Machine Learning for Bayesian Experimental Design in the Subsurface

Dissertation submitted in fulfilment
of the requirements for the Degree of
Doctor of Science: Geology
at Ghent University

Robin Thibaut

2023

Candidate	Ir. Robin Thibaut Department of Geology, Ghent University
Supervisor	Prof. Dr. Ir. Thomas Hermans Department of Geology, Ghent University
Cosupervisor	Dr. Ir. Eric Laloy Belgian Nuclear Research Centre (SCK-CEN)
Dean	Prof. Dr. Isabel Van Driessche
Rector	Prof. Dr. Ir. Rik Van de Walle
Jury member	Prof. Dr. Ir. Ellen Van De Vijver
Jury member	Prof. Dr. Ir. Frédéric Nguyen
Jury member	Prof. Dr. Kristine Walraevens
Jury member	Prof. Dr. Ty Ferré
Chair of the jury	Prof. Dr. Stephen Louwye

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

Ronald A. Fisher (1890-1962)

Abstract

Accurate modeling of the subsurface, a complex and heterogeneous environment that plays a crucial role in the Earth's water cycle, is challenging due to sparse and incomplete data. We can reduce the uncertainty associated with subsurface predictions, such as groundwater flow and contaminant transport, by conducting additional observations and measurements in the subsurface. However, practical and economic considerations frequently limit the number of measurements and their locations, such as land occupation, which may limit the number of wells that can be drilled. In this dissertation, we propose simulation-driven methods to reduce uncertainty in subsurface predictions by identifying the most informative data sets to gather. Our method, which is based on Bayesian optimal experimental design and machine learning, determines the nature and location of these data sets, which can include measurements of groundwater levels, temperature, and other parameters collected through active or passive sensing methods such as pumping tests, tracer tests, and geophysical surveys. This dissertation is the first to use Bayesian Evidential Learning (BEL) for optimal experimental design, allowing for the optimization of data source locations and the comparison of the utility of different data sources. BEL is a framework for prediction that combines Monte Carlo sampling and machine learning in order to learn a direct relationship between predictor and target variables generated by a simulation model. We demonstrate the efficacy of our methods in three groundwater modeling case studies: (i) wellhead protection area delineation, (ii) an aquifer thermal energy storage monitoring system, and (iii) groundwater-surface water interaction. The case studies show that our approach can significantly reduce the uncertainty in subsurface predictions and guide further subsurface exploration. The first case study, in particular, uses the Traveling Salesman Problem to introduce a novel approach to wellhead protection area delineation. The second case study, which compares well and geophysical data for temperature monitoring, introduces a new method for combining observations from multiple data sources in a latent space of the original data. The third case study introduces the Probabilistic Bayesian neural network (PBNN) method to BEL and transitions from a static experimental design framework to a sequential experimental design framework, which estimates groundwater-surface water interaction fluxes from temperature data. We have also developed a Python package, SKBEL, that implements our methods and can be used for a variety of Earth Science applications. Overall, this dissertation demonstrates the utility of BEL for optimal experimental design in groundwater modeling, highlights the potential of BEL for predictive modeling in Earth Sciences, and opens up new avenues for data and simulation-driven subsurface modeling.

Samenvatting

Nauwkeurige modellering van de ondergrond, een complexe en heterogene omgeving die een cruciale rol speelt in de watercyclus van de aarde, is een uitdaging vanwege schaarse en onvolledige gegevens. Door aanvullende waarnemingen en metingen in de ondergrond te doen, kunnen we de onzekerheid die samenhangt met voorspellingen in de ondergrond, zoals grondwaterstroming en transport van verontreinigingen, verkleinen. Praktische en economische overwegingen beperken echter vaak het aantal metingen en hun locaties, zoals landgebruik, wat het aantal putten dat kan worden geboord kan beperken. In deze thesis stellen we simulatiegestuurde methoden voor om de onzekerheid in ondergrondse voorspellingen te verminderen, door de meest informatieve datasets te identificeren die moeten worden verzameld. Onze methode, die is gebaseerd op Bayesiaans optimaal experimenteel ontwerp en machine learning, bepaalt de aard en locatie van deze datasets, waaronder metingen van grondwaterstanden, temperatuur en andere parameters die zijn verzameld via actieve of passieve detectiemethoden zoals pomptesten, tracertests en geofysische onderzoeken. Deze thesis is de eerste waarin Bayesian Evidential Learning (BEL) wordt gebruikt voor een optimaal experimenteel ontwerp, waardoor de locaties van gegevensbronnen kunnen worden geoptimaliseerd en het nut van verschillende gegevensbronnen kan worden vergeleken. BEL is een raamwerk voor voorspelling dat Monte Carlo-sampling en machine learning combineert om een directe relatie te leren tussen voorspellende variabelen en doelvariabelen gegenereerd door een simulatiemodel. We demonstreren de doeltreffendheid van onze methoden in drie casestudy's voor grondwatermodellering: (i) afbakening van het beschermingsgebied van de boorput, (ii) een monitoringsysteem voor thermische energieopslag in watervoerende lagen en (iii) interactie tussen grondwater en oppervlaktewater. De casestudy's laten zien dat onze aanpak de onzekerheid in ondergrondse voorspellingen aanzienlijk kan verminderen en verdere verkenning van de ondergrond kan begeleiden. Met name de eerste casestudy gebruikt het Traveling Salesman Problem om een nieuwe benadering van de afbakening van boorputbeschermingsgebieden te introduceren. De tweede casestudy, die put- en geofysische gegevens voor temperatuursmonitoring vergelijkt, introduceert een nieuwe methode voor het combineren van waarnemingen uit meerdere gegevensbronnen in een latente ruimte van de oorspronkelijke gegevens. De derde casestudy introduceert de probabilistische Bayesiaanse neurale netwerkmethode (PBNN) in BEL en gaat over van een statisch experimenteel ontwerpkader naar een sequentieel experimenteel ontwerpkader, dat de interactiefluxen tussen grondwater en oppervlaktewater schat op basis van temperatuurgegevens. We hebben ook een Python-package ontwikkeld, SKBEL, dat onze methoden implementeert en kan worden gebruikt voor verschillende aardwetenschappelijke

toepassingen. Over het algemeen demonstreert deze thesis het nut van BEL voor optimaal experimenteel ontwerp in grondwatermodellering, benadrukt het potentieel van BEL voor voorspellende modellering in aardwetenschappen en opent nieuwe wegen voor data- en simulatiegestuurde ondergrondse modellering.

Acknowledgements

Professional Acknowledgements

I would like to start by acknowledging my advisor, Prof. Dr. Ir. Thomas Hermans, and co-advisor, Dr. Ir. Eric Laloy, for their guidance and constructive criticism throughout the making of this dissertation. Their expertise and enthusiasm for the topic have been an immense source of motivation and inspiration, and they both deserve praise and recognition for providing me with the tools, resources, and knowledge I required to complete this project successfully.

Additionally, I would like to express my gratitude to the entire Geology Department for their assistance and support throughout my time at Ghent University. Marion Braeckeleire, Kurt Blom, Wim Lievens and Marc Faure deserve special recognition for their assistance with administrative tasks and for keeping the Department of Geology afloat.

I am also sincerely grateful to Ghent University for the financial support provided by the Bijzonder Onderzoeksfonds (BOF; Special Research Fund), the Department of Geology, and the FWO (Fonds Wetenschappelijk Onderzoek; Research Foundation Flanders) for funding my travel expenses to conferences and workshops.

Personal Acknowledgements

Separating the acknowledgements into professional and personal sections was not easy, as my professional and personal life have become increasingly intertwined over the past few years. I came to truly understand the meaning of the adage "find a job you love, and you will never have to work a day in your life" during the past few years, and this dissertation is the final result of that experience.

It all started four years ago when my advisor Thomas Hermans invited me to start working on this project, while I was still hired by an offshore geophysical company, which is why I like to say that he "pulled me out of the sea." He has been a tremendous source of knowledge and support throughout my research, and he has been a real mentor to me. It is remarkable how when asked a vague question about a very specific topic, he is one of the few researchers who can recall the title, authors, and line number of a paper published 15 years ago on the subject without consulting a computer.

I am forever thankful to him for believing in my potential and allowing me to contribute to the scientific community.

Going even further back in time, I would like to acknowledge my former advisor, Frédéric Nguyen, who witnessed my first steps into the worlds of engineering and geophysics at the University of Liège. He did not hesitate to give me a chance and allow me to work on a life-changing project in Cambodia. This travel shaped the person I am today and set me on the path that led to my research career.

Other travels have also been instrumental in my development as a researcher and a person. The last few years have been enriched by my immersion in the language and culture of Vietnam. I had the chance to travel to Vietnam two times for work during my PhD, and the warm welcome I received from everyone I met during my stay made me feel like I had come home. Eating and drinking snake blood and liver extract with my advisor was one of the highlights of my PhD. I would like to express my gratitude to my friends and colleagues from the Vietnam Institute of Geosciences and Mineral Resources, Hieu, Nghi, Ly and Chien, for the wonderful time we had together in the plains of the Binh Thuan province. Special thanks to Linh and Diep, who made my time in Vietnam unforgettable and whom I was fortunate to see again in Belgium.

Another trip I'll never forget was my 2021 trip to the United States, where I attended the AGU Fall Meeting in New Orleans. There were no snakes on the menu, but alligator burgers were readily available. Prior to the conference, I paid a visit to Professor Ty Ferré at the University of Arizona, where we exchanged ideas and discussed the future of my research. His knowledge and enthusiasm for the subject were a great source of inspiration for me, and they helped shape the direction of my research in the coming years.

Closer to home, I would like to thank my dear PhD colleagues for their support and friendship. Special thanks to my officemate Marieke Paepen, who has been there since the beginning and has been an invaluable source of support. The physics expert in our office, Wouter Deleersnyder, deserves special recognition as well. Both of them are excellent researchers, and it was a real pleasure to share ideas, knowledge and experiences with them. During the last few years, I've also had the chance to tutor and work with a number of talented students, some of whom have since become my colleagues, like Lore Vanhooren, who helped improve my software, and Luka Tas, who completely filled my hard drive with her data. I also want to thank Guillaume Vandekerckhove, who will graduate soon, for giving me the chance to share my Python knowledge and for helping me improve and test my software. I also want to acknowledge our newest colleague, Le Zhang, whose enthusiasm is contagious, and it's already been a pleasure working with him.
My gratitude extends to the entire Geology Department, where I have had the privilege of working alongside many talented and passionate researchers from around the world, such as Melissa, Alice, Coralie, Cristiana, Lotte, Ana, Carolina, Tom, Pjotr, Stijn.

Once upon a time, there was life outside of work. In order to maintain a healthy

work-life balance, it was fortunate for me to have lived in a variety of locations with a wide range of roommates throughout my PhD studies. It's always nice to come home to a warm and welcoming environment, and I'd like to thank all of my roommates for their friendship: Rijuta, Francesca, An-Sofie, Ruben, Joanna, Annelien, Elizabeth, Tim, Lies, and Anahita. A very special thank you goes out to Gieles and Seya, an incredible couple who were my roommates during most of the pandemic that affected our lives in 2020.

Over the past few years, I've also met a huge number of people whom, if only for a short time, I could consider friends, and occasionally even more than that. To all of you, you know who you are, I thank you for the many wonderful memories. I hope that our paths will cross again in the future. I must mention Alberto, who has been a wonderful friend (and colleague) and has welcomed me into his lovely family and stunning homeland of Baja California Sur, Mexico. Another special mention to my friend Vladimir, who introduced me to the beauty that can be found in Switzerland. I would also like to thank Wojciech, Amaury, Paddy, Sara, Michael, and the rest of the PhD community for all of our laughs and good times. In addition, I'd like to say a big thanks to Shruti, who has stuck with me through the final stages of my PhD.

Last but not least, I would like to thank my family for their unconditional love. Easily the most important people in my life, they have always supported me, no matter what. My mother, Pascale, who has been my biggest fan, deserves special recognition for her unwavering support and encouragement. I would also like to thank my father, Charles, for his subtle words of appreciation and the occasional whiskey shot. I would like to thank Charlotte, my first younger sister, who is an exceptional artist who taught me how to correctly design figures and has always been there for me. My second sister, Zoé, third sister, Lucie, and younger brother, Jules, have always been an immense source of joy and inspiration to me. I would like to thank Marie as well for keeping this little family together.

I have the most wonderful cousins in the world, and all of them deserve recognition for their support and love during my PhD. I would like to thank my cousins Clément, Lucas, Flore, Camille, Ivan, Macha, Jonas, Caroline, Hugo, Amélie and Delphine for their presence in my life. These last few summers have been marked by vacations with my cousins in Quiberon, where we have shared memorable moments spearfishing octopuses and enjoying sunsets on the beach. I would also like to thank all of my aunts and uncles for their kindness and the occasional gift of delectable food. My grandparents hold a special place in my heart. Yvan, my grandfather, is an incredible cook, and he undoubtedly provided me with much-needed energy during my PhD. I would also like to thank my grandmother, Yvie, for her love (another kind of food).

I would like to thank my family of heart, including Lara, Karl, Pablo, Martin and Chantal, for the wonderful moments and experiences we have had together, and for the many more to come. A special thanks goes to Chantal, who was the conductor of my academic orchestra during my first years at the Université Libre de Bruxelles.

Finally, I want to acknowledge the people who are not here anymore, but who have left an indelible mark on my life. I would like to thank my beloved grandmother, Gisèle, who will always be in my heart. I would also like to remember my grandfather, Jo, who was a great source of inspiration for me. My thoughts go out to my uncle Arthur, who left us too soon, but whose fighting spirit has served as an inspiration.

Contents

1. Introduction

The ability of scientists to predict the outcomes of external events or their actions in complex environments—particularly their reliance on models—is critical for researchers and decision-makers concerned about upcoming challenges. This ability to predict the short or long term future will have an impact on managing groundwater resources for future generations, as well as answering questions about the environmental impact of climate change, groundwater resource protection, and the transition to renewable energies, such as:

- "How will groundwater levels change in the future?", e.g., Aquilina et al. (2015)
- "Will this contaminant reach the nearby drinking water well?", e.g., MacDonald et al. (2016)
- "How much energy can we recover from this aquifer thermal energy storage system?", e.g., Kammen and Sunter (2016)
- "Where to drill a water abstraction well?", e.g., Robert et al. (2011)
- "Is this area safe for building certain structures?", e.g., Van Hoorde et al. (2017)
- "Where to drill to cross a deposit's mineralizations?", e.g., Evrard et al. (2018); Thibaut et al. (2021a).
- "What is the safest extent of the wellhead protection area?", e.g., Thibaut et al. (2021b).
- "What is the optimal sensor combination for monitoring a groundwater system?", e.g., Thibaut et al. (2022).
- "How to safely store CO_2 in the subsurface?", e.g., Corso et al. (2022).

It is therefore critical to obtain the most reliable subsurface information in order to facilitate human interpretation and decision-making (Hall et al., 2022). However, due to the subsurface's complexity and nature, there are significant uncertainties in any prediction. As a result, it is clear that a prediction alone is insufficient; rather, we argue in this dissertation that a full quantification of uncertainty that takes into account all potential outcomes is required for an appropriate risk analysis and subsequent decision-making (Renard, 2007). The term "risk" needs to be defined. Risk is commonly used to refer to a hazard or peril, as well as the possibility of loss as a result of that hazard

or peril (Davis et al., 1972). In this dissertation, we favor Klausner (1969) definition: "[...] risk is considered to be the consequential effect of possible uncertain outcomes." The purpose of this dissertation is reflected in Klausner's definition, which builds on the conventional definition and forms the basis for its extension.

The "forecaster" or "modeler" is typically not a single entity; rather, it is frequently a complex system consisting of both human experts and computers (Ramos et al., 2010; Weijs et al., 2010). The forecaster uses data from various sources to provide an estimate for some quantities of value to the user. This value is related to a decision problem associated with the prediction and is more closely related to engineering than science. It is thus dependent not only on the future predictions and direct observation, but also on who is using the prediction (Weijs et al., 2010).

For example, hydrological prognostications may be useful for reservoir operation, evacuation decisions, and agriculture. High quality dam operation forecasts, for example, can result in more hydropower, less flood damage, and fewer unnecessary pre-releases for flood protection (Weijs et al., 2010; Wood and Rodríguez-Iturbe, 1975). In Belgium, the need for accurate forecasts became especially clear in the summer of 2021, when severe floods affected the south of Belgium, and the need for accurate forecasts was highlighted (Brussels Times, 2021, 2022a,b,c).

Several factors contribute to the source of uncertainty when looking at the subsurface. For instance, the subsurface is subject to varying scales of uncertainty along its spatial and temporal axes. Geological information can be drawn from a variety of sources; however, the reliability of that information cannot always be established. Drilling data, borehole data, and geophysical data, for example, can be used to inform geological models, but the validity of this data is limited by our ability to collect and interpret it. Though the amount of data has increased over the past decades, larger data sets do not necessarily guarantee better predictions; moreover, the data may be incomplete or inaccurate due to limited sampling. Sampling by definition results in an incomplete dataset, but this can be compounded by inaccurate data, which is data that does not accurately represent the subsurface. This can be due to limitations in the sampling technique, or errors in the collection or interpretation of the data.

Additionally, we need to consider the inherent variation of the subsurface and the relatively short amount of time for which records are available and valid. As subsurface prediction models are intended to facilitate the decision-making process, the uncertainty analysis of these predictions is crucial. In the best situation, a modeler can provide an exact probabilistic prediction; in the worst situation, no prediction will be made. In this context, "exact" refers to the accuracy of the prediction.

Consider a decision problem involving the management of a contaminated site. The decision is whether or not the site should be cleaned up. The contamination can be assessed using computer models, but the model is bound to be uncertain; it is impossible to know for certain whether or not a specific area is contaminated without sampling a

number of locations. If the monitoring is poorly designed, the model may be unable to diagnose the contamination, resulting in an overly uncertain prediction; the decision-maker will likely reject the cleanup of the location, fearing a waste of time and money. In this case, the contamination remains despite the fact that the likelihood of contamination is not zero and cleanup would be beneficial. This is where a reliable uncertainty analysis comes in: it is not enough to simply predict a specific outcome; rather, it is critical to provide the user with an explanation of the level of confidence that can be placed in the prediction's accuracy.

The goal of this study is to provide information to increase the likelihood of obtaining reliable predictions by understanding the sources of uncertainty faced by the modeler and quantifying their effects. The principles of Bayesian inference, also known as Bayesian statistics or Bayesianism (e.g., Jaynes 2003), are the underpinnings of this study. Bayesianism is not just a statistical approach to learning from data, but rather a system of reasoning based on statistical ideas (Bayes, 1763). It offers a well-defined framework for making optimal use of evidence in the decision-making process under uncertainty. As such, it is a natural step forward from traditional methods to a promising approach for addressing subsurface uncertainties. Caers (2011) and Scheidt et al. (2018) provide a comprehensive treatment of the topic by presenting case-study-specific strategies and workflows based on the Bayesian philosophy and tailored to the specifics of the subsurface realm; they also provide the foundation for this dissertation. We will build on these foundations and extend them to include various aspects of subsurface prediction and uncertainty assessment. In particular, this dissertation revolves around experimental design (ED) and uncertainty quantification (UQ). Novel methods for ED and UQ in the subsurface based on Bayesian Evidential Learning (BEL; Scheidt et al. 2018) are developed and applied to three distinct subsurface cases.

Our primary focus is subsurface hydrology, but the subsurface's nature makes it difficult to separate it from other disciplines, particularly surface hydrology and geophysics (e.g., Ferré et al. 2009; Linde et al. 2017). As a result, each case often verges on the other disciplines, and the literature is frequently shared between them. The three cases presented in this dissertation are as follows:

- **Case 1:** Wellhead protection area (WHPA) delineation, e.g., to protect drinking water wells from contamination. We introduce the use of BEL for complex hydrological prediction under uncertainty and its integration within optimal experimental design.
- **Case 2:** Aquifer thermal energy storage (ATES) monitoring, e.g., to recover energy from the subsurface. In this case study, we extend BEL to allow for the integration and comparison of data of various types and origins in order to optimize experimental setups.
- **Case 3:** Groundwater - surface water interaction, e.g., to understand the impact of climate change on groundwater levels. We extend the concept of BEL for experimental design from static to sequential optimization in this case study,

demonstrating the need for physical simulation-based ED with the appropriate complexity level.

1.1 A primer on experimental design

Once upon a time, there was a scientist. R.A. Fisher is said to have pioneered experimental design at the Rothamsted Agricultural Experiment Station around 1918 (Box et al., 1978). Fisher was then hired to extract information from decades of crop yield records, and it was his work that led to the development of experimental design (Box et al., 1978). Since then, experimental design has become a major tool in the scientific arsenal, and its applications have expanded far beyond agriculture.

Definitions. In this section, we outline the fundamentals of experimental design and discuss considerations for designing field experiments in the geosciences. We also discuss the benefits of using machine learning and Bayesian inference (via BEL) in experimental design.

We first need to define the terms that will be used throughout this dissertation.

Definition 1 (Experiment) *Test or series of runs in which deliberate changes are made to the input variables of a process or system in order to observe and identify the causes of output response changes (Montgomery, 2019).*

We may wish to determine which input variables are responsible for the observed changes in the response, develop a model relating the response to the significant input variables, and then use this model for process or system improvement or other decision-making purposes (Montgomery, 2019).

To fully comprehend the cause-and-effect relationships in a system, it is necessary to manipulate the system's input variables and observe how these modifications affect the system's output (Montgomery, 2019). In other words, experiments must be conducted on the system. Observations of a system or process can lead to theories or hypotheses about what makes the system function, but experiments of the type described above are necessary to prove the validity of these theories (Montgomery, 2019).

Definition 2 (Experimental design) *The strategy for carrying out an experiment, including the factors (variables manipulated in an experiment), levels (value of a factor), and number of replicates required to produce the required information with the least amount of effort (Box et al., 1978; Montgomery, 2019).*

These are the three pillars of a "classical" experimental design (Box et al., 1978; Montgomery, 2019):

- **Randomization:** Randomization is the process of assigning experimental units to treatments in such a way that each experimental unit has an equal chance of being assigned to any treatment.

- **Replication:** Replication is the process of repeating the experiment on multiple experimental units.
- **Blocking:** Blocking is the process of grouping experimental units into blocks, where each block is assigned to a single treatment.

Although spatial data is distinguished by the fact that each observation is supported by a set of coordinates indicating the location of the respective data collection site, these fundamental concepts of the classical experimental design framework can be applied to the design of subsurface field experiments (Müller, 2007). A field experiment, for example, could be used to investigate the effects of subsurface injection of heated water in a shallow aquifer. The factors in this experiment would be injection temperature, injection rate, and well spacing, with the response variable being the temperature of the aquifer at various depths. Randomization entails randomly assigning injection wells to different treatments, replication entails repeating the experiment on multiple injection wells, and blocking entails grouping the injection wells into treatment-specific blocks. In order to account for the differences in subsurface properties near the wells, it is important to use a robust experimental design. This means that the factors being tested in the experiment should be placed in a way that minimizes the effects of the subsurface heterogeneity on the results. Randomization is a key part of this process, as it ensures that the different treatments are assigned to wells in a way that minimizes the likelihood of bias due to differences in subsurface properties. Additionally, replication should be used to ensure that any differences observed in the results are due to the treatments, and not to natural variability in the subsurface. Finally, blocking can be used to group wells with similar subsurface properties together, so that any differences in the results are due to the treatments and not to the subsurface.

Aside from the three pillars of experimental design, there are several other factors to consider when crafting an efficient experimental design. For example, the experiment's spatial and temporal scales must be carefully considered (Müller, 2007). The experiment must be designed in such a way that it can be accurately monitored and data collected on time. Finally, the experiment should be designed in such a way that the results are generalizable and can be used to draw conclusions about the entire system.

The experimental design's conclusions are synonymous with its *objectives*. For example, the experiment described above could be designed to determine the optimal monitoring well spacing for monitoring aquifer temperature over time. The goals of spatial experiments are frequently to determine the spatial distribution of a variable of interest, such as groundwater flow distribution in a heterogeneous aquifer. The experiment's objectives may be exploratory, to estimate spatial trends or dependence, or multipurpose, depending on the prior information available (Müller, 2007).

A complex system. Our system, the subsurface medium, is the stage for infrequently well-understood physical and geological processes; it is also the product of such processes, as illustrated in Figure 1.1. Decisions resulting from our lack of knowledge of the physical and geological processes in the subsurface need to be informed with data or tools

that can help reduce the uncertainty. For example, when designing an experiment to acquire data to inform decision-making, one of the main challenges is to manage uncertainty associated with spatial data and subsurface processes, which can be difficult due to the complexity of the system (Müller, 2007; Ramgraber et al., 2021). The difficulty of making decisions related to the subsurface medium is exacerbated by the fact that stakeholders must balance conflicting objectives, models are large and complex, and data are heterogeneous (Scheidt et al., 2018).

To forecast the behavioral patterns of processes in the subsurface medium (e.g., groundwater flow, solute transport, geomechanical behavior), forward modeling is used to produce a prediction defined on a grid of locations in space and/or time, which is conditioned on the uncertainty of the model parameters (cf. §1.2). The computational expense rises with the complexity of the mathematical models used to describe the phenomenon and the number of parameters used to describe the subsurface models. This makes it challenging to do uncertainty analyses or sensitivity analyses of hydrological models.

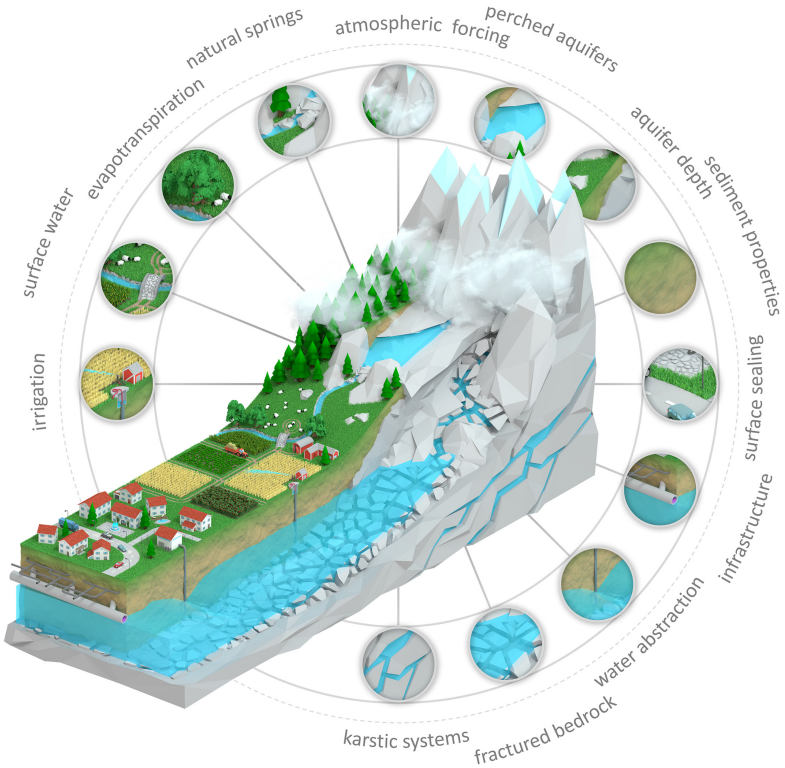

Figure 1.1: A mountainous hydrogeological system characterized by its complexity and interdependence. These aspects become sources of uncertainty for hydrogeological models when their presence, properties, and extent are insufficiently quantified (Ramgraber et al., 2021). Image taken from Ramgraber et al. (2021).

Deterministic approaches. In deterministic approaches, model outputs are compared to field data via forward modeling in order to calibrate the model unknown parameters. In probabilistic approaches, such as BEL, the prediction is directly estimated from the data (Hermans, 2017). In both strategies, data must be collected through an experimental process. When working with data that must be collected through an experimental process, as is often the case in geoscience, one of the most important considerations is where to gather new observations: designing a monitoring network necessitates careful consideration of many factors, especially when measurements are costly or resources are limited (cf. Chapter 4; Moghaddam et al. 2022). Controlling the experimental conditions for data acquisition is essential to maximize resource utilization and information gain because this process is typically expensive and/or time-consuming (Attia et al., 2018; Vilhelmsen and Ferré, 2018). This is known as optimal experimental design (OED), and it is commonly regarded as an optimization problem (Müller, 2007; Ryan et al., 2016).

Probabilistic approaches. Deterministic approaches have the drawback of not directly accounting for the uncertainty associated with model parameters and data. Probabilistic approaches, on the other hand, explicitly account for the uncertainty associated with model parameters and data by leveraging our prior knowledge of the system. In a system as complex as the subsurface medium, where uncertainty is high, probabilistic approaches are preferable to deterministic ones. Using the Bayesian paradigm to define the objective function in OED results in a so-called Bayesian OED (BOED) approach. Bayesian statistics generate the posterior distribution by combining prior knowledge of the model's unknown parameters with likelihood—the data's contribution to those parameters—from which conclusions about the model's unknown parameters can be drawn (cf. Chapter 4). The BOED strategies can be divided into static and sequential strategies: static BOED selects the optimal design when all experiments are conducted at once; sequential BOED, on the other hand, takes into account the fact that some experiments will be conducted in the very near future and others at later stages; the difference is that the latter takes into account the information already obtained from earlier experiments (Eidsvik et al., 2018, 2015; Hall et al., 2022).

How to leverage Bayesian statistics. There are several approaches to solving (B)OED problems, all of which involve maximization or minimization of a utility function (Chaloner and Verdinelli, 1995; Lindley, 1956; Ryan et al., 2016). However, the computational burden is significant because BOED requires solving the inverse problem (cf. §1.2) for each possible dataset, as the posterior distribution must first be estimated to compute the Bayesian utility function.

How should such problems be approached? The dataset that contains the most information, as determined by the amount of uncertainty that can be reduced via a data utility function, is considered to be the most informative. To find this dataset, an inverse problem must be stochastically solved for each candidate dataset. The computational expense of solving the inverse problem for each candidate dataset, often due to the computational expense of the forward model, makes using Markov chain Monte

Carlo (McMC) for BOED typically intractable. On the other hand, BEL provides a computationally efficient substitute for BOED (cf. Chapters 3; 4; 5).

A glance at the past. Previous examples of experimental design in hydrology include Kikuchi et al. (2015), who proposed a method called Discrimination-Inference, that can be used for both conceptual and predictive discrimination. The latter relies on the Kullback-Leibler divergence (cf. §1.3), which gauges the impact of accumulating more data on the prior and posterior probability distributions. To fill an input matrix using McMC sampling, Kikuchi and colleagues first calibrated numerous sets of model parameters conditioned on the available data. After obtaining a set of randomly sampled parameters, they used forward modeling to create data realisations in order to estimate the data utility function with Bayesian Model Averaging (BMA), which entails performing multiple simulations of the forward problem. In their approach, the posterior distribution of the prediction, which is at the heart of the data utility function, is estimated rather than computed directly by averaging the predictions from multiple forward problem simulations. In another study, Zhang et al. (2015) used a McMC algorithm to solve the BOED of sampling well locations and source parameters identification of groundwater contaminants, and defined a surrogate for the contaminant transport equation to reduce the computational burden of McMC. Tarakanov and Elsheikh (2020) developed a BOED methodology aimed at subsurface flow problems relying on a polynomial chaos expansion surrogate model for the utility function embedded in a McMC algorithm. These are a few examples of standard experimental design approaches employing McMC sampling, model surrogates, or a combination of the two.

A look ahead. In this contribution, we propose a methodology for solving BOED problems capable of incorporating the predictive power of different data types in a low-dimensional latent space, which alleviates the "curse of dimensionality" (Bellman, 1961) and reduces computational and memory demands. We model the relationship between predictor and target using a machine learning regression algorithm, which once trained, can predict the posterior distribution of the target for any given predictor. Similar to BMA, we use multiple simulations of the forward problem to estimate the posterior distribution of the target, but the posterior distribution is estimated using the regression model as opposed to averaging the results of multiple simulations.

This data-driven BOED strategy is therefore expected to be both time and memory efficient, as the forward simulation only needs to be run once in the case of static experimental design, where the sensor locations are fixed (e.g., Chapters 3; 4), or a few times in the case of dynamic experimental design, where the sensor locations are not fixed (e.g., Chapter 5). Moreover, our approach is more generalizable, which makes it possible to extend the proposed methodology to a larger number of problems than a specific, parameterized model can reflect.

This methodology consists of two steps: (1) learning the relationship between the predictor variables and the target variable from available training data, and (2) defining

an objective function based on the experiment goals and the predicted posterior distribution. There is no universal objective function that can be applied to all problems; the objective function must be defined by the user and tailored to the specific problem at hand (cf. Chapter 2). Objective functions may consist of anything that can be quantified from the posterior distribution, such as the logistic costs of the experiment, the time required to observe the data, and, of course, a measure of the predictive power of the data, or information gain. At the most basic level, the objective function can be defined as the predictive power of the input data, which can be quantified by simple metrics such as the mean squared error (MSE) or the mean absolute error (MAE) between few samples of the posterior distribution and true values (e.g., Chapter 4). More complex metrics rooted in information theory, such as the Kullback-Leibler divergence, can also be used to quantify the predictive power of the data, by comparing the posterior distribution of the target with the prior distribution (e.g., Chapter 5).

To avoid bias in the estimation of the optimal design, we use a k cross-validation method with multiple (k) folds in order to compute the objective function value for each fold. The objective is to examine the consistency of the experimental design across folds (cf. Chapters 3; 4; 5). If the inferred experimental design varies significantly from one fold to the next, there is a problem with our methodology: we may have chosen a poor prior distribution, the training data may be insufficient, the machine learning algorithm's hyperparameters might be inadequate, or the objective function may be poorly defined. If we have 250 training data, for example, we can use 200 data instances for training and 50 data instances for testing, and repeat this process $k = 5$ times, each time using a different set of 200 data instances for training and 50 data instances for testing. Cross-validation has the obvious drawback of increasing the number of training runs by a factor of k, which can be problematic for models whose training is computationally intensive (Goodfellow et al., 2016).

Monitoring network vs. experimental design. The astute reader may have noticed a similarity between the design of an experimental design for a field-scale geoscience problem and the design of a monitoring network for a related problem. Monitoring studies, also known as observational studies, are a type of research that involves observing and recording data about a specific phenomenon or group of subjects over time without attempting to manipulate or intervene in any way. The primary distinction between an experiment and a monitoring study is that an experiment involves manipulating one or more variables to see how they affect the outcome, whereas a monitoring study simply observes and records data with no attempt to influence the outcome. Monitoring studies are useful for comprehending the natural evolution of a phenomenon or condition, as well as identifying trends and patterns over time. They cannot, however, be used to establish cause-and-effect relationships because they do not involve the manipulation of variables.

However, there is a twist. In this contribution, we perform experimental design *in silico* to establish an optimal experimental design or an optimal monitoring network *in situ*. We accomplish this through the use of synthetic data generated by a forward

model, which is a simplified representation of the real-world problem over which we have complete control and can manipulate the variables at will. As a result, we can conduct experimental design *sensus stricto*, i.e., manipulate input variables to observe and identify the causes of output response changes.

For instance, in Chapter 3, we use synthetic data generated by a forward model to design an optimal experimental design for a tracer-injection experiment. We accomplish this by performing *in silico* experimental design to determine the best locations for the injection wells. In Chapter 4, we use synthetic data generated by a forward model to design an optimal experimental design for (i) an electrical resistivity tomography (ERT) survey and (ii) an optimal monitoring network for heat sensors in the subsurface. Therefore, Chapter 4 is a particular case where we perform experimental design *in silico* to design both an optimal experimental design and an optimal monitoring network *in situ*. Finally, in Chapter 5, we use synthetic data generated by a forward model to design an optimal monitoring network for a heat sensor network in the subsurface. In this case, we perform experimental design *in silico* to design an optimal monitoring network *in situ*.

Paradoxically, the very same reasons that make experimental design and monitoring studies similar also make it challenging to unify these disciplines under a single framework. This topic will be discussed further in the discussion (Chapter 6).

1.2 Inverse problems: a theory to infer unknown parameters from observations

What is an inverse problem? Inverse problems are ubiquitous in Earth Sciences, such as hydrology, geophysics and climate science, where the observation of a phenomenon requires starting from a model and inferring the values of its parameters. The goal of solving an inverse problem is to estimate the parameters of a model from a set of observations. For example, in hydrology, the inverse problem of interest could be to infer the hydraulic conductivity field from water level observations. Such a task can be challenging because it requires the solution of a partial differential equation (PDE), and more importantly, the model uses a numerical discretization of the underlying domain. It is usually not possible to obtain a closed-form solution, thus numerical solvers are used to predict water levels (e.g., Langevin et al. 2017). In addition, many properties of the underlying reservoir (e.g., porosity, permeability) depend on the lithology and are typically not homogeneously distributed. The latter is a particular challenge for Earth Sciences as real-world phenomenon are often complex, high-dimensional and nonlinear.

The predominant paradigm to solve an inverse problem is to use the observed data to estimate the parameters of a model that best fits the data, which is usually referred to as *calibration* in the hydrology community (Gupta and Nearing, 2014). This is usually done by minimizing an objective function usually defined as a metric of the discrepancy between simulated and observed data (Tikhonov and Arsenin, 1977). Although this approach is widely used, we show in Appendix B that it is not always the best strategy

as it frequently oversimplifies the heterogeneity of the subsurface, unless advanced constraints are introduced (e.g., Lopez-Alvis et al. 2021, 2022). For a detailed overview of the subject, we refer the reader to seminal works such as (Aster et al., 2013; Tarantola, 2005, 2006; Zhdanov, 2015).

How is it related to hydrology? In the field of hydrology, where many phenomena are based on the intricate interactions of numerous variables and processes, inverse problems are frequently ill-posed (e.g., Hsu and Yeh 1989; Stallman 1965a; Wallis 1965), which means that a solution to the inverse problem is not uniquely defined, or else cannot be recovered exactly from a given data set, i.e., different parameter values can produce the same output. The solution may also be unstable, which means that slight variations in the input data can cause large changes in the solution. Finally, the solution may not be continuously dependent on the data, which means that arbitrarily small changes in the data can cause large changes in the solution. This is typically due to a lack of observations, a lack of understanding of model "input" data such as forcing parameters, and the inherent nonlinearity of most physical processes. The problem then becomes determining the *best* solution, which must include our a priori beliefs and assumptions. This causes ill-posed inverse problems to be notoriously difficult to solve as they are subject to many degeneracies and trade-offs.

How to solve such problems? In practice, to solve an inverse problem, one needs to find a solution that best fits the data by adding extra information in the form of numerical constraints or *prior* information, which is referred to as *regularization* (Tikhonov and Arsenin, 1977) to make the inverse problem "well-posed." This is often accomplished by incorporating a stabilizing or regularization functional in the objective function to minimize, in which prior geological information can be used, as well as a data misfit term in a global cost function to minimize. If the problem is non-linear, minimization is an iterative process in which the data misfit is balanced with the *model functional* via the regularization parameter. It is typical to select model parameter estimates that allow the model to "adequately" replicate the observed input-state-output behavior of the actual system (Gupta and Nearing, 2014). Therefore, algorithms tend to depend on Occam's principle, which implies that the simplest solution is sought using a regularization functional. Occam's razor is frequently invoked to justify the use of a gradient or roughness operator, which results in smoothness constraint inversion. It can also be incorporated into advanced operators seeking compactness or minimal structure (e.g., Deleersnyder et al. 2022, 2021, Appendix B). Such approaches will produce either smooth or "simple" solutions (Loke et al., 2013), which are rarely geologically plausible (Linde et al., 2015; Zhdanov and Tolstaya, 2004); see also Appendix B.

These methods typically involve calibration of model parameters to a set of observations, and the resulting calibrated model is then used to perform predictions and decision-making. This deterministic approach of inverse modeling through model calibration is also termed model "tuning." It may be possible to specify plausible ranges for these parameters' values, but "correct" values are not a concept that has any real meaning given that they are largely abstractions related to the particular mathematical

forms chosen (Gupta and Nearing, 2014).

The aforementioned *model functional* refers to methods based on a set of assumptions regarding the physical processes that govern the system of interest: the subsurface medium, which is formed by geological processes that are not always well understood (Lavin et al., 2021). These processes are typically described by PDE systems solved numerically, which is known as *forward modeling*. Whitten (2018) reviews the history and challenges of forward modeling in geosciences, and Pham and Tsai (2017) and Singh (2018) review the subject in groundwater modeling.

Challenges and outlooks for inverse problems. There are a number of problems that arise when using such deterministic inversion techniques for decision-making. First, they do not offer any uncertainty quantification, which means that it is difficult to assess the reliability of the model's predictions (Kikuchi et al., 2015; Zhou et al., 2014). This can lead to make suboptimal decisions. Secondly, these techniques can be sensitive to outliers and small changes in the data, which can again lead to suboptimal decision-making. The modelers need to be aware of the potential problems with using these techniques so that they can make the best possible decisions. For example, Appendix B shows that the choice of the regularization parameter can have a significant impact on the results, and that the choice of the regularization parameter should be based on the uncertainty of the data. It illustrates how challenging and time-consuming it can be to obtain a single optimal image of the subsurface, considering the geological information, the data, and the regularization parameter.

In the hydrology community, Gupta and Nearing (2014) state that the emphasis started to shift away from "optimality" toward characterizing and reducing uncertainty (e.g., Beven and Binley 1992; Thiemann et al. 2001) and achieving consistency between the model and the system (Gupta et al., 1998, 2008; Martinez and Gupta, 2011). In hydrology, uncertainty is frequently viewed as stemming from a small number of primary sources: model parameters, measurements of boundary conditions, model structures, and observations used for model evaluation/inference (Nearing et al., 2016). In this paradigm-shifting wave, there has been a growing interest in using probabilistic methods for decision-making (e.g., Ferré 2020; Hermans 2017; Weijs et al. 2010). These methods offer a number of advantages over deterministic methods, including the ability to quantify uncertainty, to robustly handle outliers and small changes in the data, and to quantify bias. These stochastic approaches can and arguably should be utilized to evaluate the complete range of potential outcomes, enabling a thorough risk analysis to serve as the decision-making foundation (de Barros et al., 2012; Linde et al., 2017; Wang et al., 2022c; Zhou et al., 2014).

There is a chasm between how information is used in "classic" inverse problems and the way it is used in a probabilistic context. In the former, several ways exist to incorporate a priori information into the inversion process as described above (e.g., Caterina et al. 2014, Appendix B). One of the simplest approaches to incorporate prior information is to use a reference model in the regularization functional (Oldenburg and Li,

1994). However, an over-confidence in the reference model can lead to an erroneous solution (Caterina et al., 2014) and subsequent misinterpretation.

In probabilistic approaches, the information is incorporated through the prior distribution, which is a probability distribution that represents the modeler's beliefs about the parameters of the model. The Bayesian approach is one of the most popular approaches for doing this, in which the prior distribution is used to compute the posterior distribution, which is the probability distribution of the model's parameters given the observed data. In this approach, the posterior distribution quantifies the uncertainty in the model parameters and can be used to make decisions (e.g., Hermans et al. 2019; Schübl et al. 2022).

The main advantage of probabilistic inversion is that it can provide an estimate of the uncertainty in the model parameters, allowing for more informed decision-making. The main disadvantage is that these methods are computationally expensive and can necessitate significant computing resources. This can be a problem in the subsurface, where the parameter space is frequently high-dimensional and the forward model is computationally costly. Furthermore, due to the complexity of the methods, interpreting the results can be difficult. For example, when dealing with a high-dimensional parameter space, the results of a probabilistic inversion are typically presented in the form of a probability distribution, which can be difficult to interpret. In this context, it is important to find a way to interpret the results in a meaningful way so that reliable decisions can be made.

1.3 Information: a key ingredient for decision-making

Focal to this dissertation is the notion of *information*. This notion is too often overlooked in the Earth Sciences, especially in the context of inverse problems. However, there is a growing interest in this topic within the hydrology community, sparking debates about the application of information theory to the field (Goodwell et al., 2020; Gupta and Nearing, 2014; Kumar and Gupta, 2020; Nearing et al., 2020; Perdigão et al., 2020; Weijs and Ruddell, 2020). Some researchers have pioneered the use of information theory in hydrology, such as, e.g., Alfonso et al. (2010); Castillo et al. (2015); Franzen et al. (2020); Goodwell and Kumar (2017); Nearing and Gupta (2015); Nearing et al. (2013b); Weijs et al. (2010), to name a few.

Definitions. The concept of *information* is central to the Bayesian paradigm, which is the basis of the probabilistic methods used in this dissertation. However, the concept of information is too broad for a single definition to encapsulate it entirely. Information theory states that there is a quantity called the *entropy* that can be calculated for any probability distribution, and it has many characteristics that match the intuitive idea of what an *information measure* should be (Cover and Thomas, 2006). Nearing et al. (2016) recognize information as "the property of a signal that effects a change in our state of belief about some hypothesis."

In the context of Cox (1946) axioms (or, equivalently, Kolmogorov (1956) axioms, as demonstrated by proof), Shannon (1948) proposed what is arguably the most famous measure of information: that the expected amount of information in one random variable **d** about the value of another random variable **h** is measured as the expected Kullback-Leibler (KL) divergence to the marginal distribution over **h** from the distribution of **h** conditional on **d** (this is called the mutual information between **h** and **d**), as mentioned by Nearing and Gupta (2015). This *mutual information* is therefore a measurement of the amount of information that one random variable contains about another random variable. This concept of information differs from the common concept of information in Earth Sciences, where information is usually referred to as the sum of all data, including their interpretation.

Definition 3 (Shannon entropy) *Let $d = \{d_1, d_2, \ldots, d_n\}$ be a finite set of possible outcomes of a random variable* **d**. *The Shannon entropy of* **d** *is defined as*

$$\mathcal{S}(d) = -\sum_{i=1}^{n} p(d_i) \log p(d_i), \tag{1.1}$$

where $p(d_i)$ is the probability of outcome d_i (Shannon, 1948).

Definition 4 (Mutual information) *Let* **h** *and* **d** *be two random variables. The mutual information between* **h** *and* **d** *is defined as*

$$\mathcal{I}(h; d) = S(h) + S(d) - S(h, d), \tag{1.2}$$

where $S(h)$ is the Shannon entropy of **h**, *$S(d)$ is the Shannon entropy of* **d**, *and $S(h, d)$ is the joint entropy of* **h** *and* **d**. *The mutual information between a random variable and itself is the random variable's entropy. Entropy is sometimes referred to as self-information for this reason (Cover and Thomas, 2006).*

Definition 5 (Kullback-Leibler divergence) *Mutual information is a variant of the more general quantity known as relative entropy, also known as the Kullback-Leibler (KL) divergence, which measures the distance between two probability distributions. Let* **h** *and* **d** *be two random variables. The Kullback-Leibler divergence between the distribution of* **h** *and the distribution of* **d** *is defined as (Cover and Thomas, 2006)*

$$\mathcal{D}_{KL}(h||d) = \sum_{i=1}^{n} p(h_i) \log \frac{p(h_i)}{p(d_i)}. \tag{1.3}$$

Application to Earth Sciences. How does this relate to the inverse problems in Earth Sciences? In an inverse problem, the goal is to estimate the parameters or outputs of a model from a set of observations. The unknown model parameters or outputs can be considered as the random variable **h**, and the observed data can be considered as the random variable **d**. The entropy of the model parameters or output quantifies the quantity of information it contains, whereas the entropy of the observed data quantifies the quantity of information contained in the observed data. The amount of shared

information is quantified by the mutual information between the two. More information can be gleaned about the model parameters or outputs the more information is contained in the observations. Similarly, the greater the amount of information contained in the model parameters or outputs, the greater the amount of information that can be captured by observation.

From mathematical theory to practical application. From a more practical point of view, Eidsvik et al. (2015) present a decision-theoretic framework in which is defined the Value of Information (VOI), which is the expected utility of the information gained from an experiment. It encompasses the notions that the information gained must be (i) informative, (ii) result in material (practical) decisions, and ultimately (iii), lead to economic decisions. These notions are sequentially conditional, such that, in order to make an economic decision, the information must first lead to a material decision, which depends on the information gained. The idea of VOI was applied and developed in works such as Liu et al. (2012) to the remediation of groundwater sites, Trainor-Guitton et al. (2013) to monitor the detection of CO_2 leaks, Nenna and Knight (2014) to evaluate the benefits of acquiring geophysical data as part of a groundwater management strategy, and a geophysical perspective on VOI is presented by Trainor-Guitton (2014). In addition, a new concept called efficacy of information (EOI), which is similar to VOI but without financial rewards or costs, was recently introduced by Caers et al. (2022).

As a result, the concept of information is central to this dissertation, and it is necessary to quantify the information contained in the model parameters, model outputs, observed data, and mutual information in order to obtain an objective measure of the amount of information shared by the model and observed data; indeed, quantifying the information is the first step in the VOI framework in order to lead to a practical decision. Furthermore, because the random variables in a Bayesian analysis represent known or unknown information, the concept of information is critical to the application of Bayesian methods to inverse problems. The section that follows introduces the methods used in this dissertation to extract, manipulate, and analyze information.

1.4 Machine Learning as a tool for making predictions in hydrology

Introduction. Machine Learning (ML) is a broad field of research that aims to develop algorithms that can learn from *data*–referring to information contained in observations, measurements, or other knowledge that is available–making it a logical choice for probabilistic inverse problems in light of the arguments presented above. We can first define our terminology. In this dissertation, we focus on the application of ML to regression problems in hydrology and geophysics, but the interested reader is referred to Murphy (2012), Goodfellow et al. (2016) or Géron (2022) for a more comprehensive introduction to ML. We specify that Deep Learning (DL) algorithms are a specific class of ML algorithms that use multiple layers of mathematical operations to learn the patterns in data (Schmidhuber, 2015), but we will refer to them as ML algorithms for simplicity.

History. Today, the amount of attention that ML receives in Earth Sciences is increasing, and the number of papers that apply ML to Earth Sciences problems is paralleling the number of papers that apply ML to other scientific fields. There is a plethora of papers that apply ML to geoscience problems, as reviewed by Bergen et al. (2019); Dramsch (2020); Karpatne et al. (2019); Lary et al. (2016); Sun et al. (2022).

As noted by Sun et al. (2021), ML has been used in hydrology for decades (e.g., Coulibaly et al. 2000; Dawson and Wilby 2001; lin Hsu et al. 1995; Maier et al. 2010; Sun 2013; Zealand et al. 1999). Furthermore, Sun et al. (2021) emphasize that the current wave of hydrological ML applications has greatly benefited from more accessible cyber-infrastructures and a new breed of deep learning algorithms, bolstered by the exponential growth of Earth observation data (Peters-Lidard et al., 2017; Shen, 2018; Sun and Scanlon, 2019). These developments allow for the importation of priors or domain knowledge into ML models, the exportation of knowledge from learned models back to the scientific domain, the generation of a vast amount of synthetic data, the quantification and analysis of uncertainty in models and data, and the inference of causal relationships from the data (Lavin et al., 2021).

Discord within the community. Kratzert et al. (2019a) point out that relative advantages of data-driven versus process-driven models have been the subject of a protracted debate in the field of hydrology (see Klemeš 1986, for example). Sellars (2018) noted that "Many participants who have worked in modeling physical-based systems continue to raise caution about the lack of physical understanding of machine learning methods that rely on data-driven approaches" in their summary of "Big Data and the Earth Sciences: Grand Challenges" workshop. Kratzert and colleagues note that it is commonly argued that data-driven models may perform worse than models with explicit process representations under conditions different from training data (e.g., Kirchner 2006; Milly et al. 2008; Vaze et al. 2015).

Despite such contentions, Kumar and Gupta (2020) note that several studies have previously shown that machine-learning-based models can effectively outperform conventional Earth system models at the task of estimating key states and fluxes for which those models were intended, where information loss typically results from a combination of parameter error and model structural inadequacy (Nearing et al., 2018b).

I opine that ML is the driving force behind the advancement of data-driven, probabilistic, and information-theory-centered approaches to inverse problems in Earth Sciences. As will be shown in the following sections, the methods presented in this dissertation are a direct application of ML to inverse problems in Earth Sciences, but they are also hybrid in nature, as the ML models are fed by physically-based simulations outputs.

How can BEL leverage ML? Central to the BEL framework is the use of an appropriate ML algorithm to learn the relationship between the input and output variables. There are numerous regression techniques widely available in mainstream ML/DL packages, e.g., `scikit-learn` (Pedregosa et al., 2011), `Tensorflow` (Abadi et al., 2015),

`PyTorch` (Paszke et al., 2019), and providing a comprehensive review of all of them is beyond the scope of this dissertation. Additionally, ML classification techniques, which are used to predict a discrete label, are not considered here, as opposed to regression techniques, which are used to predict a continuous value. Furthermore, we are examining the algorithms themselves and not their use as surrogate models in the context of uncertainty quantification.

Regression techniques are used to predict continuous values by establishing a mathematical relationship between inputs and outputs. The aim is to find a function f that can accurately approximate the relationship between the inputs $\mathbf{d}$ and the outputs $\mathbf{h}$, such that f can predict the output $\mathbf{h}$ when given an input $\mathbf{d}$. The function f is often referred to as a *model*, and it is typically assumed to be a parametric function, so that the parameters of the model can be estimated from the observations. The observations are often called *training data*, and they consist of pairs of inputs and outputs. The goal is to find the parameters of the model that minimize the difference between the outputs predicted by the model and the observed outputs.

Present and future of ML in hydrology. In recent years, applications of the Bayesian paradigm to ML have been intensively studied in response to the growing need for principled uncertainty reasoning in machine learning systems as they are progressively implemented in safety-critical domains (Meinert et al., 2022). Uncertainties encountered when making predictions can be classified into two categories: aleatoric and epistemic uncertainties. Typically, aleatory uncertainty is attributed to data noise, whereas epistemic uncertainty is attributed to model parameter and model structure uncertainty (Gal, 2016; Kiureghian and Ditlevsen, 2009; Nearing et al., 2016). Epistemic uncertainty can be reduced by increasing the amount of data available for training, but aleatoric uncertainty can only be reduced by using higher precision sensors, for example (Gal, 2016).

In this dissertation, we address the problem of probabilistic regression in subsurface hydrology, where the goal is to learn the conditional probability distribution $p(\mathbf{h}|\mathbf{d})$ of a dependent variable $\mathbf{h}$ given a set of input variables $\mathbf{d}$. Conditional density estimation (CDE) generally refers to the process of inferring a probability density function (*pdf*) from a set of empirical observations (Rothfuss et al., 2019), thereby allowing for a full description of the uncertainty associated with the predicted output given a certain input. CDE seeks to capture the statistical relationship between a conditional variable $\mathbf{d}$ and a dependent variable $\mathbf{h}$ by modeling their conditional probability $p(\mathbf{h}|\mathbf{d})$ given a set of empirical observations $\mathbf{d} = d_1, d_2, \ldots, d_N$ and $\mathbf{h} = h_1, h_2, \ldots, h_N$, where N is the number of observations.

An overview of ML algorithms. ML algorithms are particularly well suited to modelling nonlinear data, and are capable of modelling complex interactions between inputs and outputs. The adaptability of ML algorithms permits the modeling of nonlinear and even non-monotonic functions, which is particularly advantageous for hydrological applications due to the inherently nonlinear nature of the hydrological system. Furthermore,

it is feasible to fit ML models in such a way that strongly nonlinear interactions between model inputs and outputs are taken into account and the models' internal representations intrinsically support estimating distributions (Klotz et al., 2022).

Only a few regression techniques allow for direct probabilistic predictions (e.g., Gaussian Process Regression). However, some ML algorithms used for (deterministic) regression, such as gradient boosting, neural networks, and Long Short-Term Memory (LSTM) networks, have been modified, hybridized, or extended to provide probabilistic predictions thereby providing a mechanism to quantify the uncertainty (e.g., Dillon et al. 2017; Duan et al. 2019; Klotz et al. 2022; Li et al. 2021).

In fact, Canonical Correlation Analysis (CCA), which has been widely used in BEL, is not *per se* a probabilistic regression technique. To allow for probabilistic predictions, it is typically hybridized with algorithms that allow conditional probability sampling on computed canonical variates, as will be detailed in Chapter 2.

The choice of appropriate ML techniques is application-dependent and heavily contextualized. No ML algorithm is the best in all contexts, and the "no-free-lunch" theorem (Wolpert, 1996) states that there is no single ML algorithm that is the best for all circumstances. The theorem is consistent with the fact that certain ML algorithms may work better than others for specific problems. In addition, the "no-free-lunch" theorem can be interpreted in a number of different ways, but one interpretation is that it asserts that every supervised learning algorithm contains some form of implicit prior (Jospin et al., 2022). If they are applied appropriately, Bayesian methods will at the very least make the prior explicit (Jospin et al., 2022).

We briefly review the most relevant ML techniques for regression, and discuss their advantages and disadvantages.

Artificial Neural Networks (ANNs): An ANN is a machine learning algorithm used to model complex data patterns (Goodfellow et al., 2016). Neural networks are made up of a large number of interconnected processing nodes, or neurons, that can learn to recognize patterns in input data. A neural network's architecture is typically a series of layers, with each layer composed of a set of interconnected neurons. The input data is received by the first layer of neurons, and each subsequent layer transforms the data before passing it on to the next layer (this is referred to as a *feed-forward neural network*). The neural network's output is produced by the final layer of neurons. A neural network's training procedure is typically a supervised learning process in which the neural network is presented with a set of training data and the desired output for each data point. The neural network adjusts the weights of the connections between neurons to minimize an objective function (usually referred to as the *loss* function), which can be defined as a distance metric between the predicted and desired outputs.

Advancements in recent years have made it possible to leverage Bayesian Inference in their training process. Bayesian Neural Networks (BNNs; Jospin et al. 2022) are able to

quantify uncertainty by assigning a probability distribution to the weights of the neural network, yielding a posterior distribution over the neural network's output. The layers of a BNN are connected by weights with a pre-defined probability distribution, typically Gaussian. Once the parameters of the optimization problem are learned, the model can be sampled multiple times to generate different outputs. This, in turn, can be used to construct the marginal posterior distribution of the model's output, which can be used to quantify the uncertainty of predictions. Khan and Coulibaly (2006); Zhang et al. (2011); Zhang and Zhao (2012), for example, used BNNs for surface hydrological forecasting. It is a powerful tool for probabilistic predictions of both multi- and univariate datasets. A disadvantage of BNNs and neural networks in general, is that they are computationally intensive compared to other ML algorithms and require the choice of several hyperparameters.

Canonical Correlation Analysis: CCA (Hotelling, 1936), is a statistical technique that can be used to assess the linear relationship between two sets of variables. CCA can be used to generate a predictive distribution for the model's output, which can then be used to quantify the uncertainty of predictions. CCA is strongly related to mutual information (Borga, 1998), making it a natural choice for BEL. CCA, in addition to BEL, has long been recognized as a useful technique for hydrological forecasting (Ouarda et al., 2001; Ribeiro-Corréa et al., 1995; Rice, 1972; Torranin, 1972). However, CCA is especially suited to multivariate dataset, and its use for univariate dataset is not advised.

Gaussian Process Regression (GPR): GPR is a non-parametric ML technique that can be used to learn a latent function from data. GPR can be used to generate a predictive distribution for the model's output, which can then be used to quantify the uncertainty of predictions. Sun et al. (2014) used GPR to forecast monthly streamflow, and Yang et al. (2018) used it in a hydrologic model to perform a Bayesian sensitivity analysis. One major downside is that although they can be trained with a multi-dimensional dataset, mainstream GPR implementations (e.g., `scikit-learn`) can only be used to predict a single variable at a time. Multi-output GPR (MOGP) is more complex than single-output GPR and is the subject of active research (de Wolff et al., 2021; Liu et al., 2018). It is worth noting that GPR is commonly referred to as kriging in the geostatistics community, while MOGP is referred to as co-kriging (Goovaerts, 1997; Müller, 2007). Additionally, Neal (1996) demonstrated that GPR is equivalent to single-layer feed forward ANNs with infinite hidden nodes but is resistant to overfitting (Nearing et al., 2013a).

LSTM: LSTM is a neural network architecture that is well-suited for modeling time-series data, and may be an excellent candidate for modeling dynamical systems such as watersheds (Kratzert et al., 2019b). Lees et al. (2022) and Nearing et al. (2022) investigated the information captured by the LSTM state vector and compared two approaches for ingesting near-real-time streamflow observations for rainfall-runoff modeling. Li et al. (2021) suggested using Bayesian LSTM with stochastic variational inference to estimate model uncertainty in process-based hydrological models. Klotz et al. (2022) trained their models by maximizing the log-likelihood function of the observations according to the

predicted mixture distributions, and benchmarked different model setups. One downside of LSTMs is that they are computationally expensive and best suited for time-series (sequential) data.

Natural Gradient Boosting for Probabilistic Prediction (NGBoost): NGBoost (Duan et al., 2019) is a gradient boosting framework that can be used to quantify the uncertainty of predictions. Shen et al. (2022) applied NGBoost for probabilistic runoff predictions, and remarked its "suitable performance," although their approach was hindered by model limitations and complexity of the actual runoff process. Başağaoğlu et al. (2021) developed and hybrid NGBoost-XGBoost (extreme gradient boosting, Chen and Guestrin 2016) to predict evapotranspiration. One downside of NGBoost is that it can only be used to predict a single variable at a time.

In general, there is a trade-off between the capability to generate multidimensional predictions and the computational expense of training the model. For example, NGBoost trains quickly but can only predict one variable at a time, whereas BNNs are computationally intensive but can predict multiple variables at the same time.

Machine learning is not a panacea for uncertainty quantification. There are in fact many similarities between training a ML model and calibrating a hydrological model:

- Both hydrological and ML models often require extensive tuning of hyperparameters in order to achieve accurate predictions.
- Both hydrological and ML models models can overfit the data.

Tying it all together. But then, why use machine learning at all? One alternative to estimate the posterior distribution of the model's output is to use Markov Chain Monte Carlo (McMC) methods (Laloy and Vrugt, 2012; Vrugt, 2016), or ensemble methods (e.g., Emerick and Reynolds 2013; White 2018). The posterior distribution is approximated in McMC methods by sampling over the parameter space, that is, by running the forward model multiple times with different parameter values and rejecting or accepting the results based on the likelihood function. Full-scale hydrological processes are computationally expensive to simulate, especially when statistical summaries require multiple realizations. Training a machine learning model on a large dataset and incorporating multiple realizations can be much faster.

Therefore, we do not advocate replacing traditional statistical and/or physics-based methods with machine learning. Instead, we propose that it should be used in conjunction with these techniques in order to better comprehend the underlying processes, extract additional data, and (hopefully) improve the accuracy of predictions.

Once an uncertainty quantification is available, it is possible to investigate how different input data (predictors) influence the output (target) and how much uncertainty is associated with the prediction, resulting in the experimental design. It is important

to emphasize at this point that the uncertainties' estimations are not the uncertainties themselves, but rather estimates of the uncertainties, and thus subject to uncertainties themselves (Klotz et al., 2022).

1.5 Making applications more tractable for experimental design

> *Modern statisticians are familiar with the notion that any finite body of data contains only a limited amount of information on any point under examination; that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination: that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue.*
>
> Ronald A. Fisher

The process of Bayesian experimental design is often made difficult by the large number of parameters that must be considered. This is particularly true for hydrological applications, which frequently involve a large number of parameters. Markov Chain Monte Carlo, the standard method of Bayesian inference for ED, is not always appropriate due to its computational complexity, which increases with the number of parameters. Tractable in this context means that the application of experimental design methods is feasible given the limitations of computational resources. Dimensionality reduction techniques can be used to decrease the number of parameters, thereby making ED more tractable. Additionally, surrogate models can be used to reduce the need for iterative runs and increase the efficiency of ED. Finally, Bayesian model averaging (BMA) can be used to reduce prior distribution complexity and test multiple scenarios. In this section, we will examine the most prevalent techniques for dimensionality reduction and model simplification.

1.5.1 Overview of dimensionality reduction techniques

Dimensionality reduction is the process of reducing the number of variables in a dataset while retaining as much information as possible. This can be done for a variety of reasons, such as reducing the amount of data required to store or process a dataset, or improving the interpretability of the data by reducing the number of variables that need to be considered. Furthermore, dimensionality reduction can be performed prior to the application of machine learning algorithms to reduce overfitting and improve the performance of the algorithms (Srivastava et al., 2014).

As noted by Laloy et al. (2015), high-parameter dimensionality makes inversion of groundwater flow and transport data difficult, but dimensionality reduction techniques can help overcome these difficulties, as also pointed out by Zhou et al. (2014). This realization is not new, as dimensionality reductions techniques have been used in hydrology since the advent of the digital computer in the 1960s (e.g., Haan and Allen 1972; Wang and Huber 1967).

There are a variety of dimensionality reduction techniques, each with its own advantages and disadvantages (Cook, 2022):

Principal component analysis (PCA): PCA is a linear dimensionality reduction technique that projects the data onto a lower-dimensional space in a way that maximizes the variance of the data. It has been used in several BEL applications in hydrology (e.g., Hermans et al. 2019; Satija and Caers 2015; Thibaut et al. 2021b). However, PCA may not be able to capture non-linear structures in the data. It has been extensively applied across fields and decades, including research on water resources, e.g., Diaz et al. (1968); Haan and Allen (1972); Keating et al. (2010); Saad and Turgeon (1988); Tripathi and Govindaraju (2008); Wang et al. (2022b); Zhao et al. (2022); Zhao and Luo (2020).

Independent component analysis (ICA): ICA is a non-linear dimensionality reduction technique that projects the data onto a lower-dimensional space in a way that maximizes the statistical independence of the variables. Examples of applications include Westra et al. (2007) who compared ICA with PCA to model multivariate hydrological series, Westra et al. (2008) to forecast multivariate streamflow, and Middleton et al. (2015), who used ICA to assess sediment-water interface temperature variability.

Factor analysis (FA): FA consists in decomposing a multivariate dataset into a set of underlying factors that explain the observed variables. It is a linear technique similar to PCA, and has been used in hydrology since the 1960s (e.g., Dawdy and Feth 1967). However, Wallis (1968) pointed out that "factor analysis, if used in the classical manner, will never be of great value for hydrologic analysis," and they may have been right. In fact, its use in hydrology was already critiqued in Matalas and Reiher (1967). It found some application in reservoir losses estimation Knisel (1970) and groundwater chemistry analysis (Ashley and Lloyd, 1978; Dawdy and Feth, 1967). More recently, Sarker et al. (2022) used FA to identify the major hydrogeochemical factors governing groundwater chemistry in the coastal aquifers of Southwest Bangladesh.

Linear discriminant analysis (LDA): LDA is a linear dimensionality reduction technique that projects the data onto a lower-dimensional space in a way that maximizes the separability of the classes. It has been used in hydrology to predict groundwater redox status on a regional scale (Close et al., 2016), to predict groundwater redox conditions resulting from denitrification (Wilson et al., 2018), and to assess the spatiotemporal groundwater quality in a coastal aquifer (Amiri and Nakagawa, 2021).

Isometric feature mapping (Isomap): Isomap (Tenenbaum et al., 2000) is a non-

linear dimensionality reduction technique that projects the data onto a lower-dimensional space in a way that preserves the geodesic distances between the points. Böttcher et al. (2014) presented a novel approach to investigate ground and surface water dynamics. They mention that "Isomap is a promising technique due to its flexibility to map both linear and non-linear relationships." I agree with this statement, however the use of isomap in hydrology is limited. Kanishka and Eldho (2017) used isomap to classify watershed in combination with hydrometerological data, and Liu and Hu (2019) for environmental data analysis.

t-distributed stochastic neighbor embedding (t-SNE): t-SNE is a non-linear dimensionality reduction technique that projects the data onto a lower-dimensional space in a way that preserves the local structure of the data. Tang and Carey (2022) believe t-SNE is underutilized in hydrological applications and has significant potential for extremely large datasets. According to their findings, t-SNE outperforms PCA in terms of separability of annual daily hydrographs from different flow regimes when it comes to grouping watersheds. Mazher (2020) and Liu et al. (2021) have previously demonstrated the utility of t-SNE for visualization of high-dimensional spatiotemporal hydrological data, as well as for identifying clusters and defining geochemical zones.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP): UMAP (McInnes et al., 2018) is a non-linear dimensionality reduction technique that projects the data onto a lower-dimensional space in a way that preserves the local structure of the data. It was adopted in some hydrology applications (e.g., Kratzert et al. 2019b; Mazher 2020; Prakaisak and Wongchaisuwat 2022). According to Mazher (2020), when non-linear trends are expected and local trends are much more significant, t-SNE and UMAP are better to use; however, UMAP is computationally more efficient. For linear datasets, on the other hand, PCA captures global trends better.

Multidimensional Scaling (MDS): MDS is a non-linear dimensionality reduction technique that projects the data onto a lower-dimensional space in a way that preserves the pairwise distances between the points (Cox and Cox, 2008). Lopez-Alvis et al. (2019) used MDS to extract salient features from geological scenarios for representing uncertainty in subsurface connectivity. They found that the extracted features had a strong correlation with the structural parameters, indicating that MDS could be used to reduce uncertainty in subsurface features affected by structural parameters. Hermans et al. (2015a) used MDS in a workflow to assess model and geological scenario posterior uncertainty.

Summary of dimensionality reduction techniques. To summarize, dimensionality reduction in hydrology is a powerful technique that can be used to improve the interpretability of data, or improve the performance of machine learning algorithms. There is no one best dimensionality reduction technique for all datasets and all goals, and it is important to select the dimensionality reduction technique that is most appropriate for the specific dataset and goal.

1.5.2 Model simplifications approaches

Utilizing surrogate models is another strategy for decreasing the computational expense of experimental design. Surrogate models are forward model approximations that are used to speed up the simulation process and bypass the numerical simulations. Razavi et al. (2012) proposed a comprehensive review of surrogate modeling in the water resources field. The surrogate model of a process that requires a lot of computing power is used to estimate an objective function, constraints, or both. According to Razavi et al. (2012), there are two main types of surrogate modeling: high-fidelity response surface models and low-fidelity models. Fidelity refers to the degree of realism of simulation models.

High-fidelity response surface models replicate the original model's numerical results. This approach employs techniques like kriging (Baú and Mayer, 2006; Garcet et al., 2006), artificial neural networks (Kourakos and Mantoglou, 2009; Yan and Minsker, 2006), radial basis functions (Regis and Shoemaker, 2005), and polynomial chaos expansion (Laloy et al., 2013; Tarakanov and Elsheikh, 2020) to determine a relationship between model parameters and one or more model response variables using statistical models or empirical data-driven models.

Low-fidelity models are physically based. In essence, they are streamlined, less accurate versions of their computationally intensive parent models. Low-fidelity models must be reasonably close to the response of the original model in order to be used in practice.

According to Razavi et al. (2012), surrogate modelling loses its usefulness or even becomes impractical as the number of model variables rises, such as in heterogeneous subsurface reservoirs, which lowers the accuracy of the analysis. Once a suitable surrogate model is identified, stochastic inversion and experimental design issues can be effectively resolved at a minimal computational cost. The approximation, however, may result in a sizable bias in the prediction, leading to an inaccurate estimation of the data utility function (Asher et al., 2015; Babaei et al., 2015; Laloy et al., 2013; Razavi et al., 2012; Tarakanov and Elsheikh, 2020; Zhang et al., 2015, 2020).

1.6 Bayesian Evidential Learning–a new tool for experimental design

> *By a small sample, we may judge of the whole piece.*
>
> Miguel de Cervantes, Don Quixote

Overview. At the heart of this dissertation is the Bayesian Evidential Learning (BEL) framework, which is a probabilistic prediction framework originally developed for predic-

tion problems in Earth Sciences (Scheidt et al., 2015), and formally introduced in Scheidt et al. (2018). Prediction problems in Earth Sciences have traditionally been linked with inversion problems, as it was long assumed that before making a prediction with a model, one had to derive a set of model parameters (deterministically or probabilistically). BEL has introduced a paradigm shift in solving prediction problems by avoiding inversion and replacing it with a machine learning approach fed by physically-based simulations.

This section intends to provide a brief overview of the BEL framework, while the technical details and mathematical formulations will be described in Chapter 2. The interested reader is referred to Scheidt et al. (2018) for a more comprehensive introduction. It is a framework that integrates Bayesian statistics, multivariate statistics, and machine learning, with the goal of quantifying the prediction's uncertainty, as introduced in sections §1.4 and §1.5.

BEL is not an algorithm, but rather a set of rules for estimating the uncertainty of predictions. It performs the Bayesian inference (usually within a low-dimensional latent space) using a direct relationship between predictor $\mathbf{d}$ and target $\mathbf{h}$ learned from a training set sampled from a user-defined prior distribution of model parameters we denote by ω.

We rarely have multiple realisations of $\mathbf{d}$ and $\mathbf{h}$ in real-world geoscience applications, as it would be prohibitively expensive to collect such data. Nonetheless, we can rely on realistic distributions of the subsurface model's parameters ω to generate, via forward modeling, the prior distribution of $\mathbf{d}$ and $\mathbf{h}$, from which a statistical relationship can be learned (Hermans et al., 2016; Scheidt et al., 2018), as illustrated in Figure 1.2. The main benefit of BEL is that the inferred predictor-target relationship can be applied to any dataset that is consistent with the prior distribution.

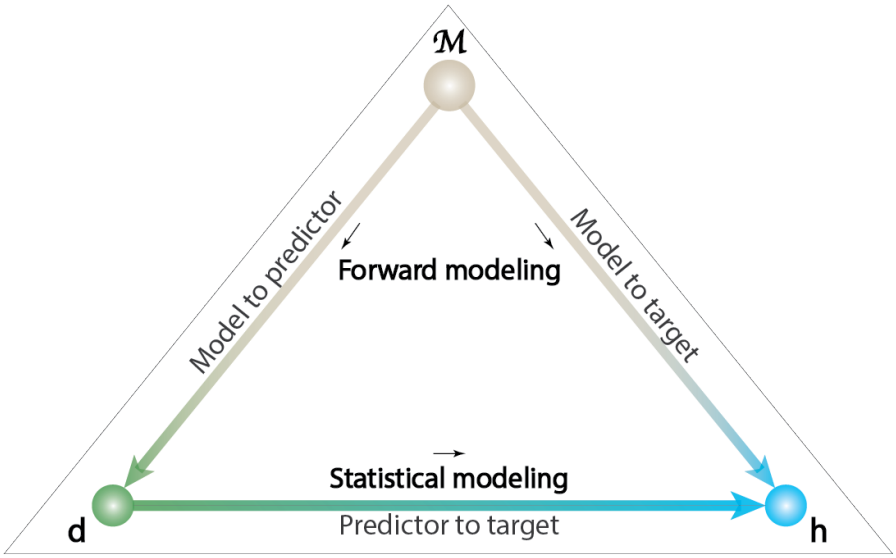

Figure 1.2: The BEL paradigm. **M** is a conceptual model, **d** is the set of all possible predictors, and **h** is the set of all possible targets.

The BEL framework is similar to Simulation-Based Inference (SBI), also known as "likelihood-free" inference, which has been applied mostly in physics and astrophysics, but not in Earth Sciences (e.g., Cranmer et al. 2020; Hermans et al. 2021a; Lavin et al. 2021; Lueckmann et al. 2021). Other techniques, such as Approximate Bayesian Computation (ABC), the method of simulated moments, and indirect inference, also fall under the umbrella term "likelihood-free" (Gutmann et al., 2018). Those methods all share the same premise, which is to perform inference about ω by identifying values that produce simulated target **h** that resemble observed predictor **d** (Gutmann et al., 2018).

BEL is not limited to the inference of ω from **h**, but can also be used to infer **h** from **d**, or **d** from **h**. What those methods have in common with BEL is that they use a generative model, or simulator, to generate synthetic targets **h** from ω.

The BEL framework as applied in this dissertation consists of the following steps:

1. Sample the prior distribution of the parameters ω to generate a training set of pairs of **d** and **h** via forward modeling.

2. Learn the statistical relationship between **d** and **h** using a machine learning algorithm.

3. Use the learned relationship to estimate the posterior distribution of **h** given **d**.

Datasets encountered in geoscience can be high-dimensional, sparse, plagued by noise, and possibly by multicollinearity. Due to the curse of dimensionality (Bellman, 1961), estimating the entropy or mutual information between probability distributions of high-dimensional random variables is challenging (Nearing et al., 2013b). To alleviate these

problems, the essence of BEL is to work in a reduced-dimensional space, although this is not required. To reduce the dimensionality of the problem, we can use a dimensionality reduction technique such as Principal Component Analysis (PCA) to project the data into a lower-dimensional space (Meloun and Militký, 2012).

A natural choice for the machine learning algorithm is Canonical Correlation Analysis (CCA; Meloun and Militký 2012). It is a relatively simple and linear method that can be used to find the relationship between two sets of variables using a linear combination of the variables in each set. While PCA seeks to maximise the reconstruction of original variables, CCA finds underlying correlations between pairs of $\mathbf{d}$ and $\mathbf{h}$ while transforming them to new, maximally correlated Canonical Variates (CVs). Given that the CVs' relationships describe the behavior of each target dimension for each predictor dimension, an observed data point can be used to infer the posterior distribution of each unknown target dimension.

To accomplish this, we project the observed data $\mathbf{d}_{obs}$ onto the data CV axes using the same transformations we derived for the predictor samples. Then, for each target dimension, we use the corresponding bivariate distribution and the observed CV to derive the posterior distribution of that dimension. Chapter 3, for example, demonstrates how this can be easily accomplished using multivariate Gaussian inference, provided that the CVs' bivariate distributions are both Gaussian and linearly correlated. Kernel Density Estimation (KDE) is another method for approximating the bivariate distribution for each CCA dimension without requiring such assumptions to be verified (e.g., Hermans et al. 2019; Michel et al. 2020b). KDE, on the other hand, has two parameters that must be adjusted: the kernel type, which defines the shape of the distribution at each coordinate, and the kernel bandwidth, which describes the size of the kernel at each position. Transport maps (Spantini et al., 2018) form another method that allows to calculate the posterior distribution.

BEL's efficiency has been demonstrated with extensive synthetic validation, and also against rejection sampling (Scheidt et al., 2015), McMC algorithms (Michel et al., 2022a, 2020b), field data (Hermans et al., 2019), and experimental design (cf. Chapter 3 to 5). Previous research has shown that BEL can estimate the posterior distribution of targets in a variety of contexts, including geothermal systems (Athens and Caers, 2019; Hermans et al., 2019, 2018), contaminant transport (Satija and Caers, 2015; Scheidt et al., 2015), geophysical inversion (Hermans et al., 2016; Michel et al., 2020b). In addition, the BEL framework has been successfully applied to a range of subsurface field cases, such as groundwater, shallow and deep geothermal and oil/gas predictions (Park and Caers, 2020; Pradhan and Mukerji, 2020; Tadjer and Bratvold, 2021).

Software implementation. Since its inception (Scheidt et al., 2015), the BEL framework has been implemented in few open-source software packages. Thibaut and Ramgraber (2021) developed a Python package called `SKBEL` (scikit-BEL) that implements the BEL framework using the `scikit-learn` library (Pedregosa et al., 2011). It is dis-

tributed under the 3-Clause BSD license, and can be found on GitHub[1] and on the Python Package Index (PyPI)[2]. `SKBEL` is a modular framework that allows users to easily implement the BEL framework in their own applications. In Chapter 3, we use `SKBEL` to estimate the uncertainty of predictions in wellhead protection area prediction, Chapter 4 applies it to compare the information content of different experimental designs in a four-dimensional inverse problem, and Chapter 5 uses it to predict exchange fluxes between groundwater and surface water in a sequential Bayesian optimal experimental design framework.

`pyBEL1D` is another open-source software package that implements the BEL framework (Michel et al., 2022b). It is a Python program that uses geophysical data to perform stochastic 1D imaging of the subsurface, and has been used to estimate the uncertainty of model parameters in a variety of geophysical problems, including surface nuclear magnetic resonance (Michel et al., 2020b) and interpretation of surface waves dispersion curves (Michel et al., 2022a).

Yin et al. (2020) shared `AutoBEL`, a Python implementation of the BEL framework that provides an automated method for quantifying uncertainty and updating geological models using borehole data for subsurface developments.

Moreover, datasets and code used in some papers that implemented BEL are available online. The datasets and code are published on Kaggle (Lesparre et al., 2022; Thibaut, 2021), allowing users to easily reproduce the results of Chapters 3 and 4.

1.7 Objectives and research questions

> *The grand assertion is that you must see the world through probability and that probability is the only guide you need.*
>
> Dennis Lindley

The objectives of this dissertation are to (i) devise methodological approaches to quantify and reduce the uncertainty associated with subsurface predictions, and (ii) explore the use of machine learning in the context of Bayesian optimal experimental design (BOED) in the subsurface.

The research questions that arise from these objectives are:

1. What are the methodological approaches that can be used to quantify and reduce the uncertainty associated with subsurface models?

2. What is the best method to measure the information content of the data sets?

[1] `https://github.com/robinthibaut/skbel`
[2] `https://pypi.org/project/skbel/`

3. How can an efficient and effective data-driven methodology be developed to quantify the uncertainty in the subsurface predictions and identify the most informative data sets for reducing the uncertainty?

To answer these research questions, we will present three case studies in groundwater modelling applications. The first case study focuses on the wellhead protection area delineation, the second case study focuses on an aquifer thermal energy storage monitoring system, and the third case study focuses on groundwater-surface water interaction. For each case study, we provide an overview of the application and its associated uncertainties, and then use data-driven approaches to quantify the uncertainty, identify the most informative data sets, and reduce the uncertainty. Finally, we discuss the performance of the proposed methods and the implications of the results.

1.8 Overview of the dissertation

In light of the preceding introduction, **Chapter 2** continues with the methodology used to answer the research questions, namely the BEL framework (cf. §2.1), Principal Component Analysis (PCA; cf. §2.2), Canonical Correlation Analysis (CCA; cf. §2.3.1) with its different inference methods, Probabilistic Bayesian Neural Networks (PBNNs; cf. §2.3.2), and the Bayesian optimal experimental design framework (BOED; cf. §5.2.1). The BEL framework software implementation is presented at the end of **Chapter 2** (cf. §2.5).

The next three chapters present three case studies in groundwater modelling applications:

Chapter 3: Wellhead protection area delineation. The Wellhead Protection Area (WHPA) is defined as the zone surrounding a pumping well where human activities are restricted in order to protect water resources (Goldscheider, 2010), generally based on the amount of time harmful contaminants within the area will take to reach the pumping well (according to local regulation). It is determined by the flow velocity in the subsurface surrounding the well, and it can be calculated numerically using particle tracking or transport simulation, or in practice, using tracer tests (Dassargues, 2018; Goldscheider, 2010). Typically, a groundwater model is calibrated against field data before calculating the WHPA using the calibrated model. The establishment of such zones can have a significant socioeconomic impact in densely populated areas where land occupation is a major concern. The WHPA's uncertainty should be quantified, and the most informative data set for reducing the uncertainty should be identified.

