Technical Report
TR-22-04
April 2024

White paper essays on model validation

Task 10 of SKB Task Force GWFTS – Validation approaches for groundwater flow and transport modelling with discrete features

G William Lanyon
Philippe Davy
William S Dershowitz
Stefan Finsterle
Björn Gylling
Jeffrey Hyman
Ivars Neretnieks
Masahiro Uchida

SVENSK KÄRNBRÄNSLEHANTERING AB

SWEDISH NUCLEAR FUEL
AND WASTE MANAGEMENT CO

Box 3091, SE-169 03 Solna
Phone +46 8 459 84 00
skb.se

SVENSK KÄRNBRÄNSLEHANTERING

ISSN 1404-0344
SKB TR-22-04
ID 2011434
April 2024

White paper essays on model validation

Task 10 of SKB Task Force GWFTS – Validation approaches for groundwater flow and transport modelling with discrete features

G William Lanyon, Fracture Systems Ltd

Philippe Davy, University of Rennes

William S Dershowitz, GeoFractal LLC

Stefan Finsterle, Finsterle GeoConsulting LLC

Björn Gylling, Gylling GeoSolutions

Jeffrey Hyman, Los Alamos National Laboratory

Ivars Neretnieks, Royal Institute of Technology, KTH

Masahiro Uchida, Fracture Flow Solutions

Keywords: Model validation, Model evaluation, Groundwater flow, Solute Transport.

This report is published on www.skb.se

Abstract

Safety assessments of radioactive waste repositories rely heavily on results obtained by numerical models that assess the long-term performance of the engineered and natural barrier systems. Given that important engineering and public policy decisions are based on these models, it is essential that we critically evaluate their abilities and limitations, and thus justify the level of confidence we have in the inferences drawn from the modelling. In this report, the authors discuss some of the issues that should be considered in the modeler's attempts to test, corroborate, confirm, and verify numerical models. This process is here referred to as model validation and a pragmatic approach is chosen for this important topic.

This report presents a series of essays concerning the pragmatic validation of numerical models of flow and solute transport in fractured crystalline rock. The different essays were contributed by members of a "White Paper Group" formed within Task 10 of the SKB Task Force on Modelling of Groundwater Flow and Transport of Solutes (GWFTS) to consider how to approach the issue of pragmatic model validation.

The basis of a proposed approach is summarised in Chapter 5. The ideas described are intended to provide a basis for discussion within the Task Force GWFTS and for the further development of pragmatic approaches to model validation.

Sammanfattning

Säkerhetsbedömningar av förvar för radioaktivt avfall och använt kärnbränsle bygger i hög grad på resultat som erhålls från numeriska modeller som kan användas för att bedöma den långsiktiga funktionen hos de tekniska och naturliga barriärsystemen. Med tanke på att för samhället viktiga tekniska och offentliga beslut baseras på dessa modeller är det väsentligt att man kritiskt utvärderar deras möjligheter och begränsningar samt på så sätt motiverar nivån av det förtroende vi har för de slutsatser som kan dras från modelleringen. I den här rapporten diskuterar författarna några av de frågor som bör beaktas när modellören försöker att testa, kontrollera, bekräfta och verifiera numeriska modeller. Denna process kallas här modellvalidering, dvs en förtroendeskapande modellering. Ett pragmatiskt tillvägagångssätt har valts för att uppnå detta.

Denna rapport presenterar en serie essäer angående validering av modeller för flöde och transport i sprickigt kristallint berg. De olika essäerna har skrivits av medlemmar i en ”White Paper Group” bildad inom Task 10 av SKB Task Force för modellering av grundvattenflöde och transport av lösta ämnen för att överväga hur man bäst kan närma sig frågan om modellvalidering.

Grunden för det föreslagna tillvägagångssättet sammanfattas i kapitel 5. De beskrivna idéerna är avsedda att ge underlag för en diskussion inom Task Force GWFTS och för vidareutveckling av pragmatiska angreppssätt inom modellvalidering.

Contents

1 Introduction

Safety assessments of radioactive waste repositories rely heavily on results obtained by numerical models that assess the long-term performance of the engineered and natural barrier systems. Given that important engineering and public policy decisions are based on these models, it is essential that we critically evaluate their abilities and limitations, and thus justify the level of confidence we have in the inferences drawn from the modelling. In this report, we discuss some of the issues that should be considered in the modeler's attempts to test, corroborate, confirm, and verify numerical models. This process is here referred to as model validation and a pragmatic approach is chosen for this important topic (Finsterle and Lanyon 2022). The purpose of this report is to lay the foundations for pragmatic validation of groundwater flow and transport models. Also, it may serve as a basis for discussion and further work within Task 10.

This report presents a series of essays around the theme of validation of and confidence building in models of groundwater flow and transport in crystalline rock. They were developed as part of Task 10 of the SKB Task Force on Groundwater Flow and Transport of Solutes (TF GWFTS; www.skb.se/taskforce) to set out our understanding of how best to validate and build confidence in our models. The original intention had been to develop a relatively short White Paper to guide the modelling work in Task 10; a White Paper Group was set up to develop the paper. The group includes:

- Philippe Davy (University of Rennes)
- Bill Dershowitz (GeoFractal)
- Stefan Finsterle (Finsterle GeoConsulting, LLC)
- Jeffrey Hyman (Los Alamos National Laboratory)
- Ivars Neretnieks (Royal Institute of Technology, KTH)
- Masahiro Uchida (Fracture Flow Solutions)
- Bill Lanyon (Fracture Systems Ltd)
- Björn Gylling (Gylling GeoSolutions)

The group produced a summary paper (Lanyon et al. 2021) setting out a preliminary view on a pragmatic validation approach, which was submitted to the DFNE 2021 conference[1]. The scope of the task and length of some of the individual contributions have resulted in the decision to also produce this report as a series of individual essays together with an updated and modified version of the summary paper.

In Chapter 2, Ivars Neretnieks considers the nature of flow and transport processes in crystalline rock and how it has been modelled within SKB's programme. The chapter includes a range of experimental observations of channelling of flow and transport from SKB's work in underground facilities since the Stripa Project, together with results from safety assessments. The chapter also suggests a set of validation principles and stresses the need for consistency within and between models.

In Chapter 3, Philippe Davy presents work from the SKB DFN Handbook (Selroos et al. 2022) that illustrates how to build model confidence by model space reduction. The chapter provides a formal view of the different modelling activities (model selection, conditioning, calibration and prediction) and their impact on the "model space". Example applications to fractured rock are also discussed.

In Chapter 4 Stefan Finsterle presents some of the philosophical issues around validation together with a proposed pragmatic model evaluation method. This chapter provides the basis for the "recipe" set out in Chapter 5.

[1] DFNE 2021 (Discrete Fracture Network Engineering) was originally planned for June 2021 but was postponed; the paper was presented at a special virtual DFN session of ARMA (American Rock Mechanics Association) 2021.

Chapter 5 is an updated and modified version[2] of the DFNE 2021 paper outlining a summary of the pragmatic validation approach based on discussions within the group and the longer essays in this report. The chapter presents the proposed general approach to validation based on the use of multiple models and a structured model evaluation. It provides the basic recipe for the approach.

In Chapter 6, Masahiro Uchida and Bill Dershowitz present proposals for how the pragmatic validation approach might be used within Task 10. They use the concept of "necessary and sufficient conditions" from mathematical logic to determine criteria for model validation and suggest a range of possible requirements on models that could be used as "necessary conditions" in different applications. The proposed content of a validation workflow document is presented.

Chapter 7 provides some final remarks on the ideas presented and their relationship to established "good practise" in hydrogeological modelling.

[2] Some changes to the paper are based on comments from Anders Winberg that highlighted the importance of potential long term processes and qualitative differences in the available data when proceeding from early site characterisation to underground exploration.

2 Confidence building in flow and solute transport modelling

Ivars Neretnieks, Royal Institute of Technology, KTH

2.1 Aim

Nuclear waste repositories in fractured crystalline rocks are in early stages of construction in Sweden and Finland. Copper clad canisters with spent fuel surrounded by compacted bentonite clay will be deposited in tunnels in the rock at depths of around 400 to 500 m. Water continuously seeps in the fractures in the rock. The water can potentially carry corrosive agents to the canisters. Should a canister be damaged so that radionuclides dissolve in the water, it could transport nuclides to the biosphere. Modelling of flow and solute transport is used to assess the protective strength of the bedrock in so-called performance assessment (PA) studies.

The models are used to simulate flowrate distributions and transport properties of the hydraulically conductive fracture network. The models are used to "predict" flowrate distributions in regions that have been explored by mainly very sparse deep boreholes and to make predictions over at least 100 000 years. This obviously cannot be made with large detail and certainty. Nevertheless, models based on sound scientific reasoning and accepted laws of nature together with general information and data from different disciplines can make bounding estimates that are sufficiently credible to convince the scientific community that the final repository can fulfil the task to protect biosphere and humans from the radioactivity of deposited spent nuclear fuel.

This chapter describes how building of the models and modelling has been done over the last three decades.

2.2 Flow and transport processes in fractured rock

Chapter 5 presents and illustrates by examples how flow and solute transport is and can be modelled in fractured rock. Experiences from modelling these processes in recent PA projects in Sweden and Finland are used to concretise some of the problems addressed. A review on hydrogeological issues for PA can be found in Tsang et al. (2015). An additional valuable source on the treatment of geosphere retention phenomena in safety assessments is the report SKB (2004), sponsored by the European Atomic Energy Community (Euratom).

2.2.1 Fracture networks in crystalline rock – Formation and properties

Fracture generation and deformation mechanisms

The Fennoscandian rock is nearly 2 billion years old. It has been subjected to stress with strongly varying magnitudes and directions. Shear stresses have created shear induced fractures and have also broken up the rock into smaller blocks and fragments within the fracture zones, the largest of which can extend over many kilometers and often reach ground level. Tensile fractures, the majority of the smaller individual fractures, are formed when the stress in one direction becomes much larger than that in another direction. Tensile fractures are oriented more or less in the direction of the minor stress at the time of formation. Often these fractures can be grouped in two or three major directions.

Immediately after the formation of a tensile fracture the stress in the surrounding rock and thus on the fracture changes. Some minor, sideways displacement of the two rough undulating faces of the fracture takes place, causing a mismatch of the two sides. The distance between the two faces, the aperture, varies. In many locations the faces are in contact.

It is likely that the faces in large fractures are more displaced than in small fractures because displacement is zero at the boundary and largest near the centre of the fracture. The elastic properties of the rock and the stress field will influence the magnitude of the displacement. When the fracture is re-compressed by changing pore pressure or rock stress magnitude and direction, the aperture decreases. In some locations, crystal fragments hinder the faces to be in direct contact.

Hyman et al. (2016a) discuss the geomechanical processes and mechanisms of fracture formation and the use of the finite-discrete element method to simulate the process of hydraulic fracturing and the hydraulic properties of the resulting fractures. The large number of references also cover various approaches to model flow and transport in fractures and networks. Maillot et al. (2016) compare the predictions of Poisson DFNs to new DFN models where fractures result from a growth process defined by simplified kinematic rules for nucleation, growth, and fracture arrest. This so-called "kinematic fracture model" is characterized by a large proportion of T intersections, and a smaller number of intersections per fracture, which agrees with what is often found in field observations.

Describing and modelling the network

To describe and model the geometry of the network, information of the fracture location, shape, size and orientation is needed. The fracture shape is commonly assumed to be a two-dimensional circular or elliptic planar structure. It is not expected that many more fractures will form during the next hundred thousand to million years even during the expected recurring glaciations. However, the rock movement and later melting of kilometer thick ice masses above the rock will change the magnitude and direction of the stress field and will induce compression and relaxation and may alter existing shear fractures. This may change fracture apertures and permanently influence transmissivities and solute dispersion. Infiltrating water during ice melting and temperate periods can cause dissolution of minerals and precipitation of new minerals in the fractures and change the aperture distribution and transmissivity of the fractures.

There is at present considerable interest in modelling fracture networks and flow and solute transport in them in different areas including oil and gas extraction, carbon dioxide sequestration and nuclear waste management. Here, only a few models are mentioned. Fracman is a widely used commercial program (code) (Dershowitz et al. 2019). Connectflow (AMEC 2014) was used by SKB and Posiva in their recent simulations for PA (Joyce et al. 2010, 2014, Hartley and Roberts 2012). Hyman et al. (2015) used a similar model with a powerful parallelised computational suite to generate DFNs to simulate flow and transport with extremely high spatial resolution. Several other modelling tools have been developed over the last 40 years and are referenced in the publications above.

Structure of geological fracture networks

The fractures form a complex three-dimensional network. Large fracture zones usually are seen on the ground surface where their size and orientation can be measured. The width (the distance between the two boundaries of the zones) and the sizes of rock blocks and fragments inside the large zones can be assessed at the surface or in shallow trenches. At depth, some of the fracture zone location and properties can be confirmed by observations in deep boreholes and tunnels. Smaller fracture zones and fractures intersecting outcrops can also give information on size and orientation as well as density (mean spacing P_{10}, mean length/area, P_{21} and indirectly of the mean surface area per volume P_{32}). These entities and their metrics can also be partly assessed by observations in deep boreholes and on borehole cores. Boreholes and cores also supply information on rock types, fracture apertures and mineralogy, providing information on past chemical alteration in and near the fracture.

To describe and quantitatively model the geometric fracture network, except for major fracture zones, the smaller zones and fractures in the deeper portions of the rock must at present be described by stochastic distributions derived from surface observations and borehole information and possibly also from tunnels. The stochastic models can be locally conditioned at outcrops, tunnel surfaces and borehole intersections where measured data is available (Bym and Hermanson 2018, Appleyard et al. 2018).

The water conducting network

The major and minor fracture zones potentially contain breccia zones and zones of strongly chemically altered minerals, including clays. These generally seem to have hydraulically conductive regions that may be connected over large distances. Smaller fractures that are fully closed over all or most of their entire size will not be part of the hydraulically conducive network. Other fractures have variable aperture and may be partly closed or sealed. Flow will seek out different preferential paths depending on the direction of the overall hydraulic gradient. Strong heterogeneity in permeability within the fracture and the large contrast to rock matrix permeability generate complex flowpaths. These essentially two-dimensional paths are often called channels although they are not channels in the sense that they may have fixed physical borders but are a dynamic consequence of the interplay between aperture heterogeneity and hydraulic head boundary conditions. It may be noted that often a "fracture" consists of a series of shorter sub-fractures in-echelon in essentially the same plane. The sub-fractures can be hydraulically connected. This makes it difficult to characterise fractures by borehole observations. In network modelling, they are considered to act as a single conductive feature.

At fracture intersections, water can flow from one fracture to the next where the intersection is not sealed or when the water can seek out a path around the sealed location. Fracture intersections can themselves also be quite conductive and facilitate flow from fracture to fracture *along* the intersections, making the flow network even more complex.

The hydraulic network can be conceived as a three-dimensional channel network (ChN), where the channels are located in the fracture "planes".

2.2.2 Flow and solute transport in fractures and in networks of fractures

Overview

Flow in conduits and in porous media is driven by hydraulic head differences, or more generally by hydraulic head gradients in each location. The flux (flowrate per cross sectional area) at very low Reynolds number is proportional to the gradient. These conditions apply to flow in narrow fractures in low permeability crystalline rock.

It is not possible with present techniques and tools to obtain exact information on every fracture in the network. Most properties, fracture intensity, location, size, orientation, transmissivity distribution etc, must be treated as stochastic entities. A large number of different methods are used to derive the information needed for modelling. This is discussed in Section 2.2.4. The information is highly site-specific. The data are summarised in the form of probability distributions, pdf's. This information is used to generate stochastic DFN's using the pdf's for the different parameters and correlations between them. A large number, typically many thousands, of different stochastic networks is generated, each giving a different result. The solution of the model flow equations gives information on the water fluxes in every fracture in the network. If there is information of how aperture and transmissivity vary in the fractures also details of local fluxes everywhere can be determined. In a similar manner ChN's can be constructed.

Flowpaths can be traced out and also the total water flowrate to some discharge location of interest e.g., a lake can be assessed as well as the amounts of radionuclides carried. The multitude of different stochastic realisations gives information on the flowrates, paths and how they may differ and vary in space and time. This will be further exemplified, detailed, and illustrated in the following sections in the present section. Next section gives examples of how this information can be used in PA applications.

Properties of "simple" single fractures

Bodin et al. (2003a, 2003b) give an overview of fundamental mechanisms and mathematical relations for modelling of flow and transport in single fractures. With simple we mean fractures that essentially have two surfaces separated by a narrow aperture that may vary over the fracture area. The aperture variations over the surface affects the transmissivity, the solute transport and channelling properties. These properties have been studied by laboratory experiments on non-stressed and stressed fractures in granites and other rocks with induced fractures. Witherspoon et al. (1980) found in large scale laboratory fracture experiments that the transmissivity- hydraulic aperture relation for non-sheared fractures (joints) was well described by the cubic law, whereas Ishibashi et al. (2015), who experimentally studied non-sheared as well as sheared fractures (faults) found more complex relations. They also found that the mechanical aperture was correlated to the fracture size in different ways for joints and faults. They note that faults can have different transmissivity and channelling properties for flow perpendicular and parallel to the shear direction and that faults are considerably more transmissive than joints.

Larger fractures that have been subjected to more shear contain broken fragments that form an in-fill. The latter may have been degraded by chemical interaction with seeping water and clays may have formed. The rock close to the fracture surface may also have developed additional microfractures. These are called complex fractures. This is exemplified and discussed later and illustrated in Figure 2-6.

The details of the fracture behaviour can affect the modelling in different ways. One is that undulation of direction around the mean fracture direction influences the uncertainty of determining the direction from borehole observations. Another reason is that the aperture variations give rise to, sometimes severe channelling, which is one of the main concerns in modelling solute transport and one of the important issues of the work in Task 10. We illustrate first how the undulation impacts on the assessment of mean fracture direction.

Correlations between apertures at locations at different distances from each other can be described by fractal relations. The aperture difference between two locations at close distance is more likely to be smaller than the difference at larger distances. For the aperture distribution it can be described by a self-affine relation that gives the relation between the apertures and the distance between locations in a stochastic sense (Ishibashi et al. 2015). The aperture distribution has been found to be well described by a truncated self-affine structure with a Hurst exponent roughly between 0.6 and 0.8 (Stigsson 2019). This has been found to be valid over many orders of magnitude. It implies that at small distances between two points on the fracture the amplitude (aperture) difference can be very large compared to the distance between the points but that at larger distances the amplitude to distance ratio decreases. The facture seems to be more planar if a plane were to be fitted to it.

This has important implications for the accuracy of fracture orientation when assessed from observation of its intersection with small boreholes. This is illustrated in Figure 2-1. Stigsson (2019) found that several tens of degrees difference (error) can be expected. This is easily conceived by looking at Figure 2-1b. Consider a fracture larger than a few tens of metres, a typical fracture size of interest in our case. The borehole size, e.g., 0.076 m, compared to that of the fracture may be on the order of one pixel or less in the enlarged figure. The fracture orientation described by the normal direction in a pixel would indicate the direction of the fracture in that location. It could easily vary between minus and plus several tens of degrees in both x and y-direction. This suggests that the uncertainties of fracture orientations derived from borehole information are very large

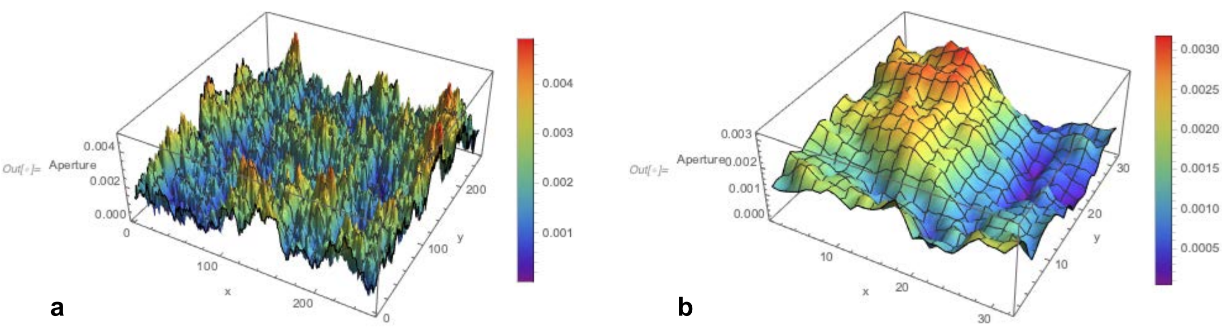


Figure 2-1. *a) Example of aperture distribution; b) an enlargement of the lower left-hand corner of a) with a lower resolution of details. The aperture scale is arbitrary and relates to the extension of the fracture 256 × 256 length units.*

Modelling flow in single simple fractures

Another important effect of aperture variation is that it causes flow channelling. Figure 2-2 and Figure 2-3 show simulations of steady flow in self-affine fractures. In the examples the aperture distribution used to generate the synthetic fractures was made with a Hurst exponent of 0.7. The scale is arbitrarily normalised, and the aperture field is discretised with 256×256 pixels. The simple algorithm used to generate the variable fracture surfaces is based on the method presented in Peitgen and Saupe (1988, Sections 2.4.3 and 2.5.3). The figures are taken from an ongoing study by Neretnieks, the author of the present section. The local transmissivity in the variable aperture fractures was taken proportional to the aperture to the third power, the so-called “local cubic law”. Figure 2-2 illustrates the transmissivity distribution and flowrate distribution for constant head difference across the fracture in the y-direction and closed boundaries on the other two sides. Preferential flow path(s) are clearly seen in Figure 2-2b and in the further examples in Figure 2-3.

The preferential flowpaths, or channels, vary considerably in width along the paths. In the white-reddish locations a given “packet” of water passes through a small “bottleneck” and then spreads out over a larger width, before converging again. The pictures suggest that the width of channels in a fracture ending at a tunnel face, where the head is constant would be a very small fraction of the size of the fracture and that the distance between “main” channels within a fracture are on the order of the size of the fracture. Flowrates in the “minor” channels vary over a very large range. The self-affine property of the fractures suggests that smaller fractures in a network of channels will tend to have narrower channels with smaller flowrate. This is supported by many observations in drifts and tunnels, Tsang and Neretnieks (1998), Abelin et al. (1991a, 1991b), Neretnieks et al. (2017). Section 2.2.8 show several examples of field experiments and observations of channelling.

Large normal stress compresses the fracture void space and decreases flow through the fracture. This process is illustrated in Figure 2-4 which shows the flowrate distribution in a fracture before and after compression by 0, 50 % and 100 % of the original mean aperture. At a compression of 90 to 100 % the flow decreases to less than 1 % of that of the uncompressed fracture with about half the fracture area closed to flow. This agrees with the findings in the Witherspoon et al. (1980) experiments.

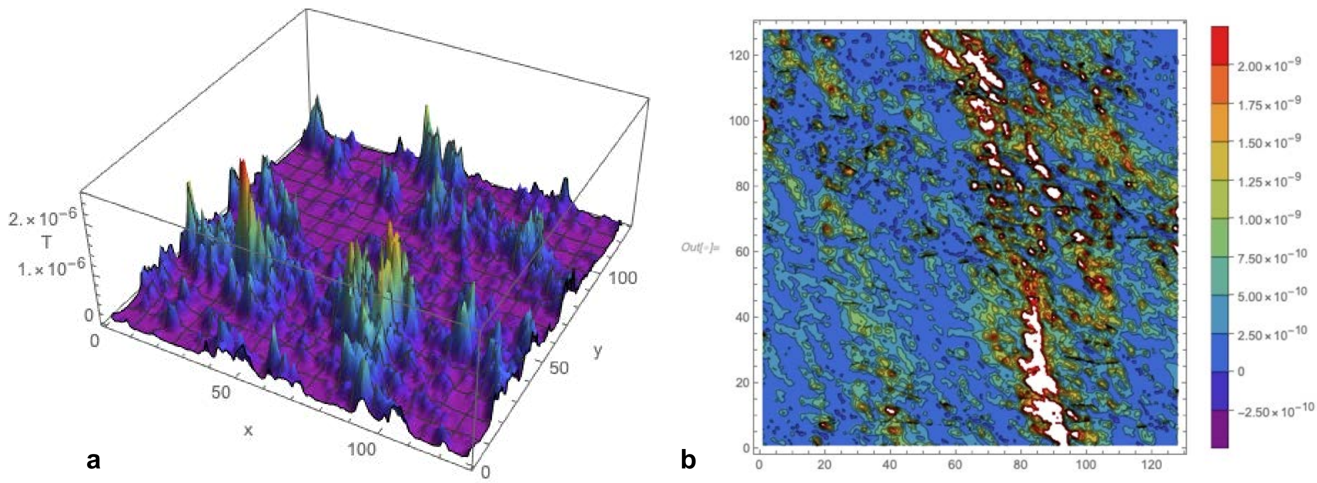


Figure 2-2. *Simulations of steady flow in self-affine fractures: a) transmissivity; b) frame and contour plot of flowrate in the fracture. Flow is in the y-direction.*

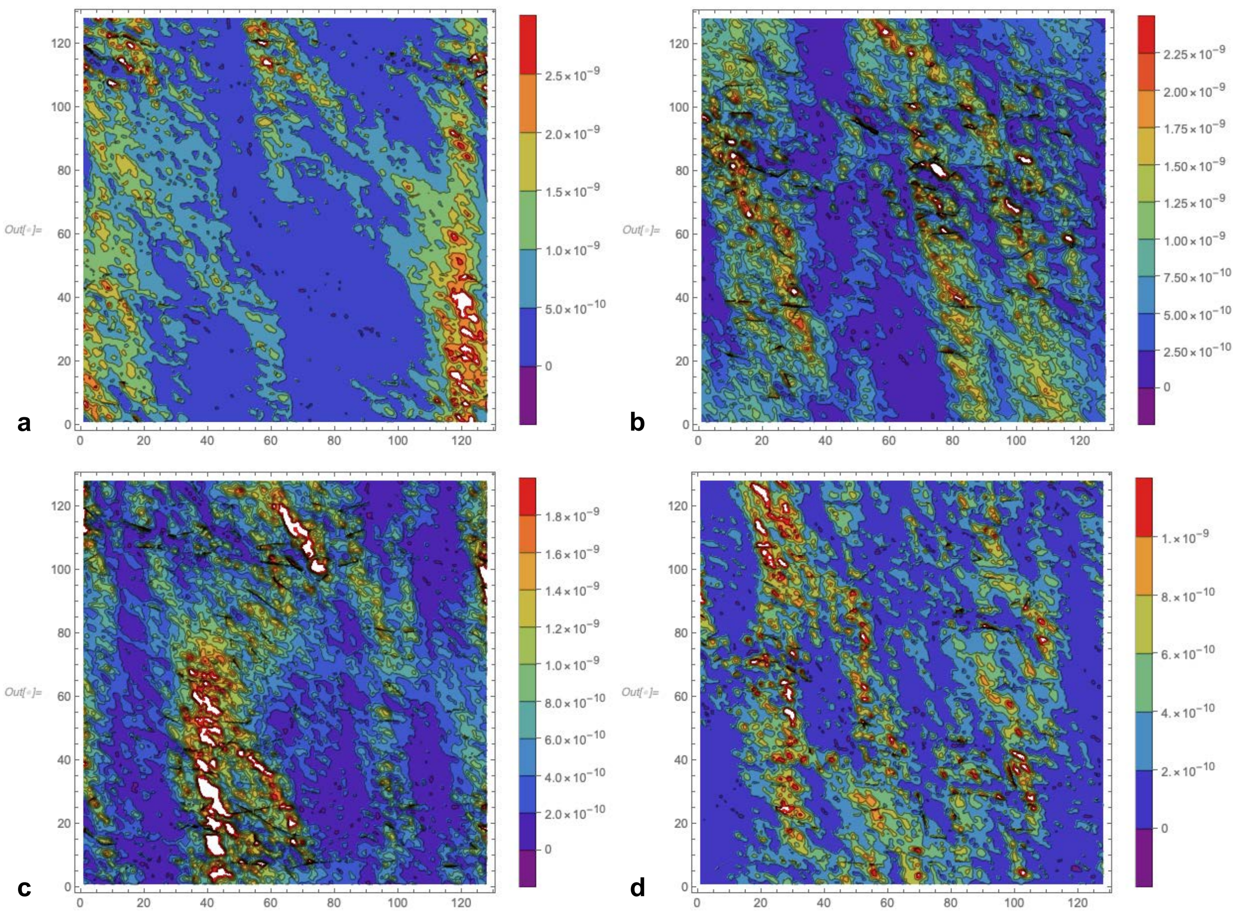


Figure 2-3. Simulations of steady flow in self-affine fractures; four (a-d) other realisations of flowpaths. Flow is vertical. The white-reddish locations act as local "bottlenecks" for flow.

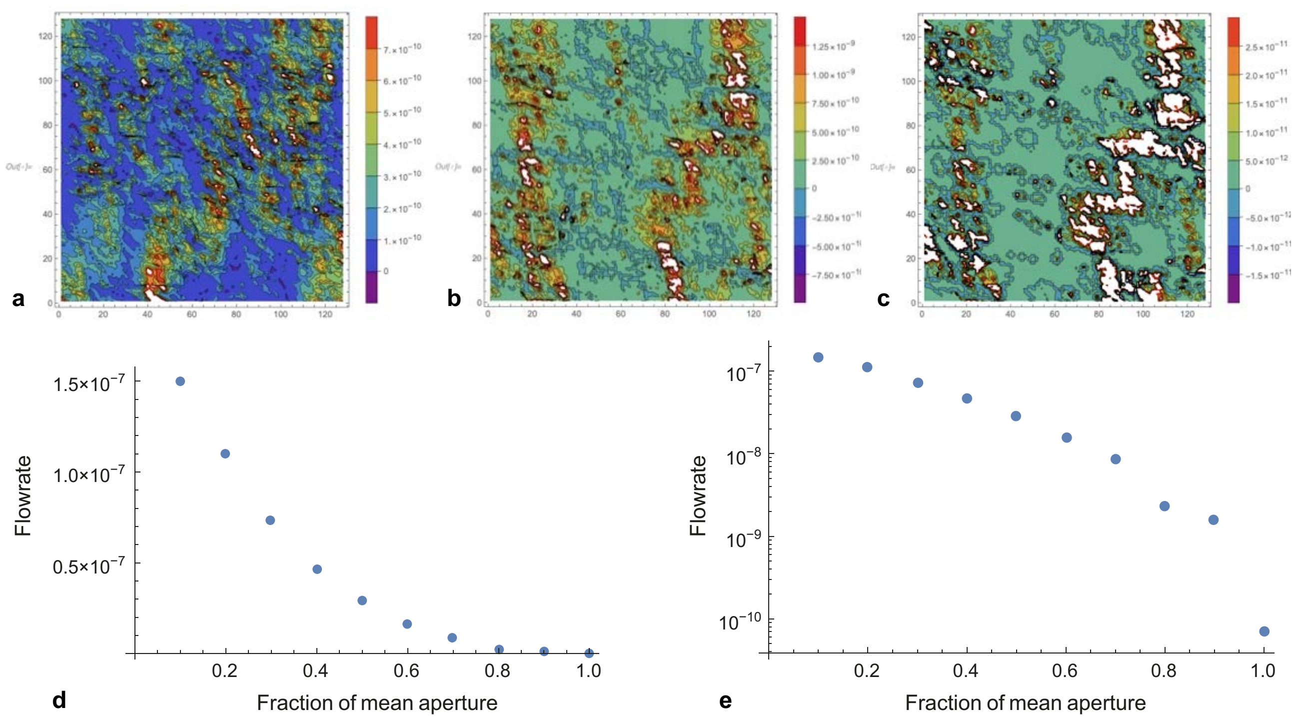


Figure 2-4. Flow pattern in an uncompressed fracture and after compression by: a) 0 times the original mean aperture; b) 0.5 times the original mean aperture; and c) 1 times the original mean aperture. Plots show how the flowrate decreases with increasing closure: d: linear scale; e) logarithmic scale.

Flow from one fracture to another

In the fracture network the processes at the intersections will impact flowrates, channelling and solute transport. We illustrate some issues that may be warranted to study more.

Flow from fracture to fracture is influenced by the transmissivity at and along the fracture intersection. Figure 2-5 illustrates the flow pattern in two intersecting fractures with different properties. Figure 2-5a shows the transmissivity of the two fractures. In Figure 2-5b the fracture intersection is a high permeability "tube" that allows flow *along* the intersection as well as across it. In Figure 2-5c flow is allowed across the fracture intersection but not along it. It can be seen that there is no dramatic difference between the flow patterns. However, had there been an overall head difference also in the left-right direction, the high transmissivity of the intersection itself in Figure 2-5b would have a considerable impact on the flow pattern in the network. Transmissive fracture intersections with high flowrates are often observed in drifts and tunnels (Stanfors 1987, Palmquist and Stanfors 1987, Abelin et al. 1991a, 1991b) and their impact needs to be considered in three-dimensional fracture networks.

Flow in complex fractures and fracture zones

The larger the fractures are the more complex they become. Fracture zones contain crushed rock with a very wide particle size distributions ranging from the size of the width of the zone down to a micrometre and less. Fracture zones have been found to contain patches or regions in which mineral alteration has produced large amounts of clay. Very low to very high conductivity regions exist, which may promote the formation of preferential flowpaths. There are indications that there can exist a "core" of very low permeability in the central part of large fracture zones that at least locally hinder flow across the zone but facilitate flow in the outer regions of the zone. This may lead to "compartmentalisation" meaning that a hydraulic head change, e.g., pressure pulse on one side of the zone negligibly is transmitted to the other side of the zone, although it is transmitted over long distances parallel to the zone. Large zones are more transmissive than smaller features (minor zones and fractures) and they form a "backbone" for the high-flow paths over long distances.

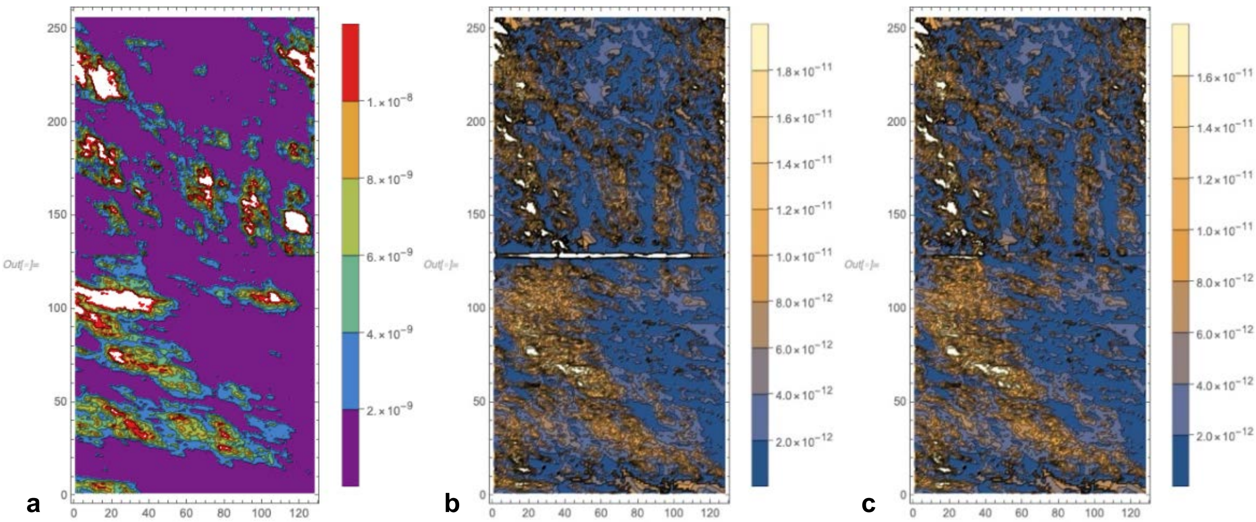


Figure 2-5. *Flow pattern in two intersecting fractures with different properties: a) transmissivity; b) flowrate distributions in two coupled compressed fractures when sideways flow is allowed at intersection; c) flowrate distributions in two coupled compressed fractures when sideways flow is not allowed at intersection. White areas represent high flowrates.*

The heterogeneous structure of complex fractures and zones is especially difficult to model because there on the one hand is little field data available and that on the other hand the correlation structure between transmissive locations along the fracture/zone is suspected to be able to generate extensive channelling over long distances. Follin and Stigsson (2014) analysed the data from the Forsmark site and a strong correlation with depth was found. At the same depth the transmissivity typically spanned three orders of magnitude but there is practically no information on the structure of the porosity and conductivity within the complex fractures. This implies that it at present is not possible to quantify channelling effects in these structures that extend over long distances although there are indications that it can be important, see Section 2.2.8.

Modelling flow in networks

Three-dimensional fracture networks can be generated using the data obtained from outcrops and deep boreholes in the rock. Orientations and extents of large fracture zones can be treated as known entities whereas data for smaller zones and individual fractures are treated as stochastic with data taken from stochastic distributions. Correlations between parameters such as fracture size and transmissivity and depth dependence can be accounted for, if known. In PA applications the elevation of the ground surface above and below the sea and in lakes is used for the upper boundary. The vertical boundaries are typically located at water divides far from the repository location and major regional fracture zones (deformation zones). Depending on what is to be simulated in a specific study, details of drifts and tunnels and even individual canister locations can be included.

The geometry of the network is crucial for all simulations of flowrates and flow distribution in the rock mass. The stochastic nature of many properties, e.g., orientations, fracture sizes and transmissivities within the fractures imply that in PA a large number of realisations must be made to assess the variability of the outcomes.

An example of how local information of flowrates are used in PA is given below to illustrate some considerations that could be used to build a simple but sufficiently good model to estimate if the rock mass that is considered could host a sufficient number of canisters. In this example it has a priory been decided that each canister deposition hole must not be intersected by a channel/fracture with lager flowrate than Q_{max}. Each canister needs a surrounding area of A_{can} and the total area of the repository, limited by the major zones and rock types must be able to host all the canisters. For the purpose of this example all canister locations must be located in the same horizontal plane.

With the stochastic data of fracture orientations, lengths, densities, transmissivities etc., the DFN is built in the computer. Even the effects of damaged zones around drifts and tunnels are accounted for. The fluxes are determined in detail. Canisters are placed at constant spacing along tunnels. Locations with too large flowrates are discarded. A number of different realisations are made and if large number of them shows that there is a good chance that a repository can be located in this rock the site is deemed to be acceptable in this respect. This was actually done in SKB's recent performance for the Forsmark site (SKB 2011).

Considering that the site's fracture properties have been estimated by a very sparse set of deep boreholes, approximately one borehole per four km^2, it should be emphasised that such a conclusion does NOT imply that the site IS suitable, only that it may possibly be. The suitability can only be confirmed during the excavation phase when the actual fractures, zones and network properties are determined. This implies that there is a need to develop suitable, practical, and rapid techniques by which information is gathered and evaluated as excavation proceeds to make predictions of the rock properties ahead. This will continuously improve the site model and increases general confidence in it. It is an important task to investigate whether data acquired from investigations match the forecast (model predictions) that have been built using data collected via surveys on the ground surface and in boreholes drilled from ground surface.

Transient flow

The network model can also be used to simulate transient flow caused e.g., by infiltration into the rock mass around the repository after construction when the rock partly has been drained of water. This can also be used to increase confidence in the hydraulic model.

Transients are also caused by viscosity and/or the water density changes e.g., by infiltration of fresh meteoric water. Fluxes change because the driving force for flow depends on local density differences, which evolve over time. Another process that generates density differences is that more/less saline water diffusing out from the rock matrix into the water seeping in the fractures continuously affects the local flowrate and flow pattern. This can potentially generate rapid vertical flow and affect the mass transfer rate from a contaminant source (Neretnieks and Winberg-Wang 2018).

The presence of heat sources e.g., due to radioactive decay in the waste generates changes in density and viscosity. This can be accounted for by simultaneously modelling heat flow by advection and conduction. The heat transport will influence the flow and vice versa. Complex local convection cells may develop. Similarly sources or sinks of chemical species from engineered barriers e.g., concrete and bentonite clay influence the water density. This can be accounted for in a similar way. In the following section solute transport in fractures and networks is described and some examples presented that may be necessary to consider in PA applications.

2.2.3 Solute transport in fractures and networks of fractures

Solute transport in both individual fractures and in fracture networks is influenced by molecular diffusion and chemical reactions such as sorption, in addition to that by the advection of the water. The dispersion of solutes along the flowpaths can have detrimental as well as beneficial effects on the transport of nuclides to the biosphere. Dilution could be beneficial in some situations but early arrival of a fraction of a plume may not allow a sufficient decay of even shorter-lived nuclides.

Dispersion is caused by a number of different and interacting mechanisms. A small pulse of solute that is injected in stagnant water will spread slowly and be diluted over time by molecular diffusion. Given time, a non-interacting solute will eventually diffuse into all accessible voids. The diffusion distance increases proportionally to the square root of time. In seeping water, the spreading pulse is carried by the water and some solute will arrive earlier than at the velocity of the streamline in which it is carried because of the diffusion in the flow-direction. The overall dispersion of a solute carried by flowing water in a fracture network is the result of a number of different mechanisms and processes. Attempts to model and quantify the spreading by an overall lumped parameter, a "dispersion coefficient" that would be a materials property of the rock have been attempted. This is one example of where model simplifications and abstractions have encountered severe problems. This is because in this case seldom one of the underlying processes and/or mechanisms has been the only or overwhelmingly dominating process for the final result over the range of conditions of interest. Also, it has not been successful to untangle the different dispersions mechanisms from tracer brake through curve, BTC, information only by curve fitting.

In conduits, tubes or slots the streamlines in the middle move more rapidly than those near the walls. For flow in *narrow* slots the diffusion across the slot will even out the concentration between different streamlines. The pulse spreading (dispersion) can be described and quantified by the so-called Taylor dispersion equation (Taylor 1953). The longitudinal spreading behaves as if it were caused by molecular diffusion but with a (much) larger diffusion coefficient. This longitudinal spreading increases proportionally to the square root of time. In PA application this effect would be negligible in constant aperture fractures. However, in *moderately* wide, tapered channels, with larger aperture in the middle, there will be diffusion between the rapid flow in the centre and the slow flow where the aperture becomes small. The dispersion in such channels can also be similar to Taylor-Aris dispersion but may not be negligible even on PA scales (Liu et al. 2018).

In contrast, in large fractures, meters or more in size, the distances between the channels may be so large that negligible mixing by molecular diffusion takes place between the channels. They behave as independent paths. A solute pulse injected evenly at the inlet of such a fracture arrives at different times to the outlet by the different channels. The joint residence time distribution at the outlet, where the waters in the different channels mix, e.g., a pumping hole or a lake, is said to subject to "velocity dispersion". The spreading by this process is proportional to the travel distance in contrast to that caused by molecular diffusion and Taylor-Aris dispersion, in which it is proportional to the square root of the travel distance. In modelling solute transport, it is essential to account for the correct dispersion mechanism when extrapolating to greater distances because the two different mechanisms will predict very different dilution and spread in arrival times, Neretnieks (1983).

Another important cause of dispersion is "matrix diffusion". In fractured crystalline rock the matrix is porous, albeit with a low porosity. This porosity is sufficient to considerably influence and delay the residence time distribution, RTD, of a solute pulse carried by the water. The solute molecules can diffuse in and out of the pores. Some of the solute may then diffuse back into the seeping stream, while the remainder will diffuse further into the matrix. A similar effect is observed in fractures with variable aperture in which there exist locations with essentially stagnant waters. The solute from the flowing channels can diffuse into the stagnant locations and be considerably delayed. This effect is compounded with that of the matrix diffusion as the solutes that diffuse into the stagnant zones meet with additional fracture surface area into which to diffuse, Shahkarami et al. (2016).

Many radionuclides are positively charged ions and complexes. They sorb (ion exchange and attach by surface complexation) to the minerals on the fracture surfaces and to the minerals in the porous rock matrix accessed by diffusion. The solutes that enter the matrix pores from the seeping water reside there for some time before they re-enter the water. Their migration is retarded compared to the mean residence time of the seeping water. The retardation in a fast path can be small but can be very large in slower paths. Many sorbing nuclides can be retarded so much that they decay to insignificant concentration before the water reaches the biosphere, Neretnieks (1980).

The matrix close to the fracture is often altered by chemical reactions, stress, and intrusion of reactive solutes. Multiple layers or regions with varying mineralogy and porosity exist. Parts of fractures are in-filled to varying degrees with small rock fragments, mineral grains, and clays. A schematic of the fracture void space developed for the TRUE-1 site at the Äspö URL is shown in Figure 2-6. These different porosity and mineralogical heterogeneities add to the uncertainties that must be considered in simulating water flow and solute transport.

The total water volume in the matrix pores of the rock is as large or even much larger than the volume of the mobile water in the fractures. The exchange of solutes between flowing and matrix water is very slow compared to the rate of change by the flowing waters in the fractures. The present-day composition of porewater can give information on water bodies that infiltrated the rock a long time ago. Such information together with independent information on ice evolution during and after an ice-age can be used to increase the confidence in the modelling.

In a fracture network in which multiple independent flowpaths link locations A to B, a pulse injected at A will spread by velocity dispersion when collected at B, in addition to the spread within each path by diffusion and Taylor-Aris dispersion. The spreading of the original pulse will increase because the pulse has travelled different distances with different velocities and flowrates. Velocity dispersion and dispersion due to matrix diffusion are the two most important dispersion mechanisms in fractured rocks for PA time scales.

