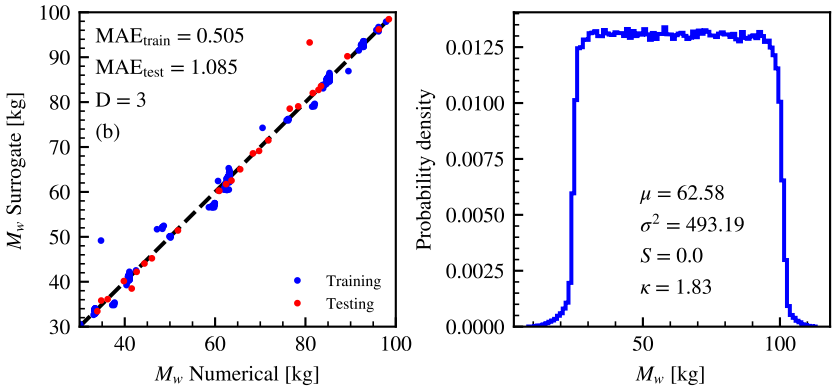

Figure 3.8: (a) Scatter plot comparing the outputs of the full numerical model and the outputs of the most accurate surrogate model for M_w, MAE for training and testing data, as well as the degree of the surrogate model are also reported. (b) pdf of M_w rendered by several evaluations of the surrogate model, the value of the first four statistical moments is also reported.

by evaluating the surrogates with $1{\times}10^6$ random parameter combinations are presented in Figures 3.7 (b) and (d), respectively. The mean, variance, skewness, and kurtosis of the surrogate estimations of $M_{h,1}$ and $M_{f,1}$ are also reported. The pdfs computed with the outputs of the surrogate model exhibit ranges of variability of $M_{h,1}$ and $M_{f,1}$ in concordance with those obtained via numerical simulations (see Figures 3.5(b) and (d)). The shape of both pdfs is unimodal, featuring a slight left skewness and platykurtosis.

Figure 3.8 presents for M_w a scatter plot (Figure 3.8(a)) comparing the outputs of the full numerical model and the outputs of the most accurate surrogate model. The corresponding MAE of the surrogate estimates is

reported for both the training and testing data, alongside the order of the polynomial utilized for developing the surrogate model. The similarity of the training and testing error metrics suggests that the PCE surrogates do not suffer from overfitting. Additionally, the pdf of M_w obtained by evaluating the surrogate with $1{\times}10^6$ random parameter combinations is presented in Figures 3.7(b). The mean, variance, skewness, and kurtosis of the surrogate estimations of M_w are also reported. The pdf computed with the outputs of the surrogate model exhibit a range of variability of M_w consistent with the range of variability exhibited by the results of the numerical simulations (see Figure 3.6(b)). The shape of the pdf resembles a uniform distribution.

3.4.3. Global Sensitivity Analysis

The results of the $1{\times}10^6$ surrogate model evaluations are employed to perform a global sensitivity analysis (GSA) of the surrogate models. Note that the results of the GSA remain unaltered by increasing the number of model evaluations (details not shown).

Figure 3.9 presents the moment-based GSA indices related to mean (AMAE), variance (AMAV), skewness (AMAS), and kurtosis (AMAK), as well as the principal Sobol' indices (Si) of both $M_{h,1}$ (3.9a) and $M_{f,1}$ (3.9b), which are evaluated with the PCE-based surrogate models estimated in the previous section. While the influence of the model parameters on the model output are not the same for the (first four) statistical moments, the AMA indices suggest that conditioning on the injection time (t_{inj}) and fracture permeability (k_f) have the most substantial impact on the first four statistical moments of $M_{h,1}$ and $M_{f,1}$. In contrast, the overburden permeability (k_o) has a relatively minor influence on the (first four) statistical moments of $M_{h,1}$ and $M_{f,1}$, as indicated by the

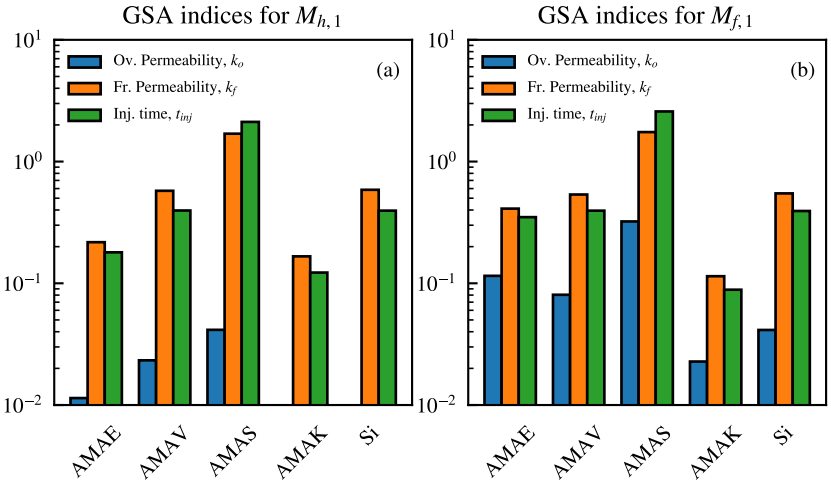

Figure 3.9: Moment-based GSA indices AMAM_{x_i} and Sobol' principal indices S_{x_i} for all x_i uncertain parameters included in the surrogate of $M_{h,1}$ (a) and $M_{f,1}$ (b), i.e., overburden permeability - k_o (blue), fracture permeability - k_f (orange), and injection time - t_{inj} (green).

very low values of the AMA indices associated with k_o. It is worth noting that k_o does exert a relatively larger impact on the statistical moments of $M_{f,1}$ when compared to $M_{h,1}$. However, the strength of its influence can be deemed negligible when compared to the above-mentioned quantities, which are recognized as the primary drivers of the principal features of the $M_{h,1}$ and $M_{f,1}$ pdfs. Values of the Sobol' principal indices are consistent and corroborate the results stemming from the moment-based GSA.

Figure 3.10 depicts the first two statistical moments of $M_{h,1}$ (top row) and $M_{f,1}$ (bottom row) conditioned on values of the surrogate model uncertain parameters. These are normalized to span the unit interval, for ease of interpretation. Unconditional moments are also depicted as a reference. Conditioning on values of the injection time (t_{inj}) and the fracture permeability (k_f) yields the most marked effects to the statistical moments of $M_{h,1}$ and $M_{f,1}$ (see green and orange markers in Figure 3.10). The mean of $M_{h,1}$ and $M_{f,1}$ increase with the values of t_{inj}, this behavior is consistent with a physical picture of the system because the larger the period in which the driving force acts on the system, a larger volume of the fluids in the rock can be displaced, and after one year the values of $M_{h,1}$ and $M_{f,1}$ are expected to increase. The mean of $M_{h,1}$ and $M_{f,1}$ also increase with the values of k_f, this behavior is also justified from a physical standpoint since the fractures in the shale rock act as preferential pathways between the fracturing well and the overburden, so, most of the fluids migrate through them and altering their permeability has a strong effect on the mass of fluids that is transferred to the overburden. Conditioning on the values of the overburden permeability (k_o) changes the mean of $M_{f,1}$ from its unconditional counterpart in a non-negligible way (See Figure 3.10(c)). This can also be explained from a physical standpoint. Since k_o defines the ability of the overburden to transmit

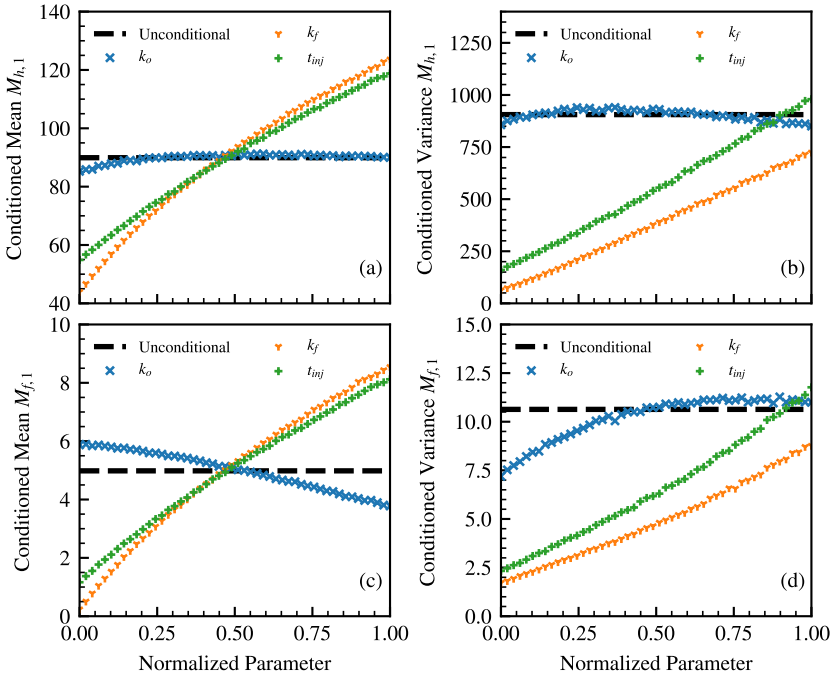

Figure 3.10: First two statistical moments of hydrocarbon mass [(a) and (b)] and fracking fluids [(c) and (d)] (kg) conditional on values of the surrogate model parameters: [(a) and (c)] expected value, and [(b) and (d)] variance. The corresponding unconditional moments are also depicted (black dashed lines). Intervals of variation of the uncertain model parameters are rescaled within the unit interval for graphical representation purposes

fluids, it also controls the strength of the gravity-driven flow in the system (that is felt after one year of simulation). Note that small values of k_o (which are associated with higher resistance to flow) should favor the migration of fluids with higher mobility (e.g., the wetting phase), on the other hand, when the permeability of the overburden increases the migration of the wetting phase is less favored. Such an effect is less marked in the conditional mean plot of $M_{h,1}$ since the magnitude of the latter is significantly larger, thus the effects of the gravity-driven flow and therefore the influence of k_o are minor in comparison to those of the pressure-driven flow. Nonetheless, low values of k_o show relatively smaller values of the $M_{h,1}$ conditional mean, suggesting that high resistance to flow also affects the magnitude of the mass transferred by a pressure-driving force.

Figure 3.11 presents the moment-based GSA indices related to mean, variance, skewness, and kurtosis, as well as the principal Sobol' indices (Si) of M_w evaluated with the PCE-based surrogate models estimated in the previous section. AMA indices suggest that conditioning on M has relatively the largest influence on the first four statistical moments of the model output. The influence of the remaining model parameters (k_o, α_o, B, and T_p) exert some influence on the first four statistical moments of the model output (as evidenced by the non-zero values of AMA indices). However, the strength of their influence can be considered minor when compared to the above-mentioned quantity. Values of the Sobol' principal indices are consistent with the results stemming from the moment-based GSA.

Figure 3.12 depicts the first two statistical moments of M_w conditioned on values of the surrogate model uncertain parameters. These are normalized to span the unit interval, for ease of interpretation. Unconditional

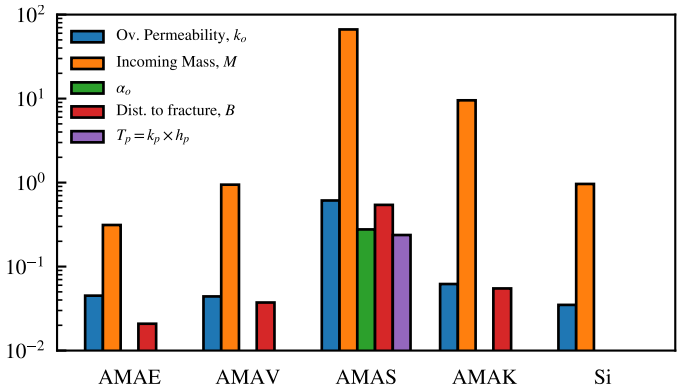

Figure 3.11: Moment-based GSA indices AMAM_{x_i} and Sobol' principal indices S_{x_i} for all x_i uncertain parameters included in the surrogate of $M_{h,2}$, i.e., overburden permeability (blue), incoming mass (orange), α_{vg} (green), distance to fracture B, and the product between the preferential pathway thickness and permeability T_p.

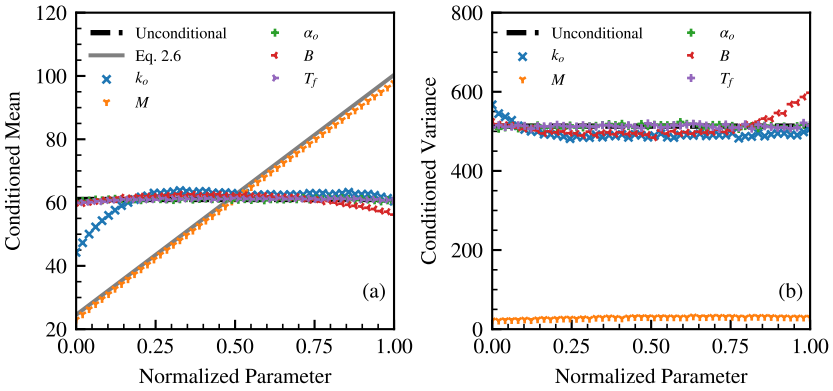

Figure 3.12: First two statistical moments and fracking fluids mass (kg) conditional on values of the surrogate model parameters: (a) expected value, and (b) variance. The corresponding unconditional moments are also depicted (black dashed lines). Intervals of variation of the uncertain model parameters are rescaled within the unit interval for graphical representation purposes. The values of $M_{f,2}$ estimated via Equation 3.6 are also included for reference (gray).

moments are also depicted as a reference. conditioning on values of the incoming mass ($M = M_{h,1} + M_{f,1}$) yields the most marked effects on both the statistical moments considered (see orange markers in Figure 3.12). The mean of the model output increases linearly with the values of M, such an effect is expected since the mass of fluids M injected at the bottom left part of the domain displaces the resident fluids in the system and larger values of M should increase the magnitude of M_w. The model output variance is significantly reduced when M is fixed, indicating that most of the variability of the model output is attributable to M, which is confirmed by its high Sobol index. As such, one can perform quick estimations of M_w as

$$M_w = \frac{\rho_n}{\rho_w} M_{h,1} + M_{f,1} \tag{3.6}$$

This is, the total mass of fluids migrating into the shallow aquifer formation is the sum of the mass of fracking fluids estimated in Scenario 1 plus the mass of hydrocarbons affected by the ratio between the densities of the fluids. Although this equation is a simplified version of reality it represents very accurately (gray line in Figure 3.12(a)) the outputs of the numerical model. Note, that even if the influence of the remaining model parameters in the variability of the model output is minor in comparison to the above-mentioned parameter, it is never zero.

3.5. Conclusions

We probabilistically quantify the risk of groundwater contamination after hydraulic fracturing activities for two contamination scenarios. Such contamination is deemed to be provoked by the migration of two types of fluids, namely, the water employed to perform hydraulic fracturing operations, and liquid hydrocarbons originally resident in the hydrocarbon-bearing formation and displaced during the hydraulic fracturing operations. The joined analysis of both scenarios enables us to estimate the risk of contamination in a shallow groundwater body connected to the overburden through a preferential pathway, however, the analysis of each of them allows us to gain insights into the physical processes occurring in the system. We based our analysis on numerical simulations considering multiphase flow and Global Sensitivity Analysis techniques. Given the computational burden of the simulations we also employed PCE-based surrogate modeling techniques. Our work leads to the following main conclusions:

1. The mass of hydrocarbons and fracking fluids reaching the overburden after hydraulic fracturing operations is probabilistically estimated. Our results indicate that the statistical moments of these quantities are mainly controlled by the injection time of the hydraulic fracturing operations (which is linked to the injected volume), and the permeability of the fractures created in the source rock. A minor contribution of the overburden permeability is also detected and the influence of all model uncertain parameters on the conditional mean and variance of the mass is assessed and satisfactorily justified with a physical interpretation of the system behavior.