Chapter 4: Aquifer thermal energy storage (ATES) monitoring system. Geothermal systems, including borehole thermal energy storage (BTES) and shallow aquifer thermal energy storage systems (ATES) are becoming more popular as the world looks for ways to reduce greenhouse gas emissions. Such systems use thermal energy extracted from the ground or groundwater to heat or cool buildings, which necessitates some electrical energy input for the heat pump, while storing the excess heat or cold underground. The goal is to re-use this thermal energy during the next season in a

cyclic utilization (Bayer et al., 2013; Duijff et al., 2021; Saner et al., 2010; Vanhoudt et al., 2011). The performance of BTES and ATES strongly depends on the subsurface properties. Many variables are involved in geothermal processes, including porosity, hydraulic conductivity, thermal conductivity, and heat capacity. Subsurface temperature fluctuations are strongly influenced by the spatial distribution of these parameters, the boundary conditions, and the aquifer's hydraulic gradient when modeling the underground response under thermal stress (Bridger and Allen, 2010; Ferguson, 2007; Sommer et al., 2014, 2013). Previous research demonstrated that time-lapse Electrical Resistivity Tomography (ERT) could monitor spatial temperature changes in the subsurface with a relatively large spatial coverage by utilizing variations in resistivity caused by temperature changes (Arato et al., 2015; Hermans et al., 2014, 2012, 2015b; Lesparre et al., 2019; Robert et al., 2019). In turn, ERT monitoring experiments can be used to predict the response of the subsurface to thermal exploitation (Hermans et al., 2018). In this case study, we develop a methodology for determining the best monitoring locations for an ATES monitoring system, which can be used to predict the subsurface's response to thermal exploitation.

Chapter 5: Groundwater-surface water interaction. The groundwater-surface water (GW-SW) exchange fluxes are driven by a complex interplay of subsurface processes and their interactions with surface hydrology (Hermans et al., 2022), which have a significant impact on the water and contaminant exchanges (e.g., Dujardin et al., 2014; Ghysels et al., 2021; Hermans et al., 2022; Irvine et al., 2016; Kikuchi and Ferré, 2017; Kurylyk et al., 2019; Moghaddam et al., 2022). Due to the complexity of these systems, the accurate estimation of GW-SW fluxes is important for quantitative hydrological studies and should be based on relevant data and careful experimental design. Therefore, the effective design of monitoring networks that can identify relevant subsurface information are essential for the optimal protection of our water resources. In this study, we present novel deep learning (DL)-driven approaches for sequential and static Bayesian optimal experimental design (BOED) in the subsurface, with the goal of estimating the GW-SW exchange fluxes from a set of temperature measurements. We apply probabilistic Bayesian neural networks (PBNN) to conditional density estimation (CDE) within a BOED framework, and the predictive performance of the PBNN-based CDE model is evaluated by a custom objective function based on the Kullback-Leibler divergence (Definition 5) to determine optimal temperature sensor locations utilizing the information gain provided by the measurements. This evaluation is used to determine the optimal sequential sampling strategy for estimating GW-SW exchange fluxes in the 1D case, and the results are compared to the static optimal sampling strategy for a 3D conceptual riverbed-aquifer model based on a real case study (Ghysels et al., 2021). Our results indicate that probabilistic DL is an effective method for estimating GW-SW fluxes from temperature data and designing efficient monitoring networks. Our proposed framework can be applied to other cases involving surface or subsurface monitoring and experimental design.

Subsequently, **Chapter 6** summarizes the main findings of the dissertation and discusses the implications of the results. Finally, **Chapter 7** presents the conclusion.

2. Methodology

I farm bits and pieces out to the guys who are much more brilliant than I am [...], and I just stick 'em together. But, none of them know what the project really is. So...

Jef Goldblum, *The Fly*, 1986

2.1 Bayesian Evidential Learning

The goal of BEL is to infer the posterior probability distribution $p\left(\mathbf{h}|\mathbf{d}_{\text{obs}}\right)$ of the target $\mathbf{h}$, conditioned by the observed predictor $\mathbf{d}_{\text{obs}}$, by training a regression model given a series of examples of both $\mathbf{d}$ and $\mathbf{h}$.

Bayes' rule is used to make the inference:

$$p\left(\mathbf{h}|\mathbf{d}\right) = \frac{L\left(\mathbf{d}|\mathbf{h}\right)\pi\left(\mathbf{h}\right)}{Z(\mathbf{d})}, \tag{2.1}$$

where $p\left(\mathbf{h}|\mathbf{d}\right)$ is the posterior distribution, $L\left(\mathbf{d}|\mathbf{h}\right)$ is the likelihood function, $\pi\left(\mathbf{h}\right)$ is the prior distribution, and $Z(\mathbf{d})$ is the marginal likelihood (or evidence). Both the target $\mathbf{h}$ and the predictor $\mathbf{d}$ are real, multidimensional random variables.

The posterior distribution is the probability of finding $\mathbf{h}$ given the predictor $\mathbf{d}$. The likelihood is the probability of having generated the predictor $\mathbf{d}$ given the target $\mathbf{h}$, and is the main contribution to the posterior. The prior is the probability of the parameters before the observation of the current data. The marginal likelihood is obtained after marginalizing the likelihood over the parameter space:

$$Z(\mathbf{d}) = \int L(\mathbf{d}|\mathbf{h})\pi(\mathbf{h})\ \mathrm{d}\mathbf{h} \tag{2.2}$$

In the BEL framework, the first step is to choose the prior distribution $\pi(\mathbf{h})$. The next step is to generate a training set containing N pairs of predictors $\mathbf{d}_i$ and targets $\mathbf{h}_i$, where $i = 1, \ldots, N$. For a given domain $\mathcal{M}$ parameterized by a vector ω, the target $\mathbf{h}$ is obtained by sampling the forward model $g(\mathcal{M}, \omega)$:

$$\mathbf{h} = \sim fg(\mathcal{M}, \omega), \tag{2.3}$$

where g is the forward model and $\sim fg$ is a Monte Carlo simulation that samples the forward model g N times. Similarly, the predictor $\mathbf{d}$ is obtained by sampling the forward model $s(\omega')$:

$$\mathbf{d} = \sim fs(\mathcal{M}, \omega'), \tag{2.4}$$

This pair of sampled predictor-target forms the training set. s and g are two different forward models that are linked by a common support (computational domain; $\mathcal{M}$), but differ in what they model, and the parameter vectors ω and ω' do not necessarily share the same parameter space.

For example, the computational domain $\mathcal{M}$ could be a subsurface region where the forward model g simulates groundwater flow and the forward model s simulates solute transport. The parameter vector ω could include geological, hydrogeological, and hydraulic parameters, while the parameter vector ω' could include solute transport parameters, hydraulic parameters, and initial conditions. However, for $\mathbf{d}$ to be informative on $\mathbf{h}$, there should be some overlap between the two. In this case, the predictor variable could be the drawdown at a given pumping well, and the target variable could be the position of a contaminant plume front.

2.2 Dimensionality Reduction

Principal Component Analysis

PCA is one of the most established and popular multivariate techniques (Meloun and Militký, 2012; Pearson, 1901), and has been widely use in previous BEL applications. Due to its establishment and popularity, many flavors of PCA have been proposed over the last century, e.g.,

- *Classical PCA* (Pearson, 1901), which is the standard PCA that is used in most applications.
- *Robust PCA* (Candès et al., 2011), which is an extension of PCA that is robust to grossly corrupted observations.
- *Sparse PCA* (Zou et al., 2006), the disadvantage of PCA is that each principal component is a linear combination of all the original variables. Sparse PCA circumvents this limitation by locating linear combinations with a small number of input variables.
- *Kernel PCA* (Schölkopf et al., 1997), which is an extension of PCA that is more robust to nonlinear relationships within the input, which is done by using a kernel function to map the input to a higher-dimensional space.

In this work, we use the classical PCA. The goal of PCA is to represent objects in the new principal component PC-coordinate space. The PCA achieves two goals: it transforms the data into a more relevant coordinate system that lies directly in the center of the data swarm of objects, and it reduces the dimensionality of the data by using only the few principal components that reflect the structure in the data (Meloun and Militký, 2012).

Mathematically, PCA generates multiple linear combinations of observed variables, and the resulting principal components are merely aggregations of correlated variables (Meloun and Militký, 2012).

Let's consider a dataset of N observations $\mathbf{d} = d_1, d_2, \ldots, d_p$, where d_i is a vector of N observations of the i-th variable, also called a feature or a dimension. The technique consists in transforming the original variables $\mathbf{d}$ into a new set of variables $\mathbf{z} = z_1, z_2, \ldots, z_p$, where z_i is a vector of N observations of the i-th principal component. The dimensionality of the data can be reduced by keeping only the first $\delta < p$ principal components, which is called truncated PCA. The variance of a principal component is a measure of the information it conveys (Meloun and Militký, 2012). The first principal component is the direction that maximizes the variance of the data, and the second principal component is the direction that maximizes the variance of the data while being orthogonal to the first principal component, and so on.

Assume we're looking for an orthogonal set of p linear basis vectors $\mathbf{w}_i \in \mathbb{R}^p$, and the corresponding principal components $\mathbf{z}_i \in \mathbb{R}^p$, so that we minimize the average reconstruction error (Murphy, 2012)

$$J(\mathbf{w}, \mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{d}_i - \hat{\mathbf{d}}_i\|^2 \tag{2.5}$$

where $\hat{\mathbf{d}}_i = \mathbf{W}\mathbf{z}_i$ under the condition that $\mathbf{W}$ is orthonormal. This objective can also be written as follows (Murphy, 2012):

$$J(\mathbf{W}, \mathbf{Z}) = ||\mathbf{d} - \mathbf{w}\mathbf{z}^T||_F^2 \tag{2.6}$$

where $\mathbf{z}$ is an $N \times p$ matrix with the principal components as columns, and $||A||_F$ is the Frobenius norm of the matrix A (Murphy, 2012):

$$||A||_F = \sqrt{\mathrm{tr}(A^T A)} \tag{2.7}$$

where $\mathrm{tr}(A)$ is the trace of the matrix A. Setting $\hat{\mathbf{w}} = \mathbf{V}_p$ yields the optimal solution. The matrix $\mathbf{V}_p$ is the $p \times p$ matrix whose columns are the eigenvectors of the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ of the data $\mathbf{d}$, ordered by decreasing eigenvalues (Murphy, 2012):

$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \mathbf{d}^T \mathbf{d}$$

Finally, the principal components are obtained by projecting the data onto the eigenvectors of the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ (Murphy, 2012):

$$\hat{\mathbf{z}} = \mathbf{W}^T \mathbf{d}_i$$

The principal components constitute a set of new variables that summarize the information contained in the original multivariate dataset. The technique is simple to interpret, even though the resulting directions are sometimes difficult to interpret physically.

Working with PCA has a number of advantages. PCA reduces the number of dimensions in the data, lowering the computational complexity of subsequent analyses, and it eliminates potential multicollinearity in variables, which can be problematic for some statistical techniques such as Canonical Correlation Analysis (Meloun and Militký, 2012). PCA is widely available in statistical software and is relatively easy to implement. It is quick to compute and has no limit on the number of variables.

PCA has some drawbacks. It must be made clear that the technique is not a method of supervised learning, so it cannot be used for prediction. The formulation of the objective function in Equation 2.5 clearly shows that PCA minimizes the reconstruction error in a least-squares sense, implying that PCA is unsuitable for categorical data. However, as demonstrated in Chapter 3, categorical or binary data can be transformed to make it suitable for PCA.

2.3 Learning and inference

2.3.1 Canonical Correlation Analysis

Canonical correlation analysis can identify and quantify the relationships between two sets of variables (Härdle and Simar, 2019; Hotelling, 1936). CCA, like PCA, reduces the original variables' dimensionality. However, whereas PCA seeks to minimize reconstruction error, CCA seeks to maximize the correlations of two variables, making it better suited for regression tasks (Meloun and Militký, 2012).

CCA is used in situations where regression techniques are appropriate and there are multiple **d** and **h** variables (Meloun and Militký, 2012). The terminology involved in CCA is a major source of confusion. There are original variables, canonical variates, and canonical variate pairs. Variables are the original variables that were measured in survey, or in most of our applications: the variables that were generated by the PCA (e.g., Chapter 3 to 5). Canonical variates are linear composites of original variables, and a pair of canonical variates is formed by these two composites. However, more than one reliable pair of canonical variates may exist. Canonical correlation combines sets of variables on each side to produce a predicted value for each side that has the highest correlation with the predicted value on the other side, where *side* refers to the **d** or **h** variables (see Figure 2.1). The combination of original variables on each side can be viewed as a dimension connecting the original variables on one side to the original variables on the other (Meloun and Militký, 2012).

CCA has a number of drawbacks that contribute to the reason for its paucity in the literature. The most important limitation is interpretability: procedures that maximize correlation do not always maximize interpretation of pairs of canonical variates. As a result, canonical solutions are frequently mathematically elegant but uninterpretable (Meloun and Militký, 2012).

CCA investigates the linear relationships between a set of **d** variables $d_1, d_2, \ldots, d_p$, i.e., $U = a_1d_1 + a_2d_2 + \cdots + a_pd_p$, and a set of more than one **h** variables $h_1, h_2, \ldots, h_q$, i.e., $V = b_1h_1 + b_2h_2 + \cdots + b_qh_q$ (Meloun and Militký, 2012). The technique entails finding several linear combinations of the **d** variables and the same number of linear combinations of the **h** variables that best express the correlations between these two sets. These linear composites V and U are referred to as canonical variates, and the correlations ρ between corresponding pairs of canonical variates are referred to as canonical correlations.

Suppose we wish to investigate the relationship between a set of variables $d_1, d_2, \ldots, d_p$ and a set of variable $h_1, h_2, \ldots, h_q$. We assume that the mean of each variable has been subtracted from the original data in any given sample, so that the sample means of all **d** and **h** variables are zero. The canonical correlation analysis serves two primary purposes: (1) the identification of dimensions among the dependent and independent variables that (2) maximize the relationship between the dimensions (Meloun and Militký, 2012). In CCA, the coefficients **a** and **b** are determined so as to maximize the correlation between U_i and V_i.

The resulting linear combination U_1 is known as the first canonical variate of the **d**'s and V_1 is known as the first canonical variate of the **h**'s. The resulting correlation between U_1 and V_1 is known as the first canonical correlation. Therefore, the first canonical correlation ρ_1^2 is the highest possible correlation between a linear combination of the **d**'s and a linear combination of the **h**'s.

The number of variables in the smallest data set equals the maximum number of canonical variates (c) that can be extracted from the sets of variables **d** or **h**, i.e., $c = min(p, q)$, as illustrated in Figure 2.1.

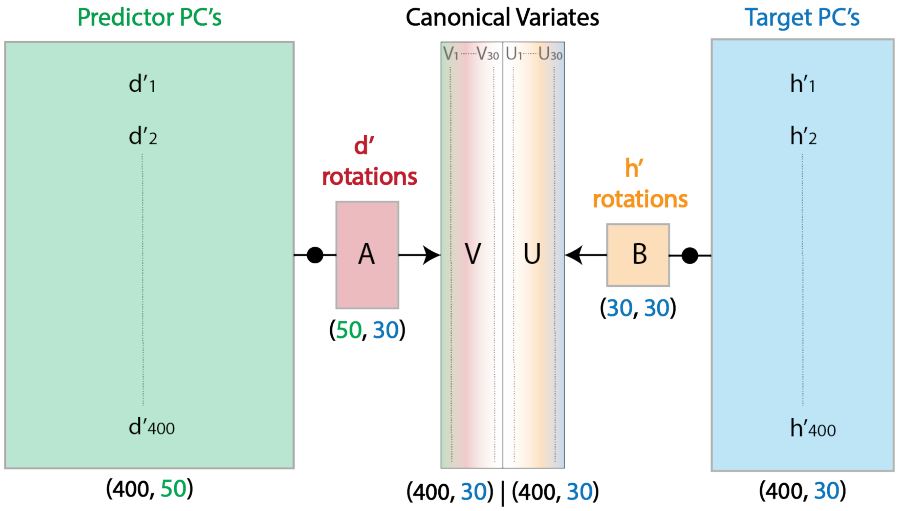

Figure 2.1: Illustration of canonical correlation analysis. The left panel shows the predictor variables (d), and the right panel shows the response variables (h) in PC space, with 400 examples. The predictor variable has 50 dimensions, and the target variable has 30 dimensions. The canonical correlations are the correlations between the canonical variates, which are obtained by projection of the original variables via the learned *rotation* matrices.

To derive the canonical coefficients $\mathbf{a}$ and $\mathbf{b}$, we first compute the covariance matrix of the $\mathbf{d}$ and $\mathbf{h}$ variables. The covariance matrix $\mathbf{\Sigma}$ of all the variables is divided into four parts (Meloun and Militký, 2012):

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{dd} & \mathbf{\Sigma}_{dh} \\ \mathbf{\Sigma}_{hd} & \mathbf{\Sigma}_{hh} \end{bmatrix} \tag{2.8}$$

- Σ_{dd} is the covariance matrix of the $\mathbf{d}$ variables.
- Σ_{hh} is the covariance matrix of the $\mathbf{h}$ variables.
- Σ_{dh} is the covariance matrix of the $\mathbf{d}$ and $\mathbf{h}$ variables.
- Σ_{hd} is the covariance matrix of the $\mathbf{h}$ and $\mathbf{d}$ variables.

We can then define (Härdle and Simar, 2019):

$$\mathcal{K} = \Sigma_{dd}^{-1/2} \Sigma_{dh} \Sigma_{hh}^{-1/2}, \tag{2.9}$$

and $\mathcal{K}(p \times q)$ is subjected to singular value decomposition (SVD) (Härdle and Simar, 2019):

$$\mathcal{K} = \mathbf{\Gamma \Lambda \Delta}^T,$$

with (Härdle and Simar, 2019):

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_1 & \gamma_2 & \dots & \gamma_c \end{bmatrix}$$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_c \end{bmatrix}$$

$$\mathbf{\Delta} = \begin{bmatrix} \delta_1 & \delta_2 & \dots & \delta_c \end{bmatrix}$$

$$c = rank(\mathcal{K}) = rank(\Sigma_{dh}) = rank(\Sigma_{hd}) = min(p, q)$$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_c$ are the singular values of $\mathcal{N}_1 = \mathcal{K}\mathcal{K}^T$ and $\mathcal{N}_2 = \mathcal{K}^T\mathcal{K}$. γ_i and δ_i are the standardized eigenvectors of $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively. The canonical coefficients are defined as (Härdle and Simar, 2019):

$$a_i = \Sigma_{dd}^{-1/2}\gamma_i$$
$$b_i = \Sigma_{hh}^{-1/2}\delta_i$$

Significant results indicate that there is a significant overlap in the variance of variables between the two sets (Meloun and Militký, 2012). In addition, when SVD is applied, the canonical variate pairs are orthogonal to one another and can therefore be inferred independently.

Therefore, CCA is a useful tool for investigating the relationship between two sets of variables, and it is particularly useful when the number of variables in each set is large. The resulting canonical variate pairs can be used to construct a scatter plot, and the resulting canonical correlations can be used to assess the strength of the relationship between the two sets of variables (e.g., Chapters 3 and 4). Since the canonical variates are linear combinations of the original variables, the canonical variates can be used to construct a regression model for the **h**'s as a function of the **d**'s. This can be done by estimating the posterior distribution of the **h**'s canonical variates given the **d**'s canonical variates. Several methods have been proposed for estimating the posterior distribution of the **h**'s canonical variates given the **d**'s canonical variates, such as multiple Gaussian regression, Kernel Density Estimation, and Transport Maps (cf. §2.3.1).

In BEL, we're interested in uncovering the mutual information between the **d**'s and the **h**'s. Assuming Gaussian variables **d** and **h**, the relationship between mutual information and canonical correlation is given by (Borga, 1998):

$$\mathcal{I}(d; h) = \frac{1}{2} \log \left(\frac{1}{\prod_i (1 - \rho_i^2)} \right) \tag{2.10}$$

where ρ_i^2 is the ith canonical correlation. This link between mutual information and canonical correlation makes CCA a natural choice in BEL. Another advantage is that CCA doesn't require any hyperparameters to be tuned.

However, CCA also has some limitations. Due to the requirement for inverse matrices during analysis, the first issue to examine is multicollinearity and singularity: Multicollinearity exists when one variable is nearly a weighted average of the others, whereas singularity exists when this relationship is exact (Meloun and Militký, 2012). In addition, although normal distribution is not required when descriptive canonical correlation is used, it enhances the analysis if the variables have a normal distribution (Meloun and Militký, 2012). In BEL, PCA is typically performed on the original variables prior to CCA. PCA will handle multicollinearity and singularity, and the principal components are easier to work with; for instance, normality can be ensured by applying a transformation to the principal components, such as the Box-Cox transformation (Box and Cox, 1964).

These constraints result from the assumption that governs CCA. The analysis is carried out on covariance matrices that only reflect linear relationships, although the use of linear combinations can identify weakly non-linear relationships. If the relationship between two variables is nonlinear, these statistics will not capture it (Meloun and Militký, 2012). Furthermore, CCA maximizes the linear relationship between a variate from one set of variables and a variate from the other set, and misses potential nonlinear components of relationships between canonical variate pairs (Meloun and Militký, 2012). CCA works best when relationships between variables are homoscedastic, i.e., when the variance of one variable is roughly the same at all levels of the other variable (Meloun and Militký, 2012).

Recent research has suggested that the shortcomings of linear CCA can be addressed by modifying the learning procedure and introducing some iterative updating of the prior (Hermans et al., 2019; Michel et al., 2022a, 2020a,b; Park and Caers, 2020). However, such adaptation increases the computation cost and reduces the adaptability of BEL as the iterative process is inherently dependent on the dataset, making it less efficient for experimental design. To keep the experimental design efficient, the learning phase could benefit from more advanced approaches, such as deep learning techniques. Neural networks, for example, can serve as an alternative to CCA and can capture nonlinear relationships between variables at the price of tuning hyperparameters, possibly a larger training set, and a longer training time (e.g., Chapter 5).

CCA demonstration. In this section, we demonstrate the application of CCA to two synthetic multivariate datasets with a Python snippet (Snippet A.1).

We consider two cases: (1) a linear case, where the variables are linearly correlated, and (2) a nonlinear case, where the variables are correlated through a nonlinear function (sinusoidal). For each case, we generate a dataset of 1000 samples of 4-dimensional variables (`X` and `Y`). We then add nonlinearity to the data by independently transforming each pair of variables (`X[:, 0]` and `Y[:, 0]`, `X[:, 1]` and `Y[:, 1]`, and so on). Finally, we plot the original 4-dimensional data (Figures 2.2A-B).

We then fit a CCA model (`cca = CCA(n_components=4)`), specifying the number

of canonical variates as 4 (the same number as the original variables, since we are not doing dimensionality reduction in this example). We then transform the data using the CCA model (`X_c, Y_c = cca.transform(X, Y)`) and compute the correlations between the pairs of transformed variables (`corr = np.corrcoef(X_c.T, Y_c.T)`). We plot the transformed data (Figures 2.2C-D).

For the linear case, the correlation between the canonical variate pairs is close to one (Figure 2.2C). For the nonlinear case, the correlation is lower but still significant (Figure 2.2D). This demonstrates the efficacy of CCA in capturing linear and nonlinear relationships between multivariate datasets.

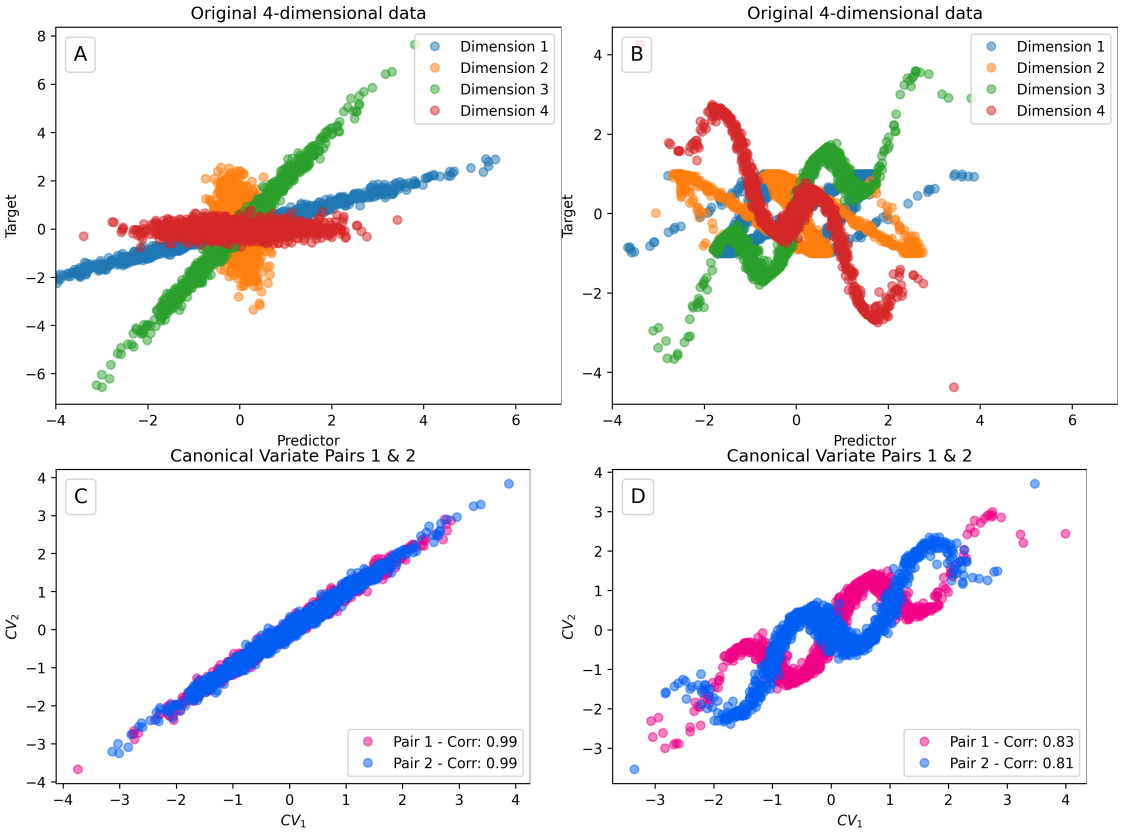

Figure 2.2: Canonical correlation analysis (CCA) applied to two synthetic multivariate datasets. The first row shows the original 4-dimensional data for the linear (**A**) and nonlinear (**B**) cases. The second row shows the transformed canonical variate pairs 1 and 2 for the linear (**C**) and nonlinear (**D**) cases. The Pearson correlation between the canonical variate pairs is shown in the legend.

Conditional sampling of canonical variates

Multivariate Gaussian Inference. Multivariate Gaussian inference (MGI) is a method for approximating the posterior distribution of a Gaussian process, and requires the verification of normality and linear assumptions. It is thus particularly well adapted to

estimating the posterior distribution of the **h**'s canonical variates given any **d**'s canonical variates in the BEL framework if CCA is used for learning. If the linear correlation between canonical variate pairs is sufficiently strong, then analytic MG inference can be performed to directly infer $p(\mathbf{h}|\mathbf{d}_{\text{obs}})$ in canonical space. If the two sets of variables are jointly Gaussian, the posterior distribution of **h** conditioned on **d** is Gaussian (Murphy, 2012), which is a useful property for sampling from the posterior distribution.

Let $\mathbf{G} \in \mathbb{R}^{q \times q}$ be the Ordinary Least Square (OLS) solution mapping **h** to **d**, such that $\mathbf{Gh} = \mathbf{d} + \epsilon$, and suppose $p(\mathbf{h}|\mathbf{d}_{\text{obs}})$ is jointly Gaussian with covariance matrix (Murphy, 2012):

$$\Sigma = \begin{pmatrix} \Sigma_{hh} & \Sigma_{hh}G^T \\ G\Sigma_{hh} & G\Sigma_{hh}G^T + \Sigma_d^* \end{pmatrix} = \begin{pmatrix} \Sigma_{hh} & \Sigma_{hd} \\ \Sigma_{dh} & \Sigma_{dd} \end{pmatrix} \tag{2.11}$$

The precision matrix $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ is

$$\Lambda = \begin{pmatrix} G^T\Sigma_{hh}G + \Sigma_d^{*-1} & -G^T\Sigma_{hh}^{-1} \\ -\Sigma_{hh}^{-1}G & -\Sigma_{hh}^{-1} \end{pmatrix} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \tag{2.12}$$

The posterior conditional $p(\mathbf{h}|\mathbf{d_{obs}}) = \mathcal{N}\left(\mathbf{h}|\mu_{\mathbf{h}|\mathbf{d_{obs}}}, \mathbf{\Sigma}_{\mathbf{h}|\mathbf{d_{obs}}}\right)$ has parameters

$$\Sigma_h|_{d_{obs}} = \Lambda_{11}^{-1} \tag{2.13}$$

$$\mu_h|_{d_{obs}} = \Sigma_h|_{d_{obs}}(\Lambda_{11}\mu_h - \Lambda_{12}(d_{obs} - \mu_d)) \tag{2.14}$$

Note that, interestingly, the posterior covariance (Equation 2.13) does not depend on the observed value $\mathbf{d}_{\text{obs}}$, and that computing the posterior mean (Equation 2.14) is simply a linear operation, given the precomputed posterior covariance.

The $\mathbf{\Sigma}_{\mathbf{d}}^*$ and $\mu_{\mathbf{d}}^*$ variables are a covariance term and a term representing deviations from the mean, respectively, resulting from the imperfect OLS fitting and noise present in the predictor set. It should be noted that histogram transformations can be used to ensure that the distributions are normal (Satija and Caers, 2015), but that such transformation might impact the linearity of the canonical variate pairs (Hermans et al., 2019). For information on how to approximate $\mathbf{\Sigma}_{\mathbf{d}}^*$ and $\mu_{\mathbf{d}}^*$, see Scheidt et al. (2018).

Once the first two moments of the MG $p(\mathbf{h}|\mathbf{d}_{\text{obs}})$ are known, sampling from it is straightforward, as it is a Gaussian distribution.

In summary, Multivariate Gaussian Inference is an analytical solution to the problem of inferring the posterior distribution in the BEL framework, used in combination with CCA to estimate the posterior distribution of the **h**'s canonical variates given any **d**'s canonical variates. Its simplicity is a double-edged sword in that it necessitates strict assumptions about the relationship between the canonical variates, which are frequently not met. It also has a tendency to overestimate the variance of the posterior distribution, which can put the model in a state of under-confidence in some cases (a lower-confidence

can lead to going on the safe side), but detrimental in others.

The MGI algorithms are implemented in the `SKBEL` Python package (Thibaut and Ramgraber, 2021), and will be illustrated in the *Software implementation* section §2.5.

Kernel density estimation. Kernel density estimation (KDE) is a simple non-parametric density estimation method. This is in contrast to parametric techniques that make strong assumptions about the underlying distribution, such as the normal distribution. KDE, as a non-parametric technique, makes no assumptions about the random variable's underlying distribution. Let's consider a 1D case where the random variable is denoted by $d = \{d_1, d_2, \ldots, d_N\}$. The empirical distribution function (EDF) of d is an equally weighted mixture of N Dirac delta functions, which is a continuous function that is zero everywhere except at the points d_i:

$$\hat{F}_N(d) = \frac{1}{N} \sum_{i=1}^{N} \delta(d - d_i) \tag{2.15}$$

The KDE of d is then computed as

$$\hat{f}_N(d) = \frac{1}{Nb} \sum_{i=1}^{N} \mathcal{K}\left(\frac{d - d_i}{h}\right) \tag{2.16}$$

where $\mathcal{K}(\cdot)$ is the kernel function and b is the bandwidth which is a parameter that controls the smoothness of the estimated density (Papamakarios, 2019). The most commonly used kernel is the Gaussian kernel

$$\mathcal{K}(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{2.17}$$

where u is the normalized distance between the point d and the point d_i. KDE can be extended to higher-dimensional data by using a multivariate Gaussian kernel.

KDE is a simple yet powerful tool for non-parametric density estimation. It is relatively easy to implement and can be used to estimate complex distributions without making strong assumptions about the underlying data. One advantage of a non-parametric model over a parametric model is that no model fitting is required, except for tuning the bandwidth, usually done by cross-validation (Bishop, 2007).

However, it has some drawbacks. First, the choice of the bandwidth b is crucial for the performance of KDE. If the bandwidth is too small, the estimated density will be too noisy (overfit); if it is too large, the estimated density will be too smooth (underfit).

In practice, there are other methods to estimate the bandwidth than cross-validation. Popular rules of thumb include the Silverman's rule of thumb (Silverman, 1986) and the Scott's rule of thumb (Scott, 1992). In addition, Michel et al. (2020a) proposed estimating the bandwidth by using the number of training samples in the vicinity of the

observed data in order to reduce computational expense.

The second drawback is that KDE is computationally expensive. It requires a summation over all of the data points, and storing and evaluating the model necessitates a significant amount of memory and time (Bishop, 2007). In high-dimensional data, the computational cost can be prohibitive.

Nevertheless, if the linear correlation and normality assumptions in the canonical variate pairs are violated too severely for the application of an analytic MGI (cf. §2.3.1), KDE can be used to approximate $p(\mathbf{h}|\mathbf{d}_{\text{obs}})$ instead of MGI (e.g., Hermans et al. 2019; Michel et al. 2022a, 2020a,b).

The KDE algorithms are implemented in the `SKBEL` Python package. It includes automatic bandwidth selection using Grid Search via `scikit-learn` (Pedregosa et al., 2011), and the conditional posterior distribution $p(\mathbf{h}|\mathbf{d})$ can be estimated for any $\mathbf{d}$'s canonical variates. It is highly modular and can also be used to produce the complex figures in this dissertation. The method will be illustrated in the *Software implementation* section §2.5.

Transport methods. In essence, transport methods seek a monotone, invertible transport map $\mathbf{S}$ that transforms samples from a target distribution π_j into samples from a simpler, user-specified reference distribution η, typically a standard multivariate normal distribution $N(0, I)$, where I is the identity matrix. This map allows us to sample conditionals of the target distribution, and thus implements the conditioning operation we are principally interested in. Transport methods are a nuanced topic, and the interested reader is referred to Villani (2009), El Moselhy and Marzouk (2012) and Spantini et al. (2018) for a more detailed discussion of their theoretical properties. In this study, we focus only on the parts necessary for the conditioning operation. Following Spantini et al. (2022) and Baptista et al. (2022), the map $\mathbf{S}$ is triangular, meaning it has as many parameterized map components as there are dimensions in π_j (two in this case, so $\mathbf{S} : \mathbb{R}^2 \to \mathbb{R}^2$). Each map component depends on one more dimension of the target's probability density function (pdf) than its predecessor. In our setting, the map is structured as:

$$\mathbf{S}\left(d^c_{:,j}, h^c_{:,j}\right) = \begin{bmatrix} S_1\left(d^c_{:,j}\right) \\ S_2\left(d^c_{:,j}, h^c_{:,j}\right) \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \mathbf{z} \tag{2.18}$$

where $\mathbf{z}$ are samples distributed according to the reference pdf η. These map component functions are made up of user-specified combinations of polynomials or radial basis functions, and must be monotone in their last argument. This means that $\partial_{d^c_{:,j}} S_1\left(d^c_{:,j}\right) > 0$ and $\partial_{h^c_{:,j}} S_2\left(d^c_{:,j}, h^c_{:,j}\right) > 0$. This will ensure that the entire map $\mathbf{S}$ remains monotone and thus invertible. This monotonicity requirement can be ensured with diligent parameterization of the map components (e.g., Equation 2 in Baptista et al. (2022)).

If we are only interested in conditioning, we only need to define, optimize, and evaluate the second map component $S_2 : \mathbb{R}^2 \to \mathbb{R}$ (Spantini et al., 2022). We optimize this map component S_2 by minimizing the Kullback-Leibler divergence (Definition 5) between the target pdf π_j, presumed to be known only through samples, and the map's approximation to the target pdf, which is obtained by sending the standard Gaussian reference η through the inverse map $\mathbf{S}^{-1}$. From this, we can derive the following objective function $\mathcal{J}$ (Spantini et al., 2018):

$$\mathcal{J}(S_2) = \sum_{i=1}^{n_{training}} \left(\frac{1}{2} \left(S_2(d_{i,j}^c, h_{i,j}^c) \right)^2 - \log \frac{\partial S_2(d_{i,1}^c, h_{i,j}^c)}{\partial h_{i,j}^c} \right). \tag{2.19}$$

The forward evaluation of S_2 is a simple evaluation of its constituent basis functions. We can also invert S_2 efficiently through an appropriate one-dimensional root finding algorithm. Note that while the forward map components S_1 and S_2 can be evaluated independently, their inverses must be evaluated in sequence:

$$\mathbf{S}^{-1}(\mathbf{z}) = \begin{bmatrix} S_1^{-1}(z_1) \\ S_2^{-1}\left(d_{:,j}^c, z_2\right) \end{bmatrix} = \begin{bmatrix} d_{:,j}^c \\ h_{:,j}^c \end{bmatrix} \tag{2.20}$$

where the inverse of the second map component S_2^{-1} depends on the outcome $d_{:,j}^c$ of the first inversion S_1^{-1}. This dependence makes triangular transport highly useful for Bayesian inference. In fact, the inverse map $\mathbf{S}^{-1}$ factorizes the target pdf π_j according to the ordering of the variables in Equation 2.18 as $\pi_j\left(d_{:,j}^c, h_{:,j}^c\right) = \pi_j\left(d_{:,j}^c\right) \pi_j\left(h_{:,j}^c | d_{:,j}^c\right)$, where the second term on the right-hand side ($\pi_j\left(h_{:,j}^c | d_{:,j}^c\right)$) corresponds to the posterior and is sampled by the map component inverse S_2^{-1} (e.g., Section 7.1 of Spantini et al. (2018)).

In other words, this means that we can sample conditionals of π_j by simply replacing the argument $d_{:,j}^c$ during the map component inversion S_2^{-1} with any values on which we want to condition. For instance, supplying duplicates of the observation $d_{obs,j}^c \mathbf{1}^\top$, where $\mathbf{1}$ is a column vector of 1s, we can sample the desired conditional $\pi_j(h_{i,j}^c | d_{obs,j}^c)$:

$$\mathbf{t}_{:,j}^{c,\text{cond.}} = S_2^{-1}(z_2; d_{obs,j}^c \mathbf{1}^\top) \sim \pi_j(h_{i,j}^c | d_{obs,j}^c). \tag{2.21}$$

the required reference samples z_2 can be either drawn from a standard Gaussian distribution, or (better) obtained from the forward map $z_2 = S_2\left(d_{:,j}^c, h_{:,j}^c\right)$ (see Spantini et al. (2022)). With the conditioned samples $\mathbf{t}_{:,j}^{c,\text{cond.}}$ for each pair of covariates $\left(d_{:,j}^c, h_{:,j}^c\right)$, we can then back-transform the posterior samples into the original target space by ascending from CCA, PCA, and undoing any transformations.

The transport map algorithm suite is also implemented in the `SKBEL` package. Similar to MGI and KDE, it can be selected as the conditional sampling method when employing CCA to find the relationship between the predictor and target variables. It is also modular and can be used as a stand-alone conditional sampling method. For more complex bivariate distributions, the application of transport map methods can require

hyperparameter optimization, such as, e.g., the structure of the monotone and non-monotone part of the transport map component functions, the scaling factors for the map components, and the type of regularization. The method will be illustrated in the *Software implementation* section §2.5.

2.3.2 Probabilistic Bayesian Neural Networks

The proposed probabilistic Bayesian neural network (PBNN) method is a hybrid of Artificial Neural Networks (ANN), Bayesian Neural Networks (BNN), and Probabilistic Neural Network (PNN). Each of these methods is briefly described below, and illustrated at the end of this section with a Python snippet (Snippet A.2) and figures (Figures 2.3 to 2.6).

Artificial neural networks (ANNs). ANNs are a category of machine learning algorithms inspired by the structure and operation of the human brain. They consist of interconnected units called "neurons" that process and transmit data. Each neuron receives input from other neurons or external sources and generates output that is transmitted to other neurons or the network's final output layer.

The impact of the input received by each neuron is determined by the weights, or strength, of the connections between neurons. During training, these weights are modified, enabling the network to learn from input data and make predictions or decisions based on this learning.

There are typically three types of layers in ANNs: an input layer, one or more hidden layers, and an output layer. The input layer receives the raw input data, while the output layer generates the network's final output. As the name implies, the hidden layers are not directly connected to the input or output layers, and their function is to extract and abstract features from the input data to support the network's prediction or decision making.

The learning process of an ANN involves presenting the network with a set of input-output pairs, known as a training set, and adjusting the weights to minimize the difference between the predicted output and the true output. This is typically accomplished by applying an optimization algorithm, such as gradient descent, to determine the set of weights that produces the lowest error. The learning process is repeated multiple times, referred to as epochs, and is also divided into batches, or small subsets of the training data. The learning rate (a hyperparameter) regulate the rate of learning and is usually set at the start of the training process.

Mathematically, an ANN can be represented as a function $f(\mathbf{d}; \mathbf{w})$, where $\mathbf{d}$ is the input predictor and $\mathbf{w}$ is the set of weights. Each layer of the network can be represented as a function $g(\mathbf{d}; \mathbf{w})$, where $\mathbf{d}$ is the input to the layer and $\mathbf{w}$ are the weights of the connections between neurons in the layer. The final output of the network is obtained by composing these functions, as follows:

$$\mathbf{h} = f(\mathbf{d}; \mathbf{w}) = g_L(g_{L-1}(\dots g_2(g_1(\mathbf{d}; \mathbf{w_1}); \mathbf{w_2}); \dots); \mathbf{w_L}) \tag{2.22}$$

where $g_i(\cdot)$ is the function of the i-th layer, $\mathbf{w_i}$ is the set of weights of the connections in the i-th layer, and $\mathbf{y}$ is the final output (target) of the network.

However, ANNs are not capable of capturing uncertainty in their predictions, which is a crucial capability when dealing with real-world applications.

Bayesian neural networks (BNNs). BNNs are a type of deep-learning approaches that leverage Bayesian methods to learn probability distributions over the NN parameters and to quantify uncertainty in predictions (Kendall and Gal, 2017), making them a promising solution for applying DL in situations where it is not permitted for a system to make inaccurate predictions without warning (Jospin et al., 2022). BNNs allow aleatory and epistemic forms of uncertainty to be captured in the model outputs, which makes BNNs "data-efficient" as they can learn from a small dataset without overfitting (Depeweg et al., 2018; Jospin et al., 2022). This paradigm offers a rigorous framework for analyzing and training uncertainty-aware neural networks (Jospin et al., 2022), and provides a principled way to integrate prior beliefs into DL models (Khan and Coulibaly, 2006).

Let $\theta = (\mathbf{w}, \mathbf{b})$ be the parameters of a BNN f_θ, where $\mathbf{w}$ are the weights of the network connections and $\mathbf{b}$ the biases. Let $\mathbf{d}$ be the input (predictor) variable, and let $\mathbf{h} = f_\theta(\mathbf{d})$ be the output (target) variable. We assume that $\mathbf{d}$ and $\mathbf{h}$ are independent and identically distributed (i.i.d.). The goal of a BNN is to infer the posterior distribution of the parameters θ given a data set $\mathcal{D} = \{\mathbf{d}_n, \mathbf{h}_n\}_{n=1}^N$. The posterior distribution can be computed using Bayes' theorem (Gal, 2016; Sharma et al., 2022):

$$p(\theta|\mathbf{d}, \mathbf{h}) = \frac{p(\mathbf{h}|\mathbf{d}, \theta)\, p(\theta)}{p(\mathbf{h}|\mathbf{d})} \tag{2.23}$$

Here, $p(\mathbf{h}|\mathbf{d}, \theta)$ is the likelihood; $p(\theta)$ is a prior distribution, which is typically chosen to be a zero-mean isotropic Gaussian (Jospin et al., 2022); and $p(\mathbf{h}|\mathbf{d})$ is the evidence, which can be computed by marginalizing the likelihood over the parameters:

$$p(\mathbf{h}|\mathbf{d}) = \int p(\mathbf{h}|\mathbf{d}, \theta)\, p(\theta) d\theta \tag{2.24}$$

The true posterior $p(\theta|\mathbf{d}, \mathbf{h})$ is usually intractable. Rather than sampling from the exact posterior, the variational distribution $q_\gamma(\theta)$ is used, which belong to a tractable family of distributions (e.g., Gaussian distribution) and is parametrized by a set of parameters γ. The parameters γ are then learned so that the variational distribution $q_\gamma(\theta)$ is as close to the true posterior $p(\theta|\mathbf{d}, \mathbf{h})$ as possible (Gal, 2016; Jospin et al., 2022). The Kullback-Leibler divergence (KL-divergence; Kullback and Leibler 1951), based on Shannon's information theory (Definition 5), is a commonly used measure of closeness between probability distributions:

$$\mathrm{KL}(q_\gamma(\theta)||p(\theta|\mathbf{d},\mathbf{h})) = \int q_\gamma(\theta) \log \frac{q_\gamma(\theta)}{p(\theta|\mathbf{d},\mathbf{h})} d\theta. \tag{2.25}$$

Minimizing the KL-divergence is equivalent to maximizing the evidence lower bound (ELBO) w.r.t. the variational parameters γ:

$$\mathcal{L}(\gamma) = \mathbb{E}_{q_\gamma(\theta)} \left[\log p(\mathbf{h}|\mathbf{d},\theta)\right] - \mathrm{KL}(q_\gamma(\theta)||p(\theta)) \leq \log p(\mathbf{h}|\mathbf{d}). \tag{2.26}$$

This procedure is referred to as variational inference (VI; Blundell et al. 2015; Gal 2016; Jospin et al. 2022). For a BNN, the optimization of the ELBO with respect to the parameters of a variational distribution requires the adaptation of VI. Stochastic variational inference (SVI; Hoffman et al. 2013), which is the stochastic gradient descent (SGD) method applied to VI, is the most widely used method for optimizing the ELBO (Jospin et al., 2022).

To efficiently approximate the gradients of the ELBO with respect to the variational parameters γ, the Bayes-by-Backprop (BBB; Blundell et al. 2015) algorithm is used, which combines the variational and reparameterization gradient estimators (Kingma and Welling, 2014) to obtain unbiased and low-variance gradients (Blundell et al., 2015). The BBB approach updates the variational parameters γ using the following optimization procedure:

$$\gamma_{t+1} = \gamma_t - \alpha \nabla_\gamma \mathcal{L}(\gamma) \tag{2.27}$$

where α is the learning rate of the optimizer (e.g., Adam; Kingma and Ba 2014).

For regression tasks, a BNN can be trained by minimizing the mean squared error (MSE) loss function between the predictor $\mathbf{d}$ and the target $\mathbf{h}$:

$$E = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{h}_n - f_\theta(\mathbf{d}_n))^2 \tag{2.28}$$

This loss function is then used to minimize the KL-divergence by calculating the gradients of the ELBO.

Once the BNN is trained, it can be used to make predictions on unseen data. The term "Bayesian" in BNN refers to the fact that the network's weights are sampled from the posterior distribution $p(\mathbf{w}|\mathbf{d},\mathbf{h})$ approximated by the variational distribution $q_\gamma(\mathbf{w})$– but at each evaluation, the trained network returns a single prediction. To estimate the uncertainty of a predicted target $\mathbf{h}^*$, the network must be evaluated multiple times with the same input $\mathbf{d}^*$, and statistics computed over the predictions.

Probabilistic neural networks (PNNs). For problems involving the prediction of continuous variables, however, such conditional averages provide a very limited description of the target variable's properties (Bishop, 1994). In inverse problems, the mapping

is frequently multivalued (non-unique), with input values having multiple valid output values. A neural network with an MSE loss function will roughly represent the conditional average of the target data when applied to such inverse problems, which frequently yields extremely subpar performance—it is not guaranteed that the average of several correct values is itself a correct value (Bishop, 1994). This problem can be solved by a mixture density network (Bishop, 1994), which can represent arbitrary distributions in the same way that a conventional neural network can represent arbitrary functions (Hornik et al., 1989).

A mixture density network (MDN) is a type of neural network that outputs a conditional probability density function $p(\mathbf{h}|\mathbf{d})$ instead of a single prediction for a given input (Bishop, 1994). The MDN is composed of two parts, a neural network and a mixture model. The neural network is used to map the input data to the parameters of the mixture model, which is then used to compute the conditional PDF. The mixture model is composed of K Gaussian components, and the PDF is given by:

$$p(\mathbf{h}|\mathbf{d}) = \sum_{k=1}^{K} \pi_k(\mathbf{d})\, \mathcal{N}(\mathbf{h}|\mu_k(\mathbf{d}), \sigma_k(\mathbf{d})) \tag{2.29}$$

where $\pi_k(\mathbf{d})$, $\mu_k(\mathbf{d})$, and $\sigma_k(\mathbf{d})$ are the mixing weights, means, and standard deviations of the k-th Gaussian component, respectively. The parameters $\pi_k(\mathbf{d})$, $\mu_k(\mathbf{d})$, and $\sigma_k(\mathbf{d})$ are the outputs of the neural network, which is trained by minimizing the negative log-likelihood (NLL) of the data with respect to the parameters $\pi_k(\mathbf{d})$, $\mu_k(\mathbf{d})$, and $\sigma_k(\mathbf{d})$:

$$E = -\sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k(\mathbf{d}_n)\, \mathcal{N}(\mathbf{h}_n|\mu_k(\mathbf{d}_n), \sigma_k(\mathbf{d}_n)) \right] \tag{2.30}$$

Once trained, an MDN can be used to obtain a comprehensive description of the target data, as well as to generate samples from the conditional PDF, which can be used to generate new data points or to explore the space of possible outputs for a given input. The main challenge of training a MDN is that the mixing coefficients (π_k), means (μ_k) and standard deviations (σ_k) of each Gaussian component must be jointly optimized, which is computationally expensive. Moreover, optimizing all parameters simultaneously is numerically unstable in higher dimensions, can lead to degenerate predictions, and MDNs are not exempt to the issue of overfitting (Hjorth and Nabney, 1999; Makansi et al., 2019). Over-fitting happens when a model has been overly optimized on the training dataset and no longer generalizes well to new datasets. An over-fitted model, in other words, has a low error on the training set but a high error on the test set, indicating that the model has memorized the training data points rather than generalizing the underlying patterns.

Demonstration. To demonstrate the performance of the presented networks, we generate a synthetic dataset shown in Figure 2.3, using the Python code available in the appendix (Snippet A.2).

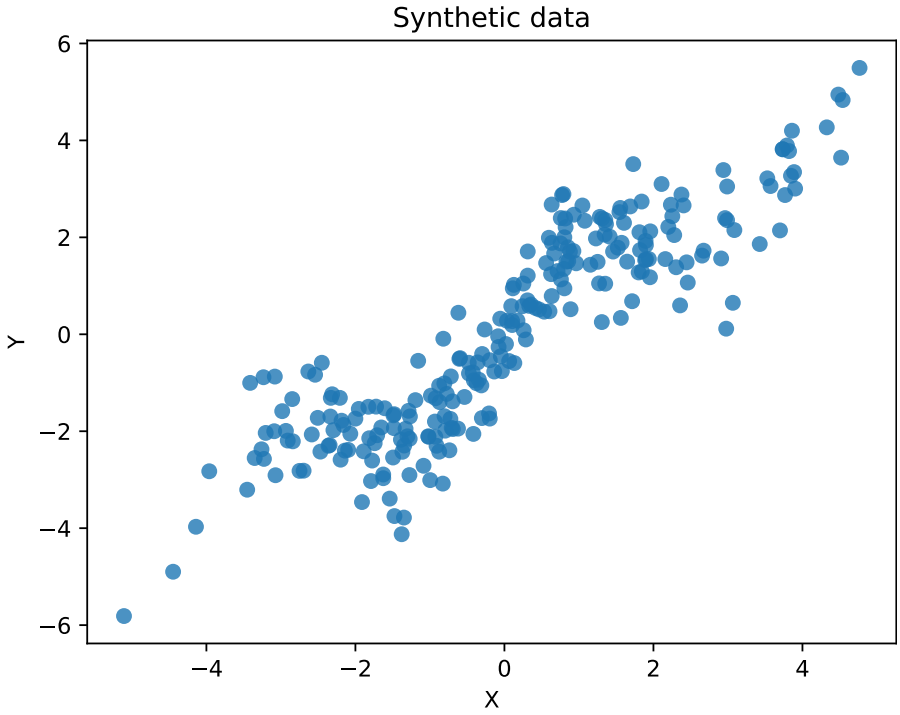

Figure 2.3: Synthetic dataset used to demonstrate the performance of the different networks.

Through these examples, we can see how to use `TensorFlow probability` to construct a deterministic neural network, a Bayesian neural network, a probabilistic neural network, and a probabilistic Bayesian neural network (combining the previous two). The results of the four different neural network architectures are shown in Figure 2.4. In the deterministic neural network (ANN) example, we define a three-layer model using Keras' `Sequential` API (Application Programming Interface), then compile and fit the model using the `Adam` optimizer and the Mean Squared Error (MSE) loss function. The predictions are made using the `model.predict` method, and the results are shown in Figure 2.4A. By maximizing the likelihood of the training data, the ANN is able to effectively fit the training data, but it cannot capture the data's uncertainty.

In the Bayesian neural network (BNN) example, we use the `DenseVariational` API to define our two layer model. We then compile and fit the model using the `Adam` optimizer and the MSE loss function. Finally, to obtain predictions from a BNN, we need to make multiple predictions, as each prediction contains a measure of uncertainty. The results are shown in Figure 2.4B. Although the BNN can capture some uncertainty in the data, each prediction is still a single point estimate, similar to the ANN predictions.

In the probabilistic neural network (PNN) example, we use the `Dense` and the `MixtureNormal` APIs to define our two layer model. We then compile and fit the model using the `Adam` optimizer and the negative log-likelihood loss function. Finally, to obtain predictions from a PNN, we need to sample from the posterior distribution, as each

prediction contains a measure of uncertainty. The results are shown in Figure 2.4C. The PNN can capture data uncertainty, and given enough samples, the predictions cover the entire range of the data.

In the probabilistic Bayesian neural network (PBNN) example, we define our two layer model using the `DenseVariational` API. We then use the `MixtureNormal` API to define a "mixture" of Gaussians with only one component. Finally, we compile and fit the model using the `Adam` optimizer and the negative log-likelihood loss function. To obtain predictions from a PBNN, we need to sample from the posterior distribution, which provides a measure of the uncertainty of the prediction. The results are shown in Figure 2.4D and are similar to the PNN predictions.

To justify the use of a PBNN over a PNN, we must compare the training histories of the various models with a validation set. The level of overfitting can be determined by comparing the training and validation losses; if the validation loss is greater than the training loss, the model is overfitting. The training history of the four models trained on the same data but with a validation set of 10% of the data is shown in Figure 2.5. The training history of the deterministic neural network (ANN) and the Bayesian neural network (BNN) are based on the Mean Squared Error (MSE) loss function, while the training history of the probabilistic neural network (PNN) and the probabilistic Bayesian neural network (PBNN) are based on the negative log-likelihood loss function. For a fair comparison, we subtract the loss from the training loss of each model, standardize the values, and apply a smoothing function to the data (Savitzy-Golay filter). The results are shown in Figure 2.6. The PNN model overfits the data the most, followed by ANN. The BNN model exhibits the least overfitting, while the PBNN model falls between the BNN and ANN models. As a result of combining a Bayesian network and a probabilistic network, we can obtain a model that is less overfitting than a PNN while still allowing us to quantify the full uncertainty of the predictions.

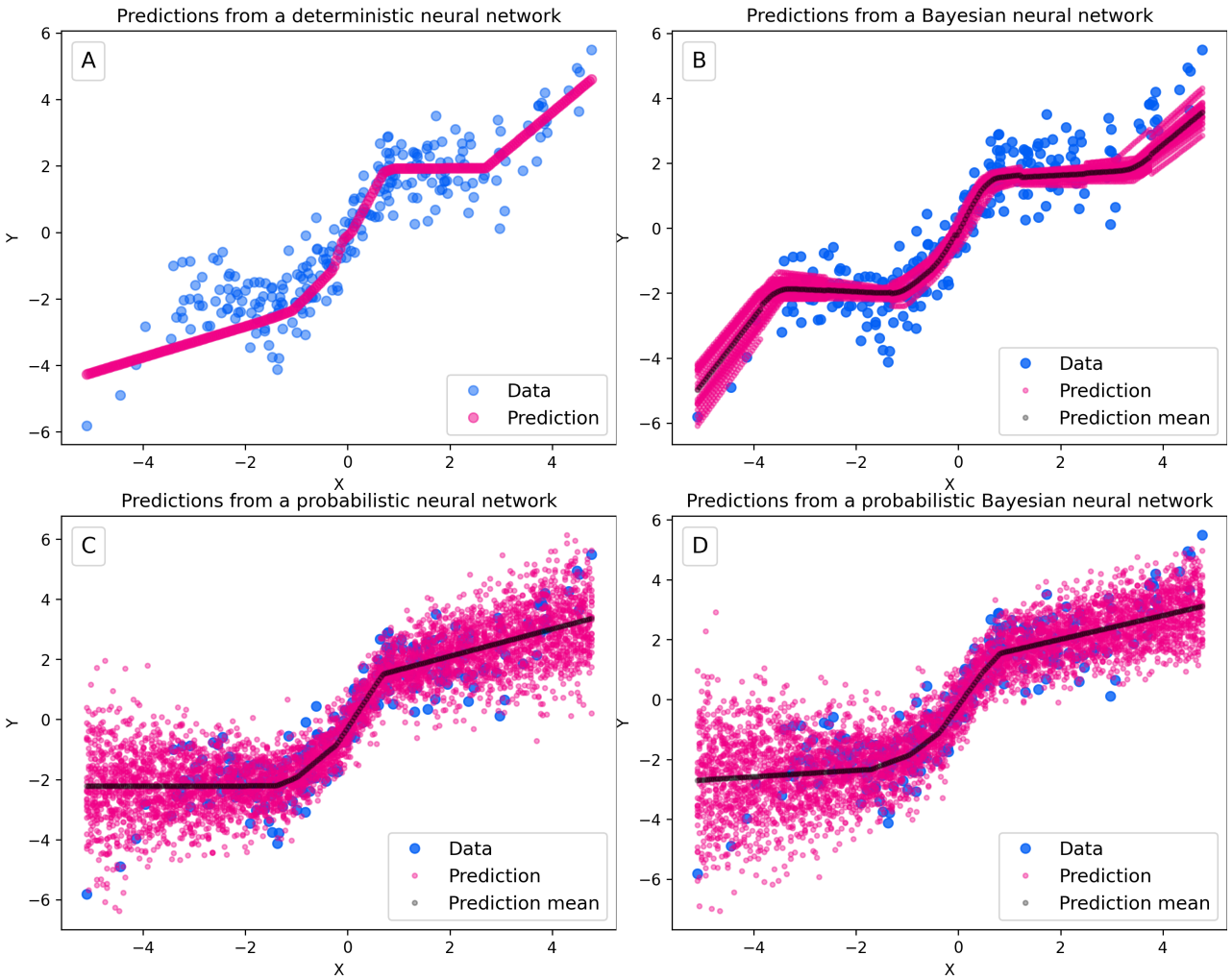

Figure 2.4: Comparison of the (**A**) deterministic neural network (ANN), (**B**) the Bayesian neural network (BNN), (**C**) the probabilistic neural network (PNN), and (**D**) the probabilistic Bayesian neural network (PBNN) on a synthetic dataset.

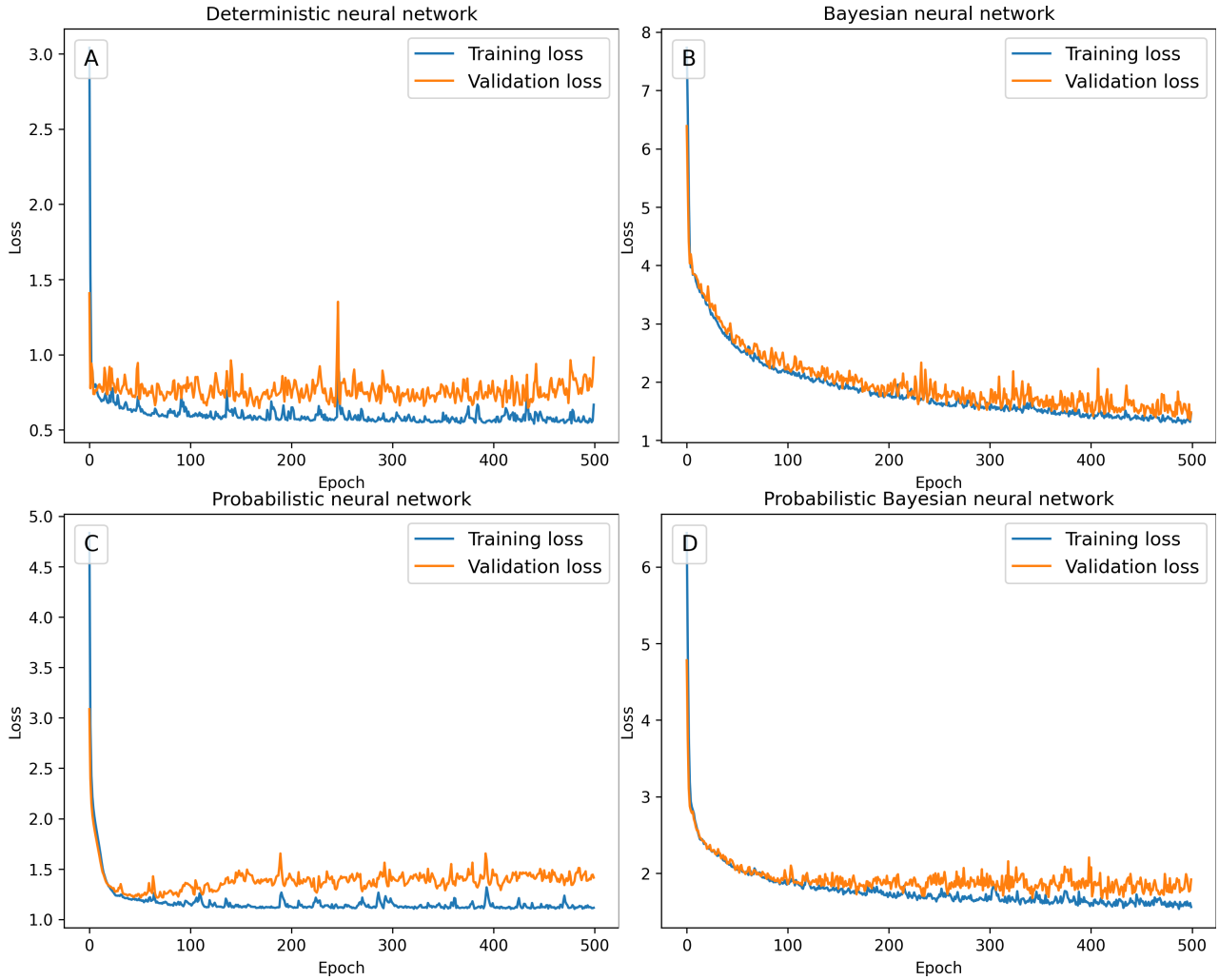

Figure 2.5: Training history of the (**A**) deterministic neural network (ANN), (**B**) the Bayesian neural network (BNN), (**C**) the probabilistic neural network (PNN), and (**D**) the probabilistic Bayesian neural network (PBNN) on a synthetic dataset.

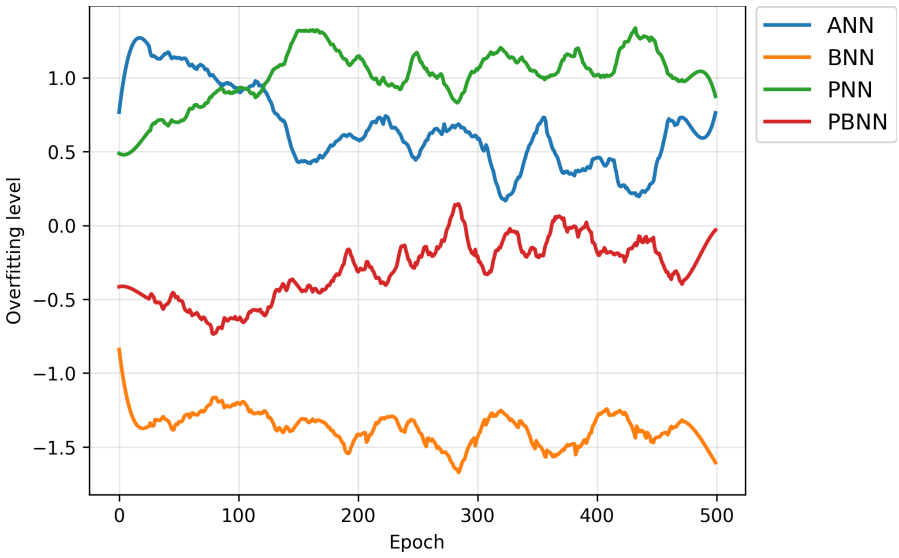

Figure 2.6: Comparison of the scaled and standardized differences between validation and training losses of the (**A**) deterministic neural network (ANN), (**B**) the Bayesian neural network (BNN), (**C**) the probabilistic neural network (PNN), and (**D**) the probabilistic Bayesian neural network (PBNN) on a synthetic dataset.

2.4 Bayesian optimal experimental design

It is better to be roughly right than precisely wrong

John Maynard Keynes

Note. In order to avoid confusion in the mathematical notation in this section, we will use the classical notation of $\mathbf{x}$ and $\mathbf{y}$ to represent the previous notation of $\mathbf{d}$ and $\mathbf{h}$ for the predictors and the targets, respectively.

2.4.1 Introduction

Bayesian optimal design entails defining a design criterion, or a utility function $\mathcal{J}(\xi, x, y)$, which describes the value (based on the experimental goals) of selecting the design ξ from the design space Ξ yielding predictor x, with target y. A probabilistic model $p(x|\xi)$ is also needed, which includes a likelihood $p(x|\xi, \theta)$ for observing a new set of measurements x under the design ξ, given parameter values y, and a prior distribution $p(y)$ for the parameters y (Lindley, 1972). Typically, the prior distribution $p(y)$ is assumed to be independent of the design ξ, i.e., $p(y|\xi) = p(y)$ (Ryan et al., 2016).

The optimal design ξ^* minimizes the expected utility function $\mathcal{J}(\xi)$ over the design space Ξ with respect to the future predictors x and the target y:

$$\begin{aligned}\xi^* &= \arg\min_{\xi\in\Xi} \mathbb{E}_{x,y}\left[\mathcal{J}(\xi,x,y)\right] \\ &= \arg\min_{\xi\in\Xi} \int_x \int_y \mathcal{J}(\xi,x,y)p(y,x|\xi)p(y)dydx \\ &= \arg\min_{\xi\in\Xi} \int_x \underbrace{\int_y \mathcal{J}(\xi,x,y)p(x|\xi,y)}_{\text{posterior expected utility}}\, p(x|\xi)dx\end{aligned} \tag{2.31}$$

Thus, given the observed data, the optimal design minimizes the posterior expected utility (Ryan et al., 2016). Equation 2.31 does not usually have a closed form solution unless the likelihood and prior are specifically chosen to allow analytic evaluation of the integration problem. To solve the minimization and integration problem, numerical approximations or stochastic solution methods are required, e.g., Monte Carlo methods (see Ryan et al. (2016) for a complete review). In this dissertation, we use the BEL framework to solve the minimization and integration problem in Equation 2.31.

In the BEL framework, the likelihood $p(x|\xi,y)$ is estimated using the forward model (cf. §2.1). Our utility function is a functional of the posterior distribution $p(y|x,\xi)$, which is directly estimated from the predictor x using the BEL framework.

In practice, this entails the following steps:

1. Estimate the posterior distribution $p(y|x,\xi)$ for each possible data set;
2. Calculate the data utility function for each posterior distribution;
3. Sum up the data utility functions over all possible data sets (i.e., integrate over all possible data sets).

Step 1 requires solving the inverse problem for each possible data set. For this reason, classic inversion techniques might become intractable. The BEL framework can be used to efficiently solve the integration problem using a forward model and machine learning (cf. §2.1).

2.4.2 Utility functions

The objective function $\mathcal{J}(\xi,x,y)$ is a utility function that describes the value of selecting the design ξ from the design space Ξ yielding predictor x, with target y.

Information theory-based utility functions. The utility function $\mathcal{J}(\xi,x,y)$ can be defined as the gain in Shannon information (cf. §1.3) between the prior distribution $p(y)$ and the posterior distribution $p(y|x,\xi)$, which is given by:

$$\mathcal{J}(\xi,x,y) = \mathcal{J}_{IG} = -\left[H(p(y)) - H(p(y|x,\xi))\right], \tag{2.32}$$

where $H(p(y))$ is the entropy of the prior distribution $p(y)$ and $H(p(y|x,\xi))$ is the entropy of the posterior distribution $p(y|x,\xi)$. The minus sign is used to respect the convention

that utility functions are minimized. Another common utility function is the KL divergence between the prior distribution $p(y)$ and the posterior distribution $p(y|x,\xi)$, which is given by:

$$\mathcal{J}(\xi,x,y) = \mathcal{J}_{KL} = -D_{KL}(p(y) \,||\, p(y|x,\xi)). \tag{2.33}$$

The KL divergence is a measure of the distance between the two probability distributions $p(y)$ and $p(y|x,\xi)$. When using $\mathcal{J}_{KL}$, we are looking for the design ξ that maximizes the distance between the prior distribution $p(y)$ and the posterior distribution $p(y|x,\xi)$. As a result, we enforce the minus sign once more to stick to the convention that utility functions be minimized.

While these utility functions are widely used, they have one significant flaw for our purposes. The issue is that they will only measure the information gain associated with an increase in the precision of the posterior $p(y|x,\xi)$ relative to the prior $p(y)$, but they will not account for the increase in accuracy regarding the true value y^*. We are interested in the information gain related to the true value because we are using BOED through supervised learning and the true targets are known.

To address this issue, we propose a new utility function that measures the information gain associated with an increase in the precision of the posterior $p(y|x,\xi)$ relative to the prior $p(y)$, but also accounts for the increase in accuracy regarding the true value y^*, as applied in Chapter 5. We first consider an ideal distribution $p(y^*)$ that is centered at the true value y^* and has a small variance, arbitrarily defined as $\sigma^2_{y^*} = 0.0001$:

$$p(y^*) = \mathcal{N}(y^*, \sigma^2_{y^*}). \tag{2.34}$$

We then define the utility function $\mathcal{J}(\xi,x,y)$ as the KL divergence between the ideal distribution $p(y^*)$ and the posterior distribution $p(y|x,\xi)$, which is given by:

$$\mathcal{J}(\xi,x,y) = \mathcal{J}^*_{KL} = D_{KL}(p(y^*) \,||\, p(y|x,\xi)). \tag{2.35}$$

This new utility function has two advantages over the existing utility functions. First, it accounts for the increase in accuracy regarding the true value y^*. Second, it severely penalizes overconfident models, which have a very low variance in the posterior distribution $p(y|x,\xi)$, but are far from the true value y^*.

In order to illustrate how these utility functions perform, we consider a simple example where the prior distribution $p(y)$ is a uniform distribution over the interval $[-4,4]$. The true target value $y^* = 1.4$, and we consider three posterior distributions $p(y|x,\xi)_i$, where $i = 1,2,3$:

$$\begin{aligned} p(y|x,\xi)_1 &= \mathcal{N}(y; 1.4, 1.5) \\ p(y|x,\xi)_2 &= \mathcal{N}(y; 1.4, 0.1) \\ p(y|x,\xi)_3 &= \mathcal{N}(y; -2, 0.4) \end{aligned} \tag{2.36}$$

where $\mathcal{N}(y; \mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ. $p(y|x,\xi)_1$ and $p(y|x,\xi)_2$ are both centered around the true value y^*, but $p(y|x,\xi)_2$ has a much lower standard deviation than $p(y|x,\xi)_1$. $p(y|x,\xi)_3$ is centered around a different value than y^*, but it has a much lower standard deviation than $p(y|x,\xi)_1$ (Figure 2.7).

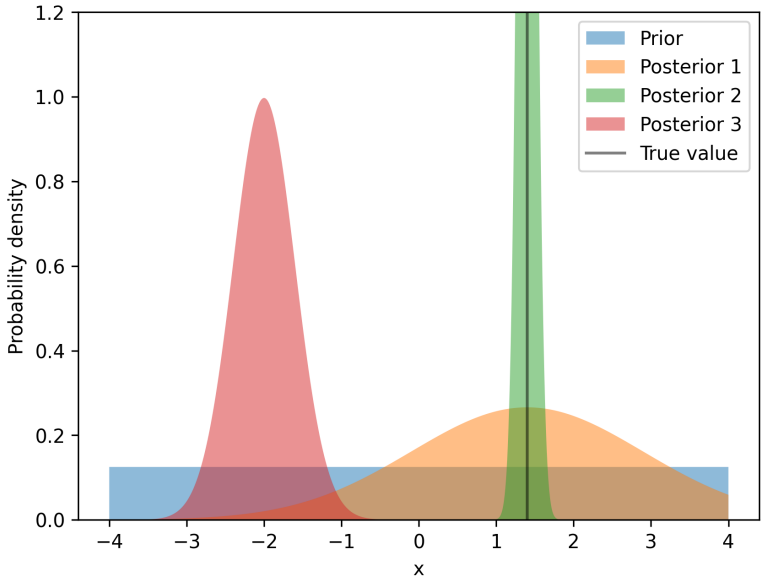

Figure 2.7: Illustration of the different Gaussian posterior distributions $p(y|x,\xi)_i$ given a uniform prior distribution $p(y)$ and a true value $y^* = 1.4$.

The score of each posterior distribution $p(y|x,\xi)_i$ is computed using the utility functions $\mathcal{J}_{IG}$, $\mathcal{J}_{KL}$, and $\mathcal{J}^*_{KL}$, which are given by:

Posterior	$\mathcal{J}_{IG}$	$\mathcal{J}_{KL}$	$\mathcal{J}^*_{KL}$
$p(y\|x,\xi)_1$	-0.26	-0.84	4.5
$p(y\|x,\xi)_2$	-2.96	-368	1.8
$p(y\|x,\xi)_3$	-1.58	-25.72	39.3

Table 2.1: Note. Utility scores of the three posterior distributions $p(y|x,\xi)_i$ using the utility functions $\mathcal{J}_{IG}$, $\mathcal{J}_{KL}$, and $\mathcal{J}^*_{KL}$. Since the natural logarithm is used, the unit of the utility scores is *nats*.

To allow for a better comparison, the scores are standardized over posteriors $p(y|x,\xi)_i$, and the matrix plot in Figure 2.8 shows the scores of the three utility functions for each posterior distribution $p(y|x,\xi)_i$. Blue colors indicate less penalty, while red colors indicate more penalty.

Compared to $\mathcal{J}_{IG}$ and $\mathcal{J}_{KL}$, $\mathcal{J}^*_{KL}$ does not either over-penalize or under-penalize the posterior distributions $p(y|x,\xi)_1$. It is also less sensitive than $\mathcal{J}_{KL}$ to the variance of the posterior distribution $p(y|x,\xi)_2$, which is centered around the true value y^*. Finally,

$\mathcal{J}^*_{KL}$ detects the overconfident posterior distribution $p(y|x,\xi)_3$, which is centered on a different value than y^*, and penalizes it much more severely than $\mathcal{J}_{IG}$ and $\mathcal{J}_{KL}$, which instead favor this information gain regardless of its accuracy.

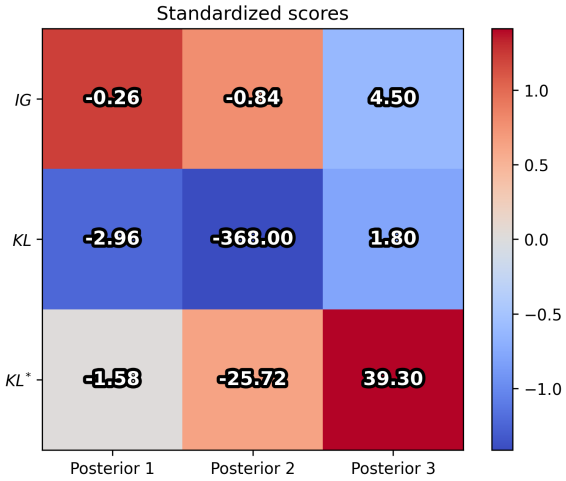

Figure 2.8: Matrix plot of the utility scores of the three utility functions for each posterior distribution $p(y|x,\xi)_i$.

A simpler utility function. We consider a simpler utility function that is based on the Euclidean distance between samples from the posterior distribution $p(y|x,\xi)$ and the true value y^*. Popular metrics such as the root mean squared error (RMSE) and the mean absolute error (MAE) are suitable for this purpose. We can therefore define the utility function $\mathcal{J}_{RMSE}$ as:

$$\mathcal{J}_{RMSE} = \mathbb{E}_{p(y|x,\xi)}\left[(y-y^*)^2\right]^{1/2} \tag{2.37}$$

where $\mathbb{E}_{p(y|x,\xi)}$ is the expectation over the posterior distribution $p(y|x,\xi)$. It is trivial to deduce the utility function $\mathcal{J}_{MAE}$. The utility function $\mathcal{J}_{RMSE}$ can act as a flexible and easy-to-implement alternative to $\mathcal{J}^*_{KL}$. Chapter 4 will demonstrate that $\mathcal{J}_{RMSE}$ can be applied and weighted across the numerous dimensions of the posterior distribution $p(y|x,\xi)$ to assess the information gain of multiple designs in a time-efficient manner.

Image-difference utility function. Some applications require the utility function to be based on the difference between images, which typically have a high dimensionality. For instance, see Chapter 3 for an application in which the utility function is based on the difference between images.

2.5 Software Implementation

> *"a computer is not just useful but indispensable for finding a Bayes solution for a design problem."*
>
> Davis et al. (1972)

We implemented the proposed framework in the Python programming language. The proposed package is called `SKBEL` (scikit-BEL; Thibaut and Ramgraber 2021) and is based on the `scikit-learn` package (Pedregosa et al., 2011). Just as BEL, it is a modular framework that allows users to easily combine different components.

In general, the user can choose three different types of components: (1) the data preprocessor, (2) the regression algorithm, and (3) the postprocessor. These components are referred to as `Pipeline` objects in `scikit-learn`. From a user's perspective, the BEL object itself is a `Pipeline` object that combines the preprocessor, regressor, and postprocessor, through which will flow the data.

Note that passing either no preprocessor or no postprocessor is also a valid option, the data will simply flow through the BEL object without any modifications.

The data preprocessor is responsible for transforming the data into a form that is suitable for the regressor algorithm. It can include any of the following steps, or their combination: (1) data normalization, (2) data standardization, (3) data imputation, (4) data discretization, (5) data encoding, (6) data filtering, and (7) data transformation, including dimensionality reduction.

The regression algorithm is responsible for learning the mapping between the input and output variables. It can include any of the previously described regression algorithm allowing uncertainty quantification. Some of these algorithms are implemented in `scikit-learn` (Pedregosa et al., 2011), while others are implemented in other popular packages, such as `TensorFlow probability` (Dillon et al., 2017).

The postprocessor is responsible for transforming the output of the regressor algorithm into a form that is suitable for the user. It can include the similar steps as the preprocessor.

The strength of the `Pipeline` paradigm is that it allows to easily combine, transform, scale, back-transform, and visualize the data at each step of the pipeline.

The main BEL object has the following methods:

- `fit(d, h)`: Fit all the pipelines. Learn the model parameters from the predictor `d` and target `h`. Required before calling `transform`, `predict`, and `random sample`.

- `transform(d, h)`: Transform the predictor `d` and target `h` using all the pipelines.
- `fit_transform(d, h)`: Combine the `fit` and `transform` methods. Fit and transform the predictor `d` and target `h` using all the pipelines.
- `predict(d)`: Determines the posterior distribution of the target `h` given the predictor `d`. Requires `fit(d, h)` to be called before. It can return posterior samples of the target `h` by passing `return_samples=True`.
- `random_sample()`: Use the learned model parameters to generate random samples for the input predictor `d`. Requires `predict(d)` to be called before. It can be called within the `predict(d)` method, by setting the `return_sample` parameter to `True`.
- `inverse_transform(d, h)`: Inverse transform the predictor `d` and target `h` through all the pipelines. Requires `fit(d, h)` to be called before.

If CCA is used as the regression algorithm, the user will have to choose between Multivariate Gaussian Inference, Kernel Density Estimation, and Transport Maps to sample the posterior distribution of the target `h` given the predictor `d`.

SKBEL demonstration

`SKBEL` can be installed on any operating system (Windows, Linux, Mac OS) using the Python package manager `pip`:

```
pip install skbel
```

The other required packages are `scikit-learn` and `numpy`, which can be installed using the same command.

We demonstrate the proposed framework using the training data from Chapter 3. An online, interactive version of this demonstration is available at `https://www.kaggle.com/code/robustus/whpa-get-started`. For this demonstration, we are not yet concerned with the nature of the data, as it will be described in great detail in the next chapter; we are presently only interested in the framework. The predictor variable `d` has 1200 dimensions, while the target variable `h` has 8700 dimensions.

We start by creating a BEL object with a preprocessor, a regressor, and a postprocessor. The preprocessor is a `StandardScaler` object, which standardizes the data by scaling to unit variance, followed by a `PCA` object, which performs dimensionality reduction by projecting the data to a lower dimensional space. The regressor is a `CCA` object, which performs canonical correlation analysis. The postprocessor is a `PowerTransformer` object, which transforms the canonical variate pairs to make it more Gaussian-like.

The following code snippet shows how to use the proposed framework to learn the mapping between the predictor `d` and the target `h` using the CCA algorithm with the three different inference methods.

We start by importing the required packages:

```python
import numpy as np  # NumPy

# Canonical Correlation Analysis
from sklearn.cross_decomposition import CCA
# Principal Component Analysis
from sklearn.decomposition import PCA
# Pipeline object which allows to combine different components together
from sklearn.pipeline import Pipeline
# Split the data into training and testing sets
from sklearn.model_selection import train_test_split
# Preprocessing methods
from sklearn.preprocessing import StandardScaler, PowerTransformer

# Our custom package
from skbel import BEL
```

Listing 2.1: Importing the required packages.