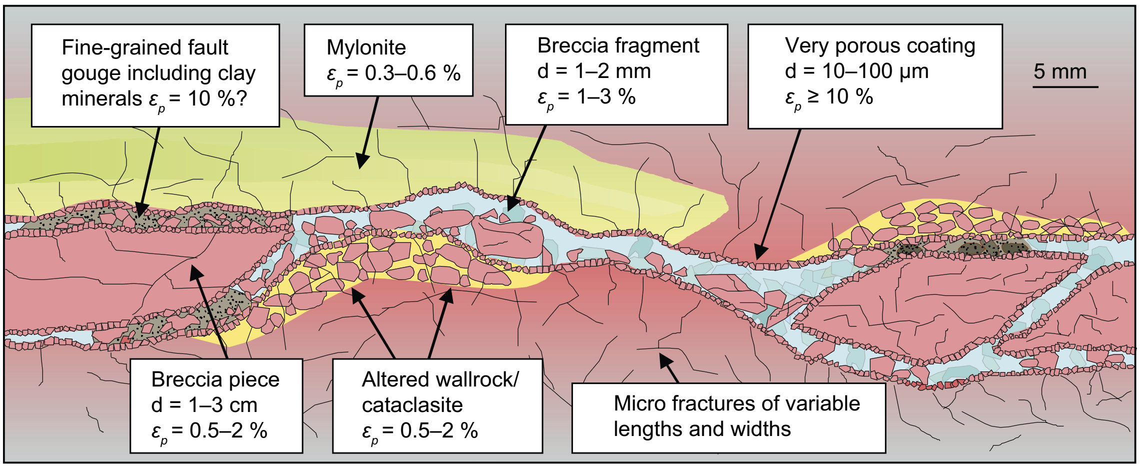


Figure 2-6. *Microstructural model of fracture porosity from TRUE site (Andersson et al. 2002).*

2.2.4 Characterisation of fractured rock for hydrology and transport

Approaches and methods to determine hydraulic and network properties

Many properties of the rock mass can be characterised in the laboratory. Some tests must be made in the field. A short overview of the characterisation performed during recent investigations for the Forsmark site is given as an example (see references in SKB 2008a, 2008b).

Fracture size, orientation and spatial distribution can be estimated from measurements at outcrops, but these data may not be representative for fractures at depth. Rock type, mineralogy, mechanical and physical properties, stress conditions and orientations can all to some extent be measured in boreholes at depth and on cores taken from the holes. Alteration zones and secondary minerals near fractures are characterized and their thickness is measured on the cores. Matrix diffusion properties are typically determined both in situ and on cores in the laboratory.

At Forsmark there are ca.15 km-long boreholes that were inspected by borehole camera to obtain fracture orientations and information on whether the fractures were sealed or open. Hydraulic heads were monitored over extended periods. The specific yield, flowrate, of individual fractures were determined for each fracture intersected by the boreholes by differential pumping tests, PFL-tests, Öhberg and Rouhiainen (2000). The specific yield in the fractures is not proportional the transmissivity because the flowrate in very high transmissivity fractures is often limited by less transmissive fractures in the vicinity. The transmissivity distribution was obtained by calibration of the specific yield by simulations using the hydraulic network model (Follin et al. 2014). Hydraulic packer tests over long borehole sections with many fractures were also made to measure the integrated transmissivity of features intersecting the test interval and compared with specific flowrate data of the individual fractures.

Measurements of water chemistry obtained in water extracted from fractures at different depths and locations were made for comparison with flow and transport simulations. This allowed deductions of historic infiltration of water under different climatic conditions including glaciation cycles, during which, depression of the rock by km's of ice occurs. Leaching of drill core samples were made to support simulations of how solutes diffuse in the rock during the earlier periods of different infiltrating waters. Such information is important to increase the credibility of simulations and increase our confidence in the models and modelling.

Solute transport and dispersion phenomena

A number of field tests and observations have been made in fractured rock at different sites within the Swedish SKB program. They include between hole tests, tracer tests in individual fractures and in networks, and tests and observations of flowrate distributions and channelling in drifts and tunnels. Some of these are shortly described in Section 2.8.

Solute transport behaviour in a *homogeneous* medium, e.g., a porous bed or a slot is commonly characterised by the mean residence time and standard deviation of the RTD. In tracer experiments the breakthrough curve, BTC, of a tracer pulse is typically fitted to the advection-dispersion equation, ADE. In experiments with non-interacting tracers in capillaries, planar slots and porous beds with small equally sized particles, the BTC is usually bell-shaped and fits well to the ADE. The accessible volume to flow in the path and a dispersion coefficient can be determined. These parameters can be conceived as material properties of the system. They will typically depend on some characteristic size, e.g., slot aperture or particle diameter in the porous bed and on fluid velocity. The underlying physics are well understood as e.g., Taylor dispersion in which there is mixing by molecular diffusion between the streamlines. Diffusion is a fundamental, well-understood process. In *homogeneous* systems, extrapolations to different flowrates or different characteristic size can usually be done with confidence.

In contrast, in *heterogeneous* materials such as porous beds with widely varying particle sizes, variable aperture fractures and fracture zones the BTC's often have multiple peaks that are more or less separated. This is due to the presence of independent pathways, each of which could be characterised with its own volume and dispersion properties. In a heterogeneous medium the properties of the different pathways will strongly influence the overall flowrate and dispersion of the ensemble of paths. Any mixing *between* paths in the medium will also have a major impact on dispersion and spread of a tracer pulse. Although a dispersion coefficient can often be estimated from the BTC's, such a coefficient is not a characteristic of the medium as it can be in homogeneous media because the spreading of a tracer pulse is influenced by different mechanisms with different impact, in different proportions, at different locations at different times. Nevertheless, sometimes some valuable information can be derived from the BTC to support or to in-validate some assumption in the model.

When the solutes can diffuse in and out of stagnant waters along the flowpaths, such as the pores in the rock matrix, this retards the solute and also changes the shape of the BTC by generating a drawn-out tail. In principle the shape of the BTC contains information about the mechanisms that generate it. For example, a sharply rising BTC suggests that at early times plug flow is an important component of the transport. A very long tail with a constant slope −3/2 when plotted in a log-log diagram suggests that matrix diffusion dominates for long times and that the matrix has a very long extension (thickness). However, very seldom parameters for the different contributing mechanisms can be derived from the BTC's from tests in heterogeneous media.

In tracer experiments in underground fractures additional difficulties are encountered. In such tests several boreholes are drilled in the site to find a suitable fracture. The fracture selected for the experiment is explored by a small number of boreholes. The borehole intersections with the fractures cover a very small fraction of the fracture surfaces. The fracture sizes, boundaries and intersection with other fractures are mostly unknown. Different hydraulic tests are used to characterise the transmissivity variations. Pressure pulse responses indicate hydraulic connections between different borehole intersections with the fractures. The target fracture is "identified" from information on local transmissivity estimates in the boreholes, the pressure pulse responses, orientations etc. It is acknowledged that there may be considerable uncertainties because fractures are not ideal planar features and because they belong to a network of conductive fractures.

For the tracer tests, injection and extraction/sampling locations are typically selected at high transmissivity locations in order to ensure sufficiently high flowrates to dominate over other flow disturbances and prevailing regional flow. The actual flowpath(s) and their properties cannot be determined from the observations in the boreholes. Changing the flowpaths by using different injection and extraction holes can be used, as for example in the TRUE-1 experiment (Winberg et al. 2000). The BTC's can differ hugely between different pairs of holes.

The chosen feature is one in a network and a part or most of the water flowrate pumped in one borehole may have entered the feature from other features intersecting between the injection and extraction holes. In such a situation the residence time of water in the chosen feature is not given by the pumping flowrate. This was found to be a major uncertainty in the evaluation of the comprehensive field testing performed as part of the TRUE 1 experiment (Winberg et al. 2000).

The literature abounds with papers in which tracer BTC's are characterized by a dispersion coefficient, more or less implying that this is a characteristic of the medium. Sometimes several tests have been made in the same medium, even in nearby locations. Widely different values of the dispersion coefficient are found. Gelhar et al. (1992) compiled data from a large number of experiments in different rock types and found that the longitudinal dispersion coefficient increases with observation distance. The data cover distances of more than five orders of magnitude. This implies that the main dispersion mechanism is hydrodynamical (velocity) dispersion, i.e., caused by mixing of flow from independent pathways/channels at the *observation location* but with essentially negligible mixing between different paths underway (Neretnieks 1983). The dispersion is not caused by a large number of mixing incidents underway as in a homogeneous porous medium. This implies that very long continuous pathways with little mixing between them is important and must be accounted for in a flow and transport modelling in fractured rocks.

2.2.5 Conceptualisation and characterization of channelling in fractures and networks of fractures

Two important themes in this chapter are flow and transport in networks of fractures and channelling both in individual fractures and in networks. They are intimately coupled.

The transmissivity distributions of the fractures are assessed from measurements of injection or withdrawal of water in the borehole intersection of the fractures in the network. There are various ways to identify or isolate individual fractures to find the flowrate from each fracture. From the flowrate/head relations a first estimate of the transmissivity distribution of the fractures in the network can be derived.

The data must be conditioned to account for the effect of the network. It must also be noted that with the (very) sparse boreholes used in characterization of a site a fracture is in practice only intersected by one borehole, if at all. Furthermore, it was found in the extensive site investigations at Oskarshamn and Forsmark that at repository depths, typically more than 90 % of all fractures in boreholes have transmissivities below the measurement limit, (Rhén and Hartley 2009, Rhén et al. 2008). At Forsmark in 13 of the about 1 km long cored boreholes only a few % of all fractures were transmissive at depths below 100 m, Follin et al. (2014).

This information does not allow us to distinguish between whether the e.g., 90 % non-flowing fractures are closed over their entire area and the remainder are transmissive over the entire area or, whether all fractures have open "channels" over 10 % of their area. A stochastic distribution of the aperture distribution would give similar results. Assuming the first case (fractures either are closed or open) one would construct a network considering only the 10 % "open" fractures, neglecting the presence of the "closed" fractures. For the second case another network with all fractures, each having one channel covering 10 % of the fracture surface will result. In the first case a sparse DFN results. In the "channel network", ChN, there are ten times more channels in a much denser network. It is not obvious how this influences channelling. It may be noted that Ishibashi et al. (2015), based on a considerable number of experiments found that the flow wetted surface FWS was between a few to 20 % of the fracture area. Sheared fractures had the larger values. About 10 % of the fracture was found to conduct the flow in a large scale flow and tracer experiment in the Stripa URL, Abelin et al. (1991a, 1991b).

It was illustrated in the examples in Section 2.2 that continuous flowpaths develop without the need to assume the "channel" in one fracture physically meets that in the next fracture because the channels are not fixed features. They develop where passage is easiest for a given situation.

The basic DFN and ChN for flow modelling are conceptually quite different but may not differ dramatically for simulations of flowrate distributions as in principle the DFN can generate a channelling pattern if each fracture is modelled as a variable aperture fracture with good resolution. This demands very much larger computer power compared to the conventional DFN approach in which every fracture is in principle assigned one conductance. For solute transport there are differences. To account the flow wetted surface, FWS, over which the diffusion in and out of the rock matrix takes place would be the same. For matrix diffusion effects the models might not differ drastically. However, the ChN can account for diffusion in and out of stagnant waters in the fracture plane adjacent to the channels because channel widths are inherently modelled in the ChN model and can be expected to result in stronger retardation of solutes. In contrast to ChN modelling, so far DFN modelling does not seem to have formally considered the presence of specifically conducive channels at fracture intersections, which have been observed in drifts and tunnels.

There are indications that the volume of (practically) stagnant water in the fractures can make up a considerable fraction of the total volume of fractures. Ishibashi et al. (2015) found that 40 to 60 % of the fracture surface is in contact for faults and joints respectively and not available to flow. They also found that about 10 % of the surface is accessed by the channels. This implies that there remains 30 to 50 % of the fracture surface that possibly is in contact with the water in the channels and could be accessed by diffusion species from the flowing channels. This can potentially increase the access to the rock matrix porosity and the retardation of solute transport (Shahkarami et al. 2016).

It may also be mentioned that simulations using the two different network models using the same data have been compared although no diffusion into stagnant water was invoked (Selroos et al. 2002).

2.2.6 Some aspects on building confidence in models, modelling and interpretation of experimental results and observations

General

Confidence in that the models and modelling sufficiently well describe the desired object and the processes is promoted in different ways. For PA applications the confidence building includes the quality of the models, the data, its acquisition and interpretation, the argumentation, review of and revisiting of assumptions and many other aspects in which firm quantification is not possible. Geohydrology has evolved enormously over the last half century driven by oil and gas exploration, mineral mining as well as studies of water resources and water contamination. The models used are continuously improving and becoming increasingly complex and elaborate and combine the interaction of scientific areas of many disciplines.

Common engineering practice for a new design of e.g., a bridge, is to use previous experience and test the newly designed components by experiments. Corrosion tests can be made under more severe conditions than expected to ensure that corrosion will not become a problem during the desired lifetime of the bridge. Supporting structures can be subjected to much larger stress than they will be subject to in the bridge. Such methods alone are not sufficient for a repository that must retain its integrity and keep the waste isolated over a 100 000 year time scale and more. An additional difficulty is that the heterogeneity of the system, the rock mass, can only be explored by a limited number of sparse boreholes before selecting a site. When the site is selected excavations and tunnels will give much more information. Different types of models and supporting investigations and observations must be used that make predictions credible without ever having the possibility to ever test the outcome.

There are some essential requirements for the models that are used to make predictions of processes far outside the time span and space that can be tested. Violating these will invalidate the model.

Models must not violate accepted laws of nature

A model must be rejected if it e.g., implies that that energy or matter is either created or destroyed, if heat spontaneously moves from a low temperature location to a high temperature location, etc. This is formulated in the laws of thermodynamics. The violation of these rather obvious demands can sometimes be subtly hidden, and violations may not be noticed in modelling of e.g., complex chemical and biochemical modelling. We will later exemplify some such cases.

Models must not be incompatible with common, well-documented and accepted observations

This again seems to be an obvious demand, but violations are sometimes not noticed (early) in multi-disciplinary models and modelling.

Models in different disciplines describing the same object from different viewpoints must be commensurable

In the present problem, the site description (modelling) of the geology, hydrology, mechanical properties and chemistry and biochemistry must not be incompatible.

Solutions of the model equations must be correct and robust

This is an obvious demand but the ways to achieve it is not always obvious, especially when different coupled processes have very different time constant and act over different space scales.

Conceptualisation and levels of model abstraction

In the present case, we try to describe and quantify how water flows and how solutes migrate in fractured rock. The simplest conceptual model could be that the rock is seen as a homogeneous porous medium and that water flow is driven by the pressure gradient. The next level might be to account for the differences in conductivity in different locations. A further level may be to include the observations that the conductive regions are limited to the fractures and to include the fracture network structure in the model. A further step may to include the variability of the conductivity within and between the fractures etc. Each level demands new and more detailed data and leads to a more complex model that needs larger computing effort and more detailed data that is more costly to obtain by qualified manpower as well as money. The gains in model accuracy will have to be balanced against the larger effort. The added gains will probably diminish with each added detail and a decision will be reached that the model and data suffice for the task at hand.

Approaches to solve the model equations

The sub-models that describe the different processes and mechanisms are often highly non-linear and rely on stochastic distributions of the input parameters. The output of the simulations results in a large number of possible cases, often covering a wide range of outcomes with large variations. Different numerical methods are used to solve the systems of algebraic, ordinary and partial differential equations by which the conceptual models are described mathematically. Sometimes it is possible to derive analytical solutions to some simpler models that describe how some key parameters of processes impact the results. Analytical solutions often facilitate understanding and gain insights into how changes in different parameters influence the results. Monte-Carlo based methods sometimes can solve the equations faster and more efficiently. Together, by using the different tools in the toolbox by different researchers and comparing the results, the probability that uncertainties in model concepts are brought out, highlighted and that errors are detected increase. The Äspö Task Force and the Decovalex project (Birkholzer et al. 2019) are examples of how this approach is used systematically.

Comparison of alternative models and model assumptions

One example of such a study is found in Selroos et al. (2002) in which three different models were used to simulate the same performance measures for a hypothetical KBS-3 repository design for nuclear waste in a site with 18 fracture zones near the Äspö hard rock laboratory in Sweden. Data on fracture zones and rock mass from deep boreholes in the site and from the laboratory were supplied to three teams that used different models, DFN, ChN and a stochastic continuum SC model. These data were translated to input data by the teams themselves. They predicted the performance measures distributions of up to 945 different canister positions using up to 34 realisations for the stochastic flow realisations. The mean results of the three performance measures differed by up to one order of magnitude between the models. Early and late arrival measures were also compared. The paper concludes that " The three modelling approaches predict similar median travel times and median canister fluxes, but dissimilar variability. The three modelling approaches also predicted similar values for minimum travel time and maximum canister flux and predicted similar locations for particles exiting the geosphere."

Some further observations and experiments that can be used to give support of the models of flow and transport in fractured rocks

The water composition in the rock, both mobile in the fractures and stagnant in the rock matrix measured in the deep boreholes in site investigation can be and is used to support several different aspects of processes, mechanisms and models. The present-day water composition has resulted from flow and solute transport over very long times during which the rock has been subject to changing rock stress, temperature, chemical reactions such as mineral dissolution and precipitation etc. It is not possible to reconstruct the main evolution of the conditions with reasonable confidence for more than the last 10 000 years after the glaciation ice receded from north Europe. The (still ongoing) land-rise has changed the extent and shape of sea and lakes above the rock. The infiltrating waters have changed in salinity over time and their imprints are seen in the water compositions in the waters in fractures and matrix, see e.g., Salas et al. (2010).

Water with high density collects in the lower portions of the rock and is not readily lifted and displaced by less dense meteoric water infiltrating from higher locations. The denser water can flow more or less horizontally in permeable regions but will resist lifting against gravity. This is supported by the presence of the accumulation of salt and geo-gases such as helium that has accumulated since the formation and solidification of the crystalline rock by uranium and thorium decay over billions of years. The observations that the geo-gases remain and are lost very slowly at depth, support the model predictions that at repository depth, released solutes migrate very slowly upward (Neretnieks 2013).

Matrix diffusion and flow wetted surface

Matrix diffusion is the by far most important mechanism that retards solutes as they are carried by the seeping water. Sorption in the matrix further enhances the retardation. It had not been much studied until the early 1980's and is still sparingly treated in textbooks. Several questions have been raised over the years about the phenomenon as such and its impact on PA. How far from the surface of the fracture into the matrix is the pore system connected and how far can the solutes penetrate into the rock that is compressed subject to the rock stress at repository depth? Will mineral precipitation and alteration clog the pores over time. Are results from laboratory experiments representative of stressed in-situ rock. These and other questions have been addressed in different ways since early 1980's. Skagius and Neretnieks (1986a,1986b, 1988) made numerous through diffusion measurements using non-sorbing and sorbing solutes in the laboratory, developed a technique to measure matrix diffusion coefficients by electrical conductivity and also measured the effect on compressed samples to stress levels comparable to repository depth. Birgersson and Neretnieks (1990) in an in-situ experiment at Stripa at 360 m below the ground, let non-sorbing tracers diffuse into the rock under natural stress for 3.5 years, over-cored the rock and analysed more than 2000 small samples in three directions from the narrow injection hole. Tracers were found in decreasing concentrations up to 30 cm from the source. Ohlsson et al. (2001) developed a method by which the electrical conductivity method of Skagius and Neretnieks (1986b) could be used in boreholes. Löfgren and Neretnieks (2003) and Löfgren (2007a, 2007b) used it in the site investigations at Oskarshamn and Forsmark. André et al. (2008a, 2008b, 2009) developed the method further so that direct current could be used to propagate sorbing and non-sorbing tracers over distances of several tens of cm long cores. These and other investigations showed that the rock matrix porosity is connected over at least decimetres to meters in the rock investigated and that it is accessible to solutes under the undisturbed conditions at depth.

Coupled processes and modelling THMC(B)

There is considerable interaction between how temperature, water flow, stress, chemistry and even biology, summarised by the acronym THMC(B), affect the evolution of the processes in crystalline rock of importance for PA. For example, changes in temperature will affect water flow by changing viscosity and density of water. Temperature will affect chemical equilibria, reaction rates, diffusion, and rock stress. Rock stress changes will close or open fractures and shear them and change their transmissivity. Microorganisms catalyse chemical reactions e.g., redox and pH. These are just a few interactions that must be considered. The international Decovalex project, which has been running for more than 25 years has addressed a number of such interactions related to PA. Seven different tasks have so far been addressed and analysed (Birkholzer et al. 2019).

Correlations

Some correlations e.g., depth dependence of conductive fracture frequency and fracture transmissivity can be assessed in the deep boreholes in the site investigations. Other correlations are suspected to exist but may be difficult to measure and to quantify. It is likely that there is some correlation between fracture size and transmissivity, but it is difficult to quantify the correlation by experiments because of the heterogeneity of the rock. Especially the larger fractures and fractures zones with their complex internal structures are poorly characterized and it is likely that they have a dominating impact on flow and on long-range channelling.

Hyman et al. (2016b) in a series of simulations found that a correlation between a fracture size and its transmissivity leads to earlier breakthrough times and higher effective permeability when compared to networks where no correlation is used. In a recent numeric study of channelling, it was concluded that "Fracture size parameters are very challenging to derive from the field data but are crucial to the model behaviour. In our study we neglected this large uncertainty. We showed that even with this geometrical constraint it is challenging to reproduce the effect of channelling observed in in-situ conditions by using simplified homogeneous models. Further research on this problem is therefore "required" (Bym and Follin 2019).

2.2.7 Some examples of interpretation of experiments by models and some uses of models that have been challenged

Use of the advection dispersion models, ADE, to extrapolate to larger distances

The use of the ADE requires that a value of the dispersion coefficient can be determined. Field experiments over distances over five orders of magnitude show that the hydrodynamic dispersion coefficient D_h increases in proportion to observation distance x, and thus is not a materials constant

(Gelhar et al. 1992). On the other hand, the ratio $\frac{D_h}{ux}$, where u is water velocity, is found to be much less dependent on x and mostly has values between 0.01 and 1, cantering around 0.1. Modelers using the ADE then would choose a D_h such that $\frac{D_h}{ux}$ is within this range and then solve the ADE. However, this equation requires that D_h is independent of distance. This approach thus violates the validity of the equation for use to simulate dispersion in fractured rocks.

Interpretation of in situ tracer tests

Another example is the interpretation of some in-situ multi-tracer experiments between boreholes at Äspö, i.e. the TRUE-1 experiments. In one interpretation of the BTC's the underlying assumption was that all flow and transport took place in one fracture, the so-called Feature-A (Winberg et al. 2000). Fair fits to the BTC's were obtained. However, the model assumption that the Feature-A is not a ”leaking aquifer” needed the fracture aperture over the 5 m distant boreholes to be about 3 mm on average. So large mean apertures over at least 5 meters have not been reported in the drifts at Äspö at depths of 400 to 500 meters. Later over-coring of a small area near one of the boreholes found a small patch with at most about one mm aperture (Byegård et al. 2017). Another team that studied the experiment showed that Feature-A belonged to a conducting fracture network with similar transmissivities as Feature-A, with mean conducting fracture distance of about 0.5 m. This implies that the pumped flowrate most likely, to a large extent, had been supplied by the fracture network between the injection and pumping hole and not supplied only from flow through Feature-A. This implies an overestimate of the fracture aperture by about a factor 30. It also implies that the performance measure, flow wetted surface to flowrate, FWS/Q, in the experiment is overestimated by the same amount. This is supported by the sorbing tracer results. The sorbing tracers were also used in the between-hole experiments. The BTC's of these tracers could be *predicted* using only laboratory sorption and matrix diffusion data together with the measured flow wetted surface of the network surrounding the experimental location together with the pumped flowrate i.e., the FWS/Q for the network case (Neretnieks and Moreno 2003). It may be noted that for tracers that are retarded so much that the water travel time can be neglected it is sufficient to have information of the FWS/Q, to predict the residence time distribution of the sorbing tracer. This is true also for non-sorbing solutes under conditions when these are retarded by matrix diffusion in moderately slow flowpaths in the network. This is why this entity FWS/Q is so useful as a performance measure in PA.

In most evaluations of tracer tests in between-hole pumping experiments it is inherently assumed that the surrounding network does not lead to a “leaking aquifer” in the target fracture. In a compilation of a number of tests in Sweden in which the surrounding network is not considered at all, the apertures were also found to be about 30 times larger than what the “cubic law” gives (Hjerne et al. 2009). These aperture data are therefore questionable by the same reason as the 3 mm aperture in Feature-A in the TRUE-1 experiment, see also Section 2.2.8.

Closing of fractures by chemical dissolution of stressed crystals

Another example of interpretation difficulty due to not considering an important mechanism and possibly also of violation of a thermodynamic principle is from experiments in which a fracture is gradually closing caused by dissolution of the minerals on the fracture surfaces in contact, subject to compression of the fracture. In this example water seeps slowly in the fracture and carries away the dissolved minerals. In the experiments with the mineral Novaculite, different temperatures were used, and closure was simulated based on laboratory measurements of mineral solubilities and rates. To obtain better fit it was assumed that low stress does not affect the solubility but that there is a sudden increase in solubility at a certain stress level. This assumption had been used in several previous publications. However, it violates a thermodynamical principle that the stress effect should give a gradual solubility increase with pressure, not a sudden jump.

One finding in evaluation of the experiment was that the laboratory derived dissolution rates had to be adjusted by factors of tens of thousands to one million at the different temperatures (Bond et al. 2016). An independent interpretation of the experiment suggested that there may be an additional sink for the dissolving mineral than that of the seeping water carrying it away. The compression force is concentrated on the “few” mineral grains at the contact between the two sides of the fracture. These grains have higher solubility than the grains inside the rock matrix in which the compression force is much less because it is distributed evenly over all grains. The crystals stressed between the two fracture surfaces dissolve, the dissolved mineral, now a solute, diffuses into the porous rock matrix and precipitates on the less stressed crystals in the matrix. When account is taken for this effect the "adjustment factor" for the dissolution rate decreases many thousandfold (Neretnieks 2014). This is still by far not enough for agreement between experiment and model. However, this example is meant to illustrate two things. One should not invoke some artefact that violates thermodynamic principles to get a better fit. The need for a large correction factor suggests that some additional mechanism or process may need to be sought.

Some further comments regarding model credibility

"Models cannot be validated; they can only be in-validated" is a common opinion in the discussions on the subject. However, we have to trust that models can be and have been credibly used to interpret experiments and observations to give arguments to those who will have to make decisions based upon the models and modelling and on other softer arguments. This is marred by the complexity and interplay of different processes and mechanisms. It is especially difficult when modelling heterogeneous materials that can only be described by stochastic data that are difficult to obtain and to correlate.

The credibility of the arguments will increase when they have been subjected to scrutiny by independent researchers and teams who have not raised serious objections after thorough deliberations. The credibility increases the longer time the arguments have been discussed in the open literature and the more the "models" have been used by independent researchers.

It is important that publications in peer reviewed journals are stimulated and made and that underlying research results and documentation is made easily and generally available.

2.2.8 Observations of channelised flow and transport in fractured rocks

This section briefly describes a number of experiments and observations of flow and transport in fractured rocks performed over the last 40 years that have influenced the modelling used in PA. More details of these and other experiments can be found in Bear et al. (1993).

Experiments in Stripa URL

The Stripa underground laboratory in Sweden was the first such laboratory solely dedicated to research of various aspects of processes needed for PA of a high level nuclear waste deep in crystalline rock.

An in-situ experiment to study migration in a single fracture in granitic rock was performed to study flow distribution and tracer transport with both sorbing and non-sorbing tracers in fractures. Figure 2-7 shows layout of the site and water and tracer flowrates to the drift. Both water and tracer inflow to the drift was unevenly distributed and channelised flow was evident (Abelin et al. 1985).

Subsequent over-coring of the fracture up to and including the tracer injection location revealed that the sorbing tracers had mostly been diverted into fractures that intersected the target fracture near the injection location.

Tracer tests have also been made for flow into drifts. In a specially excavated drift in the Stripa URL a 75 m long drift with a 25 m crossing drift was covered with 375 plastic sheets by which water was collected. The location is 360 m below ground surface in water saturated rock. The flowrate distribution in the sheets is shown in Figure 2-8. One sheet carried 10 % of all water, 12 sheets carried half of the flow. Nine different tracers were injected in three different vertical boreholes in transmissive locations ranging from 11 and 41 m above the drift. The tracers were collected over several years and were used to evaluate transmissivities, flow porosities, RTD, dispersion and retardation caused by matrix and/or other stagnant water zones (Abelin et al. 1991a, 1991b). Figure 2-8 shows the drift and injection holes, the water inflow distribution and the plastic sheeted drift.

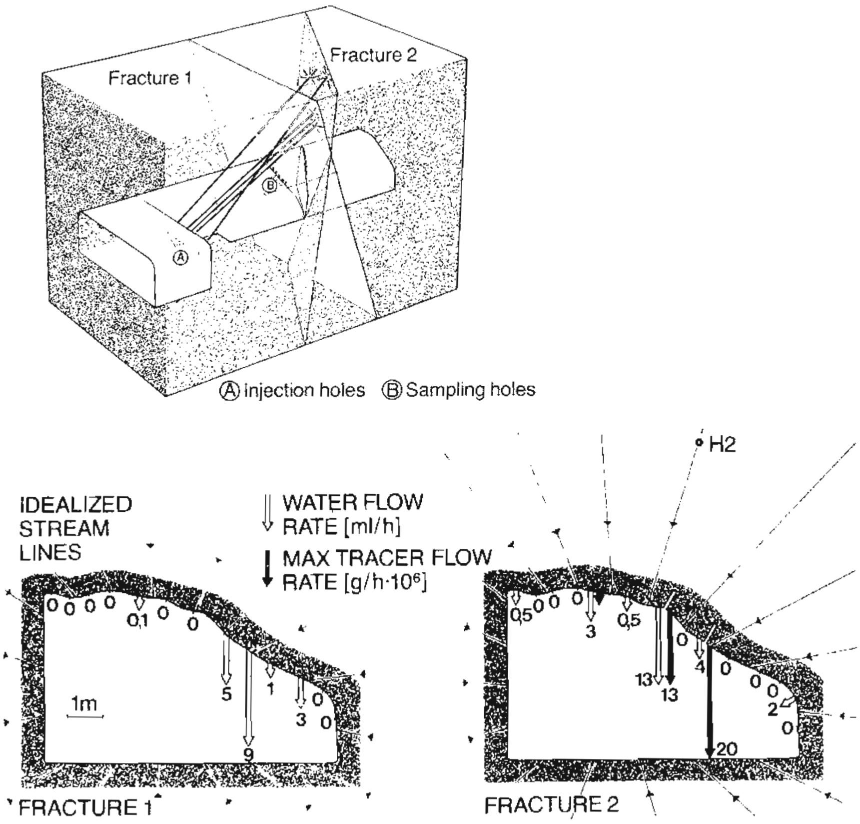


Figure 2-7. *Layout of the site and water and tracer flowrates to the drift. H2 is the tracer injection location (Abelin et al. 1985). Upper figure shows 3D situation, while lower figures show schematics of the intersections of the two main fractures with the drift.*

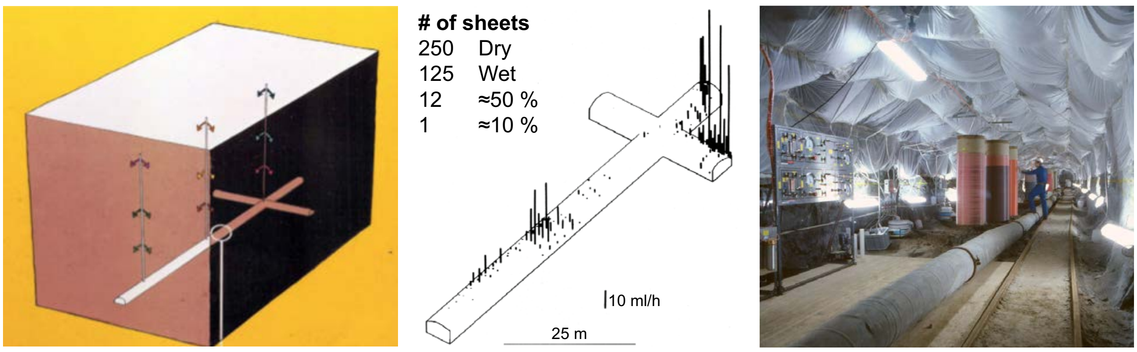


Figure 2-8. *Water inflow locations and flowrates in the 3D drift in Stripa URL. Nine different tracers were injected in boreholes above the drift. Left figure shows 3D layout including injection boreholes, central figure shows water inflows and right shows a photograph of the drift during sampling.*

Several hundred BTC's were evaluated. Most showed multiple peaks. The channel network was complex. Many tracers that passed from an injection location in the right-most injection hole to collection sheets to the "left", did not mix with the tracers coming from the left-most injection hole and emerged in sheets to the "right". Surprisingly the high flowrate locations in the right arm of the cross did not carry tracers from any of the nine injection locations above the drift. The tracers were mostly found in the mid-section of the drift. A detailed description of the tracer migration is given in (Abelin et al. 1991a). Possible causes for such behaviour are discussed by (Black et al. 2016).

In seven locations the collected water was tested for tritium, which was found in one sheet with a concentration of 6 tritium units, TU. This implies that it must be "bomb" tritium from the atmospheric hydrogen bomb tests in 1960's and that it has travelled the 360 meters from the surface down to the drift in less than 30 years with very little dilution and retardation. Considering the low rock permeability and flowrate this must be an exceptionally fast channel. One of the tracers injected 37 m above the drift was found in large amounts in a location 150 meters distant in another drift in the URL. Much less of this tracer was found in the collection sheets. This suggests the presence of another exceptionally fast channel. This was deemed to have been caused by blasting of the drifts.

Visually the 100 main mapped fractures did not seem much different. The larger flowrates were found in the parts of drift that had most fracture intersections. Tracer inflow locations showed that some of the tracers cross through the fracture network without mixing with each other. This has been interpreted in a recent study that revisits the data that the channel networks are very complex. Black et al. (2016) revisited the data and also found that there is more to it than what is expected from just a simple channel network. They conclude that: "… the commonly observed feature, 'compartmentalization', only occurs when channel density is just above the percolation threshold. It is suggested that compartments and skin are observable in the field, indicate sparse channel systems, and could form part of site characterization for deep nuclear waste repositories."

Birgersson et al. (1993) describe an experiment that studied flow and solute transport in a fracture zone and adjacent "averagely" fractured rock in the Stripa underground research laboratory. The experimental site is located in granitic rock at a depth of 385 m below the ground surface. A 50-m-long drift with a diameter of 3 m was excavated. The drift was intersected by a 6-m-wide fracture zone. The upper part of the drift was covered by 150 plastic sheets in which water was collected. The water in the lower part of the drift was collected in sump holes. Different tracers were injected in seven locations at distances between 9.5 and 25 m from the drift. More than 50 % of the water was found in one sampling area in the zone, and more than 90 % emerged in eight sampling sheets. The recovery of the tracers was concentrated in a few sheets. Most of the flow preferentially takes place in a few paths. Non-sorbing tracers were used to estimate porosity. The flow in the fracture zone is obviously highly channelised flow. Figure 2-9 shows the arrangement of the test and the flowrate distribution in the collecting sheets.

Abelin et al. (1994) in detail characterised the aperture distribution of a fracture into which 2 m long boreholes were drilled from the drift in the fracture plane. The fracture was tightly sealed off at the face of the drift. Using a specially designed "multipede" packer system that allowed simultaneous injection in 5 cm sections of traced water and collection in another "multipede" at 2 m distance in the same fracture, tracer tests were performed. Distinct channelling was observed. The tested fracture had been selected from about 100 other candidates because it was not seen to be intersected by other fractures and seemed to be more "prominent" than other fractures seen on the face of the drift. Detailed high-resolution photographs were taken of the fracture from inside the holes. The multipede packer and an example of opening distribution are shown in Figure 2-10.

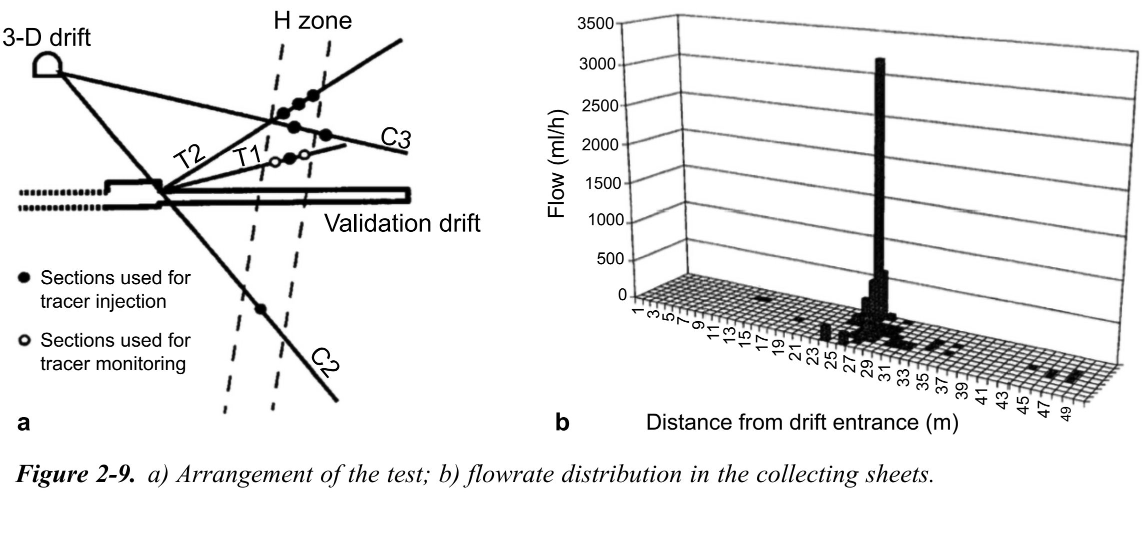


Figure 2-9. *a) Arrangement of the test; b) flowrate distribution in the collecting sheets.*

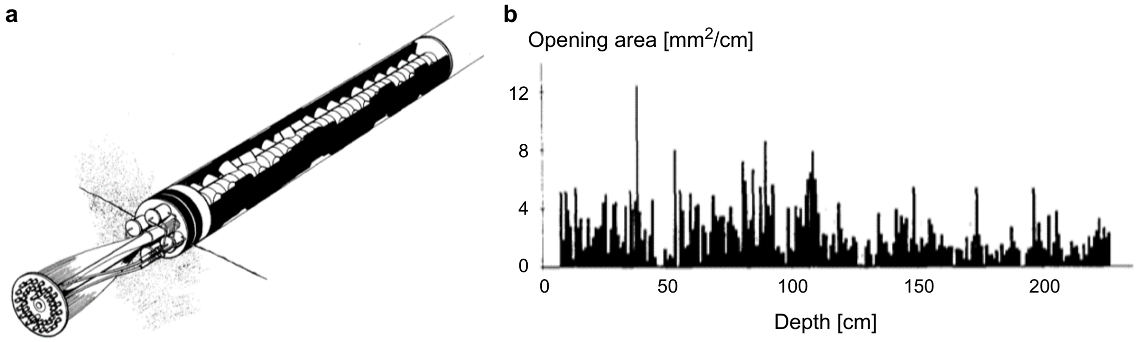


Figure 2-10. *a) Multipede packer; b) an example of distribution of openings derived from photographs taken inside the hole.*

In the between-hole tracer tests only a fraction of the tracer mass was recovered. Using UV light to look for the fluorescent Uranine tracer it was found emerging in 16 narrow spots on the face of the drift at distances up to 4 m from the injection holes. This suggests that there is a "dense" channel network in seemingly minor fractures intersecting the chosen experimental fracture.

TRUE-1 experiments

The between-hole tracer experiments at Äspö URL are reported in Winberg et al. (2000) and Byegård et al. (2017). Distances between injection and pumping holes are around 5 m. Assuming an "isolated " fracture the mean mechanical aperture of Feature-A by tracer tests is found to be about 3 mm. The cubic law suggests around 0.3 mm and less. Boreholes in the rock surrounding Feature-A found transmissive fractures about every 0.5 m. When modelling the fracture as a leaking aquifer with many fractures intersecting Feature-A (see Figure 2-11c,d), a mechanical aperture of less than 0.3 mm is needed. The notion of a leaking aquifer is also supported by the good prediction of the BTC's for sorbing tracers, which needs only data on FWS/Q (Neretnieks and Moreno 2003).

The experimental layout and some data are shown in Figure 2-11. Boreholes drilled from a drift intersect a number of fractures (Figure 2-11a). One fracture, Feature-A is used for between-hole experiments for non-sorbing and sorbing tracers. The transmissivity distribution in all boreholes is shown in Figure 2-11b. Mean distance between transmissive fractures in the boreholes is about 0.5 m. Flowpaths and distances in the experiments are shown in Figure 2-11c. Figure 2-11d is an illustration of a leaking aquifer.

Other between-hole tracer tests in Sweden

Hjerne et al. (2009) evaluated 74 between-hole tracer tests from different sites. Tracer residence times were used to evaluate mean mass balance fracture apertures, (void volume/area). These were found to be 10 to 100 times larger than cubic law apertures (Figure 2-12). This is a surprisingly large difference not observed in experiments in which there exist no intersecting fractures, Witherspoon et al. (1980). Assuming leaking aquifers in the same way as for the TRUE-1 experiments, the mass balance aperture can well be equal to the cubic law aperture, see also Section 2.2.7 regarding interpretation of tracer tests and Appendix in Neretnieks (2018).

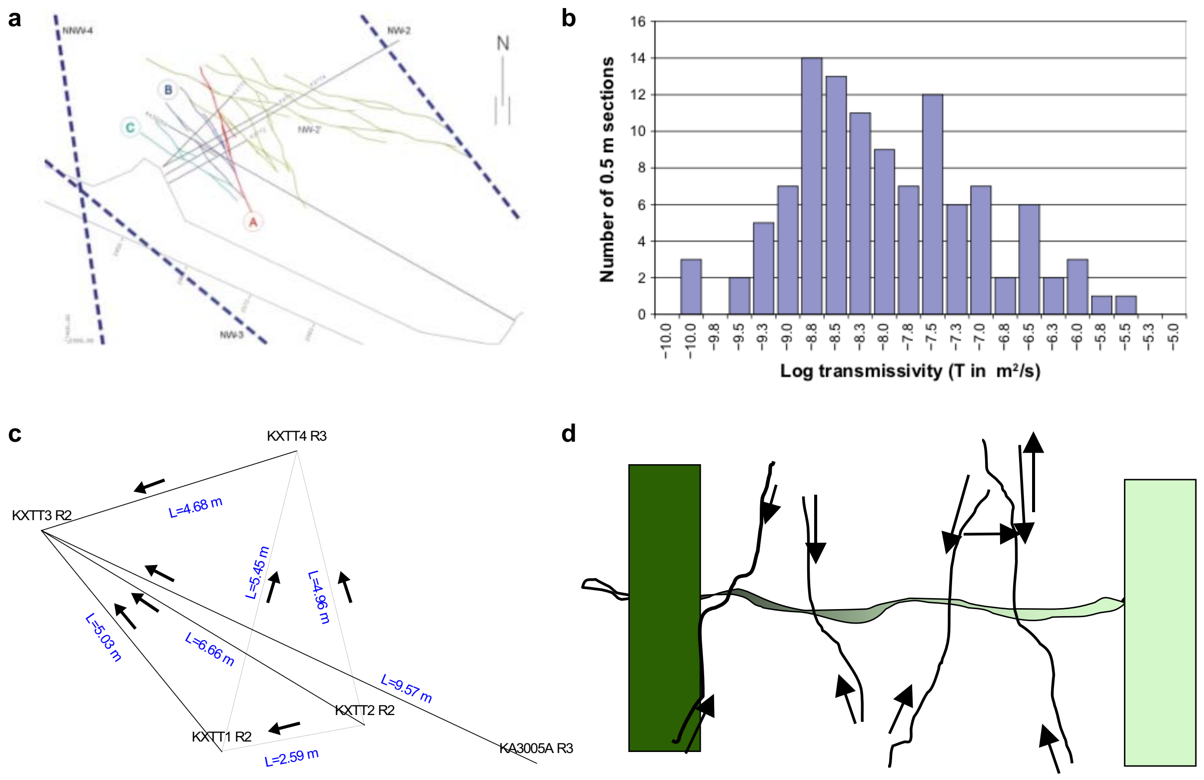


Figure 2-11. *Experimental layout and data for tracer tests in Feature-A: a) boreholes and fractures; b) transmissivity distribution in all holes in site; c) flowpaths; d) the leaking aquifer. Water is pumped in the right-hand hole and a small flow of tracers are injected in the left-hand hole. Most of the pumped water enters Feature-A from the fracture intersections (vertical black arrows).*

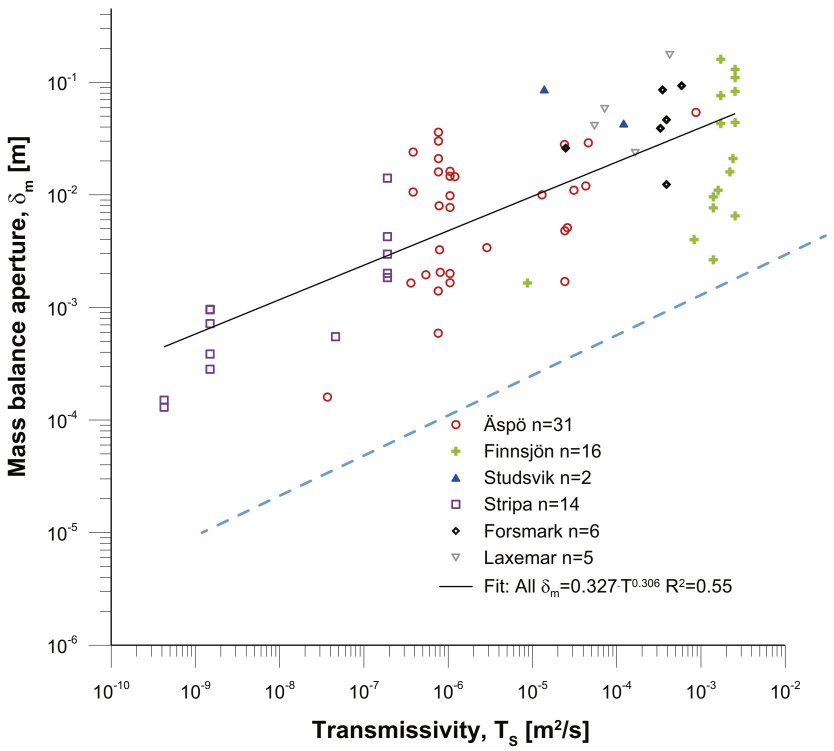


Figure 2-12. *Mass balance aperture as function of transmissivity from different Swedish sites. Dashed line shows cubic law relation (Hjerne et al. 2009).*

2.2.9 Some other observations of water inflow rates and channelling

Inflow rates at SFR

In a site, which hosts the low and intermediate waste repository at Forsmark, Sweden, after excavation, flowrates were measured. Figure 2-13 shows inflow rates to the low and intermediate waste repository at Forsmark (Tsang and Neretnieks 1998).