2. The mass of hydrocarbons reaching the target formation is zero for

all the combinations of uncertain model parameters considered in this study, our findings suggest that this behavior is attributable to the low mobility of such type of fluid, to the overburden high resistance to flow, and to the relatively large distance to the closest fault in the analyzed system. The mass of water displaced to the target formation is probabilistically estimated, however, its magnitude is significantly larger than the water that migrates from the hydrocarbon-bearing formation to the overburden after hydraulic fracturing operations, this indicates that most of the water arriving at the target formation is resident water of the overburden.

3. As expected, the results of Scenario 1 suggest that the movement of fluids from the hydrocarbon-bearing formation to the overburden is mainly controlled by the pressure gradient imposed by the injection of fluids from the fracturing well. However, the results indicate that gravity-driven flow plays a minor but noticeable role that becomes important for large times. Such gravity-driven flow is mainly controlled by the overburden resistance to flow, which is in turn related to the inverse of the overburden permeability.

4. The results obtained in this study are limited because the effects of hydrodynamic dispersion and molecular diffusion are not considered. However, given that the hydrocarbons and a portion of the fracking fluids remain trapped in the overburden after the stimulation of the fracking operations, the migration of contaminants can be only associated with molecular diffusion. Thus, in order to extend the generality of the results additional simulations can be performed considering diffusive and dispersive effects. Alternatively, a third scenario studying the diffusive/dispersive transport of the results can be analyzed.

4 | An Original Deconvolution Approach for Oil Production Allocation Based on Geochemical Fingerprinting

4.1. Abstract

Production allocation involves identifying the proportions of different components in an oil mixture, and it plays a vital role in resource management, efficient resource utilization, equipment optimization, and equitable revenue distribution among stakeholders. This allocation process typically consists of two steps: (i) determining the chromatograms associated with the mixture and the end members representing the fluids in it, and (ii) using a deconvolution algorithm to estimate the mass fraction of each end member. In this study, we introduce a novel deconvolution algorithm for oil production allocation that streamlines the process without requiring extra lab work and demonstrates its superior accuracy compared to traditional methods. Additionally, we extend established deconvolution algorithms within a stochastic Monte Carlo framework to

enhance their reliability[1].

4.2. Introduction

Several production allocation methodologies and procedures are available in the literature. Key differences among them are chiefly related to the **experimental methodology** for Gas Chromatography (GC) fingerprinting and the **deconvolution algorithm** employed to analyze the ensuing data. Three main GC experimental methods can be identified: (i) High-resolution gas chromatography (HRGC), targeting the aliphatic and aromatic peaks lying between dominant n-alkanes; (ii) Multidimensional gas chromatography (MDGC) based on the quantitative target analysis of C8-C9 alkylbenzenes; and (iii) Saturate and aromatic fraction gas chromatography-mass spectrometry (GC-MS) analysis.

HRGC [Kaufman et al., 1987, 1990] is based on the GC analysis of whole-oil samples. The chemical species considered are typically in the range C_8-C_{20} and the acquired chromatograms provide peak heights of a variety (sometimes hundreds) of molecules (which mostly remain unidentified). The main element guiding peak selection for the quantitative deconvolution is the possibility of finding a set of components that can enable one to discriminate between End Members (EMs) and can therefore be used to deconvolute commingled oil samples. Otherwise, peak selection can be somehow arbitrary and largely a subjective (and operator-dependent) element. Moreover, HRGC is often plagued by poor chromatographic resolution. Accuracy can then deteriorate with time, as it might be significantly affected by changes in detector response and baseline drift.

[1] Results of this study were published in a research article which can be consulted at https://doi.org/10.1016/j.fuel.2022.124715.

MDGC [Mohamed, 2000] focuses on a limited number of molecules (11 alkylbenzenes). These can be considered as a constrained, while representative, dataset capable of explaining much of the variability between oil groups. The MDGC technique selectively detects a limited number of molecules, all of them chromatographically well separated, a feature which positively affects the accuracy, repeatability, and reproducibility of the analysis. Therefore, peak heights can be readily determined, since either external or internal standard calibrations can be conveniently carried out on a constrained number of components. This element is markedly relevant when analyses are performed across wide temporal windows (e.g., during production monitoring activities). In such cases, newly assimilated samples can be analyzed over time without the need to re-analyze previously acquired data.

Finally, GC-MS analyses [Jweda et al., 2017; Liu et al., 2017] have the notable advantage of unambiguously identifying an extensive set of geochemical features in the chromatogram. The resulting database can then be used to identify basin-specific indicators which are directly linked to up-to-date production/performance data. This approach might require that multiple analyses (one for each type of component) be carried out on the same sample. As such, this makes (in principle) the methodology significantly more expensive and time-consuming than MDGC. Furthermore, some of these analyses require pre-analytical steps (for example fractionation through open column/medium pressure GC). These can introduce further biases, thus potentially affecting the accuracy of the deconvolution results.

When considering deconvolution algorithms for data processing, a variety of approaches have been described in the literature [Barrie et al., 2020; Baskin et al., 2013; Jweda et al., 2017; Liu et al., 2017; McCaffrey et al.,

2011; Mohamed, 2000; Nouvelle et al., 2012; Zhan et al., 2016]. These approaches can be framed within the general context of system identification, also leveraging on statistical signal processes analysis [Spagnolini, 2018]. In this framework, a system is excited by one or multiple known signals (e.g., EM chromatograms). The objective is then to estimate the target response (i.e., EMs mass fraction in the mixture) from available observations of the system output (i.e., mixture chromatogram).

Deconvolution algorithms can be grouped into two main categories (i) methods based on peak heights (or actual concentrations) and (ii) methods based on peak ratios, evaluated from peaks associated with molecules eluting at close times. The use of peak ratios instead of peak heights is consistent with the possible change of baseline of the GC detection due to the use of multiple GC devices or improper equipment calibration. Note that a baseline change can significantly affect the monitored peak height whereas the peak ratios remain unaltered [Dembicki-Jr., 2017].

The first algorithm for geochemical production allocation has been proposed by Kaufman et al. [1987], who exploited peak ratios in the C_{15}-C_{20} molecular range. This approach considers the fractional composition of an EM in a mixture to be related to the difference between the peak ratios of the EM and of the mixture. As highlighted by McCaffrey et al. [2011], the approach suffers from two main drawbacks: (i) ratios of mixture chromatogram are not linear combinations of the ratios of EMs, so that the use of artificial mixtures with known EM contributions is required; (ii) the method is typically restricted to the allocation of mixtures composed by (at most) three EMs, as the mixing curves are not associated with a simple graphical representation. McCaffrey et al. [2011] proposed an approach that makes use of peak heights (rather than peak ratios) where the relative amount of each molecule (that is proportional to peak heights)

in commingled samples is the result of a (weighted) linear combination of the concentration of molecules in each of the EMs. Relying on this approach circumvents the need for artificial mixtures, and deconvolution is not limited to the aforementioned three EMs. The approach introduced by Nouvelle et al. [2012] is based on peak ratios and is aimed at (i) overcoming errors associated with baseline change of GC and (ii) enabling production allocation for mixtures with virtually no limitation on the number of EMs. Finally, recent efforts have been directed towards the development of approaches conducive to production allocation without strictly requiring information on EMs. A promising method is based on the Alternating Least Squares (ALS) algorithm [Amendola et al., 2017; Barrie et al., 2020].

In this broad context, the distinctive aim of our study is to introduce an original deconvolution algorithm that (i) makes use of peak ratios, thus providing high flexibility against possible sampling errors caused by baselines changes and/or improper equipment calibration and (ii) does not require relying on synthetic mixtures, thus reducing efforts and resources, in terms of laboratory time and investments. We do so upon framing our methodology within a technical and theoretical assessment of the deconvolution approaches discussed above and including extensions of (a) the method proposed by Nouvelle et al. [2012] and (b) the ALS algorithm, which we view in a stochastic context. All of these approaches are then considered against a suite of new laboratory-based commingling scenarios.

4.3. Methodology

Here, we briefly introduce in Section 4.3.1 two mixing models which are traditionally used for production allocation and rely on peak heights and peak height ratios. We then present an appraisal of key elements of two widely used deconvolution algorithms (Section 4.3.2), including their area of application, advantages, and limitations. Section 4.3.3 is devoted to the introduction of an original deconvolution algorithm that overcomes some of the limitations detected in the traditional approaches. We conclude the analysis by discussing (in Section 4.3.4) the ALS algorithm. The latter is used for the deconvolution of commingled fluids in cases where the absence of information on EMs hampers the applicability of the previously considered approaches. Here, we propose to extend ALS by viewing it in the context of a stochastic (Monte Carlo) framework.

4.3.1. Mixing Models

Approaches to deconvolution in the context of production allocation are grounded on mixing models of either peak heights (or peaks) of a chromatogram or peak height ratios (hereafter termed ratios) evaluated from peaks associated with molecules eluting at close times.

Mixing models associated with peak heights rest on the assumption that peaks in the GC of a mixture are linear combinations of peaks of the GCs associated with each EM according to

$$\underline{\underline{\mathbf{A}}}\mathbf{x} = \mathbf{b} + \varepsilon; \qquad \text{with} \sum_{k=1}^{K} x_k = 1, \tag{4.1}$$

where, $\mathbf{x} = (x_1, ..., x_K)^T$ is a vector containing the (unknown) mass frac-

tions (x_k, $k = 1, ..., K$) of the K EMs in a mixture; $\mathbf{b} = (x_1, ..., x_{N_p})^T$ is a vector whose entries correspond to the N_p peaks of the mixture GC; vector $\varepsilon = (\varepsilon_1, ..., \varepsilon_{N_p})^T$ embeds GC peak measurement errors as well as model errors; and $\underline{\underline{\mathbf{A}}}$ is a N_pxK matrix, whose entry $a_{n,k}$ is the nth peak of the kth EM of the mixture (i.e., column k of matrix $\underline{\underline{\mathbf{A}}}$ contains the N_p peaks detected for the kth EM).

Production allocation methods relying on ratios, R_{ij}, are based on the following formulation [Nouvelle et al., 2012]

$$R_{ij} = \frac{b_i}{b_j}; \qquad \text{with} b_n = \sum_{k=1}^{K} x_k a_{n,k} \frac{m_K}{m_k}; \; n = i, j. \tag{4.2}$$

Here, m_k is the mass of the kth EM injected into the GC device. Note that m_k and x_k are (usually) unknown model parameters, to be estimated by making use of a deconvolution method (as described in Sects 2.2 and 2.3) on the basis of GC peak or ratio data.

4.3.2. Deconvolution Algorithms

McCaffrey algorithm

The deconvolution algorithm proposed by McCaffrey et al. [2011] rests on the peak heights mixing model 4.1 and is characterized by K unknowns (i.e., the elements of vector $\mathbf{x}$). The algorithm renders an estimate of $\mathbf{x}$, $\hat{\mathbf{x}}_{Mc}$, as

$$\hat{\mathbf{x}}_{Mc} = \left(\underline{\underline{\mathbf{A}}}^T \underline{\underline{\mathbf{A}}}\right)^{-1} \underline{\underline{\mathbf{A}}}^T \mathbf{b}. \tag{4.3}$$

Equation 4.3 descends from minimization of the generalized Least Squares (LS) criterion, assuming that elements of ε in 4.1 are zero-mean Gaussian random variables. To improve the accuracy of estimates based on

Equation 4.3, McCaffrey et al. [2011] propose the following procedure:

1. Evaluate 4.3 normalizing $a_{n,k}$ and b_n by $\max a_{n,1}, ..., a_{n,K}$; this enables one to properly consider the information content embedded in each peak value (even for small values of b_n).

2. Determine ε by cross-validation and verify that its entries are characterized by a zero-mean Gaussian distribution and reject from the analysis peaks where entries of ε do not satisfy this condition.

Note that the McCaffrey's deconvolution algorithm corresponds to a least-squares estimation approach. The latter is widely used in several areas such as, e.g., machine learning [Suykens and Vandewalle, 1999], genomics [Johansson et al., 2003], econometrics [Yeniay and GÖKTAŞ, 2002], as well as petroleum engineering (e.g., Bao and Dai [2009]; de Lima Furtado et al. [2021]; Wang et al. [2022]. The main advantage of the algorithm is its conceptual simplicity. However, it might yield unphysical or inaccurate results when GCs of poor quality (in terms of measurement accuracy and reliability) are employed.

Nouvelle Algorithm

Nouvelle et al. [2012] introduce a deconvolution algorithm based on the peak ratios mixing model 4.2. The approach can be applied when (i) GCs of the EMs and of the mixtures to deconvolute as well as (ii) GC of at least one mixture with known EMs mass fractions (hereafter termed as synthetic mixture) are available. The main advantage of this algorithm is that common GC detection errors can be neglected since they can be significantly shadowed by relying on ratios. Additionally, considering that all components in a fluid sample are equally affected by improper storage, the mixture model relying on peak ratios is highly adaptable also to

GC related to samples that have not been properly handled. Nevertheless, the application of the approach requires having at our disposal at least one synthetic mixture. Such a constraint is otherwise not needed by the McCaffrey deconvolution algorithm. This requirement implies, in turn, that efforts associated with laboratory analyses significantly increase, thus potentially limiting its application.

The mixing model 4.2 is characterized by a total of $(2K - 1)$ unknowns. These are subdivided into two groups: (i) the ratios of EM masses injected into the GC device, which form the entries of vector $\mathbf{MR} = (m_k/m_1, ..., m_k/m_{K-1})$; and (ii) the mass fractions of EMs in a mixture, i.e., $\mathbf{x}$. The Nouvelle algorithm is structured according to two steps: First (Step 1), $\mathbf{MR}$ is estimated by making use of available synthetic mixtures; then (Step 2), estimates of $\mathbf{x}$ are provided by relying on $\mathbf{MR}$ determined in Step 1.

Step 1 and Step 2 are performed by minimizing the function, $L()$

$$\begin{aligned} L(\mathbf{MR}, \mathbf{x}|\mathbf{R}^*) &= \sum_{\forall(i,j)} \ln\left[1 + Q^2_{R_{ij}}\left(R^*_{ij} - R_{ij}\right)^2\right] \\ \text{with } Q_{R_{ij}} &= \sqrt{\frac{2}{\sigma^2_{R_{ij}}(\eta_i + \eta_j)}}. \end{aligned} \tag{4.4}$$

Here, $\mathbf{R}^*$ is a vector of components R^*_{ij}, the latter corresponding to the experimentally observed ratio values, R_{ij}, defined in Equation 4.2; and η_i and η_j are the number of times that peaks i and j are used in the set of ratios of N_R elements, respectively. The quantity $\sigma^2_{R_{ij}}$ is the variance of R_{ij}, which is approximated in Nouvelle et al. [2012] as

$$\sigma^2_{R_{ij}} = \frac{1}{\bar{b}_j^2}\left(\sigma_i^2 + \frac{\bar{b}_i^2}{\bar{b}_j^2}(\sigma_j^2\right), \tag{4.5}$$

$\bar{b}_n$ and σ_n^2 (with $n = i,\ j$) being mean and variance of peak n, respectively. According to the procedure highlighted by Nouvelle et al. [2012], $\bar{b}_n$ and σ_n^2 can be estimated on the basis of replicates of laboratory experiments. However, the number of available replicates, N, is usually very limited (typically, N = 3-5). This renders the accuracy and reliability of statistical moments evaluated in such a small ensemble highly questionable (note that the error associated with estimates of mean and variance decreases as N^{-1} and $N^{-0.5}$, respectively).