For simplicity, we define a function `init_bel` that returns a BEL object initialized with the CCA algorithm and the specified preprocessor and postprocessor.

```python
def init_bel():
    """Set all BEL pipelines.
    This is the blueprint of the framework.
    """

    # Pipeline before CCA
    X_pre_processing = Pipeline(
        [
            ("scaler", StandardScaler(with_mean=False)),
            ("pca", PCA(n_components=50)),
        ]
    )
    Y_pre_processing = Pipeline(
        [
            ("scaler", StandardScaler(with_mean=False)),
            ("pca", PCA(n_components=30)),
        ]
    )

    # Canonical Correlation Analysis
    cca = CCA(n_components=30)  # we choose the maximum number of
    components

    # Pipeline after CCA
    X_post_processing = Pipeline(
        [("normalizer", PowerTransformer(method="yeo-johnson",
    standardize=True))]
    )
    Y_post_processing = Pipeline(
```

```
        [("normalizer", PowerTransformer(method="yeo-johnson",
    standardize=True))]
    )

    # Initiate BEL object
    bel_model = BEL(
        X_pre_processing=X_pre_processing,
        X_post_processing=X_post_processing,
        Y_pre_processing=Y_pre_processing,
        Y_post_processing=Y_post_processing,
        regression_model=cca,
    )

    return bel_model
```

Listing 2.2: Function that returns a BEL object initialized with the CCA algorithm and the specified preprocessor and postprocessor.

We then load the training data and initialize the BEL object. For this demonstration, we extract one sample from the training data to use as the test data:

```
# Load training data
d_train = np.load("d_train.npy")  # predictor (200 rows, 1200 columns)
h_train = np.load("h_train.npy")  # target (200 rows, 8700 columns)

# Split training data into training and test sets
d_train, d_test, h_train, h_test = train_test_split(d_train, h_train,
    test_size=1, random_state=42)
```

Listing 2.3: Loading the training data and initializing the BEL object.

In this demonstration, we consider very small dataset, with only 200 samples, i.e., 199 samples for training and 1 sample for testing. The training data is used to fit the BEL object with the CCA algorithm and the three different inference methods:

```
# Initiate BEL object
bel_model = init_bel()
# The BEL object only needs to be fitted once
bel_model.fit(d_train, h_train) # fit the model
```

Listing 2.4: Training the BEL model

We can now use the trained BEL object to predict the test target $\mathbf{h}_{\mathbf{test}}$ given the predictor $\mathbf{d}_{\mathbf{test}}$. We choose an arbitrary number of samples to generate, in this case 200. We specify `inverse_transform=False` to return the samples in the canonical space instead of the original space.

```python
# Predict on test data using the trained model for the three different modes
bel_model.mode = "mvn"  # Multivariate Gaussian Inference
samples_mgi = bel_model.predict(d_test, n_posts=200, inverse_transform=False) # predict samples

bel_model.mode = "kde"  # Kernel Density Estimation
samples_kde = bel_model.predict(d_test, n_posts=200, inverse_transform=False) # predict samples

bel_model.mode = "tm"  # Transport Maps
samples_tm = bel_model.predict(d_test, n_posts=200, inverse_transform=False) # predict samples
```

Listing 2.5: Predicting the target given the predictor

We can now compare the predicted samples with the actual test target $\mathbf{h_{test}}$.

```python
# Let's first transform the train and test target to the canonical space
d_test_canon, h_test_canon = bel_model.transform(d_test, h_test)
d_train_canon, h_train_canon = bel_model.transform(d_train, h_train)

# Let's produce a histogram of the predictions
import matplotlib.pyplot as plt

# Compare the histograms of the predictions for the three different modes
comp_n = 0  # we select the first canonical variate pair
plt.hist(samples_mgi[0][:, comp_n], alpha=0.5, label="MGI", density=True)
plt.hist(samples_kde[0][:, comp_n], alpha=0.5, label="KDE", density=True)
plt.hist(samples_tm[0][:, comp_n], alpha=0.5, label="TM", density=True)
plt.axvline(h_test_canon[0, comp_n], color="k", label="True value")
plt.legend()
plt.xlabel("First canonical dimension of $\mathbf{h}$")
plt.ylabel("Density")
plt.show()
```

Listing 2.6: Comparison of the predicted samples with the actual test target.

The results are shown in Figure 2.9. The three histograms are very similar, with the MGI method producing slightly more samples in the tails of the distribution.

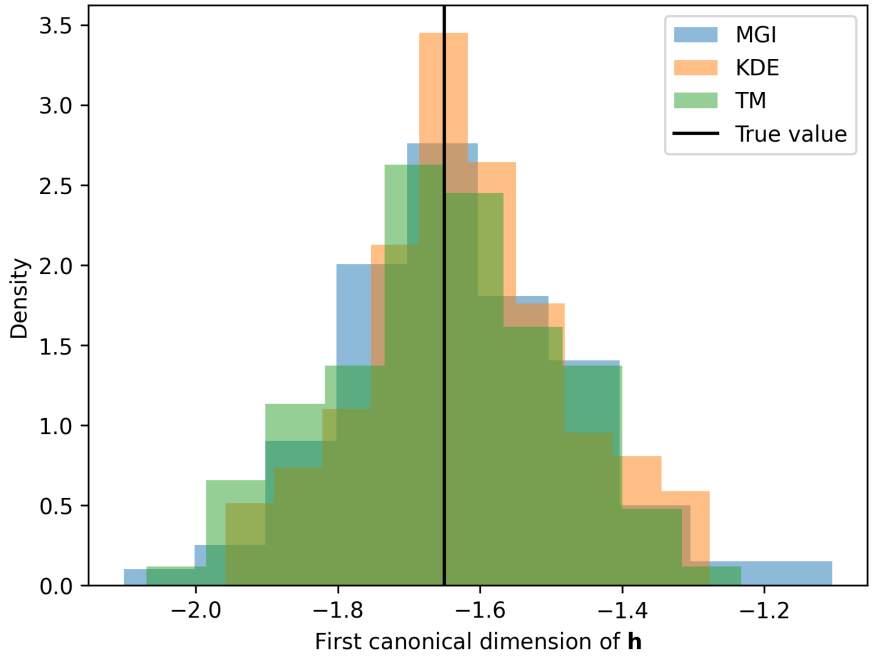

Figure 2.9: Histogram of the predicted samples for the three different inference methods.

We can also compare the predicted samples with the training data.

```
# Compare the predictions with the training data
plt.hist(h_train_canon[:, comp_n], alpha=0.5, label="Train", density=True
    , color="gray", log=True)
plt.hist(samples_mgi[0][:, comp_n], alpha=0.5, label="MGI", density=True,
     log=True)
plt.hist(samples_kde[0][:, comp_n], alpha=0.5, label="KDE", density=True,
     log=True)
plt.hist(samples_tm[0][:, comp_n], alpha=0.5, label="TM", density=True,
    log=True)
plt.axvline(h_test_canon[0, comp_n], color="k", label="True value")
plt.legend()
plt.xlabel("First canonical dimension of $\mathbf{h}$")
plt.ylabel("Density")
plt.show()
```

Listing 2.7: Compare the predicted samples with the training data.

The results are shown in Figure 2.10. The log scale is used to better visualize the differences between the histograms.

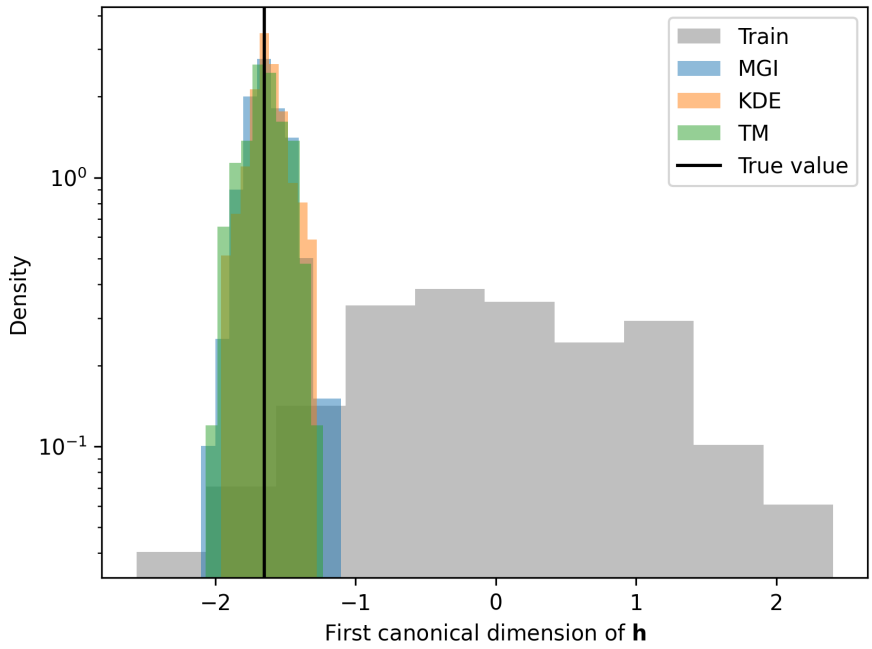

Figure 2.10: Histogram of the predicted samples for the three different inference methods compared with the training data in canonical space.

3. A new framework for experimental design using Bayesian Evidential Learning: the case of wellhead protection area

This chapter was published in Journal of Hydrology (Thibaut et al., 2021b):

Thibaut, Robin, Eric Laloy, and Thomas Hermans (Dec. 2021). "A new framework for experimental design using Bayesian Evidential Learning: The case of wellhead protection area". In: Journal of Hydrology 603, p. 126903. issn: 00221694.
doi: v10.1016/j.jhydrol.2021.126903.

CRediT author statement. Robin Thibaut: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Eric Laloy**: Conceptualization, Methodology, Writing - Review & Editing, Supervision. **Thomas Hermans**: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Abstract

Groundwater management practices can have significant socioeconomic impacts, such as sustainable drinking water extraction or contamination protection. A complete uncertainty analysis should ideally be performed to anticipate all possible outcomes and assess any risk. Uncertainties arise due to our limited understanding of the physical processes involved, as well as a scarcity of measurement data, whether directly or indirectly related to the physical parameters of interest. In this chapter, we use a small number of tracing experiments (predictor) to predict the wellhead protection area (WHPA, target), the shape and extent of which are influenced by the distribution of hydraulic conductivity (K). Our first goal is to make stochastic predictions of the WHPA within the Bayesian Evidential Learning (BEL) framework, which uses machine learning to find a direct relationship between predictor and target. This relationship is learned using a small number of training models (400) drawn from the prior distribution of K. Forward modelling is used to obtain the 400 pairs of simulated predictors and targets. Newly collected field

data can then be used directly to predict the approximate posterior distribution of the corresponding WHPA, obviating the need for the traditional step of data inversion. The number and location of data sources (injection wells) influence the posterior WHPA distribution's uncertainty range. Our second goal is to extend BEL to determine the optimal design of data source locations that minimises the WHPA's posterior uncertainty. Because the BEL model, once trained, allows the computation of the posterior uncertainty corresponding to any new input data, experimental design can be done explicitly, without averaging or approximating. We estimate the WHPA's posterior uncertainty range using the Modified Hausdorff Distance (MHD) and Structural Similarity (SSIM) index metrics. Because the breakthrough curves store information on a large area of the K field surrounding the pumping well, increasing the number of injection wells reduces the derived posterior WHPA uncertainty. Our method can also estimate which injection wells are more informative than others, as demonstrated by a k-fold cross-validation procedure. Overall, the application of BEL to experimental design allows for identifying data sources that maximise the information content of any measurement data while keeping budget constraints and computational costs to a minimum.

Plain language summary

Freshwater extraction from the ground requires careful pumping in order to preserve the resources. The area surrounding a pumping well, known as a Wellhead Protection Area (WHPA), must be protected from human contamination. This area is determined by the velocity of groundwater in the subsurface. In this chapter, we present a method for predicting the WHPA while accounting for the uncertainty inherent in our understanding of the subsurface. We also demonstrate how to apply our method to experimental design, i.e., finding the best locations for measurement data to reduce uncertainty in WHPA prediction. Our proposed method is applicable to other types of predictions and is especially useful in groundwater management applications.

Key points

- BEL is used to determine the optimal design of data source locations that minimises WHPA posterior uncertainty;
- The Traveling Salesman algorithm is used to obtain the WHPA delineation;
- We use the Modified Hausdorff Distance (MHD) and Structural Similarity (SSIM) index metrics to quantify the WHPA uncertainty;

3.1 Introduction

The Wellhead Protection Area (WHPA) is defined as the zone surrounding a pumping well where human activities are restricted in order to protect water resources (Goldscheider, 2010), generally based on the amount of time harmful contaminants within the area

will take to reach the pumping well (according to local regulation). It is determined by the flow velocity in the subsurface surrounding the well, and it can be calculated numerically using particle tracking or transport simulation, or in practice, using tracer tests (Dassargues, 2018; Goldscheider, 2010). Typically, a groundwater model is calibrated against field data before calculating the WHPA using the calibrated model. The establishment of such zones can have a significant socioeconomic impact in densely populated areas where land occupation is a major concern. As a result, in order to best support decision-making, the outcome of each possible event should be quantified.

Traditional deterministic calibration methods for computing the dimensions of this area may not be appropriate (Kikuchi et al., 2015; Zhou et al., 2014) because they do not account for the uncertainty inherent in such prediction problems, which stems primarily from our limited knowledge of the subsurface's heterogeneity. Instead, stochastic methods should be used to assess the entire range of possible outcomes, allowing for a thorough risk analysis to serve as the foundation for decision making (de Barros et al., 2012; Linde et al., 2017; Zhou et al., 2014). The drawback of stochastic methods is their computational cost. Generally, the solution is obtained by iterative methods requiring many runs of the forward problem such as Markov chain Monte Carlo (McMC) methods (Laloy and Vrugt, 2012; Vrugt, 2016) or stochastic optimisation (Hermans et al., 2015a; Hu et al., 2001) to fit the observed data. The number of iterations increases with the complexity of the mathematical models used to describe the phenomena at hand and the number of parameters used to describe the subsurface models, discouraging uncertainty analysis or sensitivity analyses of groundwater models in practical applications.

These stochastic methods, by design, not only allow for the computation of a forecast range, but they can also lead to experimental (or optimal) design for the optimisation of operations within limited budget constraints. Experimental design is generally defined as finding the data set that minimises the uncertainty of a specific prediction, which can be expressed as maximising or minimising a data utility function (e.g., Kikuchi et al. 2015). However, the computational burden is even more important because experimental design assumes that the observed data is not yet known, requiring the search for the stochastic solution to the inverse problem for any possible outcome of the unknown data (e.g., Leube et al. 2012; Neuman et al. 2012a). Two main simplifications have been proposed to make practical applications tractable: (1) Bayesian Model Averaging (BMA) combined with preposterior estimation (Kikuchi et al., 2015; Neuman et al., 2012b; Pham and Tsai, 2016; Raftery et al., 2005; Samadi et al., 2020; Tsai and Li, 2008; Vrugt and Robinson, 2007; Wöhling and Vrugt, 2008), and (2) surrogate modelling (Asher et al., 2015; Babaei et al., 2015; Laloy et al., 2013; Razavi et al., 2012; Tarakanov and Elsheikh, 2020; Zhang et al., 2015, 2020).

The BMA approach is extensively described in Kikuchi et al. (2015); Raftery et al. (2005); Samadi et al. (2020). In essence, BMA was designed to address structural uncertainty in subsurface problems by estimating the posterior distribution by averaging over a large number of Monte Carlo simulations. Because the method quickly becomes intractable for experimental design, BMA often applies a pre-posterior estimation tech-

nique with simplifying assumptions to estimate the expected value of the chosen data utility function by computing the average of several ensemble-generated realisations of the prospective data set, rather than computing the full posterior (Leube et al., 2012; Lu et al., 2012). Kikuchi et al. (2015) provide an example of how the approach was extended to propose a novel experimental design approach in hydrology called Discrimination-Inference (DI) that can be used for conceptual and predictive discrimination. The latter is based on the Kullback-Leibler divergence, which measures the effect of additional data collection on prior and posterior probability distributions. Kikuchi and colleagues began by calibrating N sets of model parameters conditioned on existing data in order to populate an input matrix using McMC sampling. After obtaining a set of randomly sampled parameters, they used forward modelling to generate data realisations in order to estimate the data utility function with BMA. The posterior distribution of the prediction, which is at the heart of the data utility function, is never computed in their approach.

Another approach to reducing the computational cost of experimental design is to use surrogate models. Razavi et al. (2012) proposed a comprehensive review of surrogate modelling in the field of water resources. The surrogate model of a computationally intensive process is used to estimate either an objective function, constraints, or both. Razavi et al. (2012) state that surrogate modelling can be divided into two main categories: high-fidelity response surface models and low-fidelity models, with fidelity referring to the level of realism of simulation models. High-fidelity response surface models emulate the numerical outputs of the original model. This method uses statistical models or empirical data-driven models to determine a relationship between model parameters and one or more model response variables, using techniques such as kriging (Baú and Mayer, 2006; Garcet et al., 2006), artificial neural networks (Kourakos and Mantoglou, 2009; Yan and Minsker, 2006), radial basis functions (Regis and Shoemaker, 2005), and polynomial chaos expansion (Laloy et al., 2013; Tarakanov and Elsheikh, 2020), to name a few.

Low-fidelity models are physically based. They are, in essence, simplified, less faithful versions of their computationally demanding parent models. Low-fidelity models must be "reasonably close" to the response of the original model in order to be used in practice. According to Razavi et al. (2012), as the number of model variables increases, surrogate modelling becomes less beneficial or even impractical, resulting in a reduction in analysis accuracy. Once an appropriate surrogate model is found, it can efficiently solve stochastic inversion and experimental design problems and at a low computational cost. However, the approximation can cause significant bias in the prediction, resulting in incorrect estimation of the data utility function (Asher et al., 2015; Babaei et al., 2015; Laloy et al., 2013; Razavi et al., 2012; Tarakanov and Elsheikh, 2020; Zhang et al., 2015, 2020).

In this chapter, we propose an alternative method for solving experimental design studies using the Bayesian Evidential Learning (BEL) framework (Hermans et al., 2018; Scheidt et al., 2018) in the context of WHPA estimation from tracers' breakthrough

curves.

The goal of this chapter is twofold: first, we demonstrate BEL's prediction capabilities for WHPA estimation. Secondly, we present a BEL-based experimental design approach for identifying informative hydrological data. Both of these parts are based on the same synthetic study.

The rest of this chapter is structured as follows. Section §3.2 develops the theoretical foundation of our framework, followed by Section §3.3, which demonstrates BEL's capabilities for prediction and uncertainty quantification. Section §3.4 then shows how we design the optimisation using the BEL framework to make the best injection well location choice. It investigates the integration of multiple-well combinations, as well as the influence of structural uncertainty in the approach. Finally, Section §3.5 discusses the approach's limitations, and Section §3.6 provides a short conclusion.

3.2 Methodology

3.2.1 WHPA prediction

In this section, we propose a method for predicting the posterior distribution of an unknown WHPA given observed breakthrough curves (BCs) using BEL. Different tracers originate from a maximum number s of data sources (injection wells) located at various locations around the pumping well. Both the target and the predictor are high-dimensional, and their relationship is non-linear. Both variables are pre-processed, and their dimensionality is reduced before training the model and performing multivariate regression in order to learn a direct relationship.

Pre-processing

Predictor The predictor is a set of BCs obtained from N solute transport simulations and measured at the pumping well. Each BC is a one-dimensional concentration time series, with one concentration value for each time step of each simulation. The N models do not produce BCs of the same dimension because the used groundwater solute transport software (MT3D-USGS; Bedekar et al. 2016) has an adaptive time-step. As a consequence, at k predefined time steps, we interpolate the breakthrough curves and extract the corresponding concentration, and each simulation yields a data set of dimensions $(\lambda \times k)$. λ is the number of data sources employed $(1 \leq \lambda \leq s)$. For N simulations, the predictor matrix is $\mathbf{B}^{\lambda}\,(n \times \lambda \times k)$. The last dimension of $\mathbf{B}^{\lambda}$, with k elements, contains only concentration values $(\frac{kg}{m^3})$. The time dimension is irrelevant as long as they share the same time steps. The shape of the curves and the magnitude of the concentration values contain the information that must be stored. For each sample in $\mathbf{B}^{\lambda}$, individual curves from each well are concatenated. The shape of $\mathbf{B}^{\lambda}$ becomes $(n \times \lambda \cdot k)$, with each row containing the raw predictor in physical space for each simulation. PCA can then be performed to obtain the principal components (PCs). Let $\beta = \min\,(N, \lambda \cdot k)$ denote the maximum possible number of PCs stored in the $(N \times \beta)$-matrix $\mathbf{D}_{\beta}$. The final pre-processing step for the predictor is to select an appropriate

PC number $1 \leq \delta < \beta$ to reduce the dimension of the predictor while preserving enough variance to adequately represent the original dataset. The final predictor training set is the $(N \times \delta)$-matrix

$$\mathbf{D}_\delta = \{\mathbf{d}_{i,j} | 1 \leq i \leq N; 1 \leq j \leq \delta\}\,. \tag{3.1}$$

PCA is only applied to the training set. The PCA coefficients are saved so that the observation set can be projected to the PC space later on during the prediction phase, assuming that the test set is consistent with the training set. The verification of this assumption falls out of the scope of this chapter. However, in field cases, the experimenter must check for inconsistency. For more information, see Scheidt et al. (2018) and Hermans et al. (2018).

Target The targets of this study are wellhead protection areas, which are determined by the 30-days endpoints of b backtracked particles that end up in the pumping well. For N forward model simulations, the raw output is the $(N \times b \times 2)$-matrix

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2 \ldots \mathbf{e}_{i-1}, \mathbf{e}_i | 1 \leq i \leq N\}\,, \tag{3.2}$$

$$\mathbf{e}_i = \left\{\{\mathrm{x}, \mathrm{y}\}_j | 1 \leq j \leq b\right\}, \tag{3.3}$$

with (x, y) the coordinate of a particle endpoint. The latter is actually three-dimensional, but because this work only considers a single layer, the depth dimension is ignored. As the number of particles b in the discretized space increases, so does the resolution of the two-dimensional WHPA delineation.

The pattern of a WHPA is controlled by the heterogeneities of the hydraulic conductivity field near the pumping well. The area of a polygon whose vertices are arranged in a specific order of the particle's endpoints defines it. The particle tracking software's output file (MODPATH 7, Pollock 2017) does not provide such a list of sorted coordinates.

As a result, we compute the WHPA polygon by solving the travelling salesman problem (TSP), a combinatorial optimisation problem (COP) aimed at finding the tour between particles such that each particle is visited only once and the total distance travelled is minimum (Diaby and Karwan, 2016; Gutin and Punnen, 2006). We used Google OR-Tools, an open-source COP software written in Python (Perron and Furnon, 2019).

Given an arbitrary starting particle endpoint, the algorithm provides the explicit WHPA border representation, which are the connected endpoints corresponding to the vertices of the closed curve. After TSP is applied to each simulation, WHPAs are then stored in the $(N \times b \times 2)$-matrix $\mathbf{E}'$, with the second dimension of $\mathbf{E}'$ sorted. To the best of the author's knowledge, this is the first time TSP has been used to determine WHPA delineation. A WHPA can also be represented by a variable φ over the grid such that $\varphi = 1$ for cells within the polygon defined by the TSP solution and $\varphi = 0$ for outer

cells, defining a binary matrix.

Our target is represented by this binary WHPA representation. PCA is not suitable for binary-valued data because it implicitly minimises the least-square error of the distance between data points and their projections in the PC subspace. To apply dimension reduction to the target, a suitable operator must be chosen to convert this Boolean matrix to a real-valued one, with smoothness being a desirable property. Yin et al. (2020) demonstrated how to successfully apply PCA to a two-dimensional discrete lithology model using the signed distance function (SDF) before dimension reduction. This approach is also adopted in this work. In two spatial dimensions, let $\mathcal{H}$ be the region of interest and $\mathrm{d}()$ a Euclidean distance function defined as

$$\mathrm{d}(\vec{\mathbf{p}}) = \min\left(|\vec{\mathbf{p}} - \vec{\mathbf{p}}_b|\right) \forall\, \vec{\mathbf{p}}_b \in \partial\mathcal{H}, \tag{3.4}$$

entailing that $\mathrm{d}(\vec{\mathbf{p}}) = 0$ on the boundary where $\vec{\mathbf{p}} \in \partial\mathcal{H}$. A SDF is an implicit function Ψ with $|\Psi(\vec{\mathbf{p}})| = \mathrm{d}(\vec{\mathbf{p}})$ for all points $\vec{\mathbf{p}} \in \mathbb{R}^2$. Hence, $\Psi(\vec{\mathbf{p}}) = \mathrm{d}(\vec{\mathbf{p}}) = 0 \forall\, \vec{\mathbf{p}} \in \partial\mathcal{H}$, $\Psi(\vec{\mathbf{p}}) = \mathrm{d}(\vec{\mathbf{p}}) \;\forall\, \vec{\mathbf{p}} \in \mathcal{H}^-$ (interior region) and $\Psi(\vec{\mathbf{p}}) = -\mathrm{d}(\vec{\mathbf{p}}) \;\forall\, \vec{\mathbf{p}} \in \mathcal{H}^+$(exterior region). In this case, the one-dimensional 0-contour of Ψ is chosen to represent the WHPA delineation that separates $\mathbb{R}^2$ into two separate subdomains with nonzero areas. An implicit representation means that the WHPA interface is represented as a one-dimensional isocontour of the higher-dimensional SDF, and not explicitly by all the points defining the curve. Signed distance functions have the property

$$|\nabla\Psi| = 1. \tag{3.5}$$

$$\text{with } \nabla\Psi = \left(\frac{\partial\Psi}{\partial \mathrm{x}}, \frac{\partial\Psi}{\partial \mathrm{y}}\right). \tag{3.6}$$

The gradient $\nabla\Psi$ is perpendicular to the isocontours of Ψ. $|\Psi(\vec{\mathbf{p}})|$ gives the shortest distance from all points $\vec{\mathbf{p}} \in \mathbb{R}^2$ to level sets of Ψ (Osher and Fedkiw, 2003). The Fast Marching Method (FMM), a scheme for solving the Eikonal equation (3.5) is a fast numerical method for approximating SDFs (Sethian, 1996). The FMM is implemented in the Python module scikit-fmm, which is limited to regular Cartesian grids. As a result, the implicit representation is discretized into a uniform grid, with the same approximation errors in both directions. Given l_x and l_y the total length of the computational grid in the x and y directions, respectively, and by selecting an appropriate dimension of the cells in both axes, $\Delta\mathrm{x}$ and $\Delta\mathrm{y}$ ($\Delta\mathrm{x}$=$\Delta\mathrm{y}$ in a uniform grid), the target matrix $\mathbf{V}$ has the shape$\left(n \times \frac{l_\mathrm{y}}{\Delta\mathrm{y}} \times \frac{l_\mathrm{x}}{\Delta\mathrm{x}}\right)$. Let $rows = \frac{l_y}{\Delta\mathrm{y}}$ and $columns = \frac{l_x}{\Delta\mathrm{x}}$.

The target is now represented as a closed surface, with the $(N \times rows \times columns)$-shaped binary matrix $\mathbf{V}$ clearly defining the interior and exterior regions. The signed distance (SD) algorithm takes each sample of $\mathbf{V}$ as an input array and computes their SD field.

After completion, individual samples are the real, smooth $(rows \times columns)$ SD images, with each pixel representing a single feature. To apply dimension reduction, features must be concatenated in the updated $(N \times (rows \cdot columns))$-matrix $\mathbf{V}'$ for each

sample. The transformed pixel values are in the PCs subspace after PCA and are stored in $\mathbf{H}_\rho$ of shape $(N \times \rho)$, with $\rho = \min(N; rows \cdot columns)$. The final step in effectively reducing the target dimensions is to select the appropriate PCs number $1 \leq \upsilon < \rho$ to keep in the $(n \times \upsilon)$ target training set matrix

$$\mathbf{H}_\upsilon = \{\mathbf{h}_{i,j} | 1 \leq i \leq N; 1 \leq j \leq \upsilon\}. \tag{3.7}$$

PCA is performed only on the training set, as in the predictor pre-processing step, and the computed PCA coefficients are used to project the test target to the PC space. It should be noted that this is only possible in the synthetic case because the true target would be unknown in reality.

Training

Following pre-processing, the model is trained with the CCA algorithm to establish a multivariate relationship between predictor $\mathbf{D}_\delta$ and target $\mathbf{H}_\upsilon$. The number of components η is set to $\min(\delta, \upsilon)$, which is the maximum number possible (Meloun and Militký, 2012). The resulting canonical variates (CVs) pairs are stored in the $(N \times \eta)$ matrices

$$\mathbf{D}_\eta^c = \left\{\mathbf{d}_{i,1}^c, \mathbf{d}_{i,2}^c \ldots \mathbf{d}_{i,\eta-1}^c, \mathbf{d}_{i,\eta}^c | 1 \leq i \leq N\right\} \tag{3.8}$$

$$\mathbf{H}_\eta^c = \left\{\mathbf{h}_{i,1}^c, \mathbf{h}_{i,2}^c \ldots \mathbf{h}_{i,\eta-1}^c, \mathbf{h}_{i,\eta}^c | 1 \leq i \leq N\right\}. \tag{3.9}$$

Normality is ensured in $\mathbf{D}_\eta^c$ and $\mathbf{H}_\eta^c$ by applying a Yeo-Johnson transform to each row (Yeo and Johnson, 2000).

Regression

Consider the predictor $\mathbf{d}_*$, a set of breakthrough curves not used in the training set. $\mathbf{d}_*$ is processed in the same way as the examples used to train the model, and the canonical weights computed during the training step are used to project the processed $\mathbf{d}_*$ to the canonical space.

This data is then used to infer $p(\mathbf{h}_*^c | \mathbf{d}_*^c)$, the posterior probability distribution in canonical space of the unknown target in light of the observed $\mathbf{d}_*^c$, according to Equation 2.11.

If a linear correlation exists between the first CV pairs, MG inference can be performed to estimate the posterior multivariate normal distribution $p(\mathbf{h}_*^c | \mathbf{d}_*^c)$, i.e., infer its mean vector $\mathbf{m}_\eta^c$ of size η (Equation 2.12) and the $(\eta \times \eta)$ positive-definite covariance matrix $\mathbf{C}_\eta^c$ (Equation 2.13).

Sampling

Samples are drawn from $p(\mathbf{h}_*^c | \mathbf{d}_*^c)$ to predict the distribution of the unknown WHPA given the trained regression model and the observed value. A sufficient number of sam-

ples ζ is chosen to allow proper graphical interpretation and uncertainty quantification. Each sample of the resulting $(\zeta \times \eta)$ matrix

$$\mathbf{H}^{c}_{*\eta} = \left\{ \tilde{\mathbf{h}}^{c}_{*i,j} | 1 \leq i \leq \zeta; 1 \leq j \leq \eta \right\} \tag{3.10}$$

is back-transformed and reshaped to the corresponding $(\zeta \times rows \times columns)$ SD field, to obtain

$$\mathbf{H}_{*SD} = \left\{ \tilde{\mathbf{h}}_{*i,j,k} | 1 \leq i \leq \zeta; 1 \leq j \leq rows; 1 \leq k \leq columns \right\}. \tag{3.11}$$

3.2.2 Experimental design

The observed data set is unknown during the design stage of an experiment. Data sources can be placed anywhere across the grid, and the actual data can have any value within the prior data space. In this section, we show how BEL can be used to quantify the amount of information delivered by each possible data source, allowing us to make the best choice for injection well locations. We only test a finite number of predefined well locations to reduce the computational burden, but the framework can be extended to all possible spatial configurations involving one or more wells on a given grid. For any chosen configuration, our method requires simulating the tracing experiments for each sample drawn from the prior model distribution.

One data-utility function must be maximised or minimised in order to identify highly informative data sets based on their spatial origin.

Informative data sets identification and data utility function

We estimate which data source yields the greatest reduction in uncertainty from the prior to the posterior to determine which data source is the most informative among a set of s possible locations. However, because WHPAs are high-dimensional, quantifying the uncertainty variation is not straightforward. A meaningful approach to assessing the uncertainty reduction for each of the s data sources, given ζ drawn samples, is to compare the true SD images with the generated ones for each test model, such that

$$\upsilon = \left\{ \sum\nolimits_{j=1}^{\zeta} \mathrm{H}\left(\mathbf{h}^{\eta}_{*}, \tilde{\mathbf{h}}_{*i,j}\right) | 1 \leq i \leq \lambda \right\} \tag{3.12}$$

is the vector of length λ whose entries are the sum of computed discrepancies over each posterior target sample for each data source and a given true (test) model or image. If $\mathbf{h}_*$ is the true SD image $(rows \times columns)$ corresponding to the predictor $\mathbf{d}_{*i}$, that is, the single breakthrough curve of tracer i, then $\mathbf{h}^{\eta}_{*}$ is the back-transformed version of $\mathbf{h}_*$ using η components. $\tilde{\mathbf{h}}_{*i,j}$ contains the ζ drawn samples using a unique data source i.

Each entry of $\tilde{\mathbf{h}}_{*i,j}$ has the same dimension as $\mathbf{h}_*$. $\mathrm{H}()$ is the operator used to compare the similarity of two images. The distribution of υ can then be examined to identify the data source index with the smallest or greatest value range, depending on

the measure of similarity H (), indicating the most informative source.

Dubuisson and Jain (1994) and Scheidt et al. (2018) recommend using the Modified Hausdorff Distance (MHD) as the data-utility function. It is especially well suited to the problem at hand because it involves measuring the distances between points forming the edges of the features of interest in the images. In this case, WHPAs are implicitly represented by the SD images' 0-contours edges, the coordinates of which can be extracted. Let $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{m \times w}$ be the pairwise Euclidean distance matrix between two vectors $\vec{\mathbf{p}}_1 \in \mathbb{R}^{m \times 2}$ and $\vec{\mathbf{p}}_2 \in \mathbb{R}^{w \times 2}$. The MHD between $\vec{\mathbf{p}}_1$ and $\vec{\mathbf{p}}_2$ is

$$\mathrm{MHD} = \max \left[\overline{\min_i \left[\mathcal{D}_{ij} \right]}, \overline{\min_j \left[\boldsymbol{\mathcal{D}}_{ij} \right]} \right] \in \mathbb{R}, \tag{3.13}$$

$$\min_i \left[\mathcal{D}_{ij} \right] \in \mathbb{R}^w, \tag{3.14}$$

$$\min_j \left[\mathcal{D}_{ij} \right] \in \mathbb{R}^m, \tag{3.15}$$

where $1 \leq i \leq m; 1 \leq j \leq w$ are the row and column index of $\boldsymbol{\mathcal{D}}$, respectively. The overbar denotes the mean operator. $\mathcal{D}_{ij}$ is the Euclidean distance between the i^{th} point of $\vec{\mathbf{p}}_1$ and the j^{th} point of $\vec{\mathbf{p}}_2$. The max operator is used in Equation 3.13 because $\boldsymbol{\mathcal{D}}$ is not symmetric. The MHD is robust and is monotonically increasing as the dissimilarity between two contours increases. Thus a smaller value of υ indicates the most similarity between images (Dubuisson and Jain, 1994).

Another candidate for the data-utility function is the Structural Similarity (SSIM) index, a metric that measures the similarity between two continuous images (Wang et al., 2004). It is symmetric ($\mathrm{SSIM}\left(\mathrm{im}_1, \mathrm{im}_2\right) = \mathrm{SSIM}\left(\mathrm{im}_2, \mathrm{im}_1\right)$) and bounded from 0 to 1 ($\mathrm{SSIM}\left(\mathrm{im}_1, \mathrm{im}_2\right) = 1$ if $\mathrm{im}_1 = \mathrm{im}_2$) for two images $\mathrm{im}_1, \mathrm{im}_2$ of the same region in space. This index can be computed directly from the SD images of the reference (true) target and the sampled ones. It takes the form

$$\mathrm{SSIM}\left(\mathrm{im}_1, \mathrm{im}_2\right) = \mathrm{F}(\mathrm{L}\left(\mathrm{im}_1, \mathrm{im}_2\right), \mathrm{C}\left(\mathrm{im}_1, \mathrm{im}_2\right), \mathrm{S}\left(\mathrm{im}_1, \mathrm{im}_2\right)). \tag{3.16}$$

The three terms $\mathrm{L}, \mathrm{C}, \mathrm{S}$ are the luminance, contrast, and structures components. F is the combination function. For more details, see Wang et al. (2004).

Equation 3.12 provides the uncertainty range for a single observation set, i.e., a single test example. In experimental design, the actual data is unknown and could take any value in the prior range. Assumptions on the prior model distribution are made by assigning the N hydraulic conductivity $\mathbf{K}$ fields to the computational grid. Thus a single performance measure is not enough to define the value of information of each injection well. In machine learning, it is common practice to use 80% of the total dataset for training the remaining 20% as a test set (Géron, 2022). Therefore, we combine the uncertainty of each data source for a set of observations, such that

$$\boldsymbol{\Upsilon} = \{\upsilon_i | 1 \leq i \leq N_{20\%}\} \tag{3.17}$$

is the matrix of shape $(\lambda \times N_{20\%})$ containing the summed MHD of each data source for each member of the test set of size $N_{20\%}$. In order to fully exploit the available dataset and to assess the sensitivity of the results to the chosen training and test sets, k-fold validation is performed. The k-fold procedure randomly splits the available dataset into k distinct subsets called folds. The technique then selects a new fold for testing k-times and trains with the remaining $k-1$ folds in sequential order (Géron, 2022). Finally, a total of k $\boldsymbol{\Upsilon}$ matrices, one for each k-fold combination, are obtained.

We then determine the most informative data sources by examining the statistical summary of their distribution, which is represented by boxplots. The interquartile range (IQR) and/or the median for each data source between the k $\boldsymbol{\Upsilon}$ matrices may be different, i.e., inconsistent with one another, indicating that the dataset utilised is too small. This assessment is mostly a visual task dependent on the researcher; however, one may automate this process by using, for example, the mode of each distribution, the interquartile range, or a combination of criteria as a proxy to evaluate informative data sources.

In contrast to other experimental design approaches, the advantage of BEL is that the training step allows for the fast calculation of the posterior distribution for any observed data. As a result, solving the optimal design problem is as straightforward as running the forward model for the training set and evaluating the information content of various sources with minimal additional computational costs. It also allows for the unrestricted definition of several data utility functions based on posterior prediction ranges. As a result, we think BEL is especially well suited to experimental design.

3.3 Application

We compute the WHPA from tracing experiments in a single-layer aquifer. Following BEL, we sample the prior distribution of the model parameters by generating N models $\{\mathbf{M}_1, \mathbf{M}_2 \dots \mathbf{M}_{N-1}, \mathbf{M}_N\}$ and simulating both the WHPA using particle backtracking and BCs using solute transport modelling. As a result, we solve N solute transport problems as well as N backtracking problems.

A two-dimensional grid is used for the experiments and the MODFLOW-2005 code is used for groundwater flow modelling (Harbaugh, 2005). The entire research code is written in Python and includes MODFLOW-2005, MT3DMS-USGS, and MODPATH 7 support (Bakker et al., 2016).

A single pair of forward modelling (transport and backtracking) took 37 minutes to compute on a 2.3 GHz 8-Core Intel Core i9 processor.

3.3.1 Groundwater model

The different steps of the experiment (flow, transport, and particle tracking) depend on the unknown hydraulic conductivity **K** field. Using SGEMS (Remy et al., 2009), N samples from the prior distribution are generated by Sequential Gaussian Simulation (SGS; Goovaerts 1997). The spatial correlation of the $\log_{10}\mathbf{K}$ field is defined by a variogram model (see Table 3.1). Initially, only the mean value of the $\log_{10}\mathbf{K}$ field is considered unknown, but we later extend the approach to include structural uncertainty via the variogram parameters (e.g., Hermans et al., 2018, 2019).

Parameter	**Value**
Grid x-extent (m)	1500
Grid y-extent (m)	1000
Grid z-extent (m)	10
n_{row}	157
n_{col}	207
n_{lay}	1
SP_1 (days, steady state)	1
SP_2 (days, transient)	0.08
SP_3 (days, transient)	100
Pumping well rate $\left(\frac{m^3}{d}\right)$ [SP_1, SP_2, SP_3]	-1000
Injection wells rate $\left(\frac{m^3}{d}\right)$ [SP_2]	24
Tracers mass loading $\left(\frac{kg}{d}\right)$	1.5
S_s (m^{-1})	10^{-4}
S_y (-)	0.25
α_L (m)	3
K mean $\left(\frac{m}{d}\right)$	[25, 100]
$\log_{10}$ K standard deviation $\left(\frac{m}{d}\right)$	0.4
Kriging type	Simple
Nugget effect (-)	0
Structure	Spherical
Range (m) [min, max]	[25, 100]

Table 3.1: Note. Model parameters.

A number s of injection wells are placed around a pumping well. Their role is to inject individual tracers to model their transport and record their breakthrough curves at the pumping well location. Several particles are artificially placed around the pumping well and their origin after a given amount of time is backtracked to delineate the corresponding WHPA. To ensure that the tracers flow to the pumping well, the hydraulic conductivity values at their location are specified as hard data in the SGS, with their values equal to the maximum value of the prior distribution, to which a small random number between 0 and 1 is added. Otherwise, the tracers may become stuck in a low-conductivity zone directly around the pumping well. It should be noted that in a

real-world scenario, the hydraulic conductivity field around a pumping well is expected to be relatively high because it allows for water pumping, so this assumption is reasonable and has no effect on the experimental design results.

The x, y, z axes of the structured grid have dimensions of $1500\,m$, $1000\,m$, and $10\,m$, respectively. The base cell dimension in the x, y axes is $10\,m$, incrementally refined around the pumping well $(1000\,m, 500\,m)$, down to $1\,m$ by $1\,m$ cells. The pumping rate is $1000\,\frac{m^3}{d}$. The dimensions $(n_{row}, n_{col}, n_{lay})$ of the grid are (157, 207, 1). The total flow simulation period is discretized into 3 stress-periods (SP_1, SP_2, SP_3) during which pumping occurs at all time steps. SP_1 is steady-state to compute the hydraulic heads in pumping conditions. Injection occurs during the transient SP_2 which is subdivided into 300 time steps of 24 seconds each. SP_3 is transient, 100 days long, and is discretized into 100 time steps to model tracer transport from the injection wells and record their BCs at the pumping well location. The flow boundary conditions are a fixed head of $0\,m$ along the western boundary and a fixed head of $-3\,m$ along the eastern boundary. The North and South boundaries correspond to no-flow boundary conditions. They are kept constant across all stress periods to create a negative gradient in the x-direction. It should be noted that the framework could also include boundary conditions uncertainty (Hermans et al., 2019, 2018). Six injection wells are placed around the pumping well (Figure 3.1A).

They inject individual tracers with mass loading of $1.5\,\frac{kg}{d}$, at the rate of $24\frac{m^3}{d}$ for two hours (SP_2).

This aquifer layer is modelled by a confined aquifer model. It has constant specific storage $\mathbf{S_s}$ and a constant specific yield $\mathbf{S_y}$ of $10^{-4}m^{-1}$ and 0.25 (no units), respectively. To solve for advection in transport modelling, the hybrid method of characteristics (HMOC) is used. The porosity is set to the $\mathbf{S_y}$ field and the longitudinal dispersivity $\alpha_\mathbf{L}$ is set to $3\,m$. The tracers used in this chapter are synthetic conservative tracers introduced in the model domain at various wells and following the flow field. The tracers' characteristics and behavior are not described in detail because they are not required for the purpose of this example, which is only meant to demonstrate the capabilities of the Bayesian experimental design framework. To reduce computational costs, the active zone for transport simulations is reduced to about $\frac{1}{4}$ of the total model area, which includes the pumping and injection wells.

To define the WHPA in pumping conditions with particle tracking, 144 particles are placed around the pumping well, and their backward endpoints are computed after 30 days. The porosity parameter of the particle-tracking algorithm is set to 0.25. The preceding parameterization is held constant.

The $\log_{10}\mathbf{K}$ mean is randomly varied between 1.4 and 2, so that the $\mathbf{K}$ mean is ranging from 25 to $100\,\frac{m}{d}$, which are values commonly observed in alluvial aquifers (Dassargues, 2018). The $\log_{10}\mathbf{K}$ standard deviation is set to $0.4\,\frac{m}{d}$ and remains constant. To ensure pumping and tracer injection, sufficiently high $\mathbf{K}$ values are fixed at all well locations, randomly chosen between 100 and $1000\,\frac{m}{d}$. The output of such a simulation is

illustrated in Figure 3.1A, and the resulting computed hydraulic heads in Figure 3.1B.

3.3.2 WHPA prediction

This section illustrates how to predict a single WHPA using all sources of information, i.e., the 6 injection wells. We demonstrate that BEL can efficiently estimate the posterior distribution of the target WHPA. The total size of the dataset in this section is $N = 500$. The size of the training set is then $N_{80\%} = 400$, and the remaining models are used to validate BEL's ability to predict the target. Because the prediction is much simpler than the underlying model, such a small training size is sufficient. Previous BEL applications have shown that this order of magnitude is adequate for making accurate predictions (Athens and Caers, 2019; Hermans et al., 2019, 2018, 2016; Michel et al., 2020a,b; Park and Caers, 2020; Yin et al., 2020). The influence of the size of the training set is discussed in Section §3.3.2. For demonstration purposes, 4 examples will be picked out of the remaining 100 samples to demonstrate the BEL framework's prediction capabilities. Only one of them, however, will be used to demonstrate the various steps.

Pre-processing

Predictor First, the BCs are interpolated through 200 predefined equidistant time-steps, significantly reducing the dimensionality of the predictor (Figure 3.2). The 6 curves from each set are then concatenated into the predictor matrix of shape $(N_{80\%} \times \lambda \cdot k) = (400 \times 1200)$ and PCA is applied to obtain the square PC matrix $\mathbf{D}_{400}$. To retain a sufficient amount of information while reducing the original dataset dimension, an adequate number of PCs must be chosen to truncate the columns of $\mathbf{D}_{400}$. Setting δ to 50 allows explaining 99.87% of the variance of the data while reducing its size by a factor of 4. The final product of the predictors pre-processing is the (400×50)-$\mathbf{D}_\delta$ matrix 3.1. The remaining 0.13% of the variance is not considered useful for predicting the target. Figure 3.3A illustrates the full predictor for the chosen example (6 concatenated curves) and its reconstruction using the 50 first PCs. The back-transformation effectively recovers the original structure while slightly smoothing out the curves.

The raw BCs of the chosen example have a size of (6×18444) (there are 18444 time steps for each curve) and the dimensionality of the original data is thus reduced by a factor of 369 by the joint interpolation-PCA pre-processing while preserving most of the information. The range of values of the model's coefficients narrows as the number of PC increases (Figures 3.3C-D). This reflects the fact that the coefficients store less information about the curves, as shown in Figures 3.3E-F which show the cumulative explained variance as the number of PCs increases. In Figure 3.3C, the PCs of the test predictor are compared to those of the training set, illustrating that this test predictor is consistent with the prior.

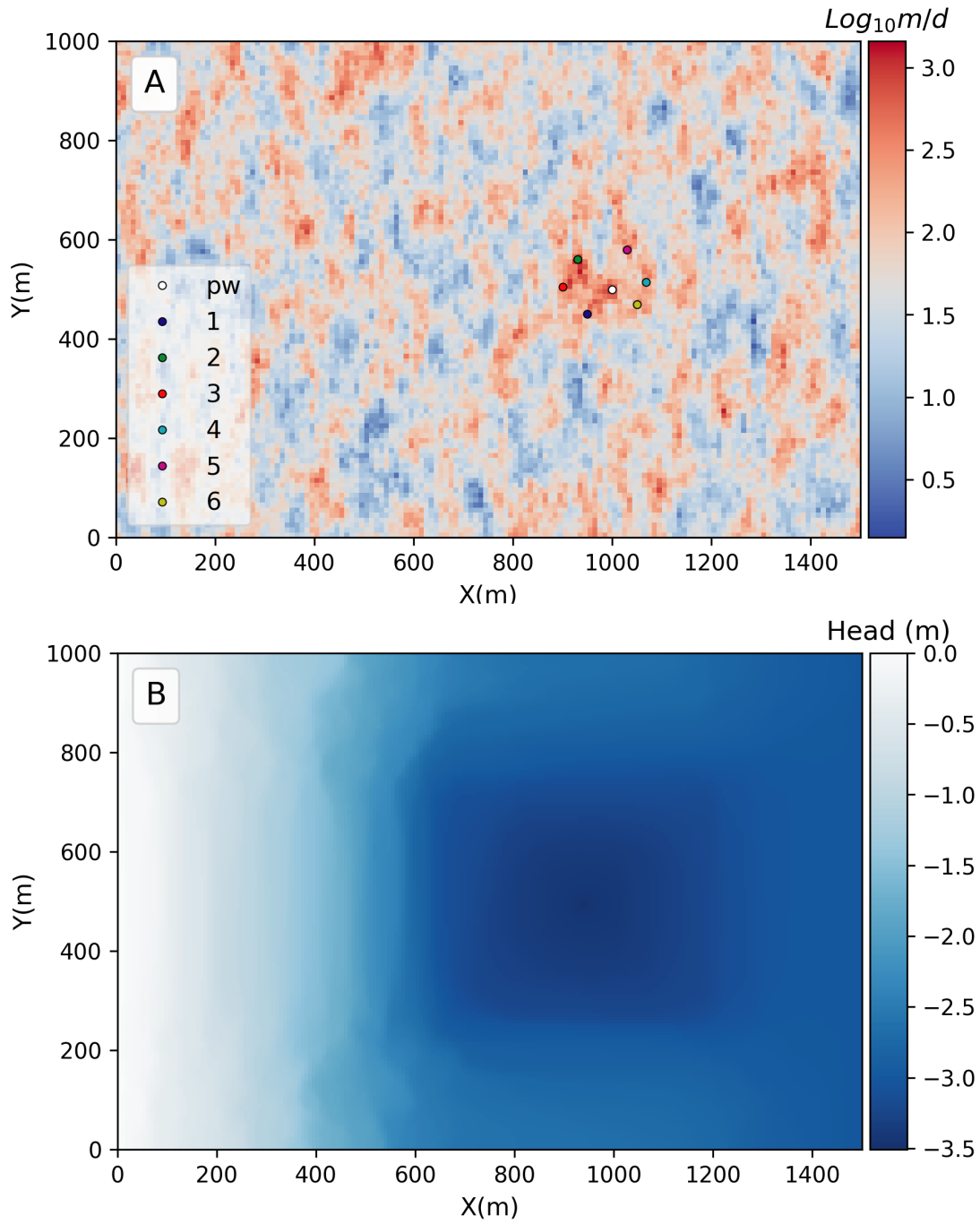

Figure 3.1: **A.** Hydraulic conductivity field in 10-logarithmic base. The pumping well (pw) is located at (x, y) = (1000m, 500m) and is surrounded by 6 injection wells. **B.** Flow solution. The direction of the natural gradient is from West to East.

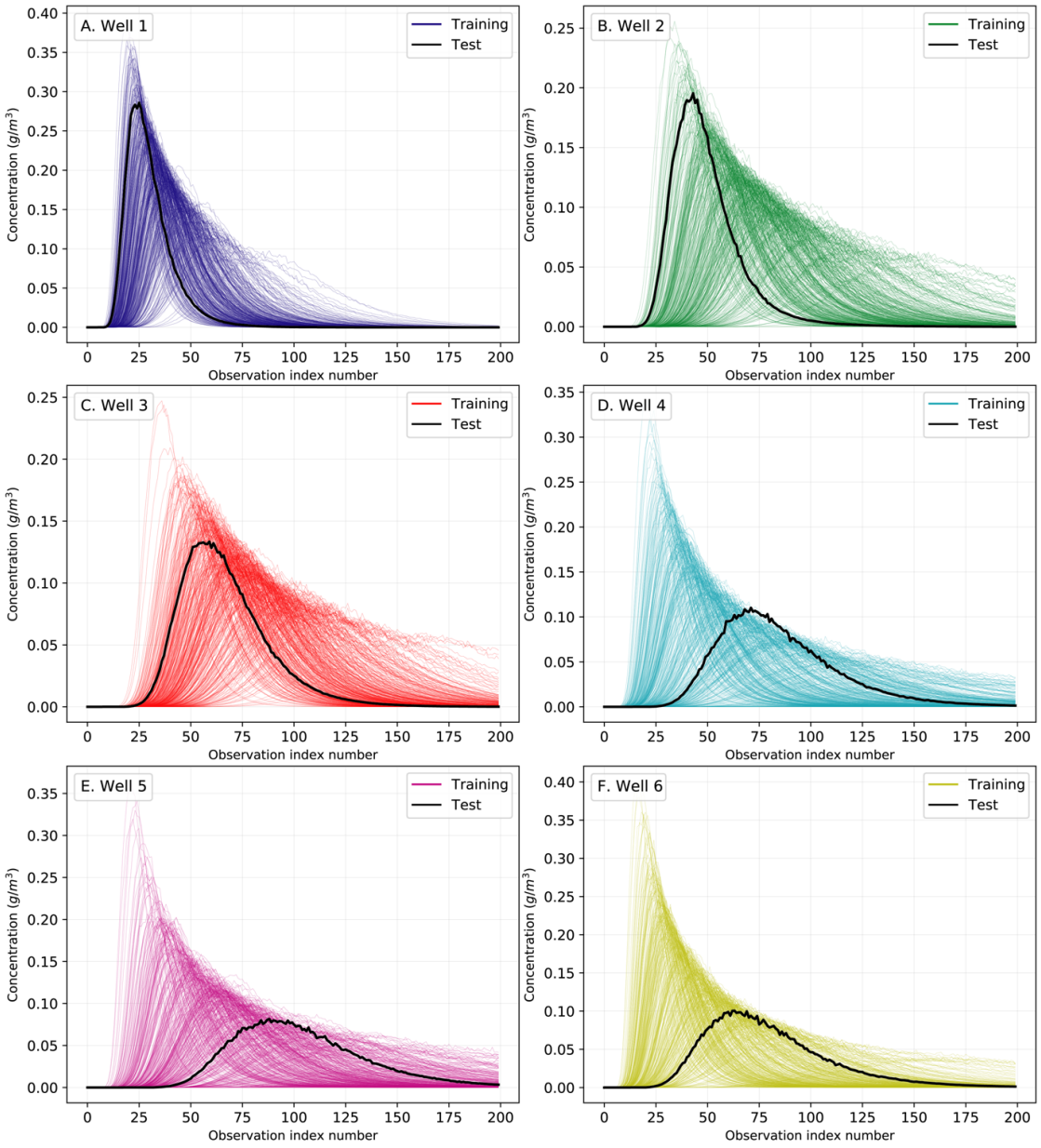

Figure 3.2: Breakthrough curves of tracers from each injection well, for both **training** and **test** (single sample) sets. They are all discretized and interpolated in 200 steps.

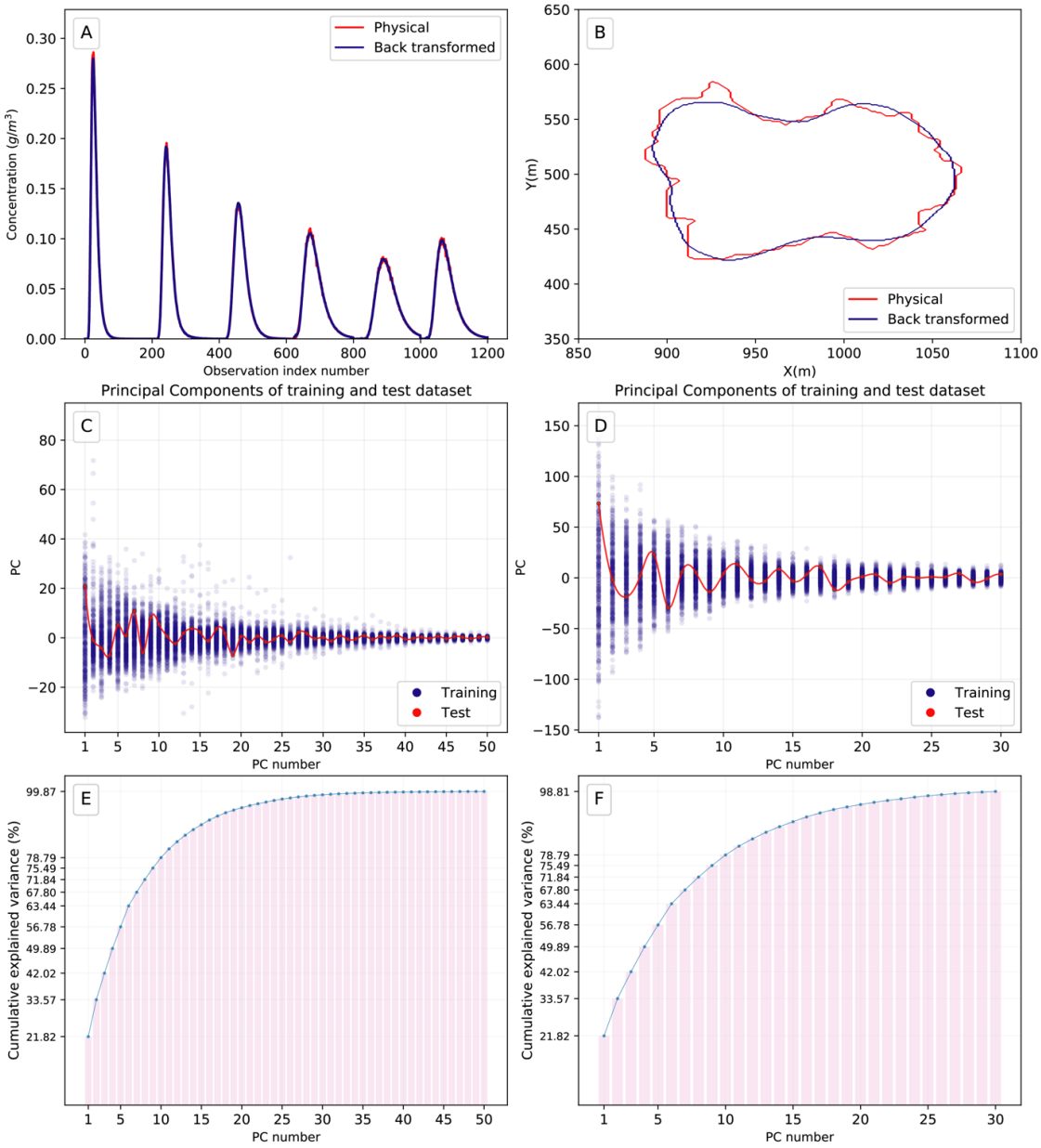

Figure 3.3: **A.** The full predictor of the **test** set is the concatenation of all breakthrough curves (black curves in Figure 3.2). PCA with 50 PCs allows to recover the original curves while smoothing out noise present in the original dataset. **B.** Zero-contour of the test target's original SD compared to the zero-contour of the test target's projected SD, which was then back-transformed with its 30 PCs. **C.** Predictor training set and projected test set PCs. **D.** PCs of the target training set's SD and the projected test set's SD. **E.** Cumulative explained variance for the PCs of the predictor **training** set. **F.** Cumulative explained variance for the PCs of the SD of the target **training** set.

Target The raw output of the backtracking simulation for the randomly chosen test example is shown in Figure 3.4A. The 144 two-dimensional coordinates are scattered around their origin, which is the location of the pumping well. A few points are chosen at random, and their index in the output file is displayed. The SD algorithm is then used to convert each item to an image. Because endpoints have limited travel extents, a focused area is defined to reduce the computational cost of SD estimation and storage using $(\mathrm{x}_{\min}, \mathrm{x}_{\max}) = (800\,m, 1150\,m)$ and $(\mathrm{y}_{\min}, \mathrm{y}_{\max}) = (300\,m, 700\,m)$. This subdomain is subdivided into $4 \cdot 4\,m^2$ cells, resulting in 100 rows and 87 columns for a total of 8700 cells. The cell dimensions have been adjusted to keep enough information on the WHPA without storing an image that is too large. The WHPA delineations are now represented by the zero-contour of the binary field. This Boolean matrix is then fed in to the SD algorithm, which computes the SD value of this isocontour for each cell in the domain, as illustrated in Figure 3.4B for the chosen test example. Figure 3.4C shows the 0-contours of the training and test WHPAs SD. Each SD image is then flattened and PCA is applied on the resulting (400×8700)-matrix. The number of PCs, $\upsilon = 30$, is chosen based on the reconstruction of the WHPA image as illustrated in Figure B.

Because the WHPA is so complex due to the **K** distribution, attempting to predict all of the nooks that cannot be captured by the tracing experiment would be futile. Instead, the slightly smoothed-out reconstructed WHPA is a good candidate for prediction. The resulting target is the (400×30)-$\mathbf{H}_\upsilon$ matrix 3.7 as a result of these choices. The content of $\mathbf{H}_\upsilon$ is displayed in Figure 3.3D for the training and test set, and the cumulative explained variance in Figure 3.3F, reaching the amount of 98.81% for the last PC. Note that the remaining 1.19% that are not predicted could be added as random noise after prediction (Park and Caers, 2020). However, since the final purpose is experimental design, there is no interest in adding the same uncertainty component to all samples.

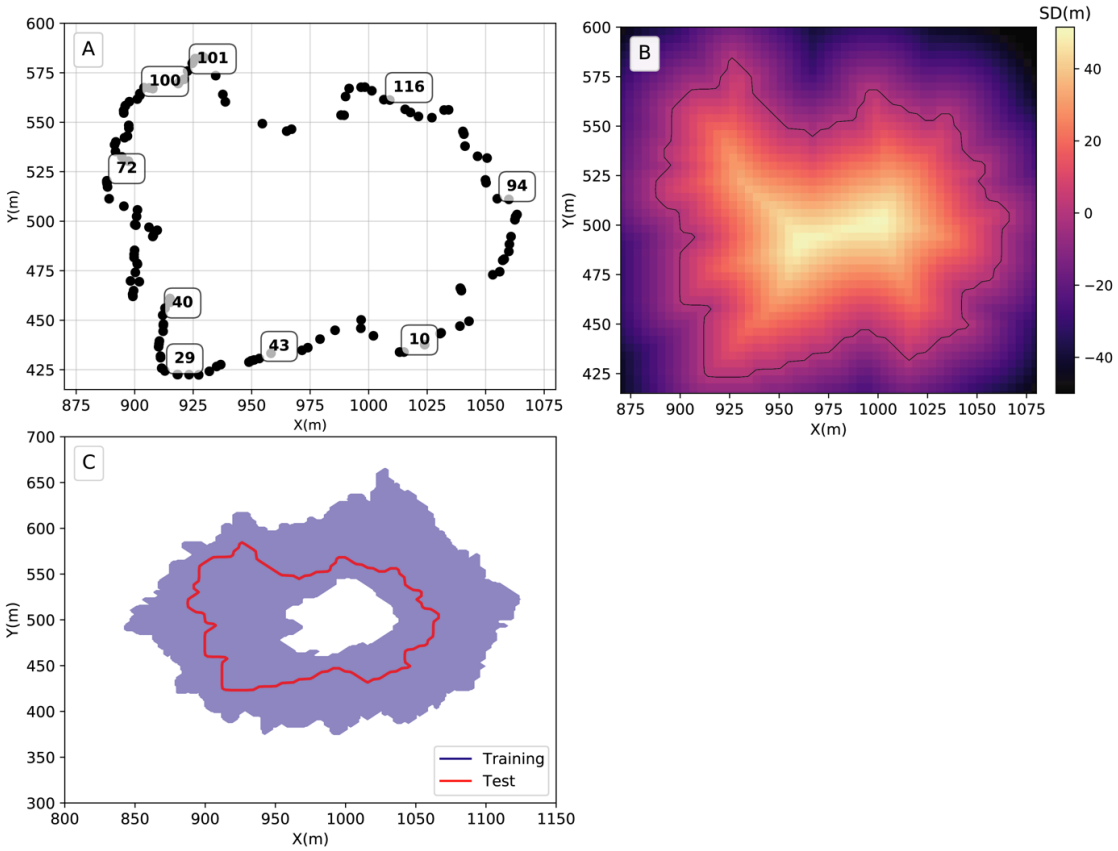

Figure 3.4: **A.** The chosen test target is in its raw form. To illustrate the meaningless ordering of the endpoints output, a few point indexes are randomly highlighted among the total of 144 particles. **B.** The test WHPA is implicitly represented on a discretized grid. The WHPA delineation corresponds to the SD field's 0-contour, which is computed for each cell as the closest distance from its centre to the boundary. **C.** Target **training** set and chosen **test** example.

Training

CCA is applied between $\mathbf{D}_{50}$ and $\mathbf{H}_{30}$. As a result, the maximum number of CVs is 30. The resulting Canonical Pairs matrices are $\mathbf{D}_{30}^{c}$ (matrix 3.8) and $\mathbf{H}_{30}^{c}$ (matrix 3.9). The first three CV pairs are illustrated in Figures 3.5A-C, showing high linear correlations whose strengths are reflected by the Canonical Correlation Coefficient r (Meloun and Militký, 2012), whose decrease with the number of CVs is shown in Figure 3.5D. For each pair, KDE is performed, and its density plotted behind the CVs point cloud. It is done here solely for visualisation of the joint probability distributions, but it could also be used for sampling as an alternative to linear regression in the case of non-linearity or non-Gaussianity (Michel et al., 2020b), as shown by the y marginal plots of Figure 3.5A-C.

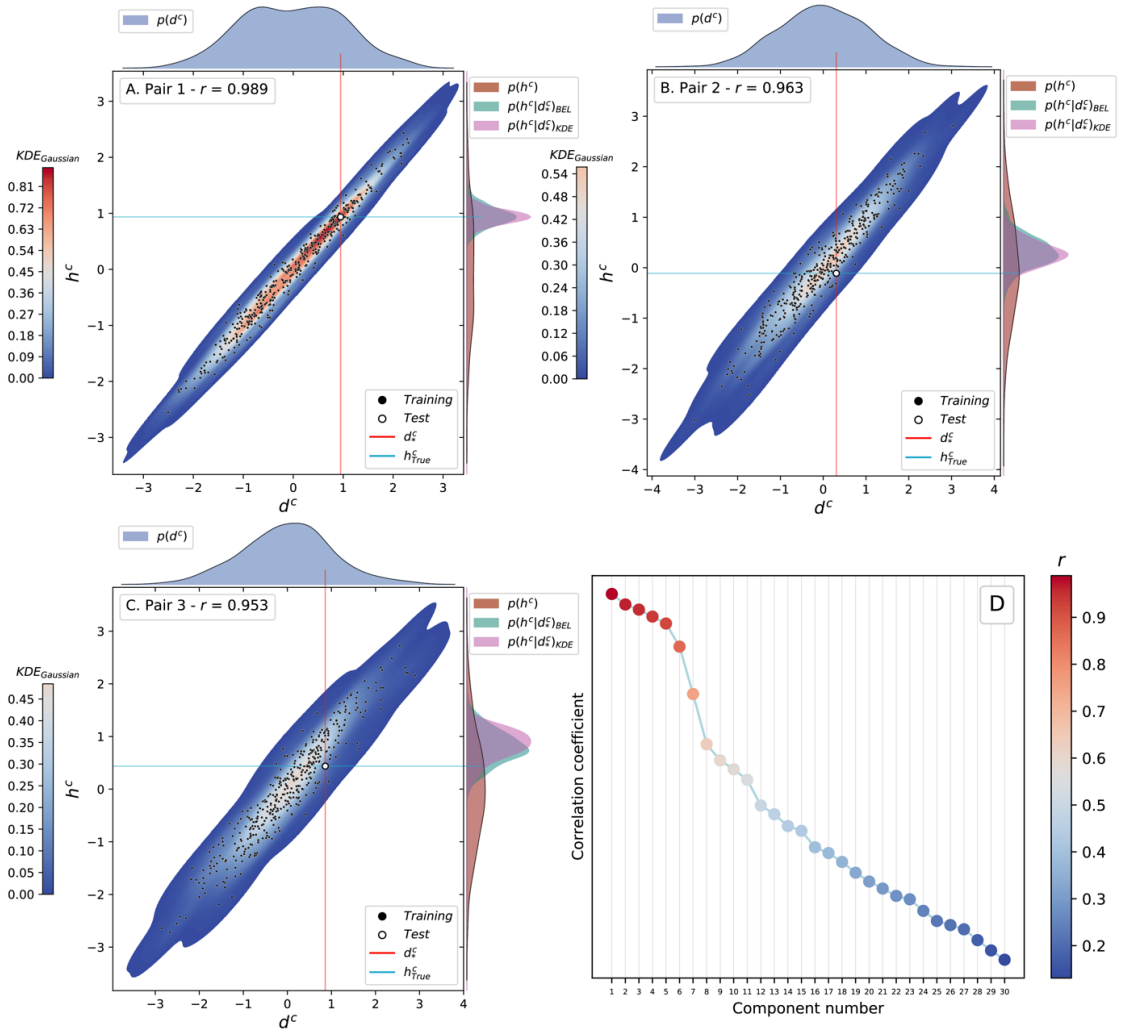

Figure 3.5: **A, B, C.** Canonical variates bivariate distribution plots for the 4 first pairs of the **training** set, and the canonical space projection of the selected **test** predictor and associated **test** target (see notches). The posterior distribution of $\boldsymbol{h}^c$ computed according to BEL and KDE can be compared on the y marginal plot. **D.** Decrease of the canonical correlation coefficient r with the number of CV pairs for the **training** set.

Regression

The test predictor depicted in Figure 3.3C is projected to the canonical space as illustrated in Figures 3.5A-C. MG inference is then applied to infer the posterior covariance (Equation 2.13) and mean (Equation 2.14).

Sampling

As explained in Section §3.2, samples are drawn from the MG described by its two first moments and back-transformed to the original space. Figure 3.6A depicts 400 WHPAs

sampled from the inferred posterior distribution of the chosen example. The 400 samples englobe the true prediction, and the uncertainty reduction is visible by comparing them with the footprint of the prior prediction range. The uncertainty reduction around injection wells is visible as knots where the WHPA distribution tightens. The methodology output is illustrated for three other (randomly chosen) test examples (Figures 3.6B-D), using the same training set. In each case, BEL successfully predicts the WHPA with its uncertainty range. Of course, given the limited amount of data, the range of uncertainty in the posterior remains important.

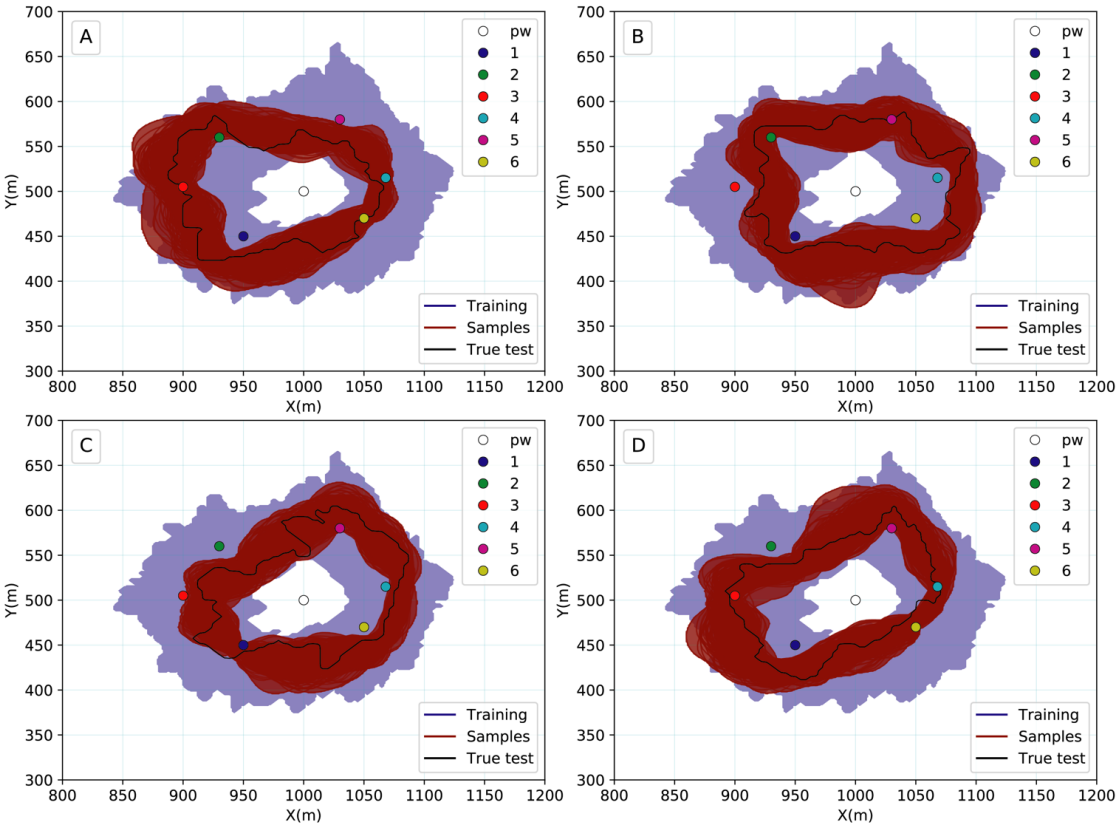

Figure 3.6: BEL-derived posterior predictions for four different tests WHPAs. Subfigure **A** displays the chosen test example associated with Figures 3.3–3.5.

Influence of the training set size

The size of the training set is an essential component of the learning process (Michel et al., 2020b). A sufficient number of samples must be used to capture the target's variability. Figure 3.7 shows how we validated our decision to use a training set of 400 models to compute the posterior. The graph depicts how the average SSIM index values change with different training set sizes ranging from 125 to 900, each calculated with 400 samples from the posterior distribution of randomly chosen targets (each line corresponding to a different WHPA). For visualisation purposes, only 20 different targets are displayed. It demonstrates that the SSIM index varies only slightly (notice the scale

for the y-axis). The average metric value begins to stabilise from the 400 training size mark. Because the SSIM index is calculated from a small number of samples in the posterior, some minor variations are still possible. As a result, a training set of this size is adequate for predicting the posterior distribution of the target for a wide range of different WHPAs. The risk of not learning the relationship well enough through BEL is reduced by increasing the number of sampled models in the prior. We can begin with a small number of prior models and progressively increase the number until the predicted range is stable. This number is determined primarily by the target's complexity (Hermans et al., 2018; Michel et al., 2020b; Scheidt et al., 2018).

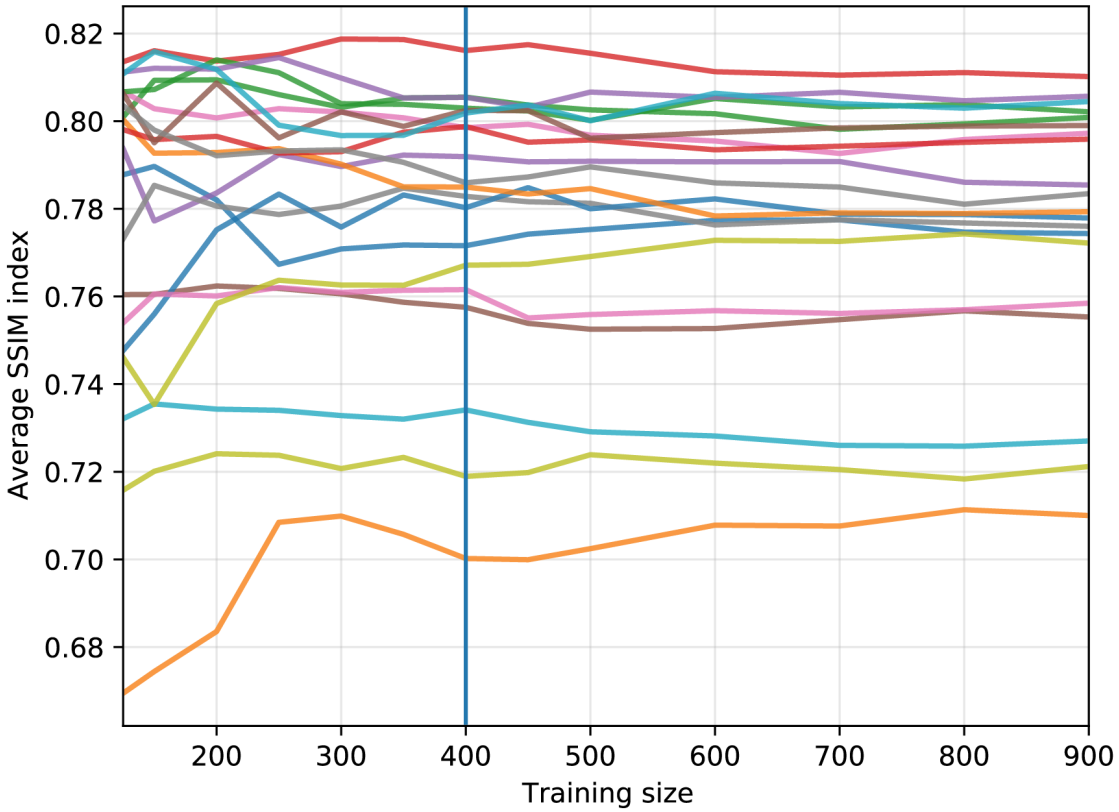

Figure 3.7: The impact of training set size (125–900) on the average SSIM index for 20 different targets. Each line represents a single target that is being predicted using training sets of increasing size. At around $N_{training} = 400$, the average metric value stabilises. For identical images, the SSIM index reaches its maximum value of 1.

3.4 Experimental design

3.4.1 Most informative well

In this section, the optimum location of a single well is estimated by making a single prediction on each source, i.e., $\lambda = 1$ with subsets [1], [2], [3], [4], [5], [6]. The

same number of training samples as in the regression case is used (Section §3.2.1); $N_{training} = N_{80\%} = 400$ and $N_{test} = N_{20\%} = 100$.

Figures 3.8A-F illustrate the effect of using a single data source at a time to predict the example WHPA. As expected, the uncertainty is reduced around the sources and the drawn models encompass the training set in further regions of the grid. In order to find out which injection well's location is the most informative, sampling is carried out in the same fashion as for the $N_{20\%} = 100$ test set (not shown). Each of them has 400 samples drawn for each well, and the MHD and SSIM index between drawn and true target are computed. The MHD and SSIM index of each sample are summed over the 100 test examples and the resulting boxplots are shown in Figures 3.9A-B, respectively. The chosen convention is that a higher metric value indicates a greater distance from the true target. To enforce this, the opposite of the SSIM index value is used. The MHD and SSIM index metrics are standardised by removing their mean and scaling to unit variance. Both metrics then yield very similar results: from Figures 3.9A-B, it appears that, globally, wells downstream the pumping well, i.e., tracers going against the natural gradient (Figure 3.1B) contain a greater amount of information (wells 4, 5, and 6), as indicated by their IQR bounds lower in magnitude than the IQR of the upstream wells 1, 2 and 3. The boxplot of the single test example considered previously is shown as an example. Figures 3.9C-D show that well 5 is the most informative in this case. This demonstrates that the results of a single data set do not allow for identifying the most informative well globally.

In order to validate the conclusions made on the most informative wells as inferred from Figures 3.9A-B, k-fold cross-validation is performed on the entire set of $N = 500$ considered samples. A total of five splits are chosen, successively dividing the set of samples into 400 training and 100 test samples. Figure 3.10 shows that in the given case, a set of 100 test samples is insufficient for performing experimental design. Indeed, the boxplots of the 5 different splits from Figures 3.10A-E are inconsistent with one another. Thus, the 5-fold cross-validation procedure is repeated with larger datasets until the boxplots begin to be consistent across splits. It is shown in Figure 3.11 that a set of size $N = 1250$ (successively split into 1000 training samples and 250 test samples) produces consistent metrics across all 5 splits and validates the interpretation that injection wells numbers 4, 5, and 6 are the most informative data sources (downstream wells). Upstream wells are consistently ranked lower in terms of information content, with well number 1 being the worst in all cases. On the other hand, the most informative data source can be deduced as well number 6, constantly showing low-bounds and narrow IQR, as opposed to well 4 showing broad IQR bounds in the 4th split (Figure 3.11D). Therefore, in this case, we find that the test dataset should contain at least 250 samples for consistent experimental design for WHPA prediction.

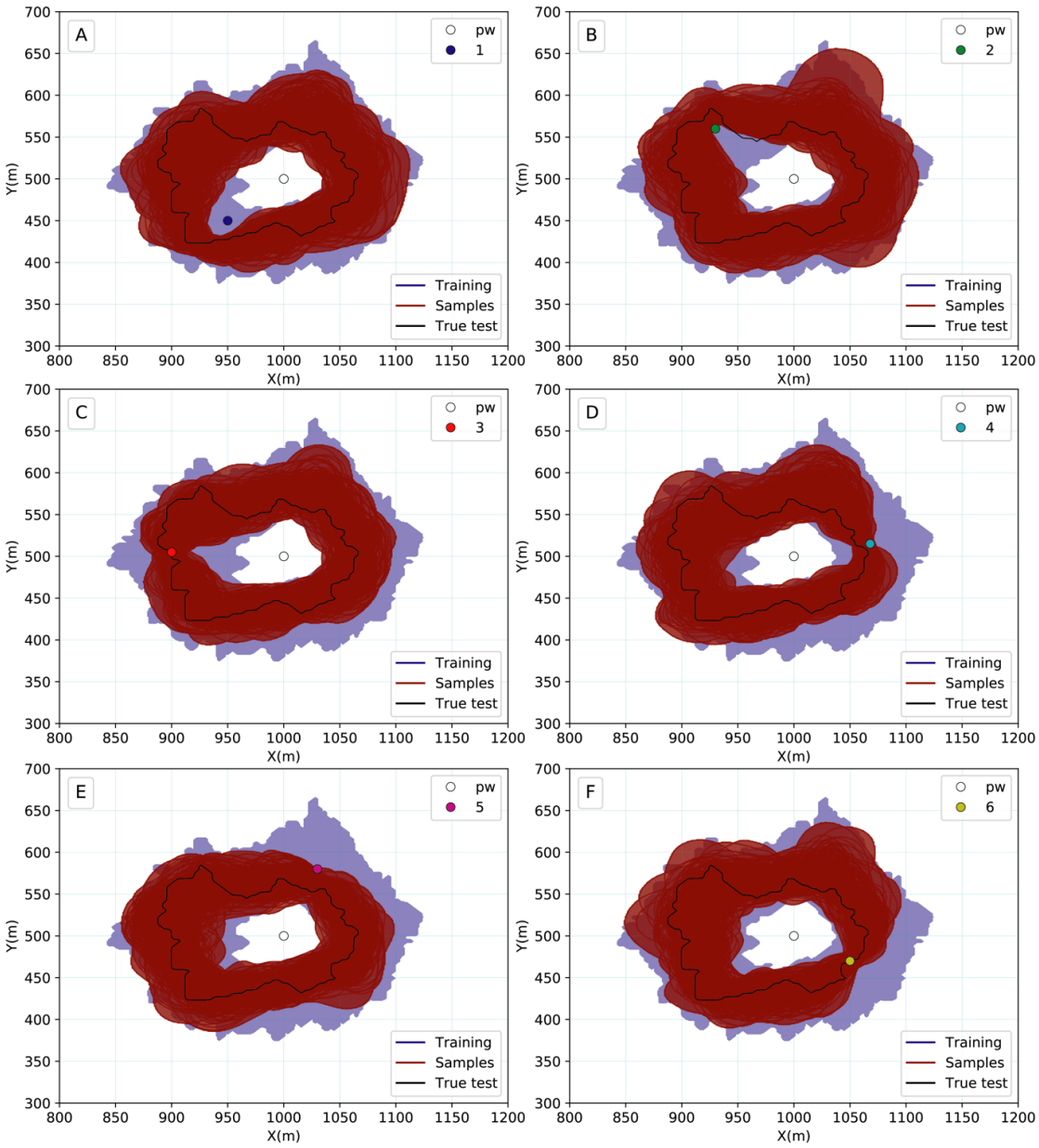

Figure 3.8: WHPA predictions for each well 1, 2, 3, 4, 5, 6 for the chosen test example.