Q-drift at Äspö

In an 80 m long drift at Äspö 166 inflow locations with low flowrates were identified by infrared photography, IR (Neretnieks et al. 2017). Only a few spots could be seen as dripping. All other locations were dry to the eye because the water evaporated when it emerged into the ventilated drift. The size and temperature of the cooled spots seen on the IR photos was used to calculate the seepage rate on each spot.

A histogram shows the flowrate distribution in Figure 2-14. The three rightmost spots were seen by the eye to be seeping and flowrate could be measured. The widths of the spots were assessed to mostly be on the order of a few dm. The seepage rates are similar to those at the Stripa 3D drift, described earlier.

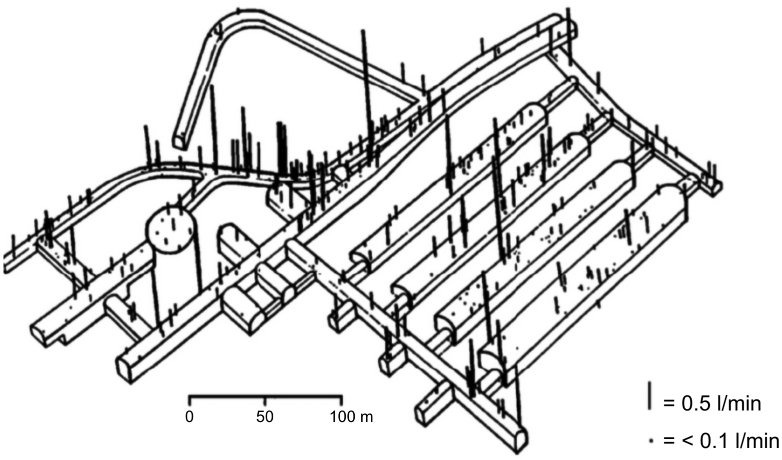


Figure 2-13. Inflow rates to the low and intermediate waste repository at Forsmark.

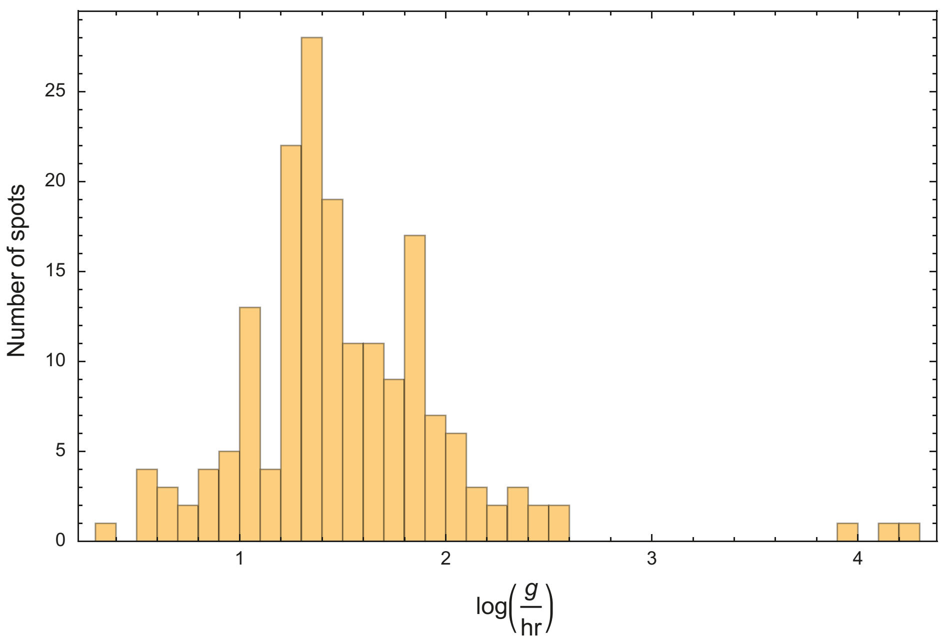


Figure 2-14. Seepage rate distribution in the Q-drift.

Bolmen and Kymmen observations

Figure 2-15 shows two pictures of channelling in a tunnel in crystalline rock not long after excavation. Water with dissolved ferrous ion seeps out from narrow channels. When it meets oxygen in the air it is oxidised to ferric iron, mediated by microorganisms, and forms the reddish precipitate. These pictures are from the Bolmen tunnel in south Sweden. Similar patterns were seen over several kilometres, sometimes as clusters, sometimes as an isolated spot at irregular intervals. Not seldom the channels were seen at fracture intersections.

Similar observations were made in the Kymmen tunnel, a drilled tunnel, Palmquist and Stanfors (1987). Here, over a distance of 5.5 km, also inflow rates were measured. Seepage in 169 locations in crushed zones and in 189 locations outside of zones were assessed. The flowrates in the zones were considerably larger than outside zones. The inflow rates were classified in five categories ranging from less than 0.01 l/min to more than 6.5 l/min.

Comments on and implications of channelling observations

Flowpath topology and flowpath properties in fractured crystalline rocks are very heterogeneous. A considerable fraction, perhaps even the majority of fractures have open narrow conduits that connect to conduits in other fractures. The density of channels is much larger than the density of transmissive fractures found in boreholes, more than tenfold. The flowrates in the channels vary over a very wide range. The channel widths vary from less than one cm to several dm. Fracture intersections can be quite conductive and contribute to form long high-flow paths over long distances. Physical channels can have formed by dissolution of more soluble minerals. This is especially pronounced in fracture zones in which the rock has been crushed to fragments with a wide particle size distribution.

Fracture zones are considerably more transmissive than "single" fractures and will facilitate the formation of long-range pathways with high flowrates and low retardation capacity FWS/Q.

It is not known if the fast pathways observed belong to the same probability distribution, pdf, as the other flowrates in e.g., Stripa 3D and Q-drift at Äspö or if they just belong to another distribution. This may be an important question to explore. It is hypothesised that flow along fracture intersections can promote long range channelling. The chance that a narrow borehole (less than 10 cm diameter) drilled into the rock will intersect narrow channels (less than one cm diameter) as those observed in drifts and tunnels is extremely small. Their frequency may be considerably underestimated.

The above and similar observations in tunnels and at the Äspö URL suggest that most of the flow takes place in flat narrow channels in the fractures and that the flow is very unevenly distributed in the rock. The channels can form connected flowpaths over long distances and can carry solutes "rapidly" compared to that of the mean flow. The understanding and quantification of channelling and its impact on PA simulations is an important task. This has been specifically pointed out by the Finnish authority, STUK (2015).

Figure 2-15. *Channelling in the Bolmen tunnel.*

The channel network concept has been implemented in the channel network model, Chan3D, (Gylling 1997, Gylling et al. 1999), in which the flowpaths in the fractures are a priori modelled as channels forming a network in much the same way as the flow in the fractures in the DFN are treated in practical implementations. One main difference is that the fracture intersections themselves are treated as potential channels and that adjacent to the channels stagnant water zones exist that can also promote solute retardation.

2.3 Use of flow and transport models in performance assessment

2.3.1 Role of host rock in performance assessment (PA)

The host rock and underground are to provide isolation from the surface environment and to ensure that the waste canisters are not damaged mechanically, thermally or chemically over very long times. Should a canister be breached, the rock will also retard the migration of radionuclides towards the biosphere.

Flow and transport models play an important part in the Performance Assessment of a site. They contribute to increasing our confidence that the "structure" of the network of water conduits adequately describes where and how much water flows in the rock around the repository. Information on water fluxes has many uses. The rock volume hosting the repository must provide enough space to host the waste canisters in locations with low enough flowrates. Negligible amounts of corrosive agents such as oxygen and sulphide must be ensured to limit corrosion to levels not endangering canister integrity. Low flowrates of very fresh water that could mobilise and carry away the swelling smectite clay surrounding the canisters must be ensured. A number of other uses of flow and transport modelling in PA of the Forsmark site can be found in (SKB 2008a, 2008b, Follin 2008, SKB 2011) and Selroos et al. (2014a, 2014b).

If a canister is damaged and radionuclides escape to the seeping water in the rock fractures, the flow and transport models are needed to estimate where, when and with what rate nuclides enter the biosphere. Models must be compatible with observations of present-day seeping water compositions and composition measurements of ancient waters found in the pores of the rock matrix. They should also be commensurate with semi-quantitative observations of flow compartmentalisation, channelling, matrix diffusion effects in the field, and other observations that might influence the repository performance. In short, models must be credible, and the scientific community must have confidence that they are reasonably accurate and fit for the purpose.

Some sub-models may be quantitatively tested against field observations. One example is the effect of matrix diffusion, which has a dominating impact on transport of all solutes. The spatial scale of interest is on the order of tens of meters or less around each flowpath, even for the hundred-thousand-year perspective. The process is well founded in theory. Laboratory and field scale tests are possible to perform. Predictions and outcomes for different sorbing and non-sorbing nuclides mostly agree well (Meng et al. 2020).

Some model outputs cannot be well supported by field observations, even when considering the stochastic nature of the underlying processes, mechanisms, and observables. This is especially pronounced for predictions of future events and to regions at large distances where no measurements can be made for direct comparison with model outputs. One example is how to assess the rock hydraulic properties in the region between sparse boreholes at several km spacing. Nevertheless, it can often be credibly argued that based on indirect observations, general knowledge, and experiences from other sites that the model is fair and reasonable for the purpose.

Model consistency, both internally and with other relevant models, is also important. It was stressed in the development of the models of the Forsmark site for PA usage that all models must be consistent with each other so that e.g., the rock mechanics model must fit in with and not contradict the hydraulic and the chemical model etc.

2.3.2 Short summary of Swedish PA work

Here follows a very short summary of how the problem was approached by the Swedish Nuclear Fuel and Waste Management Co, SKB.

The work started in late 1976 when the organisation was formed after a new law demanded that it should be shown by the utilities that an absolutely safe repository could be built before the two new reactors under late stages of construction in Sweden could be started up. First studies of a deep geologic repository had already commenced earlier. The new-born SKB sought and engaged domestic and international experts on matters related to geology, hydrology, chemistry, solute transport, clay science and other relevant areas to form a small interdisciplinary team to work intensely on the problem. Consultants from universities, research organisations and the like were engaged to make different studies in the multi-disciplinary project under coordination of SKB. A report based on reprocessed vitrified waste was published after one very intensive year. Various reviewers and an international group convened by IAEA in Vienna reviewed the report, which was then submitted to the Swedish authorities and government, who accepted that the findings complied with the law. The two reactors were permitted to commence operation. An alternative repository design, not needing reprocessing was studied and had to be completed before two additional reactors would be allowed to be built and to operate.

Designs based on spent fuel in a copper canister were presented in 1978, KBS-2, and 1983, KBS-3. This method was also found to comply with the law by the government. Over the years the designs have evolved but are essentially based on the original KBS-3 concept.

Field experiments in and between boreholes started early and an underground research laboratory, URL, in Stripa in Sweden opened in 1978 in cooperation with Lawrence Berkeley Laboratory in the USA. It operated until 1992. Another URL at Äspö near Oskarshamn was built 1990 to 1995. Here a large number of experiments by Swedish and international groups have performed different experiments resulting in numerous peer reviewed publications and reports. Extensive field investigations of potential sites at Forsmark and Laxemar by Oskarshamn over more than five years were made. In Forsmark, the site later selected for the repository, 25 cored boreholes were drilled, of which 19 were over 500 m long and 9 over 1 000 m. Altogether 16 km of drill core samples were extracted. Approximately 800 scientific reports were produced during the site investigation.

Many of the journal and conference publications as well as the SKB reports, which are available to the public have been read, scrutinised, and referred to by researchers in many countries. This feedback is and has been immensely valuable as it increases the confidence that the work and conclusions are credible and can be used in PA. Confidence building has also been attained by the reviews of the SKB PA reports by groups from among others, IAEA, US Academy of Science, the SKB Internal international "SIERG" group that continuously followed and critically reviewed and commented the progress for a decade as well as the international "INSITE" group of the Swedish radiation safety authorities. Many countries have supported their own teams in the ”SKB Äspö Task Force", which organises special forums to allow international specialists and modelling groups to collaborate on selected issues that are important for the final disposal of radioactive waste. This has continuously worked since early 1990's. Nine specific tasks have been performed and published. A similar international collaboration, DECOVALEX, was initiated by the SKI, the Swedish Nuclear Power Inspectorate, which has specifically addressed the impact of coupled processes for more than 25 years. Both the Äspö task force and the DECOVALEX have addressed and emphasised the question of how well the models can interpret and describe experiments.

The application to build a repository at Forsmark was submitted to the authorities and government in the spring of 2011. The application builds heavily on the vast amount of work and findings of the above-mentioned work performed over 35 years.

The research on the long-term safety of the repository has continued and will do so at least until it is finally closed and sealed. So far, the latest ten years work has not revealed any seriously questionable issues in the safety reports of SKB (2011) and Posiva (2012).

The Swedish government is expected to make a decision soon to allow a repository to be built at Forsmark. In Finland the same type of repository has been approved and excavation started in 2014.

Turns, twists and surprises along the road to the present description of the processes on which PA is built

This section describes some of the changes of conceptualisation of the important processes in the modelling of flow and transport in fractured rock relevant to PA and its modelling.

In the first phases of hydraulic modelling there was a fair consensus of the use of Darcy's law to model flow in the rock mass and that one could describe the flowrate based on the concept of an "representative elementary volume", REV, over which flow in the fracture in the REV could be averaged. Then the flowrates and solute transport over larger volumes could be treated as in a porous medium with some (not dramatic) variations in different locations. The rock could then be treated as an equivalent continuous porous medium. Standard numerical codes could be used.

The ideas of fractal properties of natural systems were increasingly accepted in the 1980's and it was realised that the concept of an REV would not be compatible with the observed fractal nature of the real world. The old used and accepted "textbook truths" had to be modified to comply with the new observations of heterogeneities of flow and transport in fractured rocks. Fracture network models began to emerge in which the network could also be made to have fractal properties.

2.3.3 Some examples of flow and transport models used in PA

In the examples below a DFN based flow and solute transport model was used in the simulations. Details of the models and results can be found in SKB (2010) and SKB (2011).

Groundwater fluxes

Figure 2-16 shows an example of how the average simulated water flux at Forsmark changes with depth. Such information can be obtained from the flow model and used as a starting point to calculate the influx rate of e.g., oxygen, which impacts on the groundwater chemistry, engineered barrier system evolution and radionuclide transport to the biosphere. This information can then be used to assess the radiation exposure to man. Network models can generate very detailed stochastic examples of flux patterns.

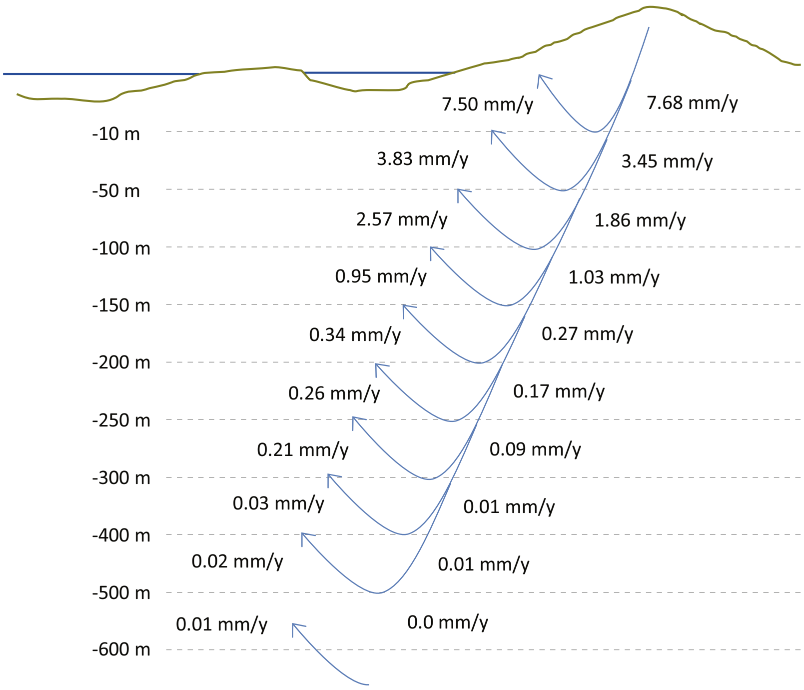


Figure 2-16. *Average vertical groundwater fluxes over a 5 km × 5 km surface area at different elevations at the Forsmark site. Since the surface discharge represents a mix of groundwaters with different advective histories, the upward flux at –100 m is chosen arbitrarily as being approximately representative of water carrying the average nuclide activity (SKB 2010).*

Radionuclide transport

Simulation of consequences of radionuclide releases to the biosphere is a central task in PA. In SKB's safety analysis for the Forsmark site a large number of scenarios were modelled including "what-if" cases. The transport rate to the biosphere was simulated for all important radionuclides and radionuclide decay chains. Numerous flowpaths were followed. In each flowpath the impact of the residence time of water as well as the retardation caused by matrix diffusion and sorption over the flow wetted surface was accounted for. Figure 2-17 shows results for a temperate period and assumed canister failure by corrosion at 100 000 years. Numerous other simulations were made, including a number of "what if" cases to explore the consequences of a few canisters failing shortly after closing the site.

Performance measures

Performance measures are used to gain general insights in how the rock may perform to influence some desired outcome e.g., how well the rock may affect radionuclide transport and retardation.

Two illustrative performance measures that indicate how well the rock may delay escaping nuclides from a repository to reach the biosphere by different pathways are presented below. One measure is the water flux U_r at the canister positions. The other is an entity that shows the "retardation power" caused by matrix diffusion, the F_r factor. After solving the flow model for the network, particle tracking is used to follow a small packet of water from canister to biosphere. F_r is a measure of how large flow wetted surface, *FWS*, that the water packet with flowrate *Q*, encounters along the path, $F_r = FWS/Q$. Small U_r at the canister position leads to small radionuclide release. Large F_r is beneficial for delaying the radionuclides in their travel toward the biosphere. The layout of the repository with its 6916 canister positions is shown in Figure 2-18 together with the flux at each canister position. The flowpaths from each individual canister position were modelled.

Figure 2-19 shows how the performance measures, U_r and F_r evolve over time when the shoreline recedes as land rebounds after the latest ice age.

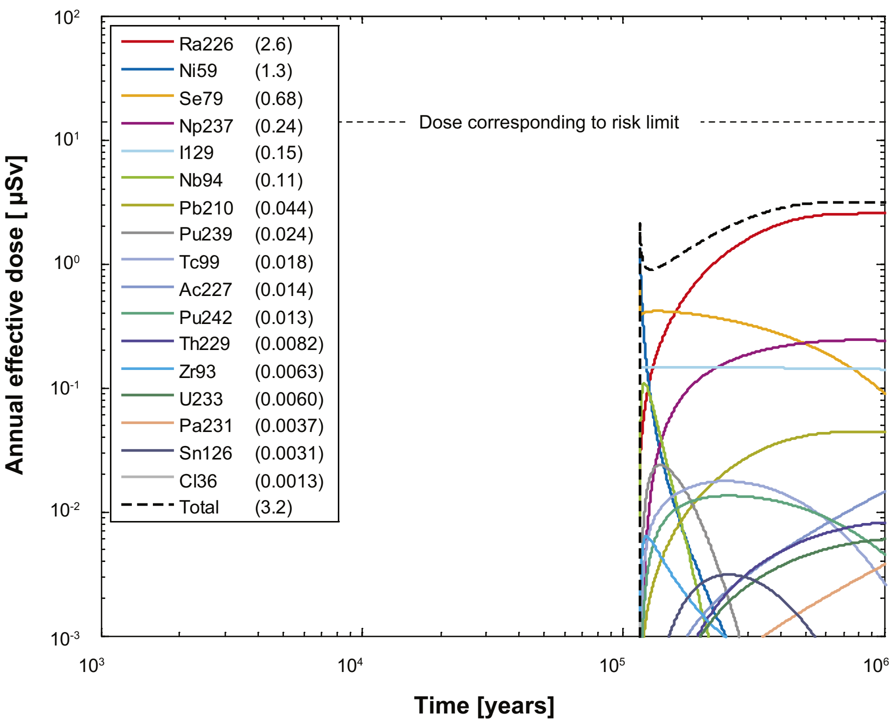


Figure 2-17. *Far-field annual effective dose for a deterministic calculation of the central corrosion case. The legend is sorted by peak (in the one-million-year period) of the annual effective dose. The values in brackets are peak dose in units of µSv, (Joyce et al. 2010).*

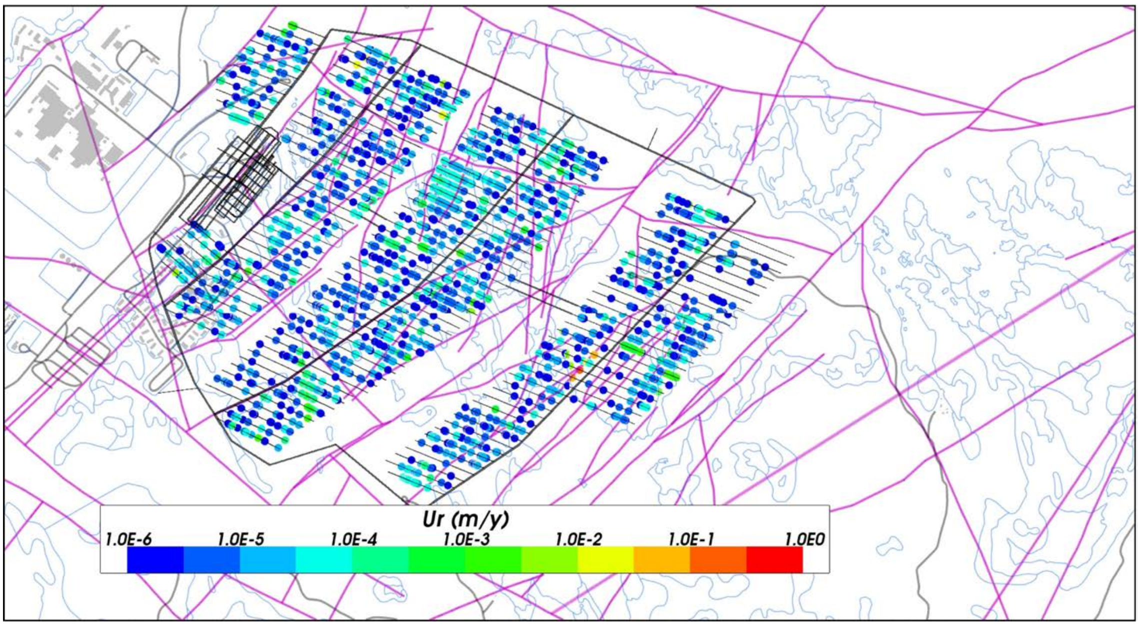


Figure 2-18. *Starting locations coloured by log10(U_r) for particles released at 2000 AD and successfully reaching the top boundary of the hydrogeological base case model. Roads (purple), buildings (black) and shoreline (blue) are also shown (SKB 2010).*

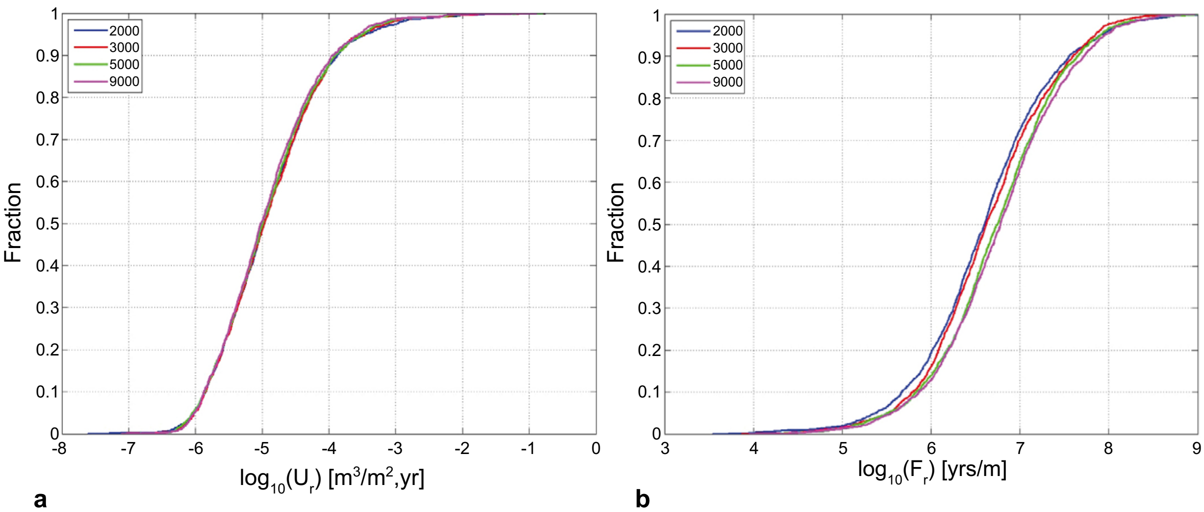


Figure 2-19. *Normalised CDF (cumulative distribution function) plots of flux a) U_r and b) Fr in the hydrogeological base case of particles successfully reaching the model top boundary (24 %).) The figures in the legend show times at which particle tracking starts (SKB 2010).*

2.3.4 Overview of important flow and transport model outputs

Flow and transport models play an important part in the Performance Assessment. In the near-field at the deposition hole (in the KBS-3V concept) scale they are used to predict:

- nature of fractures intersecting deposition holes,
- inflow-rates to deposition holes and tunnels for resaturation analyses and screening of deposition holes,
- water flowrates in fractures around deposition holes to estimate the rate of nuclide release from the bentonite buffer in the event of canister failure,
- the frequency of deposition locations receiving dilute altered meteoric or glacial meltwater as input to buffer erosion assessments.

At larger scale the key predictions include:

- retention and retardation of nuclides along flow paths,
- the evolution of groundwater at exit-locations and flow paths under temperate conditions and permafrost and ice sheet conditions,
- identification of recharge and discharge locations to the biosphere,
- interaction of natural fracture system with EDZ (excavation damaged zone) as potential pathway for flow and transport.

2.3.5 Some thoughts on credibility of and confidence in the models and modelling flow and transport in fractured rock masses and using the repository

Before an undertaking is started, be it building a bridge, a hydroelectric power plant or a final repository for spent nuclear fuel somebody will have to decide that the organisation that will do it has the necessary understanding and tools to do it. In the above-mentioned examples there is long-standing knowledge and experience that can be relied on. Modern conventional constructions are designed to last one or at most a few hundred years and can be inspected, maintained and repaired. In contrast, the final repository must function for hundreds of thousands of years without any means to remedy any mis-function. Another major difference between conventional enterprises and a repository in crystalline rock is the heterogeneous nature of the rock mass with its multitude of fractures and faults of all sizes, the locations, orientations, properties of which cannot be determined in detail. In addition to the geometric heterogeneities there are differences in geochemistry that may influence and be influenced by construction and the engineering work, to mention just a few interactions between different areas.

To design and build a final repository for spent nuclear fuel deep in fractured rock experience and understanding of many of the processes and mechanisms that will be involved was lacking when studies started in the late 1970's in different countries, among them Sweden and Finland.

It was early realized that modelling of the different processes and mechanisms involved would play an important role and that models and modelling must be based on solid scientific arguments. It was also realised that it would not be possible to make exact predictions of conditions and events over the extremely long times of interest.

Some key demands on models and modelling are listed below.

1. The models must be based on sound scientific principles and on accepted laws of nature.
2. The models shall not contradict experimental evidence and observations in nature.
3. The models must not contradict any general observations.
4. There should be a general agreement among scientist in fields that the models and model predictions in details and overall results are (not un-) reasonable.
5. Models used in different disciplines e.g., chemistry, geology, hydrology, rock mechanics etc. must be commensurate and not contradict each other.

All the work must be subjected to continuous and repeated review and scrutiny by independent experts. When the "Scientific community" over very long time has followed, scrutinised and criticised the work and when the vast majority of our peers have no more objections to the work and the conclusion that it can be used for the purpose(s) stated, the task is done. In the specific case of the spent fuel repository the documentation can be given to the authorities who may suggest to the government to go ahead and start building the repository.

2.3.6 Concluding remarks

Outlook

During the excavation of the drifts and tunnels to and of the repository there will be ample opportunities to test and refine the hydrology and solute transport models using the water inflow distributions as the excavation progresses. Prediction/outcome exercises can be made using exploration boreholes before a new section of drift is excavated. In these holes fracture hydraulic and other properties can be measured and based on the information inflow distributions in the next drift section can be predicted and then compared with what is found after excavation. It is probably not necessary to do this in the entire excavation length but could be made for few hundred m lengths occasionally, followed by a phase of model refinement.

Channelling issues

Long-range, hundreds of meters of exceptionally fast channels, have been observed, see Section 2.2.8. Such channels can carry corrosive agents to the canisters and radionuclides from a leaking waste canister. Very little is known of their frequency and properties. Such long-range channels, LRCs, are very difficult to observe and explore using only borehole information as the LRCs most probably are made of chains of shorter channels in different fractures or fracture intersections forming a continuous rapid flowpath over long distances.

High flowrate channels can be observed in the drifts and tunnels when the site is excavated, but such information says nothing about if the channel belongs to an LRC or not. One possible method to explore long-range channelling would be to inject non-sorbing UV-fluorescent tracers at some distance from the excavation and look for their emergence in the drifts by shining UV-light on the walls. This can give information of the frequency as well as on the RTD of the paths and possibly also some information on the F_r-factor (FWS/Q), which is an important performance measure.

The planning of these investigations would benefit by an early start when the "old hands" are still available and can share their experience.

3 Building model confidence by model reduction

Philippe Davy, University of Rennes

This chapter is based on the SKB DFN Modelling Methodology report (Selroos et al. 2022).

3.1 Building model confidence by model reduction

Development of a DFN methodology is a response to a need for various types of predictions of geosphere behaviour in fracture media. It relies on building models that minimize deviation from the data or assumptions the model is conditioned to, while simultaneously complying with conditions required by the purpose of the modelling activity.

It is convenient to introduce a few terms relevant to the issues discussed in this chapter. The **model space** is the envelope of possible models, when describing bounds on the system of interest, from which relevant parameterisations, idealisations and modelling principles are chosen, in a process which eventually reduces the model space. The reduction of the model space is achieved by a series of operations (Figure 3-1):

- Selection of the processes relevant to the prediction outcome (see Section 3.2.1). Essentially:
  - Identification of the physical and/or chemical processes that are of relevance for the prediction.
  - Identification of suitable qualitative or quantitative constitutive models that describe those processes and their interactions (that may involve simplifying assumptions).
- Listing of all parameters that define the model space.
- Identification of a prior probability distribution, and its moments, for each defined parameter.
- Conditioning the model space on prior data information (see Section 3.2.2) – i.e., data that are amenable for processing as model parameters (value, range or distribution).
- Calibration of the model on data (see Section 3.2.3) – i.e., assessing the probability of a chosen model parameter set to be consistent with observation.
- Validation is a demonstration that the model is an adequate representation of the real system being modelled to make the predictions required (see Section 3.5).

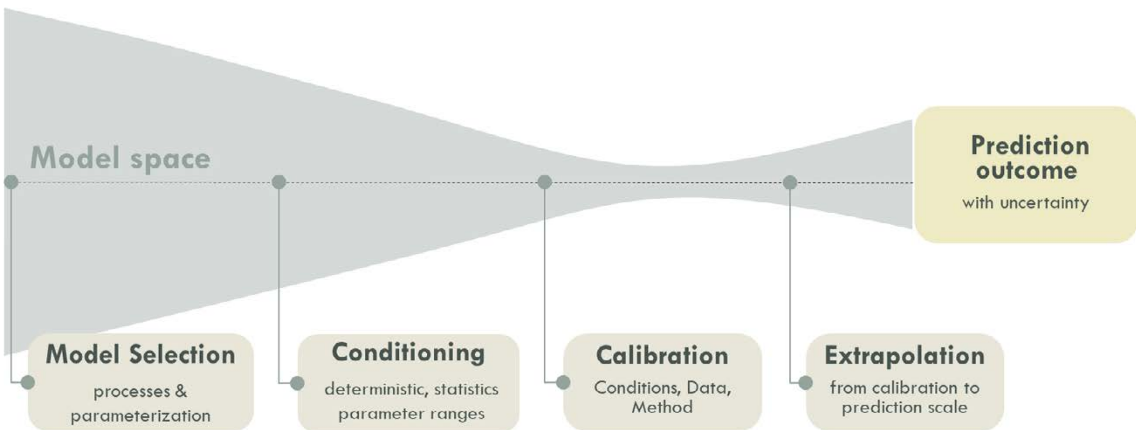


Figure 3-1. *Sketch of the processes that influence the evolution of the model space from hypothesis to prediction. The thickness of the model space gives a qualitative indication of the range of possible models or model parameters.*

The process can be iterated in different ways (see below) and the eventual output is the ensemble of Models Conditioned by the Data (MCD), where each conceptual model is characterised by a set of parameters with a probability of occurrence. The model space is expected to be reduced as compared to the initial assessment through rejection of some selected models and/or constraining parameter space.

The prediction is an application of the MCD within the prediction space that defines all the relevant conditions such as time, space, boundary conditions, etc. For operations that extrapolate the calibration conditions to much larger space and time scales, this entails not only an increase of the prediction uncertainty compared to the fitting under calibration conditions, but also likely an increase of the model space if new or modified physical/chemical processes that are not relevant under the calibration conditions are expected to occur. Examples of such processes are changes in the boundary conditions (glaciations, elevation of temperature, mechanical load, sea level, etc.), changes in the geometry of structures and/or in the material properties (e.g., due to future tectonic or glacial events, chemical fluxes), new chemical processes, etc. The extrapolation from calibration to prediction scale is the last step in Figure 3-1, although this whole process may be part of a repeated cycle during the main stages of the underground development and associated detailed site investigations and site-descriptive modelling.

The model space workflow is consistent with the Bayesian theory (Gelman et al. 2013), where the "prior", i.e. what we think about the modelled system based solely on expert judgement and no data (model selection), is combined with data (conditioning) to give the "posterior" distribution, e.g., the model space conditioned by the data, after estimating the likelihood of model parameter values given the observed data (calibration). The 'posterior' distribution is then represented by the reduced (i.e. constrained) model space after conditioning and calibration, which is then applied to the prediction space (extrapolation).

Note that conditioning and calibrating are similar ways of reducing the model space by using data. We find it convenient to differentiate a data comparison process that sets a priori parameter value or range (conditioning) with a data comparison that calibrates parameters by letting the model run under calibration conditions (calibrating). There is a wealth of published theories on calibration, also called the inverse problem, of well-posed problems (if processes and parameters are well known, a unique solution exists), with a lot of data, for which the calibration space is close to the prediction space (calibration experiments performed with conditions similar to the prediction) e.g., Tarantola (2004) and Stuart (2010). In that case, the calibration is essential in reducing the model space for prediction and it is different in nature and in techniques from the prior conditioning. In complex systems with sparse data compared to system heterogeneities, prior conditioning may be even more important than the calibration on experiments that are far from representing the natural conditions of the eventual prediction. In this case, prior conditioning and calibration are both important steps of model space reduction.

3.2 The stages of model reduction

In this section we elaborate on each of the processes of model reduction shown in Figure 3-1. to demonstrate that it comprises a cycle of iterative steps as shown in Figure 3-2. Each of these processes is described below, including the main steps: Model selection, **S**ensitivity **A**nalysis, prior **C**onditioning, calibration, **Re**jection (called SACRe process thereafter), and extrapolation.

3.2.1 Model selection

The model space (MS) is the ensemble of physical (in the broad sense) rules, mathematical equations and parameters that give a theoretical, observational and/or empirical description of the *geometrical structures and domains* (i.e., the idealized geology), *processes* and *boundary conditions* that allow simulations of a system and its evolution. This definition implies a logical and objective formalism, which makes a model of MS amenable for testing and evaluation with data or experiments. In that sense, it is a step further in the formalism compared to a conceptual model that would rely only on ideas, words and schematics. For the purpose of traceability and repeatability the rules, mathematical equations and parameters that define MS must be clearly set.

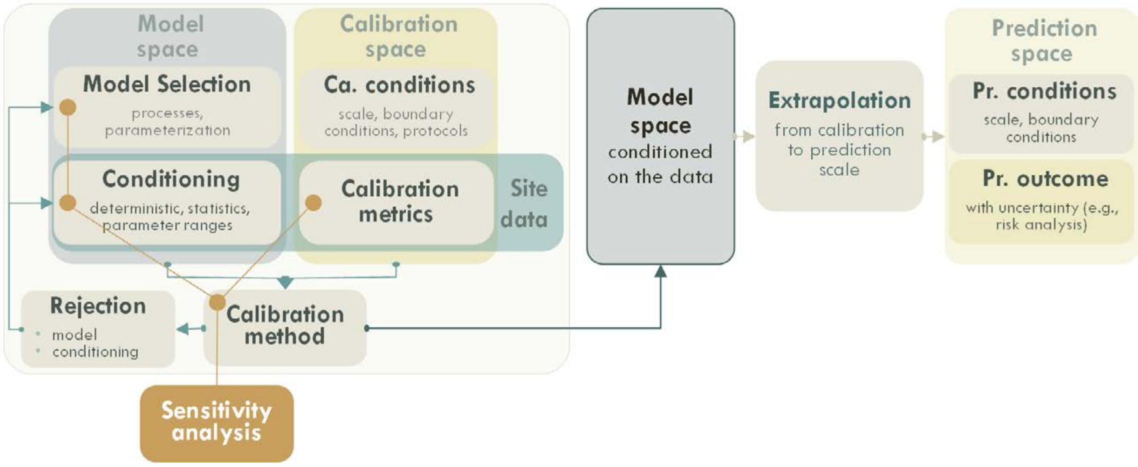


Figure 3-2. *A detailed sketch of model space reduction, including the different stages, how they combine and semantics.*

A model is obviously not the reality but a representation that contains hypotheses and simplifications of the real system. MS contains not only the ensemble of parameters that describe a model, but possibly different representations of the same geological reality (i.e., alternative models). A key issue is its amount of complexity, which is a trade-off between the complexity of the real system so far as it can be observed, and the need to remain commensurate with the requirements for the model and what the available data can support. As for statistical models or machine learning, overfitting (over-determination) is an issue that needs to be avoided; although it allows for a better fit of the data, it makes model prediction/extrapolation less reliable (Hawkins 2004). The sensitivity analysis is a useful test for assessing the right level of model complexity, in the sense that additional complexity can only be justified if it demonstrably contributes to a reduction in model predictive uncertainty (see next section).

MS evolves as new information is gathered or different requirements are placed on the model. A prior estimate of the model parameter ranges can be required to avoid starting with a too large model space, or simply to avoid adopting unphysical parameter values.

MS is not a unique ensemble of relationships since it relies on an understanding of the nature of the processes relevant to the prediction, which may not be complete. MS is underpinned by the relevant laws of physics, but it may also contain soft data, i.e., correlation between parameters of interest, established geoscientific principles or commonly used models that practitioners have found to adequately describe observations at other sites from personal experience, accepted practice or inspired by the scientific literature. Some of these empirical relationships and principles, which make up for the lack of data or knowledge, are hard to verify. To bracket the geological truth, it is recommended to test several hypotheses with different levels of scientific plausibility (e.g., if it is supported by physical principles, even simplified, or derived from data fitting, each one defining a sub-model space. It is convenient to define a *baseline MS*, which corresponds to either the most likely model from a scientific point of view or to the one most used in engineering practices of the day. The postulate is that predictions made with the *baseline MS* and *alternative MS* adequately span outcomes in the real system (see Section 3.4), which is not obvious a priori and must be questioned throughout the modelling phase.

3.2.2 Prior conditioning to site data

Prior conditioning is the process of constraining models with information about the model parameter ranges, the statistical targets, or the observations which must be respected by the models in a strict absolute or statistical sense. Part of this conditioning belongs to the normal model selection process (previous section) when it consists in collecting generic – i.e., not specific to a site – information about the physical processes. Part relies on local observations made on the site – e.g., a structural model with the main geological domains and their boundaries defined by site geologists.

For the sake of clarity, we recommend identifying the prior conditioning related to site data as a step in model space reduction different from the model selection. The point is that this conditioning is complementary to the calibration, which also uses site data, but in a more indirect sense. It clarifies the DFN methodology to know at which level of the model reduction process the data are used.

Prior conditioning to site data is often used to implement the know-how of geologists and hydrogeologists in a series of constraints with potential to significantly reduce the model space and simplify the modelling process and the models stemming thereof. The geological description provided to modelers, consisting of a set of domains whose properties are assumed to be uniform or statistically uniform, is an example of this preconditioning step. A sensitivity analysis is recommended both to evaluate the need and consequences of the prior conditioning in the prediction.

In the case of a multi-step process, where the calibration is applied several times, step by step (e.g., a calibration of the structural model space first followed by that of the hydrogeological model space), the prior conditioning of the last stage is the result of model calibration and conditioning of the previous stage. However, the conditioning may also be applied simultaneously such that the MS is constrained by e.g., combined geometrical and hydraulic conditions.

3.2.3 Calibration

Calibration space

Calibration is the process of reducing the model space by comparing model outputs to observed data (Tarantola 2004). This is not only the estimate of the best-fitting set of model parameters but the determination of an occurrence probability as in the Bayesian formalism (Bayes 1763, Stuart 2010, Gelman et al. 2013). This problem is known as "the inverse problem" (data/output → model parameter), which is the inverse of forward modelling (model parameter → output). But, as expressed by Tarantola (2004), *'while the forward problem has (in deterministic physics) a unique solution, the inverse problem does not'*. There are many reasons for this. Given uncertainties, poor or inadequate data quality and the complexity of the model itself, a number of different parameter combinations can render quite different, but equally "good" fits to the information to which the model is calibrated. A critical point, therefore, is that the conditions of calibration (e.g., the calibration space) are not necessarily adequate to explore all the model parameters which, in practice, imply that some parameters will remain poorly determined after the calibration process. In this sense, calibration may be viewed as a probability density function filtering process, in which the uncertainty in the model's parameter values is *reduced*, rather than eliminated. The extent to which the uncertainty associated with a given parameter is reduced by calibration is a function of both the information content of the observed data (as it pertains to that parameter), as well as the extent to which that information is shared with other parameters in such a way that it cannot be uniquely resolved; itself a function of the model parameterisation, amongst other factors . Therefore, defining the conditions of calibration – e.g., the observed dataset with uncertainties and the physical conditions of calibration (volume and time scale explored, boundary conditions if necessary, etc.) – is important information for any end-user of the calibrated model.

Metrics for calibration

Calibration metrics define practical measures for comparing model predictions with field observations of the fracture system and/or models with each other. When used to compare model and data, the metric must be applied to modelled conditions similar to those in which measurements are made.

Since DFN models are simplified representations of real fracture systems, i.e. aimed to show statistical equivalence, adequate metrics are not necessarily a 'local' measure, but it may be a global measure averaged over a certain volume (e.g., intensity), a statistical distribution (of e.g., sizes) or a scaling relationship (e.g., aperture scaling with fracture size). Examples of calibration metrics are:

- 'Local' metrics are defined by a quantity – e.g., transmissivity or inflow – and by the way it is measured including the position, sampling scale and uncertainty. It is only a relevant metric to calibrate DFN models, which have been heavily preconditioned on the data in the vicinity of the measurement.
- 'Density' metrics are the average of a quantity over a certain volume – e.g., trace density measured in tunnels or boreholes.

- Statistical distribution – e.g., trace size distribution, flow distribution or breakthrough curves – are relevant metrics to calibrate the spatial variability of parameters (measured/modelled on tunnel walls, and along and in between boreholes and tunnels).
- Scaling relationships characterise the structure (correlations and spatial arrangement) across scales of the target quantity. An example is the equivalent hydraulic conductivity scaling (see Chapter 8 of DFN Handbook) that can be computed from borehole intervals with PFL data to characterise the connectivity and spatial hydraulic structure.

Metrics should be chosen as a function of the prediction/outcome objective (i.e. related to properties affecting long-term safety or repository engineering) in order to increase the reliability of the calibrated model in its ultimate use. For this reason, the choice of suitable calibration metrics also reflects suitable prediction metrics. An example of prediction metrics used for validation of DFN models in the ONKALO is given in Table 3-1 classified according to their type and the components of MS they are used to validate. The issue of prediction metrics will be returned to in vol II of this DFN Modelling Methodology.

Table 3-1. Examples of prediction metrics (revised from Hartley et al. 2017).

Property	Metric	Type
Fracture size and spatial distribution	Number of full perimeter intersections fractures in tunnel section	Density
	Number of gently dipping traces >> mapping resolution of trace length[3]	Density
Flow distribution	Number $T > 1 \times 10^{-8}$ m^2/s in individual fractures	Density
	Number $T > 1 \times 10^{-9}$ m^2/s in individual fractures	Density
	Total summed flow to boreholes	Density
Inflow distribution	Difference between geometric mean of simulated values and the measured value of inflows	Local
	Median of number of inflows lying correctly above/below detection limit	Density
Fracture size and spatial distributions	P_{10}[4] in pilot holes	Density
	P_{21} for traces lengths >> mapping resolution	Density
	Trace size distribution	Statistical
Orientation	Goodness of fit on stereoplots[5]	Statistical
Terminations	Percentages of one-end and two-end terminations	Statistical
Spatial model	Trends in relative intensity by lithology	Scaling
Aperture	Correlations of geological aperture with trace size	Scaling

Many quantities – e.g., transmissivity – cannot be measured directly but are instead derived from the interpretation of an experiment or a proxy measurement (e.g., ratio between inflow and hydraulic head as a proxy for transmissivity). It must be verified that the quantities simulated in the model and in the field represent the same physical or chemical parameters and conditions.