Nouvelle et al. [2012] derived Equation 4.4 by (i) assuming that measurement errors of peaks i and j can be described through a zero-mean Gaussian distribution, so that R_{ij} follows a Cauchy distribution, and (ii) determining the weighting factor $Q^2_{R_{ij}}$ relying on an approximation of $\sigma^2_{R_{ij}}$ as given by Equation 4.5 instead of considering the scale parameter of the probability density function of R_{ij}. Note also that Equation 4.5 relies on a Taylor expansion of Equation 4.2 truncated at first order. Therefore, it is a good estimate of the variance of R_{ij} only if σ_i^2 and σ_j^2 are small.

Here, we reframe the work of Nouvelle et al. [Nouvelle et al., 2012] within a rigorous Maximum Likelihood, ML, approach. Assuming that R_{ij} follows a Cauchy distribution, ML estimates of $\mathbf{MR}$ and $\mathbf{x}$ are obtained by minimizing the negative log-likelihood function, $NLL()$

$$NLL(\mathbf{MR}, \mathbf{x}|\mathbf{R}^*) = N_R \ln(\pi) + \sum_{\forall(i,j)} \ln\left[1 + \sigma_{ij}^2 \left(R_{ij}^* - R_{ij}\right)^2\right] - \ln(\sigma_{ij})$$

with $\sigma_{ij} = \dfrac{\sigma_j}{\sigma_i}$.

(4.6)

Note that key differences between Equations 4.4 and 4.6 are (i) the weight factor, i.e., $Q_{R_{ij}}^2$ in Equation 4.4 and σ_{ij} in Equation 4.6, and (ii) the additional term, $\ln(\sigma_{ij})$, in 4.6. In the following we assume that $c_v = \sigma_n / b_n^*$ (b_n^* corresponding to the experimental value of peak n) is constant, thus implying that the relative error across peak height measurements is constant. This assumption is consistent with previous studies (e.g., Janetti et al. [2012]) linking contaminant concentration errors to measured concentration values. Thus, Equation 4.6 becomes

$$NLL(\mathbf{MR}, \mathbf{x}|\mathbf{R}^*) = N_R \ln(\pi) + \sum_{\forall(i,j)} \ln(R_{ij}^*) + \ln\left[1 + \left(1 - \frac{R_{ij}}{R_{ij}^*}\right)^2\right].$$

(4.7)

Note that minimization of Equation 4.7 is equivalent to the minimization of its last term. We further note that another possible approach is to consider σ_n as constant, i.e., independent of peak measurements. Results obtained in this case were unsuccessful and are not reported in Section 4.4.

4.3.3. Original Approach and Deconvolution Algorithm

In this Section, we introduce a novel deconvolution algorithm (hereafter termed PGM, after the initials of the authors' institutions) that enables one to overcome limitations associated with the approaches described above while maintaining operational simplicity. Our approach (i) allows the use of the key concept of the ratios mixture model 4.2, i.e., explicitly considering the objective function to be based on the difference between observed and numerically evaluated ratios; (ii) does not rely on synthetic mixtures (as otherwise required by the Nouvelle deconvolution method), thus avoiding an increase in laboratory time (as compared against the McCaffrey algorithm); (iii) is characterized by theoretical foundations that enable one to overcome the assumptions and limitations required by the Nouvelle approach (as detailed in Section 4.3.2 - Nouvelle algorithm), and (iv) does not strictly require (in principle) replicates to obtain estimates of peak measurements variance. Our approach is conducive to estimating x and MR from the mixing model 4.2 as detailed in the following.

We write

$$b_n = b_n^* + \lambda_n; \qquad \text{with } n = i, j, \tag{4.8}$$

where the peak height b_n of a mixture GC is expressed as the sum of the observed value, b_n^*, and a zero-mean measurement error, λ_n, characterized by a Gaussian distribution with (generally unknown) variance σ_n^2. Assuming that peak measurement errors are not correlated, R_{ij} in model 4.2 is a random variable characterized by the following probability density function, pdf,

$$
\begin{aligned}
&p_{R_{ij}}(r_{ij}) = \\
&= \frac{1}{2\pi\sigma_i\sigma_j}\int_{-\infty}^{\infty}|b_j|\exp\left[-\frac{1}{2}\left(\frac{r_{ij}b_j - b_i^*}{\sigma_i}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{b_j - b_j^*}{\sigma_j}\right)^2\right]db_j \\
&= \frac{\sigma_{ij}\exp\left[-\frac{1}{2}\frac{b_j^{*2}}{\sigma_j^2}({R_{ij}^*}^2\sigma_{ij}^2 + 1)\right]}{\pi(1 + r_{ij}^2\sigma_{ij}^2)}(1 + \gamma_{ij})
\end{aligned}
\tag{4.9}
$$

with

$$
\begin{aligned}
\gamma_{ij} &= \sqrt{\pi}\phi_{ij}\exp\phi_{ij}^2\mathrm{erf}(\phi_{ij}); \\
\phi_{ij} &= \frac{b_j^*}{\sigma_j}\frac{1 + r_{ij}R_{ij}^*\sigma_{ij}^2}{\sqrt{2\left(1 + r_{ij}^2\sigma_{ij}^2\right)}}; \\
\text{and } R_{ij}^* &= \frac{b_i^*}{b_j^*}.
\end{aligned}
\tag{4.10}
$$

Considering all available N_R ratios, ML estimates of model parameters (i.e., $\mathbf{x}$, $\mathbf{MR}$, and σ_n^2) can be obtained by minimizing the negative Log-Likelihood criterion, i.e.,

$$\begin{aligned} NLL(\mathbf{MR}, \mathbf{x}, \sigma_n^2 | \mathbf{R}^*) &= -\ln p_{R_{1,2},...,R_{ij}}(r_{1,2}, ..., r_{ij} | \mathbf{x}, \mathbf{MR}, \sigma_1, \sigma_2, ...,) \\ &= -\sum_{\forall(i,j)} \ln p_{R_{ij}}(r_{ij} | \mathbf{MR}, \mathbf{x}, \sigma_i, \sigma_j) \end{aligned} \tag{4.11}$$

Note that the sum in Equation 4.11 considers all of the N_R ratios in the set. Making use of Equation 4.9, Equation 4.11 becomes

$$\begin{aligned} &NLL = \\ &N_R \ln \pi + \sum_{\forall(i,j)} \ln \left(\frac{1 + r_{ij}^2 \sigma_{ij}^2}{\sigma_{ij}} \right) + \frac{1}{2} \frac{b_j^{*2}}{\sigma_j^2} \left(R_{ij}^{*2} \sigma_{ij}^2 + 1 \right) - \ln \left(1 + \gamma_{ij} \right) \end{aligned} \tag{4.12}$$

As in Section 4.3.2, we assume that $c_v = \sigma_n / b_n^*$ is constant across peak height measurements and Equation 4.12 simplifies as

$$\begin{aligned} &NLL = J + N_R \left(\ln \pi + \frac{1}{c_v^2} \right); \\ &\text{with } J = \sum_{\forall(i,j)} \ln R_{ij}^* + \ln \left(1 \frac{r_{ij}^2}{R_{ij}^{*\,2}} \right) - \ln(1 + \upsilon_{ij}) \end{aligned} \tag{4.13}$$

where

$$v_{ij} = \sqrt{\pi}\omega_{ij} \exp \omega_{ij}^2 \text{erf}(\omega_{ij}); \qquad \text{and } \omega_{ij} = \frac{1 + r_{ij}/R_{ij}^*}{\sqrt{2(1 + r_{ij}^2/R_{ij}^{*\,2})}}. \quad (4.14)$$

Parameters embedded in Equation 4.13 include $\mathbf{x}$, $\mathbf{MR}$, and c_v. If c_v is known, minimization of NLL (Equation 4.13) coincides with minimization of J. If c_v is unknown, its ML estimate can be obtained according to

$$\frac{\partial NLL}{\partial c_v} = -\frac{2N_R}{c_v^3} + \frac{1}{c_v} \sum_{\forall(i,j)} \frac{1}{(1 + v_{ij})} \left(2\omega_{ij}^2 + 2\omega_{ij}^2 v_{ij} + v_{ij}\right) = 0, \quad (4.15)$$

One can then evaluate c_v by solving the following implicit equation

$$c_v^2 = \frac{1}{N_R} \sum_{\forall(i,j)} \omega_{ij}^2 + \frac{v_{ij}}{2(+v_{ij})}. \quad (4.16)$$

Here, we propose to obtain ML estimates of $\mathbf{x}$, $\mathbf{MR}$, and c_v (denoted as $\hat{\mathbf{x}}_{PGM}$, $\hat{\mathbf{MR}}_{PGM}$ and $\hat{c_v}_{PGM}$, respectively) according to the procedure highlighted in the flowchart depicted in Figure 4.1. The latter shows that the procedure requires to (i) initialize $\mathbf{x}$ and $\mathbf{MR}$; (ii) compute the ratios of the mixture chromatogram using Eq 4.2; (iii) initialize the value of c_v and minimize J in Equation 4.13 to compute the ML estimates $\hat{\mathbf{x}}_{PGM}$ and $\hat{\mathbf{MR}}_{PGM}$; and (iv) make use of Equation 4.16 to evaluate $\hat{c_v}_{PGM}$. The procedure ends when a convergence criterion (e.g., $|c_v - \hat{c_v}_{PGM}|/c_v < \delta_0$) is satisfied. Note that δ_0 is a threshold value that must be defined at the beginning of the workflow. In our test case, we set $\delta_0 = 0.01$. Moreover,

we note that in our analyses the same mass of EMs samples is employed during GC experiments. As such, the ratios of EM masses injected into the chromatography device (corresponding to the entries of vector **MR**) are constrained to the range 0.95-1.05 during the optimization procedure.

4.3.4. Alternating Least Squares Algorithm, ALS

All deconvolution methods described in Sections 4.3.2 and 4.3.3 allow estimating the mass fraction of EMs in a mixture when chromatograms of EMs (i.e., entries of matrix $\underline{\underline{\mathbf{A}}}$) are known. The ALS deconvolution algorithm tackles a fundamentally different production allocation problem in which entries of matrix $\underline{\underline{\mathbf{A}}}$ are not known. Nevertheless, the application of ALS requires the analysis of multiple virtual mixtures (the number of which must be greater than the number of EMs) that need to be subject to the deconvolution simultaneously. To provide a proper representation of the overall system variability, these virtual mixtures must be associated with different mass fraction compositions of EMs.

As ALS relies on multiple mixtures, vectors $\mathbf{x}$ and $\mathbf{b}$ in model 4.1 are now matrices. These are hereafter denoted as $\underline{\underline{\mathbf{x}}}$ and $\underline{\underline{\mathbf{b}}}$ of size $(K \times N_M)$ and $(N_p \times N_M)$, respectively (N_M being the number of mixtures being deconvoluted simultaneously).

The first step in an ALS-based production allocation relies on estimating the number of EMs, K, making use of one of the following methodologies (or a combination thereof):

(i) perform a principal component analysis, PCA (or a Singular Value Decomposition - SVD), of the mixture GCs included in $\underline{\underline{\mathbf{b}}}$ and evaluate the minimum number of components required to explain the variance of $\underline{\underline{\mathbf{b}}}$ [Brunton and Kutz, 2022; Jaumot et al., 2015];

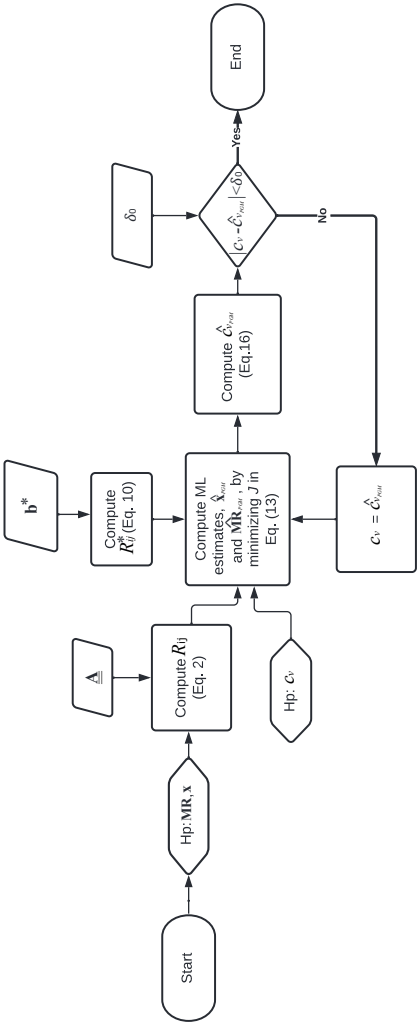

Figure 4.1: Flowchart of the PGM deconvolution algorithm.

(ii) make use of information about the natural system (e.g., the geological structure) that could assist in constraining K.

It is noted that, when the assessment of K is not unambiguous, the deconvolution process should be performed several times, upon varying K and analyzing the ability of the estimated EM spectra and mass fractions to describe $\underline{\underline{\mathbf{b}}}$ [Jaumot et al., 2015].

Once K is defined, the ALS deconvolution algorithm rests on an iterative procedure to determine a Least Square (LS) estimate of $\underline{\underline{\mathbf{A}}}$ and $\underline{\underline{\mathbf{x}}}$ according to the following workflow:

Step 1 Initialize $\underline{\underline{\mathbf{A}}}$.

Step 2 Determine an LS estimate of $\underline{\underline{\mathbf{x}}}$ as

$$\hat{\underline{\underline{\mathbf{x}}}}_{ALS} = \left(\underline{\underline{\mathbf{A}}}^T \underline{\underline{\mathbf{A}}}\right)^{-1} \underline{\underline{\mathbf{A}}}^T \underline{\underline{\mathbf{b}}}. \tag{4.17}$$

Step 3 Determine an LS estimate of $\underline{\underline{\mathbf{A}}}$ as

$$\hat{\underline{\underline{\mathbf{A}}}}_{ALS} = \underline{\underline{\mathbf{b}}}\, \hat{\underline{\underline{\mathbf{x}}}}^T_{ALS} (\hat{\underline{\underline{\mathbf{x}}}}_{ALS} \hat{\underline{\underline{\mathbf{x}}}}^T_{ALS})^{-1}; \text{ with entries } (a_{ALS})_{n,k} > 0. \tag{4.18}$$

Step 4 If $|a_{n,k} - (a_{ALS})_{n,k}|/a_{n,k} < \delta_0$, then stop; otherwise, set $\underline{\underline{\mathbf{A}}} = \hat{\underline{\underline{\mathbf{A}}}}_{ALS}$ and go to step 2.

Note that, since GCs with negative entries have no physical meaning, entries of $\hat{\underline{\underline{\mathbf{A}}}}_{ALS}$ (Equation 4.18) are constrained to be positive. Note also that N_M must be larger than (or equal to) K, to guarantee that the system is not under-constrained.

The workflow described above must be repeated multiple times with different initializations of $\underline{\underline{\mathbf{A}}}$ (Step 1) to avoid entrapment in local minima

of the objective function to be minimized [Llinares et al., 2012]. We note that, since Equation 4.17 for the evaluation of $\hat{\underline{\underline{\mathbf{x}}}}_{ALS}$ is coupled with 4.18, it is possible that the global minimum value of the objective function is not necessarily associated with the optimal $\underline{\underline{\mathbf{x}}}$ (i.e., associated with the minimum LS value) due to the action of $\hat{\underline{\underline{\mathbf{A}}}}_{ALS}$ in 4.17. To overcome this issue, we propose an original view and frame the approach within a probabilistic Monte Carlo setting. We do so by relying on multiple realizations. Each one of these corresponds to a combination of the N replicates associated with GC performed for each mixture (to be deconvoluted simultaneously). To the best of our knowledge, this is the first study exploring the potential of the ALS deconvolution algorithm in the context of a production allocation scenario under such a probabilistic framework.

4.3.5. Experimental Setup

The analysis of the relative skills of the deconvolution algorithms described in Sections 4.3.2 - 4.3.4 to assist production allocation is assessed upon considering a set of ten laboratory-based mixtures produced by commingling three EMs as listed in Table 4.1. These mixtures have been selected with the aim of reproducing at the laboratory scale typical commingling scenarios associated with field settings.