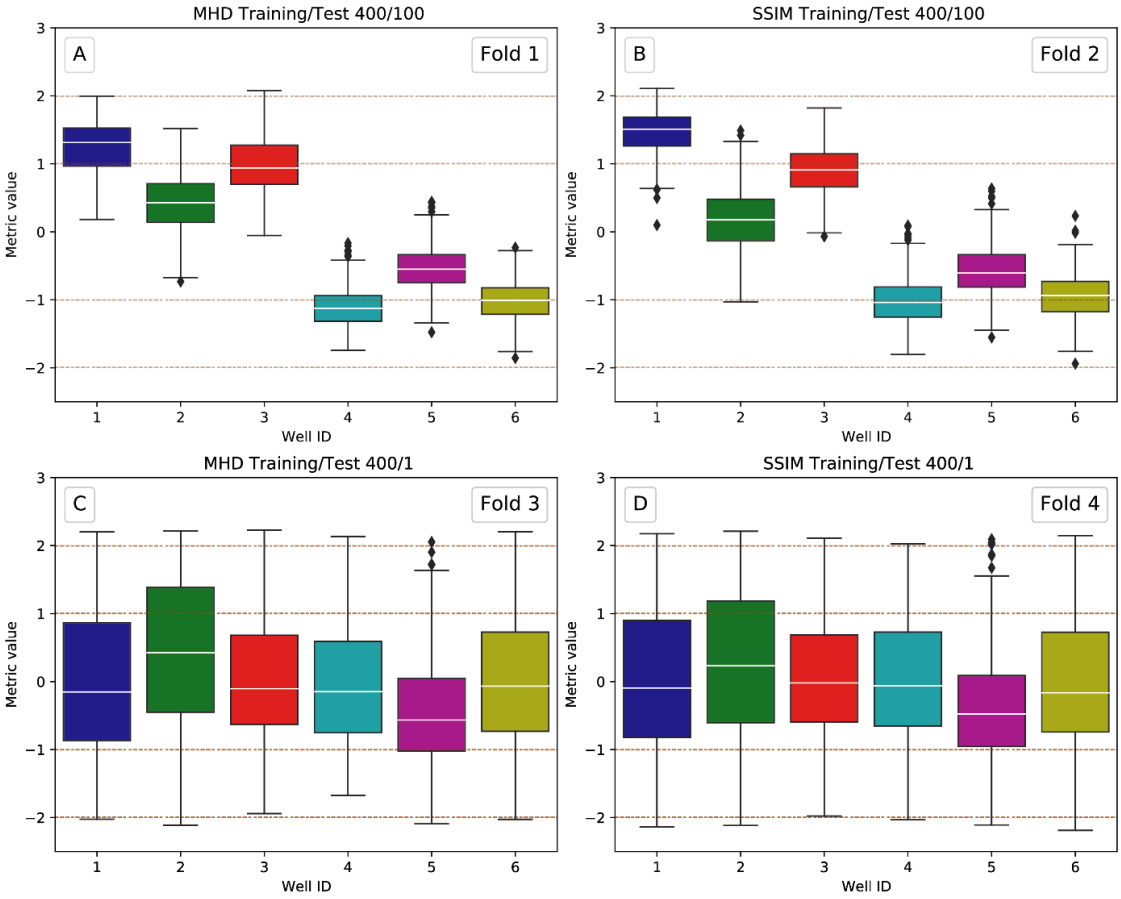

Figure 3.9: **A.** Boxplots of the standardised MHD distance for each well and 100 test samples. **B.** Boxplots of the standardised SSIM index for each well and 100 test samples. **C.** Boxplots of the standardised MHD distance for each well and the single test example. **D.** Boxplots of the standardised SSIM index for each well and the single test sample.

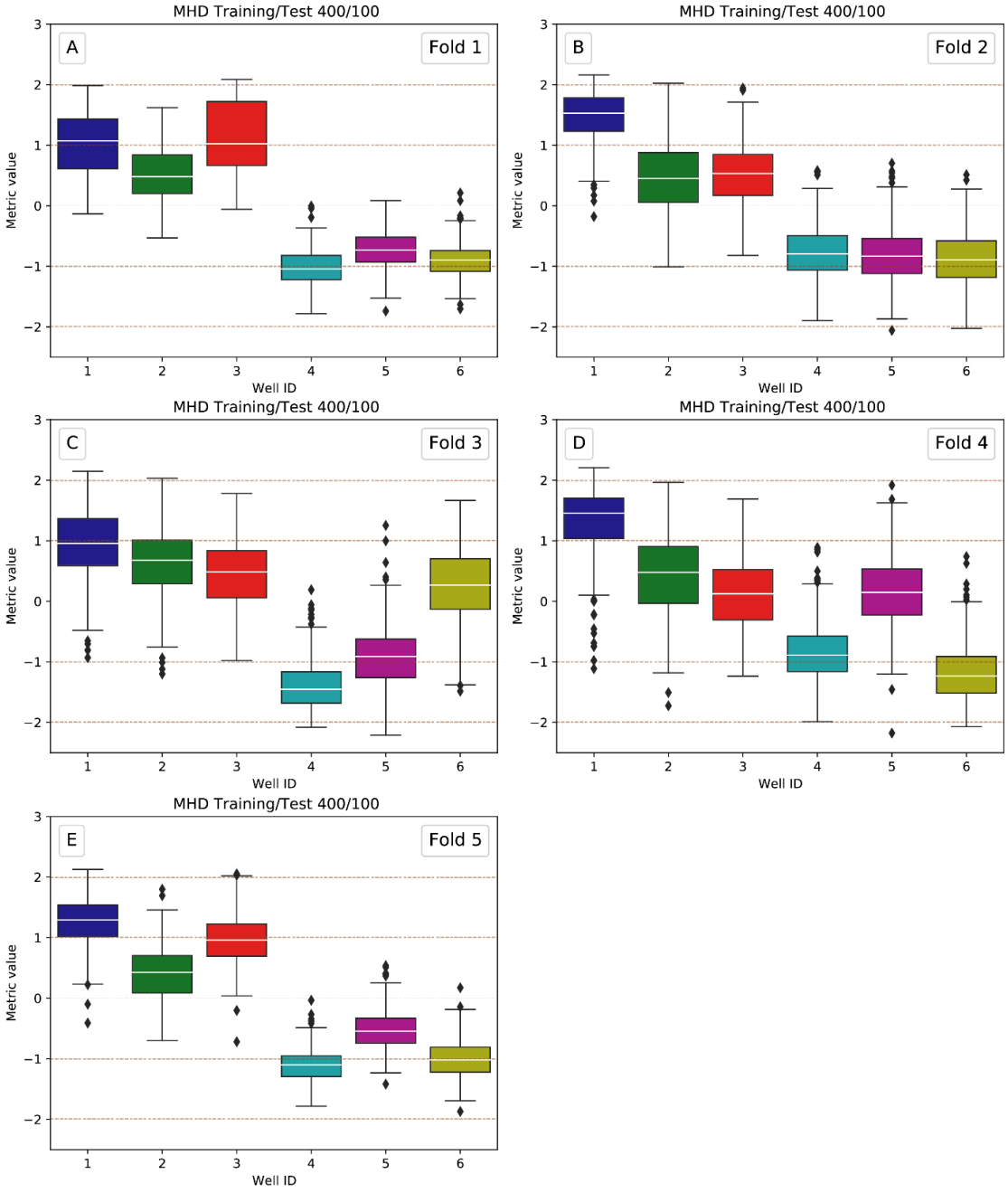

Figure 3.10: Boxplots of the standardised MHD distance for each well and the 5 successive k-fold for a 400-sample training dataset and a 100-sample test dataset. Across folds, the boxplots of each well are inconsistent.

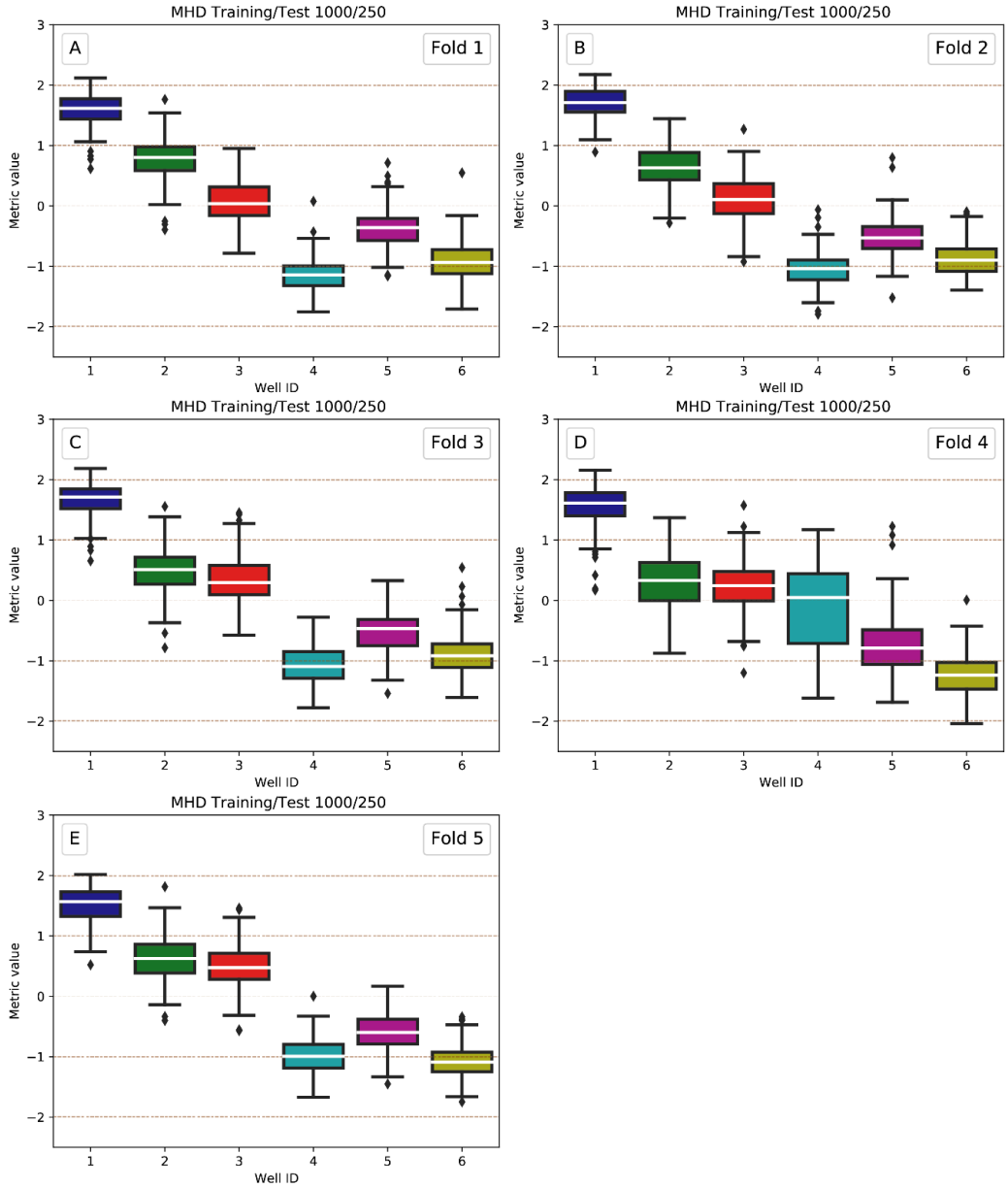

Figure 3.11: Boxplots of the standardised MHD distance for each well and the 5 successive folds for a 1000-sample training dataset and 250-sample test dataset. Across folds, the various boxplots are consistent with one another.

3.4.2 Multiple-well configuration

In this section, we search for the optimum combination of multiple wells. The framework is thus reproduced for the same experiment but with $\lambda = 2$ and 3, i.e., all combinations of 2 and 3 wells (total of 35 different combinations). The same training set as in the

previous section is used. Since the predictors and targets are already available, the methodology only needs to repeat the learning phase for each combination and then predict the posterior distribution for each tentative data set. The results are summarised in Figure 3.12 where the boxplots of the MHD are shown. The 3-well combinations are, as expected, more informative than the 2-well combinations. The most informative combinations include both downstream and upstream injection wells. The tracer curves in these cases naturally carry more information on the K field, reducing the uncertainty in the WHPA prediction, as illustrated in Figures 3.13A-B.

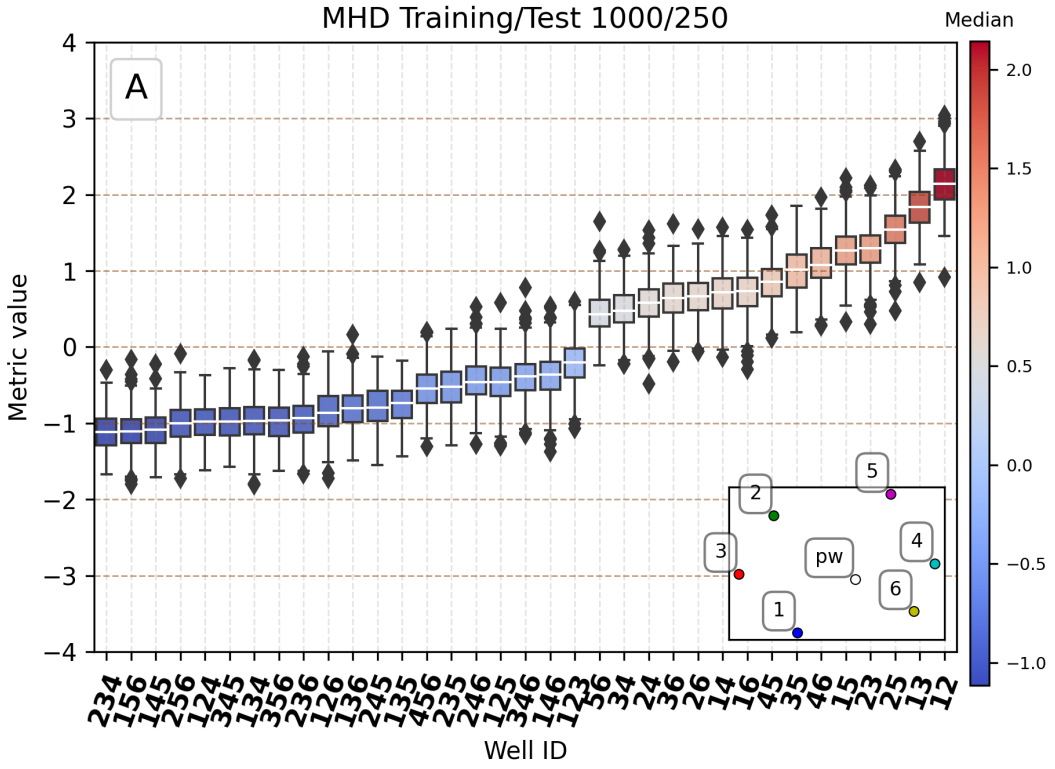

Figure 3.12: Boxplots of the standardised MHD distance for all 2 and 3 well combinations on a 1000-sample training dataset and a 250-sample test dataset.

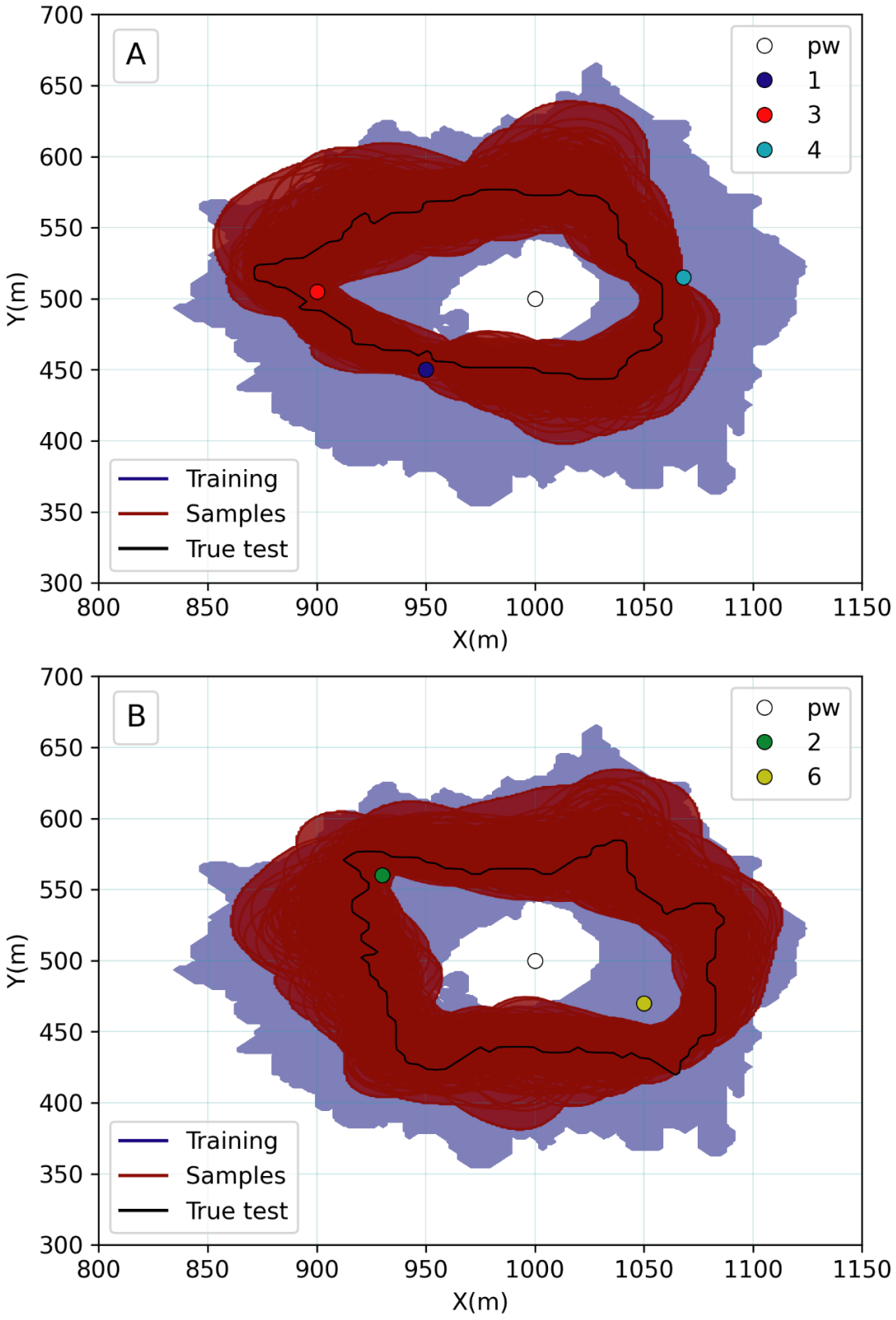

Figure 3.13: WHPA predictions for multiple-wells combinations, both performed with a training set of 1000 samples and test set of 250 samples. **A**. Prediction using wells 1, 3, 4. **B**. Prediction using wells 2, 6.

3.4.3 The case of anisotropic K field

The previously considered hydraulic conductivity fields have two degrees of uncertainty: the mean of hydraulic conductivity and the spatial distribution of the stochastically generated fields. The goal of this chapter is not to demonstrate BEL's capabilities for more complex prior distributions, as this has already been established in previous works such as Hermans et al. (2016); Satija and Caers (2015); Yin et al. (2020). However, it is important to demonstrate that the experimental design approach remains feasible when the uncertainty is greater. To show how the framework handles structural uncertainties, we include uncertainty on three additional parameters of the variogram model in order to implement anisotropy in hydraulic conductivity fields. Before running the SGSIM algorithm, we randomly select a value (Table 3.2) for the variance (or sill of the variogram), allowing us to simulate various degrees of heterogeneity, and the maximum range of the conductivity field variogram, as well as the orientation of the maximum range direction with respect to the x direction. The K mean is still randomly chosen as described in Table 3.1.

Parameter	Value
$\log_{10}$ K Standard deviation $(\frac{m}{d})$	$[0.1, 0.3]$
Angle around vertical axis (degrees)	$[-30, 30]$
Range max (m)	$[200, 400]$

Table 3.2: Note. Model parameters variation implemented to add structural uncertainty.

As expected, due to the greater prior uncertainty, the WHPA prior is broader than our initial case (Figure 3.14). The posterior uncertainty computed with the 6 wells is naturally larger as well, but the reduction in uncertainty is more significant than in the original case. The anisotropy and its effect on the WHPA, in particular, are well captured by the BCs. However, the experimental design results remain the same as in the isotropic case, with downstream wells being more informative (Figure 3.15). In this case, the greater uncertainty has no effect on the most informative well.

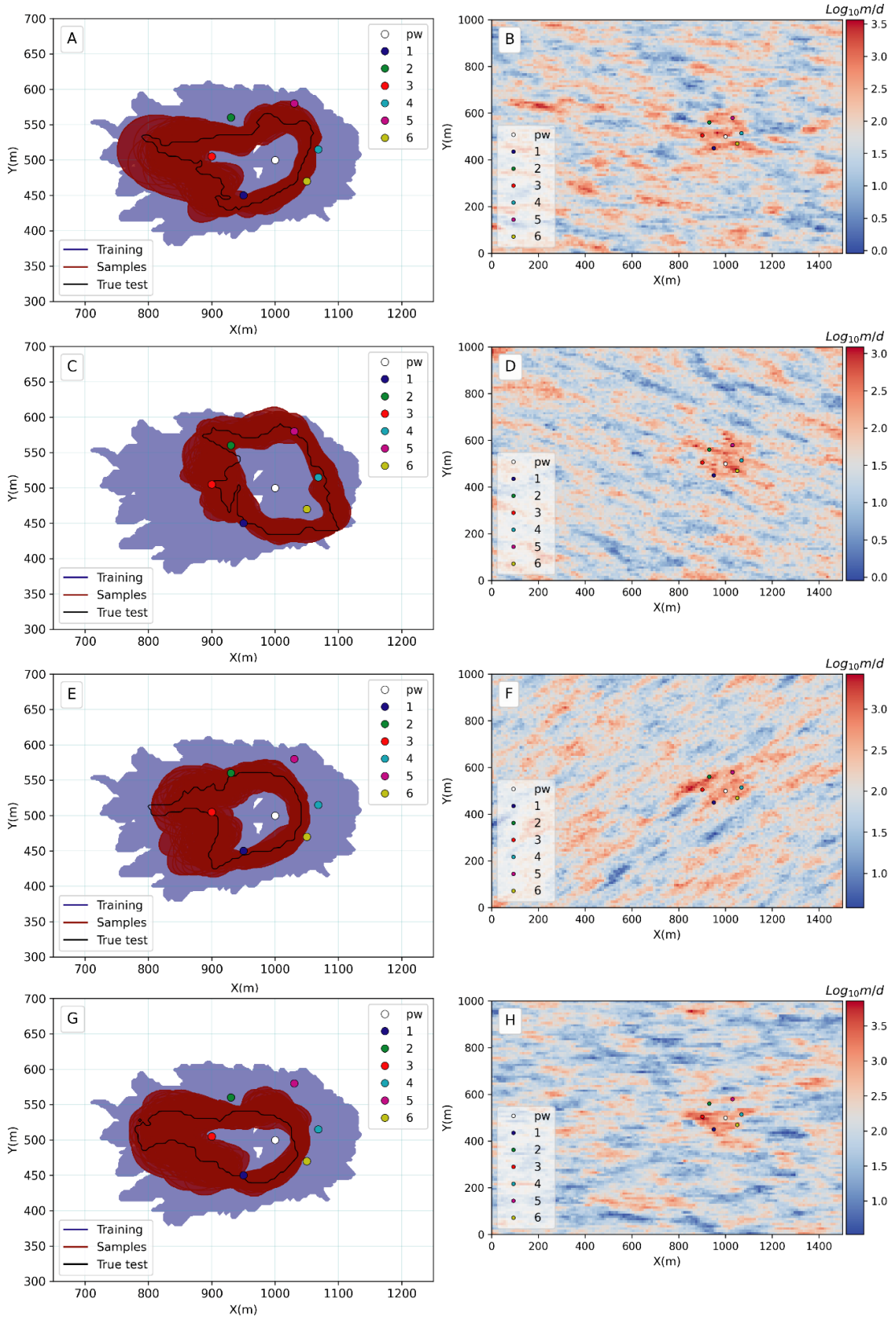

Figure 3.14: BEL-derived WHPA predictions with an anisotropic hydraulic conductivity field prior, performed with a 1000-sample training set and a 250-sample test set.

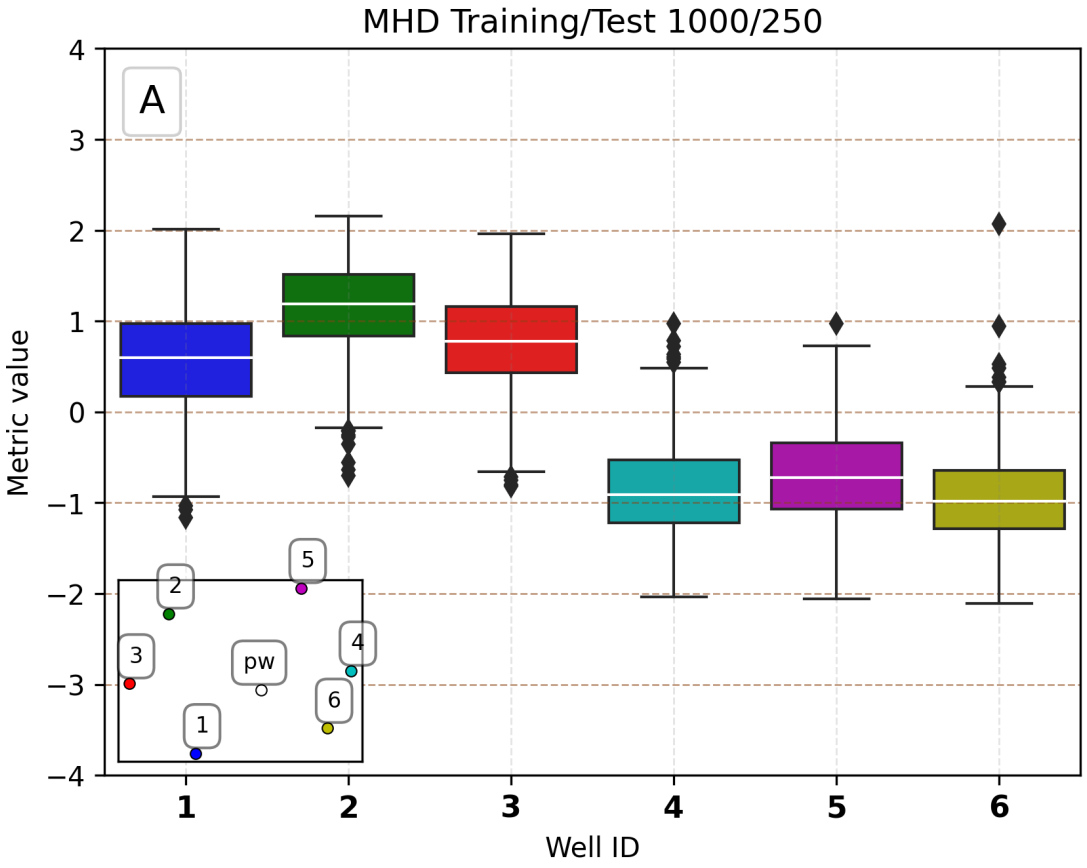

Figure 3.15: Boxplots of the standardised MHD distance for each well for a 1000-sample training set and a 250-sample test set.

3.5 Discussion

The main limitation of the BEL framework is to find an appropriate relationship between data and prediction. In contrast to traditional inversion methods, which rely on a data misfit or likelihood to constrain the solution to the data, BEL relies on statistical processes for both learning and prediction. The algorithms used in this chapter (PCA and CCA) uncover linear combinations between predictor and target variables (in reduced-dimension space) as initially designed in BEL (Hermans et al., 2018; Scheidt et al., 2018). Some PCA components of the target are ignored in CCA, so part of the variability cannot be explained. If necessary, this additional uncertainty can be added after back-transformation in the original space. Depending on the complexity of the predictor-target relationship, CCA might not be able to unravel part of the statistical relationship.

Although CCA can capture non-linear behaviours between target and predictor by combining variables, the non-linear components of these relationships are not always

properly captured, which may introduce bias in the prediction (Hermans et al., 2019) and result in overestimation of the posterior uncertainty (Michel et al., 2020b). This limitation can be mitigated in part by applying KDE instead of linear regression (Michel et al., 2020b). Recent research has suggested that the shortcomings of linear CCA can be addressed by modifying the learning procedure and introducing some iterative updating of the prior (Hermans et al., 2019; Michel et al., 2020a; Park and Caers, 2020). However, such adaptation increases the computation cost and reduces the adaptability of BEL as the iterative process is inherently dependent on the dataset, making it less efficient for experimental design. To keep the experimental design efficient, the learning phase could benefit from more advanced approaches, such as projecting the dataset into a higher-dimensional space through kernel transformation before performing CCA (e.g., Bilenko and Gallant 2016) or using non-linear CCA. This, however, is outside the scope of this chapter because linear canonical correlation analysis combined with linear regression was found satisfactory for WHPA prediction.

Since BEL prediction is based on a statistical process, it should not be applied blindly. The consistency of the data set with the prior should always be verified (Hermans et al., 2019, 2018). Otherwise, BEL runs the risk of producing an unrealistic posterior distribution. The prediction uncertainty may be slightly overestimated, due in part to the dimension reduction steps of PCA and CCA, which simplify the problem (see Figure 3.3B how the WHPA delineation is smoothed). Over-fitting, on the other hand, is avoided. Although the forward model is never computed during the prediction step in the BEL framework, the physics of the numerical model is included during the learning phase, ensuring that the performed predictions are consistent, as demonstrated by previous BEL applications, as long as the prior consistency is maintained. In experimental and optimum design, one often deals with synthetic rather than field data, and this condition is thus almost always verified.

Another limitation is the somewhat subjective choice of the size of the training set, which should be sufficient to characterise the variability in the target. This size is thus case-dependent and depends mainly on the complexity of the target (Hermans et al., 2018). Michel et al. (2020b) have shown that there is a threshold above which increasing the size has no effect on the posterior prediction. The latter is found by progressively increasing the number of samples in the prior until the posterior stabilises. Using a sufficiently large prior reduces the risk of obscuring the predictor-target relationship. In this study, we found that a training set of 400 is adequate for both prediction and experimental design purposes. Cross-validation shows that the size of the test set required for experimental design is at least 250, so the training set size was increased to 1000 samples to allow for k-fold cross-validation. The number of PCs to keep for both predictor and target, the number of posterior samples to compute the uncertainty reduction, and the definition of the data-utility function, depending on the nature of the target variable, are all additional experimental design variables that must be set by the practitioner.

A possible limitation is the heterogeneity of the medium used for forward modelling (e.g., hydraulic conductivity field). In this work, the different fields are generated by

sequential Gaussian simulations (2-points statistics) and are inherently smooth. We also investigate uncertainty in variogram parameters and demonstrate the framework's robustness when dealing with structural uncertainty. The natural variability of geological media can be far from smooth, for example, channelised media (e.g., Lopez-Alvis et al. 2021), necessitating more advanced simulation techniques such as multiple-point statistics (Mariethoz and Caers, 2014). Since the focus of this study was to demonstrate the ability of BEL in an experimental design framework, we did not investigate these more heterogeneous priors. Since BEL has been successfully used for such complex priors (Hermans et al., 2016; Satija and Caers, 2015; Yin et al., 2020), we do not anticipate any issues when applying the proposed framework to such cases. Similarly, Hermans et al. (2019, 2018) used BEL with more uncertain components such as boundary conditions and 3D cases with layers with different distributions. Although such greater prior uncertainty would inevitably increase posterior uncertainty, our experimental design approach could still be used without loss of generality.

3.6 Conclusion

In this chapter, we propose a novel approach that combines experimental design with Bayesian Evidential Learning (BEL). The targets, wellhead protection areas (WHPAs) surrounding a pumping well, are stochastically predicted by using breakthrough curves (BCs) from tracing experiments as predictors. A direct relationship between the predictor and the target is found using a small training set (400 samples) and is used to estimate the full posterior distribution of an unknown target given any new predictor not included in the training set, corresponding to observed data. The prediction procedure has a low computational cost and does not require model calibration through data inversion.

By estimating the prior to posterior uncertainty reduction for a series of prospective datasets, we assess the informative content of the considered injection wells (data sources) in order to identify the optimal location of data sources. Data-utility functions based on the Modified Hausdorff Distance (MHD) and the Structural Similarity (SSIM) index are used to quantify the prediction's uncertainty. To validate the size of the prospective data sets, we use a k-fold cross-validation procedure. It is demonstrated that a training set of 400 samples is sufficient for estimating WHPAs with BEL, whereas a test set of 250 samples is required to draw robust conclusions from the most informative data sources. In contrast to previous approaches used for similar experimental design problems (e.g., Bayesian model averaging, surrogate modelling), our approach does not require thousands (or more) forward model evaluations while estimating the full posterior distribution of the target of interest without having to simplify the mathematical model.

4. Comparing Well and Geophysical Data for Temperature Monitoring within a Bayesian Experimental Design Framework

This chapter was published in Water Resources Research (Thibaut et al., 2022):

Thibaut, Robin, Nicolas Compaire, Nolwenn Lesparre, Maximilian Ramgraber, Eric Laloy, and Thomas Hermans (Nov. 2022). "Comparing Well and Geophysical Data for Temperature Monitoring Within a Bayesian Experimental Design Framework". In: Water Resources Research 58 (11). issn: 0043-1397.
doi: 10.1029/2022WR033045.

CRediT author statement. Robin Thibaut: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Nicolas Compaire**: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation. **Nolwenn Lesparre**: Investigation, Writing - Review & Editing. **Maximilian Ramgraber**: Software, Writing - Review & Editing. **Eric Laloy**: Conceptualization, Methodology, Writing - Review & Editing, Supervision. **Thomas Hermans**: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Abstract

Temperature logs are an important tool in the geothermal industry. Temperature measurements from boreholes are used for exploration, system design, and monitoring. The number of observations, however, is not always sufficient to fully determine the temperature field or explore the entire parameter space of interest. Drilling in the best locations is still difficult and expensive. It is therefore critical to optimize the number and location of boreholes. Due to its higher spatial resolution and lower cost, four-dimensional (4D) temperature field monitoring via time-lapse Electrical Resistivity Tomography (ERT) has been investigated as a potential alternative. We use Bayesian Evidential Learning (BEL), a Monte Carlo-based training approach, to optimize the design of a 4D tem-

perature field monitoring experiment. We demonstrate how BEL can take into account various data source combinations (temperature logs combined with geophysical data) in the Bayesian optimal experimental design (BOED). To determine the optimal data source combination, we use the Root Mean Squared Error (RMSE) of the predicted target in the low dimensional latent space where BEL is solving the prediction problem. The parameter estimates are accurate enough to use in BOED. Furthermore, the method is not limited to monitoring temperature fields and can be applied to other similar experimental design problems. The method is computationally efficient and requires little training data. For the considered optimal design problem, a training set of only 200 samples and a test set of 50 samples is sufficient.

Plain Language Summary

The design of experiments is a critical step in scientific research to ensure that the data collected is of sufficient quality to answer the research question. The design of an experiment is often optimized to minimize operational costs while still fulfilling the desired accuracy in the prediction. In this chapter, we use an approach called Bayesian Evidential Learning to optimize the design of an experiment. The Bayesian philosophy is used to incorporate all available information into the design of the experiment and to quantify the uncertainty in the prediction. In particular, our approach makes it straightforward to use different data sources (sensors), and different types of data (e.g., temperature, pressure, and concentration measurements). We demonstrate the method in the design of a temperature field monitoring experiment to characterize the geothermal energy storage potential of an aquifer, a key step in the development of geothermal energy projects. This framework makes experiment design easier and faster, which is required for sound risk analysis and decision-making. The method is not limited to temperature fields and can be applied to other similar experimental problems, such as the monitoring of a contaminant plume or saltwater intrusion.

Key Points

- We propose a method for optimizing the design of a 4D temperature field monitoring experiment.
- We use Bayesian Evidential Learning (BEL) to infer the posterior distribution of the temperature field for the experiment design.
- We demonstrate how BEL can combine temperature logs and ERT data to assist in the design of experiments for a 4D temperature monitoring.

4.1 Introduction

This chapter presents a methodology for improving the design of field experiments using well and geophysical data by utilizing Bayesian Evidential Learning (BEL). In this chapter, we propose a methodology for solving a BOED problem involving two different

data types, using the BEL framework in a low-dimensional latent space, which alleviates the "curse of dimensionality" (Bellman, 1961) and reduces computational and memory demands. This study is based on the monitoring of a heat injection-storage-pumping experiment monitored by time-lapse ERT and well data mimicking the field experiment of Lesparre et al. (2019). In addition to having the potential to improve subsurface information, the fusion of various data types can also be used to plan surveys and choose the optimal set of instruments (JafarGandomi and Binley, 2013). There is a tradeoff between non-invasive data acquisition methods such as ERT and invasive data acquisition methods such as drilling, with the latter being generally more expensive. It is critical to determine whether drilling is required or if a simple geophysical survey will suffice, depending on the specific problem at hand. It can also be useful to know whether it is worthwhile to conduct a long geophysical survey and mobilize a large amount of equipment if we can use an existing borehole in the area.

This chapter is unique in five ways when compared to previous applications of BEL to experimental design:

1. We use two different type of predictors, geophysical and borehole data, which vary in time, to predict a four-dimensional target, the temperature field magnitude in the aquifer over time. It will first be demonstrated how to predict the target using each data set separately.

2. Then, it will be demonstrated how to use them together. Because the predictors are of different types, the dimensionality reduction step in the pre-processing section is applied to each instance separately, and then concatenated before being fed to the learning algorithm.

3. The Transport Map method is used to sample in the low dimensional space.

4. The data utility function is calculated in the low dimensional latent space, without the need to back-transform the data to its original space.

5. We identify the optimal choice between two ERT protocols with our proposed methodology.

4.2 Methodology

4.2.1 Experimental setup

We present a method for estimating the posterior distribution of an unknown four-dimensional temperature field during a heat injection-storage-pumping experiment. The experiment has been described in Lesparre et al. (2019). In short, hot water was injected at 3 m^3/h into an aquifer at a temperature of 42 °C for six hours, followed by another injection at 14.5 °C for 20 minutes. Then, it was stored in the aquifer for 92 hours, and then pumped back out for 16 hours and 15 minutes at a flow rate of 3 m^3/h. The goal was to track the evolution of temperature distribution in the aquifer over time. The experiment was monitored using time-lapse ERT and temperature data in

wells. The time-varying predictor (measured voltage for ERT, direct temperature for wells) and the target (temperature distribution in the aquifer) are high-dimensional, and their relationship is non-linear. By performing geothermal field experiments such as injection and hot water pumping tests, we can gain a better understanding of the aquifer's behavior Klepikova et al. (2016); Macfarlane et al. (2002); Palmer et al. (1992); Park et al. (2015); Vandenbohede et al. (2011, 2009); Wagner et al. (2014); Wildemeersch et al. (2014). Combined with these tests, geophysical and thermal monitoring can track heat transfer in the aquifer. The location of injection wells is the focal point of the hydrogeological models used in this study. These models' grids are 60 meters in length in the known direction of natural aquifer flow and 40 meters in length in the perpendicular direction to that direction (Figure 4.1A). The grid layers begin at the surface of the aquifer's saturated zone at a depth of 3 meters and conclude at the surface of the impermeable basement at 10 meters. The space step along the Z axis (depth) is 0.5 meters. The space step along the X and Y axes ranges from 2.5 centimeters at the injection point to 2.5 meters at the model's edges, with a 0.25 centimeter refinement in a 3-meter radius around the injection points in the hydraulic flow direction (Y axis).

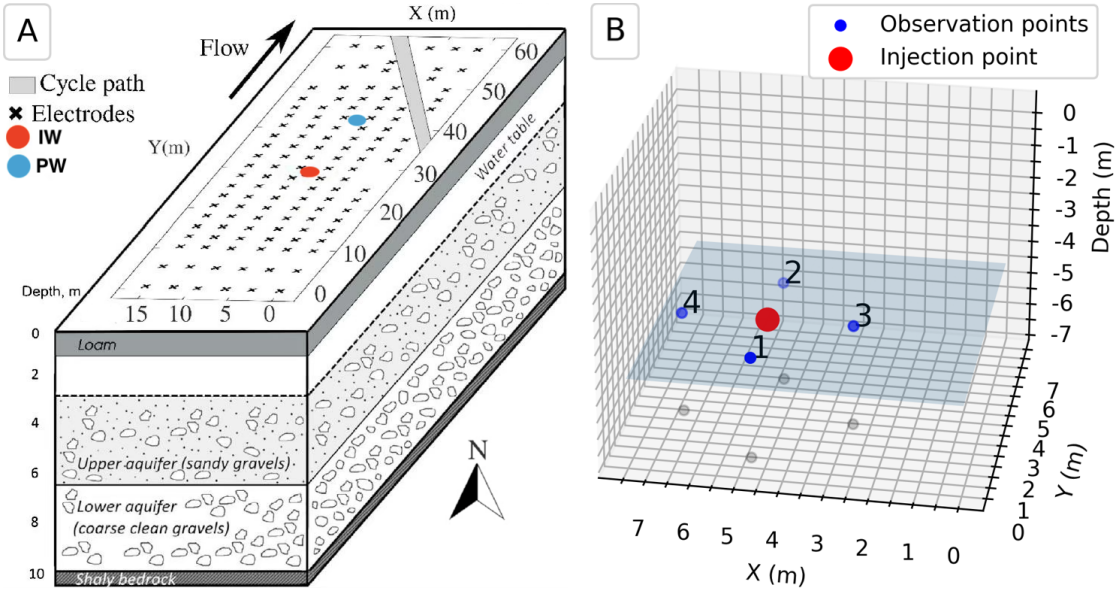

Figure 4.1: **A**. Model design (modified from Lesparre et al. (2019)). IW: Injecting well. PW: Pumping well. **B**. Positions of observation and injection wells. The boreholes are screened to around a depth of 4.95 m

The mean and variance of hydraulic conductivity (K), the anisotropy factor, and its direction in the horizontal plane and the range of the spherical variogram are used to build the models of the hydraulic conductivity distribution in the aquifer. In addition, the homogeneous porosity and the natural gradient are set as uncertain parameters. All other parameters including water and matrix thermal properties are set as constant. Bulk thermal properties are calculated based on porosity using the arithmetic average. The prior is the sum of all the definitions of the parameters under consideration. We

detail the ranges of the variables in Table 4.1.

Uncertain parameter	**Range**
Log K mean	U[-4, -1], K in m/s
Log K variance	U[0.05, 2], K in m/s
Effective porosity	U[0.05, 0.3]
Variogram main range	U[1, 10] m
Anisotropy ratio	U[0, 0.5]
Orientation	U[0, π] rad
Natural gradient	U[0.05, 0.167] %

Table 4.1: Parameters of the prior model. $U[a, b]$ refers to the continuous uniform distribution bounded by the values a and b.

Each parameter is generated randomly and independently according to a uniform distribution. The sequential Gaussian simulation algorithm Goovaerts (1997) is then used to generate the hydraulic conductivity fields. In the direction of Y, the hydraulic flow at the grid boundaries is zero, and it respects the natural gradient in the X-direction. The direction of the flow due to the natural gradient is indicated on Figure 4.1A. For this investigation, 250 hydrogeological models were built and used to generate the temperature fields resulting from the injection-storage-pumping experiment. The HydroGeoSphere code was used to simulate the temperature field Brunner and Simmons (2012). There are 106-time observations in each simulation. The temperature grids obtained are reduced to sub-grids around the injection well of size 16×16×14 of elementary volumes 0.5 m×0.5 m×0.5 m (volume affected by temperature changes within the 250 simulations) to reduce the amount of data (Figure 4.1B). The temperature is averaged when the subgrid elemental volumes contain several elemental volumes from the initial grid.

The following equations were used to transform the temperature T into conductivity Hermans et al. (2014):

$$\frac{\sigma_{f,T}}{\sigma_{f,25}} = m_{f,25}(T - 25) + 1, \text{ and} \tag{4.1}$$

$$\Delta T = \frac{1}{m_{f,25}} \left[\frac{\sigma_{b2,T}}{\sigma_{b1}} \frac{\sigma_{f1}}{\sigma_{f,25}} - 1 \right] + 25 - T_{init}, \tag{4.2}$$

where $m_{f,25}$ is the fractional change of the fluid conductivity per degree Celsius around the reference temperature of 25°C, σ_{b1} is the background conductivity, $\sigma_{b2,T}$ is the background conductivity at temperature T, σ_{f1} is the fluid conductivity at the initial state, $\sigma_{f,25}$ is the fluid conductivity at 25°C, and T_{init} is the initial temperature.

The synthetic study replicates a real ERT monitoring campaign performed on a well documented site Brouyère (2001); Dassargues (1997); Derouane and Dassargues (1998); Hermans et al. (2015b); Klepikova et al. (2016). Therefore, the initial temperature was measured directly in the field, and the background resistivity σ_{b1} was determined using the inversion of the real field data. The temperature and fluid conductivity trend from

on-site water samples was calculated to be $m_{f,25}$=0.0194 and $\sigma_{f,25}$ was estimated to be 0.0791S/m Hermans et al. (2015b); Lesparre et al. (2019). σ_{f1} = 0.0614S/m was estimated from Equation 4.2 with an initial temperature of 13.44°C.
The forward modeling process went as follows:

1. Temperature field simulations on the discretized grid (HydroGeoSphere, Brunner and Simmons (2012)).
2. Computation of the conductivity with the petrophysical law for each simulation and time step (Equation 4.2).
3. Computation of the ERT for each simulation and time step (EIDORS, Polydorides and Lionheart (2002)).

The ERT simulations are computed with the following experimental setup: 6 parallel profiles of 21 electrodes in the X direction, with a 2.5 meter spacing between the 17 central electrodes and a 5-meter spacing between the four electrodes at the profile's edge. Two protocols were used: multiple gradient (MG) and dipole-dipole (DD). The DD array was chosen due to its widespread use in ERT investigations, which is due in part to the low electromagnetic coupling between the circuits. The MG array was another appealing option due to its sensitivity distribution and high S/N ratio, as well as its robustness and multichannel compatibility Dahlin and Zhou (2006). For each of the 106 temperature simulation time observations, an ERT simulation is generated for each model. At each time step, 1100 quadrupoles are used for the DD protocol and 848 quadrupoles are used for the MG protocol. The target for each model is thus made of 16×16×14 temperature grids for the 106 observation times, and the predictor is made of the resistances measured by 1100 or 848 quadrupoles for the 106 observation times.

4.2.2 Heat prediction

Pre-processing

Target The target H is a three-dimensional temperature field subdivided into $\mathrm{n}_{rows} \times \mathrm{n}_{cols} \times \mathrm{n}_{lay}$ over n_{step} observation time steps. It is critical to reduce dimensionality because some small-scale variations of the target do not need to be perfectly reconstructed as they are already beyond the predictor's resolution. Before performing dimension reduction, the raw target is scaled to unit variance, because the dimension reduction step is sensitive to the scale of the data. The dimension reduction itself is done by linear PCA. The principal components are new variables produced by combining the initial variables in a linear way.

Predictor The geophysical predictor D_g is made up of resistance values measured by n_{quad} quadrupoles over n_{step} observation time steps, and the borehole predictor D_b is made of temperature curves measured at observation well's locations. Before performing dimension reduction by linear PCA, both raw predictors are scaled to unit variance. When working with geophysical data, filtering higher dimensions allows to reduce the effect of noise on the prediction (Hermans et al., 2016; Michel et al., 2020b).

The number of components to keep for both predictor and target is automatically determined by setting the amount of variance that needs to be explained to 99.9%. After dimensionality reduction, the PCs of both predictor and target are scaled to unit variance, because covariance matrices are sensitive to the scale of the data.

Training

Following dimensionality reduction, the next step is to use CCA to determine the relationship between the predictor and the target in the reduced space. Let δ be the number of PCs necessary to explain the required amount of variance in the predictor, and $n_{training}$ be the number of pairs of predictors and targets used for training. Our data fusion technique consists of concatenating the geophysical and borehole temperature PCs into a single matrix. Therefore, δ is the sum of the number of components of the geophysical data and the borehole temperature curves, i.e., $\delta = \delta_g + \delta_b$. The CCA algorithm projects the data onto a new set of axes that maximizes the correlation between the two data sets using the cross-covariance matrix of latent variables (PCs). The transformed variables are the CVs, representing the mutual information between the two data sets. To allow back-transformation, more components must be used for the predictor than for the target before learning the relationship between the two. However, to avoid overfitting and noise propagation, it is recommended that both have a similar number of components, although it does not have to be strictly the same. Therefore, the number of CCA components is set to δ, the maximum number that can be used (Meloun and Militký, 2012). Let the superscript c denote the canonical space. The canonical variates (CVs) pairs are stored in the $(n_{training} \times \delta)$ matrices

$$D_\delta^c = d_{i,1}^c, d_{i,2}^c, \ldots, d_{i,\delta}^c | i = 1, \ldots, n_{training} \tag{4.3}$$

$$H_\delta^c = h_{i,1}^c, h_{i,2}^c, \ldots, h_{i,\delta}^c | i = 1, \ldots, n_{training} \tag{4.4}$$

With the pairs of canonical variates $\left(d_{:,1}^c, h_{:,1}^c\right) \sim \pi_1$ to $\left(d_{:,\delta}^c, h_{:,\delta}^c\right) \sim \pi_\delta$ established, we may infer the posterior in the canonical subspaces by independently conditioning each of the resulting bivariate joint distributions $\pi_j, j = 1, \ldots, \delta$ on a new observation projected into the canonical predictor space $d_{new,j}^c$. In this study, we consider a new approach based on triangular transport, which offers a good tradeoff between computational efficiency and accuracy (cf. 2.3.1).

4.2.3 Experimental design

BEL can be used to assess the amount of information delivered by various data sources. The actual data can have any value within the prior data space, and data sources can be placed anywhere across the grid. To identify informative data sets based on their location, one data-utility function must be maximized or minimized (cf. Chapter 2). To restrict computation time and keep some realism, we will only consider four static well positions around the injection well, which is sufficient to demonstrate the approach. With the geophysical data, a total of 31 different combinations of data sets are possible. These unique combinations are combined with a total of 50 unknown ground truths

(test set). Each sample of the test set is sampled with 500 posterior samples. This number is arbitrarily chosen to be high enough to ensure that the posterior distribution is reasonably well-sampled while remaining low enough to keep the computation time manageable. We end up with an array of shape (31, 50, $n_{samples}$, δ) = (number of combinations, test set size, sample set size, number of canonical components).

One of the challenges of our case study is the high dimensionality of the problem. This is why we use the principal components to perform BOED, which is a novel approach in this context. The true target is transformed to the PC space and its RMSE with the predicted targets' PCs is computed. The advantage of working in a lower dimension is that we need to predict a smaller number of dimensions, which is computationally faster. Because the back-transformation from the PCs to the observation targets is a linear operation, it is simple to demonstrate that the prediction error is minimized if the distance between the PCs of the predicted and observed targets is also minimized. The minimization of the target prediction error is therefore equivalent to the minimization of the distance between the PCs of the predicted and the observed targets. The PCs are weighted by their explained variance ratio during distance computation because they account for the different importance of the different components.

We validate our approach with k-fold cross validation. We repeat our computations across 5-folds and average the metric results. Averaging across folds is not strictly necessary but will be used here to increase the robustness of the BOED. The methodology is summarized below:

Input: training set (D_{train}, H_{train}), test set (D_{test}, H_{test}), data utility function UF
Output: Averaged cross-validation rankings
for all Fold f in 5-fold cross-validation **do**
    $H_{train,f} \leftarrow$ Target training data for fold f
    $H_{train,f} \leftarrow PCA_h.fit\_transform(H_{train,f})$
    $H_{test,f} \leftarrow$ Target test data for fold f
    $H_{test,f} \leftarrow PCA_h.transform(H_{test,f})$
    for all Possible Combinations **do**
        $O \leftarrow Combination$
        $D_{train,f,O} \leftarrow$ Predictor training data for fold f and combination O
        $D_{train,f,O} \leftarrow PCA_d.fit\_transform(D_{train,f,O})$
        $D_{test,f,O} \leftarrow$ Predictor test data for fold f and combination O
        $D_{test,f,O} \leftarrow PCA_d.transform(D_{test,f,O})$
        $TrainedModel \leftarrow CCA.fit(D_{train,f,O}, H_{train,f})$ ▷ Training step
        for all Ground Truths in $(D_{test,f,O}, H_{test,f})$ **do**
            $D_{true} \leftarrow$ True predictor
            $D_{true} \leftarrow PCA_d.transform(D_{true})$
            $H_{true} \leftarrow$ True target
            $H_{true} \leftarrow PCA_h.transform(H_{true})$
            $H_{posterior} \leftarrow TrainedModel.predict(D_{true})$ ▷ Predicting in PC space
            $Utility \leftarrow UF(H_{posterior}, H_{true})$

```
            Results.add(Utility)
        end for
        Ranking ← add(Results)
        Results.clear()
    end for
end for
Final ← mean(Ranking)
```

The forward modeling part to generate the training and test sets took about 5 days on a standard desktop computer. The BOED method was developed in Python on a standard laptop with 16 GB of RAM and a 2.30 GHz 8-core Intel Core i9 processor. Profiling has been performed on the code. It took 27.7 seconds load the dataset, pre-process the data, and save the files. This part only has to be done once for each fold. The most time-consuming aspect of BOED is the prediction of samples (through transport map inference), which scales linearly with $n_{samples}$. For the most extensive data combination (ERT, wells 1 to 4), training the model took 0.053 seconds, and predicting $n_{samples}$ took 24.91 seconds. For 1 fold, 31 combinations and 50 examples, for a total of 7.75×10^6 predicted samples, the total time to run the BOED was 89 minutes, and the process used no more than 1.7 GB of RAM. For 5 folds, 31 combinations and 50 examples, for a total of 38.75×10^6 predicted samples, the entire BOED took about 7 hours and 30 minutes, exceeding no more RAM than previously mentioned. Given that the entire implementation was created for research purposes, the BOED implementation can be enhanced for speed and memory management (e.g., parallelization, low-level programming language implementation, etc.).

4.3 Application

4.3.1 Target prediction

This section shows how to predict a single four-dimensional temperature field, using three distinct predictors: (i) the geophysical data, made of MG-array resistance measurements from 848 quadrupoles over 106 observation time steps, (ii) a single temperature profile over 106 observation times, and (iii) the combination of the same geophysical data with all 4 borehole temperature curves. We show that BEL can accurately estimate the target posterior distribution with varying uncertainty levels according to the type of predictor used. The dataset in this section has a total size of $n = 250$. The training set is then reduced to $n_{80\%} = 200$ models, with the remaining models being used to validate BEL's ability to predict the target, and later on for BOED. Previous BEL applications have demonstrated that making accurate predictions with a dataset of this size is possible (Athens and Caers, 2019; Hermans et al., 2019, 2018, 2016; Michel et al., 2020b; Park and Caers, 2020; Thibaut et al., 2021b; Yin et al., 2020). While a small training set size is inevitable due to the time-consuming nature of the simulations, it is sufficient because the prediction is a temperature distribution that varies smoothly in both time and space and results from advection, diffusion, and dispersion processes. Such target is much simpler than the underlying K model. BEL is a Bayesian method

that incorporates uncertainty. Therefore, a large training set size is not required and the method is more robust against overfitting. Furthermore, the use of cross-validation ensures the robustness of the results.

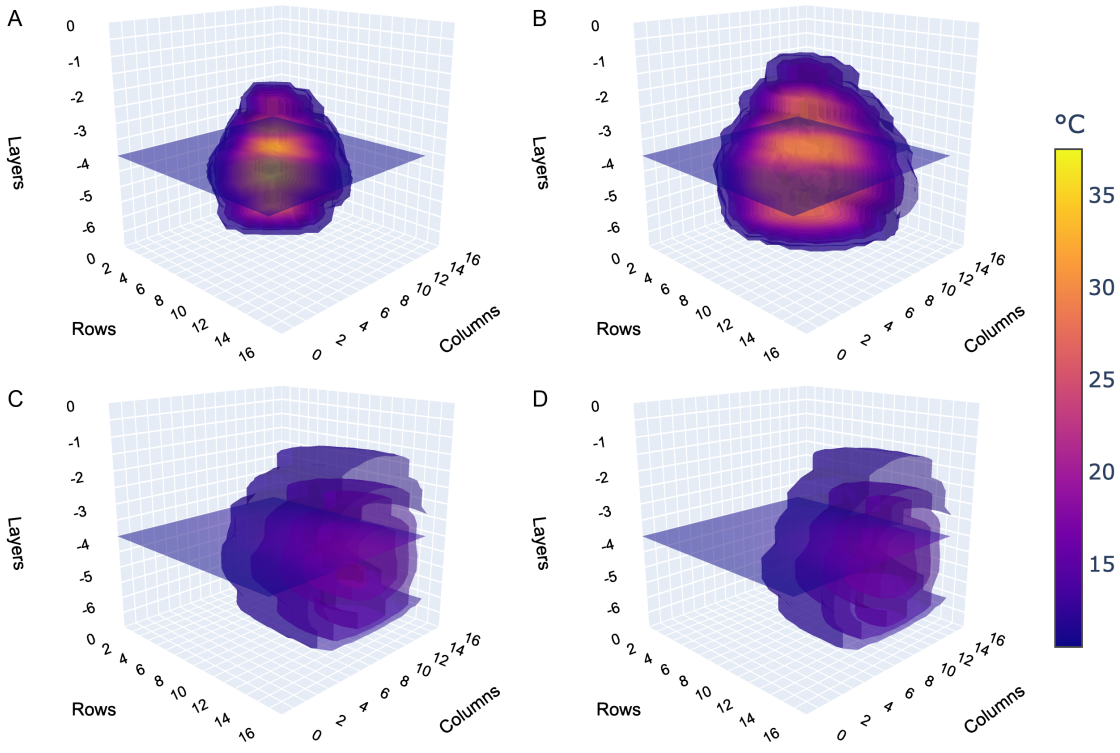

Figure 4.2: Snapshots of temperature field contours at different time steps for one example. **A**. 5th time-step (2.5h–injection phase). **B**. 15th time-step (10.5h–storage phase). **C**. 62nd time-step (99.25h–pumping phase). **D**. 75th time-step (105.75h–pumping phase). The injection well discharge is at (column, row, layer) = (9, 6, -5). The reference plane at the wells level is highlighted.

Pre-processing

The three predictors undergo the pre-processing described in Section §2.3.1. When geophysical data is combined with borehole temperature curves, the following steps are taken:

1. PCA is performed to explain 99% of the variance in the geophysical data to obtain the PCs $D_{G,99}$
2. PCA is performed to explain 99% of the variance in each of the n_b boreholes temperature curves to obtain the PCs $D^{(i)}_{B,99}$ for each borehole i. These PCs are then concatenated into a single array $D_{B,99}$.
3. The geophysical PCs are concatenated with the borehole temperature curves PCs to obtain the array $D_{G,B,99}$

We have to perform PCA separately for each predictor, because if we simply concatenated the predictor arrays and then applied PCA, the resulting predictor would be more representative of the geophysical data than the borehole temperature curves. Furthermore, the PCs are scaled to unit variance before concatenation because the PCs magnitude can vary greatly from predictor to predictor. The resulting principal components for all cases can directly be compared in Figure 4.3. The left column shows the predictor PCs, and the right column the target PCs for all cases. For each PC dimension (horizontal axis), the PC value for each of the 200 training instances is plotted (vertical axis). The test instance is plotted on top of the training instances. In line with our methodology, the number of components to keep is set to the maximum number of PCs required by the predictor, and the corresponding amount of variance explained in the target is summarized in Table 4.2.

Parameter \Combination	**G**	**1**	1, 2	1, 2, 3	1, 2, 3, 4	**G, 1, 2, 3, 4**
δ	10	3	7	10	13	23
Explained variance (target)	77%	57%	71%	77%	80%	88%

Table 4.2: Note. Effect of the number of PCs (δ) on the target PCA explained variance. **G** stands for geophysical data. **1, 2, 3, 4** stand for the borehole temperature curves. Case (i) is **G**, case (ii) is **1** and case (iii) is **G, 1, 2, 3, 4**.

Because of the smaller number of PCs, when a single temperature profile is used as a predictor (case (ii)), the target variance is naturally not explained as well as when the entire set of predictor data is used (case (iii)). When the predictor is a combination of geophysical data and all temperature profiles, some additional variance is captured. The remaining target PCs for cases (ii) and (iii) are shown in Figure 4.3B and D, respectively. These additional target components are not used in training. They are, however, saved for subsequent use in BOED.

Training and prediction

The CCA mapping allows the fusion of the low-dimensional representation of the predictor with the low-dimensional representation of the target in order to make predictions of the target distribution. It is run on all cases to find the canonical variates that define the relation between the predictor and the target, using a maximum of 23 PCs on the predictor (see Table 4.2). The first three canonical variate pairs for each case are shown in Figure 4.4. Each row corresponds to a single case, and each column to a single canonical variate pair. The variable ρ is the correlation between the canonical variates, and can be interpreted as the amount of mutual information shared between the target and predictor. In all cases, the predictor explains a large part of the target: ρ_1 (first pair) = 0.999, 0.959 and 1. for case (i), (ii) and (iii) respectively. Hence, CCA is well-suited for our purposes. The canonical variates of (ii) provide a poorer explanation for the target (ρ_2 (second pair) = 0.462). This is because the temperature curves used as predictors are one-dimensional and only convey information about a small portion of our model over time. The strong correlations in the canonical variates pairs of cases (i) and (iii) is further indication of the suitability of CCA for this study. Transport map inference

is run for each pair of each case, using the test predictor (on the horizontal axis (d^c) of Figure 4.4). The 500 samples of $p(h^c|d^c_*)$ in Figure 4.4 represent our predictions of the unknown target distribution in the canonical space. They can then be sequentially back-transformed to the principal component space (see posterior samples in Figure 4.3) and to the original space (see Figures 4.5 and 4.6 for 1D and 2D representations, respectively).

Figure 4.5 shows the temperature curves of the 500 samples at the location of the observation well number 2 for each case. This point was arbitrarily chosen to illustrate the one-dimensional temperature curve at one observation point. Across cases, the level of uncertainty, expressed by the spread of the temperature curves, is higher during the injection and storage phases than it is during the pumping phase, which is a positive development because this portion of the curve is the part that is typically inferred for ATES systems (energy recovery). The magnitude of the spread of the predicted samples in the canonical space (Figure 4.4) is sequentially transmitted through the principal component space (Figure 4.3) and the original space (Figure 4.5), since the data flows through linear transformations.

On Figure 4.6A, the true (test) temperature cross-section at layer 9 (at a depth of 4.5m) and time t=105.75h (pumping phase) is shown. Figure 4.6B shows one randomly drawn example from the 500 sampled temperature profiles when using the ERT predictor alone. The results are not only visually close to the truth, but the absolute temperature difference ranges from 0.53 to 1.36 degrees Celsius, which is the magnitude of accuracy we expect when using resistance data Hermans et al. (2015b). Figures 4.6C and 4.6D show the same results when a single borehole temperature profile and the full combination of ERT and temperature profiles are used, respectively. As expected, the single borehole image is the least accurate, while the combination of all temperature profiles and ERT is the most accurate.

Because the underlying internal behavior of the aquifer is nonlinear, given the number of parameters involved, the underlying physics of the model and the nature of the predictor (indirect geophysical data), such a level of uncertainty is expected Hermans et al. (2018, 2016, 2015b). Considering the level of complexity of the underlying physics and the relatively small training set, BEL is capable of inferring different possible outcomes with reasonable uncertainty.

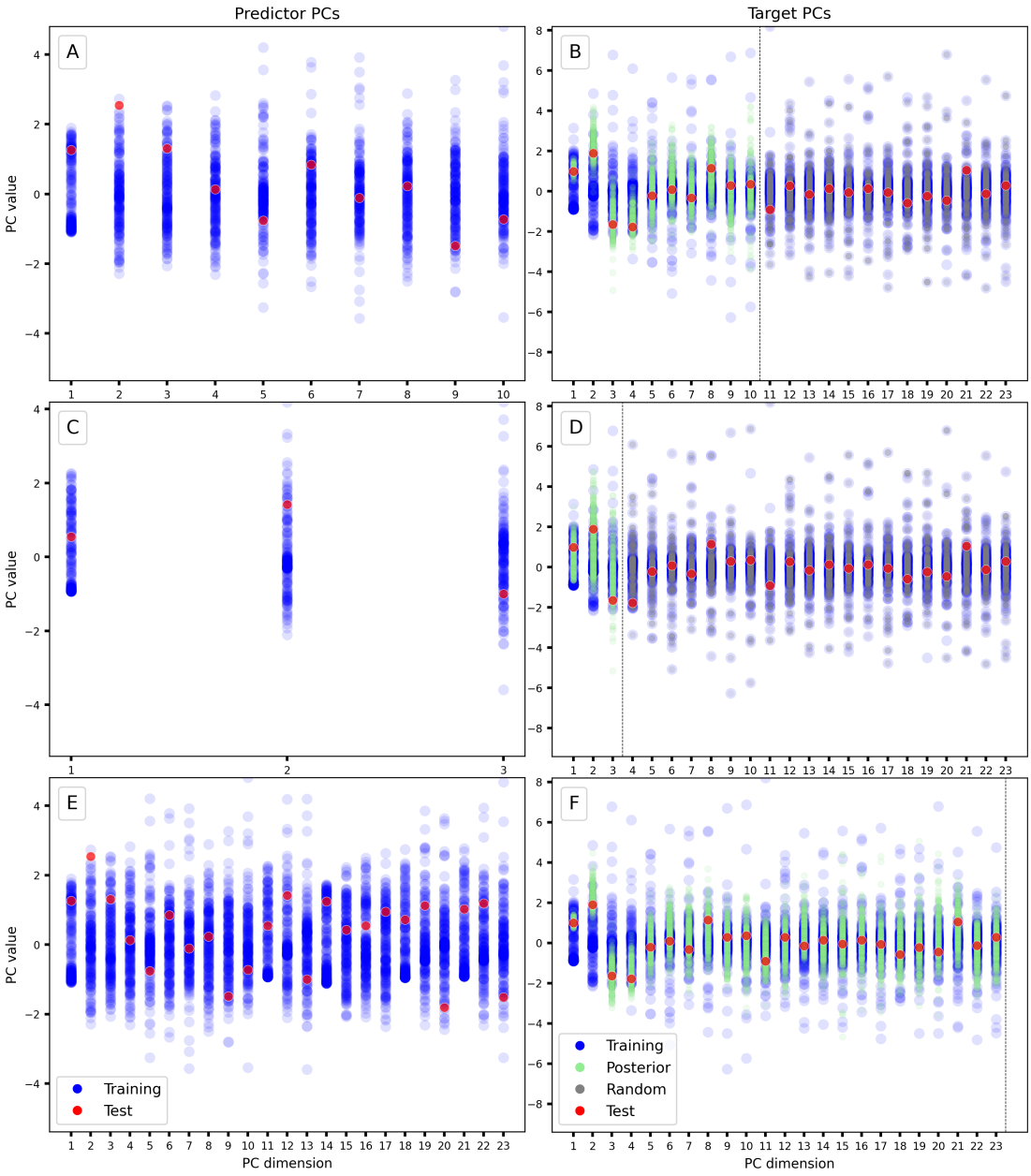

Figure 4.3: Principal Component Scores. The predicted values are distinguished from the remaining part by the vertical line in the target PC plots. The 'Random' PC samples are drawn at random from the target PC training set located to the right of the separating line. They will be used in the BOED §4.4. **A**. Case (i). Predictor: ERT data. **B**. Case (i). Target. **C**. Case (ii). Predictor: Temperature profile from borehole 1. **D**. Case (ii). Target. **E**. Case (iii). Predictor: Full combination (ERT data + four boreholes temperature profiles). **F**. Case (iii). Target.

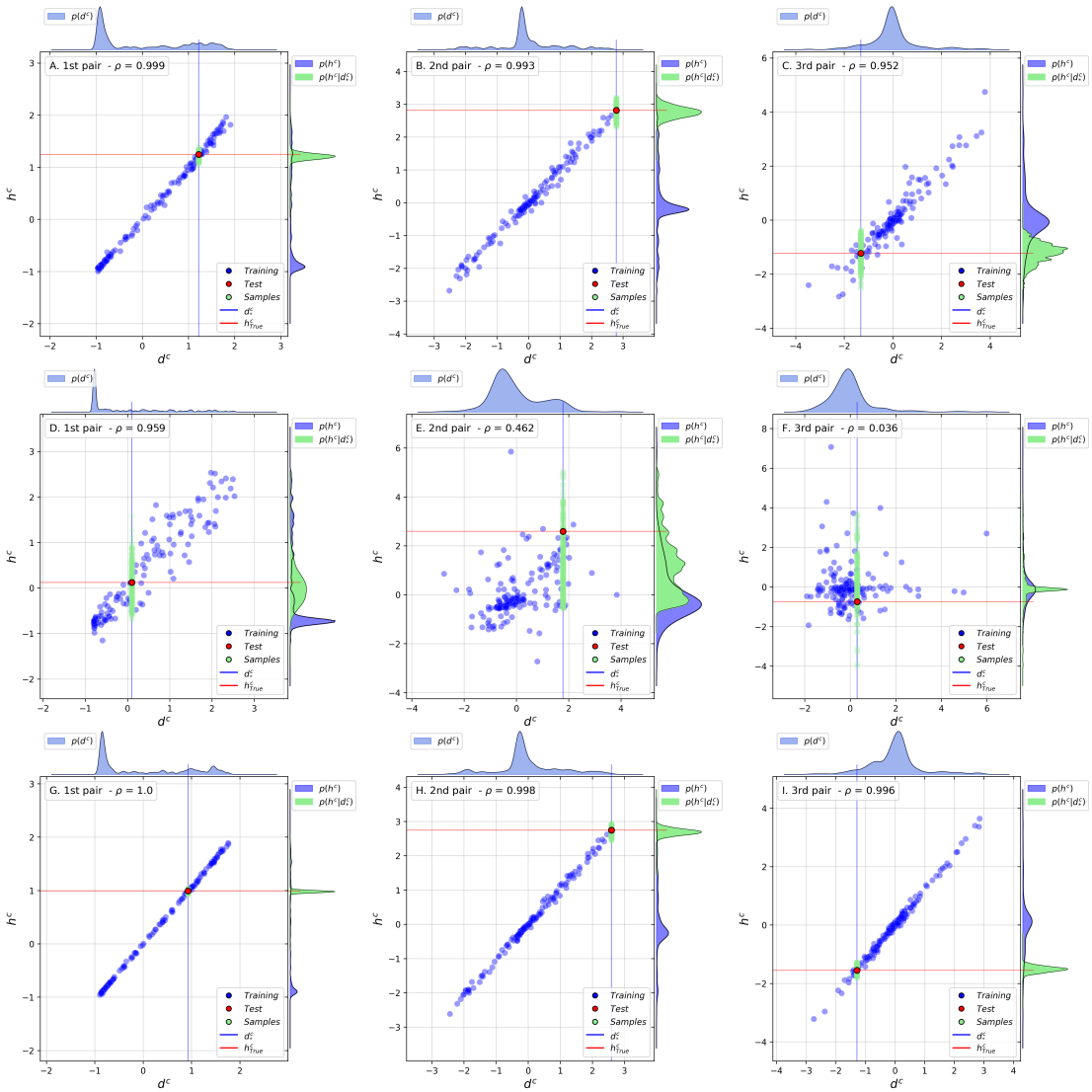

Figure 4.4: Canonical Variate pairs (1 to 3). The first row (**A**, **B**, **C**), case (i): uses the geophysical predictor, the second row (**D**, **E**, **F**), case (ii): uses the borehole predictor, and the third row (G, H, I), case (iii): uses both predictors. The true point coordinates (Test) are highlighted by the two lines in each dimension.

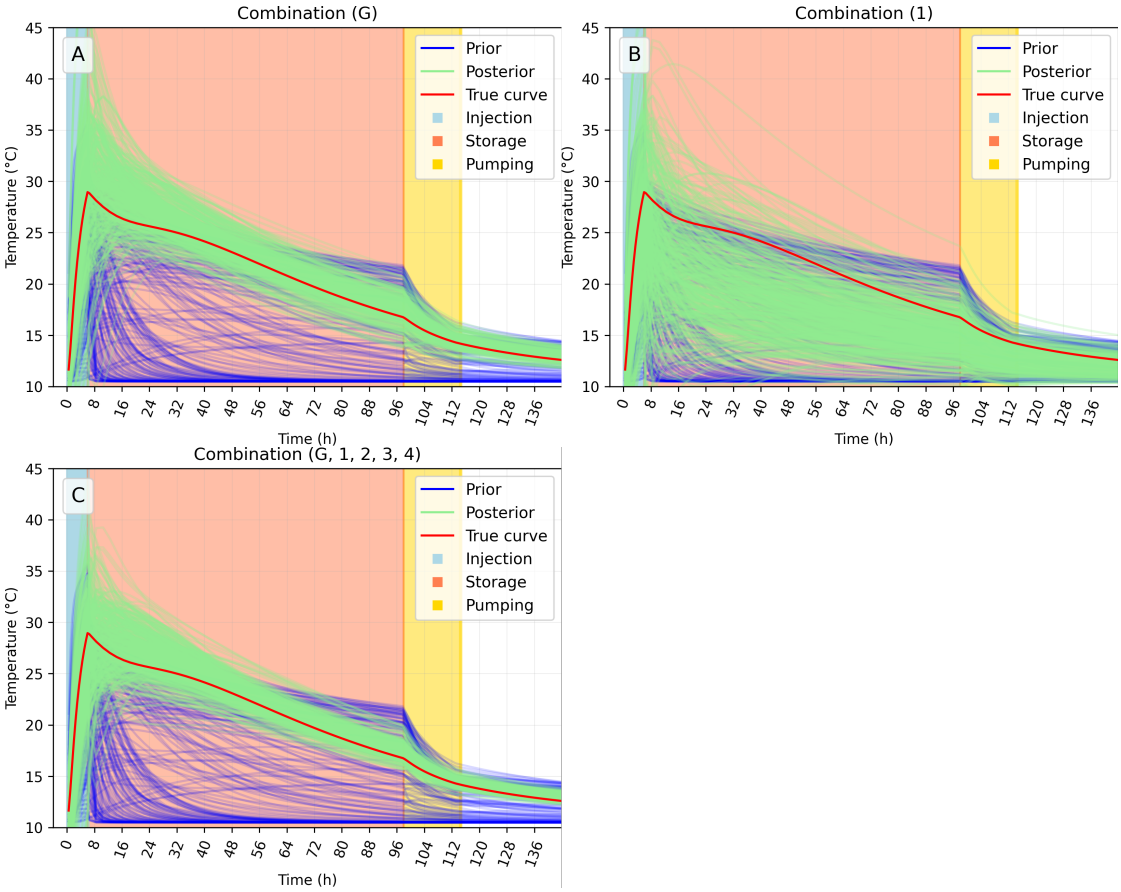

Figure 4.5: Temperature curves across all time steps, at the observation well 2. **A**. Case (i). Predictor: ERT data. **B**. Case (ii). Predictor: Temperature profile from borehole 1. **C**. Case (iii). Predictor: Full combination (ERT data + four boreholes temperature profiles).

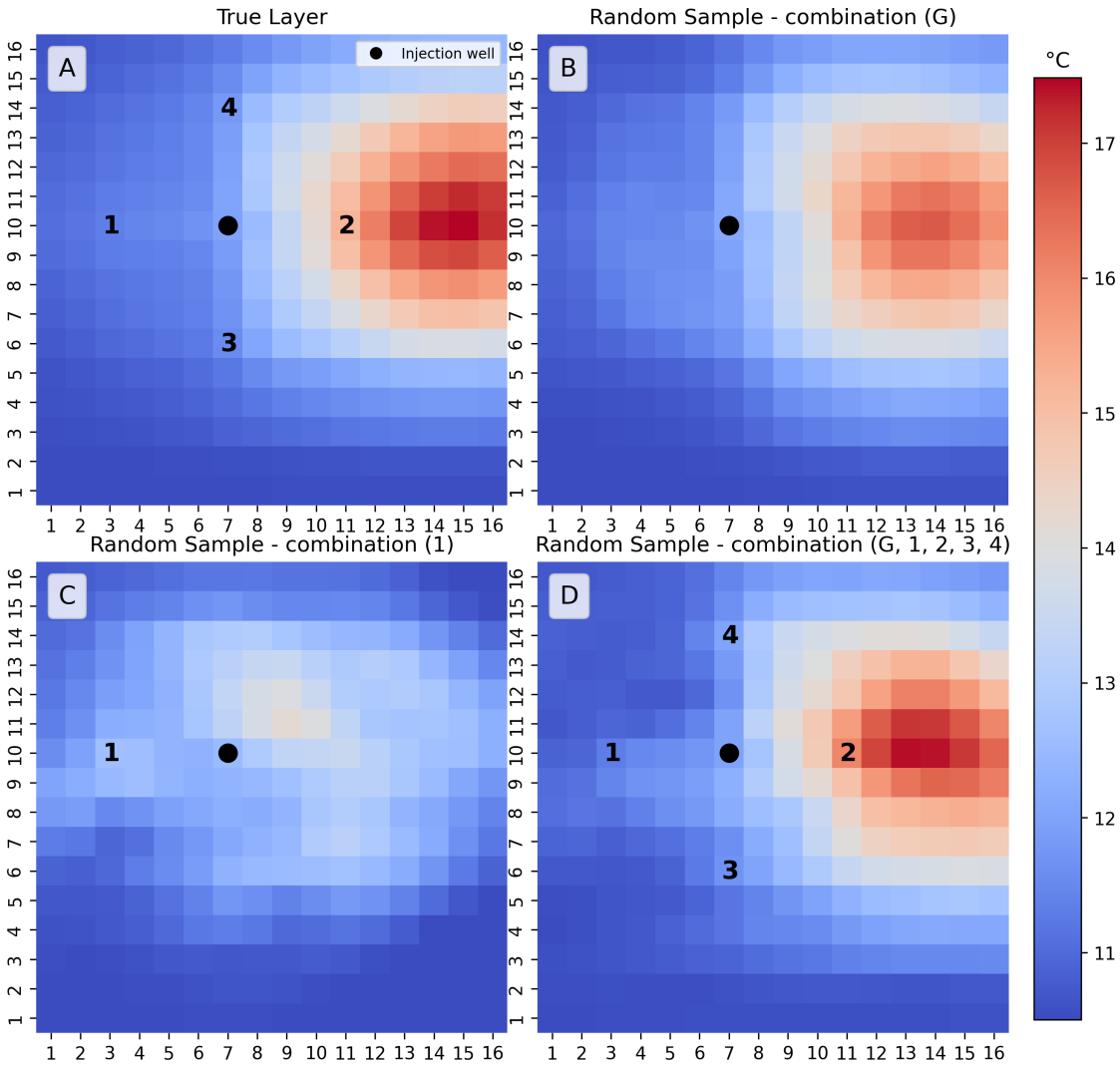

Figure 4.6: Cross-section of one predicted temperature field at time step 74 (105.75 hours–pumping phase) and layer 9–heat injection level. **A**. Ground truth. **B**. Case (i). Predictor: ERT data. **C**. Case (ii). Predictor: Temperature profile from borehole 1. **D**. Case (iii). Predictor: Full combination (ERT data + four boreholes temperature profiles).

4.4 Results

4.4.1 Optimal protocol determination

Before determining the optimal sensor combination, we can determine the optimal protocol between MG and DD using our BOED methodology. Common approaches to optimizing electrode arrays for ERT either seek the best measurement configuration on a given set of electrodes (e.g., Wilkinson et al. 2015) or select electrode locations

(e.g., Wagner et al. 2015). Both strategies are based on the resolution matrix. Uhlemann et al. (2018) propose a new approach that combines these two strategies by introducing an additional weight that penalizes the addition of electrode locations to the optimized set. Qiang et al. (2022) take a different approach. Instead of focusing on image resolution, they propose that the ERT survey be optimized for a specific target of interest by maximizing the information gained from a target area using Bayesian experimental design.

In contrast to these methods, we seek to determine the optimal protocol for a given set of electrodes, and not the electrode locations or quadrupole configurations themselves. It could, however, be extended to these uses. Furthermore, previous approaches use the resolution matrix, whereas our approach uses the predictive power of the posterior distribution in a low-dimensional latent space. Qiang et al. (2022) also makes use of the posterior distribution's predictive power, but in a physical, high-dimensional space. Lastly, these approaches are limited to a single target of interest, whereas our machine learning approach can generalize to unseen targets if the prior distribution is representative of the true distribution.

We consider four settings:

1. The MG array (848 quadrupole measures), to which PCA is applied to explain 99% of the variance.
2. The DD array (1100 quadrupole measures), to which PCA is applied to explain 99% of the variance.
3. Combination 1: concatenation of the MG and DD arrays to which PCA is applied to explain 99% of the variance.
4. Combination 2: union of the previously calculated principal components of the MG and DD configurations. Thus, it is equivalent to joining the principal components from items 1 and 2.