[3] The mapping resolution is the truncation limit dictated by the fracture sampling policy. This limit may vary within different parts of the repository, and with the purpose of the mapping, but is typically 0.25–0.5 m in niches and 0.5–3 m in the main tunnels.

[4] The borehole intensity, P_{10}, can be based either on fractures intersecting the axis of the core, i.e. a scanline proper, or as fractures making a full perimeter intersection (FPI) with the core. The choice of mapping principle will affect, among other things, the choice of bias correction.

[5] There are many different techniques to compare fracture orientations but most favoured by practitioners are based on Stereoplots.

A sensitivity analysis (see Figure 3-2) is essential to demonstrate the ability of the metrics to make the model more reliable for a given prediction objective.

A statistical metric for calibrating flow distributions is shown in Figure 3-3 based on simulating single-hole PFL hydraulic tests in 17 deep core drilled KFM holes at Forsmark compared to measurements partitioned according to fracture domain and depth. This involves calibration of connectivity and transmissivity.

An example of a metric for the scaling relationship for flow is shown in Figure 3-4. The equivalent hydraulic conductivity, $K_g(l)$, computed from the available single-hole flow tests, performed over several cored boreholes at the Forsmark site is shown by grey lines with annotation of the mean scaling behaviour and how to interpret it. It has a V-shape evolution with first a decrease and next an increase phase. At sizes L_s smaller than the minimum distance between two inflows, $K_G(L_s)$ decreases as L_s^{-1} down to a minimum value (point 1 Figure 3-4), then increases again once averaging begins of two or more inflow values within a section.

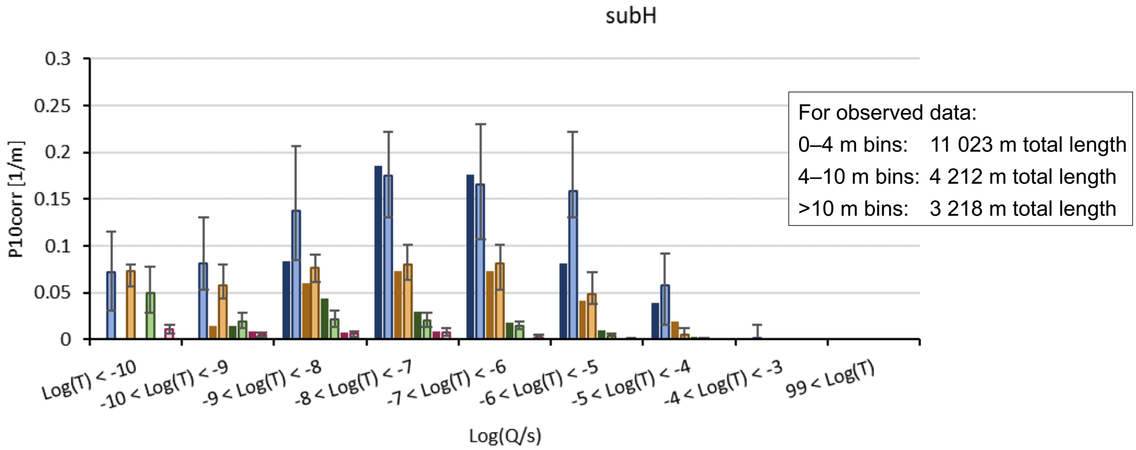


Figure 3-3. *Example statistical metric for the intensity of flowing fractures in different ranges of specific capacity by rock units FFM02(U/L) and FFM01 (above and below z = –200 m) for the sub-horizontal orientation set. Model results are the average over 10 realisations of a DFN model (simulated KFM holes only). Error bars indicate the maximum and minimum results across the realisations (Hartley et al. 2021).*

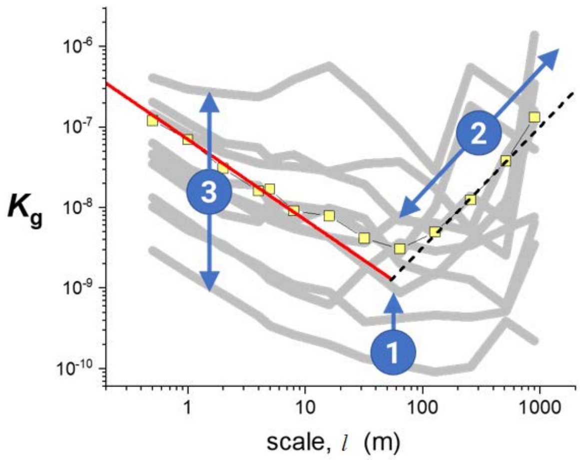


Figure 3-4. *Scaling relationships of the geometric mean of the hydraulic conductivity, K_g calculated from core-log PFL transmissivity record of the 12 KFM boreholes in Forsmark by sampling different boreholes at different scales irrespective of vertical depth. The metrics that are used to calibrate models are 1) the scale at which K_g is minimum (interpreted as the network percolation scale), 2) the scaling of K_g above the percolation scale, and 3) the variability of K_g for different boreholes). The red and dotted lines illustrate this metric for the borehole KFM08A identified by the yellow squares. Additional details and explanations are given in Davy et al. (2023).*

The calibration process and the objective/likelihood function

The calibration process is related either to a minimisation of an objective function, which calculates the difference between model outputs and observed data (Gupta et al. 1998), or to the calculation of a likelihood of the observed data given a set of model parameters (the Bayesian approach (Bayes 1763 Gelman et al. 2013). The former yields the best-fitting model with the range of plausible parameters; the latter gives a likelihood of the model space conditioned to the data.

The Bayesian approach allows the issue of over-parameterisation (e.g., the discussion in Brunetti et al. (2017): *'a parameter-rich, but geologically-unrealistic model may fit the data equally well or perhaps even better than a more parsimonious model'*) to be addressed quantitatively. Using the Bayesian approach to choose between different alternative model spaces is a possibility worth exploring. However, in the context of complex geological systems with sparse data sampling, the issue of finding the right trade-off between model complexity and goodness of fit is an open issue (see the discussion in the Section 3.4).

Calibration, reliability tests and model rejection

Reliability tests and model rejection are basic versions of the calibration process, which require defining an (un)acceptable deviation between the models and the data.

Early model rejections are important steps of the model space reduction since it allows for narrowing the options to the relevant alternative models. An example is a connectivity analysis for the open fracture system against the intensity of discrete flow measurements where some alternative intensity-size scaling and spatial models for fracture openings may be rejected prior to flow simulation (see Selroos et al. 2022).

3.2.4 Sensitivity analysis

Sensitivity analysis (SA) is defined "as the study of how the variability in the model output can be apportioned to the different sources of input" (Saltelli et al. 2006). It is a preliminary analysis of the models that underpins several important tasks of the global process:

- Model selection: Test which complexity is necessary, and which knowledge about processes and parameters is required.
- Conditioning: (i) Test which data accuracy is required (e.g., whether data can be considered deterministic or partly uncertain); (ii) test if new data bring more constraints.
- Calibration metrics: Test the ability of metrics to make the model more reliable for a given prediction objective.
- Calibration process: The sensitivity analysis basic to parameter estimates.

How SA can be carried out is an open question given the non-linearity of the models. SA can be local, one parameter at a time or global in the allowable ranges of the input space (Leamer 1985, Baroni and Tarantola 2014), the latter being preferred for complex, non-linear, models. In the context of environmental models used for decision support purposes, a range of methods have been developed for evaluating model predictive uncertainty in non-linear models. These include Markov chain Monte Carlo (MCMC) simulation , calibration-constrained Monte Carlo analysis and Bayesian ensemble-based data assimilation methods.

We recommend starting from the simplest models that are close to reality (e.g., a 'site-like' modelling that reproduces site conditions as much as possible) and to test the importance of introducing a process, a parameterisation, or a data on the eventual prediction. That is to say, additional model complexity should only be justified if it can be demonstrated (via model predictive uncertainty analysis) that it improves the model's predictive power as it pertains to the key predictions of interest. This can occur in one of two ways, namely that the additional complexity either: (1) enhances the model's ability to express expert knowledge (i.e. in the form of the prior distributions); or (2) improves the model's ability to replicate relevant system behaviour (i.e. as expressed by the likelihood function). Since the model's predictive power varies greatly with the type of prediction it is asked to make, the model selection, conditioning, and calibration processes should be tailored to the specific predictions of interest.

3.2.5 Stochasticity

Most of the models in the DFN methodology are stochastic or partly stochastic, i.e., relying on random variables constrained by stochastic processes or statistical distributions. A realization is characterized by the set of stochastic/statistical parameters used to generate it, and the set of numbers which identify the realization in the random space. For DFN, the random space contains the parameters that position fractures and describe their properties under statistical/stochastic constraints. Given the number of fractures and of fracture properties, the random space is always orders of magnitude larger than the statistical/stochastic space (i.e., the number of independent statistical parameters used to generate realizations).

For a given set of stochastic/statistical parameters, the ensemble of realizations gives the mean and variability of stochastic model outputs.

The calibration of stochastic models applies on both set of parameters, stochastic/statistical and random, with the goal to find the set of realizations that match the data. Although there is no prescribed rule, calibration can be done as a two-step process:

- First, a "statistical calibration" investigates the ensemble of statistical/stochastic parameters with adapted metric (for instance, relying on the mean and standard deviation of outputs).
- Lastly, a "full calibration" aims at identifying the subset of realizations that match data.

The former calibration is reasonably achievable since the stochastic/statistical parameters are generally well constrained by data and they exert a strong control on the model outputs. The latter is more difficult to achieve because of the size of the random space, which leaves a wide range of possible realizations. This is very often the main source of uncertainty in stochastic model predictions.

This raises the issue of assessing the mean and variability of model outputs due to the random space – i.e., the space of stochastic realizations. The model outputs are the metrics chosen to quantify the impact of each realization, and their probability distribution (pdf) determines the ensemble of possible values of the metrics. The metrics average and variability are highly dependent on the shape of the pdf. If low-probability high-impact realizations contribute significantly to the mean or standard deviation of the impact metrics, good estimates of these terms require a significant number of the high-impact realizations and therefore an even greater number of realizations. For pure random Monte-Carlo methods, the ratio between these two numbers is inversely proportional to the survival probability of the low probability events, which gives an estimate of the number of realizations to be carried out. In practice, convergence tests can be used to estimate the quality of the mean and variability of the selected metrics.

Note that the realization number depends on the purpose of modelling, i.e., on the chosen metrics; it will be different for flow, residence time or channelling characteristics, for instance. Also note that, beyond the mean and standard deviation, it could be important to evaluate the probability of some highly impacting events if, for instance, a threshold not to be exceeded has been defined.

3.2.6 Extrapolation: Applying the model space conditioned on the data to the prediction space

The prediction space is defined in terms of:

- time scales,
- spatial scales,
- spatial domains,
- boundary conditions, and
- the significant processes that may operate during the prediction time (e.g., glaciations for long term predictions).

For a repository project, prediction space has two main purposes: one in support of validation (see Section 4.4) and one in support of end users: environmental impact assessment (EIA), post closure safety assessment and repository engineering. Some predictions made for validation purposes will be made for conditions close to those used in calibration, e.g., underground mapping and hydraulic tests, while others may be on a larger scale, such as hydraulic disturbances in monitoring holes distributed across the local-scale domain. Examples that require a longer time scales are palaeo-climate simulations. Predictions for engineering will likely include extrapolation to new construction areas, as yet unsampled, while forward modelling for long-term post-closure safety will include time scales and boundary conditions not considered during calibration.

Since the prediction space may be very different from the calibration space, the projection of the models conditioned by the data is not straightforward and may lead to additional, or revised, assumptions in terms of boundary conditions, processes, etc. This results in an increase in the range of models/parameters and assumptions. The prediction-outcome must take this into account in order to deliver the likelihood prediction and its uncertainty.

Predictions used for validation are compared with outcomes derived from measurements, as part of prediction-outcome exercises using new or reserved information during ongoing investigations and site monitoring (i.e., validation date set). In this context "outcome" is not just comparing model predictions and observations, but also describing what is learned about model space from making such a comparison.

3.3 A multistep multipurpose process

The process described above can be decomposed into several successive steps as shown in Figure 3-5 and adapted to different prediction spaces. We recommend each step to be described with the required elements (model space, prior conditioning, calibration space) in order to ensure the reliability of the whole process.

For example, the top line of Figure 3-5 may represent the flow of Geometric DFN model space to Flow DFN model space to Transport DFN model space to safety assessment prediction space of flow and transport around the deposition areas. The example here thus illustrates that conditioning can include data from multiple disciplines such as geology (e.g., geometries), hydrogeology (e.g., flow) and transport (e.g., breakthrough curves, i.e., travel times).

Several prediction spaces can be defined (e.g., repository area (facility scale) and individual deposition areas (facility part scale)).

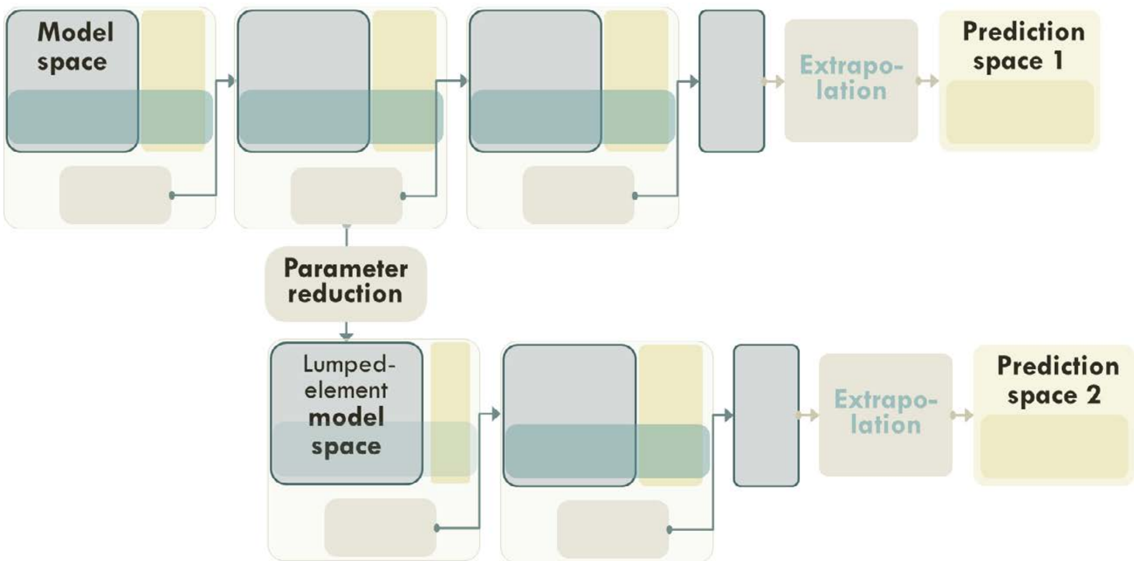


Figure 3-5. *Sketch of a multistep, multipurpose process.*

It may be necessary to use models with different degrees of complexity depending on the prediction space (especially for long-term, large-scale predictions). The new model space is then considerably reduced by integrating/averaging the local heterogeneity into a few “lumped” (effective) parameters (Incropera et al. 2007). The consistency between both models must be checked to ensure that the lumped-element model is still able to reproduce the main processes with a resolution consistent with what the prediction requires, and to evaluate the uncertainty due to the parameter reduction. For example, DFN models for flow can be up-scaled to an ECPM for performing simulation of long-term evolution of groundwater flow and chemistry.

The whole process is iterative if/when new data are acquired, or new processes identified.

3.4 A trade-off between model complexity and parsimony

An issue for validation is the trade-off between model complexity and parsimony, which deserves an explicit and motivated choice. Model complexity is usually defined as the richness of the model space, e.g., the number of free parameters (Ye 1998, Spiegelhalter et al. 2002), but also in machine learning, as a measure of how hard it is to learn from limited data.

The arguments for complex models are:

- Model confidence partly relies on their ability to reproduce geological systems in all their complexity and variability.
- Model complexity is also necessary to some extent to allow the models to be conditioned to data of different types (geological, hydrogeological, geochemical, etc.) – e.g., prediction exercises on very specific areas.
- Important processes for prediction are partly controlled by the spatial variability of fluxes, e.g., flow channelling controls transport of radionuclides. Time and space variability is hardly obtained by the simplest models; it is rather an emergent property of the model complexity, which may be required for the prediction.

However, complex models with limited (and/or noisy) data may not perform well on new data and thus on prediction. They are commonly overparameterized, which reduces their prediction capability after calibration. This is the principle of Occam’s razor, which implies that any given complex function is *a priori* less probable than any given simple function. Furthermore, the computational resources required to conduct sensitivity and uncertainty analyses on complex models can be significant.

On the contrary, simplified models are supposed to be easy to implement and fast to calculate. They are useful for several reasons, including (but not restricted to):

- to explore the model space by guiding the modeler in developing alternative conceptions,
- to make the calibration process easier by systematically identifying the main parameters and performing a complete sensitivity analysis,
- to provide a better understanding of the key processes and characteristics of system, which helps to increase confidence in the modelling through a fine understanding of the results.

Graph representations, effective field theories, sparse channels, or simplified parameter relationships are examples of simplified models.

If different kinds of models from simplified to complex are necessary to increase the scientific relevance of the modelling approach, the “best” models for prediction are a trade-off between complexity and parsimony, which consists in reducing the complexity as much as possible[6] to increase the prediction capability, but not too much so as not to lose the model relevance to describe the main geological processes. Effective theories validated by complex modelling, or simplifications by removing structures and processes of minor importance are ways to reduce model complexity. The optimum for prediction models requires frequent iterations between models of different complexity level.

[6] The debate between model complexity and parsimony is well summed up by the sentence attributed to Albert Einstein “Everything should be made as simple as possible, but no simpler”.

3.5 Validation

A vital component of the model building process is the validation of the whole process in the sense defined by the IAEA Safety Glossary (IAEA 2018): *The process of determining whether a model is an adequate representation of the real system being modelled*. IAEA (2018) also specifies that, in the case of geological disposal facilities that involves temporal scales and spatial scales for which no comparisons with system level tests are possible, models cannot be 'validated' for that which cannot be observed, so model validation implies showing that there is a basis for confidence in the model(s) by means of detailed external reviews and comparisons with appropriate field and laboratory tests, and comparisons with observations of tests and of analogous materials, conditions and geologies at the process level (IAEA 2018).

Part of the confidence is related to a clear description of the modelling workflow:

- Are the different stages (definition of the model space, calibration, …) clearly identified?
- Does the model selection and data rely on current scientific methods?
- Have the consequences for the prediction of the model selection, including the necessary simplification assumptions, of the prior conditioning, and of the calibration process, been sufficiently tested through a relevant sensitivity analysis?

The confidence also requires to understand as precisely as possible where the uncertainty in the prediction comes, and to quantify it after calibration. Most authors who have addressed this issue acknowledge that "verification and validation of numerical models of natural systems is impossible", because all such systems are open, with distributed input parameters that are incompletely known or conceptually inconsistent with their model counterparts (Oreskes et al. 1994). But models are proved to be useful to help engineers make the best decisions (de Marsily et al. 1992) if they are verified on data or if the sources of uncertainty due to assumptions and lack of data are well understood and quantified.

All the steps described above, including calibration, sensitivity analysis, fundamental understanding, and estimate of the stochastic uncertainties, are steps that contribute to giving confidence in the modelling predictions.

4 Model confirmation and validation

Stefan Finsterle, Finsterle GeoConsulting, LLC

4.1 Introduction

This chapter discusses some of the issues surrounding the attempts of scientists and engineers to corroborate, confirm, verify or validate[7] their understanding of a hydrogeological system as implement in one or multiple numerical models, specifically those that are eventually used as a basis for public policy decisions related to nuclear waste isolation.

This topic is wide-ranging. It may be approached by first examining its deep roots in the philosophy of science and hypothesis testing. The application of the fundamental principles identified by philosophers and historians of science to the problem at hand results in somewhat changed definitions as well as in a set of more pragmatic approaches that were developed with the goal to corroborate site-specific models and their usefulness for a specific purpose. Moreover, the importance of this process for a defensible safety case – a key element of a license application for a nuclear waste repository – needs to be discussed, as it may have implications for the public's confidence in the inferences and decisions made based on the model's outcomes.

The brief discussion that follows aims at providing some philosophical background. However, the remainder of this chapter focuses on the more practical aspects of model validation. We further limit our discussion to hydrogeological models that are used to examine safety-relevant features or processes of a nuclear waste repository, specifically groundwater flow and transport processes in geologic formations that exhibit discrete conductive features or features that act as flow barriers.

We will make the case that the responsible use of numerical models requires a sufficient understanding of the quality and robustness of the simulation results, with direct implications for how these results need to be interpreted, and how they can (or cannot) be used in support of important policy decisions.

4.2 Criticism of verifiability and model validation

There has been a long and interesting philosophical discourse about the fundamental (im)possibility of establishing the truth of any proposition, specifically one made about a physical system[8]; nevertheless, we only provide a very short overview of this history.

In the philosophy of science, the question of how the truth of a scientific statement, hypothesis, or theory can be verified has been expanded to the more fundamental question of whether such verification is possible even in principle. At best, a theory (or model) can only retain the status of being "not invalid". The concept of falsifiability, proposed by Popper (1934), suggests that for a theory to be considered scientific, one must be able to test it and prove it false. Falsifiability is not only formulated as an explicit opposite to verifiability, but also as a criterion for demarcation

[7] As will be made clear in the remainder of this chapter, the terms "verify", "validate", "corroborate", "confirm" and related expressions (such as "confidence building") have specific meanings in the context of this White Paper. In the scientific literature, these terms are often put in quotes to indicate that they are not meant in their literal sense, which may imply that the model makes a definite statement about the absolute truth of a physical system. We omit the quotes here as it is the intent of this discussion to clarify the meanings and limits of these terms when describing the quality of a numerical model used in the geosciences.

[8] We are not concerned here with statements made within the realm of symbolic logic or mathematics, but with conceptual and numerical models that are based on observations made in the empirical world. According to Einstein: "One reason why mathematics enjoys special esteem, above all other sciences, is that its laws are absolutely certain and indisputable, while those of other sciences are to some extent debatable and in constant danger of being overthrown by newly discovered facts."

between science and non-science, and as a methodological guiding rule for scientific research. Not only is it generally accepted that theories cannot be verified, but falsification as a method to advance science has also been criticized. In essence, taking the discrepancy between an observation and a prediction as a sufficient criterion to falsify a theory (or a model) can be misleading, because any observation is itself laden with auxiliary hypotheses – it may not even be possible to prove that a theory is false, unless we can agree that a sufficient empirical basis exists.

Thomas Kuhn argued that experiments and observations are determined by the prevailing paradigm, and – conversely – the discrepancy between observational data and a prediction do not necessarily refute the underlying theory, as competing theories could be inherently incommensurable. Consequently, scientific truth cannot be established by objective criteria, but is determined by scientific consensus, which may change rather abruptly in the course of a paradigm shift (Kuhn 1962). Kuhn (1977) also proposes to examine the following five characteristics as criteria for theory choice, noting that the evaluation of these criteria remains subjective; a theory should be:

- *accurate* – empirically coherent with observations,
- *consistent* – internally consistent; externally consistent with other theories,
- *of broad scope* – extend beyond what it was initially designed to explain,
- *simple* – the simplest explanation (Occam's razor),
- *fruitful* – disclose new phenomena or relationships among phenomena.

The debate over the verifiability or falsifiability of scientific theories, the criteria for theory choice among competing alternatives, the proper methodology for advancing science, and the actual history of scientific progress ranges from thought-provoking to contentious; it is beyond the scope of this White Paper to summarize this debate or to take a position, other than acknowledging that attempts at validating a numerical model are most likely facing similar fundamental difficulties as do scientific theories.[9]

The similarities and differences between theories and models are discussed in the influential[10] article by Oreskes et al. (1994), who examine the issue of verification, validation, and confirmation in the context of numerical models in the earth sciences – with frequent references to validation efforts for models developed to assess the safety of a nuclear waste repository.

Oreskes et al. (1994) reach the conclusion that "*(v)erification and validation of numerical models of natural systems is impossible.*" They arrive at this statement by observing that all natural systems are open, with distributed input parameters that are incompletely known or conceptually inconsistent with the definition and scale on which they might be directly measured or inferred using auxiliary hypotheses, models and assumptions. Even if not aiming for a statement that establishes truth (as the term verification implies), the legitimacy of an application-specific model cannot be established, either. Validating a numerical model by comparison of its predictions with observations only indicates consistency but does not ensure that the model represents the natural phenomena. While calibrating a numerical model may signify that it is empirically adequate, reproducing past observational data does not guarantee model performance when predicting the future, because any extrapolation requires a change in the model structure, which in turn affects processes, temporal and spatial scales, the influence of input parameters and sensitivities of output variables. Even if data not used for model

---

[9] The validation of a theory or law can be different from the validation of a numerical model. This difference may be considered fundamental or subtle. According to Oreskes (2000): "[M]odels are a complex amalgam of theoretical and phenomenological laws (and the governing equations and algorithms that represent them), empirical input parameters, and a model conceptualizalation. When a model generates a prediction, of what precisely is the prediction a test? The laws? The input data? The conceptualization? Any part (or several parts) of the model might be in error, and there is no simple way to determine which one it is." For the purposes of this White Paper, we consider the discourse about the fundamental verifiability or falsifiability of a theory sufficiently relevant also for the discussion of model validation; however, we focus on the specific challenges of model validation, specifically in Section 4.4 entitled "Pragmatic Validation".

[10] A bibliometric analysis of the scientific literature on model validation (Eker et al. 2019) identified Oreskes et al. (1994) as "a highly-cited article" which is "acknowledged not only in the general modeling literature but also in the specific validation literature."

calibration are reasonably well reproduced, the model cannot be considered validated. This is referred to as the fallacy of "affirming the consequent," in which a necessary condition – matching the data – is mistaken as being a sufficient condition – confirming the veracity of the model. While attaining empirical consistency between the model-calculated and measured data (or qualitative observations) may increase the confidence in the model, it does not confirm that a particular model captures the natural world it seeks to represent. Such confirmation is always partial, i.e., it only supports the probability of the model's utility relative to alternative models or approaches proposed to gain insights or make predictions. Oreskes et al. (1994) consider this terminology – verification, validation, and confirmation – potentially misleading, specifically when used to indicate that the results of a numerical model are reliable enough to support important public policy decisions. They acknowledge that models may be useful to corroborate a hypothesis, to reveal discrepancies in other models, to perform sensitivity analyses, and to guide further studies. They conclude that models should be used to challenge existing formulations, rather than to validate or verify their ability to make predictions about a physical system.

Another criticism comes from a direct comparison – as part of a post audit – of relevant observations with predictions made by "validated" models that were developed specifically to make these predictions. For example, Konikow and Bredehoeft (1992) and Bredehoeft (2005) found that a significant fraction of models made poor predictions because of conceptual modelling errors.[11] In these cases, new data revealed that the prevailing conceptual model was invalid, i.e., the models did not just require minor adjustments of input parameters, but a fundamental change[12] in key aspects of the conceptual model.[13]

Using alternative models may reveal the impact of such conceptual model uncertainties. For example, based on a consistent set of characterization data, multiple alternative conceptual models of fracture flow and bentonite hydration were developed as part of Task 8 of the SKB Task Force. Not only did the predicted bentonite-saturation times vary over a relatively wide range, but the modelling teams also developed different views regarding the key factors affecting the overall system behaviour and, consequently, made different recommendations about research and site characterisation needs (Finsterle et al. 2019).

Differences or inconsistencies between reality and its representation in a numerical model are inherent in the modelling process and thus unavoidable. Any model is an abstraction of the real system, which implies that it is based on conceptual decisions, the choice of simplifying assumptions and the selection of input parameters with different levels of uncertainty. Whether the errors introduced by such simplifications and deficiencies can be considered acceptable depends fundamentally on the intended purpose of the model. This is the reason why the conceptualization is the key step in model development and also the main target of a critical model validation effort.

[11] The examples provided in Konikow and Bredehoeft (1992) and Bredehoeft (2005) define wrong assumptions about future forcing terms (e.g., recharge or pumping rates) as conceptual modelling errors. De Marsily et al. (1992) argue that it is not the intended purpose of the groundwater model to predict climate variability or aquifer exploitation, so the model can indeed provide reasonable predictions if the correct external forcing terms are specified. Both the original contention and the counterargument by de Marsily et al. (1992) make some assumptions about the modelling purpose, and which aspects of the real system are considered an inherent part of the model or external to it, to be provided by boundary conditions and forcing terms.

[12] Bredehoeft (2005) describes the changes required as "paradigm shifts."

[13] Both nuclear waste repository sites in the United States were afflicted by large surprises: at the Waste Isolation Pilot Plant (WIPP), brine flow through the salt and into the disposal caverns was deemed impossible until observed; at Yucca Mountain, fast flow through the unsaturated fracture network was not considered relevant until ^{36}Cl generated from nuclear bomb testing was observed in the Exploratory Studies Facility, requiring a fundamental change in the conceptualization of unsaturated fracture flow and fracture-matrix interaction.

As illustrated in this short summary, the mere possibility of model validation is being questioned based on philosophical, historical and practical considerations. While the details of these arguments depend on the definition of the term “validation” and the claims ascribed to a “validated model”, the various critics arrive at similar conclusions and recommendations:

1. It is fundamentally impossible to confirm that a site-specific model properly represents the natural system.[14]
2. Models should not be used for predictive purposes, unless the prediction domain is commensurable with the calibration domain; however, models are useful to challenge the conceptual understanding, examine assumptions, explore what-if scenarios, and perform sensitivity analyses.[15]
3. The term “validation” and similar terms should not be used, as they give a misleading impression of predictive model capabilities.[16]

4.3 The need for validation

4.3.1 Overview

Hydrogeological process models are typically based on well-established empirical laws. Moreover, the physical and conceptual boundaries within which the given laws can be considered acceptable for practical purposes are relatively well understood. These laws and their interactions with each other are described by a mathematical model and implemented using an appropriate numerical scheme into a computer code.[17]

The mathematical model typically consists of a set of coupled partial differential equations. These governing equations contain coefficients that emerged as the empirical laws are derived or upscaled[18] from more fundamental descriptions of physical or chemical processes. These coefficients are often unknown, uncertain, problem- and site-specific input parameters to the simulator. They reflect material properties, but also geometrical aspects or initial and boundary conditions.

[14] “Ground-water models cannot be validated” (Konikow and Bredehoeft 1992); direct counterarguments are presented by de Marsily et al. (1992); “…(validation) is impossible to achieve for geo-hydrological models” (Hassanizadeh and Carrera 1992); “Verification and validation of numerical models of natural systems is impossible” (Oreskes et al. 1994); “There is some controversy about the extent to which model validation can be achieved, particularly in relation to modelling the long term migration of radionuclides from radioactive waste in repositories” (IAEA 2018).

[15] “It is naïve to believe that we will somehow validate a computer model so that it will make accurate predictions of system responses far into the future” (Konikow and Bredehoeft 1992); “(M)odels are most useful when they are used to challenge existing formulations, rather than to validate or verify them” (Oreskes et al. 1994); ”(M)odeling (is) counterproductive in that it offers the illusion of accurate predictions” (Saltelli and Funtowicz 2014); “(S)ince models can never accurately represent reality, they should not be used for predicting the future” (Eker et al. 2019).

[16] “Using the terms validation and verification are misleading, at best.” (Bredehoeft and Konikow 1993); direct counterarguments are presented in McCombie and McKinley (1993) and Bair (1994); “The terms verification and validation are now being used by scientists in ways that are contradictory and misleading” (Oreskes et al. 1994); “Validation implies an exercise in legitimation, and this is precisely what the public fears” (Oreskes 1998); “(C)urrent modeling practices (…) are a significant threat to the legitimacy and the utility of science in contested policy environments” (Saltelli and Funtowicz 2014).

[17] Testing of the correct implementation of the mathematical model into a software package is often referred to as “verification.” In addition, convergence studies are performed to confirm that the numerical discretization of space and time as well as all computational parameters are properly set to arrive at a solution that is accurate (within the bounds of its numerical approximation), i.e., does not suffer from unacceptable numerical artifacts.

[18] New parameters may emerge as the support scale is increased. These new parameters reflect a property that does not exist on the smaller scale, and smaller-scale properties may disappear as they are lumped into the new parameter. By further increasing the scale, the value of the parameter may change (specifically in heterogeneous systems), resulting in a scale-dependent parameter value; however, the meaning of the parameter remains unchanged.

Despite the use of physics-based laws, hydrogeological models include a large collection of "auxiliary hypotheses," many of which are untested or even impossible to confirm. This problem is acknowledged (and partly mitigated) by the fact that the parameters of a site-specific hydrogeological model are adjusted and determined by calibrating the model against observational data. These estimates are inherently uncertain, but – more importantly – may be ambiguous or biased for the following reasons:

- The mathematical model and/or auxiliary hypotheses are incomplete or are poor representations of the system to be modelled.
- There is a discrepancy in the definition, state, location, or scale of the calculated model output variable and the corresponding observation used for model calibration.
- Measured data have an error component that is random and/or systematic.
- The model output has an error component that is random and/or systematic; systematic errors include errors in the conceptual model, (over)simplifications in the model structure (processes and features), model truncation errors, reduction in model dimensionality, symmetry assumptions, errors in initial and boundary conditions, etc.
- Data sets are incomplete, and the inverse problem is either underdetermined or regularized using an artificial or erroneous regularization term.
- The data are not sufficiently informative about the parameters of interest, or the available data are not discriminative enough to sufficiently reduce correlations among the parameters.
- Alternative conceptual models exist that are equally capable of reproducing the calibration data.

It is important to realize that such ambiguities and biases may remain undiscovered, specifically if the model is able to accurately reproduce historical data after model calibration. As long as the model is only used for predictions with conditions that are very similar to those prevalent during the collection of calibration data, it is likely that prediction results are acceptable (referred to as "interpolative prediction"). However, this drastically limits the applicability of the model, whose main purpose is not to reproduce the system behaviour that is already revealed by the measured data, but to examine its behaviour under different conditions[19], to explore unobservable variables, or to understand the underlying processes (referred to as "extrapolative prediction"). Any of these application modes contains an extrapolation – regarding conditions, processes, states, spatial and temporal scales, etc. – and potentially also leaving the realm of established theory and fundamental understanding.

One might argue that the need for model validation arises whenever such an extrapolation from one model space to another is attempted. This pertains specifically to the step when we proceed from model calibration to model prediction: at this point, we leave the space where model development makes use of measured data of the system we want to simulate – be these deterministic or statistical conditioning data, prior information about parameters, site-characterization data, testing and monitoring data, and calibration data. Calibration produces effective parameters, i.e., parameters that are process-specific, model-related and scale-dependent. Whenever the model structure, key processes or scales are changed to adapt the calibrated model to a particular problem, the interpretation, reference frame, and numerical value of the effective parameters are likely to change as well – thus there is the need for validating the appropriateness of the prediction model for its intended use.

This notion is reflected in all validation approaches that recommend data-splitting for a prediction-outcome comparison, which in essence attempts to emulate the situation in which a model is used for predictive purposes outside its calibration space.[20] The model's ability to make reliable predictions is tested by comparing calculated quantities of interest to corresponding observation data.

[19] This includes dynamical property changes induced by future processes such as erosion, glaciation, uplift, etc.

[20] It should be noted that data-splitting is often used with time-series data, which means the calibration and validation data sets are usually of the same type, observed at the same location, and referring to the same reference scale. This similarity between calibration and prediction data limits the application range for which the model is tested, as only a minor extrapolation is examined.

While confidence in a model may be increased by critically examining the process of model development – i.e., without relying on a comparison of model results with measurable quantities of interest – a test of the model's ability to make reliable predictions is an essential part of most validation methodologies,[21] whether they are proposed as part of a philosophical argument or for pragmatic validation of numerical models used for the licensing of a nuclear waste repository.

To further elucidate the reasons why model validation is a necessary step within the overall model development process, it is useful to introduce the concept of *model space*, illustrated in the context of modelling a fractured formation. A more detailed discussion of this concept can be found in Chapter 3.

4.3.2 Evolution of Model Space

Models in geosciences, including discrete fracture network (DFN) models, are simplified representations of the pertinent features and processes affecting the characteristics or behaviour that are the object of the study.[22] This modelling outcome of interest can be a general insight or a specific prediction of a state variable.

Confidence in modelling-based inferences can be gained by constructing models that minimize deviations from relevant data sets and assumptions the model is conditioned to, while simultaneously complying with conditions required to fulfil the purpose of the model.

The *model space* is the envelope of all possible models (see Chapter 3). They describe the bounds on the system of interest, from which relevant parameterisations, idealisations and modelling principles are chosen. It is thus the ensemble of conceptualizations, physical rules, mathematical equations and parameters that give a theoretical, observational and/or empirical description of the processes and conditions that allow simulations of the system state and its evolution.[23]

The model space evolves as new information is gathered or different requirements are placed on the model. Generally, the model space is reduced during the conditioning and calibration activities but may then expand when applied to the different prediction spaces. Conditioning constrains the model space by making assumptions and tailoring it to prior information about a particular site. Calibration is the process of reducing the model space by comparing model outputs to measured data, which can be viewed as assessing the probability of a chosen model parameter set to be consistent with observations. A mismatch indicates that the specific conceptual model chosen for the analysis is an unlikely representation of the real system. In general, adding site-specific information allows us to separate more likely from less likely conceptualizations of the system, thus narrowing the model space.

---

[21] "Experiments performed expressively for the purpose of model validation are the key to validation" (ASME 2006); "The process of determining whether a model is an adequate representation of the real system being modelled, by comparing the predictions of the model with observations of the real system" (IAEA 2018).

[22] It is essential to achieve an appropriate level of model complexity, which is a trade-off between the complexity of the real system as far as it can be observed, and the need to be commensurate with the requirements for the model and the support provided by the available data. A model can be oversimplified or overly complex (i.e., overparameterized). An oversimplified model fails to capture the salient features of the system to be modelled, which likely leads to systematically wrong or overconfident predictions. Conversely, while an overparameterized model is fundamentally able to better fit the data (at the risk of overfitting), it results in highly correlated, highly uncertain parameter estimates that lead to model predictions that are also highly uncertain and unreliable (Hawkins 2004). While sensitivity analyses can help assess the appropriate level of complexity, they obviously cannot identify potentially relevant features or processes that are not implemented in an oversimplified model, and they cannot readily examine parameter correlations and their impact on estimation and prediction uncertainties as they emerge in an overparameterized model. While modellers typically start with a simple model, adding complexity as new insights or data become available, one might argue that the appropriate model complexity is best approached by starting with a relatively complex model and then using notional inversions with sub-space methods to screen out irrelevant or unsupported model components to arrive at a simpler model (Hunt et al. 2007, Christensen and Doherty 2008).

[23] Beven (2009) defines the model space as "a hyperspace defined by the ranges of feasible models and parameter values, with dimensions for each parameter within each model."

However, this narrowing also limits the application range of the model. To make a useful prediction, an extrapolation must be made from the calibration spaces to the targeted prediction space[24]; the conditions to be represented for a prediction are by definition different from those prevailing during model calibration.[25] The need to extrapolate to different spatial and temporal scales, different boundary conditions, and potentially different key processes widens the model space; it is the reason why model validation is necessary.

If the model at a given stage of a project does not satisfactorily explain or reproduce observations, showing systematic deviations or large residuals, or its calibrated parameters are inconsistent with prior information and substantiated assumptions, the first step of the modelling process, i.e., model selection (identification of new models or extension of existing models) needs to be repeated in view of the new information and experience gained from the previous analyses. Predictions made with multiple alternative models, which cover different model spaces, are more likely to adequately span outcomes in the real system, with the caveat that all of them may be nonbehavioral (Beven 2006).

4.3.3 The object of validation: What is being validated?

Validation can be viewed as a critical review of the model-development process with the aim to demonstrate that the targeted prediction space is adequately delineated by an ensemble of model outputs.

The acceptable shape and extent of the prediction space is determined by the purpose of the model – the more specific the modelling objectives are, the narrower is the targeted prediction space, and the more stringent are the validation acceptance criteria. The targeted prediction space may be represented by observations of the real system under conditions that are similar to those affecting the unknown behaviour of interest.

The validation process is intended to reduce the number of conceptual models and their associated model spaces to a set of "behavioural models" (Beven 2002), thus increasing the confidence in our assessment of the models' strengths and limits (Saltelli et al. 2020).

The concept of a model space and its evolution reveals that the object of validation is not a single numerical model, but the outcome produced by an ensemble of models. Because each alternative model has its distinct model space, the prediction space is considerably wider if multiple models and their uncertainties are considered. If predictions made with alternative conceptualisations and methods do not diverge, but instead occupy a sufficiently small prediction space, then higher confidence can be gained that the ensemble of these predictions can be used as the basis for decision making (Munafò and Smith 2018).

In this view, the object of validation is the *prediction space generated by an ensemble of models* – rather than a specific model. The short term "model validation" as used in this document shall refer to this model-development process and the interpretation outlined in this section.

4.4 Pragmatic validation

As indicated in Section 5.1, this White Paper is mainly concerned with the practical aspects of validating numerical models that are used to support important public policy decisions, specifically those related to the long-term safety of a nuclear waste repository. This might be referred to as "pragmatic validation," as the goal is not to make assertions about the ultimate truth, but merely to build confidence in the model's ability to make reliable statements about a specific aspect of the

[24] When using "prediction-outcome" exercises as part of model testing, there may be two relevant prediction spaces: (1) *Prediction Space for Model Testing*: the conditions under which the validation data set used to test the model have been obtained; and (2) *Prediction Space of Model Application*: the conditions relevant to the model purpose, where the actual outcome is unknown.

[25] Describing the conditions of calibration is essential as it provides important information for end-users of the calibrated model, including its use during model validation, an effort that needs to be defined based on the discrepancies between the calibration and prediction spaces.

repository system.[26] It is an acknowledgment that finding the truth remains elusive but that a critical fit-for-purpose assessment of a model is both crucial and valuable. Confidence can be built by showing that a model is not incorrect. This view also acknowledges the fact that any model always contains residual uncertainty.

Pragmatic validation is demanding: the effort cannot just be abandoned because achieving truth or full confidence in predictions is recognized as a futile undertaking; instead, the inherent limitations of a model and the uncertainties in its predictions must be understood, and the domain of applicability must be determined and related to the intended use of the model. Finally, this information must be effectively communicated to the end-user of the model. Conversely, pragmatic validation limits the domain of model applicability, which in turn reduces the space of influential parameters, making its exploration more tangible.

The word "pragmatic" may also refer to the fact that validation of models representing subsurface systems is constrained by the scarcity of data; this also means that the validation process can help identify which data should be collected to increase model confidence by reducing prediction uncertainty.[27] In this interpretation, the term "pragmatic" is an acknowledgment of both the validation challenge, which must be accounted for when qualifying the credibility and applicability of a validated model, and the usefulness of the validation process itself, which helps identify data-collection and research needs for an improvement of system understanding and the reliability of model predictions.

Box (1976) coined the aphorism "All models are wrong, but some models are useful". Box offered the following explanation (Box et al. 2009): "All models are approximations. Assumptions, whether implied or clearly stated, are never exactly true. All models are wrong, but some models are useful. So the question you need to ask is not 'Is the model true?' (it never is) but 'Is the model good enough for this particular application?' " These comments are relevant also for geoscientific modelling, where the abstraction process during conceptual model development introduces numerous, often strong approximations, and many assumptions are being made due to generally poor coverage of characterization data[28] and the need to limit explicit implementation of multi-scale features and processes. Box's aphorism can be viewed as a concise statement about the potential backdrop of a pragmatic approach to model validation.

Finally, because of the recognised fundamental limitations and practical constraints, pragmatic validation could also refer to the validation approach itself. It indicates that the chosen approach is clearly targeted for a specific model use or to calculate specific quantities of interest, i.e., a small subset of the state that fully describes the behaviour of the entire system. It also indirectly comments on the effort that should be expended to appraise a model. The question arises what effort is required or can be considered reasonable. For example: Is it sufficient to validate a model by benchmarking it against other models, or by testing just its individual components, or by comparing its outcome to literature data, or is it

[26] "We do not want certainty; we will be satisfied with engineering confidence" (de Marsily et al. 1992).

[27] This relation between measured data, model parameters, and confidence in predictions is also formally examined in a data-worth analysis, which evaluates the relative contribution of an actual or potential data point to uncertainty reduction in (a) parameters inferred from the data through inverse modelling, and/or (b) in a target prediction of interest, which reflects the modelling purpose. A data-worth analysis takes place in the data space as well as multiple model spaces. It propagates data uncertainty to parameter uncertainty to prediction uncertainty, a process that examines sensitivities and information content of individual data points and the influence of parameters on model predictions. The relative importance of competing target predictions is also accounted for. The workflow for a pragmatic model validation and a data-worth analysis are thus similar. In fact, a data-worth analysis should be an integral part of model validation, demonstrating that data used for a prediction-outcome model validation approach are indeed informative and related to the ultimate modelling purpose. Note that a data-worth analysis would be performed prior to the collection of validation data; however, it forces the user to think about validation acceptance criteria, and to apply them once the data become available. Some background on the data-worth analysis workflow is described in Dausman et al. (2010), Fienen et al. (2010), and Finsterle (2015).