Oil sample	EM1 x_1[%]	EM2 x_2[%]	EM3 x_3[%]
M1	33.3	33.3	33.3
M2	70	15	15
M3	10	70	20
M4	20	20	60
M5	50	30	20
M6	20	40	40
M7	45	55	0
M8	5	10	85
M9	85	10	5
M10	0	90	10

Table 4.1: Mass fractions of the three EMs in the 10 laboratory-based mixtures.

Materials

All solvents, including dichloromethane, are of analytical grade. These, as well as ethylbenzene-d10 (used as an internal standard for quantification purposes), were purchased from Merck (Darmstadt, Germany). An alkylbenzene standard mixture containing 37 molecules and used for identification, method development, and evaluation of response factors associated with the internal standard was purchased from Restek Corporation (Bellefonte, United States). Oil mixtures and EMs were properly stored in a fridge at a temperature of 4 °C to minimize potential alterations due to improper handling of the samples.

Sample preparation and GC-MS analysis

In this study, chromatograms are recovered using GC-MS since (i) GC-MS requires a simpler and more accessible instrumentation and allows for a more selective analysis of alkylbenzenes than MDGC, especially if high molecular weight molecules need to be assessed; (ii) GC-MS analyses are faster than their counterparts based on MDGC or HRGC; (iii) selectivity of GC-MS allows filtering out all signals related to non-alkylbenzene species which are typically observed in oil samples, replacing the need for a double separative column (which is otherwise required for MDGC); (iv) GC-MS allows monitoring additional molecule classes with respect to MDGC (e.g., diamondoids, alkylnaphtalenes, dibenzothiophenes, polycyclic aromatic hydrocarbons) further increasing the number of geochemical parameters that can be obtained for fingerprinting or other applications; and (v) GC-MS yields a complete baseline separation of the 11 alkylbenzenes analyzed by MDGC and extends the analytical range up to C_{12}-alkylbenzenes, thus enabling one to analyze 50-80 molecules.

EM samples are weighed and dissolved in dichloromethane to a concentration of 10 mg/mL and used to produce the 10 laboratory mixtures illustrated in Table 4.1. All EMs samples and mixtures are then analyzed 5 times through GC-MS by targeting the alkylbenzene components in the molecular range C_8-C_{12}. Fixed amounts of ethylbenzene-d10 are added to each sample to assist quantification of the various alkylbenzenes. This step is performed by applying the internal standard method, using peak heights and response factors evaluated from a standard alkylbenzene mixture of 37 molecules. Since oil samples contain a number of alkylbenzenes that is significantly higher than what is available in the standard mixture, response factors equal to those of the most closely eluting alkylbenzene available as standard are assigned to such molecules.

The analysis of alkylbenzene molecules in the oil samples is performed via gas chromatography-single quadrupole mass spectrometry. Helium is used as carrier gas and the injections are performed in split mode. The analytical separation is carried out using a Stabilwax capillary column (Restek - 60 m $\times$ 0.32 mm $\times$ 0.25 μm) in temperature gradient mode. The eluted molecules are ionized within the electron ionization source of the mass spectrometer, which operates at 70 eV and 250 °C source temperature.

The MS analyzer is operated in full scan mode only in the early stages of method development. This yields spectral data for the identification and quantification of alkylbenzene components in the standard mixture. Otherwise, sample analyses are conducted in SIM (Selected Ion Monitoring) mode. Mass-to-charge ratios (m/z) for SIM acquisition are: 91, 105, 106, 116, 119, 120, 133, 134, 147, 148, 162. A quantitative analysis is performed for each peak using the measured heights in the GC.

4.4. Results and Discussion

4.4.1. Available data

In our study, the coefficient of variation of component measurements is generally lower than 5% and never exceeds 10%. As an example, Figure 4.2 depicts the chromatogram of the five replicates associated with mixture M1. These experimental results suggest an overall high degree of repeatability of the experimental analyses. The remaining mixtures and EMs display a similar quality of repeatability (details not shown). One can see that peak responses vary across two orders of magnitude, the largest values being associated with the first 11 peaks (i.e., those related to C_9-alkylbenzenes).

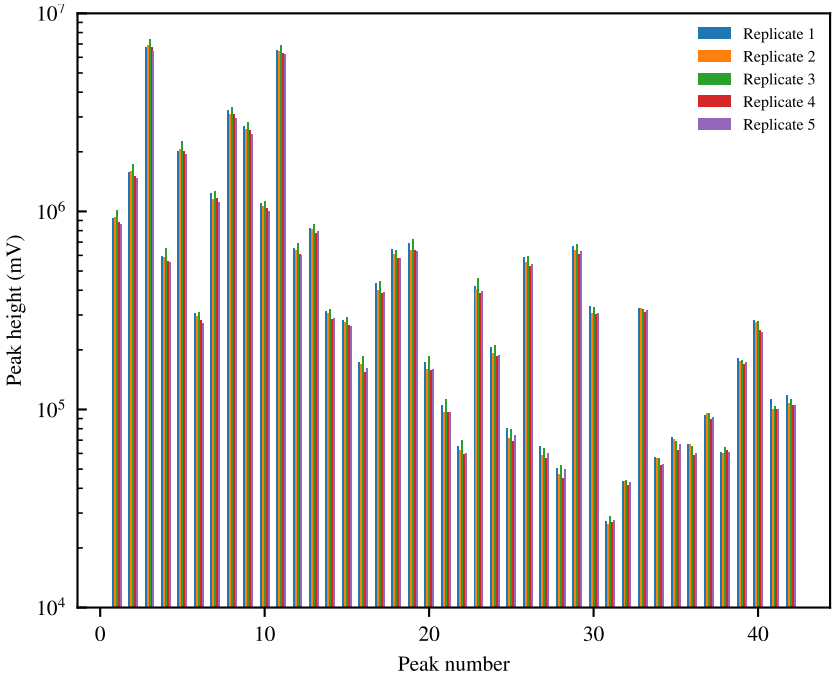

Figure 4.2: Chromatograms resulting from 5 replicates (Mixture M1).

4.4.2. Deconvolution

Here, we present a quantitative comparison of the accuracy of the various deconvolution algorithms illustrated in Section 4.3 on the basis of the laboratory dataset detailed in Section 4.4.1. For this purpose, we compute estimates of $\mathbf{x}$, $\mathbf{x}_\xi$, using (i) Equation 4.3 when $\xi = Mc$ (McCaffrey algorithm), (ii) Equation 4.4 when $\xi = Nv$ (Nouvelle algorithm), (iii) Equation 4.7 when $\xi = MNv$ (modified Nouvelle algorithm), (iv) Equations 4.13-4.16 when $\xi = PGM$, and (v) Equations 4.17-4.18 when

$\xi = ALS$. For the McCaffrey and PGM algorithms no synthetic mixtures are required. Thus, estimates $\mathbf{x}_\xi$ for each mixture have been evaluated considering all combinations of the N replicates of the given mixture and associated EMs chromatograms. This yields N^{K+1} values of $\hat{\mathbf{x}}_\xi$. On the other hand, for the ALS algorithm (where N_M is required to be larger than K) a subset of the N^{N_M} possible estimates has been randomly selected. With reference to the Nouvelle algorithm (where at least one synthetic mixture is required), we explore the goodness of the deconvolution algorithm by varying the number of synthetic mixtures, N_{SM}, from 1 to 9.

The mean associated with estimates $\hat{\mathbf{x}}_\xi$ (i.e., $\bar{\mathbf{x}}_\xi$) is then evaluated upon considering that the available data are compositional vectors (whose components express proportions or percent amount of a whole). We then leverage on the theoretical framework underlying Compositional Data Analysis (CoDa; e.g., Menafoglio et al. [2014, 2021]; Van den Boogaart and Tolosana-Delgado [2013] and references therein). For comparison purposes, we also provide an estimate of $\mathbf{x}_\xi$ (denoted $\bar{\bar{\mathbf{x}}}_\xi$) by averaging EMs and mixture replicates before implementing a deconvolution algorithm.

Finally, we assess the performance of each deconvolution method by computing the mean absolute error, MAE_ξ, and the mean absolute percentage error, $MAPE_\xi$, obtained with deconvolution method ξ and defined as

$$MAE_\xi = \frac{1}{K}\sum_{k=1}^{K}|x_k^* - \bar{\bar{x}}_{k,\xi}| \qquad \text{and } MAPE_\xi = \frac{100}{K}\sum_{\substack{k=1 \\ x_k \neq 0}}^{K}\left|\frac{x_k^* - \bar{\bar{x}}_{k,\xi}}{x_k^*}\right|. \tag{4.19}$$

Here, x_k^* is the true value associated with the kth EM in a mixture and $\bar{\bar{x}}_{k,\xi}$ is the kth element of $\bar{\bar{\mathbf{x}}}_\xi$. Note that by making use of $\bar{\mathbf{x}}_\xi$ instead of $\bar{\bar{\mathbf{x}}}_\xi$ in 4.19 we obtained almost identical results (details not shown).

McCaffrey Deconvolution Method

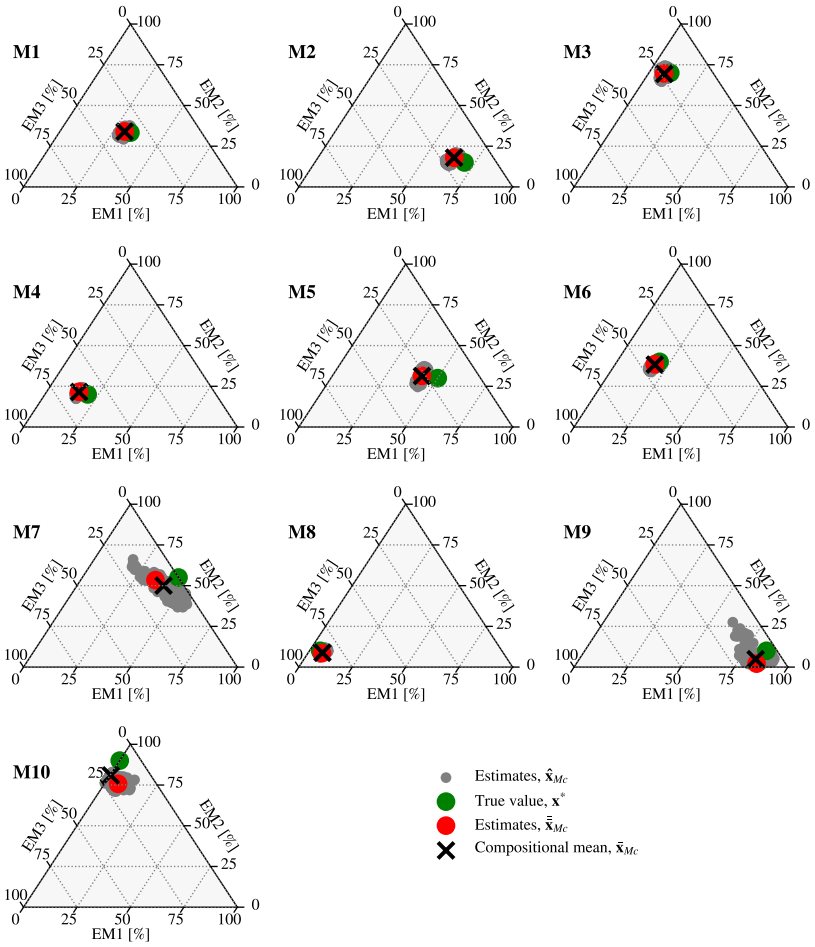

Figure 4.3: Results of the production allocation approach obtained through the McCaffrey deconvolution algorithm.

Figure 4.3 provides ternary diagrams of the N^{K+1} estimates $\hat{\mathbf{x}}_{Mc}$ (gray symbols). The variability of $\hat{\mathbf{x}}_{Mc}$ is modest, samples M7, M9, and M10 being the only exceptions. Note that M7 and M10 are formed by only two EMs (i.e., $K = 2$). The compositional mean associated with these estimates, $\bar{\mathbf{x}}_{Mc}$ (black cross), experimental values, $\mathbf{x}^*$ (green circle), as well as values of the deconvolution obtained by averaging EMs and mixture GC replicates before the use of the deconvolution algorithm, $\bar{\bar{\mathbf{x}}}_{Mc}$, are also included in Figure 4.3. Quantities $\bar{\mathbf{x}}_{Mc}$ and $\bar{\bar{\mathbf{x}}}_{Mc}$ are seen to properly represent the overall behavior of the EMs mass fractions for all of the mixtures tested. Across all mixtures, the average MAE_{Mc} (Equation 4.19) is 4.3% (its corresponding median being equal to 3%), and the average $MAPE_{Mc}$ is equal to 22.9% (median being equal to 11.9%). Figure 4.4 depicts histograms of Aitchison distances between measured mass fractions $\mathbf{x}^*$ and their estimated counterparts $\hat{\mathbf{x}}_{Mc}$. These results confirm quantitatively the observations made to Figure 4.3 according to which the largest variabilities of the estimates are associated with mixtures M7, M9, and M10, these mixtures being also associated with the largest mean estimation errors.

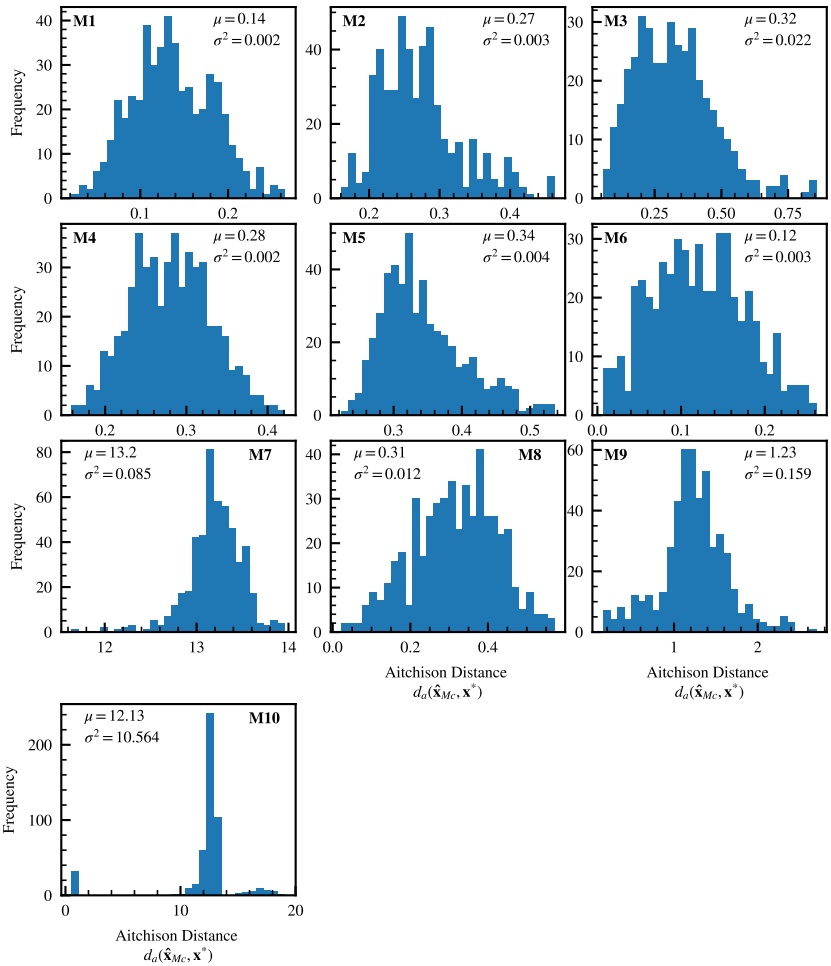

Figure 4.4: Histograms of the Aitchison distances between measured mass fractions $\mathbf{x}^*$ and their estimated counterparts $\hat{\mathbf{x}}_{Mc}$. Values of the resulting mean (μ) and variance (σ^2) are listed as reference.