Combinations 1 & 2 are not among the available options for continuing the study, but they are included here for completeness. The computation is performed on the 50 test cases over 5 different folds for each of the four cases, and the averaged results are shown in Figure 4.7. The MG array, despite having fewer quadrupoles than the DD array, is the best single protocol. In fact, of the four cases, the DD array has the lowest predictive power. Therefore, the MG array is selected as the optimal protocol for the study. Interestingly, Combination 1 has a lower predictive power than the MG array, although it contains more quadrupoles. The MG and DD configurations are not equivalent and produce different outcomes. As a result, when using PCA, joining the two configurations in this manner resulted in information loss. Combination 2, however, significantly improves the predictive power over the individual cases. It is explained by the fact that the two configurations are complementary and that the union of the previously calculated principal components of the two configurations captures the most

meaningful information for the problem. The posterior distribution is better constrained as a result of the additional information, improving predictive ability.

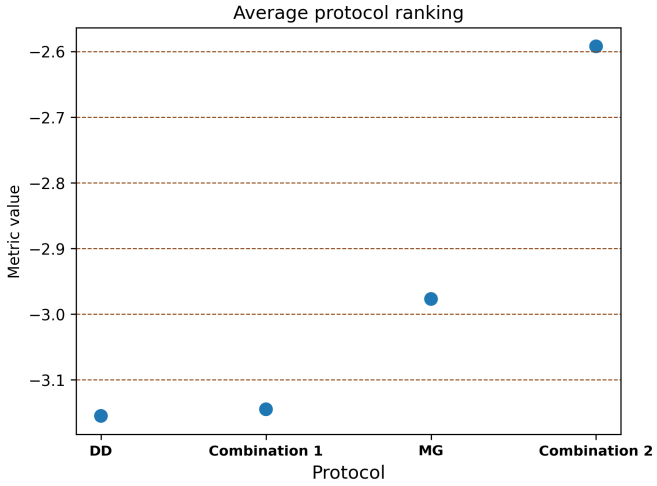

Figure 4.7: Average RMSE of the different protocols over 5 folds. The metric values opposite are displayed to show a higher score for the best combination. **DD**: PCA(Dipole-Dipole), **MG**: PCA(Multiple Gradient). **Combination 1**: PCA(Dipole-Dipole + Multiple Gradient), **Combination 2**: PCA(Dipole-Dipole) + PCA(Multiple Gradient).

4.4.2 Optimal sensor combination

To compare the different methods fairly, the same number of components must be used in each case. We therefore retained the highest number of components predicted within the combinations (iii). For combinations predicting fewer components (i and ii), the additional components are randomly sampled within the PC dimensions (Figure 4.3). The expected outcome is a ranking of combinations ranging from a small number of combinations to all considered data sources, with the most unfavourable case being the use of a single observation well and the most favourable case being the use of geophysical data and all available wells, as can be visually evaluated across cases on Figure 4.5.

To derive a robust BOED, we average the results over the test set of ground truth models. However, the results depend on the chosen ground truth models set, so a combination that is ranked as the best in one case may not be ranked as such for another set, which is why we use a 5-fold cross-validation to validate our findings. We use k-fold cross-validation to produce predictions over 5 different training and test sets, and we average the rankings across folds to obtain the final ranking. Figure 4.8A depicts the default configuration of wells. Figure 4.8B depicts the average results of all 31 combinations over the 5 folds for the default configuration of wells. The geophysical data is labeled as ‘G’ and the well data are labeled by their well ID (1, 2, 3, 4). For example, ‘G12’ is the combination of geophysical data with wells 1 and 2. The results are consistent with what was expected, and using more data sources yields the best results, which is logical

given the symmetry of our observation wells network. Our findings show that the ERT data alone provides the most information to the model, with a clear difference in metric value between the wells alone and the ERT data. To highlight that there is no overlap between the two, a darker background colour indicates when the ERT data is used, and a lighter background colour indicates when the wells alone are used. Our results corroborate previous findings that ERT is a valuable tool for monitoring the development of thermally affected zones in aquifers Hermans et al. (2016); Lesparre et al. (2019).

In terms of observation wells, wells 1, 3 and 4 are consistently ranked as the least favorable cases, and well 2 is ranked as the most favorable case. This is due to the direction of the heat plume over time. As shown on Figures 4.2, 4.5 and 4.6, the natural gradient causes the plume to move downstream, allowing well 2 to record more temperature variation from the plume over time. However, because the plume is moving away from them, wells 1, 3 and 4 are more likely to record redundant data from it (e.g., a flat temperature curve), which makes them less relevant for model training.

Similar observations were made in the case of solute transport in Thibaut et al. (2021b): downstream wells provided the most useful data. They did not investigate the information gained from any geophysical data, which in our case is sufficient to understand the evolution of the heat plume, but it is important to note that the overall information gain from the combination of all wells and ERT is the highest. Since ERT is an indirect data source, adding at least one borehole allows to provide direct information on the temperature and thus to reduce the uncertainty.

To corroborate our findings, we reproduce the experiment using four different well locations, but with the same data (the previous k-fold shuffle seed is reused). The alternative well positions are shown in Figure 4.8C and the results are shown in Figure 4.8D. The outcomes are in line with expectations. The worst-case scenario is well number 3. It is located upstream and is less affected by the heat plume. Because of its proximity to the heat injection, well number 1 is the best case. The combination of the various wells produces more insightful results than before. Wells 1 and 2 cover both upstream and downstream areas, and are far apart. As a group, they collect more data from the heat plume than the other group of two observation wells. Following the same logic, the combination of wells 1, 2, and 4 ('124') is the best case. Because alternative well locations are more dispersed and provide less redundant information, the increase in information provided by geophysical data is not as strong as in the previous case. In the case of alternative well locations, the full well combination '1234' is not the best case (Figure 4.8D). As described in §4.3, feeding to the training algorithm an array of concatenated data from various sources, some of which provide redundant information, can reduce the model's prediction power. Let us also note that, although the combination of wells '123' and '1234' contains either the same number of PCs as the ERT data alone, or even more (Table 4.2), their predictive power is lower. The amount of explained variance alone is not a reliable indicator of a data source's predictive ability.

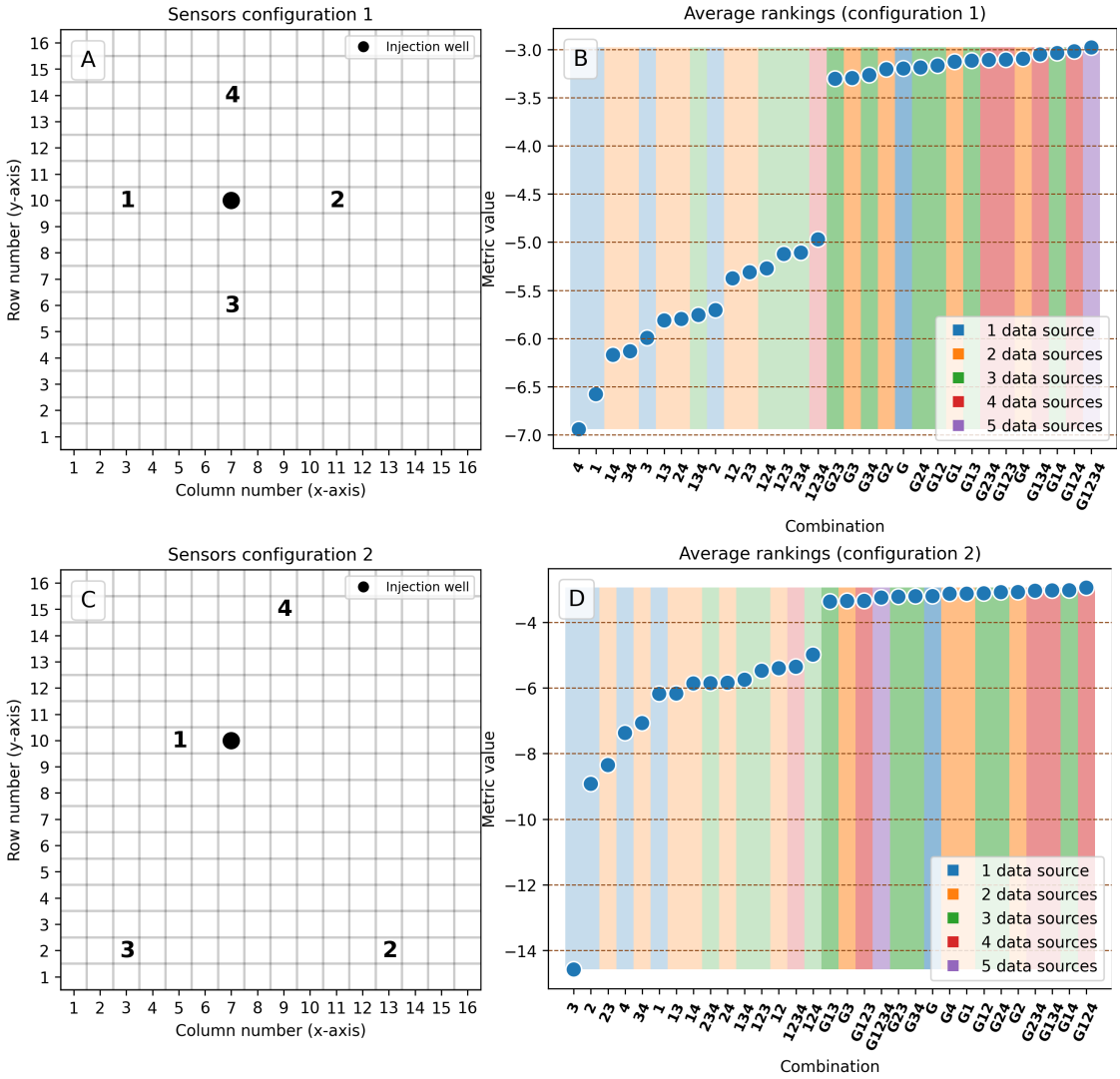

Figure 4.8: Sensors default and alternative configurations and ranking of the different combinations of data sources. To visualize a higher score for the best combination, the metric (RMSE) values opposite are displayed. The use of ERT data (G) is indicated by a darker background shade, whereas the use of wells alone is indicated by a lighter shade. **A**. Default well locations. **B**. Average ranking of 5 folds for the default well locations. **C**. Alternative well locations. **D**. Average ranking of 5 folds for the alternative well locations.

4.5 Conclusion

In this chapter, we present a method for optimizing 4D temperature field monitoring experiments using a Bayesian approach. The proposed methodology uses a combination of OED and Bayesian inference to identify informative observation well locations. We apply our method to a synthetic case study involving the prediction of a four-dimensional

temperature field from data collected by electrical resistivity tomography (ERT) and four observation wells recording the temperature over time. These predictors of different nature are combined in the principal component space to form a new predictor. Using our method, the optimal ERT electrode configuration is determined prior to optimization, and we found that, in our case, the multiple-gradient array outperforms the dipole-dipole array in terms of information gain. Following the training step, targets are sampled from the inferred posterior distribution in a low-dimensional latent space using transport map methods, which are a powerful tool in our Bayesian inference framework for sampling an unknown target given a known predictor. Applying a simple metric (RMSE) to the principal components of the predicted and true targets allows to determine the locations of observation wells that minimize uncertainty. This method can be used to optimize the design of 4D temperature field monitoring experiments and to reduce the cost of data collection by choosing a threshold between precision and number and nature of the data.

Our findings indicate that the placement of observation wells must take into account the direction of the heat plume, and that an observation network that includes wells both downstream and upstream of the original injection well is optimal. A combination of various observation wells and geophysical data always yields the best results. We also noticed that increasing the number of observation wells does not always improve prediction accuracy if the wells are not placed in the proper locations and thus provide redundant or no information. Careful consideration of the information provided by the different data types and how they should be combined is necessary prior to optimization. Our method provides a straightforward procedure for examining these trade-offs, and is thus a useful technique for determining the optimal sensor combination.

One of our framework's limitations is that it requires prior knowledge of the unknown model parameters as well as the requirement to generate training data through forward modeling, which can be computationally expensive. However, with a small number of examples (200 for training and 50 for testing), the method is able to predict the posterior distribution of the temperature field with a reasonable accuracy and to find the best combination of data sources among four prescribed well positions and their combinations, with the caveat that the training set should be as representative as possible of the real data. As a more general drawback, OED can not guarantee assumptions such as experiment effectiveness, model form validity, and criterion reflecting experiment objectives. To satisfy the aforementioned assumptions, OED must carefully consider the experiment (Smucker et al., 2018).

The proposed method is not limited to the monitoring of temperature fields and can be applied to the prediction of any type of high-dimensional target from a fusion of multiple predictors in a wide range of contexts, such as geophysical, environmental, and engineering applications.

5. Sequential optimization of temperature measurements to estimate groundwater-surface water interactions

CRediT author statement. Robin Thibaut: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Ty Ferré**: Conceptualization, Methodology, Supervision. **Eric Laloy**: Writing - Review & Editing, Supervision. **Thomas Hermans**: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Abstract

The groundwater-surface water (GW-SW) exchange fluxes are driven by a complex interplay of subsurface processes and their interactions with surface hydrology, which have a significant impact on the water and contaminant exchanges. Due to the complexity of these systems, the accurate estimation of GW-SW fluxes is important for quantitative hydrological studies and should be based on relevant data and careful experimental design. Therefore, the effective design of monitoring networks that can identify relevant subsurface information are essential for the optimal protection of our water resources. In this study, we present novel deep learning (DL)-driven approaches for sequential and static Bayesian optimal experimental design (BOED) in the subsurface, with the goal of estimating the GW-SW exchange fluxes from a set of temperature measurements. We apply probabilistic Bayesian neural networks (PBNN) to conditional density estimation (CDE) within a BOED framework, and the predictive performance of the PBNN-based CDE model is evaluated by a custom objective function based on the Kullback-Leibler divergence to determine optimal temperature sensor locations utilizing the information gain provided by the measurements. This evaluation is used to determine the optimal sequential sampling strategy for estimating GW-SW exchange fluxes in the 1D case, and the results are compared to the static optimal sampling strategy for a 3D conceptual riverbed-aquifer model based on a real case study. Our results indicate that probabilistic DL is an effective method for estimating GW-SW fluxes from temperature data and designing efficient monitoring networks. Our proposed framework can be applied to other cases involving surface or subsurface monitoring and experimental design.

Plain language summary

Groundwater-surface water exchange flux occurs when water flows from the ground to a river or vice versa. Understanding and quantifying these exchange processes is essential for safeguarding our water resources. Temperature measurement using sensors such as thermistors or fiber-optic cables is one of the most popular techniques for quantifying exchange fluxes. However, designing a monitoring network is difficult due to the numerous factors that must be taken into account. We present a novel method for estimating groundwater-surface water exchange flux from temperature data using deep learning and optimal experimental design. We show how deep learning can be used to estimate fluxes as well as design efficient and effective monitoring networks. Our findings indicate that deep learning is an effective method for estimating groundwater-surface water fluxes from temperature measurements.

Key points

- We propose a novel method for estimating the groundwater-surface water exchange flux from temperature data using deep learning.
- We introduce a novel deep learning-based Bayesian optimal experimental design framework for sequential and static optimal experimental design.
- We use a custom objective function to determine optimal temperature sensor locations based on sensor information gain.

5.1 Introduction

Methods for quantifying groundwater-surface water (GW-SW) exchange flux have long been of interest in order to better understand water and solute exchange across the sediment-water interface as well as the magnitude, spatial distribution, and temporal dynamics of fluxes between subsurface and surface compartments of the Earth (Hermans et al., 2022). Therefore, understanding and quantifying the exchange processes between rivers and groundwater is crucial for the best possible protection of our water resources: the provision of drinking water, the characterization and management of environmental flow regimes, the preservation or restoration of riverine ecosystem health and functions, the attenuation of contaminants and water storage in aquifers are some of the challenges that are impacted by processes that occur at the GW-SW interface (Dujardin et al., 2014; Ghysels et al., 2021; Hermans et al., 2022; Irvine et al., 2016; Kikuchi and Ferré, 2017; Kurylyk et al., 2019; Moghaddam et al., 2022).

Water flux has traditionally been measured using temperature tracers, shallow piezometers, tracer experiments, and differential-discharge measurements. For a comprehensive review of the various types of water flux measurements and their outlooks, see Hermans et al. (2022). Because GW and SW typically have different temperatures and advective

heat transport is the primary factor influencing subsurface temperature in riparian areas, heat is a useful natural tracer for describing GW-SW exchange patterns (Kikuchi and Ferré, 2017). Additionally, measuring temperature with conventional sensors such as thermistors or fiber-optic cables (Hare et al., 2015; Mamer and Lowry, 2013) is relatively simple and accurate (Irvine et al., 2016). As a result of their practical benefits and ability to provide flux estimates with varying resolutions, these thermal techniques have grown in popularity among alternative techniques (Anderson, 2005; Moghaddam et al., 2022).

Since the 1960s, methods for estimating vertical groundwater fluxes from one-dimensional vertical temperature profiles have been available (e.g., Bredehoeft and Papaopulos 1965; Stallman 1963, 1965b; Suzuki 1960). To that end, steady state 1D analytical solutions (Bredehoeft and Papaopulos, 1965) are a natural choice for estimating groundwater fluxes from temperature profiles because they are simple to apply and do not require any complex boundary or initial conditions (Irvine et al., 2016).

Several uncertainties and assumptions influence the estimation of exchange fluxes. One assumption is that vertical flows through the media are the primary factor influencing exchange fluxes (e.g., between river, riverbed and aquifer). In other words, it is assumed that horizontal exchange fluxes are negligible. Water will, however, flow horizontally through the media via the sediment layers, and the resulting horizontal groundwater fluxes should arguably be factored into the calculation of water flux across the riverbed (Dujardin et al., 2014). Ghysels et al. (2021) compared vertical flux estimates obtained with a 1D analytical solution to the heat transport equation with fluxes simulated with a 3D groundwater model, and compared the simulated fluxes to a real-world case study. Ghysels et al. (2021) found that the estimate based on the 3D groundwater model was roughly twice as large as the estimate based on the 1D solution for the total exchange flux between a simulated river and shallow aquifer, due to the significant contribution of non-vertical flows, particularly through riverbanks.

During the last two decades, several methods have been developed to estimate water fluxes from temperature data, as reviewed by Anderson (2005); Constantz (2008); Irvine et al. (2017); Rau et al. (2014). Recent advances in machine learning (ML) in the field of hydrology (e.g., Nearing et al. 2021) have enabled novel approaches to estimate fluxes from temperature data, as reviewed by Moghaddam et al. (2022), who also applied and compared various state-of-the-art ML algorithms for this purpose. Additionally, Moghaddam et al. (2022) examined approaches to extract information from ML methods that can be used to design efficient and effective monitoring networks.

When working with data that must be collected through an experimental process, as is often the case in geoscience, one of the most important considerations is where to gather new observations: designing a monitoring network necessitates careful consideration of many factors, especially when measurements are costly or resources are limited (Moghaddam et al., 2022; Thibaut et al., 2022). Controlling the experimental conditions for data acquisition is essential to maximize resource utilization and infor-

mation gain because this process is typically expensive and/or time-consuming (Attia et al., 2018; Vilhelmsen and Ferré, 2018). This is known as optimal experimental design (OED), and it is commonly regarded as an optimization problem (Ryan et al., 2016). Using the Bayesian paradigm to define the objective function in OED results in a so-called Bayesian OED (BOED) approach. Bayesian statistics generate the posterior distribution by combining prior knowledge of the model's unknown parameters with likelihood—the data's contribution to those parameters—from which conclusions about the model's unknown parameters can be drawn (Chapter 4). The BOED strategies can be divided into static and sequential strategies (Eidsvik et al., 2018): static BOED selects the optimal design when all experiments are conducted at once; sequential BOED, on the other hand, takes into account the fact that some experiments will be conducted in the very near future and others at later stages; the difference is that the latter takes into account the information already obtained from earlier experiments.

In recent years, applications of the Bayesian paradigm to ML have been intensively studied in response to the growing need for principled uncertainty reasoning in machine learning systems as they are progressively implemented in safety-critical domains (Meinert et al., 2022). This approach has the advantage of allowing for the quantification of the uncertainty of the model's parameter estimates. Uncertainties can be classified into two categories: aleatoric and epistemic uncertainties. Typically, aleatory uncertainty is attributed to data noise, whereas epistemic uncertainty is attributed to model parameter and model structure uncertainty (Gal, 2016; Kiureghian and Ditlevsen, 2009; Nearing et al., 2016). Epistemic uncertainty can be reduced by increasing the amount of data available for training, but aleatoric uncertainty can only be reduced by using higher precision sensors, for instance (Gal, 2016).

Quantifying and estimating uncertainty is critical for putting GW-SW interactions predictions to use because it quantifies the risk of making a poor decision by gauging our confidence in the estimate. Conversely, due to a lack of data and the complex non-linear relationships between the various processes governing GW-SW interactions, quantifying uncertainty in GW-SW fluxes is challenging. Therefore, the BOED strategy necessitates the selection of a suitable method for conditional density estimation (CDE). CDE seeks to capture the statistical relationship between a conditional variable $\mathbf{d}$ and a dependent variable $\mathbf{h}$ by modeling their conditional probability $p(\mathbf{h}|\mathbf{d})$ given a set of empirical observations $\mathbf{d} = d_1, d_2, \ldots, d_N$ and $\mathbf{h} = h_1, h_2, \ldots, h_N$, where N is the number of observations (Rothfuss et al., 2019). The posterior distribution $p(\mathbf{h}|\mathbf{d})$ covers all possible values of the unknown model parameters, given data and prior information, and the probability of each value is interpreted as the model's "credibility"; for example, a posterior mode is the most likely value. To that end, we propose a novel method for estimating GW-SW fluxes that combines two of the most advanced approaches for uncertainty quantification and estimation in geoscience: Bayesian OED and deep-learning (DL). Specifically, we investigate the use of DL in various settings to estimate water fluxes $\mathbf{h}$ from temperature data $\mathbf{d}$ within a BOED framework.

DL is a subset of ML based on Artificial Neural Networks (ANNs) that can learn

complex non-linear transformation; consequently, ANNs are frequently used to solve regression problems (Goodfellow et al., 2016), and have become a popular tool for supervised learning applications, where they are typically used to obtain an optimal mapping between the input ($\mathbf{d}$) and output ($\mathbf{h}$) variables (LeCun et al., 2015). ANNs are parameterized by a set of parameters θ, which are determined by training. Backpropagation is commonly used in training to adjust the ANN parameters until the loss function (e.g., mean squared error) reaches a minimum (Goodfellow et al., 2016). The training process requires the selection of an appropriate training algorithm and hyperparameters, such as learning rate, optimizer, and batch size. The learning rate determines the step size for adjusting the parameters θ during training. The optimizer is an algorithm for finding the optimal set of parameters θ; popular optimizers include stochastic gradient descent (SGD) and Adam (Kingma and Welling, 2014). It requires the choice of a batch size, which is the number of training samples used in one iteration of the algorithm.

Traditional DL techniques, including ANNs, are deterministic; they do not provide any measure of prediction uncertainty, which is required for BOED applications (Kingston et al., 2005). To address this limitation, we investigate the use of what we will refer to as Probabilistic Bayesian Neural Networks (PBNNs) for CDE. The term "probabilistic" refers to a wide range of neural network outputs that are parametric probability density functions instead of point estimates, which is also referred to as "neural density estimation" (e.g., Alsing et al. 2019; Papamakarios 2019). Mixture Density Networks (MDNs; Bishop 1994) are a well-known example among DL practitioners that fits our definition and are a natural choice for CDE. The term "Bayesian" refers to the use of a prior distribution over the network parameters θ, which is updated with the data to obtain a posterior distribution over the network parameters (Gal, 2016). BNNs leverage Bayesian methods to learn probability distributions over the ANN parameters and to quantify uncertainty in predictions (Kendall and Gal, 2017; MacKay, 1992; Neal, 1996), making them a promising solution for applying DL in situations where it is not permitted for a system to make inaccurate predictions without warning (Jospin et al., 2022). BNNs allow aleatory and epistemic forms of uncertainty to be captured in the model outputs, which makes BNNs "data-efficient" as they can learn from a small dataset without overfitting (Depeweg et al., 2018; Jospin et al., 2022). This paradigm offers a rigorous framework for analyzing and training uncertainty-aware neural networks (Jospin et al., 2022), and provides a principled way to integrate prior beliefs into DL models (Khan and Coulibaly, 2006).

In this contribution, we posit the question: *"How do the number, depth and sequential placement of sensors affect the accuracy of flux estimation?"* To answer this question, we begin with a synthetic dataset of 1D temperature profiles with depth and fluxes and use it to demonstrate the suitability of PBNNs for CDE and sequential BOED by deriving optimal temperature sensor locations using the sensors' information gain using a custom objective function. The 1D flux estimation problem is ideally suited for testing the newly proposed framework because it enables fast forward modeling with analytic solutions. Ghysels et al. (2021)′s 3D conceptual riverbed-aquifer model is then reused and adapted to simulate steady-state heat transport, providing us with a well-

constrained model setup based on a real-world case study. We use the optimal sensor locations derived from the 1D synthetic dataset to design a monitoring network for the real-world case study through static BOED, and we compare the 3D flux estimation to the 1D vertical flux estimation.

Previous applications of ML-driven BOED to derive optimal monitoring networks from simulated datasets include Chapters 3 and 4, which estimate the wellhead protection area of a groundwater well and to monitor underground heat flow, respectively. In both approaches, conditional density estimation (CDE) was performed in the Canonical Correlation Analysis (CCA)-derived latent-space using techniques for conditional density estimation, such as an analytical approach (Chapter 3) and a *transport method* approach (e.g., Villani (2009), Chapter 4). In Earth sciences, this framework is frequently referred to as Bayesian Evidential Learning (BEL; Scheidt et al. 2018).

In contrast, the main difference between our work and the above-mentioned contributions is that we demonstrate how to apply DL to CDE within a sequential BOED framework. Our approach also belongs within the BEL framework, albeit using PBNNs instead of CCA. Our method is among the first to employ DL for CDE in the context of BOED for estimating groundwater-surface water exchange fluxes.

In hydrology and hydrogeology, the use of BNNs is well established. BNNs have been amended to account for various forms of data uncertainty, but only for point predictions rather than probability density estimates (e.g., Humphrey et al. 2016; Khan and Coulibaly 2006; Zhang et al. 2011; Zhang and Zhao 2012). The use of PNNs, on the other hand, has been relatively scarce in hydrology, except for rainfall-runoff modeling (e.g., Carreau et al. 2009; Carreau and Vrac 2011; Klotz et al. 2022). Notably, Klotz et al. (2022) recently demonstrated the benefits of MDNs and variants for quantifying rainfall-runoff prediction uncertainty by connecting them to a Long Short-Term Memory (LSTM) network.

This is the first time in the context of BOED that probabilistic BNNs have been used for CDE in subsurface hydrology. Furthermore, unlike the previous 1D rainfall-runoff modeling applications, our target variable is multidimensional, so defining our objectives, priors, and conditions is not trivial.

The novelties of this work are threefold:

1. We demonstrate the application of DL to probability density estimation (CDE) in the context of a BOED framework, and also in cases where classical BEL with CCA wouldn't work, achieving better predictions and estimates of uncertainty.

2. We develop a custom objective function to derive optimal temperature sensor locations using the sensors' information gain.

3. We compare the performance of PBNNs for flux estimation using a 1D synthetic dataset and a 3D conceptual riverbed-aquifer model based on a real-world case

study.

Our goal is to demonstrate the potential of this approach for ensuring efficient and effective monitoring trials for characterizing GW-SW exchange fluxes, and to provide a roadmap for future research.

5.2 Methodology

The methodology is presented as follows: (i) we first start by presenting the BOED framework for each case (1D and 3D), (ii) we then introduce the PBNNs, (iii) we present the experimental setup and physical models parameterization, and finally (iv), we explain how the model outputs are pre-processed.

5.2.1 Bayesian optimal experimental design

1D case - sequential design (seqBOED)

Aims. In the 1D case, the objective is to sequentially find a number ω of optimal depths to sample the predictor variable $\mathbf{d}$ (temperature measurements) that minimize our objective function $\mathcal{J}(\mathbf{h})$. $\mathbf{h}$ is the target variable, consisting of the parameters of the physical model: the vertical flux q_v, the thermal conductivity λ, and the bottom temperature T_b. The sequentiality implies that we update the dataset with samples from the optimal solution of the previous iteration, such that $\mathcal{D}_i = \{\mathbf{d}_{i-1}, \mathbf{h}_{i-1}\}^N$, in order to converge towards the most accurate and precise prediction of the target variable $\mathbf{h}$. The number of features in the predictor variable $\mathbf{d}$ will increase at each iteration, but the number of samples N will remain constant.

To maintain tractability, we restrict the sampling space $\mathcal{Z}$ to the surface ($0\,m$) plus 10 depths evenly spaced by $20\,cm$ between 0.1 and $1.9\,m$, which are the depths at which the temperature can be measured:

$$\mathcal{Z}_{init} = \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9\}\, m$$

To maintain realism, we assume that we begin measuring from the riverbed's surface and that the temperature is always measured at the surface ($z = 0.0\,m$) and at $10\,cm$ depth ($z = 0.1\,m$). Therefore, the initial ($i = 1$) predictor is composed of two data points, $\mathbf{d}_{i=1} = \{d_{z(0.0)}, d_{z(0.1)}\}^N$ and the sampling space is reduced to $\mathcal{Z}_{i=1} = \{0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9\}\, m$. Lastly, we limit the number of additional depths for samples to three, bringing the total number of samples ω to five, which is a reasonable number of samples to collect for a single experiment.

Remark. In this problem, we have the exact analytical solution of the temperature curve, which is known from the heat equation and boundary conditions. The solution is simple, and therefore the advantage of using BOED is maybe not obvious. However, the idea is to show the potential of the BOED technique for more complex problems, where the forward model is expensive, and the analytical solution is not known. We also

note that the problem is not as simple as it initially appears due to the fact that the target variable $\mathbf{h}$ is multidimensional: it comprises three features with distinct physical meanings and is not necessarily Gaussian distributed. Moreover, the predictor consists solely of point temperature measurements that are not necessarily Gaussian distributed either.

Implementation. To begin the sequential optimization process, we first split the N-sized dataset $\mathcal{D}$ into a training and a test set using a k-fold cross-validation technique, with $k = 5$, such that the training set is composed of $N_{80\%}$ samples and the test set is composed of $N_{20\%}$ samples. The training set is used to train the PBNN, while the test set is used to determine the optimal sampling sequence of depths: an optimal sequence is determined for each test set $\{\mathbf{d}_j^*, \mathbf{h}_j^*\}$, where $j = 1, \ldots, N_{20\%}$. The objective is to validate the consistency of the optimal sequence of depths across various initial data sets and test sets (e.g., Chapters 3 and 4).

At fold number k, iteration $i = 2, \ldots, \omega$, and given a trained PBNN $f_{\theta,\,i}$ over the $N_{80\%}$-sized dataset $\mathcal{D}_{k,\,i}$, let's consider an observed predictor $\mathbf{d}_j^*$ and its corresponding target $\mathbf{h}_j^*$. The trained network $f_{\theta,\,i}$ is then used to predict the posterior distribution of the target variable: $f_{\theta,\,i}(\mathbf{d}_j^*) = p(\mathbf{h}_j^*|\mathbf{d}_j^*)$, and the objective function $\mathcal{J}(\mathbf{h}_j^*)$ is used to evaluate the score of the predicted distribution $p(\mathbf{h}_j^*|\mathbf{d}_j^*)$.

Given an "ideal" multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where μ and Σ are the mean and covariance of the ideal distribution, respectively, the objective function is defined as the KL divergence between the ideal and the predicted distribution $p(\mathbf{h}_j^*|\mathbf{d}_j^*)$:

$$\mathcal{J}(\mathbf{h}_j^*) = D_{KL}\left[\mathcal{N}(\mathbf{h}_j^*, \mathbb{I}[\sigma])) \;||\; f_{\theta,\,i}(\mathbf{d}_j^*)\right] \tag{5.1}$$

$\mathcal{J}(\mathbf{h}_j^*)$ measures the amount of information lost when $f_{\theta,\,i}(\mathbf{d}_j^*)$ is used to approximate $\mathcal{N}(\mu, \Sigma)$ (Cover and Thomas, 2006). σ is arbitrarily chosen to be small enough, e.g., $\sigma = 0.01$. It can be viewed as the degree of precision we wish to compare the predicted distribution to. The reason for using the KL divergence is that it has desirable properties (Papamakarios, 2019). For instance, it is always non-negative and its minimum value is 0 (when $\mathcal{N}(\mathbf{h}_j^*, \mathbb{I}[\sigma])) = f_{\theta,\,i}(\mathbf{d}_j^*)$).

The optimal depth z_i^* is then selected as the depth that minimizes the objective function:

$$z_i^* = \arg\min_{z \in \mathcal{Z}_i} \mathcal{J}(\mathbf{h}_j^*) \tag{5.2}$$

where $\mathcal{Z}_i$ is the set of depths at which the temperature can be measured at iteration i. At each iteration, after a new depth z_i^* is selected, the same depth is removed from the set $\mathcal{Z}_i$, and the new set $\mathcal{Z}_{i+1}$ is defined as $\mathcal{Z}_{i+1} = \mathcal{Z}_i \setminus \{z_i^*\}$.

The posterior distribution is then evaluated with $f_{\theta,\,i}(\mathbf{d}_j^* \cup \mathbf{d}_{j,\,z_i^*}^*)) = p(\mathbf{h}_j^*|\mathbf{d}_j^* \cup \mathbf{d}_{j,\,z_i^*}^*)$, where $\mathbf{d}_{j,\,z_i^*}^*$ is the new temperature measurement at the optimal depth z_i^*. To proceed to

the next iteration, the new depth z_i^* is added to the predictor $\mathbf{d}_j^*$, and the new predictor $\mathbf{d}_{j,\,z_{i+1}^*}^*$ is used to evaluate the objective function $\mathcal{J}(\mathbf{h}_j^*)$ at the next iteration $i+1$.

Threshold-Based Rejection Sampling. After each iteration, even though information on the target has been gained, the predictor set remains unchanged. To obtain an updated set of predictors for the next iteration, a Monte Carlo approach is generally used (e.g., Laloy and Vrugt 2012), and the forward model is run several times with new parameters obtained from the inferred posterior distribution, which can be computationally expensive. The posterior distribution of the target has been calculated at the end of iteration 1 and can be used as the prior distribution for the next iteration. However, a new set of predictors that captures the new knowledge from the previous iteration must be generated. Michel et al. (2022a) demonstrated that when the prior uncertainty is large, BEL may overestimate the posterior uncertainty, resulting in samples that do not fit the data, i.e., are not fully consistent with the predictor of previous iterations. To avoid this limitation, Michel et al. (2022a) recommended to use Iterative Prior Resampling (IPR) in conjunction with classical (likelihood-based) rejecting sampling.

When direct sampling is difficult, rejection sampling is used to generate samples from a probability distribution. The fundamental idea is to generate samples from a known distribution, known as the proposal distribution, and then accept or reject each sample based on its likelihood with respect to the target distribution. For simplicity, and because the forward model is inexpensive to run, we propose to use a threshold-based rejection sampling (TBRS) method in order to compare the performance of the proposed method with and without the application of a rejection technique, as well as IPR. Acceptance/rejection in this work is determined by the magnitude of the temperature difference between the predictor generated by the set of targets predicted in the previous iteration and the true temperature measurements. The new set of predictors generated using the accepted samples will be used for the next iteration. When using TBRS, different magnitude thresholds can be used for each temperature measurement and each iteration. We can select a threshold that is close to sensor precision, which is typically in the order of 0.1 Kelvin. For example, with TBRS, if the observed predictor $\mathbf{d}_j^*$ is $17.5, 16.2, 15.4\,°C$, the posterior samples $\mathbf{h}_j^*$ will be rejected if the predicted temperature are not in the range $17.4, 16.1, 15.3\,°C$ to $17.6, 16.3, 15.5\,°C$.

In addition, to ensure that the predicted values remain physically meaningful, we add constraints to the posterior samples:

- The vertical flux must be within the range of the training dataset.
- The thermal conductivity must be within the range of the training dataset.
- The bottom temperature must be greater than 0 Kelvin.

The physical parameters range are described in Section §5.2.3. The posterior is then sampled until we obtain $\mathbf{h}_i = \{\mathbf{h}_{j,\,i}^*\}^{N_{80\%}} = \{q_{v,\,j}^*, \lambda_j^*, T_{b,\,j}^*\}^{N_{80\%}}$, and the predictor variable $\mathbf{d}_{i-1}$ is updated with the new temperature measurement $\mathbf{d}_i = \mathbf{d}_{j,\,z_i^*}^*$, which results in $\mathcal{D}_{k,\,i+1} = \{\mathbf{d}_i, \mathbf{h}_i\}^{N_{80\%}}$.

Update of the prior distribution. The updated dataset $\mathcal{D}_{k,\,i+1}$ is then fed to the 1D forward model (Equation 5.11) to get a new set of predictors at the remaining unsampled depths $\mathcal{Z}_{i+1}$, with updated q_v, λ, and T_b. Finally, the updated dataset is used to retrain the PBNN $f_{\theta,\,i}$ to better predict the target variable $\mathbf{h}$. The sequential optimization process is then repeated until the target number of depths is reached, by which time the new dataset $\mathcal{D}_k$ should have converged to a reasonably accurate prediction of the target variable $\mathbf{h}$. Note that it is not strictly necessary to generate an updated dataset of the same size as the original dataset, and the number of samples can be reduced to a smaller number, such as $N_{50\%}$, which would further reduce the computational cost of the sequential optimization process.

The optimal sampling depths z_i^*, $i = 2, \cdots, \omega$ are saved for each test case $\{\mathbf{d}_j^*, \mathbf{h}_j^*\}$ across each fold, and the resulting $(k \times N_{20\%} \times c - 2)$-sized array of optimal depths is used to represent the histogram of the optimal sampling sequence.

In practice, the seqBOED algorithm is implemented as follows:

Initialise sampling space: $\mathcal{Z}_{init} = \{0; 0.1; 0.3; 0.5; 0.7; 0.9; 1.1; 1.3; 1.5; 1.7; 1.9\}\, m$
Initialise inputs: $\mathcal{D} = \{(\mathbf{d_0};\, \mathbf{d_{0.1}}); \mathbf{h}\}^N$
Initialise $i = 1$
Update $\mathcal{Z}_i = \mathcal{Z}_{init} \setminus \{0; 0.1\}$
for all Fold k in 5-fold cross-validation **do** ▷ Split into training and test sets
    Initialise $\mathcal{D}_{train} = \{(\mathbf{d_0};\, \mathbf{d_{0.1}}); \mathbf{h}\}^{N_{80\%}}$
    Initialise $\mathcal{D}_{test} = \{(\mathbf{d_0};\, \mathbf{d_{0.1}}); \mathbf{h}\}^{N_{20\%}}$
    for $j = 1, \ldots, N_{20\%}$ **do**
        Initialise $i = 2$
        Initialise $\mathbf{d}_{i-1} = \mathbf{d}_j^* \in \mathcal{D}_{test}$
        Initialise $\mathbf{h}_{i-1} = \mathbf{h}_j^* \in \mathcal{D}_{test}$
        Initialise $scores = \{\}$ ▷ empty array to store optimal depths
        while $i \leq \omega$ **do**
            for $z \in \mathcal{Z}_i$ **do**
                Update $\mathbf{d}_i = \mathbf{d}_{i-1} \cup \mathbf{d}_{i-1,\,z}$
                Train PBNN $f_{\theta,\,i}$ on $\{\mathbf{d}_i; \mathbf{h}_{i-1}\}^{N_{80\%}}$ ▷ Train PBNN
                Update $scores.append(\mathcal{J}(\mathbf{h}_j^*))$ ▷ append the score value to the array
            end for
            Find $z_{index} \leftarrow \arg\min_{z \in \mathcal{Z}_i} \text{scores}$ ▷ find the index of the optimal depth
            Get $z_i^* \leftarrow \mathcal{Z}_i[z_{index}]$ ▷ find the optimal depth
            Update $\mathbf{d}_i = \mathbf{d}_{i-1} \cup \mathbf{d}_{i-1;\, z_i^*}$ ▷ update the dataset with the new measurement
            Update $\mathbf{h}_i = f_{\theta,\,i}(\mathbf{d}_i)$ ▷ update the target variable with the new measurement
            Update $\mathbf{d}_{i+1} = forward(\mathbf{h}_i)$ ▷ forward model to update the temperature measurements at the remaining depths
            Update $i = i + 1$
            Update $\mathcal{Z}_i = \mathcal{Z}_{i-1} \setminus \{z_{i-1}^*\}$ ▷ remove the optimal depth from the set

end while
end for
end for
Outputs: $(k \times N_{20\%} \times c - 2)$-sized array of optimal depths z_i^*, $i = 2, \cdots, \omega$

3D case - static design (staBOED)

The staBOED algorithm is similar to the approaches of Chapters 3 and 4. Given the dataset $\mathcal{D}_{3D} = \{\mathbf{d}_\xi, \mathbf{h}\}^N$, the goal is to determine the optimal sampling design ξ^* from pre-selected sensor locations $\xi \in \Xi$ in order to obtain the most accurate and precise estimates of the fluxes $\mathbf{h}$ over the riverbed, such that $\xi^* = \arg\min_{\xi \in \Xi} \mathcal{J}(\mathbf{h})$.

The utility function $\mathcal{J}(\mathbf{h})$ is the same as in the 1D case (Equation 5.1), and to connect seqBOED and staBOED, the optimal sampling depths from the seqBOED algorithm are applied to the 3D scenario, i.e., the staBOED in the 3D case will determine the 2D (x, y) coordinates of the optimal sensor locations on top of the riverbed, and the optimal sampling depths (z) from seqBOED are used, fulfilling the 3D scenario.

The sequential BOED method could be applied to this scenario, but it is unnecessary due to the nature of the problem, as we are only concerned with the global mean flux. In a similar real-world scenario, the sensors would likely be pre-selected and the operator would not re-run the BOED procedure on the field for each new sensor location.

To ensure consistency and repeatability, the staBOED algorithm's performance is evaluated using k-fold cross-validation, and the results are compared and averaged across the folds.

After describing both seqBOED and staBOED, the following section will present the PBNN used for each algorithm, followed by the experimental setup to obtain the training and test datasets for each case, 1D and 3D.

5.2.2 Probabilistic Bayesian Neural Networks

Let $\theta = (\mathbf{w}, \mathbf{b})$ be the parameters of a BNN f_θ, where $\mathbf{w}$ are the weights of the network connections and $\mathbf{b}$ the biases. Let $\mathbf{d}$ be the input (predictor) variable, and let $\mathbf{h} = f_\theta(\mathbf{d})$ be the output (target) variable. We assume that $\mathbf{d}$ and $\mathbf{h}$ are independent and identically distributed (i.i.d.). The goal of a BNN is to infer the posterior distribution of the parameters θ given a data set $\mathcal{D} = \{\mathbf{d}_n, \mathbf{h}_n\}_{n=1}^N$. The posterior distribution can be computed using Bayes' theorem (Gal, 2016; Sharma et al., 2022):

$$p(\theta|\mathbf{d}, \mathbf{h}) = \frac{p(\mathbf{h}|\mathbf{d}, \theta)\, p(\theta)}{p(\mathbf{h}|\mathbf{d})} \tag{5.3}$$

Here, $p(\mathbf{h}|\mathbf{d}, \theta)$ is the likelihood; $p(\theta)$ is a prior distribution, which is typically chosen to be a zero-mean isotropic Gaussian (Jospin et al., 2022); and $p(\mathbf{h}|\mathbf{d})$ is the evidence, which can be computed by marginalizing the likelihood over the parameters:

$$p(\mathbf{h}|\mathbf{d}) = \int p(\mathbf{h}|\mathbf{d},\theta)\, p(\theta) d\theta \tag{5.4}$$

The true posterior $p(\theta|\mathbf{d},\mathbf{h})$ is usually intractable. Rather than sampling from the exact posterior, the variational distribution $q_\gamma(\theta)$ is used, which belong to a tractable family of distributions (e.g., Gaussian distribution) and is parametrized by a set of parameters γ. The parameters γ are then learned so that the variational distribution $q_\gamma(\theta)$ is as close to the true posterior $p(\theta|\mathbf{d},\mathbf{h})$ as possible (Gal, 2016; Jospin et al., 2022). The Kullback-Leibler divergence (KL-divergence; Kullback and Leibler 1951), based on Shannon's information theory (Shannon, 1948), is a commonly used measure of closeness between probability distributions:

$$\mathrm{KL}(q_\gamma(\theta)||p(\theta|\mathbf{d},\mathbf{h})) = \int q_\gamma(\theta) \log \frac{q_\gamma(\theta)}{p(\theta|\mathbf{d},\mathbf{h})} d\theta \tag{5.5}$$

Minimizing the KL-divergence is equivalent to maximizing the evidence lower bound (ELBO) w.r.t. the variational parameters γ:

$$\mathcal{L}(\gamma) = \mathbb{E}_{q_\gamma(\theta)} \left[\log p(\mathbf{h}|\mathbf{d},\theta)\right] - \mathrm{KL}(q_\gamma(\theta)||p(\theta)) \leq \log p(\mathbf{h}|\mathbf{d}) \tag{5.6}$$

This procedure is referred to as variational inference (VI; Blundell et al. 2015; Gal 2016; Jospin et al. 2022). For a BNN, the optimization of the ELBO with respect to the parameters of a variational distribution requires the adaptation of VI. Stochastic variational inference (SVI; Hoffman et al. 2013), which is the stochastic gradient descent (SGD) method applied to VI, is the most widely used method for optimizing the ELBO (Jospin et al., 2022).

To efficiently approximate the gradients of the ELBO with respect to the variational parameters γ, the Bayes-by-Backprop (BBB; Blundell et al. 2015) algorithm is used, which combines the variational and reparameterization gradient estimators (Kingma and Welling, 2014) to obtain unbiased and low-variance gradients (Blundell et al., 2015). The BBB approach updates the variational parameters γ using the following optimization procedure:

$$\gamma_{t+1} = \gamma_t - \alpha \nabla_\gamma \mathcal{L}(\gamma) \tag{5.7}$$

where α is the learning rate of the optimizer (e.g., Adam; Kingma and Ba 2014).

For regression tasks, a BNN can be trained by minimizing the mean squared error (MSE) loss function between the predictor $\mathbf{d}$ and the target $\mathbf{h}$:

$$MSE = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{h}_n - f_\theta(\mathbf{d}_n))^2 \tag{5.8}$$

This loss function is then used to minimize the KL-divergence by calculating the gradients of the ELBO.

Once the BNN is trained, it can be used to make predictions on unseen data. The term "Bayesian" in BNN refers to the fact that the network's weights are sampled from the posterior distribution $p(\mathbf{w}|\mathbf{d},\mathbf{h})$ approximated by the variational distribution $q_\gamma(\mathbf{w})$– but at each evaluation, the trained network returns a single prediction. To estimate the uncertainty of a predicted target $\mathbf{h}^*$, the network must be evaluated multiple times with the same input $\mathbf{d}^*$, and statistics computed over the predictions.

For problems involving the prediction of continuous variables, however, such conditional averages provide a very limited description of the target variable's properties (Bishop, 1994). In inverse problems, the mapping is frequently multivalued (non-unique), with input values having multiple valid output values. A neural network with an MSE loss function will roughly represent the conditional average of the target data when applied to such inverse problems, which frequently yields extremely subpar performance—it is not guaranteed that the average of several correct values is itself a correct value (Bishop, 1994). This problem can be solved by a mixture density network (MDN; Bishop 1994), which can represent arbitrary distributions in the same way that a conventional neural network can represent arbitrary functions (Hornik et al., 1989).

A MDN is a type of neural network that outputs a conditional probability density function $p(\mathbf{h}|\mathbf{d})$ instead of a single prediction for a given input (Bishop, 1994). The MDN is composed of two parts, a neural network and a mixture model. The neural network is used to map the input data to the parameters of the mixture model, which is then used to compute the conditional *pdf*. The mixture model is composed of κ Gaussian components, and the *pdf* is given by:

$$p(\mathbf{h}|\mathbf{d}) = \sum_{k=1}^{\kappa} \pi_k(\mathbf{d})\,\mathcal{N}(\mathbf{h}|\mu_k(\mathbf{d}),\sigma_k(\mathbf{d})) \tag{5.9}$$

where $\pi_k(\mathbf{d})$, $\mu_k(\mathbf{d})$, and $\sigma_k(\mathbf{d})$ are the mixing weights, means, and standard deviations of the k-th Gaussian component, respectively. The parameters $\pi_k(\mathbf{d})$, $\mu_k(\mathbf{d})$, and $\sigma_k(\mathbf{d})$ are the outputs of the neural network, which is trained by minimizing the negative log-likelihood (NLL) of the data with respect to the parameters $\pi_k(\mathbf{d})$, $\mu_k(\mathbf{d})$, and $\sigma_k(\mathbf{d})$:

$$NLL = -\sum_{n=1}^{N} \log\left[\sum_{k=1}^{\kappa} \pi_k(\mathbf{d}_n)\,\mathcal{N}(\mathbf{h}_n|\mu_k(\mathbf{d}_n),\sigma_k(\mathbf{d}_n))\right] \tag{5.10}$$

Once trained, an MDN can be used to obtain a comprehensive description of the target data, as well as to generate samples from the conditional *pdf*, which can be used to generate new data points or to explore the space of possible outputs for a given input. The main challenge of training a MDN is that the mixing coefficients (π_k), means (μ_k) and standard deviations (σ_k) of each Gaussian component must be jointly optimized, which is computationally expensive. Moreover, optimizing all parameters simultaneously is numerically unstable in higher dimensions, can lead to degenerate predictions, and MDNs are not exempt to the issue of overfitting (Hjorth and Nabney, 1999; Makansi et al., 2019). Over-fitting happens when a model has been overly optimized on the training dataset and no longer generalizes well to new datasets. An over-fitted model, in other

words, has a low error on the training set but a high error on the test set, indicating that the model has memorized the training data points rather than generalizing the underlying patterns.

To address these issues, we propose a BNN-based MDN architecture, which addresses the issue of overfitting, and the numerical instability of the MDN training procedure is further mitigated by predicting a few principal components of the target data of the 3D case. Moreover, we use an early stopping procedure during training to further prevent overfitting–the training dataset is split into a training and validation set, and the training is stopped when the validation loss starts to increase (Prechelt, 1998). For simplicity, only the neural network weights are stochastic in our application, and biases are inferred in the same way as in a conventional neural network.

5.2.3 Experimental setup

3D model description

In this study, we take advantage of a 3D groundwater model of the entire stretch of the Aa River (Belgium), which had been set up and modified in previous studies (Ghysels et al., 2021, 2019; Mohammed, 2009; Mutua, 2013). This large scale model has a horizontal resolution of 2.5 m by 2.5 m and a size of 1500 m by 1800 m. Based on the results of a regional model (Mohammed, 2009), the model domain is bounded on all sides by constant-head boundaries. Ghysels et al. (2021) developed two fine-scale models based on this larger reach-scale model, focusing on the upstream and downstream sections (see Figure 5.1). In this contribution, we focus on the downstream model, which is shown in Figure 5.1. The downstream section is a 25-meter-long stretch of river with an average width of 15 meters. The average water depth is approximately 60 cm, and the riverbed bathymetry is relatively uniform (Ghysels et al., 2021).

Ghysels et al. (2018) carried out a comprehensive field campaign with a focus on the meter-sale variability of riverbed hydraulic conductivity (K). Horizontal K (K_h) ranges from 0.1 to 23.2 m/day, with a mean of 6.4 m/day. In the downstream section, vertical K (K_v) ranges from 0.1 to 4.0 m/day, with a mean of 1.6 m/day (Ghysels et al., 2021). In addition, previous standpipe tests determined a thickness of the riverbank clogging layer of 0.5 m and an average bank hydraulic conductivity (K_{bank}) value of 5.2 m/day (Baya Veliz, 2017; Ghysels et al., 2021).

Ghysels et al. (2021) conducted a temperature survey in the downstream section of the Aa River in the summer of 2016. Temperatures profiles were measured with a T-Lance temperature probe at the downstream section at depths of 0.0, 0.1, 0.2, and 0.5 m as well as at the deepest point that was easily accessible (never deeper than 1.5 m). At a depth of 10 cm below the riverbed, the temperature ranges from 11.7 to 18.4 °C, with a mean of 16.7 °C (Ghysels et al., 2021). At the deepest measurement point, the average temperature is 12.8 °C. All temperature profiles recorded show a decrease in temperature with depth, which is consistent with higher surface water temperatures

compared to groundwater during the summer. At the time of measurement, all profiles are convex upward, indicating an upward flux towards the river (Ghysels et al., 2021).

Ghysels et al. (2021) fitted the 1D steady-state analytical solution of the heat equation (Bredehoeft and Papaopulos, 1965) to the temperature profiles in order to estimate the vertical exchange fluxes:

$$\lambda \frac{\partial^2 T}{\partial z^2} - q_z c_w \rho_w \frac{\partial T}{\partial z} = 0 \tag{5.11}$$

where $\lambda = 1.8\, J/(m\, K)$ is the thermal conductivity of the soil–water matrix, q_z is the vertical exchange flux in m/s, $c_w = 4183\, J/(kg\, K)$ is the specific heat capacity of water, $\rho_w = 1000\, kg/m^3$ is the density of water, and T is the temperature in Kelvin at depth z. Using Equation 5.11, the vertical exchange fluxes for each temperature profile were calculated, and the average flux for the downstream section was found to be -72.5 mm/day. Ghysels et al. (2021) then simulated the vertical exchange fluxes for the downstream section using their 3D groundwater model, and the mean flux was found to be -133.7 mm/day. The higher fluxes in the model are caused by lateral flows through the riverbed and banks, which are not accounted for in the analytical solution (Equation 5.11). All estimated vertical fluxes are negative, indicating that the river is gaining water from the aquifer (Ghysels et al., 2021).

In this contribution, hydrological data from Ghysels et al. (2021) are utilized to simulate river-aquifer heat exchange in the Aa River's downstream section. To demonstrate the applicability of our method, we estimate the vertical exchange fluxes using the 1D analytical solution of the heat equation (Equation 5.11) constrained by distributions of the boundary conditions and hydraulic properties from Ghysels et al. (2021). Then, we use the Ghysels et al. (2021)′s 3D groundwater model to simulate the vertical exchange fluxes. Our proposed method requires a dataset of river-aquifer heat exchange fluxes for training. The model is run several (N) times with different combinations of hydraulic properties and boundary conditions derived from field data, and the results are used to train the neural network; the problem is approached as a supervised learning problem with the temperature data points as input and the vertical exchange flux and other hydrological parameters as output.

1D Model Parameterization

The goal of the 1D case is to show the applicability of PBNNs to BOED in the context of river-aquifer heat exchange. The aim is to find optimal measurement depths along vertical riverbed profiles, which will be used to estimate vertical exchange fluxes in the 3D model. The 3D model has a maximum depth of -94.68 m. However, the maximum easily accessible depth in Ghysels et al. (2021)′s field conditions with a T-Lance temperature probe is approximately -2 m. As a result, the sampling depths available are bounded by 0.0 m and -2.0 m, and the sampling depth is discretized into 10 points. If the riverbed's surface is assumed to be 0.0 m, the first sampling depth is -0.1 m and the last sampling depth is -1.9 m. This resolution captures the vertical exchange fluxes and

is realistic for field conditions.

The 1D training dataset is generated by running the 1D analytical solution of the heat equation (Equation 5.11) with the boundary conditions and hydraulic properties from Table 5.1. Vertical fluxes are calculated with

$$q_v = \frac{K_v \left(p_f - p_i\right)}{L} \tag{5.12}$$

where K_v is the vertical hydraulic conductivity, p_f is the bottom hydraulic head, p_i is the top hydraulic head, and L is the distance between p_f and p_i.

The training dataset consists of N samples, where N is the number of combinations of hydraulic properties and boundary conditions. The predictor $\mathbf{d}$ is the temperature profile at the sampling depths, i.e., $d_i = (T_{0.1}, T_{0.2}, \ldots, T_{1.9})$, $i = 1, \ldots, N$, and the target $\mathbf{h}$ is the 3-dimensional matrix of vertical exchange fluxes, thermal conductivity, and bottom boundary temperature, i.e., $h_i = (q_z, \lambda, T_b)$, $i = 1, \ldots, N$. The thermal conductivity is a key parameter in the 1D solution of the heat equation, and the bottom boundary temperature is a boundary condition that is not measured in the field. The inclusion of thermal conductivity and bottom boundary temperature in the target is intended to better constrain the sequential BOED, resulting in more accurate flux predictions. Indeed, the essence of sequential BOED is to update the model parameters based on the predictions of the previous iteration, and prior knowledge of the existence of a model parameter or boundary condition can help to improve the efficiency of BOED in this case, as shown in the following sections.

Parameter	**Value**
Surface elevation (z_0)	0 m
Bottom elevation (z_b)	-94.68 m
Surface temperature (T_0)	U[12; 20] °C
Bottom temperature (T_b)	U[10; 12] °C
Surface hydraulic head (p_0)	0 m
Bottom hydraulic head (p_b)	U[-0.2, 0] m
Soil thermal conductivity (λ)	U[0.2; 2] $W/(m\,K)$
Water specific heat capacity (c_w)	4183 $J/(kg\,K)$
Water density (ρ_w)	1000 kg/m^3
Soil vertical hydraulic conductivity (K_v)	U[0.1; 8] m/day

Table 5.1: Note: Boundary conditions and hydraulic properties used in the 1D model. U[a, b] refers to the continuous uniform distribution bounded by the values a and b.

3D Case Parameterization

We extend and refine the 3D downstream model from Ghysels et al. (2021) to include temperature flow simulation. The objective of the 3D case is to produce N pairs

of simulated three-dimensional riverbed exchange fluxes and the corresponding three-dimensional temperature field.

Flow model. To reduce the impact of boundary conditions, we extend the spatial boundaries by 10 meters in the x and y dimensions on each side (i.e., the model is extended by 20 m in total along the x and y axis), and we keep the same cell dimensions in the x and y axis (0.5 × 0.5 m). For an accurate simulation of river-aquifer exchange fluxes, a higher vertical resolution is required: we increased riverbed resolution by subdividing the riverbed into ten layers of 20 cm each, which matches the resolution of the 1D model. Our finer-scale model has 16 layers, 89 rows and 105 columns. As a result, the x-extent is 44.5 m and the y-extent is 52.5 m. The z-extent varies due to the included topography of Ghysels et al. (2021), and the same extent is kept.

The boundary conditions at each border are defined as constant heads, the values of which are determined by the flow solution of the larger scale model. The recharge is set to 1.9×10^{-3} m/day, which is the same as in Ghysels et al. (2021). The aquifer layers are modelled as convertible, and the river is modeled as a boundary with a constant hydraulic head equal to the river stage (Ghysels et al., 2021). The Horizontal Flow Barrier (HFB) package (Langevin et al., 2017) is used to simulate the effect of a clogged riverbank layer (Ghysels et al., 2021).

In order to generate the training dataset, the hydraulic properties of interest to vary are the vertical hydraulic conductivity (K_v), the horizontal hydraulic conductivity (K_h), and the bank hydraulic conductivity (K_{bank}). Since K_v is usually defined as a fraction of K_h, i.e., $K_v = K_h/ratio$, we simply need to generate different values for K_h and use a fixed ratio of 10, which is a value inferred from Ghysels et al. (2021). By means of Sequential Gaussian Simulation (SGS; Goovaerts 1997) via SGEMS (Remy et al., 2009) and pySGEMS (Thibaut and Vandekerckhove, 2021), N K_h fields are generated. Since the spatial and magnitude variability of the hydraulic conductivity field are already accounted for in the SGS results, it is not necessary to adjust the ratio of K_v and K_h. The background hydraulic conductivity (outside the riverbed and banks), is maintained at the same level as in Ghysels et al. (2021) because we are only interested in the effect of the river-aquifer exchange fluxes and not the aquifer itself.

We add anisotropy to the hydraulic conductivity field to account for structural uncertainty by modifying the SGS variogram model. Because of the nature of 2-point geostatistics, simulations of K will remain smooth, but different degrees of heterogeneity can be obtained by varying the anisotropy. We choose a value for the K variance at random, as well as the medium range of the K variogram and its orientation with respect to the x direction, as was done in Chapter 3. The values in the simulation results of SGEMS are samples from a normal distribution with a mean of 0 and a standard deviation of 1. Simulations K_i, $i = 1, \ldots, N$ are first bounded between -1 and 1, and then scaled to the desired range by the following formula: $K_{i,\,aniso} = \exp(K_i \log K_{std,i} + \log K_{mean,i})$. $K_{mean,i}$ and $K_{std,i}$ are the desired mean and standard deviation of the K_i field, respectively. The exponential function is used to (i) to ensure K positivity and (ii) that

the magnitude of the resulting $K_{i,\,aniso}$ field varies sufficiently. Table 5.2 provides the parameter ranges for the variogram parameters.

Parameter	Value
$K_{h,\,\min}$	0.4 m/day
$K_{h,\,\max}$	80 m/day
K_h mean	U[$K_{h,\,\min}$, $K_{h,\,\max}$]
K_h std	U[1.6, 2.0] m/day
Angle around vertical axis	U[0, 360] °
Minimum range ($r_{\min}$)	5 m
Maximum range ($r_{\max}$)	20 m
Medium range (r_{med})	U[$r_{\min}$, $r_{\max}$]
Nugget effect	0

Table 5.2: Note: Hydraulic properties used in the 3D model. U[a, b] refers to the continuous uniform distribution bounded by the values a and b.

Transport model. Heat transport can be simulated using the MT3DMS-USGS software (Bedekar et al., 2016), which is a numerical model for simulating advective and dispersive transport of conservative and non-conservative solutes in three-dimensional saturated subsurface flow systems. Hecht-Méndez et al. (2010) demonstrated that MT3DMS is a suitable tool for simulating heat transport in the shallow subsurface due to the similarities between thermal and solute transport (Zheng, 2009; Zong et al., 2021).

Solute transport is simulated with the equation

$$\left(1 + \frac{\rho_b K_d}{n}\right) \frac{\partial C}{\partial t} = \nabla \cdot [D_s \nabla C] - \nabla \cdot (v_a C) + \frac{q_k C}{n}, \tag{5.13}$$

where C is the concentration of the solute (kg/cm^3), t is time (s), K_d is the distribution coefficient (m^3/kg), ρ_b is the density of the fluid (kg/cm^3), q_k is the source/sink term (represented by aquifer volumetric flow rate; $m^3/s/m^3$), and n is the porosity (-). D_s is the diffusion-dispersivity tensor:

$$D_s = \begin{pmatrix} D_m + v \cdot \alpha_L & 0 & 0 \\ 0 & D_m + v \cdot \alpha_T & 0 \\ 0 & 0 & D_m + v \cdot \alpha_V \end{pmatrix} \tag{5.14}$$

where D_m is the molecular diffusion coefficient (m^2/s), v is the fluid velocity (m/s), α_L, α_T, and α_V are the longitudinal, transverse, and vertical dispersivity coefficients (m), respectively.

The heat transport equation is given by (Hecht-Méndez et al., 2010):

$$\left(\frac{\rho_m c_m}{n \rho_w c_w}\right) \frac{\partial T}{\partial t} = \nabla \cdot [D_h \nabla T] - \nabla \cdot (v_a T) + \frac{q_h}{n \rho_w c_w}, \tag{5.15}$$

where T is the temperature (Kelvin), ρ_m is the density of the porous medium (kg/m^3), c_m is the specific heat of the porous medium ($J/kg/K$), ρ_w is the density of the fluid (kg/cm^3), c_w is the specific heat capacity of the fluid ($J/kg/K$), q_h is the heat source/sink term (W/m^3), and n is the porosity. D_h is the heat diffusion-dispersivity tensor:

$$D_h = \begin{pmatrix} \frac{\lambda_m}{n\rho_w c_w} + \alpha_L v_a & 0 & 0 \\ 0 & \frac{\lambda_m}{n\rho_w c_w} + \alpha_T v_a & 0 \\ 0 & 0 & \frac{\lambda_m}{n\rho_w c_w} + \alpha_V v_a \end{pmatrix} \tag{5.16}$$

where λ_m is the thermal conductivity of the porous medium ($W/m/K$).

The similarities between the two equations become even more apparent when stated in plain language:

Retardation factor × Transient term = Diffusion/Dispersion − Advection + Source/sink

To physically relate Equations 5.13 and 5.15, we compare the coefficients and describe each term's MT3DMS implementation following Hecht-Méndez et al. (2010)′s approach. In the solute transport equation, the retardation factor represents solute sorption. In the heat transport equation, it represent solid-water heat exchange. The heat retardation factor is the ratio of porous medium (total phase) and water heat capacity (mobile phase) (Shook, 2001). The distribution coefficient is the ratio of solids' and water's specific heat capacities. The Chemical Reaction Package in MT3DMS incorporates the new heat distribution coefficient. To maintain the temperature exchange rate between the solid and the water constant regardless of temperature changes, the type of sorption must be set to a linear isotherm (Hecht-Méndez et al., 2010).

Equation 5.13′s diffusion and dispersion term has two parts. The first part is pure molecular diffusion (D_m), a concentration gradient-driven process. In the heat transport equation, it is the temperature gradient-driven thermal diffusivity (Hecht-Méndez et al., 2010).

Hydrodynamic dispersion in Equation 5.13 is driven by pore-scale flow velocity differences. MT3DMS uses the heat dispersivity coefficient as in solute transport (Hecht-Méndez et al., 2010).

Source and sink terms represent mass entering or leaving the domain. These terms indicate energy input or extraction in the heat transport equation. To be consistent with contaminant and heat transport dimensions, Kelvin equals concentration (kg/m3), so energy input/extraction is a mass load per unit aquifer volume (Hecht-Méndez et al., 2010).

We use the third-order TVD (Total-Variation-Diminishing, ULTIMATE solver) scheme to solve for the advection term and the generalized conjugate gradient (GCG) scheme to solve for the non-advective terms, as recommended by Molina Giraldo et al. (2009) and Hecht-Méndez et al. (2010). The transport simulation is coupled to the flow field

solution obtained from the MODFLOW-2005 model (Harbaugh, 2005).

In order to simulate the temperature field, we first need to define boundary conditions, initial conditions, and the transport parameters. We assume a constant boundary condition of T_0 at the top of the model and a constant boundary condition of $T_b \leq T_0$ at the bottom of the model. We define the initial temperature of the entire model as T_b. Since the river is modeled as a component of the first layer, its temperature is assumed to be equal to T_0. This is merely a property of convenience (e.g., Cartwright 1983), as the river is not a part of the transport model and we are only concerned with the temperature field in the riverbed.

To input the proper parameters in the Reaction package to simulate heat transport, we must define the distribution coefficient K_d as follows:

$$K_d = \frac{c_m}{\rho_w c_w} \tag{5.17}$$

For the same reasons, we define the molecular diffusion coefficient D_m in the Dispersion package as follows:

$$D_m = \frac{\lambda_m}{n\rho_w c_w} \tag{5.18}$$

We define our prior knowledge of the parameters by setting realistic values for the parameters as shown in Table 5.3.

Parameter	Value
T_0	U[12, 20] $°C$
T_b	U[10, 12] $°C$
n	U[0.12, 0.25] -
$\alpha_{L,T,V}$	U[1, 2] m
ρ_m	U[1.6, 2] g/cm^3
c_m	U[1, 2] $J/g/K$
λ_m	U[0.2; 2] $W/m/K$
ρ_w	1 g/cm^3
c_w	4.183 $J/g/K$

Table 5.3: Note. Parameters used in the heat transport simulation. U[a, b] denotes a uniform distribution between a and b.

A single pair of forward modelling (flow and transport modeling) takes about 15 seconds to compute on a 2.3 GHz 8-Core Intel Core i9 processor.

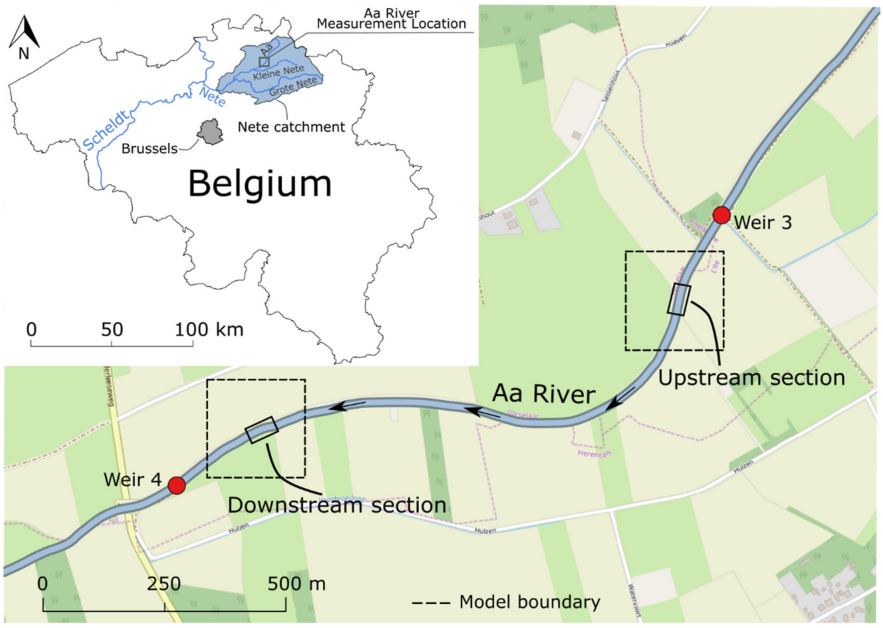

Figure 5.1: Map of Belgium depicting the Nete River catchment and its principal rivers in blue (upper left). The study area is situated along the Aa River's lower reaches. The downstream and upstream sections of Ghysels et al. (2021) are highlighted, with dashed lines indicating the boundaries of fine-scale groundwater models. The downstream model is the focus of our study. Image taken from Ghysels et al. (2021).

5.2.4 Pre-processing

Target. In the 1D case, the target variable is pre-processed by scaling each feature independently to have zero mean and unit variance, given that the scales of the three features are significantly different. This pre-processing step is essential for stabilizing the training of the neural networks (Bishop, 2007).

The target variable $\mathbf{h}$ in the 3D case is the flux field at the riverbed's top layer. The grid cells that correspond to the riverbed are contained within the model's first layer (89 rows $\times$ 105 columns). A mask is applied to the flux field to extract the fluxes at the riverbed's top layer, yielding a 1D array of 3577 fluxes.

It is essential to reduce the target dimensionality in such a way that the dimensionally-reduced variables capture the primary information contained in the original variable, such as its spatial structure, trends, and variability (e.g., Chapter 3). An obvious and natural choice would be to reduce each flux vector to a one-dimensional array containing the mean flux. However, this would inevitably destroy the fluxes' spatial structure. As a result, we choose Principal Component Analysis (PCA) to reduce dimensionality. Again, we must be realistic regarding the predictor's nature (1D temperature profiles at discrete locations), so we reduce the fluxes to only two principal components. Intuitively,

the first principal component is highly correlated with the mean flux. The pre-processed target contains two variables of size N, the first and second principal components. The explained variance ratio of the first two principal components is ≈ 0.31, which is a reasonable value for the purpose of this study.

Predictor. Before training, the predictor in the 1D case is pre-processed by scaling each feature independently to have zero mean and unit variance.

The predictor variable $\mathbf{d}$ in the 3D case is made up of discrete temperature profiles through the layers of the riverbed. The temperature profiles are taken at the layers identified by the 1D case solution. The experimental design procedure extracts the temperature profiles from the output of the 3D model at specified x and y coordinates of the surface. These profiles are then stored in a 3D array with dimension of $N \times O \times L$, where N is the number of observations, O is the number of temperature profiles per observation, and L is the number of monitored layers. For example, if we monitor 4 layers at 8 locations, the array will have a dimension of $N \times 8 \times 4$. The two last dimensions are then flattened to form a 2D array of size $N \times (O \cdot L)$ since neural networks are designed to accept only 2D arrays in our implementation. The 3D case predictor is then pre-processed by scaling each feature independently to have zero mean and unit variance.

5.3 Results

5.3.1 1D sequential BOED

Preliminary demonstration

The first step consists in generating the training dataset. The boundary conditions and hydraulic properties are sampled from the continuous uniform distributions shown in Table 5.1, and we use the 1D analytical solution of the heat equation (Equation 5.11) to generate a dataset of $N_{training} = 1250$ samples for training and validation, and $N_{test} = 3$ samples for testing and demonstrating the application of the proposed methodology. The size of the training set is chosen to be large enough to ensure that the model can generalize well to the test set, while keeping the training time reasonable in light of the computational cost of the experimental design.

The trainable parameters of an MDN are the means, standard deviations, and mixing coefficients of the Gaussian components, which depend on the output dimension. In this case, the output dimension is 3 (vertical flux, thermal conductivity, and bottom temperature), and the number of trainable parameters is $\kappa \times (\dim(\mathbf{h}) \times 2 + 1)$, where κ is the number of Gaussian components. The number of Gaussian components chosen is determined by the nature and complexity of the problem. Due to this 1D synthetic scenario's emphasis on speed and simplicity, the number of Gaussian components has been held constant at 1. The number of trainable parameters is $\kappa \times (3 \times 2 + 1) = 7$. By keeping the number of Gaussian components at one, the mixing coefficients (Equa-

tion 5.9) are not trained, reducing the number of trainable parameters to four.

The initial PBNN architecture consists of 3 input nodes for the temperatures at the sampling depths (i.e., d_0, $d_{0.1}$, d_z), $z \in \{0.3, 0.5, \ldots, 1.9\}m$. The input nodes are connected to a single hidden layer with 16 nodes, chosen empirically, because we observed that this architecture was adequate to accurately estimate the fluxes without overfitting or being computationally expensive. The activation function for the hidden layer is Rectified Linear Unit (ReLU), and the output layer has a linear activation function. The initial PBNN is trained for a maximum number of 1000 epochs using the Adam optimizer with a learning rate of 0.003, a batch size of 64, and a NLL loss function (Equation 5.10). The training dataset is split into training and validation in a ratio of 0.8:0.2, and an early stopping criterion is implemented to prevent overfitting if the validation loss does not decrease for 10 epochs.

In our application, the PBNN hyperparameters (i.e., the number of layers, nodes, components, and epochs) are kept constant during training across all sensor locations. In practice, hyperparameter tuning in DL is often done manually and can be time-consuming and tedious (e.g., Bergstra et al. 2011; Bergstra and Bengio 2012). However, the goal of this section is not to find an optimal neural network, but rather to determine the combination of sensor locations that best reduce uncertainty. Therefore, by implementing a BNN with a 1-component MDN and an early stopping criterion, we are able to bypass hyperparameter tuning beyond the trial-and-error approach and ensure that the results obtained are robust and reliable.

With the selected hyperparameters, the training process for each iteration is fast and the training time on a 2.3 GHz 8-Core Intel Core i9 processor never exceeds 10 seconds. Calculating the optimal sequence for a single test sample requires probing 9 points for the first iteration, 8 points for the second iteration, and 7 points for the third iteration, which on the same machine takes approximately two minutes.

To illustrate the 1D sequential BOED method, we select three examples from the test dataset (Figure 5.2), whose parameters are listed in Table 5.4. They represent three distinct scenarios: (1) a high-magnitude flux, (2) a low-magnitude flux, and (3) a medium-magnitude flux, each with its own initial (surface) temperature. Note that, in order to demonstrate our methodology, the experimental data (temperature) from the last case are outside the range of the training dataset (Figure 5.2C).

Light BOED. The sequential BOED method without TBRS is the lightest version, as it only requires running the forward model a number of times equal to the number of training samples desired, chosen to be equal to the original training dataset size for all cases presented in this section. The sequential BOED method is applied to the three examples, and the results are shown in Figure 5.2. Each example starts with the initial dataset with the initial sensor location at d_0 and $d_{0.1}$, and the next three optimal sensor locations ($d_{z_1^*}$, $d_{z_2^*}$, and $d_{z_3^*}$) are determined sequentially. Each time a new sensor is added, the predictor is updated through forward modeling and the training process is

repeated, as described in §5.2.1. The temperature boundary conditions are updated at each iteration. Indeed, the first sensor location is added at d_0, and the bottom temperature is part of the target. As a result, we anticipate that the bottom temperature posterior will be updated at each iteration, reducing the spread of the temperature curves.

The fluxes log prior density distribution is shown in Figure 5.2D-F (gray). In the case of a low-magnitude flux (Example 1), the uncertainty of the estimated flux is the lowest (Figure 5.2D). Conversely, the uncertainty of the estimated flux is higher for the high-magnitude flux (Example 2; Figure 5.2E), and the uncertainty is intermediate for the medium-magnitude flux (Example 3; Figure 5.2F). For each case, the uncertainty of the estimated flux decreases after each sensor is added, as expected.

Seemingly, the uncertainty of the estimated flux is not reduced as much as expected. This is due to the fact that an infinite number of flux and thermal conductivity combinations can produce the same temperature curve. For instance, a high flux and a low thermal conductivity could result in the same temperature curve as a low flux and a high thermal conductivity. As a result, even after three sensors are added, the uncertainty in the estimated flux remains high.

Threshold-based rejection sampling (TBRS). The sequential BOED method with TBRS is now considered. The forward model is run after each learning step (i.e., for each new sensor location) with flux, thermal conductivity, and bottom temperature values sampled from the posterior distribution. The temperature curves generated are then compared to the experimental data (true temperature measurements) and rejected if the difference between the two curves exceeds a 0.1 Kelvin threshold. When the desired number of posterior samples that meet the threshold is reached, the process is terminated. The results are shown in Figure 5.3. Although the temperature curves are fitted to the desired threshold, the uncertainty of the estimated flux is not reduced much more than in the previous case. This demonstrates the difficulty of inferring the flux with low uncertainty when the thermal conductivity is unknown.

TBRS and IPR. We adapt the sequential BOED method with TBRS to include IPR, as described in Michel et al. (2022a). Our implementation differs slightly from Michel et al. (2022a)′s in that we first perform the TBRS step as described above, followed by IPR, in which we add the "filtered" posterior samples to the original training dataset. At each iteration, the training dataset doubles in size, and the PBNN is retrained. Notably, the original training dataset remains in the training dataset, so we could expect the PBNN to detect more patterns in the data that would have potentially been lost if the original training dataset had been replaced with the posterior samples. As a result, the training step becomes slower with each iteration. The results are shown in Figure 5.4. The temperature curves are fitted to the desired threshold, as in the previous case. For all three examples, there is a slight improvement in the uncertainty of the estimated flux, but the improvement is not significant, and we cannot conclude that the decrease in uncertainty is due to IPR or the increased training dataset size.

Comparison with Markov Chain Monte Carlo (MCMC) sampling. We compare the performance of sequential BOED and MCMC sampling methods. We sample the posterior distribution using the DiffeRential Evolution Adaptive Metropolis (DREAM) family of sampling algorithms (Laloy and Vrugt, 2012). We implement the DREAM(ZS) algorithm (Laloy and Vrugt, 2012; Vrugt, 2016) with five chains and a burn-in of 25,000 iterations using the PyDREAM Python package (Shockley et al., 2017). The samples are accepted based on a likelihood function that generates the proposal temperature measurements at depths $\{0; 0.1, 0.5; 0.9; 1.7\}m$ for a proposal flux, thermal conductivity, and bottom temperature that are sampled from the same prior distribution as the sequential BOED method. The results are shown in Figure 5.5. The fluxes results (Figure 5.5D-L) are similar to those obtained with the sequential BOED method (Figure 5.2D-F), i.e., the uncertainty of the estimated flux is the highest for the high-magnitude flux (Example 2) and the lowest for the low-magnitude flux (Example 1). In particular, the bottom temperature for Example 1 is predicted with a higher uncertainty than the other two examples, similar to the results obtained with the sequential BOED method (Figures 5.2 and 5.5J).

The MCMC results reflect once again the fact that an infinite number of flux and thermal conductivity combinations can produce the same temperature curve. This demonstrates that the uncertainty in the estimated flux obtained with the sequential BOED method is comparable to that obtained with the MCMC sampling method. The MCMC sampling method has a higher computational cost than the sequential BOED method, as expected, because the forward model must be run more times, and requires a burn-in period to reach the desired posterior distribution. Let's further note that the MCMC result do not result from a sequential approach, and therefore excludes all potential benefits of the proposed sequential approach. Applying a sequential approach to the MCMC sampling method is possible, but this would further exacerbate the computational cost. An advantage over PBNN is that the posterior distribution is not constrained by a discrete number of Gaussian components (in this case, 1), and by the training dataset size. However, rather than obtaining an accurate posterior sample, the goal of this work is to determine the combination of sensor locations that best reduces uncertainty. As a result, in practical cases like this, the PBNN-based approach is preferred.

Comparison between sequential BOED and sequential BOED with TBRS. Overall, these results indicate that PBNNs can assess the uncertainty of estimated targets when the number of sensors and training samples are limited, and that the sequential BOED approach can reduce the uncertainty of estimated fluxes as more sensors are added. It also illustrates the difficulty of inferring the flux with low uncertainty in the absence of thermal conductivity knowledge. Reducing the prior of thermal conductivity improves the accuracy and reduces the uncertainty of estimated fluxes, as illustrated in Figure 5.6, showing the results of the sequential BOED method with no TBRS and with a prior for thermal conductivity defined by $U[1.25, 1.85]$ W/m/K.

Comparison of optimal sensor locations. The optimal sensor locations are not always the same due to the inherent stochasticity of the optimization procedure and

learning process (Figures 5.2, 5.3, 5.4, 5.5, 5.6A-C). Furthermore, because the three cases were chosen to illustrate different ends of the fluxes' spectrum, they may not be representative of all possible cases. Note that since each feature is modeled as a Gaussian distribution, we adapted the utility function to compute the KL divergence between the posterior and an ideal Gaussian distribution with a mean of 0 and a standard deviation of 0.1, for the flux only. We could instead compute the utility function on all three features, but in this instance, the flux is the most important feature. In addition, as shown in Figures 5.2, 5.3, 5.4, 5.5, 5.6J-L, the bottom temperature is updated to resemble a Dirac delta function, which is the ideal distribution for the bottom temperature. Consequently, calculating the utility function on such a distribution would be ineffective and lead to instability in the optimization procedure.

Global optimal sequence

In order to derive a meaningful global optimal sequence, the methodology is applied to a test set of size $N_{test} = 1000$, and the procedure is repeated over three successive folds, as described in §5.2.1. There is one minor difference: we generate 3750 samples for the training dataset and 3000 samples for the test dataset, which are then split into three parts each to train all three instances with a training set of 1250 and a test set of 1000. The goal is to have a large enough test set to yield a meaningful global optimal sequence while keeping the training set size small enough to reduce computational costs.