[28] It is acknowledged that extensive site investigation and testing programs generate large amounts of characterization data. Nevertheless, their spatial coverage remains sparse (partly intentionally to minimize site disturbances). If nonintrusive methods are used, the spatial resolution is relatively coarse, limiting the scale on which spatial variability can be described; if intrusive methods are used, the data are affected by local perturbations, which requires making additional assumptions during their interpretation.

necessary to perform a designated laboratory experiment or field test, or must an effort be made to demonstrate that the model is capable of performing over the entire spatial and temporal scales relevant for nuclear waste disposal? Merely addressing such questions indicates that a pragmatic approach is being taken, and the answer about what effort is considered reasonable is driven by the ultimate purpose of the model and its significance for decision making, specifically in areas where supporting information is uncertain or disputed, conclusive scientific evidence is not available, and the model outcome has important implications affecting a plurality of stakeholders.

Saltelli et al. (2013) outline a protocol to be used for a critical appraisal of a model's quality. The process they propose can be described as pragmatic in the sense that it provides practical guidelines with the aim to improve the quality of models used for the explicit purpose of supporting important policy decisions that involve considerable risks as well as unquantifiable, irreducible uncertainties. The approach is referred to as "sensitivity auditing" and goes beyond a mere evaluation of model uncertainties and parametric sensitivities. It is intended to sceptically review any inference made by the simulations. It attempts to establish whether a model is plausible regarding its assumptions, outcome and usage. It not only examines the model, but also the auditing process itself. Rules are formulated and checklists generated for the entire modelling process to achieve transparency about a prediction's reliability. The process is elaborate and multifaceted, bringing together cross-disciplinary perspectives to help address the complexity of an issue.

The approach essentially consists of seven rules and a checklist in which each claim of a model is qualified based on a comprehensive uncertainty assessment, which examines technical, methodological, epistemological, and societal uncertainties. The formal process includes a global sensitivity analysis (Saltelli et al. 2008) to identify the key factors affecting prediction uncertainty.[29] Value-laden assumptions as well as other model- and problem-related statements are then systematically qualified using the Numeral Unit Spread Assessment Pedigree (NUSAP) system (van der Sluijs et al. 2005). The first three of the five qualifiers (the numerical value of the claimed quantity, its units, and the spread, which is a measure of error) can be obtained using relatively conventional, quantitative uncertainty analysis methods. The remaining two, assessment and pedigree, constitute the more qualitative side of the approach. "Assessment" expresses qualitative judgments about the information, such as its reliability.[30] Finally, "pedigree" conveys an evaluative account of how the information was produced. Problem-specific pedigree criteria (such as theoretical understanding, empirical basis, methodological rigor, degree of validation, use of standards, quality control and safety culture, plausibility, influence on results, comparison of alternative conceptual models, agreement among peers, review process, value ladenness of estimates and assumptions and problem framing, and influence by situational limitations[31]) are evaluated by assigning a numerical value to linguistic descriptions[32] of the level to which each criterion is met.[33] The NUSAP system provides insight on two independent uncertainty aspects of a model-calculated number, one expressing its exactness, and the other expressing the methodological and epistemological limitations of the underlying knowledge base. These two aspects must be viewed together to arrive at a meaningful statement about a model's quality. For example, inexactness or even ignorance about an input parameter may not invalidate a model if the parameter is non-influential, i.e., has a negligible effect on the prediction of interest. Conversely, model predictions can be reliable even if they are highly sensitive to certain input parameters, provided that these parameters can be determined with high confidence.

---

[29] Global sensitivity analysis methods assess the relative importance of input factors in terms of the impact of their uncertainty on the model output. This assessment is done at multiple points in the parameter space to account for non-linearities, non-additivity and interaction effects. (A similar extension from a local to a global evaluation is also advisable for a data-worth analysis.)

[30] An assessment may be expressed numerically as a statistical significance level, or random and systematic error components, or linguistically as "optimistic" or "pessimistic", etc.

[31] Examples: time, money, computational resources, access to classified information, ability to travel.

[32] For example, for the pedigree criterion 'degree of validation', the description may range from 'compared with independent measurements of the same variable' to 'compared with derived quantities from measurements of a proxy variable' to 'weak, indirect, or no validation'.

[33] This leads to a problem-specific pedigree matrix used to calculate a composite score of qualitative expert judgments.

Sensitivity auditing consists of a set of rules that help to make a model, its quality, and the relative impact of uncertainties on the simulation results more transparent and accessible, so that inferences made based on the model can be evaluated by decision makers and stakeholders. The seven rules of Saltelli et al. (2013) and Saltelli and Funtowitz (2014) are briefly discussed as guidelines for practicing scepticism during pragmatic model validation.

Rule 1: "Use models to clarify, not to obscure" argues for transparency both regarding the model itself and its intended usage. Overparameterized models or models consisting of long chains of interconnected submodels are given as examples where complexity makes it difficult or impossible to evaluate the model due to non-uniqueness and a lack of transparency. Parsimony is considered essential.[34] The search for the appropriate level of model complexity can be supported by a comprehensive sensitivity analysis. The rule counters the observation that models are sometimes deliberately used to obfuscate an issue.

Rule 2: "Adopt an 'assumption hunting' attitude" means that the evaluation of a model should focus on identifying explicit and implicit assumptions[35] in the calculation chain and the potential value-ladenness of key assumptions. Sensitivity auditing should highlight implicit assumptions that are either implausible or contentious; possible alternatives should be examined, and features and processes considered irrelevant should be discussed. Assumptions with a weak pedigree and a strong sensitivity on the model-based inference require the most attention.

Rule 3: "Detect pseudoscience" warns against artificially deflating the uncertainties in the assumptions or model inputs in order to avoid that the distribution of the inference becomes so flat as to be indeterminate. Uncertainties can be deliberately under- or overestimated[36] so that preferred inferences appear to be linked to the model output.

Rule 4: "Find sensitive assumptions before they find you" suggests that a sensitivity analysis can be used to anticipate a critique. The after-the-fact identification of problematic or arguable assumptions may damage the credibility of an otherwise useful modelling exercise.

Rule 5: "Aim for transparency", as transparency enables both modellers and stakeholders to better understand how assumptions influence model outcomes.

Rule 6: "Do the right sums", i.e., rather than focusing on solving the problem very accurately ("do the sums right"), make sure the relevant problem is being solved.[37]

Rule 7: "Focus the analysis" implies that the analysis should target only one output variable, this being the relevant inference that the modelling study is trying to underpin. A global sensitivity analysis should be designed such that it exercises the entire evidential chain as opposed to covering one submodel at a time. This ensures that all interactions among factors in different compartments are being captured.

These seven rules are intended as review topics for sensitivity auditing and as minimum due-diligence requirements for the use of model-based inferences. The rules do not aim to reduce model uncertainty, but to make it transparent so that both modelling practitioners and recipients of modelling analyses are

[34] However, Oreskes et al. (1994) comment that "Ockham's razor is perhaps the most widely accepted example of an extraevidential consideration. Many scientists accept and apply the principle in their work, even though it is an entirely metaphysical assumption. There is scant empirical evidence that the world is actually simple or that simple accounts are more likely than complex ones to be true. Our commitment to simplicity is largely an inheritance of 17th-century theology." Note that while Oreskes et al. (1994) refer to truth statements, Saltelli et al. (2013) favour parsimony for the sake of increased transparency.

[35] Many implicit assumptions (such as the assumption of uniformitarianism or the applicability of basic flow and transport laws to very tight formations, fractured rock, small gradients and rates, and low concentrations) may become relevant when predicting the long-term behaviour of a nuclear waste repository.

[36] Overestimating uncertainties may lead to risk dilution. Uncertainty may also be artificially increased to render an opposing view meaningless.

[37] A diligently executed validation process may reveal our tendency to redefine (or refocus) a problem rather than to solve it.

fully aware of the conditionality of the predictions. Sensitivity auditing as described in Saltelli et al. (2013) critically examines the quality of a model used to make policy decisions – the process thus provides some guiding principles for pragmatic model validation.

In general, the notion that model validation is an auditing process guided by critical questions redirects the attention from a stringent pass-fail comparison of model-calculated and measured data to a broader evaluation of a model's adequacy through the judicious use of expert judgment as well as formal sensitivity and uncertainty analyses. Moreover, the model development and evaluation process need to be thoroughly documented and externally reviewed.

In this view, confidence is obtained by the fact that the validation process helps identify and correct obvious flaws in the model (Oreskes 1998), that hypotheses and assumptions are properly being tested (Luis and McLaughlin 1992), and that scientifically appropriate methods are being used (Neuman 1992). A rigorous validation process will ultimately improve the model and the quality of inferences and decisions made based on the model output.

The discussion of "pragmatic validation" reveals the wide range of interpretations and expectations the term "validation" evokes, with respect to both the ultimate purpose of the validation process and the most suitable method to achieve that goal. The essence of *pragmatic* validation is that it exposes a proposed modelling solution to the test of its usefulness. Expectations about what a validated model is supposed to accomplish are wide-ranging:

- A validated model provides an improved, general understanding of the system, because the model results are examined from disparate lines of evidence. However, the model results are not to be interpreted as predictions about the real system behaviour.
- A validated model is suitable to examine alternative cases and "what-if" scenarios. The model results are not accurate predictions but reveal relative changes in the expected system behaviour as a function of the chosen scenarios.
- A validated model is capable of making specific predictions that are adequate for the purpose of the model. The model does not necessarily represent the real system, but its outcomes are acceptable as they support the ultimate project purpose. For example, the model is used for conservative or bounding calculations, which – despite being unlikely, unreasonable, or even unphysical – may be adequate within a regulatory framework, e.g., as part of a performance assessment study.
- A validated model is an approximate representation of the real system and is thus capable of making specific predictions that are adequate for the purpose of the model. The degree of model fidelity is dictated by the accuracy with which the predictions need to be known to be of use for decision support.
- A validated model makes accurate statements about the absolute truth of the system.

The process of how to validate a model depends on which of the model-validation goals outlined in the preceding list shall be met. The validation process will be less elaborate and may be limited to component testing and peer review if the purpose of the model is to improve the general system understanding or to examine "what-if" scenarios; it likely requires comparison with experimental or observational data if decision-makers intend to rely on quantitative predictions; it will be an extensive, cross-disciplinary, continual research endeavour if fundamental statements about the nature of the world are to be made. The following activities may be part of a model validation exercise:

- A validated model should comply with industry-standard QA/QC procedures and have passed a formal software qualification lifecycle test ("verification").
- A validated model should have undergone a detailed review of the procedures used for the construction of the conceptual and numerical models, including the evaluation of available data, review of theoretical and empirical laws and principles, abstraction process and conceptual model development, building of the computational model, and the iterative refinement based on predictive simulations, sensitivity analyses, and uncertainty quantification.[38]

[38] See, for example, Tsang (1991) and Saltelli et al. (2013).

- A validated model should be calibrated against relevant data, with residuals being devoid of a significant systematic component, estimation uncertainties being acceptably low, and parameter correlations fairly weak. The criteria for acceptability are determined by the accuracy with which model outputs supporting the project objectives need to be calculated.[39]
- A validated model should be peer-reviewed with a general consensus among experts and stakeholders that the model qualifies for its intended use, and that limitations, the range of application, and uncertainties are sufficiently understood and documented.
- A validated model should be compared to alternative models[40] or approaches and perform equally well or superior regarding relevant validation performance criteria.[41]
- A validated model should reproduce relevant data not used for model calibration with acceptable accuracy. The type of data, the processes involved, the spatial and temporal scales, and the conditions prevailing during data collection should reflect those of the target predictions as closely as possible. The criteria for acceptability are determined by the accuracy with which model outputs supporting the project objectives need to be calculated.
- A validated model should demonstrate that it can predict emerging phenomena[42].

As indicated above, model-validation activities and related acceptance criteria vary as they are related to the demands imposed on the model. Figure 4-1 shows the approximate correspondence between the goals the validated model must meet and the required validation activities. It is self-evident that a good general system understanding is a pre-requisite also for a model that is expected to provide reliable, quantitative predictions of a previously unobserved system behaviour.[43] Similarly, following standards and best practices, and having the development of each model component checked individually and submitted to independent review are certainly ways to increase the credibility of a model, regardless of its ultimate use.

Pragmatic validation aims at demonstrating that a model is fit for purpose. This may lower the expectations of what the model needs to accomplish: it is not anticipated that the model can make accurate statements or predictions of any type and under any conditions; the model needs to perform only within a limited domain of applicability. On the other hand, the model is expected to provide useful information to solve a specific problem, not just insights about a general system behaviour.

Figure 4-1 also indicates the region within the matrix that IAEA's validation definition appears to target; this will be discussed in the next section.

[39] See, for example, Finsterle (2015), for a discussion of the relations between acceptable and attainable data uncertainty, acceptable and attainable parameter uncertainty, and acceptable and attainable prediction uncertainty.

[40] See, for example, Enemark et al. (2019) for a review of how to test alternative conceptual models.

[41] A series of international code comparison studies have been conducted with the common objective of "building confidence among experts, politicians and the general public on the use of performance assessment models and associated conceptual models and mathematical models in assessing the safety of nuclear waste repositories" (Larsson 1992).

[42] "Emergence" is a term used in the study of complex systems, where new laws, properties and phenomena emerge when the system as a whole has properties that its individual parts do not have on their own; they emerge only when the parts interact with each other. In physical systems, this mainly occurs during upscaling.

[43] This statement applies specifically to process models, which rely on an understanding of the underlying physical processes, as opposed to statistical models (including, for example, neural networks and machine-learning algorithms), which are data-driven approaches that infer input-output correlations with no or only a cursory use of physical concepts. The models of concern here are those that incorporate substantial mechanistic understanding rather than empirical correlations.

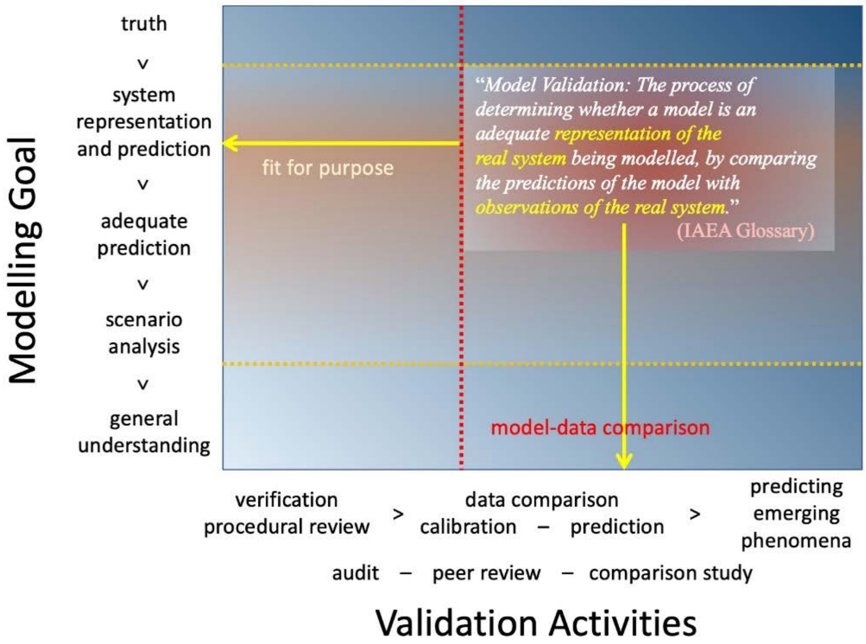


Figure 4-1. *Approximate relation between validation activities needed to reach a particular validation goal.*

4.5 IAEA Definition of model validation

The previous discussion revolves around fundamental and subtle differences in the interpretation of the term "validation". Nevertheless, definitions of validation in the context of quantitative modelling for decision support in a regulatory environment are quite consistent in the key elements they comment on. These elements include (a) the relation between the model and the real system, (b) the need for a comparison between model predictions and measured data, (c) the model's limited domain of applicability, and (d) the importance of uncertainty quantification. Some definitions emphasize one aspect over the other or are more or less prescriptive about the evidence that needs to be presented to satisfy validation acceptance criteria. This may change the requirements the model has to fulfil to be considered validated.

A frequently cited definition is the one provided by the IAEA; it is relevant for this White Paper, as it is tailored to the validation of models used to assess nuclear facilities, specifically nuclear waste repositories:

"Model Validation: The process of determining whether a model is an adequate representation of the real system being modelled, by comparing the predictions of the model with observations of the real system." (IAEA 2018).

This definition declares what the main purpose of model validation is ("*a model is an adequate representation of the real system*"), but also explicitly states what approach should be used ("*comparing the predictions of the model with observations*"). It therefore addresses models that have to meet relatively high expectations and thus require relatively extensive validation activities, as schematically indicated in Figure 4-1. Specifically, the definition does not just require that the model be fit-for-purpose (which would include bounding calculations), but that the model also adequately describes the real system.[44] Moreover, the definition implies that the model will be used for predictive purposes, and

[44] The term "adequate" may be interpreted as "sufficiently realistic", in which case the IAEA definition requests that a performance assessment model must explicitly represent all safety-relevant processes, i.e., that these processes cannot be replaced with overly simplified, potentially unrealistic or unphysical assumptions based on the assertion that the simplification will result in a conservative estimation of exposure dose. See also Davis et al. (1991). Also note that the term "validation" evokes "truth" rather than "adequacy", i.e., the term may be considered to be misleading. Oreskes (1998) recommends replacing the term "validation" with "evaluation".

that the ability of the model to make reliable predictions must be tested by comparison to observations of the real system. This holds a validated model to a higher standard than just providing broad insights into the general system behaviour; it consequently calls for a potentially elaborate validation process.

The high expectations and requirements resulting from this interpretation of the main part of the IAEA definition of validation are somewhat lessened when considering some of the notes associated with the definition. Specifically:

"Usually contrasted with model verification, although verification will often be a part of the broader process of validation." (IAEA 2018).

This note emphasizes that model validation goes beyond model verification, which is defined as "The process of determining whether a computational model correctly implements the intended conceptual model or mathematical model" (IAEA 2018).

"Modelling the behaviour of an engineered system in a geological disposal facility involves temporal scales and spatial scales for which no comparisons with system level tests are possible: models cannot be 'validated' for that which cannot be observed." (IAEA 2018).

This note appears to relieve to modeller from validating a system-level model of a nuclear waste repository because of the lack of validation data. While it is true that no observations exist that capture the long-term, large-scale response of the near field and far field to the presence of and effects from an actual nuclear waste repository, one might argue that data from sufficiently close proxy systems exist that capture relevant processes on large spatial and temporal scales. Important aspects of the conceptual model underlying a system-level performance assessment model could be tested using data from natural analogue sites or even the potential waste disposal site itself, collected by the detailed site investigations that occur during repository development. For example, isotopic data from deep groundwaters or from fluid inclusions, noble gas data, as well as geochemical and mineralogical data may indeed reflect conditions and processes occurring over geologic times and on regional scales that are comparable to those relevant for the assessment of the long-term safety of a nuclear waste repository. As with any other data used for the validation of the model, there is no complete congruence between the proxy system and the real system.[45] Nevertheless, these discrepancies or inconsistencies may be less pronounced in the case of natural analogue data than in the case of multiple short-term, small-scale data sets obtained from laboratory experiments used for the individual validation of the components of a complex performance assessment model. The use of natural analogue data for model validation is critically discussed, for example, in Ewing (1993), Côme and Chapman (1987), Miller et al. (1994), IAEA (1999, 2000), CRWMS (2000), and Marcos (2002); Alexander et al. (2015).

"'Model validation' in these circumstances implies showing that there is a basis for confidence in the model(s) by means of detailed external reviews and comparisons with appropriate field and laboratory tests, and comparisons with observations of tests and of analogous materials, conditions and geologies at the process level." (IAEA 2018)

This note expands the list of means that could be used to increase the confidence in a model. In particular, it refers to external reviews – without specifying the scope of such a review. As mentioned above, review activities are considered essential, specifically if done independently by multiple assessment groups. The note also reiterates the need to quantitatively compare model results with measured data, and it explicitly mentions natural analogues as a potential source of validation data.

"What is typically required by regulatory bodies is that such models of the behaviour of engineered systems in a geological disposal facility be shown to be 'fit for purpose'; this is typically called 'validation' in national regulations." (IAEA 2018)

This note suggests the model does not need to make statements about the absolute truth of every aspect of the repository system, but that it suffices if it addresses a particular question, which is typically reflected in one or a few composite performance metrics. The accuracy with which these answers have to be provided also depends on the ultimate purpose; in the context of public policy decisions, it is often defined by regulation, as also indicated in the note. One might argue

[45] If relevant data from the actual system of interest were available, there would most likely not be a need for extrapolation and thus modelling at all.

that this note specifically addresses the validation of system-level performance assessment models rather than submodels that examine specific subsystems or processes (such as channelling effects in fracture networks near a deposition hole). Nevertheless, the notion that a model can be validated for a particular purpose could be extended to any type of model, regardless of whether it is a process model developed to explain a physical phenomenon or the behaviour of a specific subsystem, and regardless of whether the performance metric is defined by regulatory bodies. The key is that the purpose of the model is clearly defined, preferably expressed by a readily quantifiable, composite performance metric, and that the accuracy with which that metric has to be calculated by the model is stated.

The final sentence of the note seems to indicate that this particular type of fit-for-purpose model validation is (or should be) the common interpretation of the term "validation" in the context of regulations dealing with nuclear waste disposal.

In summary, IAEA's main definition is a relatively straightforward statement of the purpose of validation, the object of validation, and the main method that should be used; it is also quite stringent in the demands placed on the models and the validation activities. The notes accompanying the definition suggest a somewhat broader and less stringent interpretation of the main definition.

Comparing the IAEA definition of model validation with the complementary definition proposed in Section 4.3.3 reveals overlaps but also differences. Both definitions recognize model validation as a (likely iterative) process rather than a simple calculation of a performance measure evaluated at the end of model development to be compared to a set of criteria. The purpose of model application is recognized as a key element that guides the acceptance criteria as well as the validation activities. The IAEA definition refers to a model's need to demonstrate its fitness for a particular purpose within the context of nuclear waste isolation, whereas the complementary definition states that the project goal defines the target prediction space, including its shape and size, which delineate the validation acceptance criteria.

However, there are also some differences in emphasis between the two definitions. It appears that the IAEA definition aims at validating individual models, with the presumption that a validated model will suitably represent the repository system and thus be able to make quantitative, predictive statements about its performance. The view presented in Section 4.3.3 places emphasis on establishing confidence in the prediction outcomes and the assessment of their uncertainties. It also suggests that this confidence is best attained by multiple realisations produced by an ensemble of models. Rather than validating the underlying models, the main focus is on delineating a prediction space that is sufficiently tight for the modelling results to be useful to decision makers.

Regarding validation activities, the IAEA definition explicitly calls out the need for comparisons of model results with observations of the real system, with external reviews mentioned in the notes. The complementary definition implies that a critical review of model conditioning and calibration as well as comparison of model results to data that represent the prediction space are essential elements of a comprehensive model audit. Note that this review extends to data used and analyses performed during the initial model development and iterative model refinement steps described in the previous essay (i.e., not just on an evaluation of predictions made by the completed model). As discussed in Chapter 3, proposed validation activities also include a critical assessment of assumptions and other modelling aspects that shed light on the uncertainties, potential biases, and limits of model the calculated modelling outcomes. The validation process thus promotes transparency and disclosure of all explicit and implicit suppositions that underly the models and potentially affect the outcomes.

4.6 Outline of a pragmatic model validation exercise

This section describes potential elements of a pragmatic validation approach that could be exercised as part of multi-team study. The overall goal of such an exercise is to examine how a formalized, pragmatic validation approach may help evaluate the quality of numerical models and potentially improve them by identifying uncertain factors – both conceptual and parametric – that significantly influence the model outcome and inferences drawn from it. While the quality of the initial models strongly depends on the available characterisation data, the validation process itself may point to aspects of the model that need to be refined, and in turn to additional characterisation data that need to be obtained. The exercise is likely to reveal fundamental aspects of model validation (including its limitations) and examines specific approaches and workflows of the process. A clear description of the context of the exercise makes it more realistic and allows accounting for the limited resources available. Finally, the exercise can take advantage of the fact that multiple modelling teams address the same system, likely using alternative or complementary conceptualisations and modelling tools.

As discussed in Chapter 3, the system of interest consists of a geological formation with discrete features that present either preferential paths or barriers to groundwater flow and radionuclide transport in the vicinity of a nuclear waste repository. Particular processes occurring within this network of discrete features have been identified as potentially relevant for the understanding and prediction of repository performance. Confidence in the appropriate representation of these features and processes within a numerical model needs to be established or increased, which can be regarded as the overall goal of the validation exercise.

It is essential to define a suitable framework for the exercise, so that full advantage can be taken of the complementary work conducted by multiple modelling teams. The framework could set the following boundaries:

- System description
  - Approximate location of model domain with respect to the repository.
  - Approximate extent of the model domain, indicating the scale of the features and processes that need to be considered.
  - Time period of interest (e.g., pre-closure, thermal period, post-thermal).
  - Approximate duration, indicating the temporal scale.
  - Subsystem components to be considered (components of the engineered and natural barrier system included in the model).
- Process description
  - Key processes to be considered (e.g., flow towards deposition hole during hydration phase, leakage processes, near-field flow and radionuclide transport, far-field flow and radionuclide transport, thermal effects, mechanical effects, chemical effects, etc.).
  - Processes within discrete features (e.g., channelling effects, flow and transport along and across feature intersections, fracture-matrix interaction).
- Data support
  - Available multi-scale characterisation data (statistics of feature geometry; property distribution within the feature and effective properties of an individual feature, large-scale feature connectivity; properties of engineered barrier components; initial and boundary conditions).
  - Actual or potential calibration data.
  - Potential validation data (from site, proxy laboratory or field experiments, natural analogues).
- Modelling purpose
  - Role of modelling in support of meeting regulatory requirements.
  - Goals to be supported by modelling.
  - Technical objectives of specific model.

Once this framework is described, the modellers need to determine which aspects of the model will require particular attention and thus warrant a targeted validation effort. These are aspects that have the greatest impact on the model outcome which define the performance metrics used to support the overall study goal. Moreover, the validation effort should focus on the subset of significant aspects that are uncertain or the modellers lack confidence in their correct or accurate representation in the model.

To address this part of the model validation process requires decisions about the following issues:

- Performance metrics and validation data
  - The performance metrics are quantities that best inform a decision maker about the subject of interest. It is calculated by the model once applied to a specific nuclear waste disposal problem (i.e., not to reproduce historical data, or to gain fundamental insights into the general system behaviour).

- The accuracy with which these target predictions need to be calculated are set by the project objectives, by decision makers, or by regulation. This accuracy can be a quantitative uncertainty measure or a linguistic expression, such as: qualitative behaviour, order-of-magnitude estimate, sufficiently sensitive to changes in key factors, prediction accuracy similar to data accuracy, of high precision, etc.
- Validation data must be both observable and computable by the model. They should be as close to the performance metrics as possible, in terms of influential factors, processes, and scale. The accuracy of both the model output and data must be sufficiently high, so they are discriminative in the evaluation of validation acceptance criteria.

• Influential factors
  - A list of factors that could significantly influence the modelling results. This list may include fundamental assumptions about the nature of the system behaviour, conceptual decisions about the representation of influential factors, modelling assumptions; reference parameters[46], parameter ranges, and uncertainties in the lower and upper bounds. A rather comprehensive and structured generic list of potentially influential factors (referred to as features, events and processes; FEPs) relevant to the assessment of the long-term safety of geologic repositories was developed by the OECD/NEA (NEA 2000, 2006). This FEPs list may be used as a starting point for the identification of potentially influential factors.
  - Accuracy with which these influential factors need to be determined so their variation within their uncertainty range allows one to make meaningful statements about the system behaviour.[47]

The identification of influential factors is an important step of any model development process; it is mainly addressed by sensitivity and scenario analyses. It is also part of the validation process, where factors influencing the performance metrics should also have a high influence on the validation data. The difference between (1) the influential factors identified during model development or model calibration and (2) the influential factors identified for model validation is an indication of the inherent differences between the two model spaces. It also denotes the degree of extrapolation undertaken when using a model for a purpose that may not have been envisioned during model development, and for which no closely related calibration data were available.

This step also includes uncertainty analyses, which involves two components:

• Uncertainty quantification
  - The uncertainty in an input factor reflects our confidence in its value (if it is a quantifiable model input parameter) or the probability or likelihood we assign to the existence and relevance a feature or process. The uncertainty of a parameter can be determined by analysing site characterisation data (by inverse modelling or sensitivity analyses); the uncertainty in assumptions may be evaluated through scenario analyses or expert elicitation.
  - The uncertainty in model predictions can be calculated using linear uncertainty propagation analysis, Monte Carlo simulations, as well as sensitivity and scenario analyses. This uncertainty is needed to determine whether the model output is consistent with or significantly different from corresponding validation data (uncertainties in the data also need to be accounted for in this assessment). Furthermore, this uncertainty, which is also referred to as the *attainable* prediction uncertainty given the current knowledge about the model and its input parameters, must be smaller than the *acceptable* uncertainty for the model to meet its objectives or to be of use to decision makers. If the attainable prediction uncertainty is greater than the acceptable prediction uncertainty, additional site characterisation data need to be collected. These data must contain sufficient information about the factors that are both influential and uncertain, so the uncertainty of these input factors can be sufficiently reduced, which consequently also reduces the uncertainty in the target predictions.

[46] Includes properties values, but more generally any aspect of the model that can be parameterised, such as initial and boundary conditions, geometrical aspects of the model structure, processes, uncertainties and systematic errors.

[47] This acceptable uncertainty determines the accuracy with which site characterization data must be measured.

Only a model that meets the accuracy requirement of the performance metrics can be declared "fit for purpose", provided that the validation data are sufficiently representative of these ultimate performance metrics. Whether this last condition is fulfilled can never be formally assessed in the context of long-term performance of a nuclear waste repository. Nevertheless, the closeness between validation data and ultimate model usage should be used as a key criterion when designing a validation study.

As all model predictions are extrapolations (spatially, temporally, parametric, and regarding the features and processes that need to be considered), and the validation data never fully correspond to the ultimate performance metrics, confidence in the model cannot solely rely on the comparison between model output and validation data. Instead, each model development step must be clearly documented in a "modeller's log" and reviewed[48]:

- Document and review the conceptual model and its assumptions, identifying potential biases in key model decisions.
- Document and review the model selection process through an audit against the FEPs list.
- Document and review the criteria used to reject a model or update it (e.g., calibration and validation acceptance criteria).
- Document the review process and its outcome, specifically regarding consensus or disagreement among the reviewers.

Finally, multiple models should be developed based on alternative conceptualisation. If these models yield consistent conclusions about the behaviour of interest, confidence can be gained that the performance metrics can be calculated in a robust manner (Munafò and Smith 2018). This indicates that the outcome does not greatly depend on uncertain factors, which may have been implemented differently in each of the models, but that the general system understanding as well as the information provided by the site characterisation data are sufficient to constrain the predictions.

Conversely, model comparison may also point to conceptual aspects that need to be revised. When combining or comparing alternative conceptual models, the performance of each model during the calibration and validation phases may be accounted for (see, for example, Neuman 2003). Such a combined analysis does not state which (if any) of the alternative models is the best representation of the real system (see also the concept of equifinality introduced by Beven (2002), making it a truly pragmatic approach. Examining validation approaches in a setting that includes multiple modelling teams is a unique opportunity to test the value of alternative conceptual models, all attempting to address a common issue based on a shared set of information.

While many computational toolsets exist that support certain steps of the validation process (for a review, see Matott et al. 2009), it is apparent that no single validation approach exists that is best regardless of the application area. Even within a specific domain, such as nuclear waste isolation, the approach has to be adapted to fit the model, the question the model is expected to answer, and the overall goal of its use. While validation has fundamental and practical limits, the exercise of trying to test a model in an attempt to find its weaknesses is a valuable if not necessary effort.

---

[48] The review criteria of Saltelli et al. (2013) described in Section 4.4 may serve as a guideline.

5 Pragmatic Validation Approach for Geomechanics, Flow, and Transport Models in Fractured Rock Masses

G.W. Lanyon, P. Davy, W.S. Dershowitz, S. Finsterle, B. Gylling, J.D. Hyman, I. Neretnieks, M. Uchida

This chapter is based on a paper (Lanyon et al. 2021) that was submitted to the DFNE session of the ARMA 2021 online conference after that DFNE 2021 was postponed. Compared to the paper this chapter is substantially updated and modified.

Abstract

The engineering application of geomechanical and hydrogeologic models for fractured rock masses requires confidence in the underlying concepts and assumptions used, and the characterization data that provides the basis for parameterization and boundary conditions. This is particularly critical for design and safety analyses of radioactive waste repositories, as they require isolation for extended periods. Although much is known about the engineering performance of fractured rock masses, there is still considerable uncertainty and variability in performance predictions, particularly when considering the heterogeneity and "channelling" of flow and solute transport in fractures and interactions with fracture infillings, coatings, and other immobile zones.

This chapter describes the development of a pragmatic validation approach for models of groundwater flow, solute transport and geomechanics considering the limitations of fractured rock site characterization. Because the intended applications include safe isolation of radioactive waste, the validation approach must consider the IAEA definition of validation (IAEA 2018). However, the validation approach must also be practical given the conceptual and characterization uncertainties involved and limitations of available data and conceptual approaches.

5.1 Introduction and scope

This chapter describes the development of a pragmatic validation approach for models of fractured rock mass flow, solute transport and geomechanics. The approach is being developed within the SKB Task Force on Groundwater Flow and Transport of Solutes (TF GWFTS; www.skb.se/taskforce) as part of Task 10, which considers how to validate and build confidence in models.

5.1.1 SKB GWFTS Task Force and Task 10

The SKB GWTS Task Force is a multi-lateral forum for modelling of groundwater flow and solute transport, focusing on issues of relevance for disposal of nuclear waste. While the early tasks were related to the construction of the Äspö Hard Rock Laboratory in Sweden and developed general modeling competence, the later tasks have served performance and safety assessment purposes in a more substantial manner. In this setting, the member organisations collaborate and interact in the area of conceptual and numerical modelling of groundwater flow and solute transport in fractured rock. The overall objective of the Task Force is "to increase the understanding of the processes that govern retention (retention here refers to both reversible and irreversible immobilisation processes) of radionuclides transported in crystalline rock and to increase the credibility in the computer models used for groundwater flow and radionuclide transport" (Gustafson et al. 2009).

The Task Force is currently working on Task 10, which focuses on validation approaches for groundwater flow and transport modelling with discrete features. This task initially considers the single fracture and small fracture network scales, including concepts of structure and heterogeneity within fractures and fracture networks (e.g., channelling and compartmentalization). The motivation for this work partly derives from the interest of the partner organisations in developing a methodology for model validation.

5.1.2 Geological repositories for radioactive waste

The management and disposal of high-level radioactive waste has been the focus of research efforts since the 1950s when the US Congress passed the 1954 Atomic Energy Act. Since then, geological disposal within a purpose-built final repository has been identified as the most promising approach (Hebel et al. 1978, USDOE Blue Ribbon Committee 2012).

In this chapter we have chosen the Swedish KBS-3V design (Figure 5-1) for the Forsmark Site (SKB 2008a, 2008b, Follin 2008, 2011) as an example of a proposed final repository for spent nuclear fuel in crystalline rock. The Forsmark repository will be located at a depth of around 500 m in the crystalline bedrock. Spent fuel rods will be encapsulated in copper canisters with cast iron inserts about 5 m long and the canisters, surrounded by compacted clay, will then be emplaced in large-diameter vertical boreholes from tunnels. The fractured rock at the site is water saturated at depths below a few meters.

The main safety functions of a KBS-3V repository are to "isolate the repository from the surface environment; contain radionuclides and to retain and retard their dispersion into the environment" (Posiva 2012, Posiva and SKB 2017). The host rock and excavations contribute to the main safety functions by providing isolation from the surface environment and by maintaining:

- favourable thermal conditions,
- mechanically stable conditions,
- chemically favourable conditions,
- favourable hydrogeological conditions to limit the transport of solutes.

Our interest here is in groundwater flow and solute transport and their influence on isolation from the surface environment, the maintenance of chemically and hydrogeologically favourable conditions, and the retardation and retention of radionuclides if containment were breached. In particular, we consider approaches to addressing discrete features such as fractures and faults and their role in flow and solute transport.

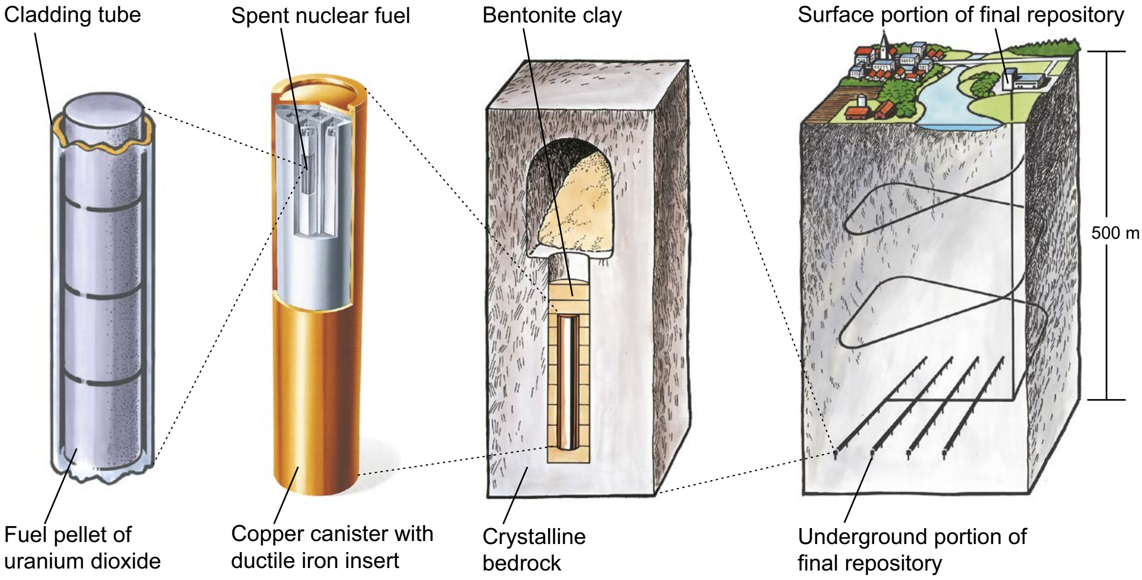


Figure 5-1. *The KBS-3V concept for disposal of spent nuclear fuel (SKB 2011).*

5.1.3 Definitions and validation of models

In this chapter, models are understood as "purposeful" representations of real-world systems (Starfield et al. 1990). Model validation is defined in the IAEA Safety Glossary (IAEA 2018) as:

"The process of testing whether a model is an adequate representation of the real system being modelled by comparing the predictions of the model with observations of the real system."

Parker (2020) proposes an account of model validation which requires assessing the relation the model holds not just to its intended target system (T)[49] but also the broader context of its use: the problem to solve or intended purpose (P), the way of use (W), intended users (U) and background circumstances (B). A model is "adequate for purpose" if it stands in a suitable relationship with all of these factors.[50] An important requirement for any model that we seek to validate is therefore that it has a clearly defined purpose and target.

Although the dominant processes of groundwater flow and transport in fractures are simple and well understood, uncertainty in the geometric structure of fault and fracture networks and the nature of the water conducting features (WCFs) within the network result in conceptual uncertainty in predictive models that needs to be addressed (Berkowitz 2002, Selroos et al. 2002, Neuman 2003, Bredehoeft 2003, 2005, Tsang et al. 2015).

Holling (1978) suggests a classification for models (or problems) based on data availability and process understanding (see Figure 5-2). In Region 1 there are good data but little understanding; in these circumstances statistics is the appropriate modelling tool. In Region 3 one has both good data and understanding; in such circumstances models can be built, validated and used with conviction. Regions 2 and 4 classify problems that are data-limited in the sense that the relevant data are unavailable or cannot easily be obtained. Starfield and Cundall (1988) argue that "problems in rock mechanics usually fall in the data-limited category; where one seldom knows enough about a rock mass to model it unambiguously". As a result, multiple conceptual models are needed to provide alternate descriptions of heterogeneity and how known processes interact with that heterogeneity. Within our approach, we consider an "ensemble of models" where all models share a common purpose, target and are likely to have similar contexts of use but may be based on different assumptions and data. We use the term "Model Space" (Beven 2006) to describe this ensemble of models and their associated parameter ranges and uncertainties, as shown in Figure 5-3.

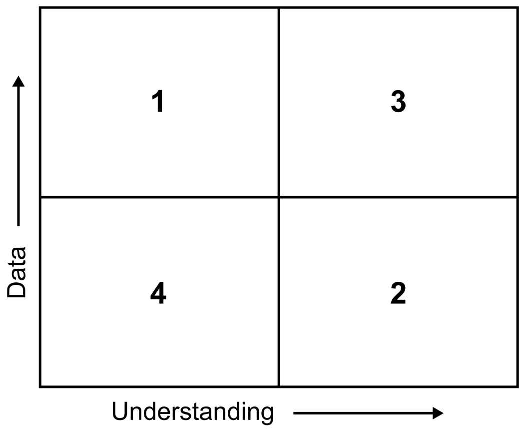


Figure 5-2. *Holling's classification of modelling problems (Holling 1978).*

[49] "Reality of Interest" in the terminology of Thacker et al. (2004).

[50] "Fitness for purpose" is an extension of "adequacy for purpose" that admits a degree of compliance such that a model may be more or less fit for a given purpose (adequacy is a discrete notion – a model is either adequate or not). Parker (2020) states that a model is fit for a given purpose if and only if it is adequate for that purpose.

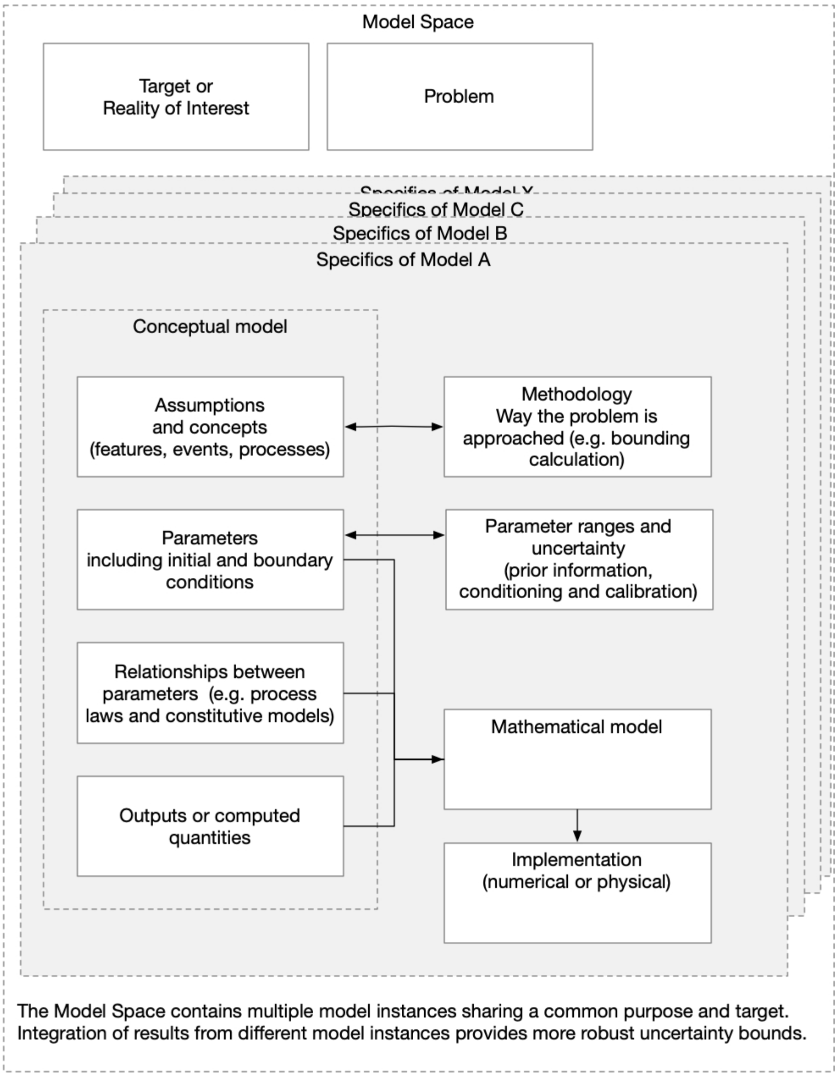


Figure 5-3. *The Model Space for a given purpose and target.*

The rest of the chapter will proceed as follows:

- In Sections 5.2 and 5.3, we discuss the geometric complexity of fractured rocks and interaction with flow and transport processes;
- In Section 5.4, we discuss modelling processes, and how the model space may be reduced by conditioning and calibration;
- In Section 5.5, we address problems surrounding model validation and set out our proposed pragmatic approach to validation.
- In Section 5.6, we characterise the activities and processes involved in pragmatic validation in more detail.

5.2 Models of flow and solute transport in fractured rock

This section presents how flow and solute transport can be modelled in fractured rock. It draws on experiences from modelling these processes in recent performance assessment (PA) projects in Sweden and Finland. A more general review of hydrogeological issues for radioactive waste repositories can be found in Tsang et al. (2015).

5.2.1 Fracture networks in crystalline rock

The Fennoscandian rock ranges in age from a little under a billion years to over 2.5 billion years. During this time it has been subjected to multiple tectonic events that have created the current fracture system including rock formation, unloading, glaciation, erosion and weathering. Different tectonic events have both reactivated existing features and created new ones (Stephens et al. 2015). This has resulted in the development of a number of well-defined fracture sets associated with common origins and orientation.

5.2.2 Modelling and describing the fracture network

Individual fractures can be described by their location, shape, size and orientation and form a complex three-dimensional network. Individual fractures may themselves be complex structures associated with multiple pore spaces, as illustrated in the microstructural model from the Äspö HRL TRUE site shown in Figure 5-4.

Smaller fracture zones and fractures intersecting outcrops at ground surface or in open excavations can also give information on size, orientation and fracture intensity, P_{10} [1/m]. Other metrics are; fracture areal intensity, P_{21} [m/m^2], fracture volumetric intensity, P_{32} [m^2/m^3], and fracture porosity P_{33} [m^3/m^3] (Dershowitz and Herda 1992). These metrics can also be assessed from core and image logs in deep boreholes. Core data also provides information on rock type, fracture aperture and mineralogy, characterising the past accumulated chemical alteration in and around fractures.