Nouvelle Deconvolution Approach

Here, we analyze four test cases (C1, C2, C3, and C4; see Table 4.2). Test cases C1 and C3 consider ratios of two consecutive peaks (e.g., $b_1/b_2, b_2/b_3, \dots$) and two formulations of the objective function employed in the procedure, corresponding to the Nouvelle et al. [2012] deconvolution method (Equation 4.4) and our suggested modification (Equation 4.7), respectively. Then, in test cases C2 and C4 we explore the benefit of relying on ratios evaluated upon using more than two peaks (e.g., $b_1/(b_2 + b_3)$) for the deconvolution procedure.

Oil sample	Ratios of two peaks	Ratios of more than two peaks
Equation 4.4	C1	C2
Equation 4.7	C3	C4

Table 4.2: Scenarios used for the assessment of the Nouvelle-based deconvolution approach..

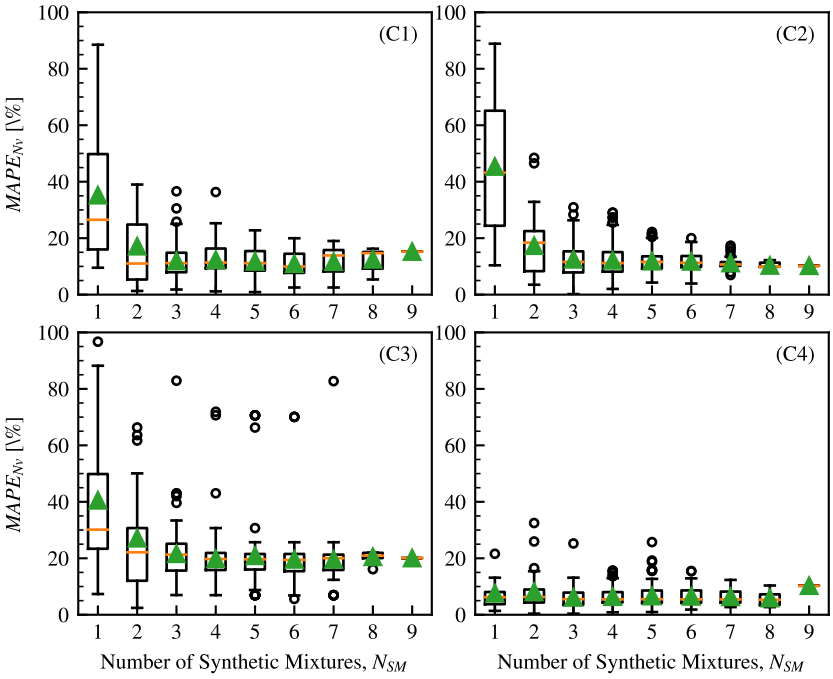

Figure 4.5: Box plots of $MAPE_{N_v}$ for mixture M3 versus the number of synthetic mixtures used in the algorithm N_{SM}. Green triangles and orange lines represent the median and the mean of the distribution, respectively.

Figure 4.5 depicts box plots of $MAPE_{N_v}$ values for mixture M3 (the remaining tested mixtures exhibit a similar behavior; details not shown) for an increasing number of the synthetic mixtures, N_{SM}, used for the deconvolution algorithm. Mean and median values of $MAPE_{N_v}$ are also included. Note that results plotted in Figure 4.5 include all possible combinations of synthetic mixtures (i.e., 9, 36, 84, 126, 126, 84, 36, 9,

and 1 combinations of synthetic mixtures when $N_{SM} = 1, 2, \ldots, 9$, respectively).

As expected, increasing N_{SM} tends to yield errors which are smaller and associated with reduced variability in all test cases. These errors are concentrated in the 5%-25% range for C4. In general, when a sufficient number of synthetic mixtures is available (at least 3 according to our analyses), all test cases are associated with production allocation estimates that are as accurate as those stemming from the McCaffrey deconvolution algorithm.

When the set formed by ratios between two consecutive peaks is used (i.e., scenarios C1 and C3), similar results are obtained employing the deconvolution algorithm proposed by Nouvelle et al. [2012] and our proposed modification (i.e., Equation 4.7). Otherwise, if the set involving ratios of more than two peaks is considered (i.e., scenarios C2 and C4), results of enhanced accuracy are achieved upon relying on the proposed objective function given by Equation 4.7 than on the original Nouvelle algorithm. Note that case C4 is characterized by the overall best accuracy of the results even when only one synthetic mixture is available. This suggests that by relying on the objective function we propose (Equation 4.7) one can potentially reduce the laboratory efforts associated with the preparation and analysis of synthetic mixtures whilst the precision of the method is enhanced. This is an important observation, as time constraints can limit the number of synthetic mixtures available in practical production allocation applications.

Original PGM Approach and Algorithm

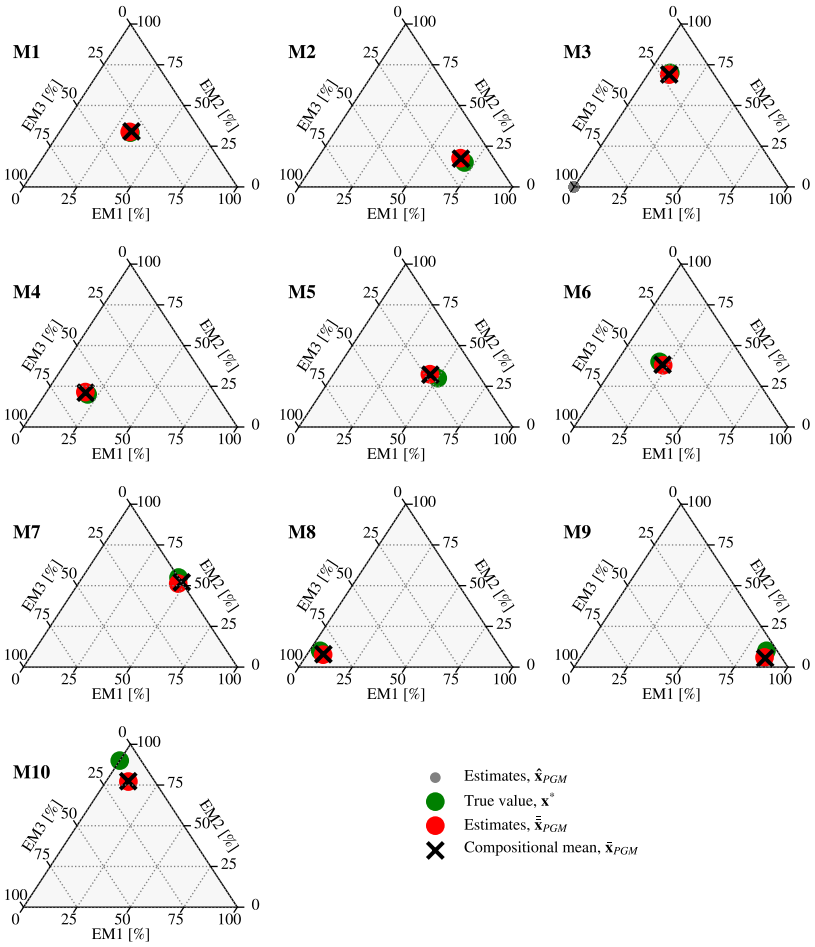

Figure 4.6: Results of the production allocation approach obtained through our original PGM approach.

Figure 4.6 provides ternary diagrams of the N^{K+1} estimates $\hat{\mathbf{x}}_{PGM}$ (gray symbols). Figure 4.6 also includes the compositional mean of $\hat{\mathbf{x}}_{PGM}$, $\bar{\mathbf{x}}_{PGM}$, experimental values, $\mathbf{x}^*$, as well as values of the deconvolution obtained by averaging EMs and mixture GC replicates before the use of the deconvolution algorithm, $\bar{\bar{\mathbf{x}}}_{PGM}$. Similar to Figure 4.3, values of $\bar{\mathbf{x}}_{PGM}$ and $\bar{\bar{\mathbf{x}}}_{PGM}$ are close to $\mathbf{x}^*$. Notably, also mixtures M7, M9, and M10 are associated with small estimation errors (as opposed to what is noted in Figure 4.3). Across all mixtures, the average MAE_{PGM} (Equation 4.19) is only 2.5% (its corresponding median being equal to 2.1%) and the average $MAPE_{PGM}$ is equal to 11.7% (its corresponding median being equal to 6.0%). The variability of the individual estimates is modest and significantly smaller than the one displayed in Figure 4.3. Figure 4.7 presents histograms of the Aitchison distances between measured mass fractions $\mathbf{x}^*$ and $\hat{\mathbf{x}}_{PGM}$. It is noted that the mean of the Aitchison distances rendered by our original PGM deconvolution algorithm is significantly smaller than its counterpart related to the McCaffrey deconvolution algorithm for almost all mixtures analyzed. Mixtures M7, M9, and M10 are characterized by the largest values of the Aitchison distance. This is similar to what has been documented for the results of the McCaffrey deconvolution algorithm, even as a reduced variability can be observed here.

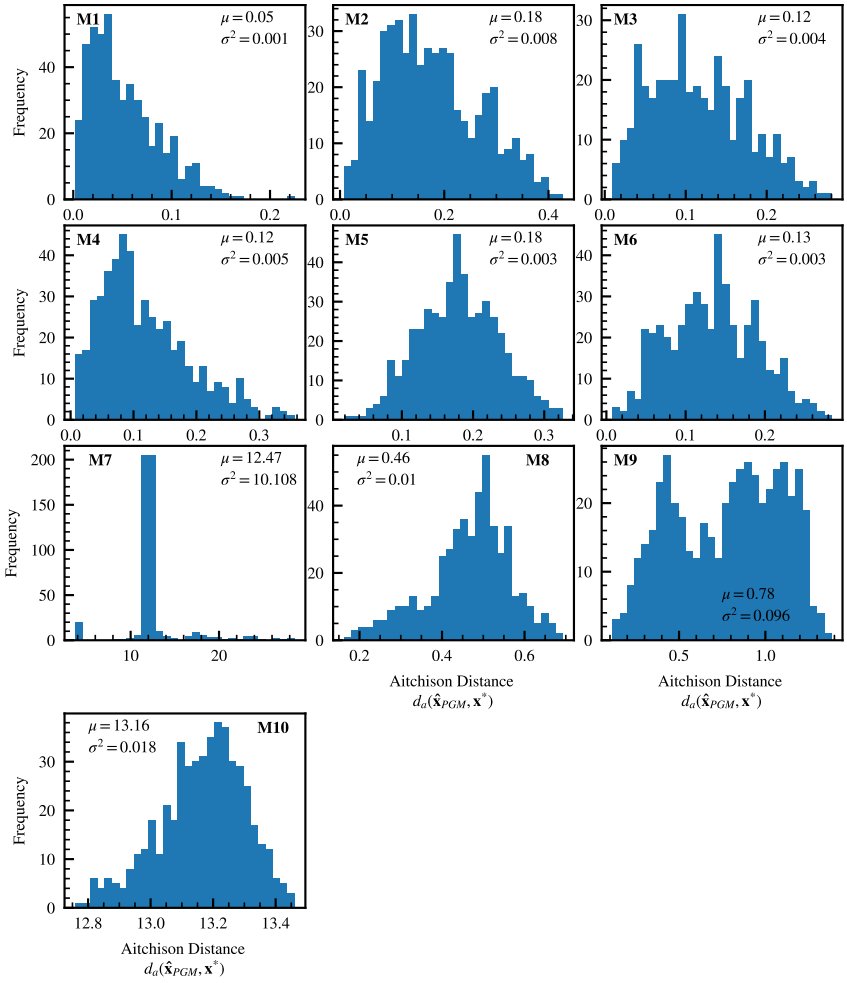

Figure 4.7: Histograms of the Aitchison distances between measured mass fractions $\mathbf{x}^*$ and their estimated counterparts $\hat{\mathbf{x}}_{PGM}$. Values of the resulting mean (μ) and variance (σ^2) are listed as reference.

4.4.3. Production Allocation Without Knowledge of End Members and Use of the ALS Algorithm

In cases where some (or all) of the chromatograms of EMs are not available and the use of the approaches and algorithms illustrated in Sections 4.3.2 and 4.3.3 is hampered, production allocation can be performed upon relying on the ALS deconvolution algorithm as described in Section 4.3.4.

To evaluate the number of EMs, K, we perform an SVD of the mixture GCs included in $\underline{\underline{\mathbf{b}}}$. Our results reveal that considering 1, 2, and 3 EMs can explain 78.9%, 96.8%, and 99.4% of the variance of $\underline{\underline{\mathbf{b}}}$, respectively (details not shown). The selection of three EMs is therefore well justified by our data, which are indeed formed by two or three EMs (see Table 4.1).

We explore the effect of the size S of a subset of the N^{N_M} possible combinations of replicates of mixtures' GC on the stability and accuracy of the estimator $\bar{\underline{\underline{\mathbf{x}}}}_{ALS}$ by plotting in Figure 4.8 the quantity $\frac{1}{K}\sum_{k=1}^{K}|x_k^* - \bar{x}_{k,ALS}|$ versus S ($K = 3$, except for M7 and M10 where $K = 2$). Our results suggest that (a) the compositional mean of the estimates, $\bar{\underline{\underline{\mathbf{x}}}}_{ALS}$, as well as the error between $\bar{\mathbf{x}}_{ALS}$ and $\mathbf{x}^*$ tend to stabilize by increasing the number of realizations; and (b) relying on about 1000 realizations yields stable results of the quantities of interest.

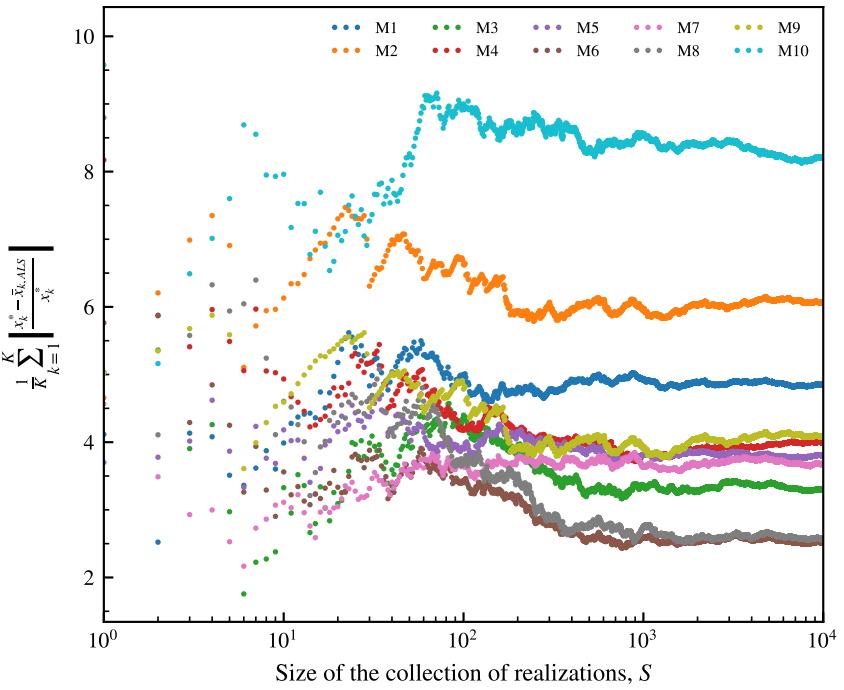

Figure 4.8: Evolution of $\frac{1}{K}\sum_{k=1}^{K}|x_k^* - \bar{x}_{k,ALS}|$ for increasing size of the collection of realizations S used for the evaluation of the compositional mean in the ALS algorithm.