The results are shown in Figure 5.7. The stacked histograms depicts the distribution of optimal sampling location sequences across all three folds for 1000 test cases, without TBRS (Figure 5.7A) and with TBRS (Figure 5.7B). As indicated by the histogram, the first location, $d_{z_1^*}$ is placed the shallowest, while the second and third locations, $d_{z_2^*}$ and $d_{z_3^*}$, are placed at the deepest location, i.e., 1.9 m below the surface. The results are similar for both cases. The interpretation is as follows: (i) shallow sensor locations provide the most valuable information for accurately predicting the flux, where the temperature gradient is steepest. These results agree with the general intuition that the temperature profile in the upper part of the subsurface is most sensitive to the flux. On the other hand, (ii) the deepest locations add supplementary information, because the temperature gradient is smoother and less sensitive to the fluxes at those depths, thereby improving the posterior distributions. The global optimal sequence is thus $d_{z_1^*} = 0.3$ m; $d_{z_2^*} = 1.7$ m; $d_{z_3^*} = 1.9$ m. The high variability in the optimal sequence is also shown in Figure 5.7. Even if intermediate depths are less interesting in general, there are still a significant number of models for which such depths are required to minimize target uncertainty, as in Example 1 with its less steep gradient. Ideally, the seqBOED would then be applied to the field to determine the best sequence locally, but this approach is frequently unrealistic.

Note that these results depend on the initial sampling locations, which are set to $d_0 = 0$ m and $d_{0.1} = 0.1$ m. The results may differ under different assumptions (e.g., an initial sensor placed at 0.5 m), and a new optimal sequence must be determined. However, our initial assumption is reasonable as the first sample is collected at the surface

(i.e., the top boundary condition) and the second sample is collected at a depth of 0.1 m with minimal additional effort. The proposed BOED method is flexible and can be used with a variety of initial sampling assumptions, training datasets, utility functions, and physical constraints.

In our case, it appears that using TBRS to reduce the uncertainty of the estimated flux has little benefit. Instead, we advise using the sequential BOED method without TBRS or IPR, as it is more computationally efficient and yields comparable results to the MCMC sampling method.

Example	q_v (mm d^{-1})	λ_m (W m^{-1} K^{-1})	T_b (°C)
1	-30	1.28	11.53
2	-170	1.70	11.86
3	-92	1.84	11.12

Table 5.4: Note. Target parameters of the three examples.

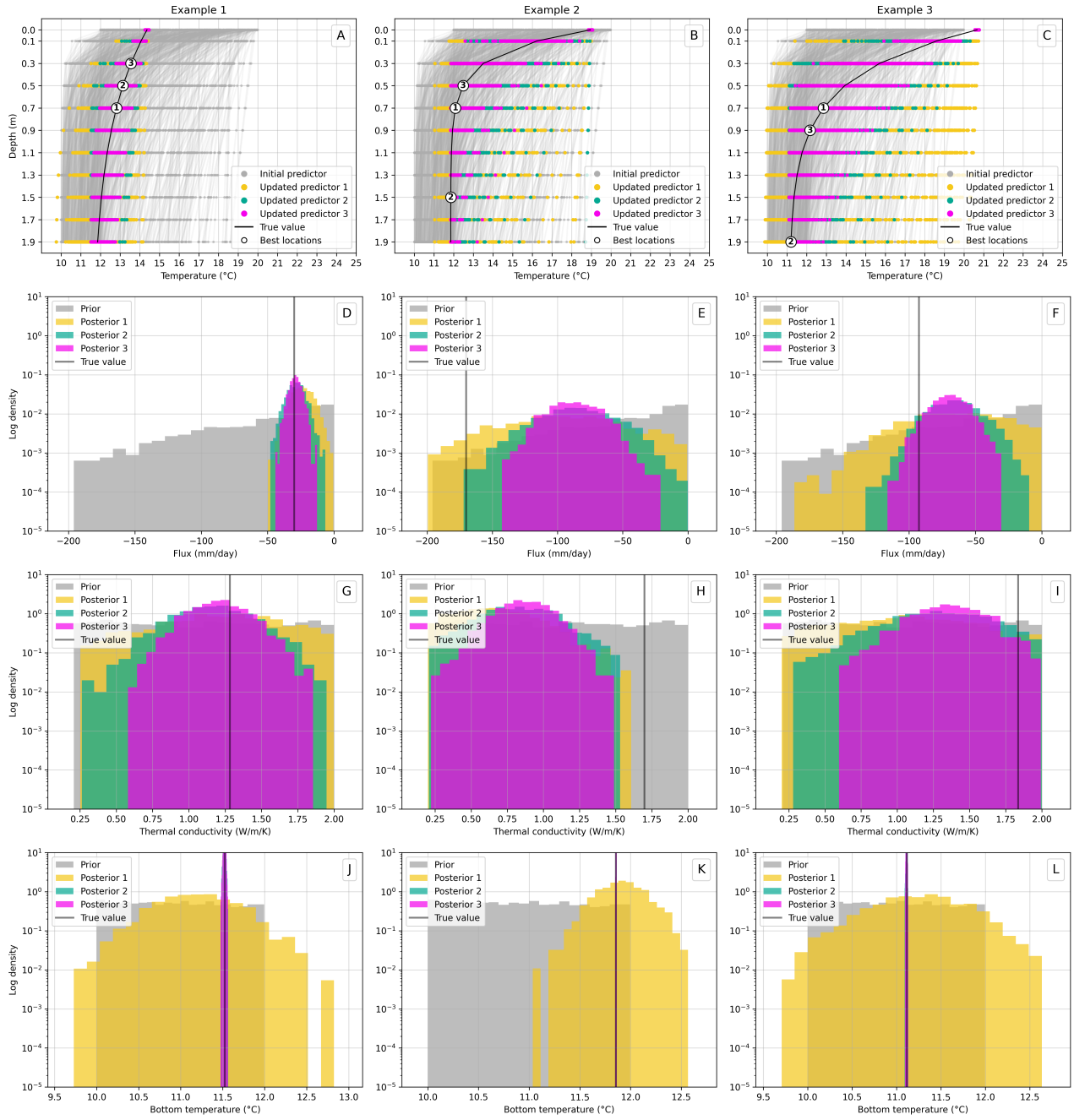

Figure 5.2: BOED sequential procedure without TBRS for the three examples and sequential optimal sampling depths (three points). First row (**A-C**): Predictors (i.e., temperature curves). The three optimal sampling depths are highlighted. Second row (**D-F**): Predicted fluxes. Third row (**G-I**): Predicted thermal conductivity. Fourth row (**J-L**): Predicted bottom temperature.

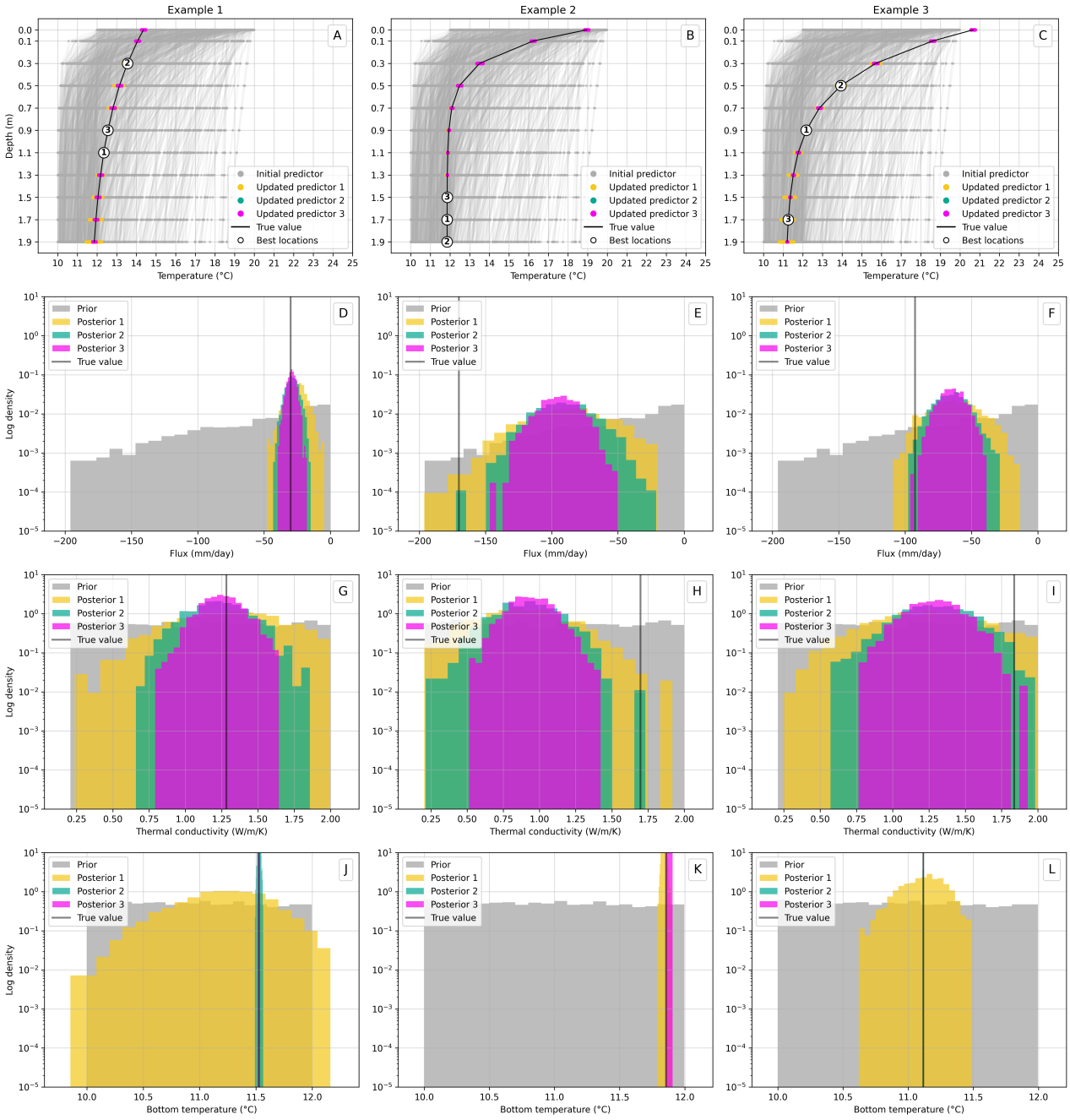

Figure 5.3: BOED sequential procedure with TBRS for the three examples and sequential optimal sampling depths (three points). First row (**A-C**): Predictors (i.e., temperature curves). The three optimal sampling depths are highlighted. Second row (**D-F**): Predicted fluxes. Third row (**G-I**): Predicted thermal conductivity. Fourth row (**J-L**): Predicted bottom temperature.

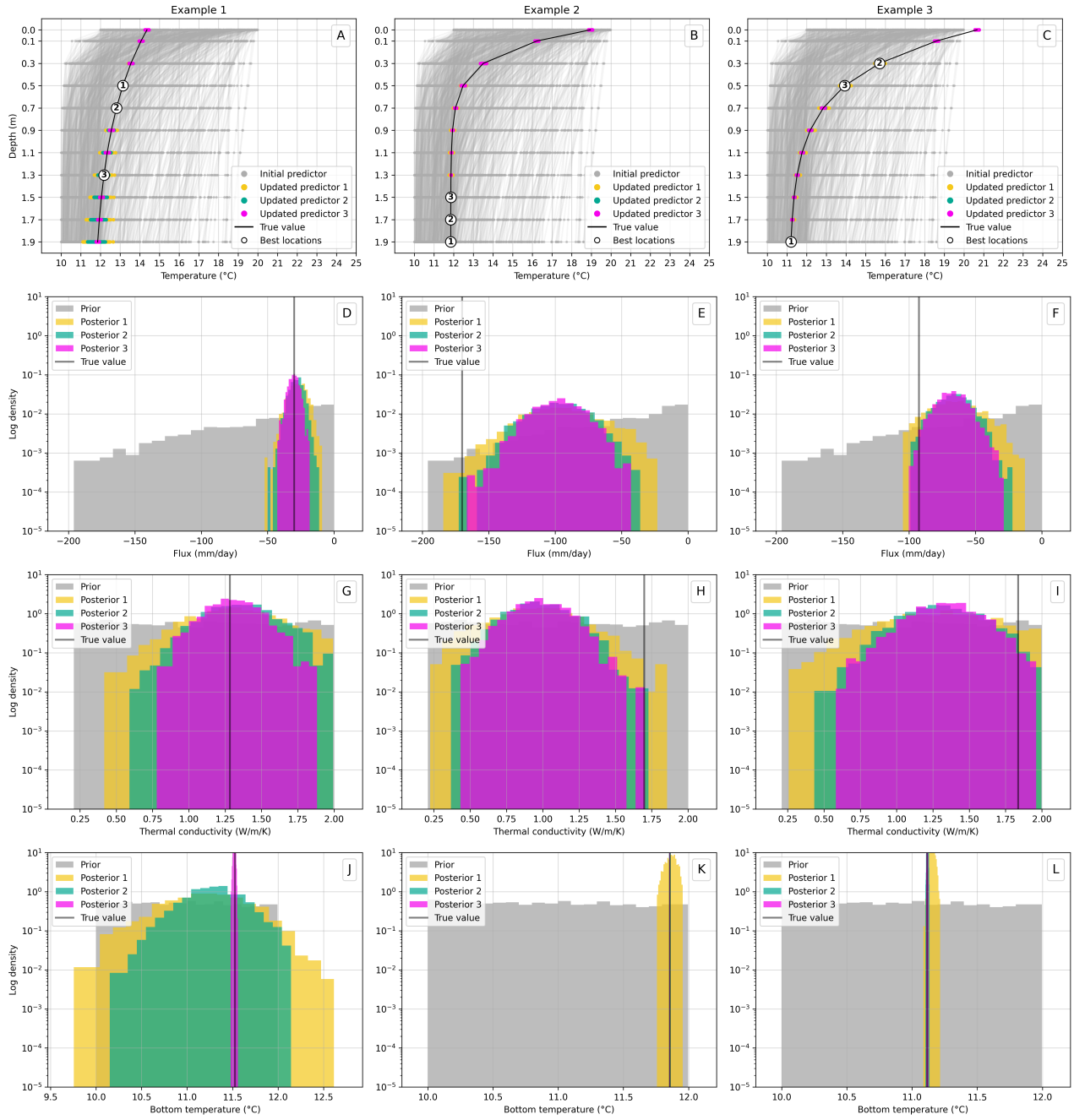

Figure 5.4: BOED sequential procedure with IPR for the three examples and sequential optimal sampling depths (three points). First row (**A-C**): Predictors (i.e., temperature curves). The three optimal sampling depths are highlighted. Second row (**D-F**): Predicted fluxes. Third row (**G-I**): Predicted thermal conductivity. Fourth row (**J-L**): Predicted bottom temperature.

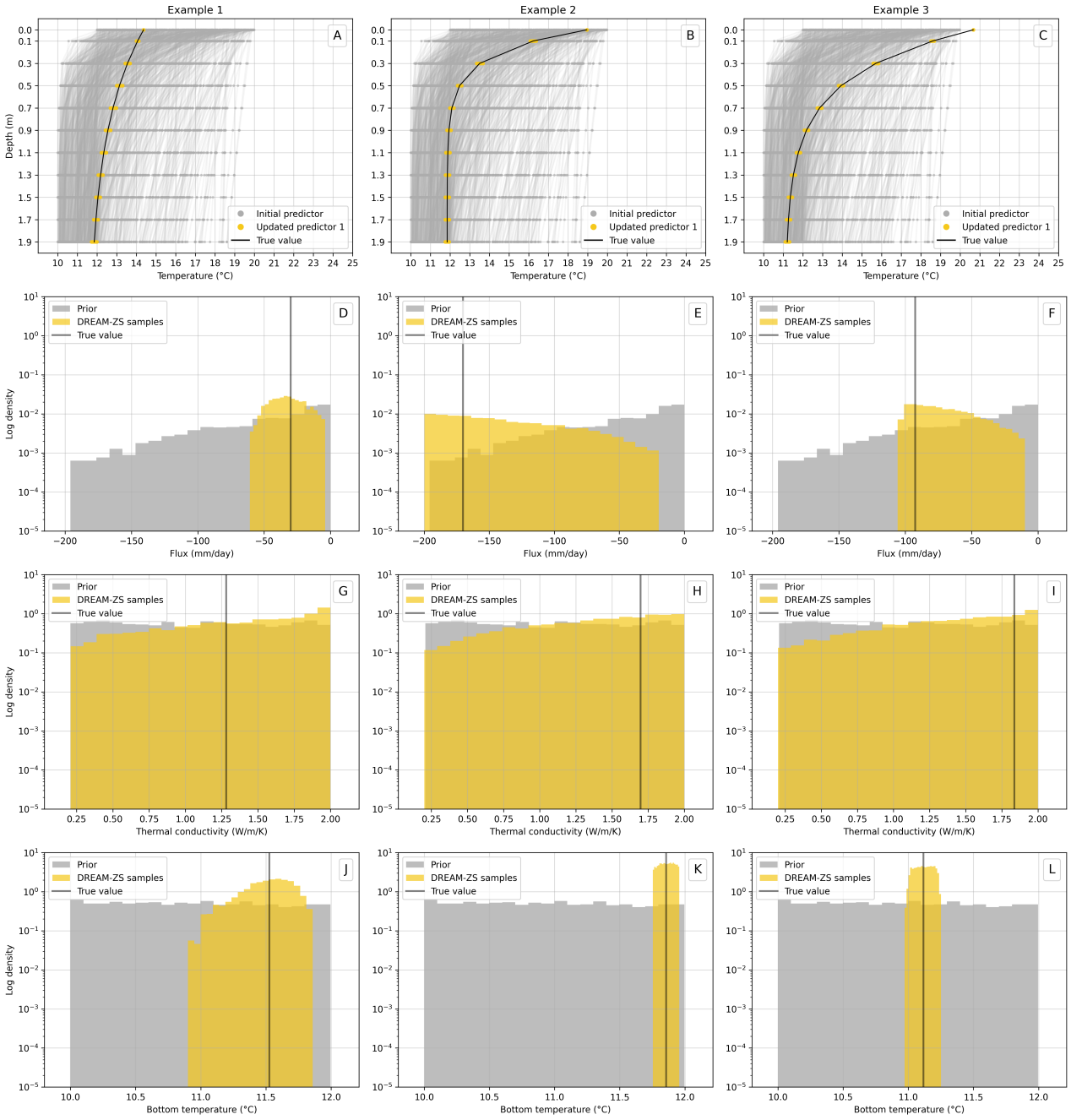

Figure 5.5: MCMC sampling results for the three examples. First row (**A-C**): Predictors (i.e., temperature curves). Second row (**D-F**): Predicted fluxes. Third row (**G-I**): Predicted thermal conductivity. Fourth row (**J-L**): Predicted bottom temperature.

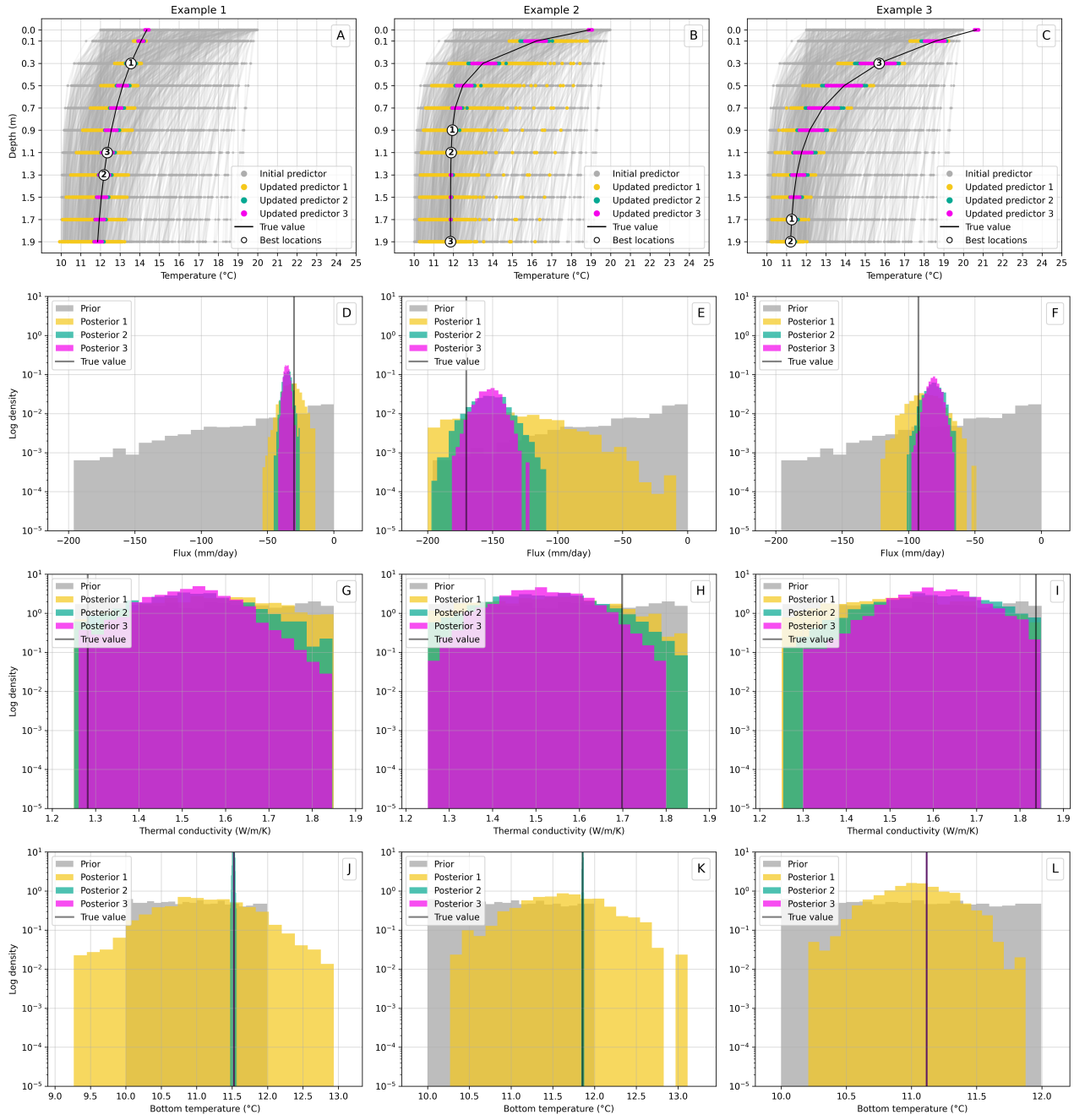

Figure 5.6: BOED sequential procedure without TBRS for the three examples and sequential optimal sampling depths (three points). In this example, the prior distribution of thermal conductivity is defined by $U[1.25; 1.85]$ W m^{-1} K^{-1}. First row (**A-C**): Predictors (i.e., temperature curves). The three optimal sampling depths are highlighted. Second row (**D-F**): Predicted fluxes. Third row (**G-I**): Predicted thermal conductivity. Fourth row (**J-L**): Predicted bottom temperature.

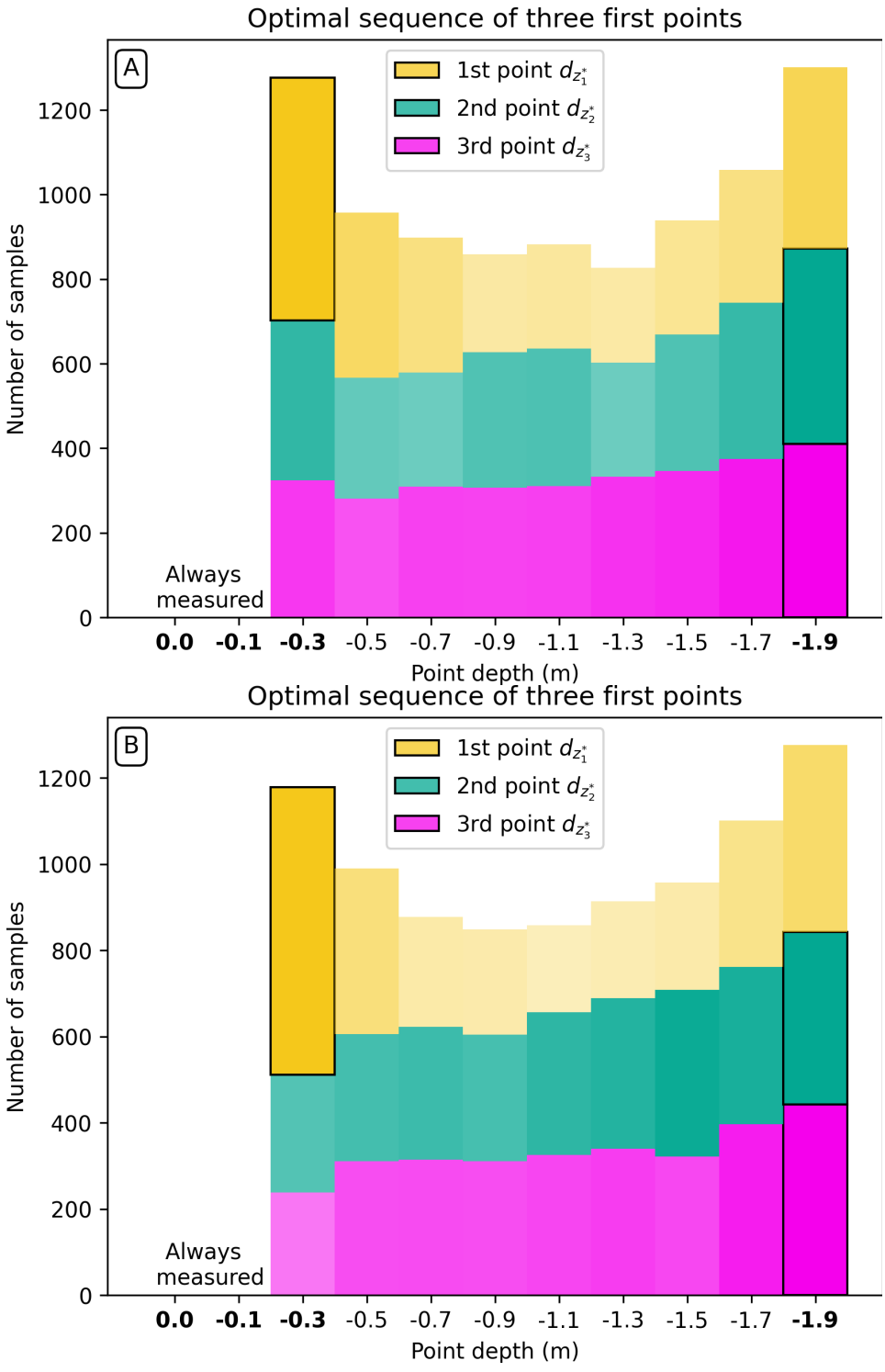

Figure 5.7: 1D scenario optimal sequential design. The stacked histogram depicts the distribution of optimal sampling location sequences across all 3 folds for 1000 test cases. The optimal depths are highlighted by a bold font on the x-axis. Each bar represents the number of times a given depth was selected as the optimal sampling location. Each color represent a different point in the sequence. For example, the yellow bar at $-0.3m$ indicates that the first optimal sampling location was $-0.3m$ in most of the test cases. **A.** Optimal sampling locations without TBRS. **B.** Optimal sampling locations with TBRS.

5.3.2 3D static BOED

In the 3D scenario, the goal is to select the best sampling strategy from pre-selected sensor locations (e.g., Chapters 3 and 4), in order to obtain the most accurate and precise estimates of the mean fluxes over the riverbed. To maintain comparability with the actual field study of Ghysels et al. (2021), we employ the same sampling locations as they did, which are depicted in Figure 5.8A. We consider three types of combinations of sensors: (i) sensors oriented perpendicularly to the river (Figure 5.8B), (ii) sensor parallel to the river (Figure 5.8C), and (iii) randomly placed sensors (Figure 5.8D). We use the same number of combinations and limit the number of sensors to 6 across all three types of combinations to ensure fair comparison.

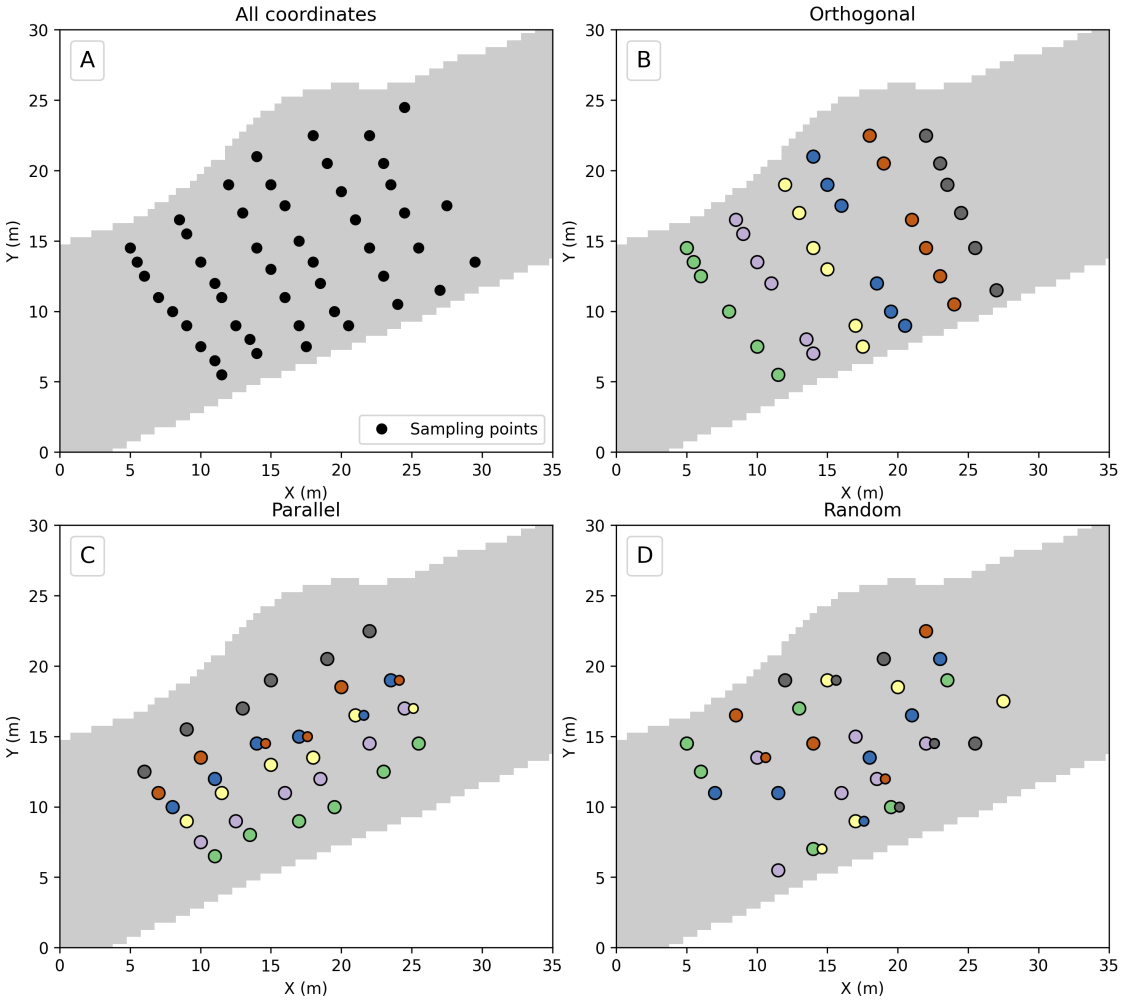

Figure 5.8: **A**. 3D scenario sensor locations. **B**. Perpendicular sensor locations. **C**. Parallel sensor locations. **D**. Random sensor locations. The gray area represents the riverbed. The different colors correspond to various tested scenarios for each sensor orientation. When two sensor groups overlap, their common points appear as a point with a smaller one on top.

First, we illustrate the results using three distinct examples, which are depicted in Figure 5.9. The first example is a case with a high flux and high hydraulic conductivity (Figure 5.9A-B). The second example is a case with a medium flux and medium hydraulic conductivity (Figure 5.9C-D). The third example is a case with a low flux and low hydraulic conductivity (Figure 5.9E-F). The temperature curves at the sensor locations depicted in the second column of Figures 5.10-5.12 are used as predictors in the prediction procedure, pre-processed as described in §5.2.4.

We use a dataset of 10,000 samples which are split into a training set (80%) and a validation set (20%), i.e., $N_{training} = 6400$ and $N_{validation} = 1600$, as described in §5.2.1 to ensure that the model is not overfitting.
We determined that a mixture of $\kappa = 8$ components was sufficient to estimate exchange fluxes accurately without sacrificing computational efficiency. It is not anticipated that the posterior density distribution will exhibit eight distinct peaks; rather, the goal is to obtain a smooth, possibly multimodal distribution that accurately represents the expected fluxes. The mixing coefficients are trained to smooth out or suppress the unnecessary individual Gaussian components, resulting in a posterior with the desired characteristics. However, excessive component count can slow down the training process and increase the number of trainable parameters, eventually leading to training instability (Wang et al., 2022a).
The results are as follows:

Example 1: (i) Perpendicular sensors: the prior and posterior distribution of the two PCs is depicted in Figure 5.10A. The first PC has a significant information gain, while the second PC has a smaller information gain (Figure 5.10A). Posterior samples are sampled from the bivariate distribution, back-transformed to the original space (mm/day), and their mean are computed (Figure 5.10B). The mode of the posterior distribution of mean fluxes is centered on the true value (depicted as the black line in Figure 5.10B).
(ii) Parallel sensors: the prior and posterior distribution of the two PCs is depicted in Figure 5.10C. The posterior distribution of the first PC is bimodal, indicating that the trained model is unable to distinguish between high and low fluxes, which is to be expected given that higher fluxes are located along river banks that are not sampled by the parallel sensors (Figure 5.10D), which is also evident in the posterior distribution of the mean fluxes (Figure 5.10D).
(iii) Random sensors: the prior and posterior distribution of the two PCs is depicted in Figure 5.10E. The posterior distribution of the first PC is wider than that of the perpendicular sensors, but it retains one mode.

Example 2: (i) Perpendicular sensors: the prior and posterior distribution of the two PCs is depicted in Figure 5.11A. The first PC's posterior distribution is bimodal, indicating that the trained model struggles to differentiate between medium and low fluxes, as shown by the posterior distribution of mean fluxes (Figure 5.11B). Nevertheless, unlike (ii) parallel sensors (Figures 5.11C-D) and (iii) random sensors (Figures 5.11E-F), the first PC's posterior distribution has one mode that is centered on the true value, indicating once more that the perpendicular sensors are the best choice.

Example 3: This case is distinct from the other two in that it is from outside the prior distribution of the training data (Figure 5.12A) and has a low flux (Figure 5.12B). (i) Perpendicular sensors: the prior and posterior distribution of the two PCs is depicted in Figure 5.12A. The trained PBNN is able to generate a sharp posterior distribution of both PCs, englobing the true value (Figures 5.12A-B).
(ii) When using parallel sensors, the trained PBNN is unable to distinguish between low and high fluxes (Figures 5.12C-D), in a situation similar to the second example (Figures 5.11C-D); the posterior distribution of the mean fluxes is bimodal (Figure 5.12D), and one of the modes englobes the true value (Figure 5.12D).
(iii) When using random sensors, the trained PBNN is the most accurate, as depicted by the concentrated posterior distribution of both PCs (Figures 5.12E), and the posterior distribution of the mean fluxes centered on the true value (Figure 5.12F).

It should be noted that in all of the preceding examples, a relatively high level of uncertainty is expected. Indeed, using only six locations does not allow for a complete coverage of the flux and K distribution variability (Figure 5.9). For example, if the cross-section is located in a low flux zone or a high flux zone, the mean flux estimation will be affected.

Testing all cases: The static BOED procedure is then applied to each combination of sensors (Figure 5.8), as described in §5.2.1, using the same dataset of 10,000 samples, which are split into 5 folds, i.e., $N_{train} = 8000$ and $N_{test} = 2000$. The training process is repeated for each of the 18 sensor combinations and for each of the 5 folds, totaling 90 training processes. The model is evaluated on the 2000 test cases for each training process, and the values of the utility function are recorded. The end result is a $5 \times 18 \times 2000$ matrix of utility function values. We then compare the performance of the different families, i.e., orthogonal, parallel, and random. The distinct families are then compared with histograms, as shown in Figure 5.13. The different histograms represent the distribution of utility function values for each of the 3 different families of sensor combinations. For comparison, we also include the histogram of the full combination of sensors, as shown in Figure 5.8A. Visually, there are few conclusions to be drawn from the histograms, as the distributions appear very similar. Therefore, in order to objectively rank the various families, we employ the same utility function defined previously (Equation 5.1 in §5.2.1). As a reminder, the minimum value of the utility function is 0. We use a Box-Cox transformation on the utility function values to ensure that the distribution is approximately normal, and then we calculate the KL divergence between the ideal and actual distributions.

The results are as follows:

Family	Divergence from ideal scores (nats)
Full	17.48
Orthogonal	29.32
Parallel	36.02
Random	32.29

Table 5.5: Note. Divergence from ideal scores for the different families of sensors.

The combination of all sensors has the best performance and is used as a benchmark for the other families. The orthogonal family has the best performance, followed by the random family, and the parallel family has the worst performance. These results are consistent with what was expected; the orthogonal family can capture temperature variations caused by high lateral fluxes, whereas the parallel family cannot, and the random family falls somewhere in between the two, capturing some of the lateral fluxes but not as consistently as the orthogonal family, as demonstrated by the previous three examples (Figures 5.10-5.12).

Due to the heterogeneity of the riverbed, the orthogonal sensor family is the optimal choice for the 3D scenario, and it makes little sense to search for the "best" sensor combination, as the optimal combination will likely vary for each riverbed.

We replicated the staBOED procedure using the same network architecture as in the 1D case, i.e., one hidden layer containing 16 neurons and one Gaussian component for each feature. The results are shown in Table 5.6, and they are consistent with the results obtained with a more complex network architecture (Table 5.5). However, the computational time is reduced because the network is significantly smaller. On the same computer, and without parallelization, training the network for each of the 18 sensor combinations for five folds took approximately 45 minutes, whereas training the more complex network for each of the 18 sensor combinations for five folds took approximately 1 hour and 15 minutes. Therefore, we suggest employing a simpler network architecture for the 3D case, as it is more computationally efficient and yields comparable results. However, the results obtained with the more complex network architecture remain valid, and the user is free to choose the network architecture.

Family	Divergence from ideal scores (nats)
Full	40.18
Orthogonal	247.21
Parallel	501.92
Random	269.67

Table 5.6: Note. Divergence from ideal scores for the different families of sensors, using a simpler network architecture.

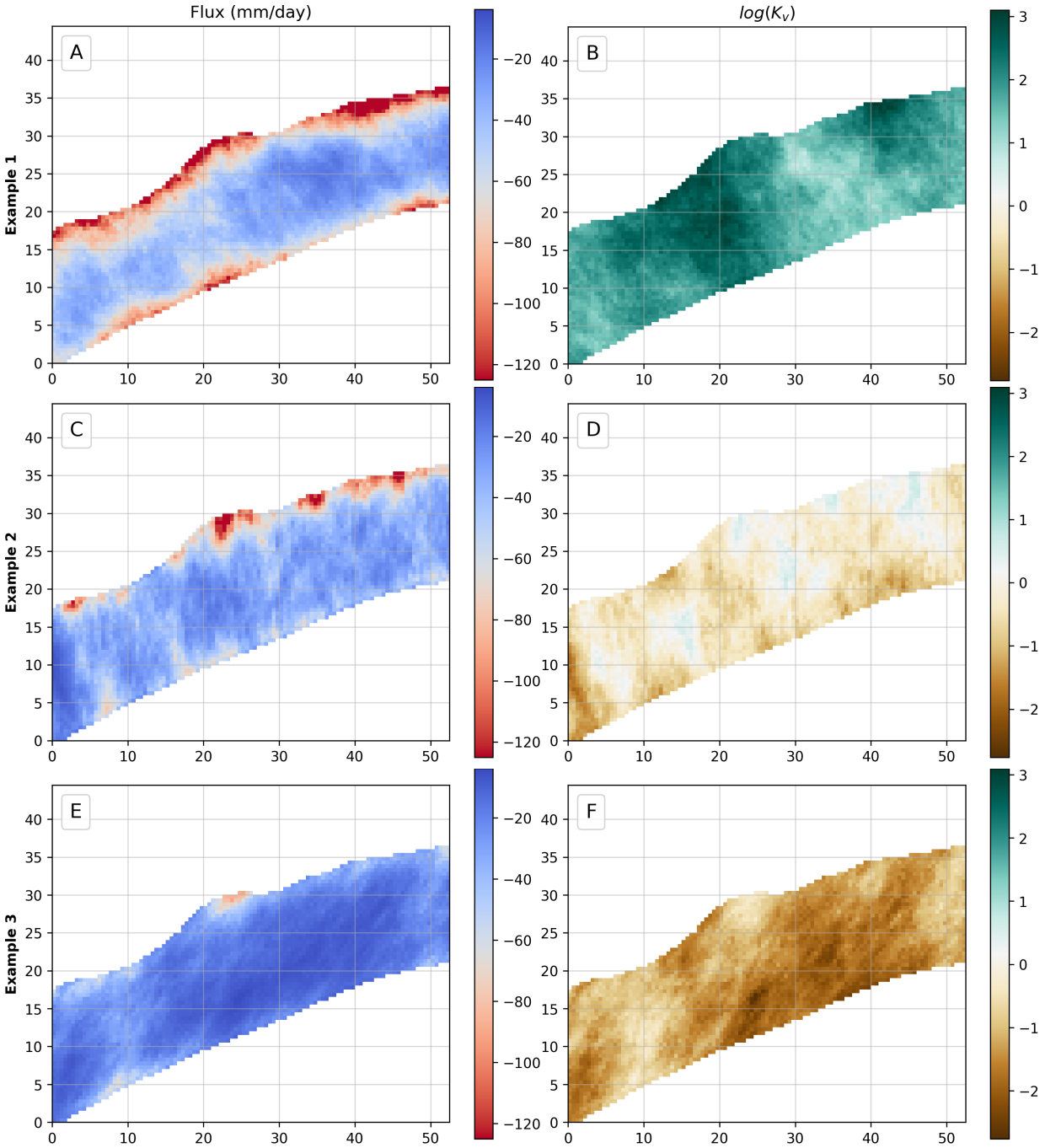

Figure 5.9: Three examples of the 3D scenario. Example 1: high fluxes (**A**) - high hydraulic conductivity (**B**). Example 2: medium fluxes (**C**) - medium hydraulic conductivity (**D**). Example 3: low fluxes (**E**) - low hydraulic conductivity (**F**).

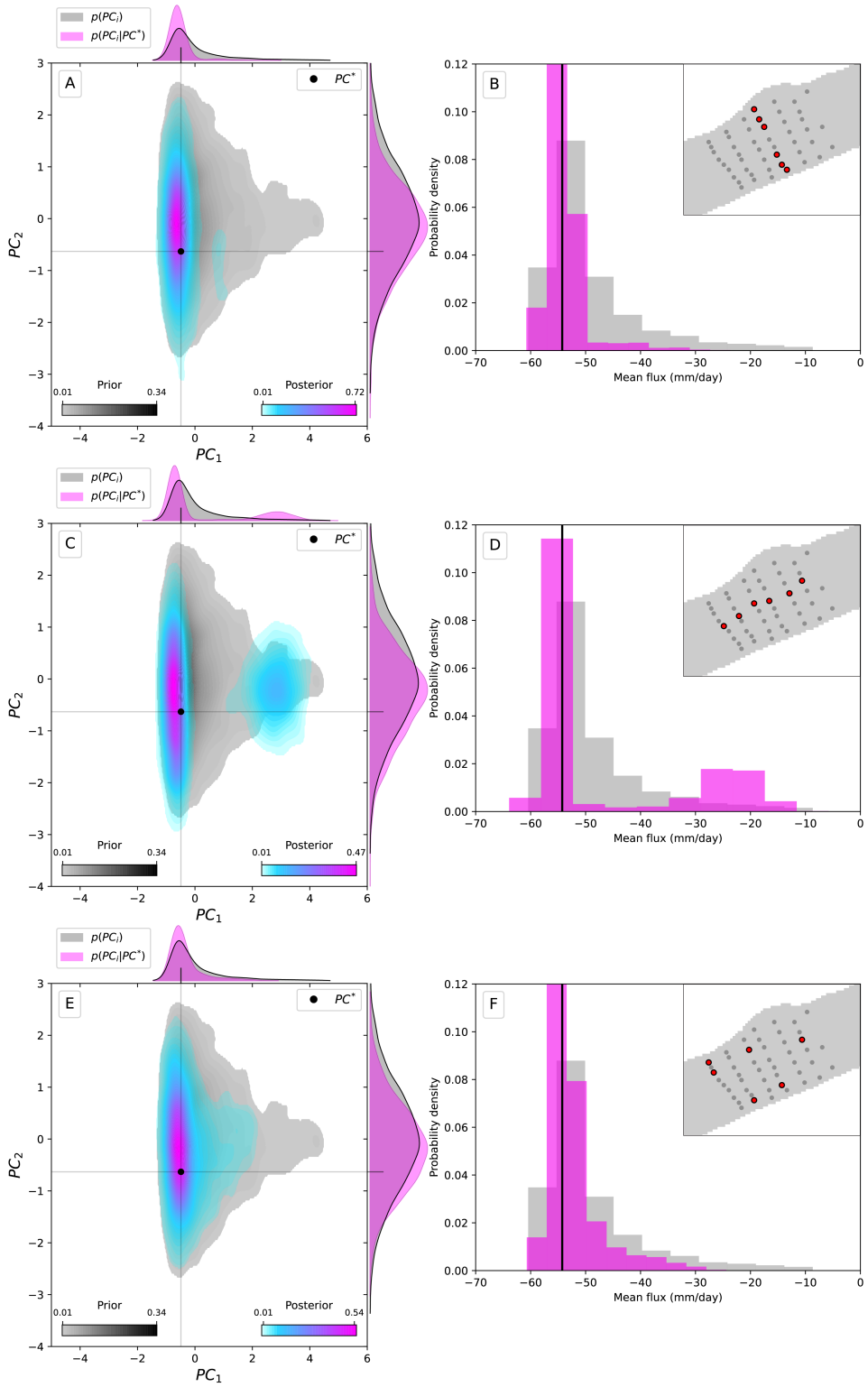

Figure 5.10: Example 1: Results for three sensor combinations. As shown in the background of the second column, each row depicts the results for a different sensor combination. The first column depicts the prior/posterior distribution of the flux field's principal components, while the second depicts the prior/posterior distribution of the mean fluxes. The vertical line in Figures **B**, **D**, **E** indicates the value of the true mean of the flux.

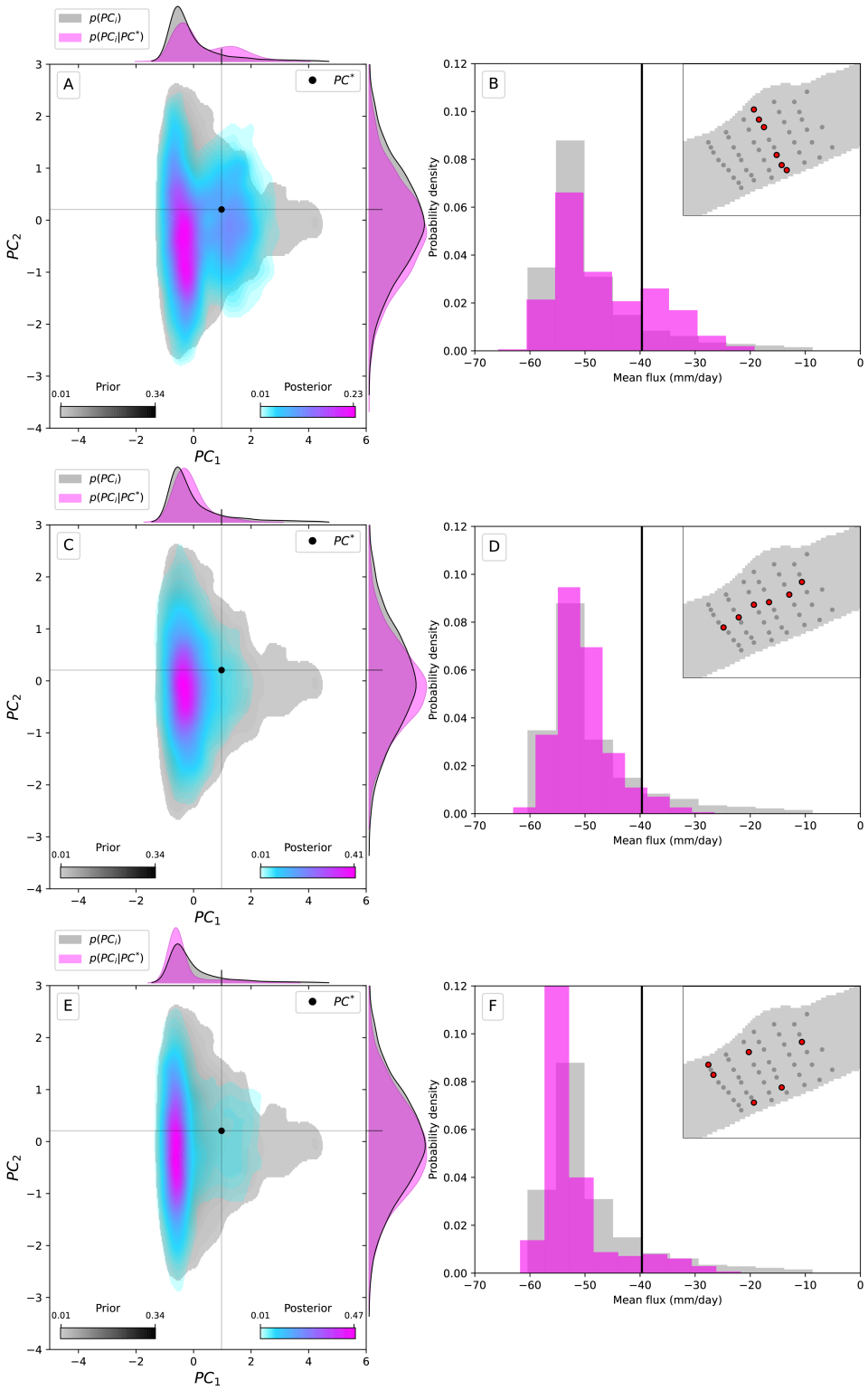

Figure 5.11: Example 2: Results for three sensor combinations. As shown in the background of the second column, each row depicts the results for a different sensor combination. The first column depicts the prior/posterior distribution of the flux field's principal components, while the second depicts the prior/posterior distribution of the mean fluxes. The vertical line in Figures **B**, **D**, **E** indicates the value of the true mean of the flux.

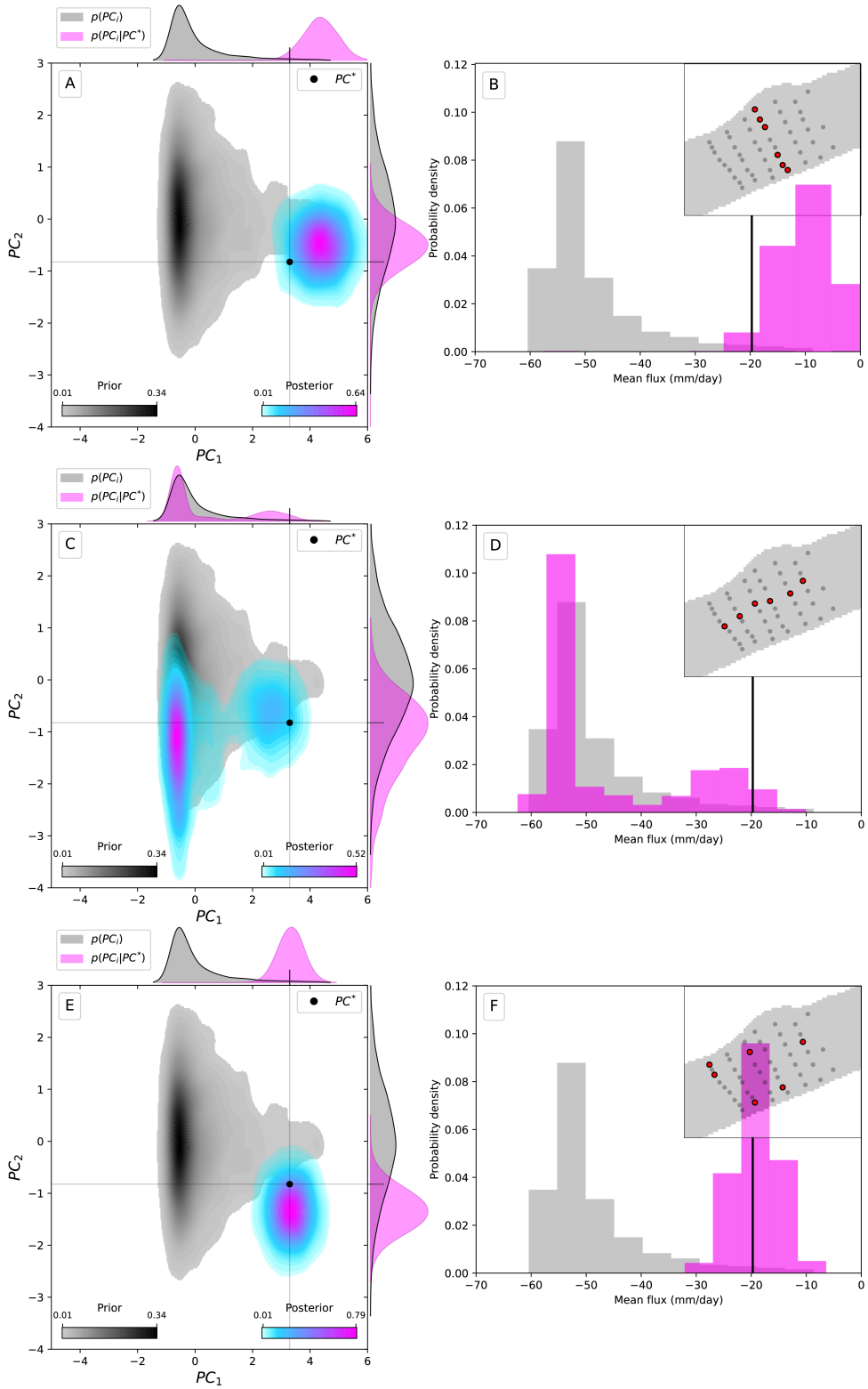

Figure 5.12: Example 3: Results for three sensor combinations. As shown in the background of the second column, each row depicts the results for a different sensor combination. The first column depicts the prior/posterior distribution of the flux field's principal components, while the second depicts the prior/posterior distribution of the mean fluxes. The vertical line in Figures **B**, **D**, **E** indicates the value of the true mean of the flux.

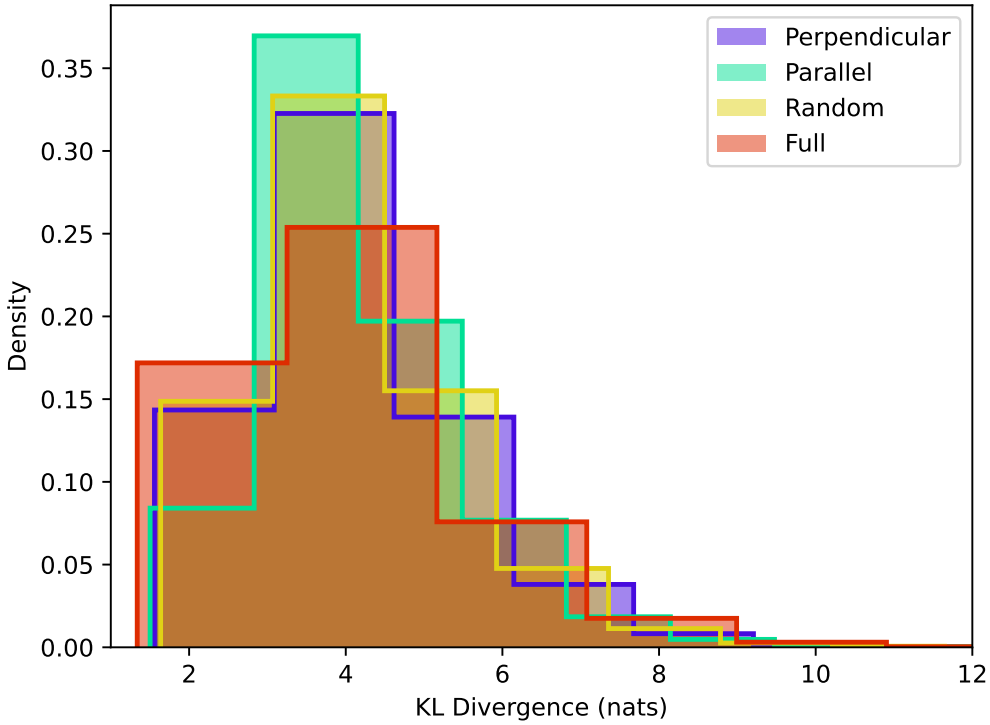

Figure 5.13: KL divergence between the ideal and actual distributions for the three sensor families and the full combination of sensors.

5.4 Discussion

In this contribution, we propose novel methodologies in the context of Bayesian optimal experimental design (BOED) for estimating river-aquifer exchange fluxes from subsurface temperature measurements. We use probabilistic Bayesian neural networks to propose data-efficient solutions to BOED. The proposed methodologies rely on iteratively training a probabilistic Bayesian neural network on experimental data and prior knowledge, which is updated at each iteration in the case of sequential experimental design (seqBOED) or remains constant in the case of static experimental design (staBOED).

The 1D case demonstrates that the proposed scheme permits an efficient sampling of the posterior distribution of hydrological parameters from subsurface temperature measurements. In this instance, the forward model is a simple analytical solution to the heat equation, which is computationally inexpensive. In a number of applications, however, the forward model is a computationally expensive numerical model, and Monte Carlo sampling of the posterior distribution of the fluxes is not feasible in a reasonable amount of time (e.g., Chapters 3 and 4). Furthermore, the proposed scheme is robust to

slightly out-of-prior experimental data, which is an advantage for real-world subsurface applications. Therefore, we believe that this work will provide new opportunities for designing data-efficient subsurface experiments.

In the 3D case, we demonstrate that the proposed methodology can find the optimal experimental design for a complex numerical model from sparse, spatially distributed temperature measurements. The results are consistent with expectations, and in particular, as demonstrated with an out-of-prior sample, the proposed PBNNs are equally efficient as the 1D case at expressing *"I don't know"* in the form of uncertainty, which is a key feature of the proposed method.

In addition, we present a novel metric for evaluating the performance of BOED schemes, though it is not limited to this purpose. The metric is based on the Kullback-Leibler divergence and is useful not only for evaluating individual experiments, but also for comparing the performance of different experimental designs. It does, however, require that the experimental results be available in the form of a probability distribution, which is not always the case. However, approximate distributions, such as the Gaussian distribution, can be used after performing a Box-Cox transformation on the experimental data if they are not normally distributed.

One limitation of our approach is that retraining the PBNN after each iteration for each test sample is time-consuming. Moreover, in seqBOED, when updating the prior with the posterior, the network may "forget" or, more accurately, "miss" patterns in the data that were previously learned. Further research is needed to leverage the concept of transfer learning to improve the performance of seqBOED. Transfer learning is a machine learning technique that allows for the transfer of knowledge from a previously trained network to a new network, and could allow for faster training times and better accuracy (e.g., Zhuang et al. 2020).

It should also be noted that for both the 1D and 3D case, the steady-state assumption is not always valid, and transient effects can either be neglected or taken into account by the numerical model. In this contribution, we focus on the steady-state assumption for the sake of simplicity, but the proposed methodologies could be applied to transient systems with appropriate numerical models. In particular, the transient case could increase the complexity of the optimal design sequences since the maximum or minimum of the temperature profiles could occur at different depths depending on the time of the year. The proposed methodologies could take such transient effects into account by using numerical models, and studying such effects is an interesting direction for future work.

Finally, we draw parallels between this work and a recent paper by Wang et al. (2022c), which treats a groundwater contamination remediation problem as a partially observable Markov decision process (POMDP), a promising avenue for subsurface applications. Sequential decision problems where the effects of our actions are uncertain can be represented by Markov decision processes (MDPs), where the goal is to find the

optimal policy that maximizes the expected reward (Wang et al., 2022c). When dealing with partially observable MDPs (POMDPs) where the system state is partially observable, the approach is to represent the system state with a "belief state," which is simply a probability distribution over the state space.

Wang et al. (2022c) proposed a decision-making framework that optimizes the trade-off between information gathering and the performance of possible future scenarios in a groundwater contamination remediation problem, their goal being to maximize the total utility of remediation wells placed in the subsurface. This is analogous to our problem, where the goal is to find the optimal sampling sequence of depths that minimizes the expected objective function $\mathcal{J}(\mathbf{h})$ (Equation 5.1). Therefore, our problem can be formalised as a POMDP; the system state is partially observable since we do not have access to the true value of the target variable $\mathbf{h}$, but we can only observe the predictor variable $\mathbf{d}$ which are temperature measurements.

In general, however, exact methods for continuous POMDPs cannot be defined; instead, some approximate methods seek optimal policies by sparsely sampling the entire set of trajectories (Wang et al., 2022c). For instance, as reviewed by Wang et al. (2022c), approximate methods have been developed, such as Monte Carlo tree search (MCTS; Chaslot et al. 2008) and determinized sparse partially observable tree (DESPOT; Ye et al. 2017) search. POMDP solutions schemes suffer both from the curse of dimensionality and the curse of history. This is because POMDPs must keep track of the entire history of actions (e.g., observations), which may be impractical for large-scale real-world problems. In comparison, our proposed methodologies are specifically tailored for BOED in the context of subsurface temperature measurements, and our approach follows the path of maximum information gain without considering the impacts of future actions. In practice, our proposed methodologies could be combined with POMDP methods; however, the purpose of this chapter was to demonstrate the potential of our methodologies in BOED, which finds the next best sampling locations given the current state of knowledge.

To summarize, potential future projects could include the following:

1. Leverage transfer learning techniques to improve the performance of seqBOED.
2. Solve transient systems with the proposed methodologies.
3. Integrate the proposed methodologies with POMDPs for sequential decision making.
4. Study the impact of the prior parameter distributions on the optimal design sequences.

5.5 Conclusion

We proposed data-efficient schemes to solve optimal experimental design problems in the subsurface by utilizing the capabilities of probabilistic bayesian neural networks. We also

introduced an information-theory-based distance function based on the Kullback-Leibler divergence to measure the distance between one ideal and one experimental distribution. This metric is not only useful for evaluating individual experiments, but also for comparing the effectiveness of different experimental designs. We demonstrated the applicability of our approach by using temperature measurements to estimate river-aquifer exchange fluxes in a 1D analytical model and a 3D numerical model, and found that the proposed methodologies are robust to out-of-prior experimental results. The results indicate that the proposed scheme reduces the computational costs of running the forward model and allows for more efficient sampling of the target parameters' posterior distribution. Our proposed methodologies provide a promising tool for designing data-efficient subsurface experiments, which is critical to informed decision-making and sustainable subsurface resource management, and we believe that the proposed methodologies will open up new avenues for designing data-efficient subsurface experiments. We intend to extend our methodologies to the combination of POMDPs for sequential decision problems in future work. In addition, we intend to implement transfer learning in seqBOED to reduce training time and enhance performance.

6. Discussion and outlooks

This dissertation has focused on the role and importance of machine learning for experimental design in the subsurface, and we have proposed a number of approaches to integrating information theory with hydrogeology-related research and decision-making. We have also shown how these approaches can be used to improve the understanding of the physical processes and their interactions on the subsurface. The general message is that there is no one-size-fits-all approach to experimental design, and there is no substitute for the expertise and judgment of a geoscientist. In this chapter, I will reflect on the results of this research and discuss how they can be applied in other domains. In doing so, I will go over the role of information and models in hydrogeology. I will also examine how machine learning and artificial intelligence can be used to inform decisions. Finally, I will discuss how experimental design can be improved in the future, and give some perspectives on Bayesian Evidential Learning.

6.1 The role of Information in hydrology

> *Information is the resolution of uncertainty.*
>
> Claude Shannon

As we have seen throughout this dissertation, the role of information in hydrology is complex and of utmost importance. It can be observed in numerous aspects of hydrological modeling, including model development, decision-making, and water resource management. The application of information theory in hydrology has a long history, and its significance in hydrological research has grown in recent years. Therefore, I want to emphasize at this point that I agree with the debate series' posed question, "Does Information Theory Provide a New Paradigm for Earth Science?" (Goodwell et al., 2020; Gupta and Nearing, 2014; Kumar and Gupta, 2020; Nearing et al., 2020; Perdigão et al., 2020; Weijs and Ruddell, 2020). As we have come to understand, information is the bedrock upon which all other foundations for decision-making are built.

Despite not possessing the ability to directly observe many important and influential water processes, we are constantly bogged down with the task of interpreting data, developing models and making decisions, communicating this decision-making. During times of climate change and rapid societal developmental growth and changing land-use,

decisions that are to be made become increasingly complex. As such, it is important to arrive at these decisions through a process based on information and understanding. When speaking of water decision-makers and the role of data, scientists and specialists in the field have long recognized the importance of information and communication: "*We hydrologists can do a better job of supporting water-resources decision-making*" (Ferré, 2017). And as posited by Grey Nearing and co-authors: "*How much information do we have, and how well do we use it? When the problem is posed in this way, we are able to clearly delineate fundamental from practical limitations to our ability to improve our knowledge of hydrological systems—and, indeed, probably all natural systems.*" (Nearing et al., 2016).

Information theory provides a holistic approach to understanding the role of information in the field of hydrology. For example, the process of model parameter estimation, which is a key component in both model development and model calibration, is essential to incorporate an uncertainty-based approach on an experimental design that considers information content. In this approach, the experiment that is being designed is concerned with the amount of information that is to be gleaned over a predefined region of space and period of time to estimate the values of a pre-defined set of model parameters or model outputs. Pre-defined set refers to the probability distribution of parameter values–this is the prior probability distribution.

The prior probabilities reflect our certainty in the model parameters (or outputs) before the experiment is conducted. As Lindley stated in 1956 in his seminal paper, *"prior distributions, though usually anathema to the statistician, are essential to the notion of experimental information"* (Lindley, 1956), and no experiment can be informative, for example, if the prior distribution is highly skewed toward a single parameter value (i.e., the state of nature is known). On the other hand, if the prior probability distribution is uniform and large, then the experiment is expected to provide a large amount of information, which can be used to improve the accuracy of the model parameters. The idea is to find the experiment that provides the most "information" and then use the information to reduce the uncertainty in the model parameters. In this sense, information theory can be used to balance the utility of available data and the cost of collecting data in various experiments.

As suggested by Nearing et al. (2013b), I believe that further research is needed to understand how information penetrates and propagates through both hydrological and machine learning model in order to enable better estimates of states that are only indirectly related to observations. Nearing et al. (2013b), for example, demonstrate how the effects of "good" and "bad" information can be explicitly measured in the context of an observing system simulation experiment. "Good information" will shift our certainty in favor of more accurate and precise simulations of system behavior, whereas "bad information" will do the opposite (Gupta and Nearing, 2014). Nearing et al. (2013b) also note that it is challenging to calculate the mutual information between probability distributions of high-dimensional random variables due to the dimensionality curse (Bellman, 1961). According to Nearing et al. (2013b), if multiple types of observations (possibly

from different sensors) or observations were available in different locations, each observation dimension would have to be analyzed separately, and in some cases, projecting sets of concurrent observations into low-dimensional space and then estimating the mutual information between these projected observations and model states may be beneficial, which is exactly what we did in this study.

The chronological sequence of papers prepared as part of this doctoral research exemplifies the importance of information in hydrology and geophysics.

Appendix B: The topic of my first publication was relating geological and geophysical data to ERT inversion results in order to obtain the most accurate image of the subsurface resistivity structure, taking into account the available information and its geological interpretation. In this study, information was viewed solely as raw data to be processed and incorporated, and its function was to lay the groundwork for model selection. Compared to the other papers, this one took a more "traditional" approach to inverse problems, in which the role of information was to provide a priori knowledge to the inversion process.

Chapter 3: The first research chapter in this dissertation was my first attempt to understand the role of information in hydrology. We assessed the information content of the observed variables by comparing the predicted targets to their "true" values, which is the most intuitive way of doing so. The approach was sound, and the results were coherent, but the curse of dimensionality was felt. To calculate the observed data's information content ranking, the predicted targets had to be back-transformed to their (high dimensional) original space, and their dissimilarity to the true values had to be calculated using an image similarity metric.

Chapter 4: The second research chapter in this dissertation dealt with a four-dimensional hydrological model, and we attempted to assess the information content of the observed variables using the same approach as in Chapter 3. However, the image similarity metric was inadequate for such a high-dimensional space, so we had to use a different tactic. Recognizing that the information content of the variables is stored in their principal components, we calculated the information content of the observed variables by computing the RMSE of the predicted targets' principal components to their "true" values. This reexamination of how information content can be measured corresponds directly to the discussion in Nearing et al. (2013b). Despite the fact that the principles of information theory were overlooked, the method is sound. As mentioned in Nearing and Gupta (2015), RMSE is a common statistics for the measure of information.

Chapter 5: The third and last research chapter in this dissertation benefits from the experience gathered in the first two papers. Information theory is used more formally in order to understand the behavior of a machine learning model. The first two papers provided a framework for calculation of the information content of the observed variables. We used the same approach as in Chapter 4, but used the KL divergence instead of RMSE to measure the dissimilarity between the predicted

targets' principal components and their "true" values. It makes the experimental design faster and more efficient, and more grounded in information theory.

In conclusion, I maintain that information is critical to understanding, designing, and choosing an accurate hydrological model, and that the role of information theory in hydrology is still in its infancy. I believe that the research presented in this dissertation represents a step forward in the use of information theory and machine learning in hydrology.

6.2 The role of models in hydrology

> *All models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.*
>
> George E. P. Box

Models are an integral part of all scientific disciplines, including hydrology. They are essential to comprehending water's physical and chemical processes. Hydrological models are necessary for incorporating physical processes, such as convection, dispersion, evaporation, infiltration, and groundwater flow, into a mathematical numerical model of water flow or solute transport in order to predict and provide decision support for improved water resource management.

We strive to be as accurate as possible when building models. However, they are rarely exact in the sense that they cannot account for all the underlying phenomena. As Cartwright (1983) points out, models always have simplifications: *"A model is a work of fiction. Some properties ascribed to objects in the model will be genuine properties of the objects modelled, but others will be merely properties of convenience."* The simplifications are a consequence of our limited knowledge about the universe, which was described by Rosenblueth and Wiener (1945) when they stated that *"no substantial part of the universe is so simple that it can be grasped and controlled without abstraction"* (Rosenblueth and Wiener, 1945). With few exceptions, the goal of modeling is to simulate a system (or a part of a system) whose behavior is already well enough understood to be simulated by the model. Modeling and simulations are frequently carried out in isolation. For example, climate simulations of coastal erosion do not take human-driven effects like urbanization and mining into account, and even then, they are only consistent over a limited geographic area and timescale (Lavin et al., 2021).

Models not only aid in our understanding of nature, but they are also used for prediction. We also build models that predict future events that are not fully understood. Nearing and Gupta (2015) provide a good basis for quantifying the role of models in science, including their role to induction and their capacity to make accurate predic-

tions. I will adopt a similar definition of model as they do: *an idealized representation of a real-world system used to concretize abstract concepts in a theory and simplify the scenario to make the theory tractable* (Cartwright, 1983; Frigg and Hartmann, 2020).

In this dissertation, we use models to determine which events are the most important to *observe* in order to gather the most information. Since models are the containers for our theories, they determine what we observe and how to interpret the observations. Consider the case of Chapter 3 and imagine using a deterministic approach instead of our BEL approach. In order to obtain the *best* or *most truthful* wellhead protection area, we would have to use all the available data in order to calibrate the hydraulic conductivity of a real-world aquifer. We would still have needed to predict the underground flow of a contaminant plume by simulating the flow of water and the transport of the contaminant. The model's parameters could be constrained by measurements in the area, and calibrated by observing the contaminant plume in the field. Any measurement that would be taken at the Earth surface is governed by physical laws that we not fully grasp, and contaminant transport theory dictates how we interpret the measurements. The concentration curve of the observed contaminant plume in the model will never be identical to the concentration curve of the observed contaminant plume in the field (e.g., Hoffmann et al. 2019), and the experimenter must determine how to account for the difference.

The problem in this case would be that the *true* model is unknown, and we must use the best available model to make predictions. In taking decisions that could have a societal impact, we must be able to quantify the uncertainty in the model. A well-trained geoscientist will be aware of the model's limitations, such as coarse spatial and temporal resolution, approximations, and regions less constrained by observations, so the modeler can decide how to proceed. This is the reason why the modeler will always implicitly treat such deterministic models probabilistically; otherwise, the models are always false. However, a complete and rigorous quantification of the uncertainty would still be lacking.

In other words, deterministic modeling or forecasting gives the modeler a lot of latitude for interpretation (Weijs et al., 2010). Nearing et al. (2016) further emphasize that the models must be probabilistic in order to be unfalsifiable. For example, if we had applied the BEL framework to the case of Appendix B, we would have had much less interpretational freedom about the model's limitations, but the uncertainty would have been much more explicit–at the cost, of course, of having to define a prior distribution for the model parameters and several forward modeling simulations. The Depth of Investigation (DOI) would have been irrelevant (although a comparison with the spatial uncertainty of the posterior distribution may be interesting). Furthermore, the difficulties of selecting the correct set of hyperparameters would have been significantly reduced. However, whether in hydrology or geophysics, the role of experts remains critical, and interpretive freedom is still required–an aspect of science that I don't see changing anytime soon.

In conclusion, hydrological models are crucial elements of our understanding of hydro-

logical phenomena. Despite their limitations, they provide the basis for decision-making and predictions based on incomplete data and knowledge. In the future, the role of models in hydrology is likely to evolve. Gupta and Nearing (2014) anticipate a shift in modeling's focus toward the more creative facets of scientific investigation. By adopting an information theory-based viewpoint, they propose a systems theoretic framework to enhance our capacity to inform the discovery/learning process (and thereby hydrologic science) from the juxtaposition of models and data. The shift in emphasis, in my opinion, is from deterministic to probabilistic modeling, and from a single model to an ensemble of models. This shift in emphasis is already taking place, and it is being driven by the increased availability of data and the increasing computing power that allows us to create larger ensembles of models. To accommodate these changes, we must reconsider our modeling tools and methods.

In any case, the role of models in hydrology continues to evolve, and it is important to consider the philosophical implications of modeling and simulations. For example, how do we identify when a model is not performing well? The modeler is responsible for determining whether a model's performance is satisfactory. This is a difficult task because the modeler must validate the theory as well as the model. In other words, the modeler must be aware of the theory's limitations, but how can we determine if the theory is incorrect or incomplete? This question leads us to another: *alternatively, rather than using an existing theory, can we infer our theory directly from observations using data-driven approaches?* This question, which has some philosophical implications (e.g., Popper 2005), is the subject of the next section.

6.3 The role of Machine Learning and AI in hydrology

An algorithm must be seen to be believed.

Donald Knuth

Machine learning brought the idea of data-driven science to the forefront of scientific research, sparking debates between data-driven and model-driven science. As mentioned in Chapter 1, some Earth scientists may believe that ML models lacking explicit process representation are untrustworthy (Nearing et al., 2021; Solomatine and Ostfeld, 2008). In my opinion, this debate is moot, and it is better to combine both approaches. I contend that the terms *data-driven* and *model-driven* are misleading. The ML approach does not replace physical modeling. Instead, it is complementary to the physical modeling.

For example, in the presented papers, we used different sort of forward modeling to generate the training data, and then used ML to predict the targets. One could argue that our approach is data-driven, but it relies on hundreds or thousands of physics-based model runs to generate training data, making BEL, in a sense, the ultimate model-driven approach. Rather, I see BEL as a hybrid approach that combines the best of both worlds. We can achieve a better understanding of the system and make more accurate

predictions–even better, we can quantify the uncertainty of the predictions–by combining knowledge from the physical model with the "data-driven" approach. In fact, BEL fits with the *"Theory-Driven Data Science"* approach described in Karpatne et al. (2019).

Purely data-driven approaches are also useful. For example, Kratzert et al. (2018, 2019a) used time series of observations to train and evaluate their LSTMs models at ungauged basins. In a large sample study spanning 30 years and hundreds of basins across the continental United States, Kratzert et al. (2019a) trained and evaluated multiple LSTMs models. The benchmarks used in that study were derived from a conceptual model calibrated independently for each basin, and from a state-of-the-art process-based model that benefited from several million dollars in development funding. When calibrated to long data records in gauged basins, LSTMs provided better average daily streamflow predictions in ungauged basins than traditional hydrology models (Kratzert et al., 2019a; Nearing et al., 2021). They concluded that the claim that process-driven models are preferable in out-of-sample situations may not be valid–machine learning methods are effective at extracting information from large, diverse data sets under a variety of hydrological conditions.

In other words, the trained LSTMs acted as *models*, and they were able to make predictions that were as good as the state-of-the-art process-based model, which answers our aforementioned question. The lack of "physics-awareness" could be one of the drawbacks of the LSTMs approach. In this context, Kratzert et al. (2019a) suggest incorporating physical catchment properties as an additional input layer into the LSTM to improve predictive power. Data for subsurface hydrology are frequently scarcer than for surface hydrology, and training such LSTM models may be difficult. The proposed approach with BEL, on the other hand, limits the need for training by using synthetic models, and the number of those models is also limited, limiting the computational time. As a result, BEL, as a hybrid approach, is particularly well suited to subsurface hydrology, where data is frequently scarce and physical processes are complex.

The last point brings us to the topic of *understanding*. To some extent, the debate between data-driven and model-driven science revolves around whether data-driven models can provide a *better understanding* of the system than model-driven approaches. The answer to this question is "it depends". As explained in this chapter and demonstrated by Kratzert et al. (2019a), data-driven models are able to make accurate predictions. However, accuracy alone does not guarantee understanding. This is due to the fact that data-driven models are "black boxes," meaning they do not provide a physical explanation for the phenomenon. In this regard, a physical model is still required to explain the system's behavior. On the other hand, Beven (2020) argues that a different perspective would be to use deep-learning models to enhance process information and *understanding*, which I think is a promising direction for future research.

What does *"understanding"* mean in this context? Only when referring to human agents, either explicitly or implicitly, is the term "understanding" appropriate: a scientist **S** understands phenomenon **P** with theory **T** in hand (De Regt and Dieks, 2005).

De Regt and Dieks (2005) go on to say that scientists (in a context **C**) can understand scientific theory **T** if they can recognize qualitatively characteristic consequences of **T** without performing exact calculations.

Let us use Darcy's law as an example. An hydrologist **S** understands that the Darcy's law **T** relates water flow **P** in a porous medium to permeability and pressure gradients **C** if they can recognize that increasing pressure gradients increases the velocity of the flow without performing exact calculations.
A machine-learning model, on the other hand, lacks this capability. The goal of machine learning is not to explain; rather, it is to accurately predict the outputs from the given inputs.

Carl Hempel argued that *understanding* is subjective and just a psychological by-product of scientific activity, so is irrelevant to the philosophy of science (Hempel et al., 1965). I think that these arguments encapsulate the essence of the debate between data-driven and model-driven science; scientists do not like to rely on what they see as "black-box" models (see the "oracle" example in De Regt and Dieks (2005)), they need a more intuitive understanding of the system. Personally, I believe the term "black-box" is misleading; machine learning models do not lack physical meaning; rather, they often encode physical relationships in a more abstract form than traditional models. The challenge is to extract physical meaning from trained models. This can be accomplished by investigating the model's input-output relationships as well as its weights and biases. These techniques can provide information about the relationships between variables as well as the physical processes that underpin the system.

In conclusion, I believe that one of the best approaches to hydrology is a hybrid approach that combines physical models and machine learning. The challenge lies in finding the ideal way to combine the two methodologies. We've seen in this dissertation how the BEL approach allows us to use both physics-based and data-driven knowledge at the same time. In this sense, BEL is a novel approach to hydrology with the advantage of providing a more intuitive understanding of the system, which can alleviate the discomfort associated with solely relying on data-driven models. In this regard, I hope that this work contributes to the debate.

Finally, given the rapid evolution of ML and AI, it is worthwhile to discuss how AI can be used to assist scientists understand complex systems such as hydrological phenomena. Recently, Krenn et al. (2022) discuss "scientific understanding," how scientists can acquire it, and how AI can assist humans in gaining new scientific understanding. They argue that there are three ways in which an AI system can advance our understanding of science (Krenn et al., 2022):

1. As a "computational microscope," it can provide data that isn't yet possible to obtain experimentally.

2. As a "resource of inspiration" or a "artificial muse," enhancing human creativity and imagination.

3. As a "agent of understanding," taking the place of humans in generalizing observations and applying these new scientific ideas to various phenomena while, crucially, sharing these revelations with human scientists.

Regarding the first point, I believe that the BEL approach is uniquely suited to this task due to its hybrid nature. As previously stated, forward modeling in BEL enables us to obtain data that would otherwise be impossible to obtain experimentally and then use the data to train and evaluate our models. In this regard, BEL can function as a "computational microscope," allowing us to gain insights into complex systems that would have been impossible to observe otherwise.

The current proliferation of Large Language Models (LLMs) could bolster the second point in the near future. For example, the GPT (Generative Pre-trained Transformer) family of language models, which includes chatGPT and GPT-3 (Brown et al., 2020), or the Galactica LLM (Taylor et al., 2022), are trained on large datasets of human-generated text and can produce human-like text as output. By suggesting creative ideas or new approaches to problems, these models can be used as a source of inspiration or an artificial muse. A scientist, for example, could use a LLM model to generate a list of potential hypotheses for a research question, which the scientist could then evaluate and test further.

LLMs can be used to help with tasks such as summarizing and synthesizing information, which can aid scientists in better understanding complex concepts or ideas. A scientist, for example, could use a LLM model to create a summary of a research paper or to review a set of experimental results, providing a concise and coherent overview of the key points.

Overall, LLMs have the potential to augment scientists' creativity and imagination, allowing them to generate new ideas and insights that would not have been possible without the help of AI.