To model the geometric fracture network (with the exception of major faults or fracture zones), small fracture zones and fractures in the less-disturbed rock are commonly described by statistical distributions, correlations and spatial relationships derived from surface observations, geophysical, geological, and borehole information. Once underground facilities have been developed, a greater range of borehole directions (e.g. tunnel pilot boreholes) can be utilised and tunnel surface mapping provide information away from any surface weathering. Stochastic models can be locally conditioned at outcrops, tunnel surfaces and borehole intersections where measured data are available (Chiles 1987, Bym and Hermanson 2018, Appleyard et al. 2018). Stochastic approaches to account for uncertainties relating to major deformation zones outside the region sampled by boreholes (and tunnels etc) are also being considered by SKB.

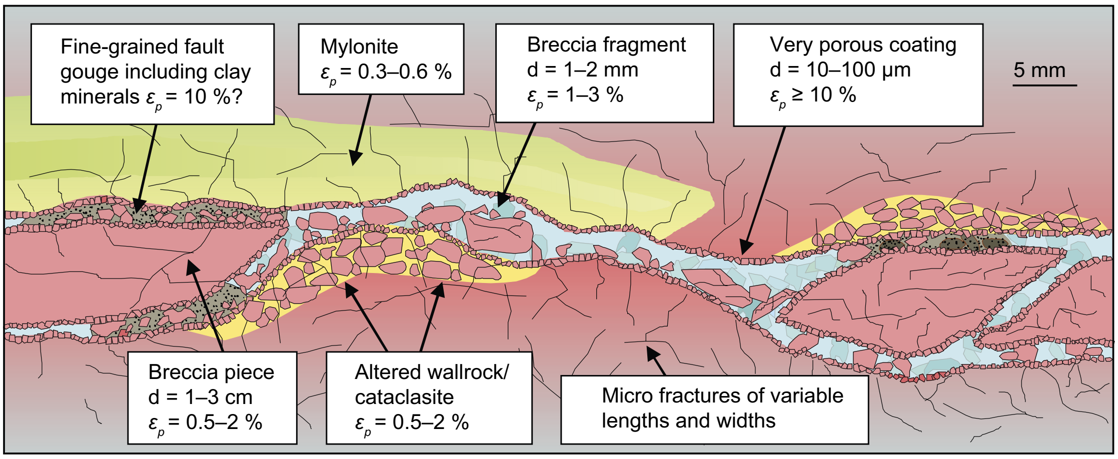


Figure 5-4. *Microstructural model of typical conductive structure and associated porosity from Äspö HRL TRUE site (Andersson et al. 2002). In this figure, ε_p is porosity and d represents the distance from the fracture surface.*

5.2.3 Modelling and describing the water conducting network

Smaller fractures may be fully closed (cemented) or partly closed or sealed with a spatially varying aperture. Flow through the fracture will seek out different preferential paths depending on the direction of the hydraulic gradient. Strong heterogeneity in permeability within the fracture and within the network, together with the large permeability contrast with the rock matrix generate complex flowpaths. These essentially two-dimensional paths are often called channels, although they are not strictly "channels" in the sense that they may have no physical borders except at the walls of the fracture and are rather a dynamic consequence of the interplay between aperture heterogeneity and hydraulic head boundary conditions (Tsang and Tsang 1989). It may be noted that often a "fracture" consists of a series of shorter sub-fractures en échelon in essentially the same plane. The sub-fractures can be hydraulically connected and it is therefore difficult to characterise fractures by borehole observations alone.

Fracture zones may contain breccia zones of strongly altered rock, including clays. They typically are associated with transmissive regions that may be connected over large distances, but clay-rich regions may also create local barriers to flow resulting in a complex conduit-barrier system (Caine et al. 1996).

At fracture intersections, water can flow from one fracture to the next where the intersection is not sealed or when the water can seek out a path around the sealed location. Fracture intersections can themselves also be conductive and facilitate flow from fracture to fracture along the intersections, making the flow network even more complex (Sanderson and Zhang 1999).

The fracture pore space within the rock can be conceived as a three-dimensional network of Water Conducting Features (WCF), where flow may occur over all or part (the channels) of a fracture for given boundary conditions. Water in other parts of the fracture pore space may be largely stagnant and accessed mainly by diffusion.

The connectivity of this WCF network controls the effective permeability of the network and exchanges between the advective flowpaths and less mobile water in stagnant zones and within the matrix. The connectivity can be constrained by observations of flowing fracture frequency (from PFL) and in specific circumstances by targeted cross-hole hydraulic testing (Follin 2008).

5.3 Flow and solute transport in fractures and in networks of fractures

Flow in conduits and in porous media is driven by hydraulic head gradients. The flux (flowrate per cross-sectional area) at very low Reynolds numbers is proportional to the gradient. These conditions typically apply to flow in narrow fractures in low-permeability crystalline rock.[51] It is impossible to image the complex WCF network at scales greater than a meter, considering the size of the rock volume to be injected by resin, over-cored and analysed, with present techniques and most WCF properties (e.g., location, size, orientation, transmissivity) must be treated as stochastic entities. A range of different characterization methods can be used to derive site-specific distributions for the properties and correlations between them.

Three main methods have been used to represent the WCF network: Discrete Fracture Networks (DFNs), Channel Network Models (CNM) and Equivalent Continuous Porous Media (ECPM). Each requires probabilistic descriptions of the WCF network and can be used to generate multiple (ideally many thousand) realisations of the network for which flow and transport simulations can be performed. The outputs from such models include volumetric or mass flowrates through volumes or across boundaries, flow and transport paths and surface discharge locations.

[51] High Reynolds numbers can potentially occur within fractures at their intersection with a borehole during well testing when large head differences are applied.

5.3.1 Flow and transport at the single fracture scale

Bodin et al. (2003a, 2003b) give an overview of fundamental mechanisms and mathematical relations for modelling of flow and transport in single "simple" fractures (or joints). By "simple" we mean fractures that essentially have two surfaces separated by a narrow aperture which may vary over the fracture area. Aperture variation affects the transmissivity, the solute transport and channelling properties. In reality, single fractures may be associated with multiple porosities (as shown in Figure 5-4). Such simple fractures have often been studied in laboratory experiments under a range of loading conditions. Witherspoon et al. (1980) found in large-scale laboratory fracture experiments that the transmissivity- hydraulic aperture relation for non-sheared fractures (joints) was well described by the cubic law (equivalent to laminar flow through a parallel plate, Zimmerman and Bodvarsson 1996), whereas Ishibashi et al. (2015) studied non-sheared as well as sheared fractures (faults) and found more complex relations. Figure 5-5 illustrates the extreme heterogeneity in transmissivity and flow that may occur within even simple single fractures.

5.3.2 Fracture intersections

Flow from fracture to fracture is influenced by the open aperture at and along the fracture intersection. Fracture intersections are likely to show similar levels of heterogeneity to the fracture surfaces but are difficult to characterise in situ or in the laboratory, such that often they are modelled using the following simple assumptions:

- The intersection is a high permeability "tube" that allows flow along the intersection as well as across it (Tsang 1991).
- Pressures in the two fractures are assumed to be continuous across the intersection and channels within the fractures.
- Channels are discontinuous across the intersection and significant pressure discontinuities may exist where flow (if any) occurs through low-transmissivity regions or through the rock matrix.

Transmissive fracture intersections with high flowrates are often observed in drifts and tunnels (Stanfors 1987, Palmquist and Stanfors 1987, Abelin et al. 1991a, 1991b), although flow in the EDZ and tunnel nearfield will inevitably occur under disturbed conditions which may influence the effect of any flow channeling. The effects of solute transfer at intersections on network-scale transport were analysed by Park et al. (2001) who found that the influence of overall network geometry was more important than the choice of mixing rule at fracture intersections (Berkowitz 2002).

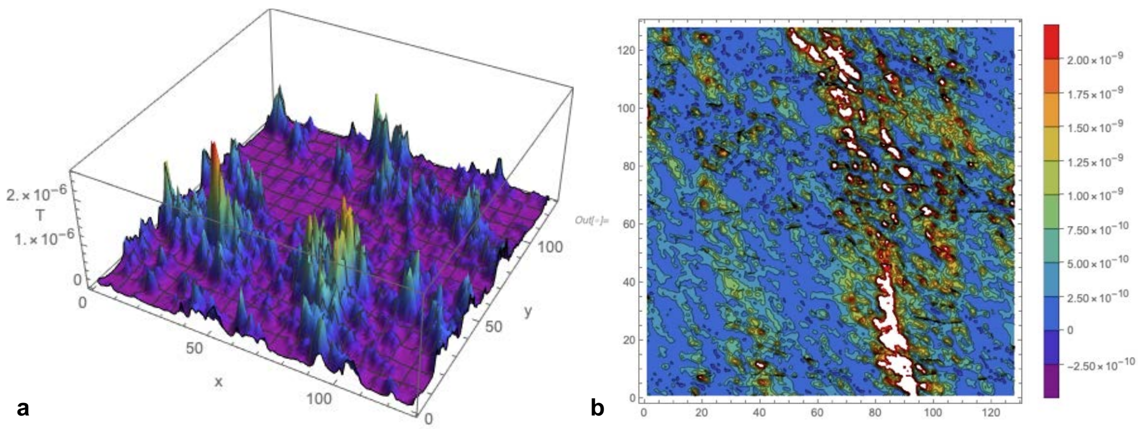


Figure 5-5. *Single fracture: a) transmissivity shown as surface plot coloured by transmissivity; b) colour contour plot of flowrate in the fracture. Flow is in the y-direction. Figures are in arbitrary units.*

5.3.3 Larger fractures and fracture zones

The larger fractures are the more complex the larger they become. Fracture (or deformation) zones contain crushed rock with very wide particle size distributions, ranging from the width of the zone down to a micron and less. Large-transmissive zones are likely to form the "backbone" for flow paths through the fractured bedrock. Hydraulic testing and chemical signatures support the assumption that fracture zones are typically more permeable than the background (less fractured) rock. Large fractures are also often assumed to be more transmissive than smaller fractures (Bonnet et al. 2001, Klimczak 2010, Hyman et al. 2016b) but field evidence is limited (e.g., observations of vein systems). The heterogeneous structure of complex fractures and fracture zones is particularly difficult to model because of the limited field data available to characterise them. Transmissive zones within large fractures and fracture zones may generate extensive channelling.

Fault and fracture zone complexity is also a strong control on fluid flow (Gartrell et al. 2004, Fossen and Rotevatn 2016, Dimmen et al. 2017, Peacock et al. 2018). Ligtenberg (2005) suggests that flow along faults may be localised in "weak zones" (step-overs between the fault segments, fault intersections and bends in fault zones). Nixon et al. (2020) studied the detailed topology of damage zone fractures in a carbonate-hosted fault and identified that locations of stress perturbations (fault bends and fault intersection points) produce complex zones within each damage zone and result in higher fracture intensities and connectivity evidenced by higher levels of fracture mineralisation.

5.3.4 Solute transport in fractures and in networks of fractures

Solute transport in both individual fractures and in fracture networks is influenced by molecular diffusion and chemical reactions such as sorption in addition to advection by water and diffusion into the rock matrix. Dispersion of solutes along the flowpaths can have both detrimental and beneficial effects on the transport of nuclides to the biosphere. Dilution could be beneficial in some situations, but early arrival of a fraction of a plume may not allow for sufficient decay even of shorter-lived nuclides.

Dispersion is caused by a number of different and interacting mechanisms (Neretnieks 1983). A small pulse of solute that is injected in stagnant water will spread slowly and be diluted over time by molecular diffusion. Given time, a non-interacting solute will eventually diffuse into all accessible voids. The diffusion distance increases proportionally to the square root of time. In seeping water, the spreading pulse is carried by the water, and some solute will travel faster than the velocity of the streamline in which it is carried because of diffusion in the flow-direction. In conduits, tubes or slots, the streamlines in the middle move more rapidly than those near the walls. For flow in *narrow* slots the diffusion across the slot will even out the concentration between different streamlines. The resulting pulse spreading (dispersion) can be described and quantified by the Taylor dispersion equation (Taylor 1953). Taylor´s basic idea can readily be extended from the original circular tube to other channel geometries. In such conduits longitudinal spreading behaves in a similar manner to molecular diffusion but with a (much) larger apparent diffusion coefficient.

In contrast, in large fractures of meters or more in extent, the distances between flow channels may be so large that negligible mixing by molecular diffusion takes place between the channels and a solute pulse injected into such a fracture arrives at the outflow at different times by different channels. The joint residence time distribution at the outflow, where waters in the different channels mix (e.g., a pumping interval[52]), shows "velocity dispersion". The spreading by this process is proportional to the travel distance, in contrast to that caused by molecular diffusion and Taylor-Aris dispersion.

Another important cause of spreading is "matrix diffusion" (Neretnieks 1980). In fractured crystalline rock the matrix is porous (containing microfractures and porous minerals, e.g. Voutilainen et al. 2012), albeit with a low porosity. This porosity is sufficient to considerably influence and delay the residence time distribution (RTD) of a solute pulse carried by the water. The solute molecules can diffuse in and out of the pores. Some of the solute may then diffuse back into the seeping stream, while the remainder will diffuse further into the matrix. A similar effect is observed in fractures with variable

[52] In situ tracer tests often indicate that the main dispersion mechanism is velocity dispersion caused by mixing of flow from independent pathways/channels at the observation location but with very limited mixing between paths in the rock.

aperture or porous infillings where solute from the flowing channels can diffuse into the stagnant water volumes and be considerably delayed.

Many radionuclides form positively charged ions and complexes. These sorb (ion exchange and attachment by surface complexation) to the minerals on the fracture surfaces and to the minerals in the porous rock matrix accessed by diffusion. The solutes that enter the matrix pores from the seeping water reside there for some time before they re-enter the water. As a result, their migration is retarded compared to the mean residence time of the seeping water. The matrix close to fractures is often altered by chemical reactions, stress, and intrusion of reactive solutes. Multiple layers or regions with varying mineralogy and porosity exist (Figure 5-4). These different porosity and mineralogical heterogeneities add to the variabilities and uncertainties that must be considered in simulating water flow and solute transport.

The overall dispersion of a solute carried by flowing water in a fracture network is therefore the result of a number of different mechanisms and processes and their interaction with complex, highly heterogeneous geological structures.

Since solute transport is to a large extent influenced by groundwater chemistry, hydrogeology, rock properties such as porosity, effective diffusivity, minerals, and other factors, one must show that we have enough knowledge about these topics to be able to build confidence in the solute transport modelling. Especially, the local groundwater composition has an influence of the interactions between the solutes and the rock. In addition, and from a validation perspective, field data on groundwater geochemistry and flowrates could be used for comparisons of in predictive modelling exercises.

5.3.5 Summary

Key features of groundwater flow and solute transport in fractured rock include:

- Geologic structure of the fracture network.
- Limited sampling (essentially 1D[53]) of the fracture network at depths of interest.
- High levels of heterogeneity within fractures.
- Observations of potentially channelled transport and geochemical conditions.

These features result in significant conceptual complexity. The need for highly parameterised models to reproduce aspects of network complexity and spatial variability together with correlations and uncertainties regarding individual model parameters requires model calibration. In these circumstances, models are likely to be data-limited and cannot be "used routinely; they have to be used far more cautiously and thoughtfully" (Starfield and Cundall 1988).

5.4 Modelling processes and model-space reduction

Scientific modelling can have many different purposes (McBurney 2012, Edmonds et al. 2019) but our interest is in prediction. Beven (2009) distinguishes between two kinds of prediction:

a) Prediction as science to "show that we do after all understand our science and its complex interrelated phenomena".

b) To make predictions that will be useful for management and decision-making processes.

Here we are concerned with both the first (confidence building) and the second (predictive) kind of prediction, particularly in relation to the assessment of the safety of geological repositories for radioactive waste in fractured rock. Confidence in b) requires a).

[53] These limitations are most severe during early surface investigations when boreholes are likely to be limited in number and sub-vertical, once underground facilities have been developed, opportunities for multiple boreholes and cross-borehole testing can provide a much improved dataset.

5.4.1 The modelling process: Real World – Model World

Starfield and Jarre (2011) describe the modelling process and its relationship to reality using the concept of the Model World. The Model World encompasses both the conceptual and mathematical model and is an intermediary between the real world and the simulations provided by model implementation. The different processes involved in model development and use are briefly listed below. In Figure 5-6 we present a schematic of these processes which relates Starfield and Jarre's Model World concept to Parker's notion of model-adequacy (Parker 2020).

- **Model Design** includes the development of the conceptual model and mathematical model (see Section 5.4.3 for a discussion of model selection). Enemark et al. (2019) review conceptual model building and testing. Design is an abstraction process: what goes into the Model World is determined by the purpose of the model and the context of its use; similar arguments determine what can be excluded from the Model World.
- **Model construction** involves the translation of the mathematical model into either a computer or physical model capable of performing simulations. In principle, the accuracy and robustness of this translation can be verified by performing appropriate testing, although the availability of benchmark solutions for complex problems may limit this process, here denoted "model verification". (Berre et al. 2020).

Model simulations require data to derive input parameter values and to condition and calibrate the model, where:

- **Conditioning** is the process of refining model structure or the distribution of parameter values as data become available. Here, conditioning refers to the inclusion of site-specific data to reduce parameter uncertainty. This includes adjusting assumptions and parameters to match specific identified geological features;
- **Calibration** is the process of adjusting parameter and state variables to obtain a better fit to observations.

Once a model has been constructed, it can be used to test the impact of uncertain assumptions and parameters. This can involve:

- **Assumption analysis:** key assumptions are tested by modifying the model to explore how each major assumption may affect the results.
- **Sensitivity analysis**[54]: analytical and numerical simulations are essential to help identify the key parameters of the model(s) and better target the key characteristics of the model(s) and determine where and how different model(s) and/or more in situ charcterization data and/or laboratory data would be required to reduce the overall modelling uncertainties.
- **Uncertainty quantification:** Uncertainty in model inputs together with uncertainty in relevant experimental conditions and any "solution error" (Christie et al. 2005) associated with the model implementation lead to uncertainties in predicted quantities. This uncertainty can be estimated using a range of techniques. Related techniques can also be used to identify the relative worth of different observations in a **Data-worth analysis** (Finsterle 2015).
- **Outcome analysis:** Simulation outputs usually require post-processing, e.g., consolidation and statistical analysis of multiple realisations (Aldrich et al. 2020), to derive relevant model outcomes. A key point is that the model outcomes relate to the model world and have to be interpreted back to the real world. This **interpretation** is dependent on the model purpose and the context of use. It needs to account for the model assumptions, identified sensitivities and uncertainties and the data that have been used to parameterise, condition and calibrate the model.

[54] Sensitivity analysis is the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input (Saltelli 2002).

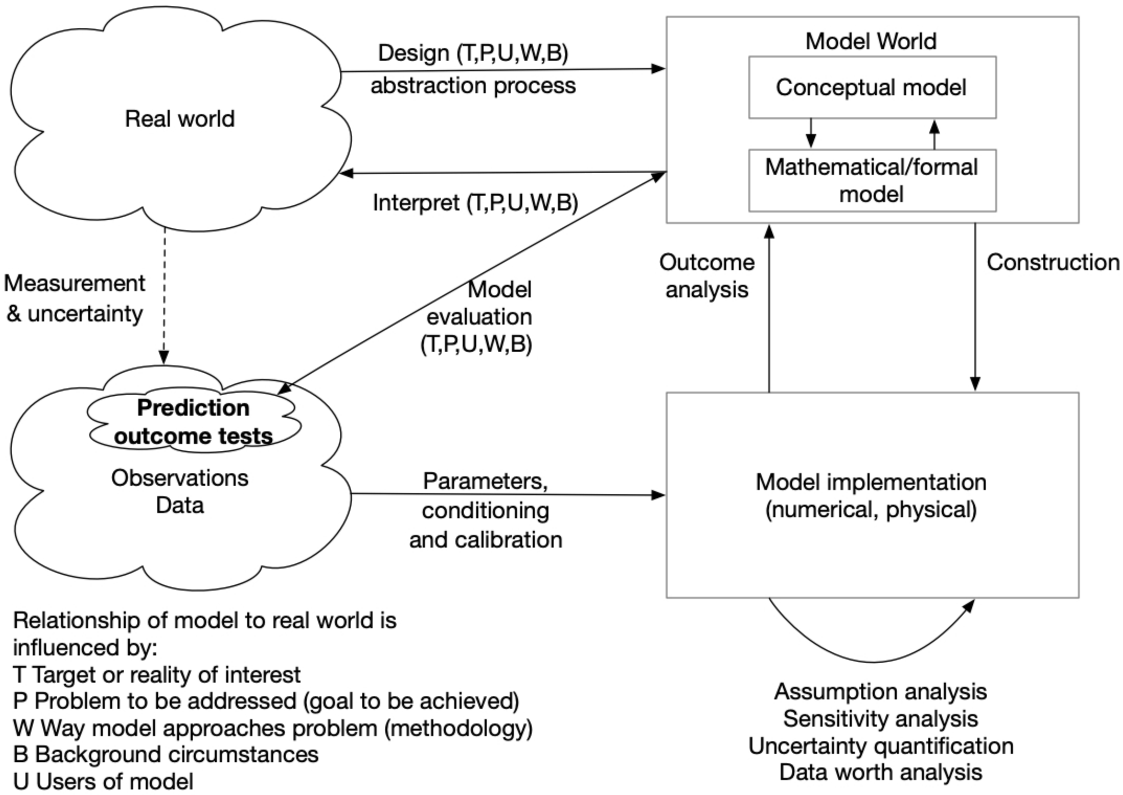


Figure 5-6. *The modelling process (adapted from Starfield and Jarre 2011).*

5.4.2 Model space reduction and extrapolation to prediction

The "Model Space" (Figure 5-3) will evolve during the process of model design, construction and use, as potentially relevant features, events and processes are included or excluded, and as parameter ranges are refined during conditioning and calibration. This is illustrated in Figure 5-7, where the width of the model space represents both the number of models considered within the ensemble and the parameter uncertainties associated with each model.

Models may be rejected (or revised[55]) at each stage of model evolution for various reasons, such as:

- Conditioning data stand in conflict with model assumptions.
- Calibration results in non-physical or otherwise unacceptable values.
- Sensitivity and assumption analyses show that important (relative to purpose) features, events or processes are not adequately addressed in the model.
- The model fails to meet performance criteria from prediction-outcome tests.

[55] Thacker et al. (2004) define **Model Revision** as the process of changing the basic assumptions, structure, parameter estimates, boundary values, or initial conditions of a model to improve agreement with experimental outcomes.

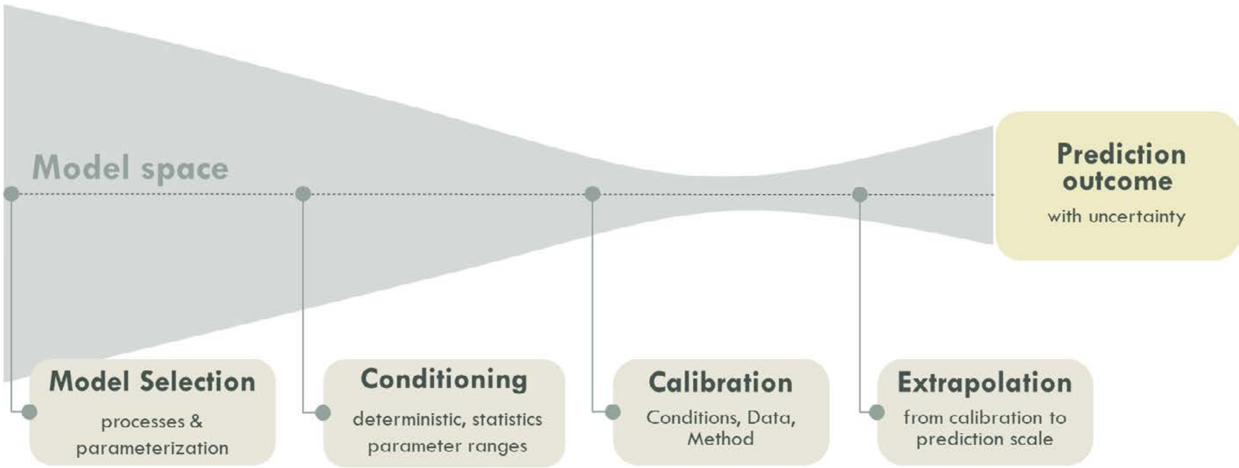


Figure 5-7. *Evolution of model space (from Selroos et al. 2022). The width of the model space (grey shading) represents the uncertainty in parameters and concepts. The thickness of the model space gives a qualitative indication of the range of possible models or model parameters.*

5.4.3 Model selection

Model selection involves:

- Identification of the physical and chemical processes that are of relevance for the purpose.
- Identification of suitable qualitative or quantitative sub-models that describe those processes and their interactions.
- Identification of the model parameters (including boundary conditions) together with suitable initial parameter ranges.

Conceptual uncertainty (or extreme parameter uncertainty) may lead to the selection of multiple models making different assumptions or simplifications about the target (Pruess et al. 1999, Selroos et al. 2002, Enemark et al. 2019). A key issue is the level of model complexity, which is a trade-off between the complexity of the real system, the purpose of the model and the available data. Overfitting needs to be avoided as it results in less reliable model predictions (Hawkins 2004). Starfield and Cundall (1988) make further arguments for the use of simple models. Model selection relies on understanding of the relevant physics, but also established geoscientific principles and commonly used constitutive models. Such models often include empirical relationships whose applicability are hard to confirm and thus require testing by including relevant alternative model assumptions. In practise, model selection may also be influenced by aspects relating to the users (e.g., software availability or the modeller's familiarity with specific models or techniques).

5.4.4 Conditioning

Conditioning is the process of constraining models with information about the parameter ranges. When this comprises of generic (i.e., not site-specific) information about the physical processes, it should be considered as part of the model selection process. However, when this conditioning relies on local observations made at the site, e.g., a structural model with the main geological domains and their boundaries defined by site geologists, we consider this a distinct step in model space reduction (see Figure 5-7 and Figure 5-8). Conditioning is often complementary to calibration, which also relies on site data, and may be used to implement the knowledge of geologists and hydrogeologists as a series of constraints (soft conditioning) with the potential to significantly reduce the model space.

5.4.5 Calibration

Calibration results in a reduction of the model space as model outputs are compared to data (Tarantola 2004). As a result of this process, we are able to estimate the best-fitting set of model parameters as well as determine an occurrence probability, as in the Bayesian formalism (Bayes 1763, Gelman et al. 2013, Stuart 2010). Given uncertainties, poor or inadequate data quality and the complexity of the models, a number of different parameter combinations can render quite different, but equally "good" fits to the information to which the model is calibrated (Beven 2006). A critical point, therefore, is that the conditions of calibration are not necessarily adequate to explore all the model parameters. This suggests that, in practice, some parameters will remain poorly determined after the calibration process. Therefore, a clear definition of the conditions of calibration (labelled 'calibration space' in Figure 5-8) – e.g., the dataset with uncertainties and the physical conditions of calibration (volume and time scale explored, boundary conditions) – is important information for any end-user of the calibrated model.

Since models are likely to be stochastic representations of real fracture systems, suitable calibration metrics are not necessarily 'local', but may be global measures averaged over defined volumes, a statistical distribution (of, e.g., fracture or channel sizes) or a scaling relationship. Example calibration metrics are listed below:

- 'Local' metrics are defined by a quantity (e.g., a pressure or flow), the measurement method and uncertainty. Such metrics are likely only to be relevant for models which have been heavily preconditioned on the data close to the measurement point.
- 'Density' metrics are the average of a quantity over a certain volume – e.g., fracture trace intensity, P_{21}, measured in tunnels or fracture frequency, P_{10}, measured in boreholes or scanlines.
- Statistical distributions and correlations – e.g., trace size distribution, flow distribution or breakthrough curves – are relevant metrics to calibrate the spatial variability of parameters (measured/modelled on tunnel walls, and along and in between boreholes).
- Scaling relationships characterise the structure (correlations and spatial arrangement) across scales of the target quantity.

Model rejections are an important step in the process of model space reduction since they help encourage the selection of relevant alternatives. An example could include calibration to hydraulic cross-hole tests which might require significant hydraulic connectivity over inter-well distances and hence be inconsistent with some representations of permeability heterogeneity.

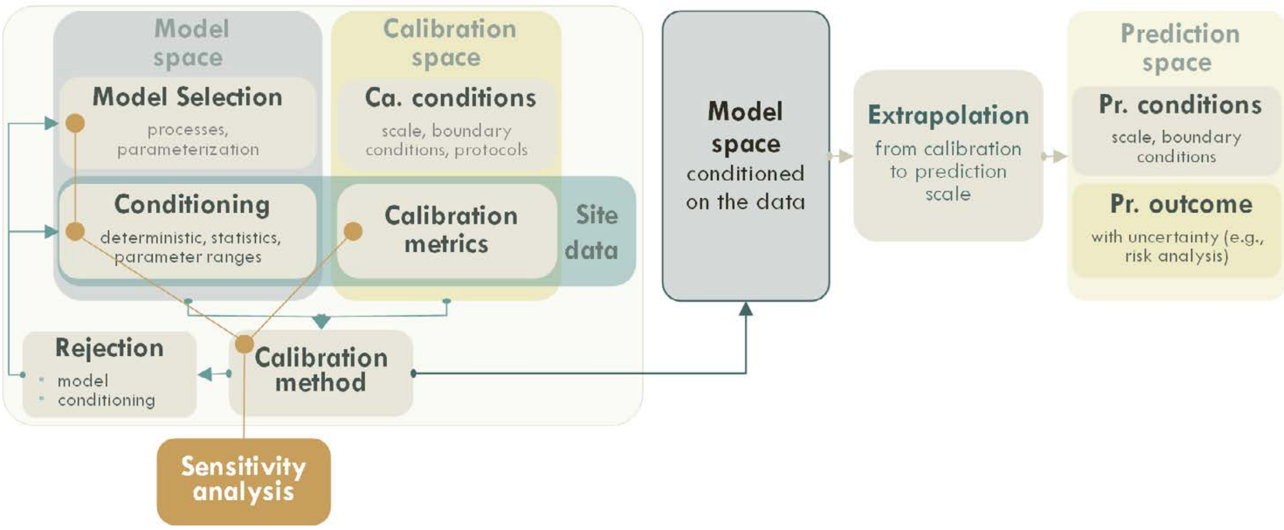


Figure 5-8. *Model Space at calibration and prediction stages (Selroos et al. 2022).*

5.4.6 Extrapolation to prediction

Two sorts of predictions are required:

- Validation predictions in support of prediction-outcome tests, for the purpose of model confirmation and demonstration of site understanding.
- Target predictions in support of end use: e.g., safety assessment and repository engineering.

Validation predictions should ideally be made for conditions as close as possible to those of the target prediction and different from those relevant to calibration. However, the conditions must still permit the required characterisation activities, e.g., underground mapping and hydraulic tests, or – at larger scale – hydraulic disturbances in monitoring holes across a site.

Target predictions for repository engineering will probably include extrapolation to new construction areas which are as yet uncharacterised, while predictions for long-term safety will include time scales and boundary conditions not considered during calibration. For long-term predictions the influence of processes that can change feature hydraulic properties and network structure need to be addressed.

Since the prediction conditions (labelled prediction space in Figure 2-8) may be very different from those at conditioning and calibration, the projection of the models conditioned by the data is not straightforward and may require additional assumptions in terms of parameters and processes. This results in an increase in the model space in terms of the range of models/parameters and assumptions (Figure 5-7). It is this extrapolation beyond the calibration and (possibly) model-testing conditions that poses a challenge for the validation of models, as discussed in the next chapter.

5.5 The validation issue and pragmatic model evaluation

Predictive use of any models requires a sufficient understanding of the quality and robustness of the simulation results, with direct implications for how these results need to be interpreted and the degree to which they can be used in support of important policy decisions.

The debate over the fundamental verifiability or falsifiability of scientific theories (Popper 1934, Kuhn 1977) is also relevant for the evaluation of numerical models and their application to radioactive waste isolation, as highlighted in Oreskes et al. (1994). These and other authors reach the conclusion that "*verification and validation of numerical models of natural systems is impossible*", because all such systems are open, with distributed input parameters that are incompletely known or conceptually inconsistent with their model counterparts. While attaining empirical consistency between the model-outputs and measurements may increase the confidence in the model, it does not confirm that a particular model captures the aspects of the natural world it seeks to represent. However, representational accuracy may not be the sole or even main criterion that renders a model adequate for its intended purpose (Parker 2020). This insight leads to a more pragmatic view of how the quality of a model can be evaluated. In this section we will lay the conceptual groundwork for our proposed account of pragmatic validation. In Section 5.6 we will set out a framework of activities and processes that are involved in pragmatic model evaluation.

5.5.1 From validation to pragmatic evaluation

The need for model evaluation arises whenever an extrapolation from one set of conditions (i.e., model space) to another is attempted. This pertains expressly to the step when proceeding from model calibration to model prediction, as at this point we leave the space where model development makes use of measured data of the target system, be these deterministic or statistical conditioning data, prior information about parameters, site-characterization data, testing and monitoring data, or calibration data (see Figure 5-9). When attempting a prediction, many elements of the model (such as structure, key processes, spatial and temporal scales and meaning of effective parameters) may have to be changed in order to adapt the calibrated model to the problem of interest. As a result, it is necessary to evaluate the appropriateness of the prediction model for its intended use. In Figure 5-9 "validation points" correspond to the conditions relevant to predictions made for model testing and validation, These conditions should be chosen to test the most important model assumptions within constraints of practicality. "Prediction points" correspond to the conditions relevant to predictions made to fulfil the model purpose.

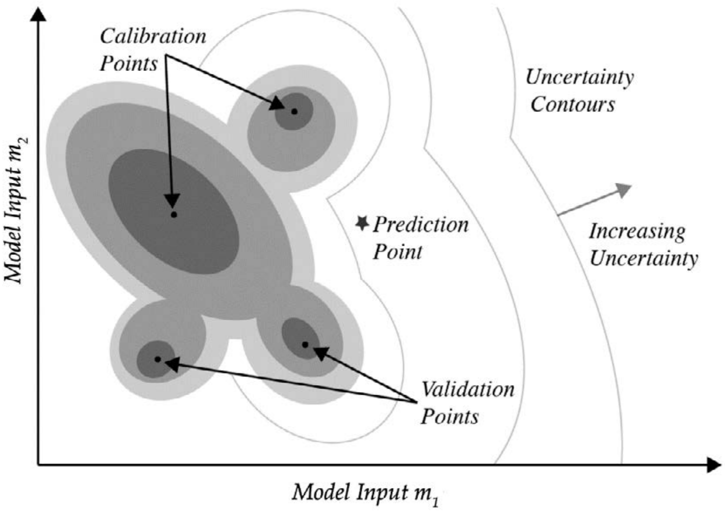


Figure 5-9. *Schematic of the calibration, validation, and prediction/application domains for two model inputs. Shaded contours show uncertainty during model calibration and validation with darker shading indicating lower uncertainty. Line contours show uncertainty in prediction. Based on Thacker et al. (2004).*

The purpose of pragmatic model evaluation is to demonstrate whether a model is fit-for-purpose (Parker 2020). Crucially, pragmatic model evaluation is contextual: it is not anticipated that the model is an accurate representation of the real system in any absolute sense or that it can make accurate statements or predictions of any type and under any condition. Instead, the model is assessed only within a limited domain of applicability, with the expectation that it provides information useful for the solution of the specific problem at hand.[56]

5.5.2 A framework for pragmatic model-evaluation

Such pragmatic model evaluation involves a critical review of the model-development process with the goal of demonstrating that the acceptable region of uncertainty is adequately delineated by an ensemble of model outputs. The acceptable level of prediction uncertainty is determined by the purpose of the study. While confidence in a model may be increased by critically examining the process of model development (Saltelli et al. 2013, Parker 2020) – i.e., without relying on a comparison of model results to observations – a test of the model's ability to make reliable predictions is an essential component of most validation methodologies, whether they are proposed as part of a philosophical argument or for pragmatic evaluation of numerical models used for the licensing of a nuclear waste repository (IAEA 2018).

Evaluation activities and acceptance criteria will vary, as they are related to the demands imposed on the models by features of the engineering or scientific problem at hand. That said, it is self-evident that a good general system understanding is a pre-requisite for building models that are expected to provide reliable, quantitative predictions of previously unobserved system behaviour. Similarly, following standards and best practices (Hill et al. 2004, Crout et al. 2008), checking the development of individual model components and undergoing independent review will increase the credibility of a model, regardless of its ultimate use. Nevertheless, assessing adequacy-for-purpose of a model includes additional considerations.

[56] One implication of this is that where models are re-used for different purposes, their applicability must be re-evaluated.

Pragmatic model evaluation helps to identify and correct flaws in the model by identifying and properly testing hypotheses, assumptions and methods. It will also expose each proposed modelling approach to a test of its usefulness. It is apparent that the evaluation approach has to be adapted to fit the model, the question the model is expected to answer, and the overall goal of its use. Our proposed approach to pragmatic model evaluation can be conceptually divided into six distinct phases or steps:

1. **Definition of the model purpose:** As discussed above, the aim of pragmatic model-evaluation is to determine whether a model is adequate-for-purpose: does the model make a valuable contribution to the solution of the problem at hand? Therefore, clearly specifying the intended purpose of the model is a crucial aspect of model-evaluation, as the model-purpose helps determine the benchmark and standards for critical evaluation.

2. **Determination of critical aspects:** The next step of pragmatic model evaluation is to determine which aspects of the models will require particular attention and thus warrant targeted review and testing effort. These aspects are likely to be specific to the intended use and are those that have the greatest impact on critical model outcomes. Moreover, model evaluation should be focussed on the subset of aspects that are uncertain or where the modellers lack confidence in their correct or accurate representation in the model.

3. **Definition of performance measures and criteria:** To be able to assess whether a model is adequate-for-purpose requires the definition of suitable performance measures and acceptance criteria. They must either be directly calculatable by the model or indirectly inferable from the modelling results, and they must be relevant to the end use (model purpose and users). Information, observations or testing data used for model assessment should be as close to the performance measures as possible, in terms of influential factors, processes, and scale. The accuracy of both the model output and data must be sufficiently high that they are discriminative in the evaluation of the acceptance criteria.

4. **Sensitivity and uncertainty analysis of influential factors:** Selecting influential factors is an important step during model development, but even more so for model evaluation. Influential factors are model specific, although they may be common to several models. The difference between the influential factors identified during model development (specifically model calibration) and the factors identified as influential for the ultimate model use is an indication of the degree of extrapolation undertaken when using a model for a purpose that may not have been envisioned during model development, and for which no closely related calibration data were available.

5. **Prediction-outcome exercises:** An important aspect of pragmatic evaluation is the testing of model predictions. Whilst direct testing of the model-predictions against the reality of interest is often not possible, the critical aspects and significant influential factors should be the basis for design and evaluation of prediction-outcome tests. Uncertainties in the influential factors need to be propagated through the model to the performance measures so that meaningful statements about system behaviours can be made that account for relevant uncertainty.

6. **Model evaluation, documentation and model audit:** As all model predictions are extrapolations (spatially, temporally, parametric, and regarding the features and processes that need to be considered) and the testing data never fully correspond to the ultimate performance metrics, confidence in the model cannot solely rely on the comparison between model output and measurements. Instead, each model development step must be clearly documented. In particular, the conceptual models and their assumptions need to be reviewed as they often have the greatest potential to bias modelling results (Bredehoeft 2005). It is also important to document and review the criteria used to reject a model or the criteria employed when calling for an update of the model. Any consensus and in particular any disagreement among model reviewers should be acknowledged.

5.6 Pragmatic validation

As discussed in the previous section an evaluation of model adequacy or fitness for purpose (Parker 2020) requires:

- A well-defined model purpose.
- Determination of aspects critical to the model purpose.
- Definition of performance measures and criteria relevant to the model purpose.
- Sensitivity analyses to determine influential factors.
- Comparison of predicted performance measures with measurements or observations (prediction-outcome tests).
- Documentation and audit of the modelling processes.

An important part of the approach is the development of multiple models based on alternative conceptualisations. If these models yield consistent conclusions about the behaviour of interest, the outcome is less likely to depend on uncertain or unidentified factors. Rather, it seems more probable that the general system understanding, as well as the information provided by the site characterisation and laboratory data, are sufficient to constrain the predictions (Munafò and Smith 2018).

The pragmatic model evaluation approach together with the development of an ensemble of multiple models addressing conceptual uncertainties allows for a pragmatic validation of the overall modelling approach to the target prediction (i.e. the purpose). The interaction between model evaluation and the model prediction is illustrated in Figure 5-10. Only those predictions from models that are judged "fit for purpose" are combined in the final ensemble target prediction. Models may be rejected or revised at various points in the process.

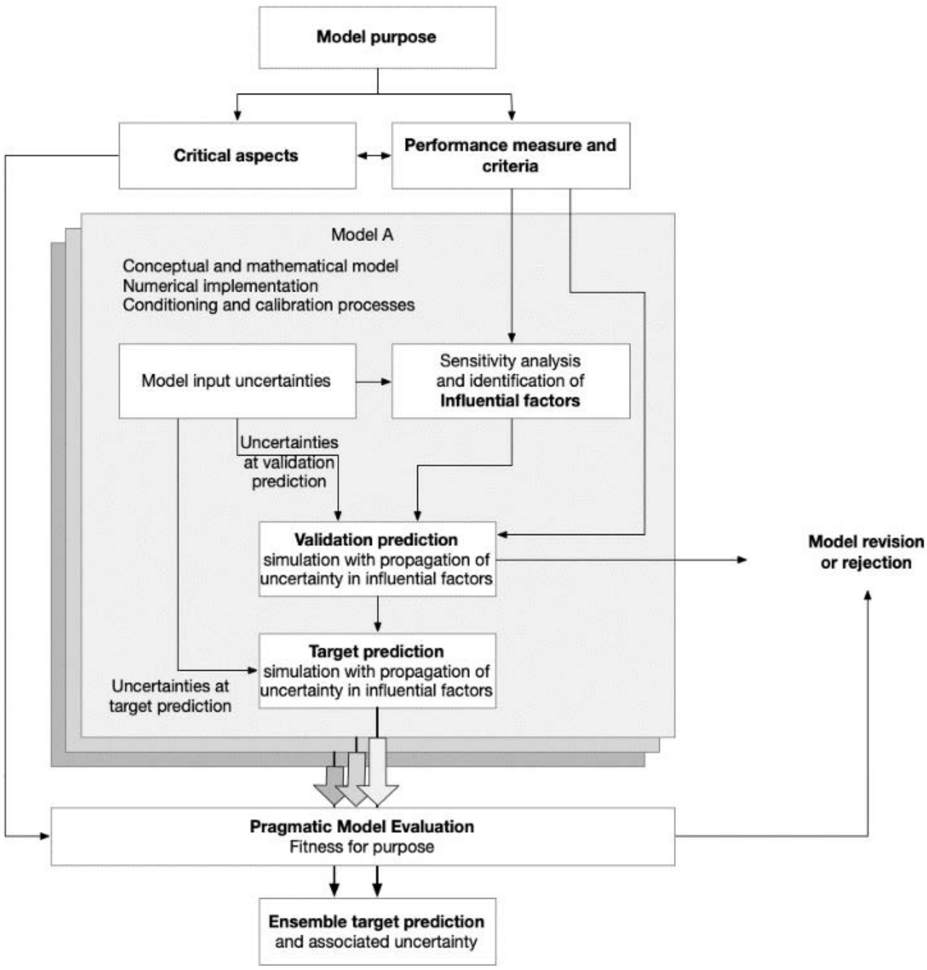


Figure 5-10. *Pragmatic model evaluation and validation.*

5.6.1 Definition of model purpose

A clear definition of model purpose is vital for any evaluation of model adequacy. Model development and evaluation should be driven by the model purpose; the more explicit the purpose, the better (Crout et al. 2008). Within our framework, the model purpose (for example, prediction of water inflow to an open deposition hole) is common to the ensemble of models, although individual models may approach the problem in different ways. Specification of the modelling purpose should address:

- Role of modelling in support of meeting regulatory requirements.
- Goals to be supported by modelling.
- Technical objectives of specific models.

5.6.2 Determination of critical aspects

Critical aspects of the model should be discussed as part of the documentation of the conceptual model. These are aspects that have the greatest impact on the model outcomes which define the performance metrics used in support of the overall study goal. Validation efforts will be concentrated on aspects that are not only critical but also uncertain, or where modellers lack confidence in their correct or accurate representation. In fractured rock, these may relate to:

- Assumptions regarding the role of different processes.
- Assumptions regarding parameter ranges and boundary conditions.
- Attribution of hydraulic or transport processes to geological structures (e.g., conductive, barrier or composite fault structures; Caine et al. 1996).
- Scale, resolution and non-uniqueness of observations and determination of connectivity (Berkowitz 2002).
- Assumptions regarding the definition of statistical populations and spatial stationarity of fracture properties.

Critical aspects may differ between models within the ensemble but are likely to share a few common aspects (e.g., reliability of key input data). Critical aspects shared by many models should be identified as key considerations in subsequent sensitivity analyses and prediction-outcome testing. Determination of critical aspects is likely to be based on scoping calculations, exploratory or insight modelling together with relevant knowledge and experience.

5.6.3 Definition of performance measures and criteria

Performance measures are those quantities that best inform a decision maker about the target system for the given purpose. The ideal performance measure would be observations of the system under the conditions for which we wish to make predictions. However, these have typically not been observed or are fundamentally not observable (thus requiring the use of a predictive model).

Pragmatic validation requires prediction-outcome testing. To make this testing relevant to the model purpose, the observable performance measures used in prediction-outcome exercises should be as closely related to the target system performance measures, with acceptance criteria determined by the accuracy with which these measures need to be known to be useful for decision making. Such measures need to:

- Be directly related to the purpose of the model.
- Be well-defined in terms of both calculation (the model world) and observation (the real world).
- Take account of errors in observations.