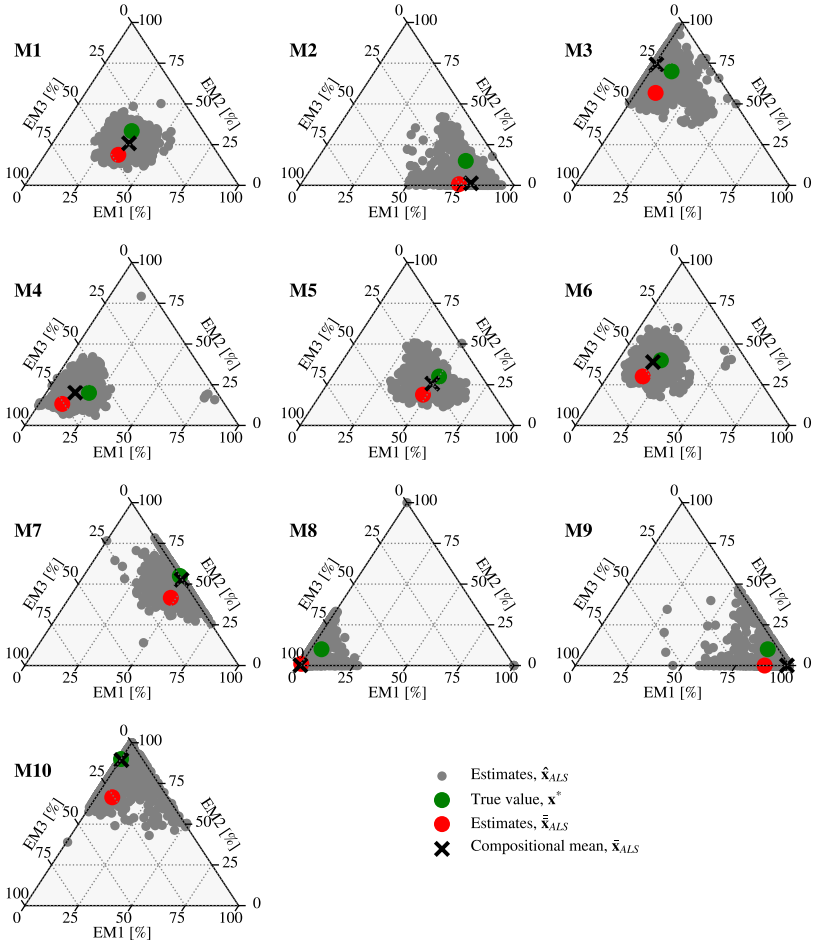

Figure 4.9: Results of the production allocation approach obtained through ALS approach.

Figure 4.9 depicts estimates $\underset{=}{\hat{\mathbf{x}}}_{ALS}$ associated with $S = 1000$ random

combinations of the mixtures' GC replicates. Each combination $\underline{\underline{\mathbf{A}}}$ is randomly initialized 10^3 times. The best one (i.e., the one that minimizes Equation 4.17) is selected and plotted in Figure 4.9 for each combination (gray symbols). The compositional mean associated with estimates $\bar{\underline{\underline{\mathbf{x}}}}_{ALS}$ (black cross), experimental values $\underline{\underline{\mathbf{x}}}^*$ (green circle) as well as values of the deconvolution obtained by averaging mixtures' GC replicates before the use of the deconvolution algorithm, $\bar{\bar{\underline{\underline{\mathbf{x}}}}}_{ALS}$ (red circle), are also included in Figure 4.9. Values of $\bar{\underline{\underline{\mathbf{x}}}}_{ALS}$ and $\bar{\bar{\underline{\underline{\mathbf{x}}}}}_{ALS}$ are close to the true values of mass fractions of EMs in each mixture, although with reduced accuracy when compared against results of deconvolution algorithms based on EMs' chromatograms. The average MAE_{ALS} across mixtures (Equation 4.19) is 9.8% (its corresponding median being equal to 10.2%) and the average $MAPE_{PGM}$ is 46.4% (the median being equal to 33.8%). Figure 4.10 depicts histograms of the Aitchison distances between measured mass fractions $\mathbf{x}^*$ and rows of $\hat{\underline{\underline{\mathbf{x}}}}_{ALS}$. Our results indicate that mean Aitchison distances associated with the results rendered by the ALS algorithm are in general larger than their counterparts stemming from the McCaffrey and PGM algorithms. The largest distances are associated with mixtures M9 and M8.

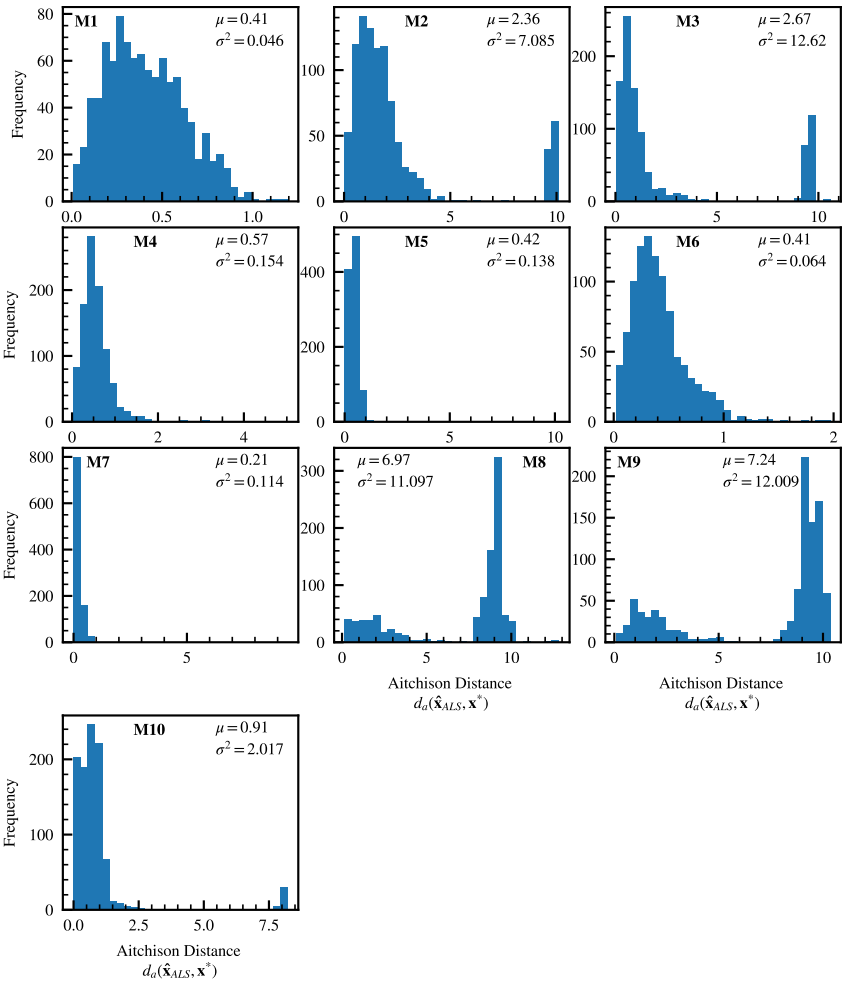

Figure 4.10: Histograms of the Aitchison distances between measured mass fractions $\underline{\underline{\mathbf{x}}}^*$ and rows of $\hat{\underline{\underline{\mathbf{x}}}}_{ALS}$. Values of the resulting mean (μ) and variance (σ^2) are listed as reference.

4.4.4. Performance of the analyzed deconvolution algorithms

Error metric	Feature setting	McCaffrey	Nouvelle based on Eq. 4.4	Nouvelle based on Eq. 4.7	Original PGM Approach	ALS
MAE_ξ	11 Peaks (39 ratios)	8.5%	20.9%	6.7%	10%	11.7%
MAE_ξ	41 Peaks (189 ratios)	2.0%	11.6%	0.8%	0.7%	9.4%
$MAPE_\xi$	11 Peaks (39 ratios)	43.9%	109.2%	49%	48.7%	53.8%
$MAPE_\xi$	41 Peaks (189 ratios)	14.5%	75.8%	5.6%	2.3%	32.6%

Table 4.3: Mean absolute error, MAE_ξ, and mean absolute percentual error, $MAPE_\xi$, of the deconvolution algorithms. The algorithms are implemented for two diverse numbers of features of the mixtures and of the EMs.

Table 4.3 lists values of the MAE_ξ and $MAPE_\xi$ metrics related to M3 and associated with production allocation estimates obtained through all deconvolution algorithms discussed in this study. The importance of the number of features (i.e., peaks) used by the deconvolution algorithms is also tested through the comparison of the resulting allocation errors by employing (i) the entire available set of 41 peaks of the mixture GCs and (ii) a subset of the first 11 peaks from the original dataset. Note that (a) the Nouvelle deconvolution algorithm (implemented by considering the original formulation, Equation 4.4, or our proposed modification, Equation 4.7), is applied using M1 as synthetic mixture and employing ratios encompassing more than two peaks; and (b) the ALS deconvolution algorithm is initialized for a total of 10^3 times.

The accuracy of the estimates tends to increase with the number of fea-

tures (either peak height or peak ratios) employed. Thus, the additional information content carried by considering various peaks (which might include enhanced information on molecular differences between oils or mitigate the effect of measurement errors) can be beneficial to enhance the accuracy of production allocation estimates. Therefore, extending the target alkylbenzene molecular range to C_{12}-species with the experimental procedure illustrated in this study leads to enhanced robustness and accuracy of all deconvolution methods.

In general, our proposed original approach, as well as our reformulation of the Nouvelle algorithm, are characterized by an improved performance (in terms of the metrics considered in this study) when compared against the traditionally employed deconvolution algorithms.

4.5. Conclusions

We introduce an original deconvolution approach for production allocation to enable effective assessment of the diverse oil types forming a mixture originating from the common practice of commingling oils associated with diverse reservoirs, wells, and/or fields. Our original approach (which we term PGM) (a) is inspired by methods resting on peak ratios and (b) does not require relying on synthetic mixtures, thus being potentially associated with reduced laboratory analyses efforts. The approach is framed in the context of typically used deconvolution algorithms, i.e., the algorithm proposed by McCaffrey et al. [2011], the method of Nouvelle et al. [2012], as well as the approach based on the Alternating Least Square (ALS) algorithm. We also present extensions of (a) the method proposed by Nouvelle et al. [2012] and (b) the ALS algorithm, which we view in a stochastic context, corresponding to a Monte Carlo framework, with the aim of improving their robustness and reliability.

The potential of the new PGM approach is shown together with an assessment of the other analyzed deconvolution algorithms against a suite of new laboratory-based three-oil commingling scenarios. These are based on the design and introduction of a novel and low-cost experimental approach. The latter rests on a direct quantitative determination of C_8-C_{12} alkylbenzene components in oil through GC-MS fingerprinting and has been developed to circumvent some limitations of the typically employed methodologies.

Results of the analyses of the controlled experiments provide a unique, comprehensive, and rigorous comparison of the traditional production allocation deconvolution algorithms and highlight the benefit of our extensions to these and of the new PGM approach and algorithm. Our study documents that the number of features used during a quantitative deconvolution is critical to enhance the accuracy of the procedure. Additionally, we found that our new PGM approach is the most accurate methodology, followed by the Nouvelle algorithm based on our modified objective function (Equation 4.7).

5 | Conclusions and future perspectives

In this Doctoral Thesis, we study and discuss, from a probabilistic standpoint, three critical topics associated with the energy sector, particularly pertaining to the extraction and management of subterranean energy resources. These topics are: (i) the utilization of robust sensitivity analysis techniques to diagnose models representing the flow of gases through low-permeability materials; (ii) the application of a methodology for conducting probabilistic assessments of groundwater contamination associated with unconventional energy resource production; and (iii) the development of an innovative deconvolution approach, firmly grounded in probabilistic methodologies, aimed at oil production allocation. The main contributions of this work are:

- We showed that the uncertainty linked to gas flow in low permeable materials is mainly affected by the imperfect knowledge of reference pore radius. Subsequently, in decreasing order of significance, the uncertainty of gas flow is primarily governed by uncertainties in reference porosity, pore pressure, tortuosity, temperature, and, to a lesser extent, the blockage migration ratio of adsorbed molecules. It is noteworthy that the remaining parameters within our model

have a negligible impact.

The results of our study also suggest that slip flow has a significantly stronger contribution to the overall gas migration in comparison with the other two migration mechanisms. Actually, Knudsen diffusion is negligible in all of the analyzed test cases. These findings are expected to provide valuable guidance for the allocation of resources during experimental activities aimed at characterizing gas flow in caprocks.

- We present a novel methodology for the probabilistic assessment of groundwater contamination associated with unconventional energy resource production. This methodology is applied to a case study in which two contamination scenarios are considered. The first one quantifies the mass of contaminating fluids displaced from the source rock into the overburden, while the second one quantifies the mass of contaminants reaching a groundwater body after migrating through a preferential pathway. These scenarios are connected, meaning that the outputs of the first scenario are considered as inputs of the second one. The methodology allows to quantify the probability distribution of groundwater contamination as well as the influence of the analyzed uncertain model parameters on the probability distribution of contaminants displaced.

  The results of our study case show that the probability of groundwater contamination is primarily dictated by the injection duration of hydraulic fracturing operations, as well as the permeability of: (i) the fractures generated within the source rock and (ii) the preferential pathway. Additionally, our analysis shows the influence of the analyzed uncertain model parameters on the conditional mean and variance of volumes of contaminants displaced, providing a

well-justified physical interpretation of system behavior.

- We introduced an original deconvolution approach, termed PGM, for the effective assessment of diverse oil types within oil mixtures arising from the commingling of oils from different reservoirs, wells, or fields. This approach was inspired by methods relying on peak ratios and distinguishes itself by not necessitating synthetic mixtures, thus potentially reducing the need for extensive laboratory analyses. Furthermore, our study encompasses a comprehensive assessment of the accuracy of this PGM approach, as well as traditional deconvolution algorithms, by employing data of controlled experiments. The results prove the superiority –in terms of accuracy– of the novel deconvolution approach over traditional methods, particularly when the number of features employed during quantitative deconvolution is limited.

The contributions of this work are expected to provide valuable guidance in applications at different scales of the energy sector. For instance, the novel PGM methodology presented in Chapter 4, is currently being employed in the private sector to assess the proportions of oil mixtures at field scale, which allows for more efficient and less expensive energy production, a key aspect in the context of energy security. However, the contributions of this technology are not limited to the energy production. They can also be employed for the environmental protection since the PGM approach can be used to unambiguosly identify the source of oil spills, which is a key aspect in the context of environmental protection.

The probabilistic assessment of groundwater contamination associated with unconventional energy resource production, presented in Chapter 3, is expected to provide valuable guidance for the design of hydraulic fracturing operations, as well as the development of risk mitigation strategies.

Such a methdology is key in keeping water bodies clean and minimizing the risk of methane leakage into the atmosphere, which is a relevant aspect in the context of environmental protection and climate action.

Finally, the global sensitivity analysis of gas flow in low-permeability materials, presented in Chapter 2, is expected to provide valuable guidance for the design of experimental activities aimed at characterizing gas flow in low permeable materials. This physical phemonenon is key in the context of carbon capture and storage, but also in the geology storage of other gases like hydrogen and methane, thus contributing to the development of a more sustainable energy sector.