The third point is the most controversial: can AI really be used as an *agent of understanding*, taking the place of humans in generalizing observations and applying new scientific ideas (e.g., George and Walsh 2022)? The answer, in my opinion, is no: AI is not yet capable of comprehending complex scientific phenomena in the same way that humans can. Rather, I believe that AI systems should be seen as tools that can aid human understanding by performing tasks or analyzing data automatically, or by providing new insights or ideas from large datasets. The first point is therefore, in my opinion, currently the most pertinent in the context of hydrology and applied sciences in general. In this context, BEL serves as a tool for better understanding hydrology phenomena and assisting in the application of physical models.

The success of machine learning models relies heavily on having a good representation of the underlying patterns in the data, that are likely to be too subtle for our human brain to detect, but are real nonetheless. Our best theories (e.g., Darcy's law (Darcy, 1856),

Newton's law of gravity (Newton, 1833), Einstein's theory of relativity (Einstein, 1905), quantum mechanics (Feynman et al., 2010)) are the result of trial and error throughout the history of human civilization, each based on previous theories, and gradually converging closer to a better understanding of reality, in a manner similar to how machine learning models are trained–but this analogy shouldn't be stretched too far (Bostrom, 2003). The success of the current theories is proof that there are underlying patterns, and that they are worth pursuing (e.g., Krenn et al. 2022). Some of these patterns may have an absolute, inherent existence, while others may be an artefact of our limited human perception. The role of machine learning is to learn the underlying patterns and discern which of them are relevant, and it can only do so with a good representation of these patterns in the training data, which brings us back to the importance of defining both a suitable prior and a suitable learning algorithm in BEL.

One aspect of BEL that is relevant to this discussion is that we have always used a physical model to generate training data. I believe the developments made in this work have great potential when using pure data. We now have a more flexible approach to learning patterns in data thanks to the incorporation of Probabilistic Bayesian neural networks into the BEL framework in this dissertation, and we can use the learned patterns to make predictions with uncertainty quantification. Before training the model with deep learning, it would be interesting to use CCA as a tool to estimate the amount of mutual information that exists between the predictor and the target variables.

According to Nearing et al. (2021), the future of hydrology will be a mix of AI and physics-based approaches, and Kratzert et al. (2019a) opine that theory-driven data science (Karpatne et al., 2019) will likely be the most successful approach moving forward. I concur with these opinions, and I believe that the BEL framework is a step in the right direction.

6.4 Experimental design in the subsurface

Those who ignore Statistics are condemned to reinvent it.

Bradley Efron

The backbone of this dissertation has been determining the best way to extract the most information from the available data and developing metrics that can be used to evaluate the performance of the models within the constraints of the available data. Once such a framework has been established, we can use it to evaluate the performance of models given spatially distributed data in a variety of designs–a process known as experimental design.

In the hydrological sciences, there is discussion regarding the need for improved experimental design to address problems in data-scarce environments (Beven et al., 2020; Chacon-Hurtado et al., 2017). For example, Beven (2019) asserts that we must be much

more cautious in considering the value of available data in model evaluation, and that better observational techniques are required. The spatial aspect, which includes spatial uncertainty, spatially distributed information, and spatial decisions, distinguishes the application of experimental design to the earth sciences from other disciplines (Eidsvik et al., 2015; Müller, 2007). From my experience over the past four years, Earth sciences have a shaky experimental design tradition and a bewildering array of technical terms.

Eidsvik et al. (2015)′s textbook is a comprehensive book on the subject in Earth sciences, although focusing on the decision theoretic notion of value of information (VOI), which they use to evaluate and analyze various sources of data. However, links to the roots of experimental design are absent, and the terminology is not always consistent: for example they use terms such as "data gathering schemes," "data acquisition schemes," "information gathering scheme," and "optimal strategy". Another textbook written by Müller (2007) entitled "Collecting Spatial Data" is an excellent resource for learning more about the topic at hand. However, when it comes to experimental design or "data gathering schemes," this reference and the theory it describes are frequently overlooked in the hydrology literature.

In my search of the scientific literature, I also found terms such as "Bayesian decision theory" (Davis et al., 1972), "optimal decision making" (Ogie et al., 2017), "optimal allocation decisions" (Kaune et al., 2017), "optimal monitoring network" Chen et al. (2022). The unifying concept of these terms seems to be deriving optimal (under a specific criterion) locations for an experimental campaign or strategy for observing a phenomenon, be it spatial observations, space-based observations, or field experiments such as those aimed at data assimilation. The issue is that the terms are not used consistently, and the literature is scattered across a variety of disciplines, which make it difficult to find relevant information (see also Rahman et al. 2022; Stein et al. 2022). It is most likely due to the fact that it is a multidisciplinary field, with experiment goals frequently varying greatly from one problem to the next.

In surface hydrological sciences, the design and evaluation of hydrometric networks best exemplifies the use of "experimental design" or some variant thereof, as reviewed by Chacon-Hurtado et al. (2017). Remarkably, in this context, Chacon-Hurtado et al. (2017) begin by acknowledging that experimental design and sensor network design employ the same concepts. Chacon-Hurtado et al. (2017) further state that due to the variety of cases, criteria, assumptions, and limitations, the scientific community does not appear to have reached an agreement on a unified methodology for sensor network design. As hydro(geo)logical science advances, I concur that we should adopt more consistent terminology and a more unified approach to experimental design in hydrology, as well as improve observational methods for testing process representations and gaining a better understanding (as also expressed by Beven et al. (2020)).

I can move in this direction by proposing guidelines for what constitutes a good Bayesian optimal experimental design in the subsurface using concepts from other disciplines. I am well aware of statistician Bradley Efron's adage that "those who ignore

statistics are condemned to reinvent it" (Friedman, 2001), and I am not trying to reinvent the wheel—just to make it roll more smoothly. The goal is to come closer to a unified approach, to increase the awareness of experimental design for hydrological data gathering within the hydrological sciences, and to illustrate the benefit of applying experimental design in hydrological studies.

Murphy (1993) defines a "good forecast" in meteorology as consisting of **consistency**, **quality**, and **value**. Weijs et al. (2010) also used these ideas for the evaluation of hydrological forecasts, and I believe that these three criteria can be applied to experimental design in the subsurface as well. The definitions proposed by Murphy (1993) are as follows:

1. **Consistency**: *"Correspondence between forecasts and judgments"*
2. **Quality**: *"Correspondence between forecasts and observations"*
3. **Value**: *"Incremental benefits of forecasts to users"*

Here, I will adapt these ideas to experimental design in hydrology.

1. Consistency. In Murphy's lore, consistency is rooted in the information available to the forecaster and their best *judgment* based on that information. It is therefore a somewhat subjective concept. I think that some level of subjectivity is unavoidable in experimental design, but we can still strive to make the design as objective as possible. I will therefore divide consistency into two parts:

- **Internal consistency**: Correspondence between the optimal design and expectations from scientific understanding.
- **External consistency**: Consistence of experimental results across different studies in similar circumstances.

Internal consistency is more objective than external consistency because it requires scientific agreement with our current understanding of the problem. Internal consistency can be achieved by including scientific understanding and modeling in the experiment design. It requires communication and discussions with colleagues who have expertise in related fields (as also emphasized by Box et al. (1978)). At the most fundamental level, we want to ensure that the optimal design discovered did not result from simulations that disregard physical laws, which is ensured by including a physical model in the design process, as was done in Chapters 3 to 5. Furthermore, we want to make sure that the optimal design is consistent with our understanding of how information is transferred in the system under consideration. In Chapter 3, for example, the most informative wells to place were found to be those located downstream of the sensor. Logically, the injected tracers stayed in the modeled aquifer longer and took up more space in the synthetic cells, making the breakthrough curves more informative. If the results had been different, it would have indicated that the model was not correctly capturing the physical processes. Similarly, in Chapter 4, the difference in information content between the

geophysical data and the boreholes was found to lessen as the boreholes were spread out rather than clustered in the initial design. As a result, they were able to capture more information, and their overall information content increased. In addition, in Chapter 5, for the 1D case, the level of uncertainty decreased as new measurements were selected and taken within the zone of highest target variance (i.e., the shallow zone with high temperature gradient). The highest ranked sensor locations in the 3D case were placed perpendicular to the river, which is consistent with the fact that exchange fluxes through the riverbanks are expected to be greater than those in the river's middle area.

I can imagine that in more complex cases, the intricacies of phenomena make assessing consistency regarding information flow difficult, but I believe that the experimenter should strive for "internal consistency" by incorporating scientific understanding into the design process and information flow in the system being modeled (e.g., Ruddell et al. 2013).

External consistency is attained when the outcomes of different experimental designs with different initial conditions agree. Replicability studies are essential in experimental design to ensure the validity of the optimal design. In fact, *replicability* is one of the three pillars of experimental design (Montgomery, 2019). At the most fundamental level, as we have done in Chapters 3 to 5, consistency of the optimal design was ensured through a cross-validation study. To ensure consistency on a larger, community-wide scale, we can consider conducting a meta-analysis of the findings of various studies. The creation of a common knowledge base for experimental design in hydrology would help to promote external consistency (e.g., Clarke 2008). Beven et al. (2020) recognizes the need for improved observational technologies and network designs to support hypothesis testing in real catchments of interest that go beyond current monitoring capabilities, necessitating information from other studies as well as direct observations, remote sensing, intensive field campaigns, or other strategies. In a recent discussion about the future of hydrogeology, Hermans et al. (2022) also advocate for the sharing of data in an open format.

2. Quality. In Murphy (1993), high-quality forecasts closely match observations. In experimental design, I believe a same concept applies. I contend that comparing the posterior distributions of the design's predictions to the observations can assess the quality of the optimal design, or at least one quality thereof. In other words, the experimental design process should not be done blindly, and the information metric of each design should be displayed and evaluated by the experimenter at each iteration.

The goal of experimental design is to maximize some objective function which depends on the information content of the observations, and the quality of the design is thus dependent on the observational information content. For the sake of simplicity, we will assume that the objective function is solely dependent on the information content of the observations in the following. In static designs (Chapters 3 and 4), the total information content varies between designs, with the design containing the most information being the optimal design. In sequential designs, like the 1D case of Chapter 5,

the information content is evaluated sequentially from one sensor to the next, and the optimal design is the one with the highest information content. At each step, the information content of an observation is estimated using a specific metric in both cases, and the metrics of the designs are compared. If a found optimal design produces large posterior distributions that are not centered on the observations after this process, the design is most likely not optimal. It is the experimenter's responsibility to investigate the posterior distributions of the predicted parameters used to estimate the information content. Without this check, non-informative designs might be mistakenly deemed optimal; experimental design cannot guarantee assumptions such as experiment effectiveness, model form validity, and criterion reflecting experiment objectives (Smucker et al., 2018).

The selection of an appropriate information metric, and thus the quality estimation of the experimental design, is possibly the most difficult task in experimental design. In Chapter 3, an image similarity metric was used to assess the information content in the target's original space. In Chapter 4, we found a more model-agnostic metric that operates in the parameter's low-dimensional space, which I believe is a promising direction for future research. Finally, in Chapter 5, we devised a metric grounded in information theory that is model-agnostic and can be used in any experimental design.

3. Value. The concept of value of information has already been discussed in §1.3. The value of an optimal design relates to the benefits of the design to the people who will deploy it. Such benefits can be measured in terms of cost, time, and other resources, more broadly referred to as *utility* (Weijs et al., 2010). In practice, determining the value of an optimal design is frequently difficult or subjective, and the user and decision-maker must rely on their own judgment. The reason for this is that once the optimal design is found and executed, there is usually no way to know what would have happened had another design been chosen, since the experiment will have already been executed. As a result, the value criterion is less objective than the other two criteria, but "good value" can be ensured by validating the other two criteria to the best of our ability.

General limitations of experimental design. We strive to create reliable simulation models and computer systems by combining a variety of unreliable components, such as noisy and defective sensors, human and algorithmic error, unknown or incomplete material properties, boundary conditions, and so on (Lavin et al., 2021). In this regard, experimental design should not be seen as a silver bullet or a panacea for all our problems, but as a tool to make the best of the available data.

In addition to the constraints mentioned by Smucker et al. (2018), Beven et al. (2020) claim that because simulation-generated data is dependent on the structural assumptions of the model that generated it, and there is a mismatch between the complexity of the relevant processes' perceptual model and the relative simplicity of current model structures, simulations will only go so far in determining what type of measurements should be prioritized.

The experimental design framework in this dissertation is also limited in other ways: for example, testing for all sensors positions and combinations is possible in our synthetic studies, but time-consuming, because each combination requires a training and testing phase.

Let's take Chapter 4 as an example; as a reminder, the number of possible combinations is $\frac{n!}{w!(n-w)!}$, where n is the number of possible positions and w is the number of wells. For a number of 3 wells and 256 possible positions (number of cells for one plane in the grid), the number of possible combinations is $2,763,520$. This is assuming uniform grid cells and no hydrogeological property scaling issues and all screen intervals at the same depth level. Indeed, petrophysical laws may not be viable at different scales (Beven et al., 2020; Blöschl et al., 2019; Singha et al., 2015). Kratzert et al. (2019a) also point out that the problem on how to use one model, or one set of models, to provide spatially continuous hydrological simulations across large areas has been a long-standing challenge in the hydrological sciences (e.g., regional, continental, global).

However, regardless of which combination is selected, the computations will be based on the same simulations. This ensures that the most time-consuming component using BEL will not be performed more than once. As a result, running the experimental design for a representative subset of these combinations is feasible. Constraints such as at least some minimum distance between sensors and other location constraints can be used to reduce the number of possible combinations (e.g., Wang et al. 2022c).

One ideal situation in which to apply our framework is when potential sensor locations are predetermined and limited by factors such as land occupation. The situation where the sensors are already positioned, and we are looking for the single next-best sensor position is an alternative ideal situation. It entails sequentially updating our prior knowledge from the existing sensor(s), as each new measurement would necessitate rerunning the forward model with updated parameters. This topic was beyond the scope of Chapter 3 and Chapter 4, since the forward model was so expensive to run. Sequential BOED, solving the problem of allocating new sensors placement adaptively such as Haan et al. (2021), Lykkegaard and Dodwell (2022), Wang et al. (2022c), was developed in Chapter 5.

General outlooks for experimental design in the subsurface. As Beven et al. (2020) claim, we need data streams for water fluxes, water storage, water quality, and catchment properties to improve hydrological science by providing better inputs for hydrological predictions and supporting better hypothesis testing. Within the context of Bayesian statistics, optimizing observational improvements can be investigated through the use of a form of pre-posterior prior analysis through simulations to test the value new observations or new types of observations (Beven et al., 2020), which is what we have done in this dissertation.

The technological aspect of observational improvements is not the focus of this dissertation, but the experimental design aspect is. I believe that the work presented in this

dissertation is a step in the right direction in this regard, as demonstrated in Chapter 3, where we demonstrated the suitability of the BEL framework with a very small dataset, Chapter 4, which presented a novel way to combine data gathered from different sources to better constrain the uncertainty in predictions, and Chapter 5, which presented a novel approach for sequential experimental design.

6.5 Perspectives on Bayesian Evidential Learning

> *Whatever way uncertainty is approached, probability is the only sound way to think about it.*
>
> Dennis Lindley

Wrap-up. Bayesian evidential learning is a machine learning approach that combines Bayesian inference and evidential learning to make predictions based on data. In Bayesian inference, probabilities are used to represent the uncertainty associated with different possible outcomes or events. These probabilities are updated as new information becomes available, using Bayes' theorem.

Evidential learning, on the other hand, is a way of combining evidence from multiple sources to make a decision or prediction. It is often used in expert systems and other decision-making contexts where there is uncertainty or incomplete information.

Bayesian evidential learning combines these two approaches by using Bayesian inference to update probabilities based on new evidence, and using evidential learning to combine this evidence with other information to make a prediction. This approach allows a machine learning model to make more accurate predictions by taking into account the uncertainty and complexity of the data.

Given the nomenclature of BEL, however, I believe that its name does not reflect the true essence of BEL, which is to combine Bayesian inference in order to learn a statistical relationship between predictor $\mathbf{d}$ and target $\mathbf{h}$ from a training set of synthetic data generated from forward simulations.

Bayesian Evidential Learning implies a combination of Bayesian Inference and Evidential Learning, but it is simply a descriptive name for applying Bayesian inference to a machine learning problem. I believe that the term "Bayesian Simulation-Based Inference" (BSBI) is more appropriate, as it more accurately conveys the framework's purpose, or "BAyesian Simulation-Informed Learning" (BASIL) could also be used. Alternatively, we could unify BEL with the already established Simulation-Based Inference (SBI; Lavin et al. 2021). Aside from that minor quibble, BEL has been demonstrated to provide accurate predictions and reliable uncertainties in a wide range of Earth Science applications, and it is rapidly gaining popularity.

Outlooks. In addition to BOED, the BEL framework can be used to solve a wide range of data sources fusion and allocation problems with any utility function. Haan et al. (2021) mention that solving such allocation problems helps answer questions such as:

- *Which additional sensor type would yield the highest gain?*
- *What is the most effective data source combination?*
- *How many and where should sensors be activated given logistical or energy constraints?*

Chapter 4 demonstrates that the BEL framework is appropriate for subsurface monitoring as well as the fusion of hydrological and geophysical data, a problem known as hydrogeophysical coupled inversion (Hermans et al., 2021b). Additionally, the BEL framework can help to solve the challenging problem of 4D inversion of DC resistivity monitoring data, a problem that has been studied by Kim et al. (2009). By abstracting the time-dimensions of the problem, the BEL framework can be extended to true 4D problems, as done in Chapter 4.

Several other dynamic processes can be monitored using time-lapse electrical techniques (Hermans et al., 2018; Singha et al., 2015), and the methodology could be applied to cases such as the monitoring of a contaminant plume (Nazifi et al., 2022; Power et al., 2014; Robinson et al., 2020; Tso et al., 2020), saltwater intrusion (SWI) (Johnson et al., 2015; Palacios et al., 2020), or more generally, hydrological events happening in the vadose zone (Furman et al., 2004). For example, to assess the efficiency of a remediation strategy for a contaminated aquifer, the position of a well can be optimized to minimize the cost of drilling, while still fulfilling the desired accuracy in the prediction of the contaminant concentration. In SWI-prone areas, it could also be used to minimize the uncertainty in the most vulnerable areas in the prediction of the saltwater front by optimizing the number and positions of measurements.

Other geophysical monitoring applications could also benefit from the methodology. Wilkinson et al. (2015) review some examples of geophysical monitoring applications. Newer developments include applications such as CO_2 sequestration (Auken et al., 2014; Lu et al., 2015; Pezard et al., 2016; Schmidt-Hattenberger et al., 2016), soil properties for agriculture applications (Blanchy et al., 2020), or even erosion (Masi et al., 2020) and ground-loosening processes (Kim et al., 2022).

The approach can be used to combine datasets of any type and any dimension in order to stochastically solve geophysical inverse problems. Data fusion is the process of combining information from different sources (e.g., sensors, databases, human expertise), and is a key method for handling imperfect raw data to obtain useful and accurate information (Meng et al., 2020). The process of combining observations into computer models is also known as data assimilation, and it is widely used in numerous fields, including atmospheric prediction, seismology, energy and environmental applications (Lavin et al., 2021; Nearing et al., 2018a, 2022; Tso et al., 2020).

Although different geophysical methods are complementary, JafarGandomi and Binley (2013) argue that inconsistency in their acquisition geometries (i.e., 1D, 2D, or 3D) makes the data fusion difficult. The scale of sensors and systems as well as the complexity of application environments present numerous challenges (Lavin et al., 2021). In JafarGandomi and Binley (2013)′s inversion strategy, the subsurface model was recasted to a set of locally layered 1D earth-models to overcome the incompatibility of the acquisition geometries and also to reduce the computational cost of the 2D and 3D inversions. Our fusion method easily combines datasets of any dimension because it first reduces their dimensionality before concatenating them and feeding them to the learning algorithm.

Our proposed framework is the first to use BEL for data fusion in a low-dimensional latent space and Bayesian optimal experimental design. Meng et al. (2020) recently reviewed the literature on data fusion and identified a number of unresolved issues as well as future research directions that warrant further investigation, but they did not discuss the low-dimensional latent space approach.

However, using simulation and machine learning together (the heart of BEL) does not always result in solutions. Instead, it enables intelligent navigation of intricate, high-dimensional spaces of potential solutions when integrated in the ways we've described: bounding the space helps to direct the user to the optimal solution (Lavin et al., 2021).

Limitations. The consistency of new data (e.g., field observation or measurement) with the forward-model generated prior should always be verified (Hermans et al., 2019, 2018). Otherwise, BEL runs the risk of producing an unrealistic posterior distribution. In experimental design, one often deals with simulation-generated synthetic data rather than field data, and this condition is thus almost always verified. Nevertheless, Chapter 5 demonstrated that the framework is robust to slightly out-of-prior data. In general, the prediction uncertainty may be overestimated due in part to the dimension reduction steps of PCA and CCA, which simplify the problem (see Figure 3.3B how the WHPA delineation is smoothed), or due to a large prior (Michel et al., 2022a). Overestimating the prediction's uncertainty, on the other hand, prevents overfitting, or gaining too much confidence in a false prediction, which can increase the risks (cf. Chapter 1). Iterative prior resampling (IPR) followed by rejection sampling is a method proposed by Michel et al. (2022a) to overcome the limitations of a large prior. IPR combines the posterior distribution calculated in a previous iteration with the prior distribution in a subsequent iteration to improve the estimation of the final posterior distribution, whereas rejection sampling eliminates models that do not adequately fit the data. Other techniques for preventing overfitting within the BEL framework include the use of probabilistic Bayesian neural networks and an early stopping mechanism, as introduced in Chapter 5.

Another limitation regarding the prior is the somewhat subjective choice of its size, which should be sufficient to characterise the variability in the predictor and target variables. This size is thus case-dependent and depends mainly on the complexity of the target (Hermans et al., 2018; Thibaut et al., 2021b). Michel et al. (2020b) have shown

that there is a threshold above which increasing the size has no effect on the posterior prediction. The latter is found by progressively increasing the number of samples in the prior until the posterior stabilises. The use of a sufficiently large prior reduces the risk of incorrectly estimating the predictor-target relationship. In Chapter 3, we found that a training set of 400 is adequate for both prediction and experimental design purposes; cross-validation shows that the size of the test set required for experimental design is at least 250, so the training set size was increased to 1000 samples to allow for k-fold cross-validation. The number of PCs to keep for both predictor and target, the number of posterior samples to compute the uncertainty reduction, and the definition of the data-utility function, depending on the nature of the target variable, are all additional experimental design variables that must be set by the practitioner.

A possible limitation is the heterogeneity of the medium used for forward modelling (e.g., hydraulic conductivity field). In Chapters 3, 4 and 5, the different fields are generated by sequential Gaussian simulations (2-points statistics) and are inherently smooth. We also investigate in Chapters 3 and 5 uncertainty in variogram parameters and demonstrate the framework's robustness when dealing with structural uncertainty. The natural variability of geological media can be far from smooth, for example, channelised media (e.g., Lopez-Alvis et al. 2021), necessitating more advanced simulation techniques such as multiple-point statistics (Mariethoz and Caers, 2014). Since the focus of this study was to demonstrate the ability of BEL in an experimental design framework, we did not investigate these more heterogeneous priors. Since BEL has been successfully used for such complex priors (Hermans et al., 2016; Satija and Caers, 2015; Yin et al., 2020), we do not anticipate any issues when applying the proposed framework to such cases. Although such greater prior uncertainty would inevitably increase posterior uncertainty, our experimental design approach could still be used without loss of generality.

7. Conclusion

Bird's eye view. We presented a novel framework for Bayesian optimal experimental design in this dissertation that integrates simulation and data-driven methods, Bayesian inference, and machine learning. Our method is based on Bayesian Evidential Learning (BEL), a Monte Carlo method that employs machine learning to learn a direct relationship between predictor and target variables. The BEL framework is extremely adaptable and can accommodate a wide range of data sources, including field observations or measurements and simulated data, though only the latter is used in this dissertation. Three case studies in groundwater modeling have demonstrated the efficacy of our methods: wellhead protection area delineation, an aquifer thermal energy storage monitoring system, and groundwater-surface water interaction.

Innovations for Bayesian Evidential Learning. In the context of BEL, we introduced three innovations rolled into one. First, we compared three different pre-existing methods for inferring the posterior distribution of the predictor-target relationship derived from training data in the low-dimensional latent space using Canonical Correlation Analysis (CCA). The algorithms use multivariate Gaussian inference, Kernel Density Estimation (KDE), and Transport Maps (TM). The Gaussian method is the simplest and most computationally efficient method, whereas the KDE and TM methods offer greater accuracy at the expense of increased computational cost. The Gaussian and Kernel methods have been applied to BEL in the past, whereas the TM method is novel.

Second, in order to address the shortcomings of CCA, we developed a new method for inferring the posterior distribution of the target variable using a Probabilistic Bayesian neural network (PBNN). CCA maximizes the linear relationship between a variate from one set of variables and a variate from the other set, and thus misses potential nonlinear components of relationships between canonical variate pairs. Furthermore, CCA works best when variable relationships are homoscedastic, that is, when the variance of one variable is roughly the same at all levels of the other variable. The PBNNs alleviate the linear assumptions of CCA and demonstrates greater flexibility, albeit at the expense of increased computational cost and hyperparameter tuning. The PBNN technique predates BEL, but it has never been applied to BEL before.

Third, we proposed a data fusion method for combining observations from various data sources. The data fusion method combines observations from two data sources (e.g., wells monitoring data and geophysics measurements) in a low-dimensional latent

space of the predictor and target variables (e.g., via PCA). This can be enabled using any of the above-mentioned inference methods. Working in low-dimensional latent space allows for more efficient inference of the target variable's posterior distribution, at the inevitable cost of information loss, which can be controlled by the number of principal components used in the data fusion approach and careful examination of the resulting posterior distribution. The method can be used to combine data from any source, including direct measurements and simulated data. The combined observations are then used in the BEL framework to infer the target variable's posterior distribution, providing a more accurate estimate of the target's uncertainty.

Finally, we have integrated the aforementioned innovations into a single framework, which we call `SKBEL`, a Python package that is freely available on GitHub[1]. The package is built on top of the `scikit-learn` Python package and provides a unified and modular interface for the aforementioned methods.

Innovations for Bayesian optimal experimental design. We have introduced BEL4ED, a new framework for Bayesian optimal experimental design which integrates BEL and Bayesian optimal experimental design (BOED). The method is computationally efficient and requires minimal training data; unlike other BOED methods, it does not average or approximate the posterior uncertainty of the target variable. The method employs the BEL model to explicitly determine the optimal design of data sources that minimizes the posterior uncertainty of the target variable. It expands BOED's capabilities, such as comparing different data sources, optimizing data source locations, and sequentially optimizing data sources with model parameter updates. Three case studies have demonstrated the effectiveness of our method.

Innovations specific to the case studies. The BEL4ED framework was developed in the first application (wellhead protection area delineation) to optimize the location of injection wells in a wellhead protection area (WHPA) delineation problem. Using the pre-existing Modified Hausdorff Distance (MHD) and Structural Similarity (SSIM) index metrics, we estimated the WHPA's posterior uncertainty range. The use of the Traveling Salesman Problem (TSP) to delineate the WHPA from a set of backtracked particles is one novel element. The TSP is a well-known combinatorial optimisation problem that seeks the shortest path that visits all nodes in a graph. It is, however, the first time it is used in the context of WHPA delineation. Other elements brought to the community through this contribution are the implementation of the MHD metric into the well-known `scikit-image` Python package [2], the implementation of a `back-transform` function for the CCA method in the `scikit-learn` package to enable the use of CCA in the context of BEL4ED [3], and the creation of the Python packages `SKBEL` and `PySGEMS` [4], a package for interfacing with the `SGEMS` software.

[1] `https://github.com/robinthibaut/skbel`
[2] `https://github.com/scikit-image/scikit-image/pull/5581#event-5850950501`
[3] `https://github.com/scikit-learn/scikit-learn/pull/19680#issuecomment-943318723`
[4] `https://github.com/robinthibaut/pysgems`

In the second application (aquifer thermal energy storage monitoring system), we used a Bayesian experimental design framework to compare well and geophysical data for temperature monitoring. We used Bayesian Evidential Learning (BEL) to optimize the design of a 4D temperature field monitoring experiment by considering different data source combinations (temperature logs combined with geophysical data), and we presented a new method for combining observations from different data sources and quantifying the resulting uncertainty. In particular, we introduced a novel method for optimizing the Electrical Resistivity Tomography protocol. In addition, Transport Maps (TM) was introduced to BEL in this application.

We used BEL to estimate groundwater-surface water interaction fluxes from temperature data in the third application. In this application, we introduced the Probabilistic Bayesian neural network (PBNN) method to BEL and transitioned from a static experimental design framework to a sequential experimental design framework. To derive optimal temperature sensor locations using the sensors' information gain, we devised a custom objective function based on information theory. Using a 1D synthetic dataset and a 3D conceptual riverbed-aquifer model based on a real-world case study, we compared the performance of PBNNs for flux estimation.

Future work. Overall, this dissertation demonstrates the utility of BEL in groundwater modeling for optimal experimental design. We have demonstrated the potential of BEL4ED for various applications, and our method is not limited to subsurface modelling. We are developing an open-source Python package for BEL4ED, which we believe will be useful for practitioners in a variety of fields. The BEL framework is an effective tool for reducing the uncertainty associated with subsurface predictions, and its use opens up new avenues for data and simulation-driven subsurface modeling.

In the future, we anticipate that BEL will continue to be applied to a variety of Earth Science applications, such as hydrological modelling, groundwater contamination prediction, climate change prediction, geophysical inversion, geothermal energy, and more. We expect that the framework will continue to evolve as better machine learning algorithms and dimensionality reduction methods are developed. We believe that BEL represents a promising new tool for predictive modelling and that it has the potential to revolutionise the way we approach predictions in Earth Sciences. Only time will tell if this is indeed true.

A. Python snippets

```python
import matplotlib.pyplot as plt  # for plotting
import numpy as np  # for numerical routines
from sklearn.cross_decomposition import CCA

# Create a sample dataset of correlated multivariate normal distributions

np.random.seed(0)  # set seed for reproducibility
n = 1000  # number of samples
X = np.random.normal(size=(n, 4))  # 4-dimensional

# (1) Linear case:
Y = X + np.random.normal(size=(n, 4)) / 4  # Add noise

# (2) Nonlinear case:
# Y = np.sin(3*X + np.random.normal(size=(n, 4))/4)

# Transform the data to make it more interesting
X[:, 0] = X[:, 0] + Y[:, 0]
X[:, 1] = X[:, 1] - Y[:, 1]
Y[:, 2] = Y[:, 2] + X[:, 2]
Y[:, 3] = Y[:, 3] - X[:, 3]

# Plot the data
plt.plot(X[:, 0], Y[:, 0], "o", alpha=0.5, label="Dimension 1")
plt.plot(X[:, 1], Y[:, 1], "o", alpha=0.5, label="Dimension 2")
plt.plot(X[:, 2], Y[:, 2], "o", alpha=0.5, label="Dimension 3")
plt.plot(X[:, 3], Y[:, 3], "o", alpha=0.5, label="Dimension 4")
plt.xlabel("X")
plt.ylabel("Y")
plt.title("Original 4-dimensional data")
plt.legend(loc="upper right")
plt.xlim(-4, 7)
plt.show()

# Fit the CCA model
cca = CCA(n_components=4)  # specify the number of components
cca.fit(X, Y)  # fit the model

X_c, Y_c = cca.transform(X, Y)  # transform the data

# Compute the correlation between pairs
corr = np.corrcoef(X_c.T, Y_c.T)

```

```
# Plot the transformed data
colors = ["#F00087", "#005EF5"]
plt.figure(figsize=(6, 4))
plt.plot(X_c[:, 0], Y_c[:, 0], "o", alpha=0.5, label=f"Pair 1 - Corr: {
    corr[0, 4]:.2f}", color=colors[0],
)
plt.plot(X_c[:, 1], Y_c[:, 1], "o", alpha=0.5, label=f"Pair 2 - Corr: {
    corr[1, 5]:.2f}", color=colors[1],
)
plt.xlabel("X")
plt.ylabel("Y")
plt.title("Canonical Variate Pairs 1 & 2")
plt.legend(loc="lower right")
plt.show()
```

Listing A.1: CCA code snippet

```
import numpy as np  # for numerical routines

import tensorflow as tf  # for neural networks
import tensorflow_probability as tfp  # for Bayesian neural networks


tfd = tfp.distributions

# generate synthetic data
n = 250  # number of samples
rng = np.random.RandomState(0)  # set seed for reproducibility
X = rng.randn(n, 1) * 2  # X is a 1-dimensional normal distribution
x_range = np.min(X), np.max(X)  # range of X
noise = rng.normal(loc=0.0, scale=0.7, size=n)  # noise is a normal
    distribution
Y = X[:, 0] + np.sin(0.5 * np.pi * X[:, 0]) + noise  # Y is a function of
     X plus noise

X_test = np.linspace(x_range[0], x_range[1], n).reshape(-1, 1)  # test
    data

# 1. Deterministic neural network
# -------------------------------

# define the model
model = tf.keras.Sequential(
    [
        tf.keras.layers.Dense(10, activation="relu", input_shape=(1,)),
        tf.keras.layers.Dense(10, activation="relu"),
        tf.keras.layers.Dense(1),
    ]
)

# compile the model
model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=0.01),
    loss="mse")

# fit the model
```

```
history_ann = model.fit(X, Y, epochs=500, verbose=0)

# predict the test data
predictions = model.predict(X_test)

# 2. Bayesian neural network
# --------------------------
from skbel.nn_utilities import prior_trainable, posterior_mean_field,
    neg_log_likelihood

# define the model
model = tf.keras.Sequential(
    [
        tfp.layers.DenseVariational(
            units=10,
            activation="relu",
            make_prior_fn=prior_trainable,
            make_posterior_fn=posterior_mean_field,
            kl_weight=1 / n,
            input_shape=(1,),
        ),
        tfp.layers.DenseVariational(
            units=10,
            activation="relu",
            make_prior_fn=prior_trainable,
            make_posterior_fn=posterior_mean_field,
            kl_weight=1 / n,
        ),
        tf.keras.layers.Dense(1),
    ]
)

# compile the model
model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=0.01),
    loss="mse",
)

# fit the model
history_bnn = model.fit(X, Y, epochs=500, verbose=0)

# predictions
# in a BNN, we need to make several predictions to get a sense of the
    uncertainty
n_samples = 20  # number of samples to draw
predictions = np.array([model.predict(X_test) for _ in range(n_samples)])
      # make predictions

# 3. Probabilistic neural network
# -------------------------------

# define the model
num_components = 1  # number of mixture components
output_dim = 1  # output dimension
params_size = tfp.layers.MixtureNormal.params_size(
```

```
    num_components, output_dim
)  # number of parameters

model = tf.keras.Sequential(
    [
        tf.keras.layers.Dense(10, activation="relu", input_shape=(1,)),
        tf.keras.layers.Dense(params_size, activation=None),
        tfp.layers.MixtureNormal(
            num_components, event_shape=[output_dim]
        ),  # mixture of Gaussians
    ]
)

# compile the model
model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=0.01),
    loss=neg_log_likelihood,  # negative log-likelihood
)

# fit the model
history_mdn = model.fit(X, Y, epochs=500, verbose=0)
# predictions
# in a PNN, we need to sample from the posterior to get a sense of the uncertainty
n_samples = 20  # number of samples from the posterior
posterior = model(X_test)  # posterior distribution
predictions = posterior.sample(n_samples)  # sample from the posterior

# 4. Probabilistic Bayesian neural network
# ----------------------------------------

# define the model
num_components = 1  # number of mixture components
output_dim = 1  # output dimension
params_size = tfp.layers.MixtureNormal.params_size(num_components, output_dim)  # number of parameters

model = tf.keras.Sequential(
    [
        tfp.layers.DenseVariational(
            units=10,
            activation="relu",
            make_prior_fn=prior_trainable,
            make_posterior_fn=posterior_mean_field,
            kl_weight=1 / n,
            input_shape=(1,),
        ),
        tfp.layers.DenseVariational(
            units=params_size,
            activation=None,
            make_prior_fn=prior_trainable,
            make_posterior_fn=posterior_mean_field,
            kl_weight=1 / n,
        ),
        tfp.layers.MixtureNormal(num_components, event_shape=[output_dim
```

```
    ]),  # mixture of Gaussians
    ]
)

# compile the model
model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=0.01),
    loss=neg_log_likelihood,  # negative log-likelihood
)

# fit the model
history_pnn = model.fit(X, Y, epochs=500, verbose=0)

# predictions
# in a PBNN, we need to sample from the posterior to get a sense of the
    uncertainty
n_samples = 20  # number of samples from the posterior
posterior = model(X_test)  # posterior distribution
predictions = posterior.sample(n_samples)  # sample from the posterior
```

Listing A.2: A simple implementation of a probabilistic neural network.

B. A new workflow to incorporate prior information in minimum gradient support (MGS) inversion of electrical resistivity and induced polarization data

This chapter was published in Journal of Applied Geophysics (Thibaut et al., 2021a):

Thibaut, Robin, Thomas Kremer, Annie Royen, Bun Kim Ngun, Frédéric Nguyen, and Thomas Hermans (Apr. 2021). "A new workflow to incorporate prior information in minimum gradient support (MGS) inversion of electrical resistivity and induced polarization data". In: Journal of Applied Geophysics 187, p. 104286. issn: 09269851. doi: 10.1016/j.jappgeo.2021.104286.

CRediT author statement. **Robin Thibaut**: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Thomas Kremer**: Investigation, Writing - Review & Editing. **Annie Royen**: Investigation, Writing - Review & Editing. **Bun Kim Ngun**: Resources, Supervision, Project Administration, Funding Acquisition. **Frédéric Nguyen**: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition. **Thomas Hermans**: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Abstract

The current paradigm for geophysical inversions is to select the simplest solution according to Occam's principle. The implicit assumption usually made is that the parameters of interest have a smooth spatial distribution, which is rarely geologically plausible. An alternative is the Minimum Gradient Support (MGS), a functional that allows to compute a regularized inversion favoring sharp contrasts. However, solutions are highly sensitive to the selection of a variable called the focusing parameter β and the method is not very performant when many structures are present in the subsurface. Thus, we propose a new workflow to apply this functional to real case studies where heterogeneous

structures are expected: a smooth solution is first computed and used as a starting model for sharp MGS inversions. Sequentially incorporating additional prior information on resistivity (e.g., from drilling, previous geophysical surveys) is possible and further improves imaging for resistivity and chargeability structures. The new methodology is first tested on a synthetic case and then applied to ERT/IP data collected on a gold deposit. The methodology enables to compute plausible electrical resistivity spatial distributions in accordance with the vast prior geological knowledge, and reveals new insights about the mineralization key characteristics. The choice of β remains challenging and should be automated in future developments.

B.1 Introduction

Geophysical inverse problems are often ill-posed and generally require regularization to be solved (Tikhonov and Arsenin, 1977). This is done by implementing a stabilizing or regularization functional, in which prior geological information can be implemented, in addition to a data misfit term in a global cost function to minimize If the problem is non-linear, minimization takes the form of an iterative process, where the data misfit is balanced with the model functional through a regularization parameter.

In practice, algorithms often rely on Occam's principle, where in essence the simplest solution is sought for through a regularization functional taking the form of a gradient or roughness operator, leading to the so-called smoothness constraint inversion. The latter will result in smooth solutions (Loke et al., 2013), which are rarely geologically plausible (Linde et al., 2015; Zhdanov and Tolstaya, 2004).

In many cases, the end goal of geophysical investigations is not to provide an image of a geophysical property, but to make informed decisions based on results, such as where to drill a water abstraction well (e.g., Robert et al. 2011), determine if a building area is safe to support certain structures (e.g., Van Hoorde et al. 2017), or where to drill to cross a deposit's mineralizations (e.g., Evrard et al. 2018).

Obtaining the most valid subsurface image through geophysical inversion is therefore crucial to facilitate human interpretation and pertinent decision-making, and the smoothness constraint solution may not always be suitable to properly meet these expectations. Several ways exist to incorporate a priori information into the inversion process (e.g., Caterina et al. 2014). One of the simplest approaches to incorporate prior information is to use a reference model in the regularization functional (Oldenburg and Li, 1994). However, an over-confidence in the reference model can lead to an erroneous solution (Caterina et al., 2014) and subsequent misinterpretation. In practice, it is difficult to choose the weight to be be given to the reference model, especially since low sensitivity zones are generally more influenced by it, as shown for example through the DOI (Depth Of Investigation index, Oldenburg and Li (1999)). To counter this issue, Kim et al. (2014) developed an inversion algorithm that re-weights the importance given to a-priori information at each iteration, to guide the convergence process in earlier iterations with higher weights, but let the solution develop freely in subsequent iterations with decreasing weights assigned to the reference model, which reduces the influence of incorrect a-priori information. As an alternative, Dumont et al. (2016) used the inverse of the cumulative sensitivity matrix to constrain their inversion using a reference model based on borehole data.

To recover sharp contrasts, the L1 norm (sum of the absolute values of the spatial changes in the parameter distribution) can be implemented in the regularization functional. This requires an iteratively re-weighted least-squares method to solve the iterative process and returns blocky structures since less penalty is given to abrupt changes compared to a L2 norm (e.g., Loke et al. 2003). To recover sharp horizontal or vertical contrasts, one can also implement structural constraints by modifying the horizontal and vertical weights in smoothness constraint inversions for areas where such contrasts

are expected (Kaipio et al., 1999). Saunders et al. (2005) successfully applied such an approach to constrain ERT inversions using seismic refraction data, Doetsch et al. (2012) constrained the process by using structural information derived from ground-penetrating radar data, and Hermans et al. (2012) used borehole information to constrain the depth of an impermeable clay layer. However, this approach generally requires obtaining spatially continuous information from an independent source, which is not always possible. An alternative to the smoothness constraint is to incorporate spatial correlation into the model functional using for example a covariance matrix or a variogram (Bouchedda et al., 2017; Chasseriau and Chouteau, 2003; Hermans and Irving, 2017; Hermans et al., 2012; Johnson et al., 2007; Jordi et al., 2018). The vertical and/or horizontal correlation lengths can for example be estimated from borehole data. This approach has been proven superior to the smoothness constraint in many contexts but is less adapted to sharp contrasts.

Another alternative, investigated in this chapter, is the minimum gradient support (MGS), a functional favoring sharp contrasts between homogeneous blocks (Portniaguine and Zhdanov, 1999). Although it has found its place in several applications (e.g., Ajo-Franklin et al. 2007; Fiandaca et al. 2015; Last and Kubik 1983; Ley-Cooper et al. 2015; Vignoli et al. 2015; Zhdanov and Tolstaya 2004), and has been implemented in some freely available software programs (e.g., Auken et al. 2015), it has a significantly lower number of field applications compared to the more popular smoothness constraint inversion.

The MGS functional is not implemented in popular commercial softwares such as RES2DINV, Zondres2d or ERTLab, even though such a functional would be appropriate in many contexts (faults, layered structures, mineralized bodies, etc.). One probable reason for the limited transfer of the MGS to practitioners and commercial software packages is the sensitivity of the solution to the focusing parameter β, responsible for the degree of sharpness of the final image, and the difficulty in finding its optimal value (Gündoğdu et al., 2020; Nguyen et al., 2016; Zhao et al., 2016). Another reason is the relative inability of MGS to image a subsurface with multiple heterogeneous structures.

Blaschek et al. (2008) implemented a new algorithm in the complex resistivity tomography code CRTOMO (Kemna, 2000) including cell sensitivity in the MGS functional as a weighting factor. They first applied the modified functional to a synthetic case with a single isolated body, then to a three-layered model, and successfully applied their methodology to field data collected on a hydrogeological experimental site whose geology, known from many drill logs, consists of several unconsolidated sediment layers. The structures are the same for resistivity and phase. In order to choose the value of the focusing parameter β, their suggestion is to scan the model space by varying the β value until a satisfactory solution in accordance with the known geological structure is obtained. This approach is also adopted in this work.

In another contribution, Vignoli et al. (2015) used the functional in order to invert time-domain electromagnetic profiles. They applied their methodology to synthetic and

field data from quasi-layered models. In their formulation, they set the focusing parameter to 1 and instead used the data noise vector to control the sharpness of their images. They obtained plausible solutions without having to constrain the inversion by additional prior information.

In this contribution, we focus on the improvement of the MGS solution in heterogeneous geological settings by proposing a new workflow where prior information can be integrated in a robust manner. The workflow is based on incorporating prior information into the start model and the reference model, and combining the robustness of smooth inversions with the focusing nature of MGS. The reference model approach has been demonstrated to be an efficient way to improve imaging when prior information is available (Caterina et al., 2014; Kim et al., 2014), but has never been extensively investigated when combined with the MGS regularization. Compared to previous studies, we also work with large datasets and at a scale of 500 meters. Furthermore, we consider complex geological settings with both vertical and lateral structures, where electrical resistivity and phase/chargeability do not necessarily share the same structure. The choice of β is made according to the approach proposed by Blaschek et al. (2008). An objective and rigorous method of selecting this parameter is still needed, but it falls out of the scope of this study.

This appendix is organized as follows: first, we introduce the MGS functional and illustrate its limitations for imaging heterogeneous resistivity structures. Then, we propose a new workflow to improve MGS imaging for electrical resistivity tomography and induced polarization. We then test the workflow on synthetic cases constrained by various synthetic prior information. The workflow is then applied to a real case-study investigating a gold deposit. Our results demonstrate that the MGS inversion for resistivity and chargeability can be significantly improved by following the proposed workflow.

B.2 Methods

B.2.1 Inversion algorithm

The software used to perform the inversions is CRTOMO (Kemna, 2000), a two-dimensional finite-element complex resistivity inversion code. The regularized non-linear least-square inverse problem is formulated as the minimization of the following objective function

$$P^{\lambda}(m) = P_d(m) + \lambda P_m(m) \tag{B.1}$$

where $P_d(m)$ is the data misfit and $P_m(m)$ is the stabilizing functional, λ is the regularization parameter.

$$\begin{aligned} P^{\lambda}(m) = (d - G(m))^H W_d^H W_d (d - G(m)) + \\ \lambda \left(m^H W_m^T W_m m + \alpha (m - m_{ref})^H (m - m_{ref}) \right) \end{aligned} \tag{B.2}$$

where m is the model vector of the natural logarithm of the complex conductivity, W_m and W_d are the model functional and data weighting matrices, respectively. G is the non-linear forward operator, d is the data vector of the natural logarithm of apparent complex conductivity, λ is the regularization parameter, m_{ref} is the reference model chosen by the user, and α is a diagonal matrix containing closeness factors weighting the impact of the reference model cell by cell (Dumont et al., 2016). When only a starting model (and no reference model) is used in the inversion, that term is dropped from the objective function. The superscript H denotes the Hermitian. The objective function is similar to the formulation by Blaschek et al. (2008) except for the reference model term.

At each iteration, the code uses a Gauss-Newton scheme to iteratively compute the model update to minimize Equation B.2. At each iteration, before the model update, a line-search is performed to find the optimal λ value, that minimizes the data misfit. The observed data are fitted within their expected noise level, which is done using the chi-square misfit function:

$$\chi^2(m^k) = \sum_{i=1}^{n} \left(\frac{d_i^k - d_i^{obs}}{\sigma_i} \right)^2, \tag{B.3}$$

where n is the number of data points, m^k is the model parameter values matrix at iteration k, d^k, d^{obs} the vectors of computed data from model m^k and observed data, respectively. σ_i is the individual estimate of the data standard deviation for data point d_i. From statistical theory, an adequate value for the chi-square criterion to find a solution fitting the data within its noise level is close to the number of data point n Kemna (2000), which also corresponds to a root-mean square error e_{RMS} of 1,

$$e_{RMS} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{\chi^2(m^k)}{n}}. \tag{B.4}$$

With this stopping criterion, the inversions converge to fit the observed data within their estimated noise level. Once the root-mean-square error reaches a value below its target

value of 1, the model functional is maximized, fulfilling Occam's principle to reach an exact value of 1. All the inversions in this study converge to the target value and are therefore fitting the data to the same level. In complex resistivity inversions, the data misfit is commonly dominated by the real part. Therefore, an additional step for the inversion of the phase is carried out, during which the resistivity values are fixed and only the phase is updated (final phase improvement, see Kemna (2000)). This allows to refine the final phase solution by giving more weight to the phase data misfit.

In our field application, IP data were collected in the time domain. The conversion between time and frequency domain IP is made through a Laplace transform, based on the constant phase angle relaxation model (Kemna, 2000; Orozco et al., 2012). In the smoothness constraint inversion, W_m corresponds to the roughness matrix and the minimization of Equation B.1 leads to a smooth solution (Constable et al., 1987). For the MGS stabilizing functional, $P_m(m)$ is calculated as

$$P_{m,\beta}(m) = \int_V \frac{\nabla m \cdot \nabla m}{\nabla m \cdot \nabla m + \beta^2} dv \tag{B.5}$$

with β, the focusing parameter. V is the entire investigated volume, while $P_{m,\beta} \rightarrow$ support(∇m^k) for $\beta \rightarrow 0$, with support(∇m^k) being the volume in which ∇m^k is non-vanishing (Portniaguine and Zhdanov, 1999). This stabilizer enforces models with piece-wise constant functions in space. We follow the recommendation of Blaschek et al. (2008) to select β for field data, i.e., we scan the model space with several β values and choose the most appropriate value based on our prior knowledge about the geology, while avoiding the emergence of inversion artifacts. The selected β values for our inversions are all of the order of 10^{-3}.

B.2.2 Survey design

To avoid any influence of using different protocols between synthetic and field cases, we used the same multiple-gradient array in all instances. This electrode configuration is well-suited for multichannel acquisition, offers good resolution and balanced sensitivity to horizontal and vertical structures Aizebeokhai and Oyeyemi (2014); Dahlin and Zhou (2006), which is useful in the absence of a-priori information on the main directions of resistivity contrasts within the survey area (which is usually the case in field surveys). In addition, the multiple gradient array is known to be less sensitive to noise while ensuring resolution equivalent to the dipole-dipole array (Dahlin and Zhou, 2004), which is a significant advantage in ERT/IP surveys where high signal-to-noise ratios are particularly important, ensuring quality of the IP decay curves. The acquisition parameters of our multiple-gradient protocol are a separation factor s=8, and electrode spacing parameter a=1, 2, 3, 4, 5, 6, 7, 8, 9, the latter being the maximum possible value for our profile of 96 electrodes. The pre-processing and filtering of the field data is explained in section "Field case – Data acquisition and processing".

B.2.3 Limitations of MGS regularization for models with high heterogeneity

In this section, we investigate a heterogeneous resistivity model (no chargeability or phase considered) to assess the ability of MGS to recover complicated structures. A 2D resistivity model with sharp contrasts is built (Figure B.1A). The synthetic dataset is generated, and the resulting resistance values contaminated by a 3% Gaussian noise component. The smoothness constraint solution (Figure B.1C) retrieves some of the different blocks and the associated resistivity contrasts, but the MGS inversion (Figure B.1B) completely fails to produce a model depicting the expected resistive blocks. Both solutions are equally fitting the data. In this case, the MGS inversion can only resolve the sharp contrast corresponding to the synthetic soil/bedrock interface and separates the model in two homogeneous distinct parts. We hypothesize that the minimization is trapped in a local minimum due to the combination of loss of resolution with depth and the tendency of MGS to limit the number of homogeneous blocks to a minimum. Those simulations illustrate that the MGS functional suffers from important limitations when the degree of heterogeneity of the resistivity spatial distribution is high. To quantitatively appraise the solutions, model misfits can be computed with a L1-norm and weighted by the number of cells N:

$$M_{L1} = \sum_{c=1}^{N} \left| \frac{m_c - m_{c,true}}{N} \right|, \tag{B.6}$$

with m_c the logarithmic resistivity or phase value computed at cell c, and $m_{c,true}$ the logarithmic resistivity or phase value of the true model at cell c. In this study the total number of cells is N=1545.

Note that the model misfits of both solutions (Figure B.1) are equivalent (Table 1). However, the structures recovered by the smooth solution are more realistic in this case, especially the heterogeneity within the bedrock, which is more favorable for decision-making (e.g., about drilling).

Figure	Description	M_{L1}
B.1B	Sharp solution	0.42
B.1C	Smooth solution	0.43

Table B.1: Model misfits (Figure B.1)

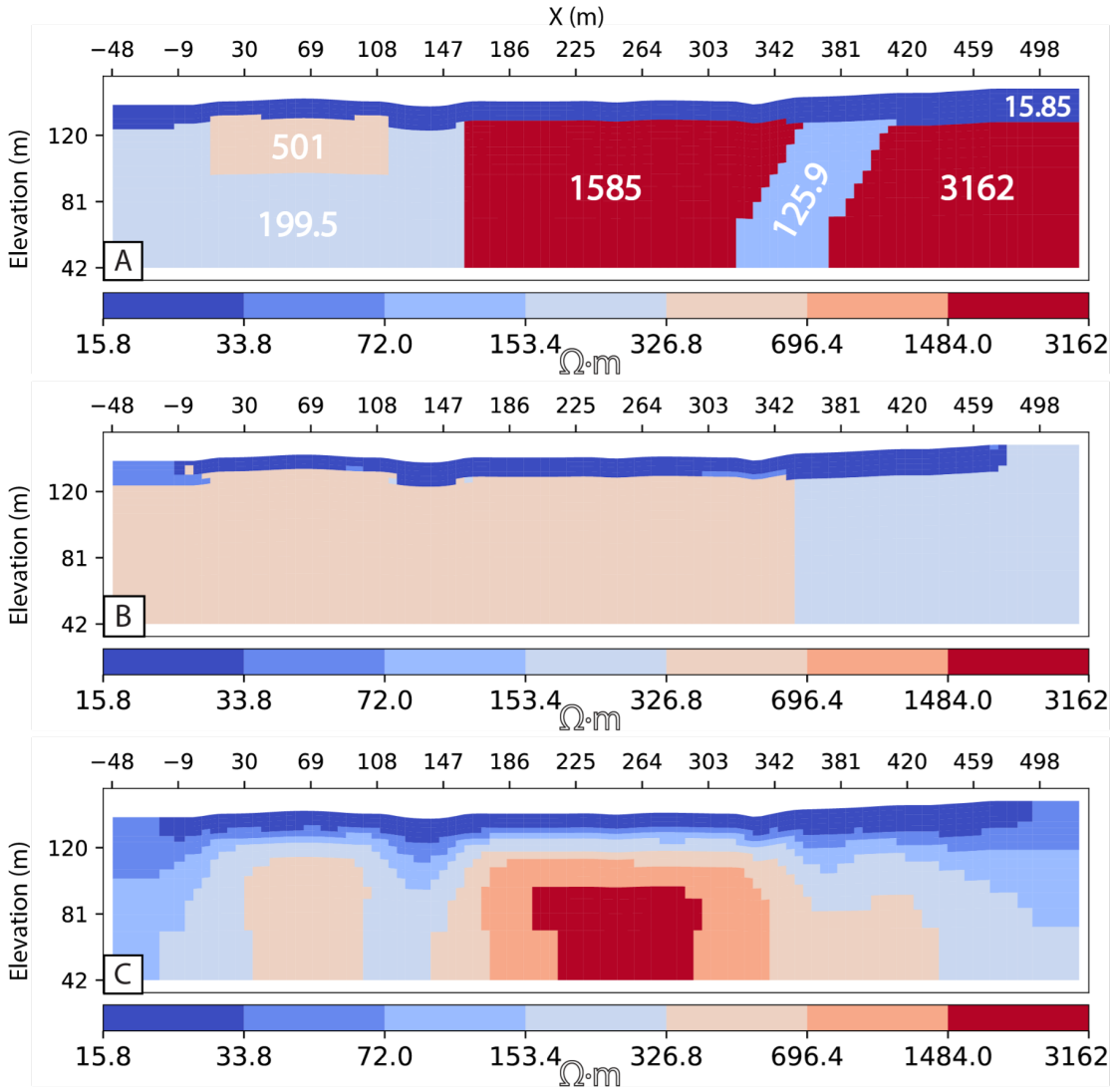

Figure B.1: Inversion results of a heterogeneous resistivity model depicting the inability of the MGS inversion to resolve highly heterogeneous resistivity fields. All sections share the same horizontal (X) and vertical (Elevation) axes in meters, as well as the resistivity color bar in Ωm. **A.** Synthetic model. **B.** MGS solution ($\beta = 70 \cdot 10^{-4}$). **C.** Smooth solution.

B.2.4 Workflow to incorporate prior information

To overcome the above limitation, we propose two new workflows, respectively named **A** & **B**, that will be applicable depending on the amount of prior information. First, for situations where little, such as only a few boreholes, to no prior information is available, we propose the workflow **A** composed of two steps: (I) Computation of a smoothness-constraint resistivity and phase solution. It is widely recognized that a smooth inversion, albeit not producing geologically realistic solutions, is robust and generates the minimum necessary structures to fit the data (Ajo-Franklin et al., 2007). (II) Using the smooth resistivity solution as a starting model, we switch regularization functionals to generate a sharp image. This switch will only create the necessary structures to maintain the data fit under the MGS functional. By using a L2-norm followed by the MGS, we aim to combine the benefits from both, preserving the robustness of the smooth inversion for detecting the necessary structure, while creating sharp contrasts.

Note that we do not input any specific reference value for the phase in the starting model (i.e., inversion starts with a constant phase of 0). The first reason is that the resistivity model will be the main contributor to the data misfit throughout the inversion. At the beginning of the final phase improvement, the phase model obtained during the previous step is used as starting model, so there is no need to add it at an earlier stage is not necessary. The second reason is that this approach limits the amount of prior information incorporated in the phase inversion, as the former is generally lacking. As we will see, this approach is sufficient to improve the results both in resistivity and in phase.

Usually, the combined effects of switching functionals, the convergence approach of CRTOMO and the initial null phase model suffice to compute a sharp solution, even if the starting model in resistivity is already fitting the data. In some rare cases, however, the convergence with the MGS may not be satisfactory because the solution is trapped in the minimum of the objective function provided by the smooth inversion. In such cases, we suggest increasing the smooth solution's resistivity values by a low percentage (e.g., 5%). This preserves the structures but artificially increases the data misfit and enables the sharp inversion to readily converge.

Second, if drill logs data or other kind of prior knowledge on the underlying geology is available, we propose the workflow **B** which consists of four steps: (I) Computation of the smooth resistivity and phase solution. (II) Design of an interpreted resistivity model based on the interpretation of the previous solutions and the geological knowledge. (III) Computation of the resistivity smooth solution using the resistivity interpreted model as a reference model. (IV) MGS inversion for the resistivity and phase, using the resistivity smooth solution from the previous step as the starting model as in workflow **A** (the phase is kept constant at zero).

It should be noted that the proposed methodology **B** uses a reference model for the final smoothness constraint inversion, but the resulting resistivity model is only used as a starting model for the MGS inversion, i.e., there is no further constraint during the

inversion (the last term of Equation B.2 is dropped). Therefore, the constraints on the MGS inversion are not strong. However, this approach allows to start from a model already close to the solution, so that the functional switch will only generate the model changes necessary to align with the MGS assumptions in the final solution (Nguyen et al., 2016). The final MGS step will favor piecewise constant functions from the starting model and penalize unnecessary ones. Furthermore, if the targets correspond to chargeability structures, we avoid introducing too much bias in the inversion process for chargeability because we never impose any specific chargeability structure in the reference or starting model (e.g., Caterina et al. 2014).

B.3 Application on a realistic synthetic case

A synthetic scenario inspired by field conditions is designed with both resistivity and chargeability structures (Figure B.2A and Figure B.2B), on which we apply the proposed workflows **A** (Figure B.2I and Figure B.2J) and **B** (Figures B.2C to B.2H) in the absence and presence of prior knowledge, respectively. Note that the resistivity and chargeability structures are different. All inversions were fitted to an estimated noise level of 1% resulting in a e_{RMSE} of 1 for the final iteration (Equation B.2). Prior information is simulated through virtual boreholes located at key locations as shown in Figure B.2C. Such adequate prior information is generally not realistic for real field cases. In this illustration case, the prior information is only intended to demonstrate that our approach can significantly improve imaging taking into account enough prior knowledge. For workflow **B** with prior information, at step I, the smooth solution of the resistivity and chargeability model is computed (Figures B.2C and B.2D).

We then build an interpreted resistivity model (Figure B.2E), including the information of the 5 virtual boreholes providing information on the lithology and the dip indicated by the delineation marks (see Figure B.2C) (step II). Although the global resistivity is lower than the true model, that interpreted model has a lower model misfit than the previous smooth solution. However, it is completely inconsistent with the dataset as indicated by a e_{RMSE} of 51.15. In order to find a model consistent with the dataset, the interpreted model is used in step III as the reference model to compute a smooth resistivity solution (Figure B.2F). Given the hypothesized presence of drill logs, the closeness factor α (Equation B.2) is defined by assigning weights to cells inversely proportional to the orthogonal distance between the cell centers and the drill holes, such that the solution is highly constrained only within the width of a few meters around the drill logs (not shown). For the last step (IV), the latter smoothness constraint inversion (Figure B.2F) is used as the starting model for the MGS inversion for both resistivity and phase.

Compared to the true model, in terms of structures, the workflow-based MGS solution (Figure B.2G and Figure B.2H) is superior to its smooth counterpart (Figure B.2C and Figure B.2D), which is further supported by the model misfits (see Table B.2). In particular, the MGS solution captures the main chargeable body while the smooth solution demonstrates clear artifacts. If we assume that no prior information is available and workflow **A** is applied, the MGS solution (Figure B.2I and Figure B.2J), computed using the smooth solution (Figure B.2C) directly is logically less performant than workflow **B**. However, it is more realistic and produces a lower model misfit than the original smooth solution, both in resistivity and chargeability (Table B.2). The absence of drill holes does not allow to recover the resistivity structures on the sides of the section where the sensitivity is lower. Note that the main dipping chargeability structure is hard to resolve since the mesh is structured. This limitation could be tackled similarly to Farquharson (2008), whose method includes a different roughness filter matrix with components that link the model cells that are diagonally adjacent, in order to recover oblique interfaces. Another possibility would be to refine the grid to let the regularization functional play an even more important role. However, this would come at the expense of a longer computation time.

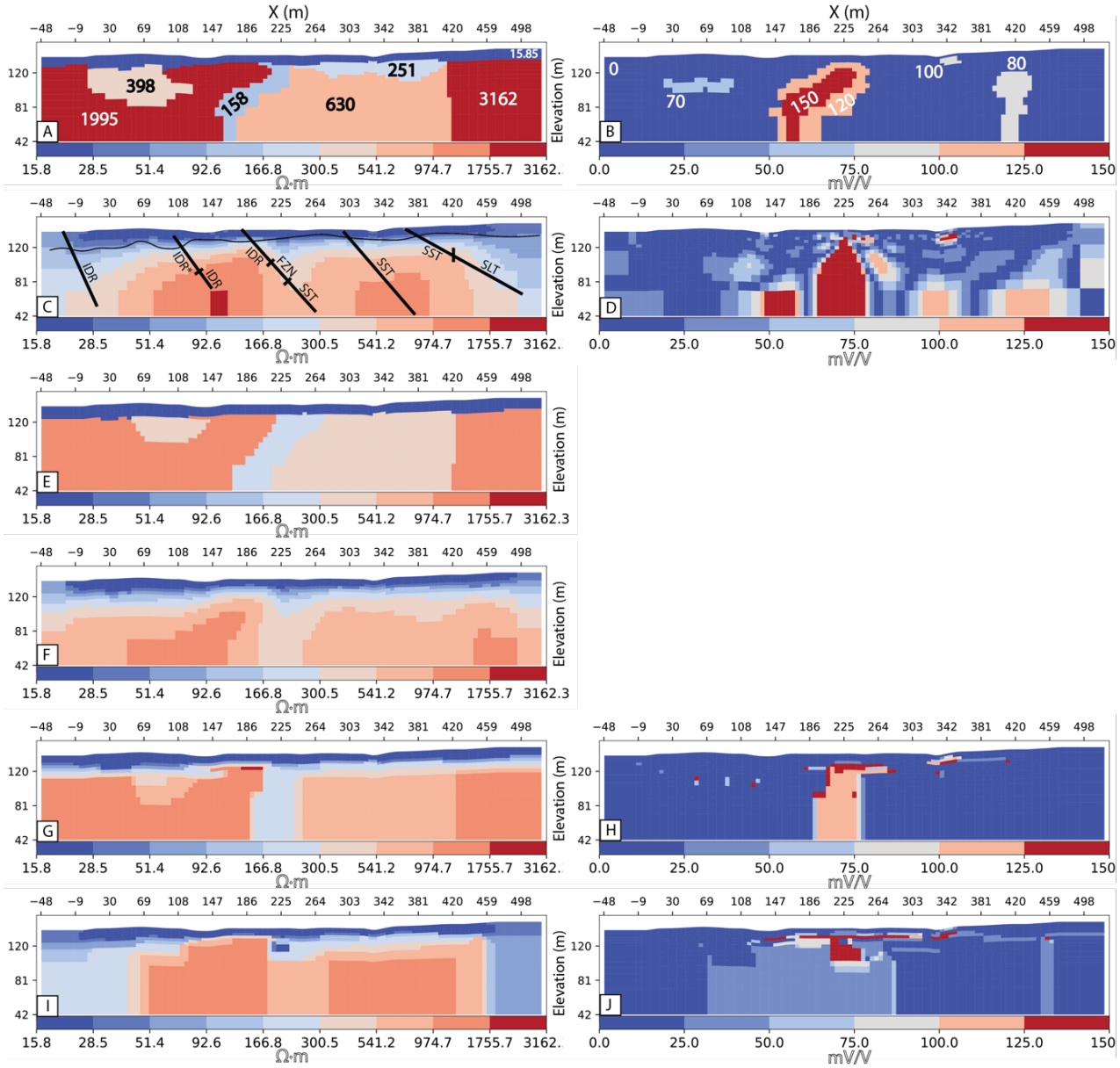

Figure B.2: Synthetic models and inversion results for a realistic heterogeneous synthetic scenario. **A.** Resistivity model (in $\Omega.m$). **B.** Chargeability model (in mV/V). **C.** Step B.I. Resistivity smooth solution with superimposed synthetic prior information. The letters refer to realistic lithological facies: IDR=diorite, IDR*=altered diorite, FZN=fault zone, SST=sandstone, SLT=siltstone. **D.** Step B.I. IP smooth solution. **E.** Step B.II. Interpreted resistivity model based on the drill logs data and smooth solutions (fig. C & D). **F.** Step B.III. Resistivity smooth solution using step B.II (fig. E) as a reference model. **G.** Step B.IV. Resistivity MGS solution ($\beta = 7 \times 10^{-4}$) constrained by the smooth solution from step B.III (fig. F) as a start model. **H.** Step B.IV. IP MGS solution constrained by the smooth solution from step B.III (fig. F) as a start model. **I.** Step A.II. Resistivity MGS solution ($\beta = 16 \times 10^{-4}$) using only the first smooth solution (fig. C) as a start model. **J.** Step A.II. IP MGS solution using only the first smooth solution (fig. C) as a start model.

Figure	**Step**	$\mathbf{M_{L1,\rho}(\log_{10}\Omega m)}$	$\mathbf{M_{L1,\Phi}(mV/V)}$
B.2C-D	B I	0.47	66
B.2E	B II	0.31	-
B.2F	B III	0.39	-
B.2G-H	B IV	0.30	47
B.2I-J	A	0.40	61

Table B.2: Note. Model misfitis (Figure B.2).

B.4 Application on a real case study

B.4.1 Geological context and petrophysical information

In March and April 2017, resistivity and time-domain induced polarization profiles were acquired to study a gold deposit in Cambodia. The deposit is a member of the newly defined Intrusion Related Gold System class of deposits (Hart, 2005). The geology of the study area is dominated by magmatic intrusions within bedded sedimentary formations. The targets of the investigation are mineralizations linked to chargeable sulfides, mainly pyrite, arsenopyrite and pyrrhotite. Figure B.3 summarizes a petrophysical study carried out at the laboratory scale on drill cores by Systems Exploration (NSW) Pty Limited (2008). 14 drill core samples of diorite and hornfelsed sediments were analyzed. As can be seen in Figure B.3, the highest chargeability responses are expected to emerge from these sulfides mineralizations distributed in vein networks. The laboratory results do not consider the presence of fractures and other structures that would influence the electrical conductivity and chargeability at a macro-scale, but they do indicate that zones of interest regarding mineralization are expected to be more conductive and produce a significant IP response, especially in fractured zones or in the case of well-connected mineralization networks (see also Kemna 2000). Pure diorite, sandstone or siltstone are poor conductors and cannot be differentiated on the basis of resistivity values alone. Mao et al. (2016) studied the time-domain IP response of disseminated pyrite particles in porous media and found values up to 150 mV/V. Evrard et al. (2018) conducted an induced polarization survey on a Pb-Zn-Fe sulfide deposit, and the chargeability responses of these mineralizations also ranged up to 150 mV/V, which are values consistent with our own field results.

Most of the mineralization is hosted in a Cretaceous diorite intrusion, embedded within a Triassic sedimentary sequence consisting of sandstone and siltstone, variably bedded and hornfelsed (Figure B.4). Gold occurrences seem to be associated with fractured zones within the diorite and at the contact intrusion/sedimentary host. Many drill logs are available and illustrate the complexity of the geology in the area (Figure B.4B). Unlike the synthetic case, the drill logs will not be directly put into the reference model because of their overabundance, which makes a clear interpretation difficult, but rather are interpreted globally in terms of main structures.

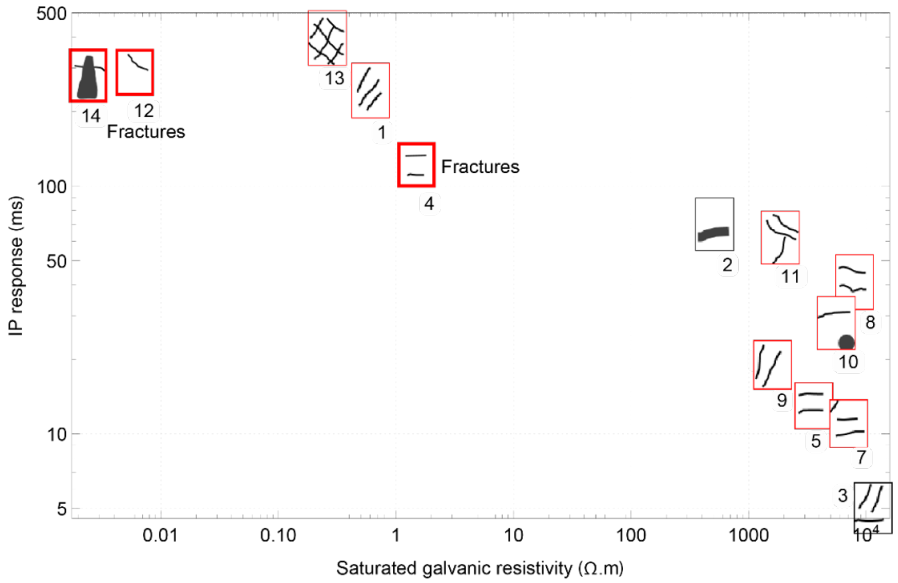

Figure B.3: Log-log plot of the IP response in ms versus the saturated galvanic resistivity in Ωm (Systems Exploration (NSW) Pty Limited., 2008). The markers illustrate the texture of each sample. The red frame indicates diorite and a gray frame sedimentary rock. Thick borders correspond to semi/massive sulfides mineralizations, and thin borders indicate veined/banded mineralizations.

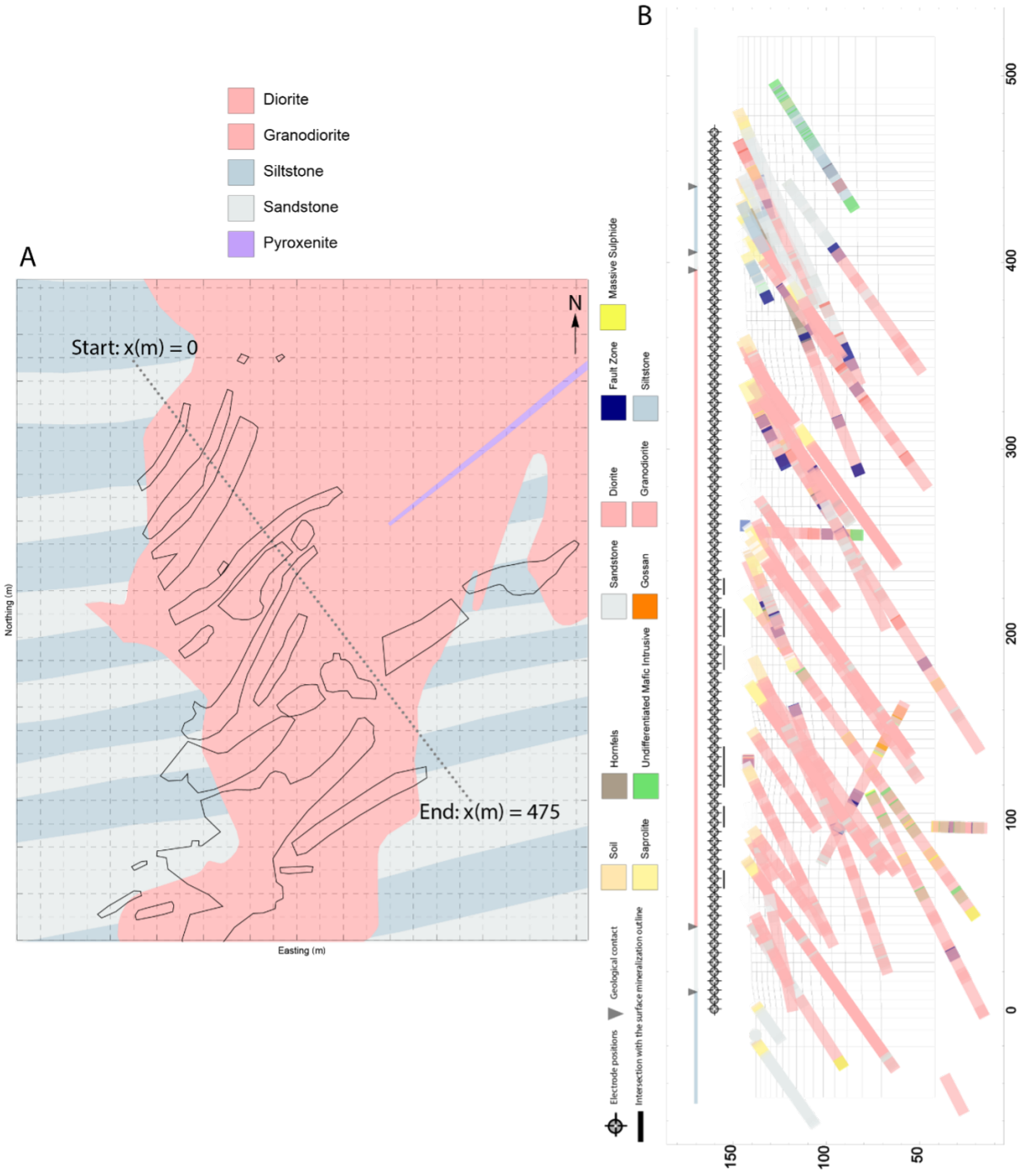

Figure B.4: **A.** Geological map of the area surrounding the selected profile (electrodes position as gray dots). The black lines delineate the surface mineralizations. Each square has the dimension 20 × 20 m and the profile is 475 m long. **B.** Cross section of the selected profile with available information.

B.4.2 Data acquisition and processing

An ERT/IP profile was done using a 475 m profile composed of 96 stainless steel electrodes with 5 m spacing, using the multiple-gradient array configuration (Dahlin and Zhou, 2006). Data acquisition first started with 64 electrodes and the profile was then extended using the roll-along technique. Due to an obstacle 150 m from the start, 4 electrodes had to be discarded, slightly reducing data coverage under this mark. The delay between the current turn-on and the start of potential measurement was 0.5 s. After current turn-off, 17 IP windows were measured for a total integration time of 3 s. The dead-time before recording the IP signal was set to 0.06 s. The instrument used was a Terrameter LS (ABEM), and the injected current was 500 mA (maximum achievable given the measuring sequence duration, and power availability at the remote field location in the middle of the jungle). Due to these harsh field conditions, it was not feasible to decouple current and potential cables as suggested by Dahlin and Leroux (2012) to limit the EM coupling effects on IP data. Therefore, to ensure good data quality, the IP data were filtered according to the methodology described in Evrard et al. (2018): only the potential curves with clear exponential decrease were kept. Figure B.5A illustrates such a curve, whereas B.5B, B.5C, B.5D are typical bad measurements which have been removed. Furthermore, outliers with a variation coefficient greater than 2% were rejected from the dataset (Figure B.6).

Out of 2150 initial data points, 100 were removed because they exceeded the repeatability threshold. Out of the 2050 remaining points, 248 points were rejected for having a bad IP curve as illustrated in Figure B.5B, B.5C, B.5D. In total, 15.8% of the original dataset was discarded. After data processing, the mean variation coefficient is 0.18% with a standard deviation of 0.29% (Figure B.6).

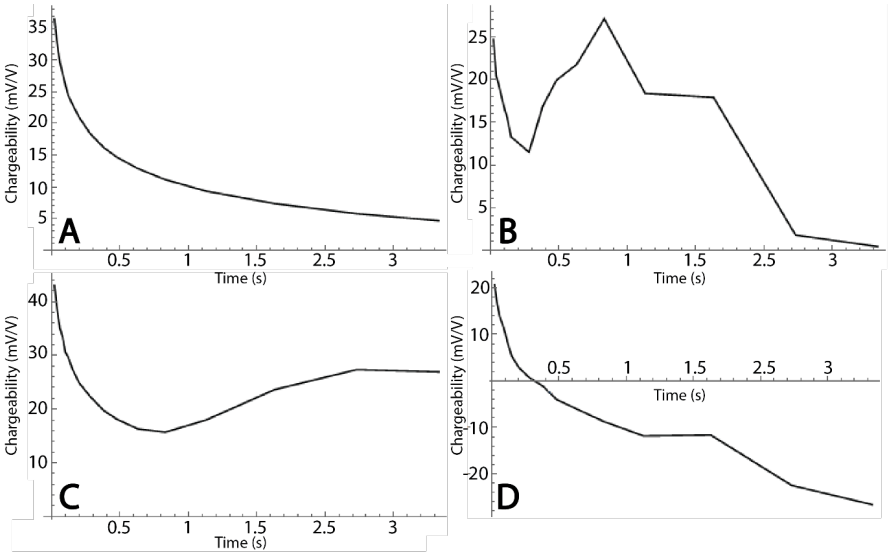

Figure B.5: IP curves (chargeability vs time) from the field dataset. **A.** Typical curve expected from the measurements. **B.**, **C.**, **D.** Bad IP curves filtered out of the dataset. **B.** Spurious oscillations occur. **C.** The curve begins to gradually decay but unexpectedly starts to rise after some time. **D.** The curve starts in the positive part of chargeability and plunges into the negative part after some time.

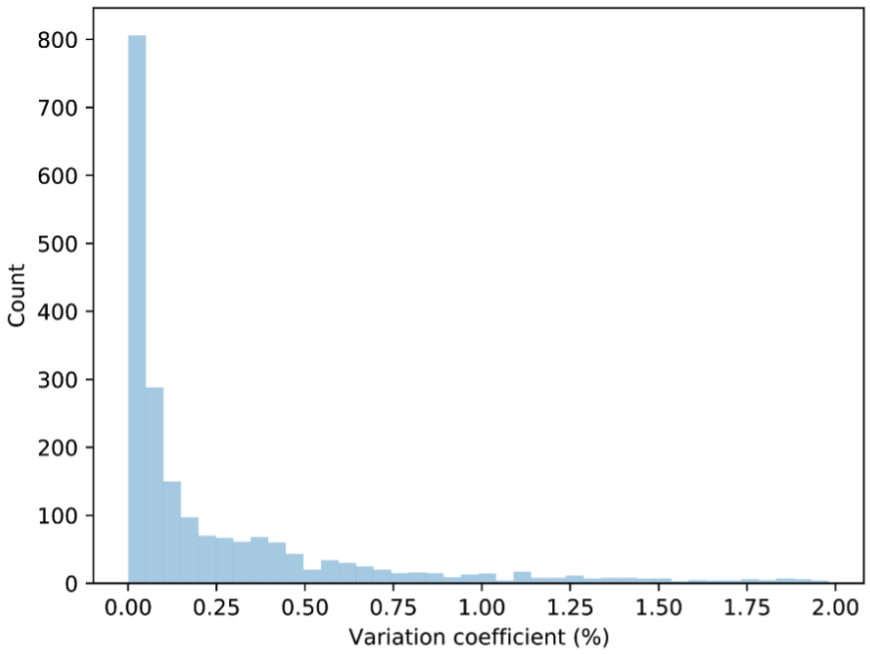

Figure B.6: Histogram of the variation coefficient for the field dataset.

B.4.3 Application of the workflow to the field data

Following the same procedure as the synthetic case, both proposed workflows **A** and **B** are applied to the field data in Figures B.7H and B.7I and Figures B.7C to B.7G respectively. As we are lacking a measure of error in the data, the relative error level was first assessed using the standard smoothness constraint inversion. Error magnitudes around 1% for the resistance and 0.2 mrad for the phase allowed the inversion process to readily converge without creating significant inversion artifacts. These error levels were then kept for the MGS inversions.

In step I, common to both workflows, the smooth resistivity and chargeability solutions are computed (Figures B.7A and B.7B). The main chargeability structure is roughly located at the interface between an inferred fractured zone and an inferred diorite block located in the middle of the profile, and extends into the latter. It also coincides with the surface mineralization locations (Figure B.4). Taking into account the available prior information from the smooth inversion and drill-holes, workflow **B** is applied, and an interpreted resistivity section is designed (Figure B.7C) in step II. On the far North-West (Figure B.4), a transition between magmatic and sedimentary rocks occurs, and is associated with a slight decrease in resistivity that we interpret as a conductive faulted zone. Based on the resistivity model, two other similar zones are interpreted on both sides of the diorite block in the middle of the profile (180m $<$ x $<$ 340m). The conductive zone depicted on the South-East of B.7C, at approximately x = 400 m, coincides with a geological contact (see Figure B.4B). The block on the far South-East is interpreted as resistive sandstone/siltstone, in accordance with the geological map and borehole data. The soil and weathered bedrock can be identified by their low and constant resistivity values along the profile and are well resolved.