The accuracy with which these performance measures need to be calculated are set by the project objectives, by decision makers, or by regulation. Accuracy may be defined in terms of a quantitative uncertainty measure or a linguistic expression. For example, accuracy of performance measures may be assessed in terms of qualitative behaviour, an order of magnitude estimate, sensitivity to changes in key factors, similarity of prediction accuracy to data accuracy, high precision, etc.

Only a model that meets the performance measure criteria can be declared “fit for purpose”.

5.6.4 Sensitivity analysis of potentially influential factors

Influential factors are likely to include:

- Parameters or parameter groups that are likely to dominate model behaviour if changed by an amount corresponding to the parameter's uncertainty.
- Assumptions within the conceptual model and mathematical models.

The identification of influential factors is an important step of any model development process; it is mainly addressed by sensitivity and scenario analyses. It is also part of the validation process, where factors influencing the performance metrics should strongly influence model audit and design of prediction-outcome tests. Any difference between (1) influential factors identified during model development or model calibration and (2) influential factors identified for model validation is an indication of the inherent differences between the two model spaces. It also denotes the degree of extrapolation undertaken when using a model for a purpose for which no closely related calibration data were available.

A sensitivity analysis procedure consists of two basic components:

i. A strategy to vary the model parameters, inputs, or states.

ii. The definition of a (numerical) measure to estimate how the model response has changed based on varying one or more parameters, inputs, or states (Wagener and Kollat 2007).

Here, the measures have already been identified in the previous step, and the strategy is driven by the critical aspects.

Uncertainty analyses involve two components: uncertainty in input factors and uncertainty in model predictions.

Uncertainty in model input: The uncertainty in an input reflects our confidence in its value (if it is a quantifiable model input parameter) or the probability or likelihood we assign to the existence and relevance of a feature or process. The uncertainty of a parameter can be determined by analysing site characterisation data (by inverse modelling or sensitivity analyses); the uncertainty in assumptions may be evaluated through scenario analyses or expert elicitation.

Uncertainty in model prediction: The uncertainty in model predictions can be calculated using linear uncertainty propagation analysis and Monte Carlo simulations as well as sensitivity and scenario analyses. This uncertainty is needed to determine whether the model output is consistent with or significantly different from corresponding validation data from prediction outcome tests (uncertainties in the data also need to be accounted for in this assessment). Furthermore, these analyses determine the attainable prediction uncertainty (given the current knowledge about the model and its input parameters), which must be smaller than the acceptable uncertainty for the model to meet its objectives or to be of use to decision makers. If the attainable prediction uncertainty is greater than the acceptable prediction uncertainty, a first step would be to review model assumptions to determine whether they can be strengthened, and if not, additional site characterisation data need to be collected. These data must contain sufficient information about the factors that are both influential and uncertain, so the uncertainty of these input factors can be sufficiently reduced which consequently also reduces the uncertainty in the target predictions (Finsterle 2015).

5.6.5 Prediction-outcome exercises

Predictions are at the very heart of the practice of civil engineering (Lambe 1973):

"*The successful engineer must identify predictions which are critical to the safety, function, and economics of the project at hand, estimate the reliability of each of his predictions, employ predictions in design and construction, assess the consequences of predictions, especially erroneous predictions, and he must select and execute appropriate actions based on comparisons of the actual situations as they unfold and his predictions*."

Lambe (1973) distinguishes between different types of engineering prediction according to the timing of prediction and knowledge of the outcome (Table 5-1) and cautions that "*one must be suspicious when an author uses type C1 predictions to 'prove' that any prediction technique is correct*."

Direct testing of the model predictions under conditions similar to those at the target is often not possible due to differences in spatial or temporal scales between what is possible to test and the target system. However, some form of prediction-outcome testing is required by our definition of validation and provides a key test of the model ensemble.

Table 5-1. Classification of prediction (Lambe 1973).

Prediction type	When prediction made	Results at time of prediction
A	Before event	-
B	During event	Not known
B1	During event	Known
C	After event	Not known
C1	After event	Known

True tests of a model's capability to make predictions about reality must be evaluated against independent data (Refsgaard and Henriksen 2004). In designing suitable model validation tests, a guiding principle should be that the model is tested to show how well it can perform the kind of task for which it is specifically intended (Klemes 1986). Ideally such tests should involve "blind predictions" (Type A or B in Lambe's terminology) where the modellers are unaware of the measurements they want to predict. Where data are known (Type C) but have not been used in conditioning and calibration, knowledge of the outcomes may influence other modelling choices. For example, if a particular quantity is known to be low, modellers might choose (consciously or unconsciously) to include processes that reduce the equivalent quantity in the simulation and exclude processes that tend to increase the predicted quantity.

Any prediction must include an assessment of uncertainty – in part this accounts for uncertainty in measurements and information used to develop the model – and this assessment will help determine limits on the degree of accuracy we can attain within our final predictions. Therefore, within our suggested approach:

- Prediction-outcome tests should be designed to target the influential factors that have been identified during the process.
- Predictions must be associated with uncertainty estimates.
- Prediction-outcome tests should ideally be of type A or B (or as close as is possible to these).
- Where predictions require knowledge of the operational conditions (e.g., test duration or environmental conditions), a "blind prediction" should be considered with an option to update the predictions when the operational conditions have been documented. This should minimise changes in the model once the outcomes potentially become known.

Where there is an abundance or at least sufficient experimental data, it can be useful to set aside a portion of the data to be used later in prediction-outcome exercise (sometimes known as cross-validation). However, there are risks associated with "reserving" data in that if it is statistically similar to the data used to condition or calibrate the models it may not represent a real test of the model under "extrapolated" conditions, if on the other hand there are significant differences between the two datasets, the reserved data may be more useful within the model development process.

5.6.6 Model evaluation

Many aspects of model development and usage cannot be tested directly but still require scrutiny. All model predictions are extrapolations (spatial, temporal, parametric and regarding the features and processes that need to be considered). Prediction-outcome tests cannot fully correspond to the target performance measures, and thus confidence in the model cannot solely rely on the comparison between model outputs and prediction-outcome results. Furthermore, the extended evaluation outlined in the previous section addresses the possibility that a model might make correct predictions for reasons other than its ability to adequately represent the processes that are significant with regard to the model's purpose (Nordstrom et al. 2012).

Good modelling practice requires being transparent in documenting the model purpose, assumptions, formulations, input data and parameters (Crout et al. 2008). Model assumptions should be clearly documented and justified. Once stated with sufficient detail, model assumptions can be evaluated. Audit of the individual models should be performed by relevant experts. The approach we advocate is the "Sensitivity Auditing" of Saltelli et al. (2013), which is based on seven rules:

1. "Use models to clarify, not to obscure"
2. "Adopt an 'assumption hunting' attitude"
3. "Detect pseudoscience"
4. "Find sensitive assumptions before they find you"
5. "Aim for transparency"
6. "Do the right sums"
7. "Focus the analysis"

These seven rules are intended as review topics for sensitivity auditing and as minimum due-diligence requirements for the use of model-based inferences. The rules do not aim to reduce model uncertainty, but rather to make uncertainties visible and transparent so that both modelling practitioners and recipients of modelling analyses are fully aware of the conditionality of the predictions. Sensitivity auditing as described in Saltelli et al. (2013) critically examines the quality of a model used to make policy decisions – the process thus provides some guiding principles for pragmatic model validation.

One output from the model evaluation process will be the subset of the model target predictions that have been accepted as fit-for-purpose. This subset forms the basis of the ensemble target prediction (Figure 5-10). It is likely that some models may be rejected during the evaluation process or that revisions to the models may have been required (e.g., consideration of additional uncertainty). It is therefore possible that several iterations of the audit will be required.

In general, the notion that model validation is an auditing process guided by critical questions redirects the attention from a stringent pass-fail comparison of model-calculated and measured data to a broader evaluation of a model's adequacy through the judicious use of expert judgment as well as formal sensitivity and uncertainty analyses. Moreover, the model development and evaluation process need to be thoroughly documented and externally reviewed.

In this view, confidence is obtained by the fact that the validation process helps identify and correct obvious flaws in the model (Oreskes 1998), that hypotheses and assumptions are properly being tested (Luis and McLaughlin 1992), and that scientifically appropriate methods are being used (Neuman 1992). A rigorous validation process will ultimately improve the model and the quality of inferences and decisions made based on the model output. However, it fundamentally cannot confirm the validity of the model and its relation to the real world.

5.6.7 Conclusion

In this work we have set out a pragmatic validation approach for models of fractured rock mass flow, solute transport based on fitness for purpose (Parker 2020). The approach attempts to address the data-limited nature of fracture flow and transport problems and associated conceptual and parametric uncertainties via the use of:

- Multiple models.
- Sensitivity analysis and uncertainty quantification.
- Prediction-outcome tests.
- Model audit based on "sensitivity auditing" (Saltelli et al. 2013).

The approach requires that key performance measures and criteria are defined and that the critical aspects of the models having greatest impact on these performance measures are identified and used to focus both model development and model evaluation. We do not expect that any individual model will necessarily be "validated" for the given purpose; rather we aim to produce an ensemble of predictions from models that:

- have been the subject of critical expert review,
- address both the conceptual and parametric uncertainties revealed by the audit process, and
- have individually met the performance criteria associated with prediction-outcome tests.

The GWFTS Task Force will develop and test this approach and has already begun to consider how this can be achieved in a practical sense within the framework of models of flow and transport across a range of scales.

6 Applications to Task 10 for pragmatic validation

Masahiro Uchida, Fracture Flow Solutions

Bill Dershowitz, GeoFractal LLC

6.1 Pragmatic validation workflow and intended purpose

The IAEA definition of validation is specifically focused on documented analyses that provide confidence that a proposed approach produces results that are suitable for the intended application. The Task Force has decided to implement the IAEA definition within a concept which is referred to within the Task Force as "pragmatic validation" as described in Chapters 4 and 5. Pragmatic validation requires demonstration that the model is suitable for the intended purpose, i.e., the model not only provides a reasonable match to observations but is also suitable for the intended purpose. Therefore, the workflow used to obtain results must meet a set of requirements (*pragmatic validation criteria)* as described in Section 6.2.

As far as we are aware, Task 10 is one of the first efforts to demonstrate pragmatic validation for geosphere flow and transport processes in fractured rock. Task 10 is therefore somewhat aspirational – to build and demonstrate a reliable procedure for pragmatic validation – a "Pragmatic Validation Workflow". As explained below, a complete Pragmatic Validation Workflow for any significant task (e.g., a safety assessment) would require a large effort to address all the complex issues that arise in fractured rock over the timescales of interest. Recognizing the limitations of time, resources and data, and the different priorities and resources of the participating teams, Task 10 will demonstrate Pragmatic Validation Workflow for submodels with limited sets of purposes. It is hoped that the Pragmatic Validation Workflows developed in Task 10 contains sufficient elements to demonstrate a way forward to achieving IAEA Validation for geosphere flow and transport in fractured rock.

According to the IAEA definition, validation must be carried out within the context of a specific application, an "Intended Purpose". Examples of Intended Purposes related to long-term safety assessment (e.g. SKB 2011, Posiva 2012, Odén et al. 2014) that might be used in Task 10 include:

- Predicting inflow rate to an emplacement borehole.
- Predicting flow-related transport resistance along flow paths.
- Predicting advective travel time along flow paths.

To reduce the level of effort needed by each team, it is expected that the Task Force Secretariat will define the "Intended Purpose" for the models used within each of the Task 10 subtasks. However, the Teams can also take the opportunity to select an alternative or limited "Intended Purpose" more appropriate to their workflow.

6.2 Pragmatic validation criteria

6.2.1 Pragmatic validation criteria and validation acceptance criteria

As described in Section 4.4, the aim of pragmatic validation is to demonstrate the usefulness of the model solutions for the intended purpose. To achieve this we can formulate pragmatic validation criteria and the specific requirements.

The terminology of "pragmatic validation criteria" corresponds to previously discussed validation acceptance criteria. However, the former organizes requirements hierarchically from higher level to lower level, as well as includes the specific requirements. The specific requirements are formulated to evaluate conceptual models for flow and transport simulation within fractured crystalline rocks. Task 10 pragmatic validation will be built on the idea that pragmatic validation must meet "pragmatic validation criteria".

The level of effort required for pragmatic validation depends on the purpose of the model application and its significance for decision making. This is illustrated in Figure 6-1. In the low requirement case, relaxed criterion C may be applied; for example less rigorous QA/QC procedure can be used and/or the coverage of the end user of the model may be limited to the experts in the same discipline (not including policy makers and public).

In Task 10, each Modelling Team will define a workflow to demonstrate pragmatic validation based on their modelling efforts for the subtasks. These workflows will be evaluated against the pragmatic validation criteria for application of the approach according to its intended purpose.

Prediction/outcome ("P/O") exercises will be used in Task 10 as one of the approaches to demonstrate that a given modelling workflow satisfies the corresponding pragmatic validation criteria. Other methods include Quality Assurance/Quality Control (QA/QC), independent review and audit, and expert elicitation (see Chapter 4 and 5).

Consideration of pragmatic validation criteria should be provided in the validation workflow. These criteria can be organized hierarchically as follows:

Pragmatic validation criteria

1. The domain of model applicability must be determined and related to the intended purpose of the model:
    a. Inherent limitations of a model must be documented.
        i. A list of factors that could significantly influence the modelling results must be provided. In making a list, specific requirements in Table 6-1 should be evaluated.
    b. Uncertainties in model predictions must be documented.
        i. Site characterization data, which reduce prediction uncertainty, must be identified.
    c. The type of data, the processes involved, the sampling intervals in space and time, and the conditions prevailing during data collection should reflect those of the target predictions as closely as possible.
2. The Prediction uncertainty of performance metrics should conform to pre-established uncertainty criteria:
    a. Ensemble uncertainty needs to be evaluated using different assumptions/alternative models.
    b. Performance metrics should be defined such that they usefully support decision making.
    c. The criteria for acceptability are determined by the accuracy with which model outputs supporting the project objectives need to be calculated.
    d. A model should be compared to alternative models or approaches and perform equally well or superior regarding relevant validation performance criteria.
    e. A validated model should reproduce relevant data not used for model calibration with acceptable accuracy. (P/O exercise).
3. Information on model limitations and uncertainty as well as the domain of model applicability must be effectively communicated to the end-user of the model.
    a. The model development and evaluation process need to be thoroughly documented and externally reviewed.
        i. A model should have undergone a detailed review of the procedures used for the construction of the conceptual and numerical models, including the evaluation of available data, review of theoretical and empirical laws and principles, abstraction process and conceptual model development, building of the calculational model, and the iterative refinement based on predictive simulations, sensitivity analyses, and uncertainty quantification.
        ii. As to conceptual model development, the specific requirements listed in Table 6-1 should be evaluated.
    b. A model should comply with industry-standard QA/QC procedures and have passed a formal software qualification lifecycle test ("verification").

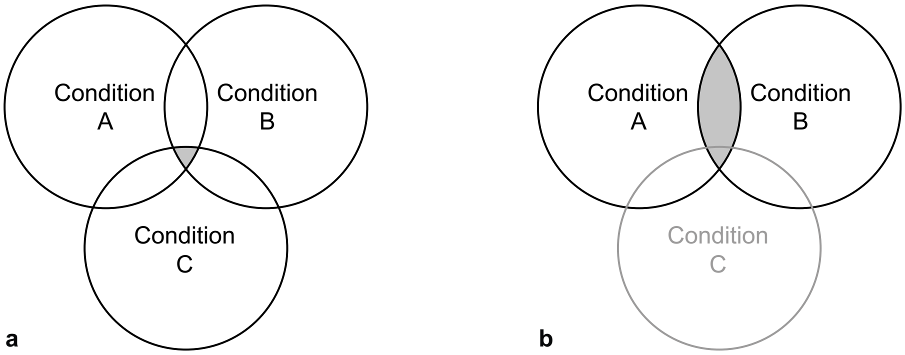


Figure 6-1. *Pragmatic validation criteria (for simplicity only three criteria are shown): a) high requirement case: This case requires that all three criteria A, B and C are met; b) low requirement case: criteria A and B need be fully met, but criteria C can be less rigorous or skipped. Each criterion is represented as a disk.*

Table 6-1. Example list of possible "Specific requirements" for conceptual model development ("Specific requirements" to be applied depends on the model purpose).

A comprehensive list of Features Events & Processes (FEPs) should be developed, screened, and performance-relevant FEPs should be properly included in the model	Proper selection of FEPs depends on the question being asked, the geologic conditions, processes and the time scale, and the specific issues being addressed by modelling. When interpreting measured data and making prediction, FEPs should be appropriately selected.
Proper conceptual model needs to be selected and documented	Conceptual models for flow and transport need to consider alternative formulations and their limitations. These might include: • Darcy and non-Darcy flow. • Laminar and turbulent flow. • Fracture-matrix interaction. • Flow channelling effects. • Streamline routing and mixing processes at fracture intersections. • Density driven flow. • Thermal convection flow.[57] • Hydrochemical evolution along pathways due to rock-water interaction during circulation of hydrothermal fluids, mixing of groundwater and other factors. • Hydromechanical effects due to stress changes and poroelastic processes.
Extrapolation of time scales should be properly considered, even where data at the time scale of the application is impossible to obtain	Extrapolation of time scales will be required to go from experimental time scales (hours to years) to the modelling/analysis time scales (tens to millions of years) for the "Intended Use" in repository safety assessment. Natural analogues using for example groundwater age can be used. Possibility of generation of new fractures and sealing/re-opening of existing fractures may need to be considered.
Site characterization should be designed to reduce uncertainty sufficiently for model purposes	Data Worth Analysis is one of the approaches to consider the value of site characterization activities. Pragmatic validation should avoid assumptions of unlimited datasets from repository construction activities and needs to assume those of the target predictions as closely as possible.
Boundary conditions should be consistently defined	• Control of boundary conditions can be particularly challenging for in situ experiments. Uncertainty and variability of experimental boundary conditions need to be considered in all stages of pragmatic validation. • For deposition hole inflow, both outer and inner boundary conditions are different between pre-closure and post-closure periods. For example, when experimental deposition hole inflows are used in support of safety assessment, the effect of the EDZ around the deposition holes, deposition hole backfill and canister material, as well as outer boundary condition after resaturation of the repository need to be considered and documented.

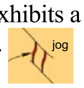
[57] This could become significant in the regional scale groundwater flow model whose depth can reach 10 km, as well as in the near-field model during the thermal period.

Multidisciplinary approach should be used (Consistency with hydrochemistry etc.)	• Geomechanical principles of fracture formation and propagation are particularly useful for extrapolating fracture data to larger scales beyond the region of characterization. • Hydrochemistry will evolve along the flow path due to rock-water interaction, mixing and other factors. Therefore, flow paths should be consistent with transient hydrochemistry processes, such as those due to long-term changes in climate, topography, surface development, etc. Note that THMC modelling capabilities need to be considered in defining the "pragmatic validation criteria" to be addressed in validation. • Groundwater age should be consistent with regional groundwater flow, taking into account mixing and rock matrix diffusion. • Electric resistivity distribution can be used to identify the distribution of saline water (example of integration of geophysical measurements).
Abstraction (simplification) workflows should be established and documented	Abstraction (simplification) workflows used to derive equivalent homogeneous fracture or equivalent porous medium flow and transport properties including matrix diffusion effects, considering the in situ heterogeneity and correlated spatial structure of properties (including aperture) within fractures and at fracture intersections need to be explicitly described and justified.
Fracture network geometry of conductive fractures including spatial variability and systematic trends need to be justified	**Key concepts to be considered include:** • Distinguishing fractures by the model purpose：Distinguishing between geologic fractures, geomechanically significant fractures, fractures capable of supporting advective or convective flow, and fractures providing storage for diffusive processes. **Fracture network geometry** • Fracture shape (e.g., polygonal, elliptical, and non-planar). • Alternative spatial models, including Markovian, fractal, geomechanically based, geostatistical, Poisson processes, and non-Poisson processes. • Spatial model to describe parameter heterogeneity including both stochastic and systematic (non-stationary, periodic, trend) variations. • Multi-fractal populations where scaling is inappropriate between fractures formed by different mechanisms: possible that higher probability of larger fractures to be conductive, since larger fractures are more likely to form pathways. • Fracture intersection and termination processes (e.g., T- and X- intersections). • Sampling bias, censoring, and truncation effects on fracture geometric measurements need to be compensated for. **Connectivity and Correlations** • Correlations between fracture sets, geometry, and properties (e.g., size-transmissivity). • Where "conservative" predictions are required (e.g. in safety assessment) pathways related to the lowest solute retention (including the lowest flow resistance and the lowest flow wetted surface), and also for the shortest transport times should be considered. • Correlations and conditioning between fracture geometric, hydrogeologic, and geomechanical properties need to be defined and justified explicitly. • Overconnectivity: Field experiments have indicated that DFN models tend to be overconnected compared to in situ conditions, even at the scale of individual fractures. This can be seen as a channelling effect, or a connectivity/compartmentalization issue. For repository safety analysis, the potential occurrence of *sparsely connected persistent (long) channels* needs to be evaluated.

Spatial models for channelling within fractures[58] and at fracture intersections need to be described and justified	**Mechanism of channelling** Selection of spatial model for aperture distribution requires consideration on the genesis[59] of channels. The spatial model for channelling also needs to consider the spatial pattern of flowing and flow barrier porosities, and porosities that are accessible to mixing and diffusive properties. Examples of the latter porosities include gouge, breccia and altered rim zone. **Channelling percentage** Alternative channelling concepts need to be clearly defined in experimental design, data interpretation, and pragmatic validation. The channelling percentage, for example, can be interpreted as (a) percentage of fracture area with an aperture above a defined threshold, (b) percentage of fractures active in flow, and (c) combination of (a) and (b). Which definition is used at each stage (data interpretation, forward modelling) needs to be clearly documented. **Correlation** The effect of correlations (e.g., between transport aperture and mechanical aperture, and between geological aperture and fracture size) on pathway formation needs to be considered and documented. **Fracture intersection** Many field observations indicate the possibility of fracture intersections being more or less conductive than other portions of fractures, such that they have a significant effect on connectivity and channelling. Wennberg et al. (2015) indicated that the dilational jog[60] portion of fracture intersections tends to be more conductive. Tetsu and Sawada (2003) quantitatively demonstrated the increased transmissivity at Fracture Intersection Zones (FIZ), as described in Appendix B.1.
Mechanical, hydrogeologic, and transport property assumptions and their derivation of fracture intersection need to be described and justified	**Document and justify assumptions regarding:** • Preferential flow channels along fracture intersections. • Flow barriers at fracture intersections. • Mixing at fracture intersection (complete mixing vs streamline routing).
Major conductive structure should be characterized and modelled.	Major conductive structures can act as both preferential pathways in the direction parallel to structure due to increased transmissivity caused by increased fracture densities and increased fracture apertures, and also as flow barriers in the direction normal to the structure due to the presence of an impermeable fault core. Major conductive structures are often comprised of numerous small fractures which sometimes produce "fault breccia" and may need to be simplified rather than being modelled as discrete fractures.

[58] It may be desirable to study the simplification from the viewpoint of simulation feasibility, since in a larger scale model, a detailed model describing aperture distribution is computationally more expensive. However such issues of model implementation should not limit the development of the conceptual model.

[59] Channeling occurs as some combinations of "Hard", "Soft", and "Structure" channeling. "Hard" channeling is defined by channel geometries which include aperture distributions and the spatial pattern of fracture intersections, precipitation, and infilling. "Soft" channeling is defined by channel geometries which include channel geometries which vary depending on boundary conditions so that channeling can change as the flow field changes. "Structure" channeling occurs where an individual structure such as a "fracture zone" is composed of network of joints and micro faults, and flow preferentially occurs in some subset of those underlying discontinuities. In "Structure channeling", channel pathways are defined by multiple pathways with a range of geometries supporting flow and a range of velocities and immobile zone interaction rates. "Shear" and "Tension" fracture formation mechanisms also influence fracture flow and transport and channeling behaviour. These could affect the spatial distribution such that "Shear" fractures, for example, may have anisotropic transmissivity related to the shear displacement direction, and may serve as flow barriers due to fracture infillings or reduced transmissivity due to aperture closure when protruding portions of fracture surfaces ride over each other during displacement.

[60] A structure commonly observed along the strike-slip fault. The structure develops where the fault exhibits a step-over structure. In large scale, it forms a "pull apart basin". An example of dilational jog is shown here. 

Fracture mechanical, hydrogeologic, and transport property assumptions and their derivation are properly assigned and documented.	**Document and justify assumptions regarding:** • Correlation of parameters (e.g., fracture size to transmissivity). • Scaling relationships.[61] • Conditioning of material properties based on experimental and site characterization data. • Resistance of hydraulically active fractures along fracture network should be accurately reproduced. • Properties assigned based on theoretical and empirical understanding of underlying mechanisms. • Correlation between effective stress state and transmissivity may depend on the material properties of fracture fillings. • Local effective stress conditions at each fracture may be different from regional stress condition due to local stress redistribution by discontinuities and other fractures, which may have weak material properties.

6.2.2 Specific requirements versus scales

Fracture networks involve a range of heterogeneities specific to each scale. At the single fracture scale, channels and channel dimension, aperture distribution, fracture fillings, as well as a network of microfractures in relatively large fractures, and other characteristics are the main sources of heterogeneities. For a simple fracture network scale, fracture intersections become an additional heterogeneity. In larger block scales (a few hundred meters), fracture network organization and sometimes fracture zones become additional heterogeneities. It is proposed that Task 10 should take a step-wise approach, proceeding from smaller scale/less complex to larger, more complex scales – progressively increasing scale from single fracture scale, the simple fracture network scale (including fracture intersections) to hundreds-of-meters block scale as shown in Figure 6-2. Each successively larger scale will integrate the heterogeneities from smaller scales. Table 6-2 illustrates how each successive scale can integrate the specific requirements relating to conceptual model at smaller scales. As Task 10 proceeds, it is important that insights from each scale and each type of model be considered in defining and executing subsequent subtasks.

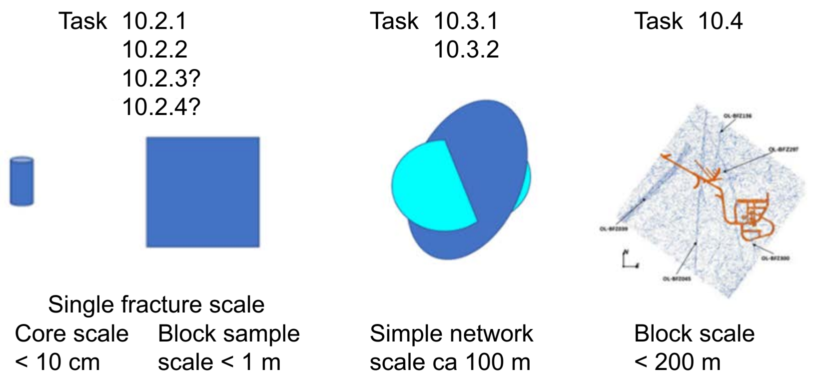


Figure 6-2. *Scale of Task 10 subtasks.*

[61] Fracture size-intensity relationship (power law) of conductive fractures may not be same as that of geologic fracture, since large fractures tend to be more conductive and have greater chance to form a network.

Table 6-2. Matrix of scales and the different issues and specific requirements relating to the conceptual model. Circles indicate in which scale specific requirements should be applied.

		Scale		
		Single fracture scale	Simple network scale	Block scale
Issues and Specific requirements	Proper selection of FEPs	○	○	○
	Selection of conceptual model	○	○	○
	Long term aspect	○	○	○
	Spatial models for channeling within fractures and at fracture intersections	○	○	○
	Network geometry of conductive fractures			○
	Fracture intersection		○	○
	Major conductive structures (geometry, material properties, barrier effect)			(○)
	Material properties	○	○	○
	Optimally designed site characterization/Data worth analysis	○	○	○
	Use of well-established boundary conditions	○	○	○
	Multidisciplinary approach	○	○	○
	Abstraction (simplification)	○	○	○

6.3 Structure of Task 10 subtasks and reporting requirements

6.3.1 Basic structure of Task 10

The basic structure of Task 10 is shown in Figure 6-3. For an actual repository, site characterization needs to be designed to minimize disturbance to the host rock. Therefore, the Task 10 demonstrations of pragmatic validation should use datasets comparable to what can be obtained at a repository site (e.g., "routine site characterization"). A model constructed from "routine site characterization" is referred to as a "SC model". In contrast, "reference models" built using the full scope of data available from off-site in-situ and laboratory experiment, are not subject to these restrictions. Reference models are useful in Task 10 to define a level of in-situ realism not possible with "SC models".

In each subtask of Task 10, models will be implemented:

1. Construct pragmatic models based on routine site characterization data.
2. Optionally, a simplified, "abstracted" model will also be implemented for each sub-task. For example, the true, complex geochemically and geomechanically defined mechanical fracture aperture pattern can be implemented as a uniform transport aperture which reproduces essential responses such as a peak concentration, peak arrival time etc. The purpose of this model is to provide an implementation more consistent with the level of detail in a typical repository PA.
3. A reference model will be constructed using all data obtained during in-situ experiment and/or laboratory experiment (where generally more data are obtained than the routine site characterization). The purpose of constructing a reference is to compare its uncertainty with that of pragmatic model. This effort can be replaced with the Data Worth Analysis by showing the worth of potential data in a routine site characterization as compared to the full data set.

For each of these models, experimental results will be predicted, and performance measures and their uncertainties will be calculated. It is very important for the Task definitions to define observable performance measures for addressing uncertainty and confidence. Predictions will be compared to the measured data as a "P/O" exercise. Results will be presented with qualitative and quantitative evaluation of bias, error, and uncertainty. In order to address uncertainty, propagation of uncertainty will be evaluated to study the impact on PA due to extrapolation (in time and spatial scale, and other conditions). This will also provide an opportunity to consider the domain of applicability of the model.

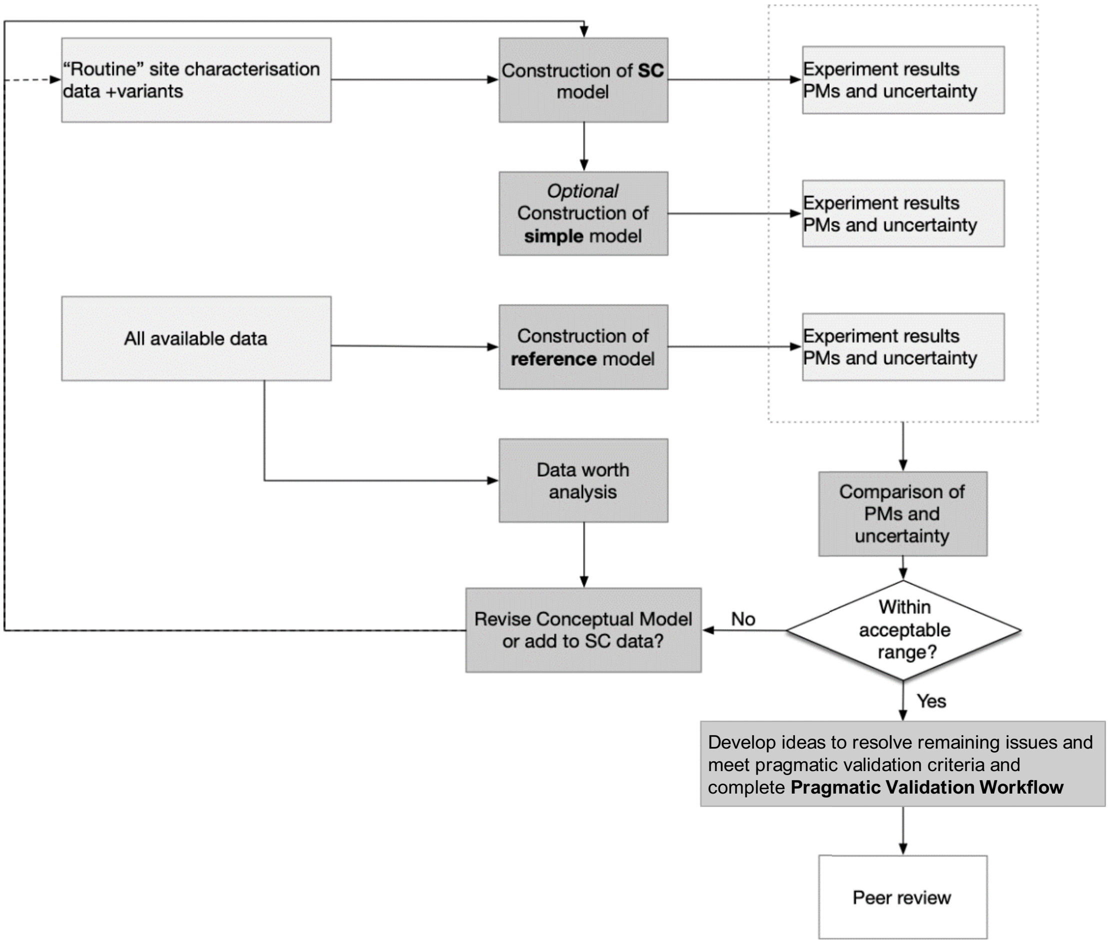


Figure 6-3. *Basic structure of Task 10 subtasks. PM refers to performance measures; SC refers to site characterization.*

Ensemble uncertainties will be evaluated from multiple realizations for aleatoric uncertainties and from multiple conceptual/numerical approaches for SC models, simple models, and reference models, respectively. Ensemble uncertainties will be compared to a pre-defined uncertainty criterion. If ensemble uncertainty of the SC model is greater than acceptable uncertainty but that of the reference model is lower than acceptable uncertainty, then additional site characterization data could be used within a revised SC model until the ensemble uncertainty has been sufficiently reduced. This iteration is intended to provide an approach to improve site characterization during repository development. Alternatively, a Data Worth Analysis can be performed using ensemble uncertainties as metrics.

Each team's Pragmatic Validation Workflow documentation should describe how the pragmatic validation criteria and the specific requirements for Pragmatic Validation are addressed, as well as the modelling procedure. This will include the model purpose, the domain of applicability of the model including limitations and prediction uncertainty, and descriptions of the datasets used, modelling procedures described below, and assumptions including boundary conditions and conceptual models. The dataset used to construct a SC model needs to be compared to what might be available from repository site characterization. If there are pragmatic validation criteria not met or addressed in the task, possible solutions meeting these pragmatic validation criteria should be proposed.

Each of the Task 10 Subtasks should follow the following modelling procedure for the pragmatic validation.

1) Clear definition of the model purpose.
2) Identification and evaluation of appropriate conceptual models.
3) The use of multiple models (conceptual and numerical) with clear methods for model selection, rejection, and update.
4) Identifications of influential factors.
5) An understanding of model sensitivity to parameter uncertainty and model assumptions.
6) Relevant prediction-outcome exercises comparing model predictions with observations of the real system.
7) Calculation of prediction uncertainty on the performance measures and comparison with the acceptable uncertainty.
8) Identification of the model inputs and required quality and scale of measurements (to confirm planned site characterization during repository construction).

Task 10 Deliverables

The Pragmatic Validation Workflow document should include the components detailed in Table 6-3. Thus, the deliverable from each Team and for each subtask will directly support the goals of developing and demonstrating pragmatic validation approaches for the sponsoring organizations. These Pragmatic Validation Workflows will be peer-reviewed from the perspective of supporting the development of pragmatic validation procedures that can be used directly by Radioactive Waste Management organisations participating in Task 10.

Table 6-3. Required components of Pragmatic Validation Workflow document.

Required content	Details
Statement of the Purpose for which the pragmatic validation is to be applied.	
Model description.	• Key model/approach assumptions. • (Generic) Scientific/engineering support for the assumptions and approach applied. • Known limitations of the approach for this application. • Key uncertain model parameters.
Definition of performance measures and criteria.	
Identification of influential factors.	• Specific requirements in Table 6-1 should be included.
Uncertainty in input parameters.	• Uncertainty quantification of influential parameters. • Uncertainty quantification on structural model (scenarios or parameter variation). • Relationship between uncertainty of the key parameters and site characterization (or Data Worth Analysis).
Probabilistic prediction for the experiment and performance measures (PMs), reflecting input uncertainties.	• Quantitative and qualitative evaluation of the prediction uncertainty and sensitivity analysis for influential parameters including material properties, boundary conditions, and geometry in the approach applied. • Determination of the domain of model applicability. • Consideration of subjective (rather than purely statistical) uncertainties, biases, and errors.
Pragmatic validation statements based on the results achieved.	
Recommendations for work needed to complete a pragmatic validation suitable to the purpose defined, if any of the pragmatic validation criteria/specific requirements relevant to intended model purpose are not satisfied.	

7 Final remarks

Safety assessments of radioactive waste repositories rely heavily on results obtained by numerical models that assess the long-term performance of the engineered and natural barrier systems. Given that important engineering and public policy decisions are based on these models, it is essential that we critically evaluate their abilities and limitations, and thus justify the level of confidence we have in the inferences drawn from the modelling. In this report, some of the issues that should be considered in the modeler's attempts to test, corroborate, confirm, and verify numerical models are described. This process is here referred to as model validation and a pragmatic approach is chosen for this important topic (Finsterle and Lanyon 2022).

This report provides background to the work within Task 10 of the Task Force GWFTS (Selroos and Gylling 2023) and documents the views of the White Paper Group which was set up at the start of the Task 10 – Validation approaches for groundwater flow and transport modelling with discrete features.

Chapter 2 provides a survey of the "state of the art" regarding flow and solute transport in fractured rock and identifies key processes and issues, together with relevant experience from the many years of the extensive Swedish Programme. While it does not provide specific guidance on validation, it outlines the "domain of expertise", i.e., the knowledge and understanding of the essential aspects of flow and transport in fractured rock that is our focus.

Chapter 3 discusses the different stages in the modelling process, framed in terms of model space (the envelope of possible models) concepts, and discusses them with examples from site-scale modelling as presented within the SKB DFN Handbook (Selroos et al. 2022). Reduction of the model space is achieved by a series of modelling activities, including:

- Identification of the significant processes and appropriate representations.
- Conditioning of the model on prior information.
- Calibration of the model on data.
- Testing/validation of the model.

The chapter discusses each of the stages and provides recommendations on good practice.

Chapter 4 provides some philosophical background to the debates around model validation in earth sciences and suggests a range of practical methods to address them. These modelling methods form the basis of the "recipe" set out in Chapter 5 and the more detailed suggested implementation within Task 10 that is discussed in Chapter 6.

The modelling approaches recommended in Chapters 3, 4 and 5 are not new but are similar to (or incorporate) existing "good practice" recommendations (Hill et al. 2004, Refsgaard and Henriksen 2004, Crout et al. 2008, Biondi et al. 2012, Nordstrom 2012, Saltelli et al. 2013).

Common aspects of the recommendations include:

- Clear specification of the model purpose.
- Consideration of alternative models.
- Well-defined steps in model development.
- Structured, transparent documentation of model development with particular emphasis on the assumptions and data on which the models rely.
- Testing of the models against relevant observations.
- Critical review by a community of experts and stakeholders.

Assessment of the reliability of the models' ability to fulfil their purposes depends on the community of experts who can provide the critical scrutiny needed. There is no "magic formula" that can be followed (Oreskes 2019) to ensure that our science and models are trustworthy. Instead, we rely on "science as a communal activity of experts who use diverse methods to gather empirical evidence, and critically vet claims deriving from it." The pragmatic validation approach that we wish to develop relies on providing support for this effective critical scrutiny of models.

It is likely that the ideas set out here will develop during Task 10 and that an update on the GWFTS Task Force's views on pragmatic validation and how it can be best accomplished will come from the application of the approach by the modelling groups within the individual tasks and the open supportive culture that exists within the Task Force.

Glossary of abbreviations

ADE	Advection dispersion equation
BTC	Breakthrough curve
ChN	Channel network
CNM	Channel network model
DFN	Discrete fracture network
ECPM	Equivalent continuous porous media
FWS	Flow wetted surface
LRC	Long-range channelling
PDF	Probability density function
RTD	Residence time distribution
TF GWFTS	Task force on modelling groundwater flow and transport of solutes
WCF	Water conducting feature

References

SKB's (Svensk Kärnbränslehantering AB) publications can be found at www.skb.com/publications.

Abelin H, Neretnieks I, Tunbrant S, Moreno L, 1985. Migration in a single fracture: Experimental results and evaluation. Final report. Stripa Project Technical Report 85-03, SKB/OECD, Royal Institute of Technology, Sweden.

Abelin H, Birgersson L, Gidlund J, Neretnieks I, 1991a. A Large Scale Flow and Tracer Experiment in Granite: I. Experimental Design and Flow Distribution. Water Resources Research 27, 3107–3117.

Abelin H, L Birgersson L, Moreno, H. Widén, T. Ågren, and I. Neretnieks, 1991b. A large scale flow and tracer experiment in granite: 2. Results and interpretation. Water Resources Research 27, 3119–3135.

Abelin H, Birgersson L, Widén H, Ågren T, Moreno L, Neretnieks I, 1994. Channeling experiments in crystalline fractured rocks. Journal of Contaminant Hydrology 15, 129–158.

Aldrich G, Lukasczyk J, Hyman J D, Srinivasan G, Viswanathan H, Garth C, Leitte H, Ahrens J P, Hamann B, 2020. A Query-Based Framework for Searching, Sorting, and Exploring Data Ensembles. Computing in Science & Engineering 22, 64–76.

Alexander W R, Reijonen H M, McKinley I G, 2015. Natural analogues: studies of geological processes relevant to radioactive waste disposal in deep geological repositories. Swiss Journal of Geosciences 108, 75–100. https://doi.org/10.1007/s00015-015-0187-y

AMEC, 2014. ConnectFlow Technical Summary, Release 11.2, Technical Summary Document. London: Jacobs Clean Energy Limited.

Andersson P, Byegård J, Winberg A, 2002. Final report of the TRUE Block Scale project, 2. Tracer tests in the block scale. SKB TR-02-14, Svensk Kärnbränslehantering AB.

André M, Malmström M E, Neretnieks I, 2008a. Measuring sorption coefficients and BET surface areas on intact drillcore and crushed granite samples. Radiochim Acta 96, 673–677.

André M, Malmström M E, Neretnieks I, 2008b. Determination of sorption properties of intact rock samples: New methods based on electromigration. Journal of Contaminant Hydrology 103, 71–81.

André M, Malmström M E, Neretnieks I, 2009. Specific surface area determinations on intact drillcores and evaluation of extrapola- tion methods for rock matrix surfaces. Journal of Contaminant Hydrology, 110, 1–8.

Appleyard P, Jackson P, Joyce S, Hartley L, 2018. Conditioning discrete fracture network models on intersection, connectivity and flow data. SKB R-17-11, Svensk Kärnbränslehantering AB.

ASME, 2006. Guide for Verification and Validation in Computational Solid Mechanics. ASME V&V 10-2006, The American Society of Mechanical Engineers.

Bair E S, 1994. Model (in)validation – A view from the courtroom. Guest Editorial. Ground Water 32, 530–531.

Baroni G, Tarantola S, 2014. A General Probabilistic Framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. Environmental Modelling & Software 51, 26–34. https://doi.org/10.1016/j.envsoft.2013.09.022

Bayes T R, 1763. Essay towards solving a problem in the doctrine of chances. Republished in Biometrika 45, 298–315.

Bear J, Tsang C-F, de Marsily G (eds), 1993. Flow and contaminant transport. Cambridge, MA: Academic press.

Berkowitz B. 2002. Characterizing flow and transport in fractured geological media: A review. Advances in Water Resources 25, 861–884. https://doi.org/10.1016/S0309-1708(02)00042-8

Berre I, Boon W M, Flemisch B, Fumagalli A, Gläser D, Keilegavlen E, Scotti A, Stefansson I, Tatomir A, Brenner K, Burbulla S, Devloo P, Duran O, Favino M, Hennicker J, Lee I-H, Lipnikov K, Masson R, K. Mosthaf K, Nestola M G C, Zulian P, 2020. Verification benchmarks for single-phase flow in three-dimensional fractured porous media. Advances in Water Resources 147, 103759. https://doi.org/10.1016/j.advwatres.2020.103759

Beven K, 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16, 41–51.

Beven K, 2002. Towards a coherent philosophy for modelling the environment. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 458, 2465–2484.

Beven K, 2006. A manifesto for the equifinality thesis. Journal of Hydrology 320, 18–36.

Beven K, 2009. Environmental Modelling – An Uncertain Future? Routledge: London.

Birgersson L, Neretnieks I, 1990. Diffusion in the matrix of granitic rock: Field test in the Stripa mine. Water Resources Research 26, 2833–2842.

Birgersson L, Moreno L, Neretnieks I, Widén H, Ågren T, 1993. A tracer migration experiment in a small fracture zone in granite. Water Resources Research 29, 3867–3878.

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

Biondi D, Freni G, Iacobellis V, Mascaro G, Montanari A, 2012. Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. Physics and Chemistry of the Earth, Parts A/B/C 42-44, 70–76. https://doi.org/10.1016/j.pce.2011.07.037

Black J, Woodman N, Barker J, 2016. Groundwater flow into underground openings in fractured crystalline rocks: an interpretation based on long channels. Hydrogeology Journal DOI 10.1007/s10040-016-1511-y. https://doi.org/10.1007/s10040-016-1511-y

Bodin J, Delay F, de Marsily G, 2003a. Solute transport in a single fracture with negligible matrix permeability: 1. Fundamental mechanisms. Hydrogeology Journal 11, 418–433.

Bodin J, Delay F, de Marsily G, 2003b. Solute transport in a single fracture with negligible matrix permeability: 2. Mathematical formalism. Hydrogeology Journal 11, 434–454.