A | Appendix: Additional Mathematical Details Related to the Description of the Gas Flow Model Considered in Chapter 1

The correction factor ζ_{ms} is given by

$$\zeta_{ms} = \frac{\phi}{\tau}\left(1 - \frac{r_{ad}}{r}\right)^2 \left[\left(1 - \frac{r_{ad}}{r}\right)^{-2} - 1\right], \tag{A.1}$$

with

$$r_{ad} = r - d_m\theta, \tag{A.2}$$

$$r = r_o\left(\frac{p_c - p}{p_o}\right)^{-t}, \tag{A.3}$$

$$\phi = \phi_o\left(\frac{p_c - p}{p_o}\right)^{-q}, \tag{A.4}$$

where r_{ad} is thickness of the adsorbed gas layer, d_m is gas molecule diameter, p_c is overburden pressure, p_o is atmospheric pressure, and θ is

evaluated through a Langmuir equilibrium isotherm as:

$$\theta = \frac{p/Z}{p_L + p/Z}, \tag{A.5}$$

where p_L is a Langmuir pressure evaluated with

$$p_L = p_{L_o} \exp\left(-\frac{\Delta H}{RT}\right). \tag{A.6}$$

The correction factor ζ_{mb} is expressed as

$$\zeta_{mb} = \frac{\phi}{\tau}\left(1 - \frac{r_{ad}}{r}\right)^2. \tag{A.7}$$

The value of α (in Equation (2.7)) is evaluated through

$$\alpha = \alpha_0 \frac{2}{\pi} \tan^{-1}\left(\alpha_1 K_n^{\beta}\right), \tag{A.8}$$

where uncertain parameters α_0, α_1, and β allow representing the variation of α as a function of the Knudsen number (K_n). Here, α_0 represents the maximum value of α for large values of K_n, α_1 governs the values of α for small values of K_n, and β is a shape parameter.

The surface diffusion coefficient (D_s) is given by

$$D_s = D_s^0 \frac{(1-\theta) + \frac{\kappa}{2}\theta(2-\theta) + H(1-\kappa)(1-\kappa)\frac{\kappa}{2}\theta^2}{\left(1 - \theta + \frac{\kappa}{2}\theta\right)^2}, \tag{A.9}$$

with

$$D_s^0 = 8.29 \times 10^{-7} \exp\left(-\frac{\Delta H^{0.8}}{RT}\right), \tag{A.10}$$

$$H(1-\kappa) = \begin{cases} 0, & \text{if } \kappa \geq 1 \\ 1, & 0 \leq \kappa \leq 1 \end{cases}. \tag{A.11}$$

The adsorbed concentration (C_{sc}) is given by

$$C_{sc} = \frac{4\theta M}{\pi d_m^3 N_A}, \tag{A.12}$$

where N_A is the Avogadro Constant (6.02×10^{-23}/mol).

Bibliography

Amendola, A., Caldiero, L., Cerioli Regondi, A., Dolci, D., Galimberti, R. and Nali, M. [2017], Production allocation without end members: now it is possible, *in* 'Offshore Mediterranean Conference and Exhibition', OnePetro.

Arthur, J. D., Bohm, B. K., Coughlin, B. J., Layne, M. A. and Cornue, D. [2009], Evaluating the environmental implications of hydraulic fracturing in shale gas reservoirs, *in* 'SPE Americas E&P environmental and safety conference', OnePetro.

Bao, X. and Dai, L. [2009], 'Partial least squares with outlier detection in spectral analysis: A tool to predict gasoline properties', *Fuel* **88**(7), 1216–1222.

Barrie, C. D., Donohue, C. M., Zumberge, J. A. and Zumberge, J. E. [2020], 'Production allocation: Rosetta stone or red herring? best practices for understanding produced oils in resource plays', *Minerals* **10**(12), 1105.

Baskin, D., McCaffrey, M. and Kornacki, A. [2013], 'Allocating the contribution of oil from the eagle ford formation, the buda formation, and the austin chalk to commingled production from horizontal wells in south texas using geochemical fingerprinting technology', *Search & Discovery* **41268**.

Bianchi Janetti, E., Guadagnini, L., Riva, M. and Guadagnini, A. [2019], 'Global sensitivity analyses of multiple conceptual models with uncertain parameters driving groundwater flow in a regional-scale sedimentary aquifer', *Journal of Hydrology* **574**(September 2018), 544–556.

Blatman, G. and Sudret, B. [2011], 'Adaptive sparse polynomial chaos expansion based on least angle regression', *Journal of computational Physics* **230**(6), 2345–2367.

Britt, L. [2012], 'Fracture stimulation fundamentals', *Journal of Natural Gas Science and Engineering* **8**, 34–51.

Brunton, S. L. and Kutz, J. N. [2022], *Data-driven science and engineering: Machine learning, dynamical systems, and control*, Cambridge University Press.

Carati, C., Bonoldi, L., Bonetti, R., Nali, M. and Amendola, A. [2020], Production allocation of commingled reservoir fluids by on-site spectroscopic analysis, *in* 'International Petroleum Technology Conference', OnePetro.

Chiquet, P., Daridon, J. L., Broseta, D. and Thibeau, S. [2007], 'CO2/water interfacial tensions under pressure and temperature conditions of CO2 geological storage', *Energy Conversion and Management* **48**(3), 736–744.

Chuparova, E., Kratochvil, T., Kleingeld, J., Bilinski, P., Guillory, C., Bikun, J. and Djojosoeparto, R. [2010], 'Integration of time-lapse geochemistry with well logging and seismic to monitor dynamic reservoir fluid communication: Auger field case-study, deep water gulf of mexico', *Geological Society, London, Special Publications* **347**(1), 55–70.

Ciriello, V., Guadagnini, A., Di Federico, V., Edery, Y. and Berkowitz,

B. [2013], 'Comparative analysis of formulations for conservative transport in porous media through sensitivity-based parameter calibration', *Water Resources Research* **49**(9), 5206–5220.

Civan, F. [2010], 'Effective correlation of apparent gas permeability in tight porous media', *Transport in Porous Media* **82**(2), 375–384.

Colombo, I., Porta, G. M., Ruffo, P. and Guadagnini, A. [2017], 'Uncertainty quantification of overpressure buildup through inverse modeling of compaction processes in sedimentary basins', *Hydrogeology Journal* **25**(2), 385–403.

Coppens, M. O. [1999], 'The effect of fractal surface roughness on diffusion and reaction in porous catalysts – from fundamentals to practical applications', *Catalysis Today* **53**(2), 225–243.

Coppens, M. O. and Dammers, A. J. [2006], 'Effects of heterogeneity on diffusion in nanopores—From inorganic materials to protein crystals and ion channels', *Fluid Phase Equilibria* **241**(1-2), 308–316.

Cubitt, J. M., England, W. A. and Larter, S. R. [2004], 'Understanding petroleum reservoirs: Towards an integrated reservoir engineering and geochemical approach', *Geological Society, London, Special Publications* **237**, 1–5.

Darabi, H., Ettehad, A., Javadpour, F. and Sepehrnoori, K. [2012], 'Gas flow in ultra-tight shale strata', *Journal of Fluid Mechanics* **710**, 641–658.

de Lima Furtado, W., Corgozinho, C. N. C., Tauler, R. and Sena, M. M. [2021], 'Monitoring biodiesel and its intermediates in transesterification reactions with multivariate curve resolution alternating least squares calibration models', *Fuel* **283**, 119275.

De Pater, C. and Baisch, S. [2011], 'Geomechanical study of bowland shale seismicity', *Synthesis report* **57**.

Dell'Oca, A., Riva, M. and Guadagnini, A. [2017], 'Moment-based metrics for global sensitivity analysis of hydrological systems', *Hydrology and Earth System Sciences* **21**(12), 6219–6234.

Dell'Oca, A., Riva, M. and Guadagnini, A. [2020], 'Global Sensitivity Analysis for Multiple Interpretive Models With Uncertain Parameters', *Water Resources Research* **56**(2), 1–20.

Dembicki-Jr., H. [2017], *Petroleum Geochemistry for Exploration and Production*, Candice Janco.

Dong, J. J., Hsu, J. Y., Wu, W. J., Shimamoto, T., Hung, J. H., Yeh, E. C., Wu, Y. H. and Sone, H. [2010], 'Stress-dependence of the permeability and porosity of sandstone and shale from TCDP Hole-A', *International Journal of Rock Mechanics and Mining Sciences* **47**(7), 1141–1157.

Eberhardt, P., Feodoroff, T., Lui, E., Olivet, C. and Trew, S. [2013], 'The right to say no eu-canada trade agreement threatens fracking bans', *Policy* .

Edlmann, K. and McDermott, C. [2016], Hydro-geo-chemo-mechanical facies analysis relative to gas shales' of key basins, Technical report, 'FrackRisk European Project'.

Ekpo, B., Essien, N., Neji, P. and Etsenake, R. [2018], 'Geochemical fingerprinting of western offshore niger delta oils', *Journal of Petroleum Science and Engineering* **160**, 452–464.

Elsinger, R. J., Leenaarts, E. M., Kleingeld, J. C., van Bergen, P. and Gelin, F. [2010], Otter-eider geochemical production allocation: 6+

years of continuous monitoring to provide fiscal measurements for hydrocarbon accounting, *in* 'AAPG Hedberg Conference Applications of Reservoir Fluid Geochemistry, Vail, Colorado', pp. 8–11.

England, W. [2007], 'Reservoir geochemistry—a reservoir engineering perspective', *Journal of Petroleum Science and Engineering* **58**(3-4), 344–354.

European Comission [2014], 'Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions', `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52014DC0023R(01)`. Online; accessed 27 March 2023.

Folk, R. L. [1980], *Petrology of sedimentary rocks*, Hemphill publishing company.

Gläser, D., Dell'Oca, A., Tatomir, A., Bensabat, J., Class, H., Guadagnini, A., Helmig, R., McDermott, C., Riva, M. and Sauter, M. [2016], 'An approach towards a fep-based model for risk assessment for hydraulic fracturing operations', *Energy Procedia* **97**, 387–394.

Gläser, D., Helmig, R., Flemisch, B. and Class, H. [2017], 'A discrete fracture model for two-phase flow in fractured porous media', *Advances in Water Resources* **110**, 335–348.

Glazer, Y. R., Kjellsson, J. B., Sanders, K. T. and Webber, M. E. [2014], 'Potential for using energy from flared gas for on-site hydraulic fracturing wastewater treatment in texas', *Environmental Science & Technology Letters* **1**(7), 300–304.

Howarth, R. W., Ingraffea, A. and Engelder, T. [2011], 'Should fracking stop?', *Nature* **477**(7364), 271–275.

Hughes, J. D. [2013], 'Energy: A reality check on the shale revolution', *Nature* **494**(7437), 307–308.

Hwang, R., Baskin, D. and Teerman, S. [2000], 'Allocation of commingled pipeline oils to field production', *Organic Geochemistry* **31**(12), 1463–1474.

IEA [2022], 'World energy outlook 2022'.

International Energy Agency [2014], 'Gas - fuels & technology', `https://www.iea.org/fuels-and-technologies/gas`. Online; accessed 27 March 2023.

Jabbari, N., Aminzadeh, F. and de Barros, F. P. [2017], 'Hydraulic fracturing and the environment: risk assessment for groundwater contamination from well casing failure', *Stochastic Environmental Research and Risk Assessment* **31**, 1527–1542.

Janetti, E. B., Dror, I., Guadagnini, A., Riva, M. and Berkowitz, B. [2012], 'Estimation of single-metal and competitive sorption isotherms through maximum likelihood and model quality criteria', *Soil Science Society of America Journal* **76**(4), 1229–1245.

Jaumot, J., de Juan, A. and Tauler, R. [2015], 'Mcr-als gui 2.0: New features and applications', *Chemometrics and Intelligent Laboratory Systems* **140**, 1–12.

Javadpour, F. [2009], 'Nanopores and apparent permeability of gas flow in mudrocks (shales and siltstone)', *Journal of Canadian Petroleum Technology* **48**(08), 16–21.

Javadpour, F., Singh, H., Rabbani, A., Babaei, M. and Enayati, S. [2021], 'Gas Flow Models of Shale: A Review', *Energy and Fuels* **35**(4), 2999–3010.

Johansson, D., Lindgren, P. and Berglund, A. [2003], 'A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription', *Bioinformatics* **19**(4), 467–473.

Jweda, J., Michael, E., Jokanola, O. A., Hofer, R. and Parisi, V. A. [2017], Optimizing field development strategy using time-lapse geochemistry and production allocation in eagle ford, *in* 'SPE/AAPG/SEG Unconventional Resources Technology Conference', OnePetro.

Kanshio, S. [2020], 'A review of hydrocarbon allocation methods in the upstream oil and gas industry', *Journal of Petroleum Science and Engineering* **184**, 106590.

Karniadakis, G., Beskok, A. and Aluru, N. [2005], *Microflows and Nanoflows*, (Springer Science+Business Media, Inc.

Kaufman, R., Ahmed, A. and Hempkins, W. [1987], 'A new technique for the analysis of commingled oils and its application to production allocation calculations'.

Kaufman, R., Ahmed, d. S. and Elsinger, R. J. [1990], 'Gas chromatography as a development and production tool for fingerprinting oils from individual reservoirs: applications in the gulf of mexico'.

Kissinger, A., Helmig, R., Ebigbo, A., Class, H., Lange, T., Sauter, M., Heitfeld, M., Klünker, J. and Jahnke, W. [2013], 'Hydraulic fracturing in unconventional gas reservoirs: Risks in the geological system, part 2: Modelling the transport of fracturing fluids, brine and methane', *Environmental earth sciences* **70**, 3855–3873.

Koch, T., Glaser, D., Weishaupt, K., Ackermann, S., Beck, M., Becker, B., Burbulla, S., Class, H., Coltman, E., Emmert, S., Fetzer, T., Gruninger, C., Heck, K., Hommel, J., Kurz, T., Lipp, M., Moham-

madi, F., Scherrer, S., Schneider, M., Seitz, G., Stadler, L., Utz, M., Weinhardt, F. and Flemisch, B. [2020], 'DuMu^x 3 - an open-source simulator for solving flow and transport problems in porous media with a focus on model coupling', *Computers & Mathematics with Applications* .

Koks, E. E., Bočkarjova, M., de Moel, H. and Aerts, J. C. [2015], 'Integrated direct and indirect flood risk modeling: Development and sensitivity analysis', *Risk Analysis* **35**(5), 882–900.

Koolen, H. H., Gomes, A. F., de Moura, L. G., Marcano, F., Cardoso, F. M., Klitzke, C. F., Wojcik, R., Binkley, J., Patrick, J. S., Swarthout, R. F. et al. [2018], 'Integrative mass spectrometry strategy for fingerprinting and tentative structural characterization of asphaltenes', *Fuel* **220**, 717–724.

Krishna, R. and Wesselingh, J. A. [1997], 'The maxwell-stefan approach to mass transfer', *Chemical engineering science* **52**(6), 861–911.

Kumar, A., Chao, K., Hammack, R., Harbert, W., Maity, D., Eisenlord, S., Hayes, T. and Perry, K. [2018], Environmental impact analysis on the hydraulic fracture test site (hfts), *in* 'Unconventional Resources Technology Conference (URTEC)'.

la Cecilia, D., Porta, G. M., Tang, F. H., Riva, M. and Maggi, F. [2020], 'Probabilistic indicators for soil and groundwater contamination risk assessment', *Ecological Indicators* **115**, 106424.

Laloy, E., Rogiers, B., Vrugt, J. A., Mallants, D. and Jacques, D. [2013], 'Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion', *Water Resources Research* **49**(5), 2664–2682.

Lange, T., Sauter, M., Heitfeld, M., Schetelig, K., Brosig, K., Jahnke, W., Kissinger, A., Helmig, R., Ebigbo, A. and Class, H. [2013], 'Hydraulic fracturing in unconventional gas reservoirs: risks in the geological system part 1', *Environmental Earth Sciences* **70**, 3839–3853.

Lemmon, E. W., Bell, I., Huber, M. L. and McLinden, M. O. [2018], 'NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 10.0, National Institute of Standards and Technology'.

Li, Z., Dong, M., Li, S. and Huang, S. [2006], 'CO2 sequestration in depleted oil and gas reservoirs—caprock characterization and storage capacity', *Energy Conversion and Management* **47**(11-12), 1372–1382.

Liehui, Z., Baochao, S., Yulong, Z. and Zhaoli, G. [2019], 'Review of micro seepage mechanisms in shale gas reservoirs', *International Journal of Heat and Mass Transfer* **139**, 144–179.