Using the interpreted resistivity structure as the reference model for a smooth inversion in step III yields B.7E. Here, since the borehole data are not directly geometrically translated into weights as in the synthetic case, we adapt the choice of the closeness factor. We choose a closeness factor of 1 for the first meters corresponding to the conductive soil. A closeness factor of 0.5 is assigned to zones inferred as faults, and the rest of the model is constrained with a value of 0.05. B.7E is then used as the starting model for the MGS workflow **B** (Figures B.7F-7G) in step IV. For comparison, the workflow **A** solutions in which only the first smooth solution is used as the starting model are also computed and shown in Figures B.7H and B.7I. The smoothness constraint (Figure B.7A) and workflow-based MGS (Figure B.7F) resistivity solutions both allow the same interpretation in terms of resistivity distribution, with the MGS solution logically yielding sharper transitions. In contrast, in terms of chargeability, the workflow-based MGS solution (Figure B.7G) suggests that the surface contains several small chargeable structures, not identified by the smooth solution (Figure B.7B), although we cannot affirm these are not artifacts, surface mineralizations are confirmed by the geological maps and direct field observations. The MGS solution also indicates that the main chargeable bodies are located at the boundaries of a conductive zone, rather than being smoothly distributed. This result is supported by geological observations and miner-

alization processes in this area. Indeed, the main mineralizations are associated with fractured zones within the diorite intrusive body, and at the interface intrusion/sediments, such as shears, veins and breccia infill zones. In workflow **A**, the MGS solution constrained by the smooth solution (Figure B.7A) as the starting model without adding information from the boreholes (Figures B.7H-7I) resolves as well several chargeability structures in the subsurface and is imaging more focused and realistic structures compared to the smoothness constraint solution. Difference in terms of structure with workflow **B** is minimal, indicating that a workflow with limited prior information already significantly improves the solution for resistivity and chargeability.

The depth of investigation index (DOI) (Figure B.7D) is computed from the smoothness constraint solution according to Caterina et al. (2013). The DOI values are interpreted by looking at their gradient as recommended by Oldenborger et al. (2007), which is maximal for DOI values close to 0.1, confirming that the model remains well constrained by the data to depths greater than 50 m below the surface. The inferred highly resistive diorite block in the middle of the profile impedes current spread, and its image during the inversions is logically less constrained by the data, as indicated by the strong DOI values at this location. The cells on the far right and far left of the grid are, as expected, less influenced by the data, since these areas are simply not covered by any measures due to the survey design.

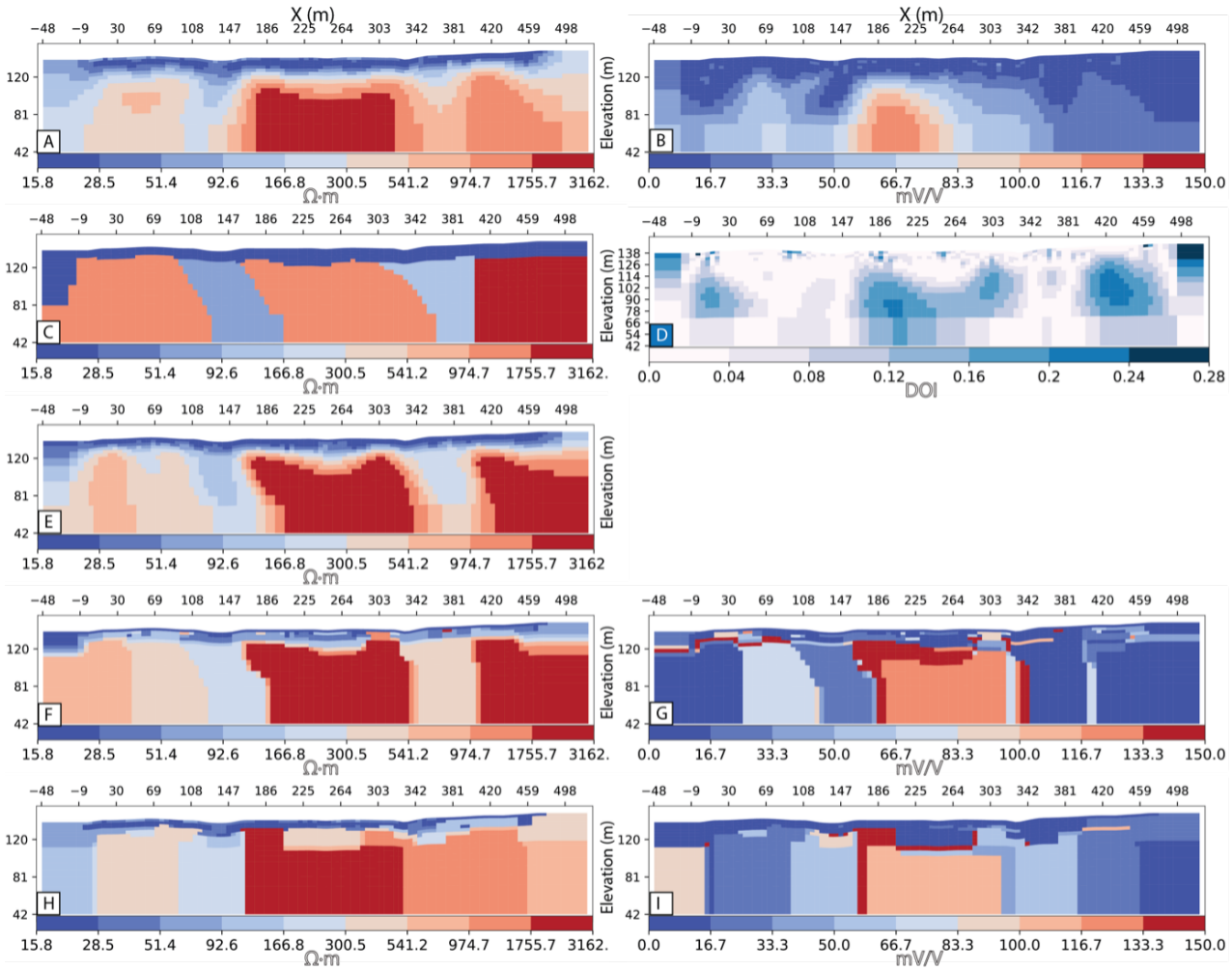

Figure B.7: Inversion results of a real field investigation. **A.** Step B.I. Smooth resistivity solution. **B.** Step B.I. Smooth chargeability solution. **D.** DOI computed with resistivity reference models of 10^1 and $10^3 \Omega m$ with constant closeness factor of 0.05. **C.** Step B.II. Interpreted resistivity model based on the smooth solutions and drill logs. **E.** Step B.III. Smooth resistivity solution constrained by the reference model defined in step II. **F.** Step B.IV. Resistivity MGS solution ($\beta = 12 \times 10^{-4}$) constrained by the smooth solution from step III. **G.** Step B.IV. MGS chargeability solution constrained by the smooth solution from step III. **H.** Step A.II. MGS resistivity solution ($\beta = 12 \times 10^{-4}$) using only the smooth solution from step I as the starting model. **I.** Step A.II. MGS chargeability solution using only the smooth solution from step I as the starting model.

B.5 Conclusion

In this contribution, we propose a new workflow combining the smoothness constraint solution and the MGS approach for the inversion of resistivity and induced polarization data. In the absence of additional prior information, we simply use the smoothness constraint solution as a starting model for the MGS inversion, which already yields significant improvements for both resistivity and chargeability imaging. When additional prior information is available, it is interpreted in conjunction with the smooth solution to create a realistic reference model. The latter is used for an improved smooth solution, the results of which are used as a starting model for the MGS solution. Our results show that workflow-based MGS inversions converge towards satisfactory results compared to ground truth for the synthetic and the field cases. The workflow-based MGS inversion improves imaging and resolves several geological structures missing from the smooth solution. We think that the proposed workflow can improve imaging results in many settings where sharp contrasts are expected, as the method can be applied both in the presence and in the absence of significant prior information. In the latter case, the smooth solution simply serves as a starting model for the MGS inversion and makes it possible to retrieve several sharp structures.

The proposed method does not require many additional computations but just a few more inversions. In the synthetic and field cases, it provides significant improvements over direct MGS inversion and standard smoothness constraint and therefore can be widely applied to actual field investigations. It combines the robustness of the smoothness constraint inversion to image the necessary structure to explain the observed data with the ability of the MGS to generate sharp contrasts.
Using a two-step inversion approach also ensures faster convergence of the MGS inversion (in a few iterations). It also opens the possibility to apply the approach developed by Nguyen et al. (2016) for the optimization of β for time-lapse data sets to static surveys. This will be investigated in future research.

Acronyms

ABC Approximate Bayesian Computation.

AI Artificial Intelligence.

ANN Artificial Neural Network.

API Application Programming Interface.

ATES Aquifer Thermal Energy Storage.

BBB Bayes By Backprop.

BC Boundary Condition.

BEL Bayesian Evidential Learning.

BMA Bayesian Model Averaging.

BNN Bayesian Neural Network.

BOED Bayesian Optimal Experimental Design.

BP Backpropagation.

BTES Borehole Thermal Energy Storage.

CCA Canonical Correlation Analysis.

CDE Conditional Density Estimation.

COP Combinatorial Optimization Problem.

DC Direct Current.

DD Dipole-Dipole.

DESPOT Determinized Sparse Partially Observable Tree.

DI Discrimination-Inference.

DL Deep Learning.

DOI Depth of Investigation.

DREAM DiffeRential Evolution Adaptive Metropolis.

ED Experimental Design.

EDF Empirical Distribution Function.

ELBO Evidence Lower BOund.

EM Electromagnetic.

EOI Efficacy of Information.

ERT Electrical Resistivity Tomography.

FA Factor Analysis.

FMM Fast-Marching Method.

GB GigaByte.

GCG Generalized Conjugate Gradient.

GHz GigaHertz.

GPR Gaussian Process Regression.

GW Groundwater.

HFB Horizontal Flow Barrier.

HMOC Hybrid Method of Characteristics.

ICA Independent Component Analysis.

ID Identification.

IG Information Gain.

IP Induced Polarization.

IPR Iterative Prior Resampling.

IQR Interquartile Range.

KDE Kernel Density Estimation.

KL Kullback-Leibler.

LDA Linear Discriminant Analysis.

LS Least Squares.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

MC Monte Carlo.

MCMC Markov Chain Monte Carlo.

MCTS Monte Carlo Tree Search.

MDN Mixture Density Network.

MDP Markov Decision Process.

MDS Multidimensional Scaling.

MG Multivariate Gaussian.

MGI Multivariate Gaussian Inference.

MGS Mimimum Gradient Support.

MHD Modified Hausdorff Distance.

ML Machine Learning.

MLE Maximum Likelihood Estimation.

MOGP Multi-Output Gaussian Process.

MSE Mean Squared Error.

NGB Natural Gradient Boosting.

NLL Negative Log-Likelihood.

NN Neural Network.

ODE Ordinary Differential Equation.

OED Optimal Experimental Design.

OLS Ordinary Least Squares.

OSSE Observation System Simulation Experiment.

PBNN Probabilistic Bayesian Neural Network.

PC Principal Component.

PCA Principal Component Analysis.

PDE Probability Density Estimation.

PDF Probability Density Function.

PNN Probabilistic Neural Network.

POMDP Partially Observable Markov Decision Process.

RAM Random Access Memory.

ReLU Rectified Linear Unit.

RMSE Root Mean Squared Error.

RS Rejection Sampling.

SBI Simulation-Based Inference.

SD Signed Distance.

SDF Signed Distance Function.

SGD Stochastic Gradient Descent.

SGEMS Stanford Geostatistical Modeling Software.

SGS Sequential Gaussian Simulation.

SGSIM Sequential Gaussian Simulation.

SKBEL scikit-BEL.

SP Stress Period.

SSIM Structural Similarity Index.

SVD Single Value Decomposition.

SVI Stochastic Variational Inference.

SW Surface Water.

SWI Salt Water Intrusion.

t-SNE t-distributed Stochastic Neighbor Embedding.

TBRS Threshold-Based Rejection Sampling.

TM Transport Map.

TSP Travelling Salesman Problem.

TVD Total Variation Diminishing.

UCT Upper Confidence Bound for Trees.

UF Utility Function.

UMAL Uncountable Mixture of Asymmetric Laplacians.

UMAP Uniform Manifold Approximation and Projection.

UQ Uncertainty Quantification.

VOI Value of Information.

WHPA Wellhead Protection Area.

List of Figures

List of Tables

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Aizebeokhai, A. P. and Oyeyemi, K. D. (2014). The use of the multiple-gradient array for geoelectrical resistivity and induced polarization imaging. *Journal of Applied Geophysics*, 111:364–376.

Ajo-Franklin, J. B., Minsley, B. J., and Daley, T. M. (2007). Applying compactness constraints to differential traveltime tomography. *Geophysics*, 72:1–18.

Alfonso, L., Lobbrecht, A., and Price, R. (2010). Optimization of water level monitoring network in polder systems using information theory. *Water Resources Research*, 46.

Alsing, J., Charnock, T., Feeney, S., and Wandelt, B. (2019). Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notices of the Royal Astronomical Society*.

Amiri, V. and Nakagawa, K. (2021). Using a linear discriminant analysis (LDA)-based nomenclature system and self-organizing maps (SOM) for spatiotemporal assessment of groundwater quality in a coastal aquifer. *Journal of Hydrology*, 603:127082.

Anderson, M. P. (2005). Heat as a ground water tracer. *Ground Water*, 43:951–968.

Aquilina, L., Vergnaud-Ayraud, V., Landes, A. A. L., Pauwels, H., Davy, P., Pételet-Giraud, E., Labasque, T., Roques, C., Chatton, E., Bour, O., Maamar, S. B., Dufresne, A., Khaska, M., Salle, C. L. G. L., and Barbecot, F. (2015). Impact of climate changes during the last 5 million years on groundwater in basement aquifers. *Scientific Reports*, 5:14132.

Arato, A., Boaga, J., Comina, C., Seta, M. D., Sipio, E. D., Galgaro, A., Giordano, N., and Mandrone, G. (2015). Geophysical monitoring for shallow geothermal applications – two Italian case histories. *First Break*, 33.

Asher, M. J., Croke, B. F. W., Jakeman, A. J., and Peeters, L. J. M. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research*, 51:5957–5973.

Ashley, R. and Lloyd, J. (1978). An example of the use of factor analysis and cluster analysis in groundwater chemistry interpretation. *Journal of Hydrology*, 39:355–364.

Aster, R. C., Borchers, B., and Thurber, C. H. (2013). *Parameter Estimation and Inverse Problems*. Elsevier.

Athens, N. D. and Caers, J. K. (2019). A Monte Carlo-based framework for assessing the value of information and development risk in geothermal exploration. *Applied Energy*, 256:113932.

Attia, A., Alexanderian, A., and Saibaba, A. K. (2018). Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems*, 34:095009.

Auken, E., Christiansen, A. V., Kirkegaard, C., Fiandaca, G., Schamper, C., Behroozmand, A. A., Binley, A., Nielsen, E., Effersø, F., Christensen, N. B., Sørensen, K., Foged, N., and Vignoli, G. (2015). An overview of a highly versatile forward and stable inverse algorithm for airborne, ground-based and borehole electromagnetic and electric data. *Exploration Geophysics*, 46:223–235.

Auken, E., Doetsch, J., Fiandaca, G., Christiansen, A. V., Gazoty, A., Cahill, A. G., and Jakobsen, R. (2014). Imaging subsurface migration of dissolved CO2 in a shallow aquifer using 3-D time-lapse electrical resistivity tomography. *Journal of Applied Geophysics*, 101:31–41.

Babaei, M., Pan, I., and Alkhatib, A. (2015). Robust optimization of well location to enhance hysteretical trapping of CO2: Assessment of various uncertainty quantification methods and utilization of mixed response surface surrogates. *Water Resources Research*, 51:9402–9424.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., and Fienen, M. N. (2016). Scripting MODFLOW model development using Python and flopy. *Groundwater*, 54:733–739.

Baptista, R., Marzouk, Y., and Zahm, O. (2022). On the representation and learning of monotone triangular transport maps. *arXiv*.

Baya Veliz, G. (2017). Influence of riverbank seepage on river-aquifer interactions at the Aa river. *Master's Thesis, Vrije Universiteit Brussel (VUB) & KU Leuven, Brussels, Belgium*.

Bayer, P., Rybach, L., Blum, P., and Brauchler, R. (2013). Review on life cycle environmental effects of geothermal power generation. *Renewable and Sustainable Energy Reviews*, 26:446–463.

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical transactions of the Royal Society of London*, (53):370–418.

Baú, D. A. and Mayer, A. S. (2006). Stochastic management of pump-and-treat strategies using surrogate functions. *Advances in Water Resources*, 29:1901–1917.

Başağaoğlu, H., Chakraborty, D., and Winterle, J. (2021). Reliable evapotranspiration predictions with a probabilistic machine learning framework. *Water*, 13:557.

Bedekar, V., Morway, E. D., Langevin, C. D., and Tonkin, M. J. (2016). MT3D-USGS version 1: A US Geological Survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW. Technical report, US Geological Survey.

Bellman, R. (1961). Adaptive control processes: A guided tour. (a rand corporation research study). Princeton, N. J.: Princeton University Press, XVI, 255 p. (1961).

Bergen, K. J., Johnson, P. A., Hoop, M. V. d., and Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 363(6433).

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Beven, K. (2019). How to make advances in hydrological modelling. *Hydrology Research*, 50:1481–1494.

Beven, K. (2020). Deep learning, hydrological processes and the uniqueness of place. *Hydrological Processes*, 34:3608–3613.

Beven, K., Asadullah, A., Bates, P., Blyth, E., Chappell, N., Child, S., Cloke, H., Dadson, S., Everard, N., Fowler, H. J., Freer, J., Hannah, D. M., Heppell, K., Holden, J., Lamb, R., Lewis, H., Morgan, G., Parry, L., and Wagener, T. (2020). Developing observational methods to drive future hydrological science: Can we make a start as a community? *Hydrological Processes*, 34:868–873.

Beven, K. and Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6:279–298.

Bilenko, N. Y. and Gallant, J. L. (2016). Pyrcca: Regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, 10.

Bishop, C. M. (1994). Mixture density networks.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition.

Blanchy, G., Watts, C. W., Richards, J., Bussell, J., Huntenburg, K., Sparkes, D. L., Stalham, M., Hawkesford, M. J., Whalley, W. R., and Binley, A. (2020). Time-lapse geophysical assessment of agricultural practices on soil moisture dynamics. *Vadose Zone Journal*, 19.

Blaschek, R., Hördt, A., and Kemna, A. (2008). A new sensitivity-controlled focusing regularization scheme for the inversion of induced polarization data based on the minimum gradient support. *Geophysics*, 73:45–54.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France. PMLR.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., de Amorim, P. B., Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Y., Chen, Y., Chifflard, P., Claps, P., Clark, M. P., Collins, A. L., Croke, B., Dathe, A., David, P. C., de Barros, F. P. J., de Rooij, G., Baldassarre, G. D., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Bevacqua, A. G., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T. H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M. L. R., Lindquist, E., Link, T., Liu, J., Loucks, D. P., Luce, C., Mahé, G., Makarieva, O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B. D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V. O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M. J., Post, D., Sierra, C. P., Ramos, M.-H., Renner, M., Reynolds, J. E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D. E., Rosso, R., Roy, T., Sá, J. H., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R. C., Skaugen, T., Smith, H., Spiessl, S. M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R. J., van der Ploeg, M., Loon, A. F. V., van Meerveld, I., van Nooijen, R., van Oel, P. R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A. J., Ward, P., Westerberg, I. K., White,

C., Wood, E. F., Woods, R., Xu, Z., Yilmaz, K. K., and Zhang, Y. (2019). Twenty-three unsolved problems in hydrology (UPH) – a community perspective. *Hydrological Sciences Journal*, 64:1141–1158.

Borga, M. (1998). *Learning Multidimensional Signal Processing*. PhD thesis, Linköping University Electronic Press.

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, 53:243–255.

Bouchedda, A., Bernard, G., and Gloaguen, E. (2017). Constrained electrical resistivity tomography Bayesian inversion using inverse Matérn covariance matrix. *GEOPHYSICS*, 82(3):E129–E141.

Box, G. E., Hunter, W. H., Hunter, S., et al. (1978). *Statistics for experimenters*, volume 664. John Wiley and sons New York.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Bredehoeft, J. D. and Papaopulos, I. S. (1965). Rates of vertical groundwater movement estimated from the Earth's thermal profile. *Water Resources Research*, 1:325–328.

Bridger, D. W. and Allen, D. M. (2010). Heat transport simulations in a heterogeneous aquifer used for aquifer thermal energy storage (ATES). *Canadian Geotechnical Journal*, 47:96–115.

Brouyère, S. (2001). *Etude et modélisation du transport et du piégeage des solutés en milieu souterrain variablement saturé (study and modelling of transport and retardation of solutes in variably saturated media)*. Phd thesis edition.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Brunner, P. and Simmons, C. T. (2012). Hydrogeosphere: A fully integrated, physically based hydrological model. *Ground Water*, 50:170–176.

Brussels Times (2021). New rainfall, new flooding - new normal? URL: `https://www.brusselstimes.com/178999/new-rainfall-new-flooding-new-normal`– Accessed: 2022-12-14.

Brussels Times (2022a). One year since deadly floods, is Belgium now better prepared for extreme weather? URL: `https://www.brusselstimes.com/254360/one-year-since-deadly-floods-is-belgium-now-better-prepared-for-extreme-weather`– Accessed: 2022-12-14.

Brussels Times (2022b). Wallonia: Over €20 million required to repair flood damage. URL: https://www.brusselstimes.com/251312/wallonia-over-e20-million-required-to-repair-flood-damage– Accessed: 2022-12-14.

Brussels Times (2022c). Wallonia to create temporary immersion basins to combat floods. URL: https://www.brusselstimes.com/belgium/228747/wallonia-to-create-temporary-immersion-basins-to-combat-floods– Accessed: 2022-12-14.

Böttcher, S., Merz, C., Lischeid, G., and Dannowski, R. (2014). Using isomap to differentiate between anthropogenic and natural effects on groundwater dynamics in a complex geological setting. *Journal of Hydrology*, 519:1634–1641.

Caers, J. (2011). *Modeling Uncertainty in the Earth Sciences*. Wiley.

Caers, J., Scheidt, C., Yin, Z., Wang, L., Mukerji, T., and House, K. (2022). Efficacy of information in mineral exploration drilling. *Natural Resources Research*, 31:1157–1173.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37.

Carreau, J., Naveau, P., and Sauquet, E. (2009). A statistical rainfall-runoff mixture model with heavy-tailed components. *Water Resources Research*, 45(10).

Carreau, J. and Vrac, M. (2011). Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 47(10).

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press.

Castillo, A., Castelli, F., and Entekhabi, D. (2015). An entropy-based measure of hydrologic complexity and its applications. *Water Resources Research*, 51:5145–5160.

Caterina, D., Beaujean, J., Robert, T., and Nguyen, F. (2013). A comparison study of different image appraisal tools for electrical resistivity tomography. *Near Surface Geophysics*, 11(6):639–658.

Caterina, D., Hermans, T., and Nguyen, F. (2014). Case studies of incorporation of prior information in electrical resistivity tomography: comparison of different approaches. *Near Surface Geophysics*, 12:451–465.

Chacon-Hurtado, J. C., Alfonso, L., and Solomatine, D. P. (2017). Rainfall and streamflow sensor network design: a review of applications, classification, and a proposed framework. *Hydrology and Earth System Sciences*, 21(6):3071–3091.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10.

Chaslot, G., Bakkes, S., Szita, I., and Spronck, P. (2008). Monte-Carlo tree search: A new framework for game ai. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 4, pages 216–217.

Chasseriau, P. and Chouteau, M. (2003). 3D gravity inversion using a model of parameter covariance. *Journal of applied geophysics*, 52(1):59–74.

Chen, J., Dai, Z., Dong, S., Zhang, X., Sun, G., Wu, J., Ershadnia, R., Yin, S., and Soltanian, M. R. (2022). Integration of deep learning and information theory for designing monitoring networks in heterogeneous aquifer systems. *Water Resources Research*, 58(10).

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Clarke, R. T. (2008). Issues of experimental design for comparing the performance of hydrologic models. *Water Resources Research*, 44.

Close, M., Abraham, P., Humphries, B., Lilburne, L., Cuthill, T., and Wilson, S. (2016). Predicting groundwater redox status on a regional scale using linear discriminant analysis. *Journal of Contaminant Hydrology*, 191:19–32.

Constable, S. C., Parker, R. L., and Constable, C. G. (1987). Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, 52(3):289–300.

Constantz, J. (2008). Heat as a tracer to determine streambed water exchanges. *Water Resources Research*, 44.

Cook, R. D. (2022). A slice of multivariate dimension reduction. *Journal of Multivariate Analysis*, 188:104812.

Corso, A., Wang, Y., Zechner, M., Caers, J., and Kochenderfer, M. J. (2022). A POMDP model for safe geological carbon sequestration.

Coulibaly, P., Anctil, F., and Bobée, B. (2000). Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *Journal of Hydrology*, 230:244–257.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Cox, M. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117:30055–30062.

Dahlin, T. and Leroux, V. (2012). Improvement in time-domain induced polarization data quality with multi-electrode systems by separating current and potential cables. *Near Surface Geophysics*, 10(6):545–565.

Dahlin, T. and Zhou, B. (2004). A numerical comparison of 2D resistivity imaging with 10 electrode arrays. *Geophysical prospecting*, 52(5):379–398.

Dahlin, T. and Zhou, B. (2006). Multiple-gradient array measurements for multichannel 2D resistivity imaging. *Near Surface Geophysics*, 4:113–123.

Darcy, H. (1856). *Les fontaines publiques de la ville de Dijon: Exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau: Ouvrage terminé par un appendice relatif aux fournitures d'eau de plusieurs villes, au filtrage des eaux et à la fabrication des tuyaux de fonte, de plomb, de tôle et de bitume*, volume 2. V. Dalmont.

Dassargues, A. (1997). Modeling baseflow from an alluvial aquifer using hydraulic-conductivity data obtained from a derived relation with apparent electrical resistivity. *Hydrogeology Journal*, 5:97–108.

Dassargues, A. (2018). *Hydrogeology: Groundwater science and engineering*. CRC Press.

Davis, D. R., Kisiel, C. C., and Duckstein, L. (1972). Bayesian decision theory applied to design in hydrology. *Water Resources Research*, 8:33–41.

Dawdy, D. R. and Feth, J. H. (1967). Applications of factor analysis in study of chemistry of groundwater quality, Mojave River Valley, California. *Water Resources Research*, 3:505–510.

Dawson, C. and Wilby, R. (2001). Hydrological modelling using artificial neural networks. *Progress in physical Geography*, 25(1):80–108.

de Barros, F. P., Ezzedine, S., and Rubin, Y. (2012). Impact of hydrogeological data on measures of uncertainty, site characterization and environmental performance metrics. *Advances in Water Resources*, 36:51–63.

De Regt, H. W. and Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, 144(1):137–170.

de Wolff, T., Cuevas, A., and Tobar, F. (2021). MOGPTK: The multi-output Gaussian process toolkit. *Neurocomputing*, 424:49–53.

Deleersnyder, W., Maveau, B., Dudal, D., and Hermans, T. (2022). Flexible quasi-2D inversion of time-domain AEM data, using a wavelet-based complexity measure.

Deleersnyder, W., Maveau, B., Hermans, T., and Dudal, D. (2021). Inversion of electromagnetic induction data using a novel wavelet-based and scale-dependent regularization term. *Geophysical Journal International*, 226(3):1715–1729.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR.

Derouane, J. and Dassargues, A. (1998). Delineation of groundwater protection zones based on tracer tests and transport modeling in alluvial sediments. *Environmental Geology*, 36:27–36.

Diaby, M. and Karwan, M. H. (2016). *Advances in combinatorial optimization: linear programming formulations of the traveling salesman and other hard combinatorial optimization problems*. World Scientific.

Diaz, G., Sewell, J. I., and Shelton, C. H. (1968). An application of principal component analysis and analysis in the study of water yield. *Water Resources Research*, 4:299–306.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M. D., and Saurous, R. A. (2017). TensorFlow Distributions. *CoRR*, abs/1711.10604.

Doetsch, J., Linde, N., Pessognelli, M., Green, A. G., and Günther, T. (2012). Constraining 3-D electrical resistance tomography with GPR reflection data for improved aquifer characterization. *Journal of Applied Geophysics*, 78:68–76.

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review.

Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A. Y., and Schuler, A. (2019). NGBoost: Natural gradient boosting for probabilistic prediction.

Dubuisson, M.-P. and Jain, A. (1994). A modified Hausdorff distance for object matching. In *Proceedings of 12th International Conference on Pattern Recognition*. IEEE Computer Society Press.

Duijff, R., Bloemendal, M., and Bakker, M. (2021). Interaction effects between aquifer thermal energy storage systems. *Groundwater*.

Dujardin, J., Anibas, C., Bronders, J., Jamin, P., Hamonts, K., Dejonghe, W., Brouyère, S., and Batelaan, O. (2014). Combining flux estimation techniques to improve characterization of groundwater–surface-water interaction in the Zenne River, Belgium. *Hydrogeology Journal*, 22:1657–1668.

Dumont, G., Pilawski, T., Dzaomuho-Lenieregue, P., Hiligsmann, S., Delvigne, F., Thonart, P., Robert, T., Nguyen, F., and Hermans, T. (2016). Gravimetric water distribution assessment from geoelectrical methods (ERT and EMI) in municipal solid waste landfill. *Waste management*, 55:129–140.

Eidsvik, J., Martinelli, G., and Bhattacharjya, D. (2018). Sequential information gathering schemes for spatial risk and decision analysis applications. *Stochastic Environmental Research and Risk Assessment*, 32(4):1163–1177.

Eidsvik, J., Mukerji, T., and Bhattacharjya, D. (2015). *Value of Information in the Earth Sciences*. Cambridge University Press.

Einstein, A. (1905). Zur elektrodynamik bewegter körper. *Annalen der physik*, 4.

El Moselhy, T. A. and Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850.

Emerick, A. A. and Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15.

Evrard, M., Dumont, G., Hermans, T., Chouteau, M., Francis, O., Pirard, E., and Nguyen, F. (2018). Geophysical investigation of the Pb–Zn deposit of Lontzen–Poppelsberg, Belgium. *Minerals*, 8:233.

Farquharson, C. G. (2008). Constructing piecewise-constant models in multidimensional minimum-structure inversions. *Geophysics*, 73(1):K1–K9.

Ferguson, G. (2007). Heterogeneity and thermal modeling of ground water. *Ground Water*, 45:485–490.

Ferré, T., Bentley, L., Binley, A., Linde, N., Kemna, A., Singha, K., Holliger, K., Huisman, J. A., and Minsley, B. (2009). Critical steps for the continuing advancement of hydrogeophysics. *Eos, Transactions American Geophysical Union*, 90:200.

Ferré, T. P. (2017). Revisiting the relationship between data, models, and decision-making. *Groundwater*, 55:604–614.

Ferré, T. P. (2020). Being Bayesian: Discussions from the perspectives of stakeholders and hydrologists. *Water*, 12:461.

Feynman, R. P., Hibbs, A. R., and Styer, D. F. (2010). *Quantum mechanics and path integrals*. Courier Corporation.

Fiandaca, G., Doetsch, J., Vignoli, G., and Auken, E. (2015). Generalized focusing of time-lapse changes with applications to direct current and time-domain induced polarization inversions. *Geophysical Journal International*, 203:1101–1112.

Franzen, S. E., Farahani, M. A., and Goodwell, A. E. (2020). Information flows: Characterizing precipitation-streamflow dependencies in the Colorado headwaters with an information theory approach. *Water Resources Research*, 56.

Friedman, J. H. (2001). The role of statistics in the data revolution? *International Statistical Review*, 69(1):5–10.

Frigg, R. and Hartmann, S. (2020). Models in Science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition.

Furman, A., Ferré, T. P. A., and Warrick, A. W. (2004). Optimization of ERT surveys for monitoring transient hydrological events using perturbation sensitivity and genetic algorithms. *Vadose Zone Journal*, 3:1230–1239.

Gal, Y. (2016). Uncertainty in deep learning. *PhD thesis, University of Cambridge*.

Garcet, J. P., Ordoñez, A., Roosen, J., and Vanclooster, M. (2006). Metamodelling: Theory, concepts and application to nitrate leaching modelling. *Ecological Modelling*, 193:629–644.

George, A. and Walsh, T. (2022). Can AI invent? *Nature Machine Intelligence*.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Ghysels, G., Anibas, C., Awol, H., Tolche, A., Schneidewind, U., and Huysmans, M. (2021). The significance of vertical and lateral groundwater–surface water exchange fluxes in riverbeds and riverbanks: Comparing 1D analytical flux estimates with 3D groundwater modelling. *Water*, 13:306.

Ghysels, G., Benoit, S., Awol, H., Jensen, E. P., Tolche, A. D., Anibas, C., and Huysmans, M. (2018). Characterization of meter-scale spatial variability of riverbed hydraulic conductivity in a lowland river (Aa river, Belgium). *Journal of Hydrology*, 559:1013–1027.

Ghysels, G., Mutua, S., Veliz, G. B., and Huysmans, M. (2019). A modified approach for modelling river–aquifer interaction of gaining rivers in MODFLOW, including riverbed heterogeneity and river bank seepage. *Hydrogeology Journal*, 27:1851–1863.

Goldscheider, N. (2010). Delineation of spring protection zones.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodwell, A. E., Jiang, P., Ruddell, B. L., and Kumar, P. (2020). Debates—does information theory provide a new paradigm for Earth Science? causality, interaction, and feedback. *Water Resources Research*, 56(2):e2019WR024940. e2019WR024940 10.1029/2019WR024940.

Goodwell, A. E. and Kumar, P. (2017). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resources Research*, 53:5920–5942.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Applied geostatistics series. Oxford University Press, New York.

Gündoğdu, N. Y., Demirci, İ., Demirel, C., and Candansayar, M. E. (2020). Characterization of the bridge pillar foundations using 3d focusing inversion of DC resistivity data. *Journal of Applied Geophysics*, 172:103875.

Gupta, H. V. and Nearing, G. S. (2014). Debates-the future of hydrological sciences: A (common) path forward? using models and data to learn: A systems theoretic perspective on the future of hydrological science. *Water Resources Research*, 50:5351–5359.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34:751–763.

Gupta, H. V., Wagener, T., and Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes*, 22:3802–3813.

Gutin, G. and Punnen, A. P. (2006). *The traveling salesman problem and its variations*, volume 12. Springer Science & Business Media.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2018). Likelihood-free inference via classification. *Statistics and Computing*, 28:411–425.

Haan, C. T. and Allen, D. M. (1972). Comparison of multiple regression and principal component regression for predicting water yields in Kentucky. *Water Resources Research*, 8:1593–1596.

Haan, S., Ramos, F., and Müller, R. D. (2021). Multiobjective Bayesian optimization and joint inversion for active sensor fusion. *Geophysics*, 86:ID1–ID17.

Hall, T., Scheidt, C., Wang, L., Yin, Z., Mukerji, T., and Caers, J. (2022). Sequential value of information for subsurface exploration drilling. *Natural Resources Research*, 31(5):2413–2434.

Harbaugh, A. W. (2005). *MODFLOW-2005, the US Geological Survey modular groundwater model: the ground-water flow process*, volume 6. US Department of the Interior, US Geological Survey Reston, VA, USA.

Härdle, W. K. and Simar, L. (2019). *Applied multivariate statistical analysis*. Springer Nature.

Hare, D. K., Briggs, M. A., Rosenberry, D. O., Boutt, D. F., and Lane, J. W. (2015). A comparison of thermal infrared to fiber-optic distributed temperature sensing for evaluation of groundwater discharge to surface water. *Journal of Hydrology*, 530:153–166.

Hart, C. J. (2005). Classifying, distinguishing and exploring for intrusion-related gold systems. *The Gangue*, 87(1):9.

Hecht-Méndez, J., Molina-Giraldo, N., Blum, P., and Bayer, P. (2010). Evaluating MT3DMS for heat transport simulation of closed geothermal systems. *Groundwater*, 48(5):741–756.

Hempel, C. G. et al. (1965). *Aspects of scientific explanation*, volume 1. Free Press New York.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. (2021a). Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*.

Hermans, T. (2017). Prediction-focused approaches: An opportunity for hydrology. *Groundwater*, 55:683–687.

Hermans, T., Compaire, N., Thibaut, R., and Lesparre, N. (2021b). Bayesian Evidential Learning: An alternative to hydrogeophysical coupled inversion. volume 2021-Septe, pages 3125–3129. Society of Exploration Geophysicists.

Hermans, T., Goderniaux, P., Jougnot, D., Fleckenstein, J., Brunner, P., Nguyen, F., Linde, N., Huisman, J. A., Bour, O., Alvis, J. L., Hoffmann, R., Palacios, A., Cooke, A.-K., Pardo-Alvarez, A., Blazevic, L., Pouladi, B., Haruzi, P., Kenshilikova, M., Davy, P., and Borgne, T. L. (2022). Advancing measurements and representations of subsurface heterogeneity and dynamic processes: towards 4D hydrogeology. *Hydrology and Earth System Sciences Discussions*, 2022:1–55.

Hermans, T. and Irving, J. (2017). Facies discrimination with electrical resistivity tomography using a probabilistic methodology: effect of sensitivity and regularisation. *Near Surface Geophysics*, 15(1):13–25.

Hermans, T., Lesparre, N., De Schepper, G., and Robert, T. (2019). Bayesian Evidential Learning: a field validation using push-pull tests. *Hydrogeology Journal*, 27(5):1661–1672.

Hermans, T., Nguyen, F., and Caers, J. (2015a). Uncertainty in training image-based inversion of hydraulic head data constrained to ERT data: Workflow and case study. *Water Resources Research*, 51(7):5332–5352.

Hermans, T., Nguyen, F., Klepikova, M., Dassargues, A., and Caers, J. (2018). Uncertainty quantification of medium-term heat storage from short-term geophysical experiments using Bayesian Evidential Learning. *Water Resources Research*, 54:2931–2948.

Hermans, T., Nguyen, F., Robert, T., and Revil, A. (2014). Geophysical methods for monitoring temperature changes in shallow low enthalpy geothermal systems. *Energies*, 7:5083–5118.

Hermans, T., Oware, E., and Caers, J. (2016). Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data. *Water Resources Research*, 52:7262–7283.

Hermans, T., Vandenbohede, A., Lebbe, L., and Nguyen, F. (2012). A shallow geothermal experiment in a sandy aquifer monitored using electric resistivity tomography. *GEOPHYSICS*, 77:B11–B21.

Hermans, T., Wildemeersch, S., Jamin, P., Orban, P., Brouyère, S., Dassargues, A., and Nguyen, F. (2015b). Quantitative temperature monitoring of a heat tracing experiment using cross-borehole ERT. *Geothermics*, 53:14–26.

Hjorth, L. and Nabney, I. (1999). Regularisation of mixture density networks. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, volume 2, pages 521–526 vol.2.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Hoffmann, R., Dassargues, A., Goderniaux, P., and Hermans, T. (2019). Heterogeneity and prior uncertainty investigation using a joint heat and solute tracer experiment in alluvial sediments. *Frontiers in Earth Science*, 7.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Hsu, N.-S. and Yeh, W. W.-G. (1989). Optimum experimental design for parameter identification in groundwater hydrology. *Water Resources Research*, 25:1025–1040.

Hu, L. Y., Blanc, G., and Noetinger, B. (2001). Gradual deformation and iterative calibration of sequential stochastic simulations. *Mathematical Geology*, 33:475–489.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., and Maier, H. R. (2016). A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology*, 540:623–640.

Irvine, D. J., Briggs, M. A., Lautz, L. K., Gordon, R. P., McKenzie, J. M., and Cartwright, I. (2017). Using diurnal temperature signals to infer vertical groundwater-surface water exchange. *Groundwater*, 55:10–26.

Irvine, D. J., Cartwright, I., Post, V. E., Simmons, C. T., and Banks, E. W. (2016). Uncertainties in vertical groundwater fluxes from 1-D steady state heat transport analyses caused by heterogeneity, multidimensional flow, and climate change. *Water Resources Research*, 52:813–826.

JafarGandomi, A. and Binley, A. (2013). A Bayesian trans-dimensional approach for the fusion of multiple geophysical datasets. *Journal of Applied Geophysics*, 96:38–54.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Johnson, T., Routh, P. S., Clemo, T., Barrash, W., and Clement, W. P. (2007). Incorporating geostatistical constraints in nonlinear inversion problems. *Water resources research*, 43(10).

Johnson, T., Versteeg, R., Thomle, J., Hammond, G., Chen, X., and Zachara, J. (2015). Four-dimensional electrical conductivity monitoring of stage-driven river water intrusion: Accounting for water table effects using a transient mesh boundary and conditional inversion constraints. *Water Resources Research*, 51:6177–6196.

Jordi, C., Doetsch, J., Günther, T., Schmelzbach, C., and Robertsson, J. O. (2018). Geostatistical regularization operators for geophysical inverse problems on irregular meshes. *Geophysical Journal International*, 213(2):1374–1386.

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., and Bennamoun, M. (2022). Hands-on Bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17:29–48.

Kaipio, J. P., Kolehmainen, V., Vauhkonen, M., and Somersalo, E. (1999). Inverse problems with structural prior information. *Inverse problems*, 15(3):713.

Kammen, D. M. and Sunter, D. A. (2016). City-integrated renewable energy for urban sustainability. *Science*, 352:922–928.

Kanishka, G. and Eldho, T. I. (2017). Watershed classification using isomap technique and hydrometeorological attributes. *Journal of Hydrologic Engineering*, 22(10):04017040.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V. (2019). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31:1544–1554.

Kaune, A., Werner, M., Rodríguez, E., Karimi, P., and De Fraiture, C. (2017). A novel tool to assess available hydrological information and the occurrence of sub-optimal water allocation decisions in large irrigation districts. *Agricultural Water Management*, 191:229–238.

Keating, E. H., Doherty, J., Vrugt, J. A., and Kang, Q. (2010). Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. *Water Resources Research*, 46:2009WR008584.

Kemna, A. (2000). Tomographic inversion of complex resistivity. *Ruhr-Universität Bochum*, page 169.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Khan, M. S. and Coulibaly, P. (2006). Bayesian neural network for rainfall-runoff modeling. *Water Resources Research*, 42.

Kikuchi, C. P. and Ferré, T. P. A. (2017). Analysis of subsurface temperature data to quantify groundwater recharge rates in a closed Altiplano basin, northern Chile. *Hydrogeology Journal*, 25:103–121.

Kikuchi, C. P., Ferré, T. P. A., and Vrugt, J. A. (2015). On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models. *Water Resources Research*, 51:4454–4481.

Kim, B., Joung, I., Cho, A., Shin, D., Han, Y., and Nam, M. (2022). Monitoring the perturbation zone near a foundation excavation with electrical resistivity tomography: Comparison between time-lapse 3D and 2D inversions in single-profile study. *Journal of Applied Geophysics*, 205.

Kim, J.-H., Tsourlos, P., Yi, M.-J., and Karmis, P. (2014). Inversion of ERT data with a priori information using variable weighting factors. *Journal of Applied Geophysics*, 105:1–9.

Kim, J.-H., Yi, M.-J., Park, S.-G., and Kim, J. G. (2009). 4-D inversion of DC resistivity monitoring data acquired over a dynamically changing Earth model. *Journal of Applied Geophysics*, 68(4):522–532.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2014). Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, page 121.

Kingston, G. B., Lambert, M. F., and Maier, H. R. (2005). Bayesian training of artificial neural networks used for water resources modeling. *Water Resources Research*, 41(12).

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, 42.

Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.

Klausner, R. F. (1969). The evaluation of risk in marine capital investments. *The Engineering Economist*, 14(4):183–214.

Klemeš, V. (1986). Dilettantism in hydrology: Transition or destiny? *Water Resources Research*, 22:177S–188S.

Klepikova, M. V., Borgne, T. L., Bour, O., Dentz, M., Hochreutener, R., and Lavenant, N. (2016). Heat as a tracer for understanding transport processes in fractured media: Theory and field assessment from multiscale thermal push-pull tracer tests. *Water Resources Research*, 52:5442–5457.

Klotz, D., Kratzert, F., Gauch, M., Sampson, A. K., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, 26:1673–1693.

Knisel, W. G. (1970). A factor analysis of reservoir losses. *Water Resources Research*, 6:491–498.

Kolmogorov, A. (1956). *Foundations of the Theory of Probability*. AMS Chelsea Publishing Series. Chelsea Publishing Company.

Kourakos, G. and Mantoglou, A. (2009). Pumping optimization of coastal aquifers based on evolutionary algorithms and surrogate modular neural network models. *Advances in Water Resources*, 32:507–521.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22:6005–6022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S. (2019a). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55:11344–11354.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019b). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23:5089–5110.

Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., Yao, Z., and Aspuru-Guzik, A. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Kumar, P. and Gupta, H. V. (2020). Debates—does information theory provide a new paradigm for Earth Science? *Water Resources Research*, 56(2):e2019WR026398. e2019WR026398 2019WR026398.

Kurylyk, B. L., Irvine, D. J., and Bense, V. F. (2019). Theory, tools, and multidisciplinary applications for tracing groundwater fluxes from temperature profiles. *WIREs Water*, 6.

Laloy, E., Linde, N., Jacques, D., and Vrugt, J. A. (2015). Probabilistic inference of multi-Gaussian fields from indirect hydrological data using circulant embedding and dimensionality reduction. *Water Resources Research*, 51:4224–4243.

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*, 49:2664–2682.

Laloy, E. and Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. *Water Resources Research*, 48(1).

Langevin, C. D., Hughes, J. D., Banta, E. R., Niswonger, R. G., Panday, S., and Provost, A. M. (2017). Documentation for the MODFLOW 6 groundwater flow model. Technical report, US Geological Survey.

Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7:3–10.

Last, B. and Kubik, K. (1983). Compact gravity inversion. *Geophysics*, 48(6):713–721.

Lavin, A., Zenil, H., Paige, B., Krakauer, D., Gottschlich, J., Mattson, T., Anandkumar, A., Choudry, S., Rocki, K., Baydin, A. G., Prunkl, C., Paige, B., Isayev, O., Peterson, E., McMahon, P. L., Macke, J., Cranmer, K., Zhang, J., Wainwright, H., Hanuka, A., Veloso, M., Assefa, S., Zheng, S., and Pfeffer, A. (2021). Simulation intelligence: Towards a new generation of scientific methods.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., Bruijn, J. D., Sahu, R. K., Greve, P., Slater, L., and Dadson, S. J. (2022). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 26:3079–3101.

Lesparre, N., Compaire, N., Hermans, T., and Thibaut, R. (2022). 4D temperature monitoring.

Lesparre, N., Robert, T., Nguyen, F., Boyle, A., and Hermans, T. (2019). 4D electrical resistivity tomography (ERT) for aquifer thermal energy storage monitoring. *Geothermics*, 77:368–382.

Leube, P. C., Geiges, A., and Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, 48.

Ley-Cooper, A. Y., Viezzoli, A., Guillemoteau, J., Vignoli, G., Macnae, J., Cox, L., and Munday, T. (2015). Airborne electromagnetic modelling options and their consequences in target definition. *Exploration Geophysics*, 46(1):74–84.

Li, D., Marshall, L., Liang, Z., Sharma, A., and Zhou, Y. (2021). Bayesian LSTM with stochastic variational inference for estimating model uncertainty in process-based hydrological models. *Water Resources Research*, 57.

lin Hsu, K., Gupta, H. V., and Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, 31:2517–2530.

Linde, N., Ginsbourger, D., Irving, J., Nobile, F., and Doucet, A. (2017). On uncertainty quantification in hydrogeology and hydrogeophysics. *Advances in Water Resources*, 110:166–181.

Linde, N., Renard, P., Mukerji, T., and Caers, J. (2015). Geological realism in hydrogeological and geophysical inverse modeling: A review. *Advances in Water Resources*, 86:86–101.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27:986–1005.

Lindley, D. V. (1972). *Bayesian Statistics*. Society for Industrial and Applied Mathematics.

Liu, H., Cai, J., and Ong, Y.-S. (2018). Remarks on multi-output Gaussian process regression. *Knowledge-Based Systems*, 144:102–121.

Liu, H., Yang, J., Ye, M., James, S. C., Tang, Z., Dong, J., and Xing, T. (2021). Using t-distributed stochastic neighbor embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data. *Journal of Hydrology*, 597:126146.

Liu, T.-Y. and Hu, M.-C. (2019). Environmental data analysis using isomap approach. In *Geophysical Research Abstracts*, volume 21.

Liu, X., Lee, J., Kitanidis, P. K., Parker, J., and Kim, U. (2012). Value of information as a context-specific measure of uncertainty in groundwater remediation. *Water Resources Management*, 26:1513–1535.

Loke, M., Chambers, J., Rucker, D., Kuras, O., and Wilkinson, P. (2013). Recent developments in the direct-current geoelectrical imaging method. *Journal of Applied Geophysics*, 95:135–156.

Loke, M. H., Acworth, I., and Dahlin, T. (2003). A comparison of smooth and blocky inversion methods in 2D electrical imaging surveys. *Exploration geophysics*, 34(3):182–187.

Lopez-Alvis, J., Hermans, T., and Nguyen, F. (2019). A cross-validation framework to extract data features for reducing structural uncertainty in subsurface heterogeneity. *Advances in Water Resources*, 133:103427.

Lopez-Alvis, J., Laloy, E., Nguyen, F., and Hermans, T. (2021). Deep generative models in inversion: The impact of the generator's nonlinearity and development of a new approach based on a variational autoencoder. *Computers & Geosciences*, 152:104762.

Lopez-Alvis, J., Nguyen, F., Looms, M. C., and Hermans, T. (2022). Geophysical inversion using a variational autoencoder to model an assembled spatial prior uncertainty. *Journal of Geophysical Research: Solid Earth*, 127(3).

Lu, C., Zhang, C., Hunag, H., and Johnson, T. C. (2015). Monitoring CO2 sequestration into deep saline aquifer and associated salt intrusion using coupled multiphase flow modeling and time-lapse electrical resistivity tomography. *Greenhouse Gases: Science and Technology*, 5:34–49.

Lu, D., Ye, M., Neuman, S. P., and Xue, L. (2012). Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs. *Advances in Water Resources*, 35:69–82.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. (2021). Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR.

Lykkegaard, M. B. and Dodwell, T. J. (2022). Where to drill next? A dual-weighted approach to adaptive optimal design of groundwater surveys. *Advances in Water Resources*, 164:104219.

MacDonald, A. M., Bonsor, H. C., Ahmed, K. M., Burgess, W. G., Basharat, M., Calow, R. C., Dixit, A., Foster, S. S. D., Gopal, K., Lapworth, D. J., Lark, R. M., Moench, M., Mukherjee, A., Rao, M. S., Shamsudduha, M., Smith, L., Taylor, R. G., Tucker, J., van Steenbergen, F., and Yadav, S. K. (2016). Groundwater quality and depletion in the Indo-Gangetic basin mapped from in situ observations. *Nature Geoscience*, 9:762–766.

Macfarlane, A., Förster, A., Merriam, D., Schrötter, J., and Healey, J. (2002). Monitoring artificially stimulated fluid movement in the Cretaceous Dakota aquifer, western Kansas. *Hydrogeology Journal*, 10:662–673.

MacKay, D. J. C. (1992). The evidence framework applied to classification networks. *Neural Computation*, 4:720–736.

Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, 25:891–909.

Makansi, O., Ilg, E., Cicek, O., and Brox, T. (2019). Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7153.

Mamer, E. A. and Lowry, C. S. (2013). Locating and quantifying spatially distributed groundwater/surface water interactions using temperature signals with paired fiber-optic cables. *Water Resources Research*, 49:7670–7680.

Mao, D., Revil, A., and Hinton, J. (2016). Induced polarization response of porous media with metallic particles—part 4: Detection of metallic and nonmetallic targets in time-domain-induced polarization tomography. *Geophysics*, 81(4):D359–D375.

Mariethoz, G. and Caers, J. (2014). *Multiple-point geostatistics: stochastic modeling with training images*. John Wiley & Sons.

Martinez, G. F. and Gupta, H. V. (2011). Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. *Water Resources Research*, 47.

Masi, M., Ferdos, F., Losito, G., and Solari, L. (2020). Monitoring of internal erosion processes by time-lapse electrical resistivity tomography. *Journal of Hydrology*, 589.

Matalas, N. C. and Reiher, B. J. (1967). Some comments on the use of factor analyses. *Water Resources Research*, 3:213–223.

Mazher, A. (2020). Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques. *Water*, 12:590.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction.

Meinert, N., Gawlikowski, J., and Lavin, A. (2022). The unreasonable effectiveness of deep evidential regression.

Meloun, M. and Militký, J., editors (2012). *Statistical data analysis: a practical guide*. WPI, Woodhead Publ. India Pvt. Ltd, New Delhi, reprinted edition. OCLC: 696087077.

Meng, T., Jing, X., Yan, Z., and Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information Fusion*, 57:115–129.

Michel, H., Hermans, T., and Nguyen, F. (2022a). Iterative prior resampling and rejection sampling to improve 1D geophysical imaging based on Bayesian Evidential Learning (BEL1D). *Geophysical Journal International*.

Michel, H., Nguyen, F., and Aigner, L. (2022b). hadrienmichel/pyBEL1D: Latest version of pyBEL1D.

Michel, H., Nguyen, F., and Hermans, T. (2020a). Improving Bayesian Evidential Learning 1D imaging (BEL1D) accuracy through iterative prior resampling. American Geophysical Union.

Michel, H., Nguyen, F., Kremer, T., Elen, A., and Hermans, T. (2020b). 1D geological imaging of the subsurface from geophysical data with Bayesian Evidential Learning. *Computers & Geosciences*, 138:104456.

Middleton, M. A., Whitfield, P. H., and Allen, D. M. (2015). Independent component analysis of local-scale temporal variability in sediment-water interface temperature. *Water Resources Research*, 51:9679–9695.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J. (2008). Stationarity is dead: Whither water management? *Science*, 319:573–574.

Moghaddam, M. A., Ferré, T. P. A., Chen, X., Chen, K., and Ehsani, M. R. (2022). Application of machine learning methods in inferring surface water groundwater exchanges using high temporal resolution temperature measurements. *CoRR*, abs/2201.00726.

Mohammed, G. A. (2009). Groundwater-surface water interaction along a lowland river. *Ph.D. Thesis, Department of Hydrology and Hydraulic Engineering (HYDR), Vrije Universiteit Brussel (VUB), Brussels, Belgium.*

Molina Giraldo, N., Hecht-Méndez, J., Blum, P., and Bayer, P. (2009). Use of MT3DMS for heat transport simulation of shallow geothermal systems. *AGU Fall Meeting Abstracts*.

Montgomery, D. (2019). *Design and Analysis of Experiments*. John Wiley & Sons, Limited.

Müller, W. G. (2007). *Collecting spatial data: optimum design of experiments for random fields*. Springer Science & Business Media.

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8:281–293.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press.

Mutua, S. M. (2013). Analysing the influence of groundwater-surface water interaction on the groundwater balance in the Aa river. *Master's Thesis, Vrije Universiteit Brussel, Brussels, Belgium*.

Nazifi, H. M., Gülen, L., Gürbüz, E., and Pekşen, E. (2022). Time-lapse electrical resistivity tomography (ERT) monitoring of used engine oil contamination in laboratory setting. *Journal of Applied Geophysics*, 197.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*, volume 118. Springer New York.

Nearing, G., Yatheendradas, S., Crow, W., Zhan, X., Liu, J., and Chen, F. (2018a). The efficiency of data assimilation. *Water Resources Research*, 54:6374–6392.

Nearing, G. S. and Gupta, H. V. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, 51:524–538.

Nearing, G. S., Gupta, H. V., and Crow, W. T. (2013a). Information loss in approximately Bayesian estimation techniques: A comparison of generative and discriminative approaches to estimating agricultural productivity. *Journal of Hydrology*, 507:163–173.

Nearing, G. S., Gupta, H. V., Crow, W. T., and Gong, W. (2013b). An approach to quantifying the efficiency of a Bayesian filter. *Water Resources Research*, 49:2164–2173.

Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., and Nevo, S. (2022). Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. *Hydrology and Earth System Sciences*, 26:5493–5513.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57.

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., and Gupta, H. V. (2020). Does information theory provide a new paradigm for Earth Science? hypothesis testing. *Water Resources Research*, 56(2):e2019WR024918. e2019WR024918 2019WR024918.

Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., and Peters-Lidard, C. (2018b). Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology*, 19:1835–1852.

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., and Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, 61:1666–1678.

Nenna, V. and Knight, R. (2014). Demonstration of a value of information metric to assess the use of geophysical data for a groundwater application. *GEOPHYSICS*, 79:E51–E60.

Neuman, S. P., Xue, L., Ye, M., and Lu, D. (2012a). Bayesian analysis of data-worth considering model and parameter uncertainties. *Advances in Water Resources*, 36:75–85.

Neuman, S. P., Xue, L., Ye, M., and Lu, D. (2012b). Bayesian analysis of data-worth considering model and parameter uncertainties. *Advances in Water Resources*, 36:75–85.

Newton, I. (1833). *Philosophiae naturalis principia mathematica*, volume 1. G. Brookman.

Nguyen, F., Kemna, A., Robert, T., and Hermans, T. (2016). Data-driven selection of the minimum-gradient support parameter in time-lapse focused electric imaging. *Geophysics*, 81:A1–A5.

Ogie, R. I., Shukla, N., Sedlar, F., and Holderness, T. (2017). Optimal placement of water-level sensors to facilitate data-driven management of hydrological infrastructure assets in coastal mega-cities of developing nations. *Sustainable cities and society*, 35:385–395.

Oldenborger, G. A., Routh, P. S., and Knoll, M. D. (2007). Model reliability for 3D electrical resistivity tomography: Application of the volume of investigation index to a time-lapse monitoring experiment. *Geophysics*, 72(4):F167–F175.

Oldenburg, D. W. and Li, Y. (1994). Subspace linear inverse method. *Inverse Problems*, 10:915–935.

Oldenburg, D. W. and Li, Y. (1999). Estimating depth of investigation in dc resistivity and IP surveys. *GEOPHYSICS*, 64:403–416.

Orozco, A. F., Kemna, A., Oberdörster, C., Zschornack, L., Leven, C., Dietrich, P., and Weiss, H. (2012). Delineation of subsurface hydrocarbon contamination at a former hydrogenation plant using spectral induced polarization imaging. *Journal of contaminant hydrology*, 136:131–144.

Osher, S. and Fedkiw, R. (2003). *Level Set Methods and Dynamic Implicit Surfaces*. Springer New York.

Ouarda, T. B., Girard, C., Cavadias, G. S., and Bobée, B. (2001). Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*, 254:157–173.

Palacios, A., Ledo, J. J., Linde, N., Luquot, L., Bellmunt, F., Folch, A., Marcuello, A., Queralt, P., Pezard, P. A., Martínez, L., Bosch, D., and Carrera, J. (2020). Time-lapse cross-hole electrical resistivity tomography (CHERT) for monitoring seawater intrusion dynamics in a Mediterranean aquifer. *Hydrology and Earth System Sciences*, 24:2121–2139.

Palmer, C. D., Blowes, D. W., Frind, E. O., and Molson, J. W. (1992). Thermal energy storage in an unconfined aquifer: 1. field injection experiment. *Water Resources Research*, 28:2845–2856.

Papamakarios, G. (2019). *Neural density estimation and likelihood-free inference*. Phd thesis edition.

Park, B.-H., Bae, G.-O., and Lee, K.-K. (2015). Importance of thermal dispersivity in designing groundwater heat pump ({GWHP}) system: {Field} and numerical study. *Renewable Energy*, 83:270–279.

Park, J. and Caers, J. (2020). Direct forecasting of global and spatial model parameters from dynamic data. *Computers & Geosciences*, 143:104567.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2:559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perdigão, R. A., Ehret, U., Knuth, K. H., and Wang, J. (2020). Debates: Does information theory provide a new paradigm for Earth Science? emerging concepts and pathways of information physics. *Water Resources Research*, 56(2):e2019WR025270. e2019WR025270 2019WR025270.

Perron, L. and Furnon, V. (2019). OR-tools. *Google.[Online]. Available: https://developers. google. com/optimization*.

Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E. C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T. E., and Woods, R. (2017). Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences*, 21:3701–3713.

Pezard, P. A., Denchik, N., Lofi, J., Perroud, H., Henry, G., Neyens, D., Luquot, L., and Levannier, A. (2016). Time-lapse downhole electrical resistivity monitoring of subsurface CO2 storage at the Maguelone shallow experimental site (Languedoc, France). *International Journal of Greenhouse Gas Control*, 48:142–154.

Pham, H. and Tsai, F. (2017). Groundwater modeling. chapter 48. *Handbook of applied hydrology. McGraw-Hill, New York*, pages 48–1.

Pham, H. V. and Tsai, F. T.-C. (2016). Optimal observation network design for conceptual model discrimination and uncertainty reduction: Observation network design for model discrimination. *Water Resources Research*, 52(2):1245–1264.

Pollock, D. (2017). MODPATH v7. 2.01—a particle-tracking model for MODFLOW: US Geological Survey release, 15 December 2017.

Polydorides, N. and Lionheart, W. R. B. (2002). A Matlab toolkit for three-dimensional electrical impedance tomography: a contribution to the electrical impedance and diffuse optical reconstruction software project. *Measurement Science and Technology*, 13:1871–1883.

Popper, K. (2005). *The Logic of Scientific Discovery*. Routledge Classics. Taylor & Francis.

Portniaguine, O. and Zhdanov, M. S. (1999). Focusing geophysical inversion images. *Geophysics*, 64(3):874–887.

Power, C., Gerhard, J., Karaoulis, M., Tsourlos, P., and Giannopoulos, A. (2014). Evaluating four-dimensional time-lapse electrical resistivity tomography for monitoring DNAPL source zone remediation. *Journal of Contaminant Hydrology*, 162-163:27–46.

Pradhan, A. and Mukerji, T. (2020). Seismic Bayesian Evidential Learning: estimation and uncertainty quantification of sub-resolution reservoir properties. *Computational Geosciences*, 24:1121–1140.

Prakaisak, I. and Wongchaisuwat, P. (2022). Hydrological time series clustering: A case study of telemetry stations in Thailand. *Water*, 14:2095.

Prechelt, L. (1998). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Qiang, S., Shi, X., Kang, X., and Revil, A. (2022). Optimized arrays for electrical resistivity tomography survey using Bayesian experimental design. *GEOPHYSICS*, 87:E189–E203.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174.

Rahman, M., Frame, J. M., Lin, J., and Nearing, G. S. (2022). Hydrology research articles are becoming more topically diverse. *Journal of Hydrology*, 614:128551.

Ramgraber, M., Weatherl, R., Blumensaat, F., and Schirmer, M. (2021). Non-Gaussian parameter inference for hydrogeological models using Stein variational gradient descent. *Water Resources Research*, 57(4).

Ramos, M.-H., Mathevet, T., Thielen, J., and Pappenberger, F. (2010). Communicating uncertainty in hydro-meteorological forecasts: mission impossible? *Meteorological Applications*, 17:223–235.

Rau, G. C., Andersen, M. S., McCallum, A. M., Roshan, H., and Acworth, R. I. (2014). Heat as a tracer to quantify water flow in near-surface sediments. *Earth-Science Reviews*, 129:40–58.

Razavi, S., Tolson, B. A., and Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, 48.

Regis, R. G. and Shoemaker, C. A. (2005). Constrained global optimization of expensive black box functions using radial basis functions. *Journal of Global Optimization*, 31:153–171.

Remy, N., Boucher, A., and Wu, J. (2009). *Applied geostatistics with SGeMS: A user's guide*. Cambridge University Press.

Renard, P. (2007). Stochastic hydrogeology: What professionals really need? *Ground Water*, 45:531–541.

Ribeiro-Corréa, J., Cavadias, G., Clément, B., and Rousselle, J. (1995). Identification of hydrological neighborhoods using canonical correlation analysis. *Journal of Hydrology*, 173:71–89.

Rice, R. M. (1972). Using canonical correlation for hydrological predictions. *Hydrological Sciences Bulletin*, 17:315–321.

Robert, Paulus, Bolly, Lin, K. S., and Hermans (2019). Heat as a proxy to image dynamic processes with 4D electrical resistivity tomography. *Geosciences*, 9:414.

Robert, T., Dassargues, A., Brouyère, S., Kaufmann, O., Hallet, V., and Nguyen, F. (2011). Assessing the contribution of electrical resistivity tomography (ERT) and self-potential (SP) methods for a water well drilling program in fractured/karstified limestones. *Journal of Applied Geophysics*, 75:42–53.

Robinson, J., Johnson, T., and Rockhold, M. (2020). Feasibility assessment of long-term electrical resistivity monitoring of a nitrate plume. *Groundwater*, 58:224–237.

Rosenblueth, A. and Wiener, N. (1945). The role of models in Science. *Philosophy of Science*, 12(4):316–321.

Rothfuss, J., Ferreira, F., Walther, S., and Ulrich, M. (2019). Conditional density estimation with neural networks: Best practices and benchmarks.

Ruddell, B. L., Brunsell, N. A., and Stoy, P. (2013). Applying information theory in the geosciences to quantify process uncertainty, feedback, scale. *Eos, Transactions American Geophysical Union*, 94:56–56.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 84:128–154.

Saad, M. and Turgeon, A. (1988). Application of principal component analysis to long-term reservoir management. *Water Resources Research*, 24:907–912.

Samadi, S., Pourreza-Bilondi, M., Wilson, C. A. M. E., and Hitchcock, D. B. (2020). Bayesian model averaging with fixed and flexible priors: Theory, concepts, and calibration experiments for rainfall-runoff modeling. *Journal of Advances in Modeling Earth Systems*, 12.

Saner, D., Juraske, R., Kübert, M., Blum, P., Hellweg, S., and Bayer, P. (2010). Is it only CO2 that matters? A life cycle perspective on shallow geothermal systems. *Renewable and Sustainable Energy Reviews*, 14:1798–1813.

Sarker, M. M. R., Hermans, T., Camp, M. V., Hossain, D., Islam, M., Ahmed, N., Bhuiyan, M. A. Q., Karim, M. M., and Walraevens, K. (2022). Identifying the major hydrogeochemical factors governing groundwater chemistry in the coastal aquifers of Southwest Bangladesh using statistical analysis. *Hydrology*, 9(2):20.

Satija, A. and Caers, J. (2015). Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. *Advances in Water Resources*, 77:69–81.

Saunders, J., Herwanger, J., Pain, C., Worthington, M., and De Oliveira, C. (2005). Constrained resistivity inversion using seismic data. *Geophysical Journal International*, 160(3):785–796.

Scheidt, C., Li, L., and Caers, J. (2018). *Quantifying Uncertainty in Subsurface Systems*. Geophysical Monograph Series. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Scheidt, C., Renard, P., and Caers, J. (2015). Prediction-focused subsurface modeling: Investigating the need for accuracy in flow-based inverse modeling. *Mathematical Geosciences*, 47:173–191.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Schmidt-Hattenberger, C., Bergmann, P., Labitzke, T., Wagner, F., and Rippe, D. (2016). Permanent crosshole electrical resistivity tomography (ERT) as an established method for the long-term CO2 monitoring at the Ketzin pilot site. *International Journal of Greenhouse Gas Control*, 52:432–448.

Schölkopf, B., Smola, A., and Müller, K.-R. (1997). Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer.

Schübl, M., Stumpp, C., and Brunetti, G. (2022). A Bayesian perspective on the information content of soil water measurements for the hydrological characterization of the vadose zone. *Journal of Hydrology*, 613:128429.

Scott, D. W. (1992). *Multivariate Density Estimation*. Wiley.

Sellars, S. L. (2018). "Grand Challenges" in Big Data and the Earth Sciences. *Bulletin of the American Meteorological Society*, 99(6):ES95–ES98.

Sethian, J. A. (1996). A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Sharma, M., Farquhar, S., Nalisnick, E., and Rainforth, T. (2022). Do Bayesian neural networks need to be fully stochastic?

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, 54:8558–8593.

Shen, K., Qin, H., Zhou, J., and Liu, G. (2022). Runoff probability prediction model based on natural gradient boosting with tree-structured parzen estimator optimization. *Water*, 14:545.

Shockley, E. M., Vrugt, J. A., and Lopez, C. F. (2017). PyDREAM: high-dimensional parameter inference for biological models in Python. *Bioinformatics*, 34(4):695–697.

Shook, G. M. (2001). Predicting thermal breakthrough in heterogeneous media from tracer tests. *Geothermics*, 30(6):573–589.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Springer New York, NY.

Singh, V. P. (2018). Hydrologic modeling: progress and future directions. *Geoscience Letters*, 5:15.

Singha, K., Day-Lewis, F. D., Johnson, T., and Slater, L. D. (2015). Advances in interpretation of subsurface processes with time-lapse electrical imaging. *Hydrological Processes*, 29:1549–1576.

Smucker, B., Krzywinski, M., and Altman, N. (2018). Optimal experimental design. *Nature Methods*, 15:559–560.

Solomatine, D. P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, 10:3–22.

Sommer, W., Doornenbal, P. J., Drijver, B. C., van Gaans, P. F. M., Leusbrock, I., Grotenhuis, J. T. C., and Rijnaarts, H. H. M. (2014). Thermal performance and heat transport in aquifer thermal energy storage. *Hydrogeology Journal*, 22:263–279.

Sommer, W., Valstar, J., van Gaans, P., Grotenhuis, T., and Rijnaarts, H. (2013). The impact of aquifer heterogeneity on the performance of aquifer thermal energy storage. *Water Resources Research*, 49:8128–8138.

Spantini, A., Baptista, R., and Marzouk, Y. (2022). Coupling techniques for nonlinear ensemble filtering. *SIAM Review, in press*.

Spantini, A., Bigoni, D., and Marzouk, Y. (2018). Inference via low-dimensional couplings. *arXiv*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Stallman, R. (1963). Methods of collecting and interpreting ground-water data. *US Geological Survey Water Supply Paper*, pages 36–46.

Stallman, R. W. (1965a). Effects of water table conditions on water level changes near pumping wells. *Water Resources Research*, 1:295–312.

Stallman, R. W. (1965b). Steady one-dimensional fluid flow in a semi-infinite porous medium with sinusoidal surface temperature. *Journal of Geophysical Research*, 70:2821–2827.

Stein, L., Mukkavilli, S. K., and Wagener, T. (2022). Lifelines for a drowning science - improving findability and synthesis of hydrologic publications. *Hydrological Processes*, 36.

Sun, A. Y. (2013). Predicting groundwater level changes using GRACE data. *Water Resources Research*, 49:5900–5912.

Sun, A. Y., Jiang, P., Mudunuru, M. K., and Chen, X. (2021). Explore spatio-temporal learning of large sample hydrology using graph neural networks. *Water Resources Research*, 57.

Sun, A. Y. and Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters*, 14:073001.

Sun, A. Y., Wang, D., and Xu, X. (2014). Monthly streamflow forecasting using Gaussian process regression. *Journal of Hydrology*, 511:72–81.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W. H., Ma, X., Rao, Y., Bednar, J. A., Tan, A., Wang, J., Purushotham, S., Gill, T. E., Chastang, J., Howard, D., Holt, B., Gangodagamage, C., Zhao, P., Rivas, P., Chester, Z., Orduz, J., and John, A. (2022). A review of Earth artificial intelligence. *Computers & Geosciences*, 159:105034.

Suzuki, S. (1960). Percolation measurements based on heat flow through soil with special reference to paddy fields. *Journal of Geophysical Research*, 65:2883–2885.

Tadjer, A. and Bratvold, R. B. (2021). Managing uncertainty in geological CO2 storage using Bayesian Evidential Learning. *Energies*, 14.

Tang, W. and Carey, S. K. (2022). Classifying annual daily hydrographs in Western North America using t-distributed stochastic neighbour embedding. *Hydrological Processes*, 36.

Tarakanov, A. and Elsheikh, A. H. (2020). Optimal Bayesian experimental design for subsurface flow problems. *Computer Methods in Applied Mechanics and Engineering*, 370:113208.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics.

Tarantola, A. (2006). Popper, Bayes and the inverse problem. *Nature Physics*, 2:492–494.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

Thibaut, R. (2021). WHPA prediction.

Thibaut, R., Compaire, N., Lesparre, N., Ramgraber, M., Laloy, E., and Hermans, T. (2022). Comparing well and geophysical data for temperature monitoring within a Bayesian experimental design framework. *Water Resources Research*, 58.

Thibaut, R., Kremer, T., Royen, A., Ngun, B. K., Nguyen, F., and Hermans, T. (2021a). A new workflow to incorporate prior information in minimum gradient support (MGS) inversion of electrical resistivity and induced polarization data. *Journal of Applied Geophysics*, 187:104286.

Thibaut, R., Laloy, E., and Hermans, T. (2021b). A new framework for experimental design using Bayesian Evidential Learning: The case of wellhead protection area. *Journal of Hydrology*, 603:126903.

Thibaut, R. and Ramgraber, M. (2021). SKBEL - Bayesian Evidential Learning framework built on top of scikit-learn.

Thibaut, R. and Vandekerckhove, G. (2021). pysgems–use SGeMS (Stanford Geostatistical Modeling Software) within Python.

Thiemann, M., Trosset, M., Gupta, H., and Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research*, 37:2521–2535.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

Torranin, P. (1972). *Applicability of canonical correlation in hydrology*. PhD thesis, Colorado State University. Libraries.

Trainor-Guitton, W. J. (2014). A geophysical perspective of value of information: examples of spatial decisions for groundwater sustainability. *Environment Systems and Decisions*, 34:124–133.

Trainor-Guitton, W. J., Ramirez, A., Yang, X., Mansoor, K., Sun, Y., and Carroll, S. (2013). Value of information methodology for assessing the ability of electrical resistivity to detect CO2/brine leakage into a shallow aquifer. *International Journal of Greenhouse Gas Control*, 18:101–113.

Tripathi, S. and Govindaraju, R. S. (2008). Engaging uncertainty in hydrologic data sets using principal component analysis: BaNPCA algorithm. *Water Resources Research*, 44.

Tsai, F. T.-C. and Li, X. (2008). Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. *Water Resources Research*, 44.

Tso, C.-H., Johnson, T., Song, X., Chen, X., Kuras, O., Wilkinson, P., Uhlemann, S., Chambers, J., and Binley, A. (2020). Integrated hydrogeophysical modelling and data assimilation for geoelectrical leak detection. *Journal of Contaminant Hydrology*, 234.

Uhlemann, S., Wilkinson, P. B., Maurer, H., Wagner, F. M., Johnson, T. C., and Chambers, J. E. (2018). Optimized survey design for electrical resistivity tomography: combined optimization of measurement configuration and electrode placement. *Geophysical Journal International*, 214:108–121.

Van Hoorde, M., Hermans, T., Dumont, G., and Nguyen, F. (2017). 3D electrical resistivity tomography of karstified formations using cross-line measurements. *Engineering Geology*, 220:123–132.

Vandenbohede, A., Hermans, T., Nguyen, F., and Lebbe, L. (2011). Shallow heat injection and storage experiment: Heat transport simulation and sensitivity analysis. *Journal of Hydrology*, 409:262–272.

Vandenbohede, A., Louwyck, A., and Lebbe, L. (2009). Conservative solute versus heat transport in porous media during push-pull tests. *Transport in Porous Media*, 76:265–287.

Vanhoudt, D., Desmedt, J., Bael, J. V., Robeyn, N., and Hoes, H. (2011). An aquifer thermal storage system in a Belgian hospital: Long-term experimental evaluation of energy and cost savings. *Energy and Buildings*, 43:3657–3665.

Vaze, J., Chiew, F., Hughes, D., and Andréassian, V. (2015). Preface: HS02 – hydrologic non-stationarity and extrapolating models to predict the future. *Proceedings of the International Association of Hydrological Sciences*, 371:1–2.

Vignoli, G., Fiandaca, G., Christiansen, A. V., Kirkegaard, C., and Auken, E. (2015). Sharp spatially constrained inversion with applications to transient electromagnetic data. *Geophysical Prospecting*, 63:243–255.

Vilhelmsen, T. N. and Ferré, T. P. (2018). Extending data worth analyses to select multiple observations targeting multiple forecasts. *Groundwater*, 56:399–412.

Villani, C. (2009). *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, 75:273–316.

Vrugt, J. A. and Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, 43.

Wagner, F. M., Günther, T., Schmidt-Hattenberger, C., and Maurer, H. (2015). Constructive optimization of electrode locations for target-focused resistivity monitoring. *GEOPHYSICS*, 80:E29–E40.

Wagner, V., Li, T., Bayer, P., Leven, C., Dietrich, P., and Blum, P. (2014). Thermal tracer testing in a sedimentary aquifer: field experiment (Lauswiesen, Germany) and numerical simulation. *Hydrogeology Journal*, 22:175–187.

Wallis, J. R. (1965). Multivariate statistical methods in hydrology-a comparison using data of known functional relationship. *Water Resources Research*, 1:447–461.

Wallis, J. R. (1968). Factor analysis in hydrology-an agnostic view. *Water Resources Research*, 4:521–527.

Wang, G.-J., Cheng, C., Ma, Y.-Z., and Xia, J.-Q. (2022a). Likelihood-free inference with the mixture density network. *The Astrophysical Journal Supplement Series*, 262(1):24.

Wang, L., Kitanidis, P. K., and Caers, J. (2022b). Hierarchical Bayesian inversion of global variables and large-scale spatial fields. *Water Resources Research*, 58.

Wang, L.-L. and Huber, A. L. (1967). Estimating water yields in utah by principal component analysis. *Report*.

Wang, Y., Zechner, M., Mern, J. M., Kochenderfer, M. J., and Caers, J. K. (2022c). A sequential decision-making framework with uncertainty quantification for groundwater management. *Advances in Water Resources*, 166:104266.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Weijs, S. V. and Ruddell, B. L. (2020). Debates: Does information theory provide a new paradigm for Earth Science? sharper predictions using Occam's digital razor. *Water Resources Research*, 56(2):e2019WR026471. e2019WR026471 10.1029/2019WR026471.

Weijs, S. V., Schoups, G., and van de Giesen, N. (2010). Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, 14:2545–2558.

Westra, S., Brown, C., Lall, U., and Sharma, A. (2007). Modeling multivariable hydrological series: Principal component analysis or independent component analysis? *Water Resources Research*, 43.

Westra, S., Sharma, A., Brown, C., and Lall, U. (2008). Multivariate streamflow forecasting using independent component analysis. *Water Resources Research*, 44.

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environmental Modelling & Software*, 109:191–201.

Whitten, E. H. T. (2018). Forward and inverse models over 70 years.

Wildemeersch, S., Jamin, P., Orban, P., Hermans, T., Klepikova, M., Nguyen, F., Brouyère, S., and Dassargues, A. (2014). Coupling heat and chemical tracer experiments for estimating heat transfer parameters in shallow alluvial aquifers. *Journal of Contaminant Hydrology*, 169:90–99.

Wilkinson, P. B., Uhlemann, S., Meldrum, P. I., Chambers, J. E., Carrière, S., Oxby, L. S., and Loke, M. (2015). Adaptive time-lapse optimized survey design for electrical resistivity tomography monitoring. *Geophysical Journal International*, 203:755–766.

Wilson, S., Close, M., and Abraham, P. (2018). Applying linear discriminant analysis to predict groundwater redox conditions conducive to denitrification. *Journal of Hydrology*, 556:611–624.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

Wood, E. F. and Rodríguez-Iturbe, I. (1975). Bayesian inference and decision making for extreme hydrologic events. *Water Resources Research*, 11:533–542.

Wöhling, T. and Vrugt, J. A. (2008). Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models. *Water Resources Research*, 44.

Yan, S. and Minsker, B. (2006). Optimal groundwater remediation design using an adaptive neural network genetic algorithm. *Water Resources Research*, 42.

Yang, J., Jakeman, A., Fang, G., and Chen, X. (2018). Uncertainty analysis of a semi-distributed hydrologic model based on a Gaussian process emulator. *Environmental Modelling & Software*, 101:289–300.

Ye, N., Somani, A., Hsu, D., and Lee, W. S. (2017). DESPOT: Online POMDP planning with regularization. *Journal of Artificial Intelligence Research*, 58:231–266.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Yin, Z., Strebelle, S., and Caers, J. (2020). Automated Monte Carlo-based quantification and updating of geological uncertainty with borehole data (AutoBEL v1.0). *Geoscientific Model Development*, 13:651–672.

Zealand, C. M., Burn, D. H., and Simonovic, S. P. (1999). Short term streamflow forecasting using artificial neural networks. *Journal of Hydrology*, 214:32–48.

Zhang, J., Zeng, L., Chen, C., Chen, D., and Wu, L. (2015). Efficient Bayesian experimental design for contaminant source identification. *Water Resources Research*, 51:576–598.

Zhang, J., Zheng, Q., Chen, D., Wu, L., and Zeng, L. (2020). Surrogate-based Bayesian inverse modeling of the hydrological system: An adaptive approach considering surrogate approximation error. *Water Resources Research*, 56.

Zhang, X., Liang, F., Yu, B., and Zong, Z. (2011). Explicitly integrating parameter, input, and structure uncertainties into Bayesian neural networks for probabilistic hydrologic forecasting. *Journal of Hydrology*, 409:696–709.

Zhang, X. and Zhao, K. (2012). Bayesian neural networks for uncertainty analysis of hydrologic modeling: A comparison of two schemes. *Water Resources Management*, 26:2365–2382.

Zhao, C., Yu, P., and Zhang, L. (2016). A new stabilizing functional to enhance the sharp boundary in potential field regularized inversion. *Journal of Applied Geophysics*, 135:356–366.

Zhao, Y., Guo, Q., Lu, C., and Luo, J. (2022). High-dimensional groundwater flow inverse modeling by upscaled effective model on principal components. *Water Resources Research*, 58.

Zhao, Y. and Luo, J. (2020). Reformulation of Bayesian geostatistical approach on principal components. *Water Resources Research*, 56.

Zhdanov, M. and Tolstaya, E. (2004). Minimum support nonlinear parametrization in the solution of a 3D magnetotelluric inverse problem. *Inverse Problems*, 20:937–952.

Zhdanov, M. S. (2015). *Inverse Theory and Applications in Geophysics*. Elsevier.

Zheng, C. (2009). Recent developments and future directions for MT3DMS and related transport codes. *Ground Water*, 47:620–625.

Zhou, H., Gómez-Hernández, J. J., and Li, L. (2014). Inverse methods in hydrogeology: Evolution and recent trends. *Advances in Water Resources*, 63:22–37.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Zong, Y., Valocchi, A. J., and Lin, Y. F. (2021). Coupling a borehole thermal model and MT3DMS to simulate dynamic ground source heat pump efficiency. *Groundwater*.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

I'll be back

Arnold Schwarzenegger