Liu, F., Michael, E., Johansen, K., Brown, D. and Allwardt, J. [2017], Time-lapse geochemistry (tlg) application in unconventional reservoir development, *in* 'Unconventional Resources Technology Conference, Austin, Texas, 24-26 July 2017', Society of Exploration Geophysicists, American Association of Petroleum . . . , pp. 1078–1094.

Liu, J., Wang, J. G., Gao, F., Ju, Y., Zhang, X. and Zhang, L. C. [2016], 'Flow Consistency Between Non-Darcy Flow in Fracture Network and Nonlinear Diffusion in Matrix to Gas Production Rate in Fractured Shale Gas Reservoirs', *Transport in Porous Media* **111**(1), 97–121.

Llinares, R., Igual, J. and Camacho, A. [2012], 'Application of regularized alternating least squares to an astrophysical problem', *Applied Mathematics and Computation* **219**(3), 1367–1374.

Lu, J., Larson, T. E. and Smyth, R. C. [2015], 'Carbon isotope effects of methane transport through Anahuac Shale - A core gas study', *Journal of Geochemical Exploration* **148**, 138–149.

Mahmudova, A., Borsi, I. and Porta, G. M. [2022], 'Model-based characterization of permeability damage control through inhibitor injection under parametric uncertainty', *Computational Geosciences* **26**(5), 1119–1134.

Mahmudova, A., Civa, A., Caronni, V., Patani, S., Bozzoni, P., Bazzana, L. and Porta, G. [2023], 'Modelling parametric uncertainty in large-scale stratigraphic simulations', *Scientific Reports* **13**(1), 817.

Maina, F. Z., Guadagnini, A. and Riva, M. [2021], 'Impact of multiple uncertainties on gravimetric variations across randomly heterogeneous aquifers during pumping', *Advances in Water Resources* **154**, 103978.

Maina, F. Z. and Siirila-Woodburn, E. R. [2020], 'The Role of Subsurface Flow on Evapotranspiration: A Global Sensitivity Analysis', *Water Resources Research* **56**(7), 1–20.

McCaffrey, M. A., Ohms, D. H., Werner, M., Stone, C., Baskin, D. K. and Patterson, B. A. [2011], Geochemical allocation of commingled oil production or commingled gas production, *in* 'SPE Western North American Region Meeting', OnePetro.

McCaffrey, M., Baskin, D., Patterson, B., Ohms, D., Stone, C. and Reisdorf, D. [2012], 'Oil fingerprinting dramatically reduces production allocation costs', *World Oil* **55**.

Mehmani, A., Prodanović, M. and Javadpour, F. [2013], 'Multiscale, Multiphysics Network Modeling of Shale Matrix Gas Flows', *Transport in Porous Media* **99**(2), 377–390.

Menafoglio, A., Guadagnini, A. and Secchi, P. [2014], 'A kriging approach based on aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers', *Stochastic Environmental Research and Risk Assessment* **28**(7), 1835–1851.

Menafoglio, A., Guadagnini, L., Guadagnini, A. and Secchi, P. [2021], 'Object oriented spatial analysis of natural concentration levels of chemical species in regional-scale aquifers', *Spatial Statistics* **43**, 100494.

Milkov, A. V., Goebel, E., Dzou, L., Fisher, D. A., Kutch, A., McCaslin, N. and Bergman, D. F. [2007], 'Compartmentalization and time-lapse geochemical reservoir surveillance of the horn mountain oil field, deep-water gulf of mexico', *AAPG bulletin* **91**(6), 847–876.

Mohamed, M. S. [2000], Obaiyed field fluid geochemical analysis, *in* 'Abu Dhabi International Petroleum Exhibition and Conference', OnePetro.

Mohd Amin, S., Weiss, D. J. and Blunt, M. J. [2014], 'Reactive transport modelling of geologic CO2 sequestration in saline aquifers: The influence of pure CO2 and of mixtures of CO2 with CH4 on the sealing capacity of cap rock at 37°C and 100bar', *Chemical Geology* **367**, 39–50.

Moore, C. W., Zielinska, B., Petron, G. and Jackson, R. B. [2014], 'Air impacts of increased natural gas acquisition, processing, and use: a critical review', *Environmental science & technology* **48**(15), 8349–8359.

Murray, A. P. and Peters, K. E. [2021], 'Quantifying multiple source rock contributions to petroleum fluids: Bias in using compound ratios and neglecting the gas fraction', *AAPG Bulletin* **105**(8), 1661–1678.

Naraghi, M. E., Javadpour, F. and Ko, L. T. [2018], 'An Object-Based

Shale Permeability Model: Non-Darcy Gas Flow, Sorption, and Surface Diffusion Effects', *Transport in Porous Media* **125**(1), 23–39.

Niederreiter, H. [1992], *Random number generation and quasi-Monte Carlo methods*, SIAM.

Nouvelle, X., Rojas, K. and Stankiewicz, A. [2012], Novel method of production back-allocation using geochemical fingerprinting, *in* 'Abu Dhabi International Petroleum Conference and Exhibition', OnePetro.

Osborn, S. G., Vengosh, A., Warner, N. R. and Jackson, R. B. [2011], 'Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing', *proceedings of the National Academy of Sciences* **108**(20), 8172–8176.

Pan, Z., Connell, L. D., Camilleri, M. and Connelly, L. [2010], 'Effects of matrix moisture on gas diffusion and flow in coal', *Fuel* **89**(11), 3207–3217.

Patience, R., Bastow, M., Fowler, M., Moore, J. and Barrie, C. [2021], The application of petroleum geochemical methods to production allocation of commingled fluids, *in* 'SPE Europec featured at 82nd EAGE Conference and Exhibition', OnePetro.

Permanyer, A., Douifi, L., Lahcini, A., Lamontagne, J. and Kister, J. [2002], 'Ftir and suvf spectroscopy applied to reservoir compartmentalization: a comparative study with gas chromatography fingerprints results', *Fuel* **81**(7), 861–866.

Peters, K. E., Ramos, L. S., Zumberge, J. E., Valin, Z. C. and Bird, K. J. [2008], 'De-convoluting mixed crude oil in prudhoe bay field, north slope, alaska', *Organic Geochemistry* **39**(6), 623–645.

Rani, S., Prusty, B. K. and Pal, S. K. [2018], 'Adsorption kinetics and dif-

fusion modeling of CH4 and CO2 in Indian shales', *Fuel* **216**(November 2017), 61–70.

Ritchie, H., Roser, M. and Rosado, P. [2020], 'Co2 and greenhouse gas emissions', *Our world in data* .

Riva, M., Guadagnini, A. and Dell'Oca, A. [2015], 'Probabilistic assessment of seawater intrusion under multiple sources of uncertainty', *Advances in Water Resources* **75**, 93–104.

Riva, M., Guadagnini, L., Guadagnini, A., Ptak, T. and Martac, E. [2006], 'Probabilistic study of well capture zones distribution at the lauswiesen field site', *Journal of contaminant hydrology* **88**(1-2), 92–118.

Rudyk, S., Spirov, P. and Sogaard, E. [2013], 'Application of gc–ms chromatography for the analysis of the oil fractions extracted by supercritical co2 at high pressure', *Fuel* **106**, 139–146.

Safari, A., Das, N., Langhelle, O., Roy, J. and Assadi, M. [2019], 'Natural gas: A transition fuel for sustainable energy system transformation?', *Energy Science & Engineering* **7**(4), 1075–1094.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. and Tarantola, S. [2010], 'Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index', *Computer Physics Communications* **181**(2), 259–270.

Saltelli, A. and Sobol', I. M. [1995], 'Sensitivity analysis for nonlinear mathematical models: numerical experience (in Russian)', *Mathematical models and computer experiment* **7**(11), 16–28.

Savaresi, A. [2016], 'The paris agreement: a new beginning?', *Journal of Energy & Natural Resources Law* **34**(1), 16–26.

Schloemer, S. and Krooss, B. M. [2004], 'Molecular transport of methane, ethane and nitrogen and the influence of diffusion on the chemical and isotopic composition of natural gas accumulations', *Geofluids* **4**(1), 81–108.

Schlömer, S. and Krooss, B. M. [1997], 'Experimental characterisation of the hydrocarbon sealing efficiency of cap rocks', *Marine and Petroleum Geology* **14**(5), 565–580.

Setzmann, U. and Wagner, W. [1991], 'A New Equation of State and Tables of Thermodynamic Properties for Methane Covering the Range from the Melting Line to 625 K at Pressures up to 100 MPa', *Journal of Physical and Chemical Reference Data* **20**(6), 1061–1155.

Singh, H. and Myong, R. S. [2018], 'Critical Review of Fluid Flow Physics at Micro- to Nano-scale Porous Media Applications in the Energy Sector', *Advances in Materials Science and Engineering* **2018**.

Smith, M. B. and Montgomery, C. [2015], *Hydraulic fracturing*, CRC press.

Sochala, P. and Le Maître, O. P. [2013], 'Polynomial Chaos expansion for subsurface flows with uncertain soil parameters', *Advances in Water Resources* **62**, 139–154.

Song, W., Yao, J., Li, Y., Sun, H., Zhang, L., Yang, Y., Zhao, J. and Sui, H. [2016], 'Apparent gas permeability in an organic-rich shale reservoir', *Fuel* **181**, 973–984.

Spagnolini, U. [2018], *Statistical Signal Processing in Engineering*, John Wiley & Sons.

Stringfellow, W. T., Domen, J. K., Camarillo, M. K., Sandelin, W. L. and Borglin, S. [2014], 'Physical, chemical, and biological characteris-

tics of compounds used in hydraulic fracturing', *Journal of hazardous materials* **275**, 37–54.

Sudret, B. [2008], 'Global sensitivity analysis using polynomial chaos expansions', *Reliability Engineering & System Safety* **93**(7), 964–979.

Sun, Z., Li, X., Shi, J., Zhang, T. and Sun, F. [2017], 'Apparent permeability model for real gas transport through shale gas reservoirs considering water distribution characteristic', *International Journal of Heat and Mass Transfer* **115**, 1008–1019.

Suykens, J. A. and Vandewalle, J. [1999], 'Least squares support vector machine classifiers', *Neural processing letters* **9**(3), 293–300.

Tan, Y., Pan, Z., Liu, J., Kang, J., Zhou, F., Connell, L. D. and Yang, Y. [2018], 'Experimental study of impact of anisotropy and heterogeneity on gas flow in coal. Part I: Diffusion and adsorption', *Fuel* **232**(15), 444–453.

Tartakovsky, D. M. [2013], 'Assessment and management of risk in subsurface hydrology: A review and perspective', *Advances in Water Resources* **51**, 247–260.

U.S. Energy Information Administration [2015], 'World Shale Resource Assessments'.
URL: *https://www.eia.gov/analysis/studies/worldshalegas/*

van Bergen, P. F. and Gordon, M. [2020], 'Production geochemistry: fluids don't lie and the devil is in the detail', *Geological Society, London, Special Publications* **484**(1), 9–28.

Van den Boogaart, K. G. and Tolosana-Delgado, R. [2013], *Analyzing compositional data with R*, Vol. 122, Springer.

Van Genuchten, M. T. [1980], 'A closed-form equation for predicting the hydraulic conductivity of unsaturated soils', *Soil science society of America journal* **44**(5), 892–898.

Vassiliou, M. S. [2009], *The A to Z of the Petroleum Industry*, Vol. 116, Scarecrow Press.

Vengosh, A., Jackson, R. B., Warner, N., Darrah, T. H. and Kondash, A. [2014], 'A critical review of the risks to water resources from unconventional shale gas development and hydraulic fracturing in the united states', *Environmental science & technology* **48**(15), 8334–8348.

Walker, R., Aminzadeh, F. and Tiwari, A. [2014], Correlation between induced seismic events and hydraulic fracturing activities in california, *in* 'AGU Fall Meeting Abstracts', Vol. 2014, pp. S54A–05.

Wang, H., Chen, L., Qu, Z., Yin, Y., Kang, Q., Yu, B. and Tao, W.-Q. [2020], 'Modeling of multi-scale transport phenomena in shale gas production—a critical review', *Applied Energy* **262**, 114575.

Wang, H., Chu, X., Chen, P., Li, J., Liu, D. and Xu, Y. [2022], 'Partial least squares regression residual extreme learning machine (plsrr-elm) calibration algorithm applied in fast determination of gasoline octane number with near-infrared spectroscopy', *Fuel* **309**, 122224.

Wang, T., Tian, S., Li, G. and Zhang, P. [2019], 'Analytical Model for Real Gas Transport in Shale Reservoirs with Surface Diffusion of Adsorbed Gas', *Industrial and Engineering Chemistry Research* **58**(51), 23481–23489.

Wu, K., Chen, Z. and Li, X. [2015], 'Real gas transport through nanopores of varying cross-section type and shape in shale gas reservoirs', *Chemical Engineering Journal* **281**, 813–825.

Wu, K., Chen, Z., Li, X., Guo, C. and Wei, M. [2016], 'A model for multiple transport mechanisms through nanopores of shale gas reservoirs with real gas effect-adsorption-mechanic coupling', *International Journal of Heat and Mass Transfer* **93**, 408–426.

Wu, K., Chen, Z., Li, X., Xu, J., Li, J., Wang, K., Wang, H., Wang, S. and Dong, X. [2017], 'Flow behavior of gas confined in nanoporous shale at high pressure: Real gas effect', *Fuel* **205**, 173–183.

Wu, K., Li, X., Wang, C., Yu, W. and Chen, Z. [2015], 'Model for surface diffusion of adsorbed gas in nanopores of shale gas reservoirs', *Industrial and Engineering Chemistry Research* **54**(12), 3225–3236.

Xiao, S., Praditia, T., Oladyshkin, S. and Nowak, W. [2021], 'Global sensitivity analysis of a CaO/Ca(OH)2 thermochemical energy storage model for parametric effect analysis', *Applied Energy* **285**(December 2020), 116456.

Yan, Y., Guan, Z., Yan, W. and Wang, H. [2020], 'Mechanical response and damage mechanism of cement sheath during perforation in oil and gas well', *Journal of Petroleum Science and Engineering* **188**, 106924.

Yang, C., Yang, Z., Zhang, G., Hollebone, B., Landriault, M., Wang, Z., Lambert, P. and Brown, C. E. [2016], 'Characterization and differentiation of chemical fingerprints of virgin and used lubricating oils for identification of contamination or adulteration sources', *Fuel* **163**, 271–281.

Yang, W., Casey, J. F., Gao, Y. and Li, J. [2019], 'A new method of geochemical allocation and monitoring of commingled crude oil production using trace and ultra-trace multi-element analyses', *Fuel* **241**, 347–359.

Yang, X. and Karniadakis, G. E. [2013], 'Reweighted l1 minimization

method for stochastic elliptic differential equations', *Journal of Computational Physics* **248**, 87–108.

Yeniay, Ö. and GÖKTAŞ, A. [2002], 'A comparison of partial least squares regression with other prediction methods', *Hacettepe Journal of Mathematics and Statistics* **31**, 99–111.

Yuan, W., Pan, Z., Li, X., Yang, Y., Zhao, C., Connell, L. D., Li, S. and He, J. [2014], 'Experimental study and modelling of methane adsorption and diffusion in shale', *Fuel* **117**(PART A), 509–519.

Zhan, Z.-W., Tian, Y., Zou, Y.-R., Liao, Z. et al. [2016], 'De-convoluting crude oil mixtures from palaeozoic reservoirs in the tabei uplift, tarim basin, china', *Organic Geochemistry* **97**, 78–94.

Zhang, Q., Su, Y., Wang, W., Lu, M. and Sheng, G. [2018], 'Gas transport behaviors in shale nanopores based on multiple mechanisms and macroscale modeling', *International Journal of Heat and Mass Transfer* **125**, 845–857.

Ziarani, A. S. and Aguilera, R. [2012], 'Knudsen's Permeability Correction for Tight Porous Media', *Transport in Porous Media* **91**(1), 239–260.

List of Figures

List of Tables