Bond A E, Brusky I, Chittenden N, Feng X-T, Kolditz O, Lang P, Lu R, McDermott C, Neretnieks I, Peng-Zhi Pan P-Z, Jan Sembera J, Shao H, Yasuhara H, Zheng H, 2016. Development of approaches for modelling coupled thermal–hydraulic–mechanical–chemical processes in single granite fracture experiments. Environmental Earth Sciences 75. https://doi.org/10.1007/s12665-016-6117-0

Bonnet E, Bour O, Odling N E, Davy P, Main I, Cowie P, Berkowitz B, 2001. Scaling of fracture systems in geological media. Reviews of Geophysics 39(3), 347. https://doi.org/10.1029/1999RG000074

Box G E P, 1976. Science and statistics. Journal of the American Statistical Association 71, 791–799.

Box G E P, Luceño A, del Carmen Paniagua-Quiñones M, 2009. Statistical Control by Monitoring and Adjustment. New York, NY: John Wiley & Sons.

Bredehoeft J D, 2003. From models to performance assessment: The conceptualization problem. Ground Water 41, 571–577. https://doi.org/10.1111/j.1745-6584.2003.tb02395.x

Bredehoeft J, 2005. The conceptualization model problem – surprise. Hydrogeology Journal 13, 37–46. https://doi.org/10.1007/s10040-004-0430-5

Bredehoeft J D, Konikow L F, 1993. Ground-water models: Validate or Invalidate. Ground Water, 31, 178–179.

Brunetti C, Linde N, Vrugt J A, 2017. Bayesian model selection in hydrogeophysics: Application to conceptual subsurface models of the South Oyster Bacterial Transport Site, Virginia, USA. Advances in Water Resources, 102, 127–141. https://doi.org/10.1016/j.advwatres.2017.02.006

Byegård J, Hakami E, Hjerne C, Nordqvist R, Cvetkovic V, Drake H, Tullborg E-L, Winberg A, 2017. TRUE-1 Completion. Final report. SKB TR-12-11, Svensk Kärnbränslehantering AB.

Bym T, Hermanson J, 2018. Methods and workflow for geometric and hydraulic conditioning. SKB R-17-12, Svensk Kärnbränslehantering AB.

Bym T, Follin S, 2019. A numerical study of channelling, in heterogeneous versus calibrated homogeneous discrete fracture network realisations. SKB R-19-24, Svensk Kärnbränslehantering AB.

Caine J S, Evans J P, Forster C B, 1996. Fault zone architecture and permeability structure. Geology 24, 1025–1028.

Chiles P, 1987. Three dimensional geometric modelling of a fracture network. Proceedings of DOE/AECL Conference on Geostatistical, Sensitivity and Uncertainty Methods for Groundwater Flow and Radionuclide Transport Modelling. Battelle Press.

Christensen S, Doherty J D, 2008. Predictive error dependencies when using pilot points and singular value decomposition in groundwater model calibration. Advances in Water Resources, 31, 674–700.

Christie M, Glimm J, Grove J W, Higdon D M, Sharp D H, Wood-Schultz M M, 2005. Error Analysis and Simulations of Complex Phenomena. Los Alamos Science 29, 6–25.

Côme B, Chapman N A, 1987. Natural Analogues in Radioactive Waste Disposal. Dordrecht: Springer: Dordrecht. https://doi.org/10.1007/978-94-009-3465-8

Crout N M J, Kokkonen T, Jakeman A J, Norton J P, Newham L T, Anderson R, Assaf H, Croke B F W, Gaber N, Gibbons J, Holzworth D, Mysiak J, Reichl J, Seppelt R, Wagener T, Whitfield P, 2008. Good Modelling Practise. In Jakeman A J, Voinov A A, Rizzoli A E, Chen S H (eds). Environmental Modelling, Software and Decision Support: State of the Art and New Perspective. U.S. Environmental Protection Agency Papers 73. 1st ed. Amsterdam: Elsevier.

CRWMS M&O, 2000. Natural Analogs for the Unsaturated Zone. ANL-NBS-HS-000007 REV 00, Civilian Radioactive Waste Management System Management and Operating Contractor, United States.

Dausman A M, Doherty J, Langevin C D, Sukop M C, 2010. Quantifying data worth toward reducing predictive uncertainty. Ground Water 48, 729–740.

Davis P A, Olague N E, Goodrich M T, 1991. Approaches for the Validation of Models Used for Performance Assessment of High-Level Nuclear Waste Repositories. Report NUREG/CR-5537, U.S. Nuclear Regulatory Commission.

Davy P, Le Goc R, Darcel C, Selroos J-O, 2023. Scaling of fractured rock flow. Proposition

of indicators for selection of DFN based flow models. Comptes Rendus. Géoscience 355, 1–23.

de Marsily G, Combes P, Goblet P, 1992. Comment on 'Groundwater models cannot be validated', by L.F. Konikow and J. D. Bredehoeft. Advances in Water Resources, 15, 367–369.

Dershowitz W, Lee G, Geier J, Foxford T, LaPointe P, Thomas A, 2019. Fracman, Interactive Discrete Feature Data Analysis, Geometric Modelling, and Exploration Simulation, User Documentation. Version: 7.5. Redmond, WA: Golder Associates, Inc.

Dershowitz, W, Herda H, 1992. Interpretation of fracture spacing and intensity. In Tillerson and Wawersik (eds). Proceedings of the 33rd U.S. Symposium on Rock Mechanics, Santa Fe, New Mexico 8–10 June 1992. Rotterdam: Balkema, 757–766.

Dimmen V, Rotevatn A, Peacock D C P, Nixon C W, Nærland K, 2017. Quantifying structural controls on fluid flow: Insights from carbonate-hosted fault damage zones on the Maltese Islands. Journal of Structural Geology 101, 43–57. https://doi.org/10.1016/j.jsg.2017.05.012

Edmonds B, Le Page C, Bithell M, Chattoe-Brown E, Grimm V, Meyer R, Montañola-Sales C, Ormerod P, Root H, Squazzoni F, 2019. Different modelling purposes. Journal of Artificial Societies and Social Simulation 22, 6.

Eker S E, Rovenskaya S, Langan, Obersteiner M, 2019. Model validation: A bibliometric analysis of the literature. Environmental Modelling & Software 117, 43–54. https://doi.org/10.1016/j.envsoft.2019.03.009

Enemark T, Peeters L J M, Mallants D, Batelaan O, 2019. Hydrogeological conceptual model building and testing: A review. Journal of Hydrology 569, 310–329. https://doi.org/10.1016/j.jhydrol.2018.12.007

Ewing R C, 1993. Long-term predictions using natural analogues. In W.M. Murphy and L.A. Kovach (eds). The Role of Natural Analogues in Geologic Disposal of High-Level Nuclear Waste. CNWRA 93-020, Center for Nuclear Waste Regulatory Analyses, U.S.A.

Fienen M N, Doherty J E, Hunt R J, Reeves H W, 2010. Using Prediction Uncertainty Analysis to Design Hydrologic Monitoring Networks: Example Applications from the Great Lakes Water availability pilot project. Scientific Investigations Report 2010–5159, U.S.A.

Finsterle S, 2015. Practical notes on local data-worth analysis. Water Resources Research 51, 9904–9924. https://doi.org/10.1002/2015WR017445

Finsterle S, Lanyon B, 2022. Pragmatic validation of numerical models used for the assessment of radioactive waste repositories: A perspective. In Energies 15, 3585. https://doi.org/10.3390/en15103585

Finsterle S, Lanyon B, Åkesson M, Baxter S, Bergström M, Bockgård N, Dershowitz W, Dessirier B, Frampton A, Fransson Å, Gens A, Gylling B, Hančilová I, Holton D, Jarsjö J, Kim J-S, Kröhn K-P, Malmberg D, Pulkkanen V M, Sawada A, Sjöland A, Svensson U, Vidstrand P, Viswanathan H, 2019. Conceptual uncertainties in modelling the interaction between engineered and natural barriers of nuclear waste repositories in crystalline rock. In Norris S, Neeft E A C, van Geet M (eds). Multiple Roles of Clays in Radioactive Waste Confinement, Geological Society, London, Special Publications 482, 261–283. https://doi.org/10.1144/SP482.12

Follin S, 2008. Bedrock hydrogeology Forsmark. Site descriptive modelling, SDM-Site Forsmark. SKB R-08-95, Svensk Kärnbränslehantering AB.

Follin S, Stigsson M, 2014. A transmissivity model for deformation zones in fractured crystalline rock and its possible correlation to in situ stress at the proposed high-level nuclear waste repository site at Forsmark, Sweden. Hydrogeology Journal 22, 299–311. https://doi.org/10.1007/s10040-013-1078-9

Follin S, Hartley L, Rhén I, Jackson P, Joyce S, Roberts D, Swift B, 2014. A methodology to constrain the parameters of a hydrogeological discrete fracture network model for sparsely fractured crystalline rock, exemplified by data from the proposed higdoih-level nuclear waste repository site at Forsmark, Sweden. Hydrogeology Journal, 22, 313–331. https://doi.org/10.1007/s10040-013-1080-2

Fossen H, A. Rotevatn A, 2016. Fault linkage and relay structures in extensional settings – a review. Earth-Science Reviews 154, 14–28.

Gartrell A, Zhang Y, Lisk M, Dewhurst D, 2004. Fault intersections as critical hydrocarbon leakage zones: Integrated field study and numerical modelling of an example from the Timor Sea, Australia. Marine and Petroleum Geology 21, 1165–1179. https://doi.org/10.1016/j.marpetgeo.2004.08.001

Gelhar L W, Welty C, Rehfeldt C, K R, 1992. A critical review of data on field-scale dispersion in aquifer. Water Resources Research 28, 1955–1974.

Gelman A, Carlin J B, Stern H S, Dunson D B, Vehtari A, Rubin D, 2013. Bayesian Data Analysis, Boca Raton FL: CRC Press.

Gupta H V, Sorooshian S, Yapo P O, 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resources Research 34, 751–763. https://doi.org/10.1029/97WR03495

Gustafson G, Gylling B, Selroos J O, 2009. The Äspö Task Force on groundwater flow and transport of solutes: bridging the gap between site characterization and performance assessment for radioactive waste disposal in fractured rocks. Hydrogeology Journal 17, 1031–1033. https://doi.org/10.1007/s10040-008-0419-6

Gylling B, 1997. Development and applications of the channel network model for simulations of flow and solute transport in fractured rock. PhD thesis. Royal Institute of Technology, Stockholm.

Gylling B, Moreno L, Neretnieks I, 1999. The Channel Network Model – A tool for transport simulation in fractured media. Groundwater 37, 367–375.

Hartley L, Roberts D, 2012. Summary of discrete fracture network modeling as applied to hydrogeology of the Forsmark and Laxemar sites. SKB R-12-04, Svensk Kärnbränslehantering AB.

Hartley L, Appleyard P, Baxter S, Hoek J, Joyce S, Mosley K, Williams T, Fox A, Cottrell M, La Pointe R, Gehör S, Darcel C, Le Goc C, Aaltonen I, Vanhanarkaus O, Löfman J, Poteri A, 2018. Discrete Fracture Network Modeling (Version 3) in Support of Olkiluoto Site Description 2018. Appendix J: Validation Plan for ODFN3. Posiva Working report WR 2017-32, Posiva, Finland.

Hartley L, Baxter S, Carty J, Follin S, Libby S, 2021. Exploratory integration of discrete fracture network models and 1D stress models with data from hydraulic tests for the shallow bedrock at the Forsmark site. SKB R-21-13, Svensk Kärnbränslehantering AB.

Hassanizadeh S M, Carrera J, 1992. Validation of geo-hydrological models. Editorial. Advances in Water Resources 15, 1–3.

Hawkins D M, 2004. The problem of overfitting. Journal of Chemical Information and Computer Sciences 44, 1–12.

Hebel L C, Christensen E L, Donath F A, Falconer W E, Lidofsky L J, Moniz E J, Moss T H, Pigford R L, Pigford T H, Rochlin G I, Silsbee R H, Wrenn M E, Frauenfelder H, Cairns T L, Panofsky W K H, Simmons M G, 1978. Report to the American Physical Society by the study group on nuclear fuel cycles and waste management. Reviews of Modern Physics 50, S1–S176.

Hill M C, Middlemis H, Hulme, Poeter P E, Riegger J, Neuman S P, Williams H, Anderson M, 2004. Brief overview of selected groundwater modeling guidelines. In Kovar K, Bruthans J, Hrkal Z (eds). Finite-Element Models, MODFLOW, and More, Solving groundwater problems. Proceedings, 105–120, Carlsbad, Czech Republic, September 13–16, 2004. Carlsbad: FEM_MODFLOW.

Hjerne C, Nordqvist R, Harrström J, 2009. Compilation and analyses of results from cross-hole tracer tests with conservative tracers. SKB R-09-28, Svensk Kärnbränslehantering AB.

Holling C S (ed), 1978. Adaptive Environmental Assessment and Management. Chichester: Wiley.

Hunt R J, Doherty J E, Tonkin M J, 2007. Are models too simple? Arguments for increased parameterization. Ground Water, 45, 254–262.

Hyman J D, Karra S, Makedonska N, Gable C W, Painter S L, Viswanathan H S, 2015. DFN WORKS: Computers & Geosciences A discrete fracture network framework for modeling subsurface flow and transport Comput. Computers & Geosciences 84, 10–19. https://doi.org/10.1016/j.cageo.2015.08.001

Hyman J D, Jiménez-Martínez J, Viswanathan H S, Carey J W, Porter M L, Rougier E, Karra S, Kang Q, Frash L, Cheng L, Lei Z, O'Malley D, Makedonska N, 2016a. Understanding hydraulic fracturing: a multi-scale problem. Philosophical Transactions of the Royal Society A, 374:20150426. https://doi.org/10.1098/rsta.2015.0426

Hyman J D, Aldrich G, Viswanathan H, Makedonska N, Karra S, 2016b. Fracture size and transmissivity correlations: Implications for transport simulations in sparse three-dimensional discrete fracture networks following a truncated power law distribution of fracture size. Water Resources Research, 52, 6472–6489. https://doi.org/10.1002/2016WR018806

IAEA, 1999. Use of Natural Analogues to Support Radionuclide Transport Models for Deep Geological Repositories for Long Lived Radioactive Wastes. Report IAEA-TECDOC-1109, International Atomic Energy Agency, Austria.

IAEA, 2000. Extrapolation of Short Term Observations to Time Periods Relevant to the Isolation of Long Lived Radioactive Waste. Report IAEA-TECDOC-1177, International Atomic Energy Agency, Austria.

IAEA, 2018. IAEA Safety Glossary – Terminology Used in Nuclear Safety and Radiation Protection. 2018 Edition. International Atomic Energy Agency, Austria.

Incropera F P, Lavine A S, Bergman T L, DeWitt D P, 2007. Fundamentals of heat and mass transfer. Hoboken, NJ: Wiley.

Ishibashi T, Watanabe N, Hirano N, Okamoto A, Tsuchiya N, 2015. Beyond-laboratory-scale prediction for channeling flows through subsurface rock fractures with heterogeneous aperture distributions revealed by laboratory evaluation. Journal of Geophysical Research: Solid Earth 120, 106–124. https://doi.org/10.1002/2014JB011555

Joyce S, Simpson T, Hartley L, Applegate D, Hoek J, Jackson P, Swan D, Marsic N, Follin S, 2010. Groundwater flow modelling of periods with temperate climate conditions – Forsmark. SKB R-09-20, Svensk Kärnbränslehantering AB.

Joyce S, Hartley L, Applegate D, Hoek J, Jackson P, 2014. Multi-scale groundwater flow modeling during temperate climate conditions for the safety assessment of the proposed high-level nuclear waste repository site at Forsmark, Sweden. Hydrogeology Journal 22, 1233–1249. https://doi.org/10.1007/s10040-014-1165-6

Klemes V, 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal 31, 13–24.

Klimczak C, Schultz R, Parashar R, Reeves D M, 2010. Cubic law with aperture-length correlation: implications for network scale fluid flow. Hydrogeology Journal 18, 851–862. https://doi.org/10.1007/s10040-009-0572-6

Konikow L F, Bredehoeft J D, 1992. Ground-water models cannot be validated. Advances in Water Resources 15, 75–83

Kuhn T S, 1962. The Structure of Scientific Revolutions. Chicago IL: The University of Chicago Press.

Kuhn T S, 1977. Objectivity, Value Judgment, and Theory Choice, in: The Essential Tension: Selected Studies in Scientific Tradition and Change. Chicago, IL: University of Chicago Press, 320–339.

Lambe T W, 1973. Predictions in soil engineering. Geotechnique 23, 151–202. https://doi.org/10.1680/geot.1973.23.2.151

Lanyon G W, Davy P, Dershowitz W S, Finsterle S, Gylling B, Hyman J D, Neretnieks I, Uchida M, 2021. Pragmatic validation approach of geomechanics, flow, and transport models in fractured rock masses. American Rock Mechanics Conference, 5–7 October 2021.

Larsson A, 1992. The international projects INTRACOIN, HYDROCOIN and INTRAVAL. Advances in Water Resources 15, 85–87.

Leamer E E, 1985. Sensitivity analyses would help. The American Economic Review 75, 308–313.

Ligtenberg J H, 2005. Detection of fluid migration pathways in seismic data: implications for fault seal analysis. Basin Research 17, 141–153.

Liu L, Neretnieks I, Shahkarami P, Meng S, Moreno L, 2018. Solute transport along a single fracture in a porous rock: a simple analytical solution and its extension for modeling velocity dispersion. Hydrogeology Journal 26, 297–320. https://doi.org/10.1007/s10040-017-1627-8

Löfgren M, 2007a. Formation factor logging in situ by electrical methods in KFM01D and KFM08C. Forsmark site investigation. SKB P-07-138, Svensk Kärnbränslehantering AB.

Löfgren M, 2007b. Formation factor logging in situ by electrical methods in KLX07A, KLX08, KLX10 and KLX12A. Oskarshamn site investigation. SKB P-06-288, Svensk Kärnbränslehantering AB.

Löfgren M, Neretnieks I, 2003. Formation factor logging by electrical methods: Comparison of formation factor logs obtained in situ and in the laboratory. Journal of Contaminant Hydrology 61, 107–115.

Luis S J, McLaughlin D, 1992. A stochastic approach to model validation. Advances in Water Resources, 15, 15–32.

Maillot J, Davy P, Goc R, Le Darcel C, 2016. Connectivity, permeability and channeling in randomly-distributed and kinematically-defined Discrete Fracture Network Models. Water Resources Research 52, 8526–8545. https://doi.org/10.1002/2016WR018973

Marcos N, 2002. Lessons from Nature – The Behaviour of Technical and Natural Barriers in the Geological Disposal of Spent Nuclear Fuel. Espoo: Finnish Academy of Technology. (Acta Polytechnica Scandinavica, Civil Engineering and Building Construction Series No. 124)

Matott, L S, Babendreier J E, Purucker S T, 2009. Evaluating uncertainty in integrated environmental models: a review of concepts and tools. Water Resources Research 45. https://doi.org/10.1029/2008WR007301

McBurney P, 2012. What Are Models for? In Cossentino M, Kaisers M, Tuyls K, Weiss G (eds). Multi-Agent Systems. Lecture Notes in Computer Science vol. 7541. Berlin: Springer, 175–188.

McCombie C, McKinley I, 1993. Validation – Another perspective. Ground Water 31, 530–531.

Meng Z, Moreno L, Neretnieks I, Liu L, 2020. Modelling matrix diffusion in Task9B -LTD-SD. SKB P-20-01, Svensk Kärnbränslehantering AB.

Miller W, Alexander R, Chapman N, McKinley I, Smellie J, 1994. Natural analogue studies in the geological disposal of radioactive wastes. Studies in Environmental Science. Vol 57. Amsterdam: Elsevier.

Munafò M R, Smith G D, 2018. Repeating experiments is not enough. Nature 553, 399–401.

NEA, 2000. Features, Events and Processes (FEPs) for Geologic Disposal of Radioactive Waste, An International Database. OECD/NEA Nuclear Energy Agency, France.

NEA, 2006. NEA Feature, Event and Process (FEP) Database. Available at: https://www.oecd-nea.org/fepdb [July 30 2020].

Neretnieks I, 1980. Diffusion in the rock matrix: An important factor in radionuclide retardation? Journal of Geophysical Research: Solid Earth 85, 4379–4397.

Neretnieks I, 1983. A note on fracture dispersion mechanisms in the ground. Water Resources Research, 19, 365–370. https://doi.org/10.1029/WR019i002p00364

Neretnieks I, 2013. Some aspects of release and transport of gases in deep granitic rocks: possible implications for nuclear waste repositories. Hydrogeology Journal 2013, 1701–1716.

Neretnieks I, 2014. Stress-mediated closing of fractures – Impact of matrix diffusion. Journal of Geophysical Research: Solid Earth, 119. https://doi.org/10.1002/2013JB010645

Neretnieks I, 2018. Density driven mass transfer in repositories for nuclear waste. A note on tracer tests, channelling and leaking aquifers. SKB R-18-05, Svensk Kärnbränslehantering AB.

Neretnieks I, Moreno L, 2003. Prediction of some in situ tracer tests with sorbing tracers using independent data. Journal of Contaminant Hydrology 61, 351–360.

Neretnieks I, Moreno L, Liu L, Mahmoudzadeh B, Shahkarami P, Maskenskaya O, Kinnbom P, 2017. Use of infrared pictures to assess flowing channel frequencies and flowrates in fractured rocks. SKB R-17-04, Svensk Kärnbränslehantering AB.

Neretnieks I, Winberg-Wang, H, 2018. Density-Driven Mass Transfer in Repositories for Nuclear Waste Density-Driven Mass Transfer in Repositories for Nuclear Waste. Nucl. Technol. 00, 1–11. https://doi.org/10.1080/00295450.2018.1537460

Neuman S P, 1992. Validation of safety assessment models as a process of scientific and public confidence building. In Proceedings of High Level Radioactive Waste Management Conference. Vol. 2, Las Vegas, Nevada, 12–16 April 1992. La Grange Park, I: American Nuclear Society, 1404–1413.

Neuman S P, 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stochastic Environmental Research and Risk Assessment 17, 291–301.

Nixon C W, Nærland K, Rotevatn A, Dimmen V, Sanderson D J, Kristensen T B, 2020. Connectivity and network development of carbonate-hosted fault damage zones from western Malta. Journal of Structural Geology 141, 104212. https://doi.org/10.1016/j.jsg.2020.104212

Nordstrom D K, 2012. Models, validation, and applied geochemistry: Issues in science, communication, and philosophy. Applied Geochemistry 27, 1899–1919. https://doi.org/10.1016/j.apgeochem.2012.07.007

Odén M, Follin S, Öhman J, Vidstrand P, 2014. SR-PSU Bedrock hydrogeology, Groundwater flow modelling methodology, setup and results. SKB R-13-25, Svensk Kärnbränslehantering AB.

Ohlsson Y, Löfgren M, Neretnieks I, 2001. Rock matrix diffusivity determinations by in-situ electrical conductivity measurements. J Contaminant Hydrology 47, 117–125.

Oreskes N, 1998. Evaluation (not validation) of quantitative models. Environmental Health Perspectives 106, 1453–1460.

Oreskes N, 2000. Why predict? Historical perspectives on prediction in Earth Science. In Sarewitz D, Pielke R A Jr, Byerly R Jr (eds). Prediction, Science, Decision Making and the Future of Nature. Washington DC: Island Press.

Oreskes N, 2019. Why trust science? Princeton, New Jersey: Princeton University Press (University Center for Human Values series).

Oreskes N, Shrader-Frechette K, Belitz K, 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263, 641–656.

Park Y J, Lee K K, Berkowitz B, 2001. Effects of junction transfer characteristics on transport in fracture networks. Water Resources Research 37, 909–23.

Parker W, 2020. Model Evaluation: An Adequacy-for-Purpose View. Philosophy of Science 87, 457–477. https://doi.org/10.1086/708691

Palmquist K, Stanfors R, 1987. The Kymmen power station. TBM tunnel. Hydrogeological mapping and analysis. SKB TR- 87-26, Svensk Kärnbränslehantering AB.

Peacock D C P, Sanderson J, Rotevatn A, 2018. Relationships between fractures. Journal of Structural Geology 106, 41–53. https://doi.org/10.1016/j.jsg.2017.11.010

Peitgen H-O, Saupe D, 1988. Algorithms for random fractals, in: The Science of Fractal Images. Berlin: Springer Verlag.

Popper K, 1934. Die Logik der Forschung. Vienna: Julius Pringer Verlag.

Posiva, 2012. Safety case for the disposal of spent nuclear fuel at Olkiluoto – Performance assessment 2012. Posiva 2012-04, Posiva, Finland.

Posiva SKB, 2017. Safety functions, performance targets and technical design requirements for a KBS-3V repository. Conclusions and recommendations from a joint SKB and Posiva working group. Posiva SKB Report 01, Posiva Oy, Svensk Kärnbränslehantering AB.

Pruess K, Faybishenko B, Bodvarsson G S, 1999. Alternative concepts and approaches for modeling flow and transport in thick unsaturated zones of fractured rocks. Journal of Contaminant Hydrology, 38, 281–322. https://doi.org/10.1016/S0169-7722(99)00018-2

Refsgaard, J C, Henriksen, H J, 2004. Modelling guidelines—terminology and guiding principles. Advances in Water Resources 27, 71–82. https://doi.org/10.1016/j.advwatres.2003.08.006

Rhén I, Hartley L, 2009. Bedrock hydrogeology Laxemar Site descriptive modeling, SDM-Site Laxemar. SKB R-08-92, Svensk Kärnbränslehantering AB.

Rhén I, Forsmark T, Hartley L, Jackson P, Roberts D, Swan D, Gylling B, 2008. Hydrogeological conceptualisation and parameterisa- tion, Site descriptive modeling SDM-Site, Laxemar. SKB. R-08-78, Svensk Kärnbränslehantering AB.

Salas J, Gimeno M J, Auqué L, Molinero J, Juárez I, 2010. SR-Site – hydrogeochemical evolution of the Forsmark site. SKB TR-10-58, Svensk Kärnbränslehantering AB.

Saltelli, A. 2002. Sensitivity analysis for importance assessment. Risk Analysis, I. https://doi.org/10.1111/0272-4332.00040

Saltelli, A, Funtowicz S, 2014. When all models are wrong, Issues in Science and Technology 30, 79–58.

Saltelli A, Ratto M, Tarantola S, Campolongo F 2006. Sensitivity analysis practices: Strategies for model-based inference. Reliability Engineering & System Safety 91, 1109–1125. https://doi.org/10.1016/j.ress.2005.11.014

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S, 2008. Global Sensitivity Analysis: The Primer. Chichester: John Wiley and Sons.

Saltelli A, Pereira A G, Van der Sluijs J P, Funtowicz S, 2013. What do I make of your latinorum? Sensitivity auditing of mathematical modelling, Int. J. Foresight and Innovation Policy 9, 213–234.

Saltelli A, Bammer G, Bruno I, Charters E, Di Fiore M, Didier W, Nelson Espeland W, Kay J, Lo Piano S, Mayo D, Pielke Jr R, Portaluri T, Porter T M, Puy A, Rafols I, Ravetz J R, E. Reinert, D. Sarewitz, P. B. Stark, A. Stirling, van der Sluijs J, Vinei P, 2020. Five ways to ensure models serve society: A manifesto. Nature 582, 482–484.

Sanderson D J, Zhang X, 1999. Critical stress localization of flow associated with deformation of well-fractured rock masses, with implications for mineral deposits. Geological Society Special Publication, 155, 69–81. https://doi.org/10.1144/GSL.SP.1999.155.01.07

Sato H, Sawada A, Takayama Y, 2020. Development of technique for acquiring fracture void structure data for a single fracture and acquisition of hydraulic/mass transport properties in a fracture, JAEA-Research 2020-012, Japan Atomic Energy Agency. (In Japanese with English abstract and table of contents.)

Selroos J-O, Follin S, 2014a. Overview of hydrogeological site descriptive modeling conducted for the proposed high-level nuclear waste repository site at Forsmark, Sweden. Hydrogeology Journal 22, 295–298. https://doi.org/10.1007/s10040-013-1077-x

Selroos J-O, Follin S, 2014b. Overview of hydrogeological safety assessment modeling conducted for the proposed high-level nuclear waste repository site at Forsmark, Sweden. Hydrogeology Journal 22, 1229–1232. https://doi.org/10.1007/s10040-014-1163-8

Selroos J-O, Gylling B, 2023. How findings from a Multi-Annual International modeling initiative are implemented in a nuclear waste management organization. Energies 16, 1–23. https://doi.org/10.3390/en16020684

Selroos J-O, Walker D D, Ström A, Gylling B, Follin S, 2002. Comparison of alternative modelling approaches for groundwater flow in fractured rock. Journal of Hydrology 257, 174–188.

Selroos J-O, Mas Ivars D, Munier R, Libby S, Darcel C, Davy P, Trinchero P, 2022. Methodology for Discrete Fracture Network Modelling of the Forsmark Site. Volume I: Concepts, Data and Interpretation Methods. SKB R-20-11, Svensk Kärnbränslehantering AB.

Shahkarami P, Liu L, Moreno L, Neretnieks I, 2016. The effect of stagnant water zones on retarding radionuclide stransport in fractured rocks: An extension to the Channel Network Model. Journal of Hydrology 540, 1122–1135. https://doi.org/10.1016/j.jhydrol.2016.07.031

Skagius K, Neretnieks I, 1986a. Porosities and diffusivities of some non-sorbing species in crystalline rocks. Water Resources Research, 22, 389–398.

Skagius K, Neretnieks I, 1986b. Diffusivity measurements and electrical resistivity measurements in rock samples under mechanical stress. Water Resources Research 22, 570–580.

Skagius K, Neretnieks I, 1988. Measurements of cesium and strontium diffusion in biotite gneiss. Water Resources Research 24, 75–84.

SKB, 2004. Treatment of geosphere retention phenomena in safety assessments, Scientific basis of retention processes and their implementation in safety assessment models (WP2). SKB R-04-48, Svensk Kärnbränslehantering AB.

SKB, 2008a. Site description of Forsmark at completion of the site investigation phase. SDM-Site Forsmark. SKB TR-08-05, Svensk Kärnbränslehantering AB.

SKB, 2008b. Bedrock hydrogeology Forsmark. Site descriptive modelling, SDM-Site Forsmark. SKB R-08-95, Svensk Kärnbränslehantering AB.

SKB, 2010. Radionuclide transport report for the safety assessment SR-Site. SKB TR-10-50, Svensk Kärnbränslehantering AB.

SKB, 2011. Long-term safety for the final repository for spent nuclear fuel at Forsmark Main report of the SR-Site project, vols. I, II and II. SKB TR-11-01, Svensk Kärnbränslehantering AB.

Spiegelhalter D J, Best N G, Carlin B P, van Der Linde A, 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B: Statistical Methodology 64, 583–639. https://doi.org/10.1111/1467-9868.00353

Stanfors R, 1987. The Bolmen tunnel project. Evaluation of geophysical site investigation methods. SKB TR-87-25, Svensk Kärnbränslehantering AB.

Starfield A M, Cundall P, 1988. Towards a methodology for rock mechanics modelling. International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts 25, 99–106. https://doi.org/10.1016/0148-9062(88)92292-9

Starfield A M, Jarre A, 2011. Interdisciplinary Modeling for an Ecosystem Approach to Management in Marine Social-Ecological Systems. World Fisheries: A Social-Ecological Analysis, 105–119.

Starfield A M, Smith K A, Bleloch A L, 1990. How to Model It: Problem Solving for the Computer Age. New York, NY: McGraw-Hill.

Stephens M B, Follin S, Petersson J, Isaksson H, Juhlin C, Simeonov A, 2015. Review of the deterministic modelling of deformation zones and fracture domains at the site proposed for a spent nuclear fuel repository, Sweden, and consequences of structural anisotropy. Tectonophysics 653, 68–94. https://doi.org/10.1016/j.tecto.2015.03.027

Stigsson M, 2019. Structural uncertainties of rock fractures and their effect on flow and tracer transport. PhD thesis. Royal Institute of Technology, Sweden.

Stuart A M, 2010. Inverse problems: A Bayesian perspective. Acta Numerica 19, 451–559. https://doi.org/10.1017/S0962492910000061

STUK, 2015. STUK's review on the construction license stage post closure safety case of the spent nuclear fuel disposal in Olkiluoto. STUK-B 197, Radiation and Nuclear Safety Authority, Finland.

Tarantola A, 2004. Inverse Problem Theory and Methods for Model Parameter Estimation. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Taylor G, 1953. Dispersion of soluble matter in solvent flowing slowly through a tube. Proceedings of the Royal Society of London: A 219, 186–203.

Tetsu K, Sawada A, 2003. Water permeability test of Rock Specimen with natural fractures using high viscosity liquid 2. JNC TN8430 2003-004, Japan Nutrition Co. Ltd, 1–65. (In Japanese with English abstract.)

Thacker B H, Doebling S W, Hemez F M, Anderson M C, Pepin J E, Rodriguez E A, 2004. Concepts of Model Verification and Validation. LANL Report LA-14167-MS. https://doi.org/10.2172/835920

Tsang C-F, 1991. The Modeling process and model validation. Ground Water 29, 825–831.

Tsang Y W, Tsang C-F, 1989. Flow channeling in a single fracture as a two-dimensional, strongly heterogeneous permeable medium. Water Resources Research 25, 2076–2080.

Tsang C-F, Neretnieks I, 1998. Flow channeling in heterogeneous fractured rocks. Reviews of Geophysic 36, 275–298.

Tsang C-F, Neretnieks I, Tsang Y, 2015. Hydrologic issues associated with nuclear waste repositories. Water Resources Research 51, 6923–6972. https://doi.org/10.1002/2015WR017641

USDOE Blue Ribbon Committee, 2012. Blue Ribbon Commission on America's Nuclear Future – Report to the Secretary of Energy. Washington, D. C: United States Department of Energy.

van der Sluijs J, Craye M, Funtowicz S, Kloprogge P, Ravetz J, Risbey J, 2005. Combining quantitative and qualitative measures of uncertainty in model based environmental assessment: the NUSAP System. Risk Analysis 25, 481–492.

Voutilainen M, Siitari-Kauppi M, Sardini P, Lindberg A, Timonen J, 2012. Pore-space characterization of an altered tonalite by X-ray computed microtomography and the 14C-labeled-polymethylmethacrylate method. Journal of Geophysical Research: Solid Earth 117. https://doi.org/10.1029/2011JB008622

Wagener T, Kollat J, 2007. Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox. Environmental Modelling & Software 22, 1021–1033.

Wennberg O P, Casini G, Jonoud S, Peacock D C P, 2015. The characteristics of open fractures in carbonate reservoirs and their impact on fluid flow: a discussion. Petroleum Geoscience 22, 91–104. https://doi.org/10.1144/petgeo2015-003

Winberg A, Andersson P, Hermanson J, Byegård J, Cvetkovic V, Birgersson L, 2000. Äspö Hard Rock Laboratory. Final report of the first stage of the tracer retention understanding experiments. SKB TR-00-07, Svensk Kärnbränslehantering AB.

Witherspoon P A, Wang J S Y, Iwai K, Gale J E, 1980. Validity of Cubic Law for fluid flow in a deformable rock fracture. Water Resources Research 16, 1016–1024. https://doi.org/10.1029/WR016i006p01016

Ye J, 1998. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association 93, 120–131.

Zimmerman, R W, Bodvarsson G S, 1996. Hydraulic conductivity of rock fractures. Transport in Porous Media 23, 1–30.

Öhberg A, Rouhiainen P, 2000. Posiva groundwater flow measuring techniques. Posiva 2000-12, Posiva, Finland.

Appendix A

Note on model space

Philippe Davy, University of Rennes

The concept of model space is not new, but it has not really been described precisely and certainly not in the context of such a broad methodology as the one we are working on. In short, the model space is the range of all possibilities, i.e., concepts, equations, parameters. When everything is fixed, there is "model" that can be compared to data and extrapolated to prediction conditions.

The model space is not an abstract concept but a decision by all parties involved with the current scientific knowledge to fix the concepts, the equations and the range of corresponding parameters. If modellers decide to focus on a limited range of parameters, or on a specific empirical concept (knowing that others exist), then they are investigating a sub-model space (SMS). The concept of SMS is very important since that is what is done in practice. It is necessary to make assumptions for many reasons (from accumulated experience to software availability) but the point is that this should be done in complete transparency, and with discussion about the consequences of our choices.

A useful concept could be the "baseline model", which is the most likely model given data and current knowledge, and/or the "baseline SMS", which is the ensemble of reasonable conceptual assumptions that allow us to derive the "baseline model" by calibrating on site data. In theory, if the model space is set, the way it is solved for a set of parameters should not affect the outcome, e.g. the eventual model. The methodology requirement is just to test the accuracy of the numerical methods. DFNs, effective continua, or channel network models are more than numerical methods; they are concepts of the underlying geometrical structure. All of them are SMS. Not all the numerical concepts have the same status. Some are developed for numerical efficiency reasons. This is the case of ECPM, which aims at replacing DFNs to address large-scale systems. ECPM is basically a reduced parameter model, and the methodology requires to assess the consequences of replacing local complexities by effective parameter. Some are developed for theoretical reasons, e.g. truly different concepts. This is the case of channel network models compared to DFNs. The model space is supposed to be at least the union of the sub-spaces, but saying that doesn't really help if the SMS lead to very different predictions. The skill of the modellers and geologists is to give a weight to each of the SMS by assessing their physical/geological likelihood. This can be extremely speculative, but hopefully can be done by analyzing the likelihood of the assumptions underlying these models.

Appendix B

Fracture intersections and a tracer experiment

Masahiro Uchida, Fracture Flow Solutions

B.1 Fracture Intersections

There are many reports that describe Fracture Intersection Zones (FIZ) may be more conductive than other parts of a fracture. However, there is only a limited number of reports, e.g., Tetsu and Sawada (2003), which quantitatively measured the transmissivity of FIZ. Tetsu and Sawada (2003) used a 50 cm cube granite rock block which includes two fractures with Y-type intersection. They attached rubber gaskets with 6 × 6 windows to 4 surfaces of rock block as shown in Figure B-1. Each gasket has two tubes, one for inject/withdraw water and the other to monitor the head. They controlled the head at injection window and extraction window, and measured flow rate. Then, they calculated transmissivity along the path between injection and extraction windows. They reported that transmissivity along the FIZ was the highest among all combinations of injection and extraction windows (Figure B-2), especially the transmissivity along FIZ was more than an order of magnitude greater than the lowest transmissivity in the normal part of fracture.

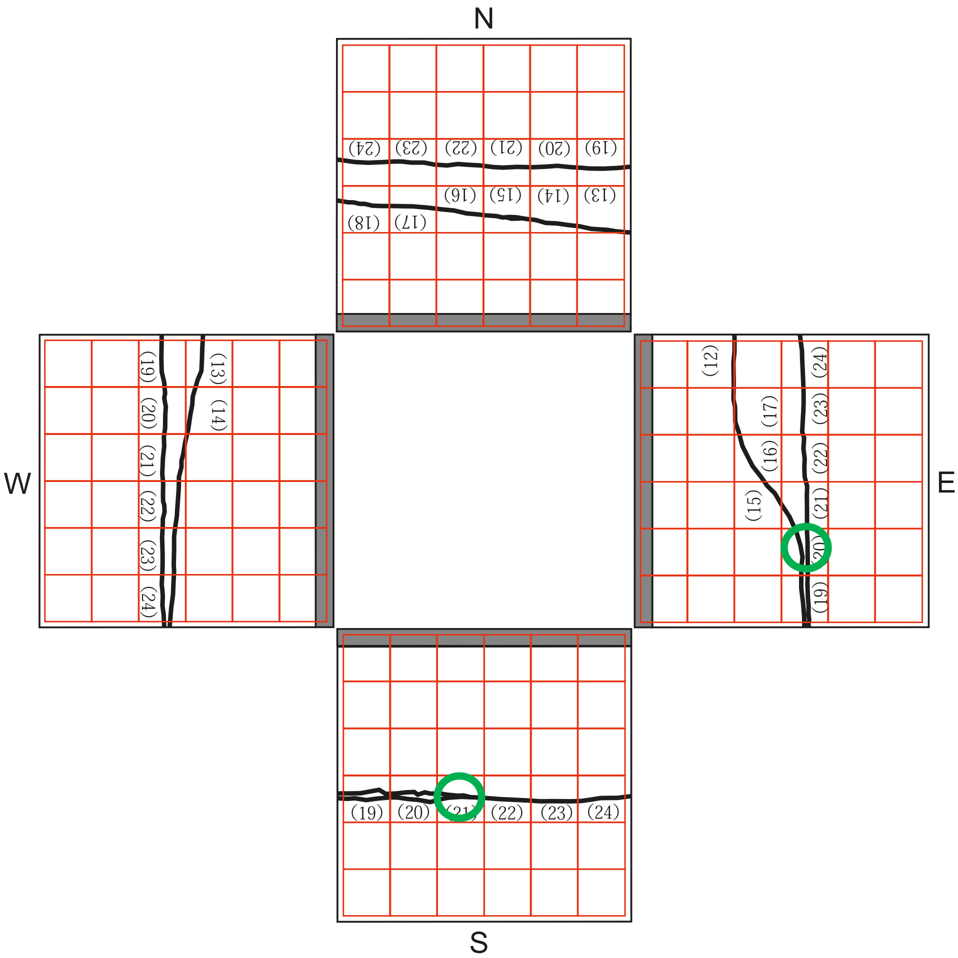


Figure B-1. Fracture intersection experiment by Tetsu and Sawada (2003). Y type fracture intersections are indicated with green circles in the figure. The upper fracture in this figure is wavy and fracture intersection is not clear in "S" surface.

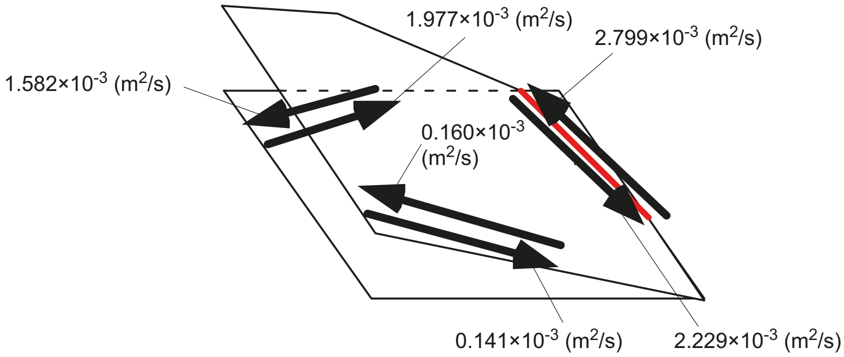
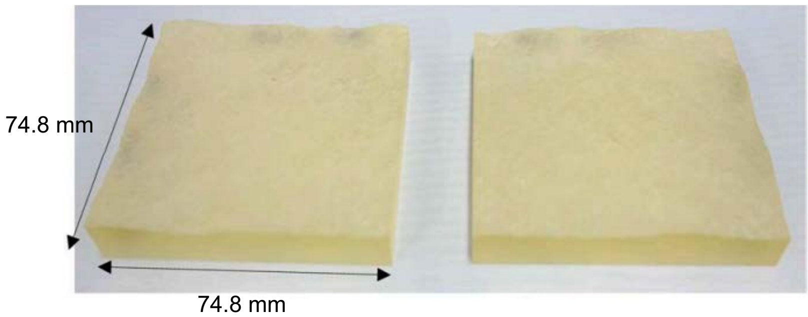


Figure B-2. *Comparison of transmissivity along fracture intersection and respective fractures. Transmissivities along fracture intersection are highest with order of 10^{-3} m²/s, whereas transmissivities in the upper fracture and the lower fracture are 10^{-5} to 10^{-3} m²/s. The transmissivity of the upper fracture is lower than that of the lower fracture. (Note that the transmissivities shown in the figure are rather high end among the large number of measurements.)*

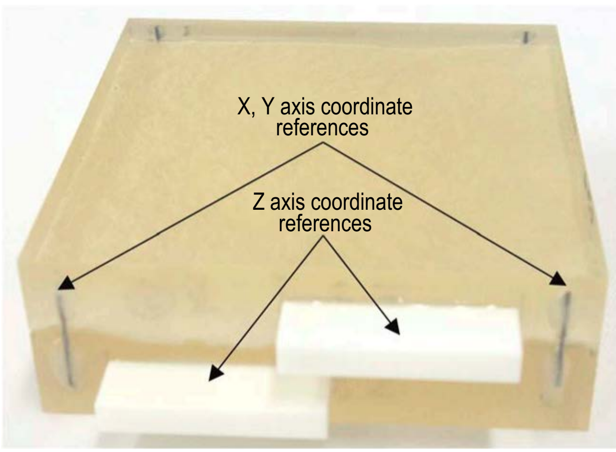
B.2 Tracer experiment using replica specimen casted from granite artificial tensile fracture and observing tracer movement with transmitting light

Sato et al. 2020, JAEA

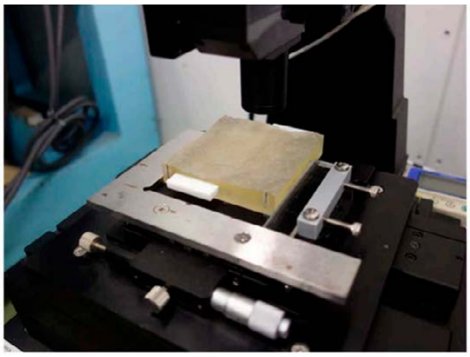
Prepare replica of fracture

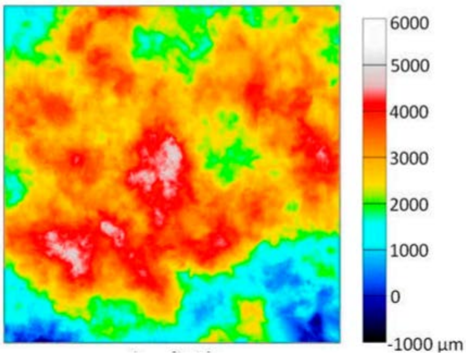


Install coordinate references



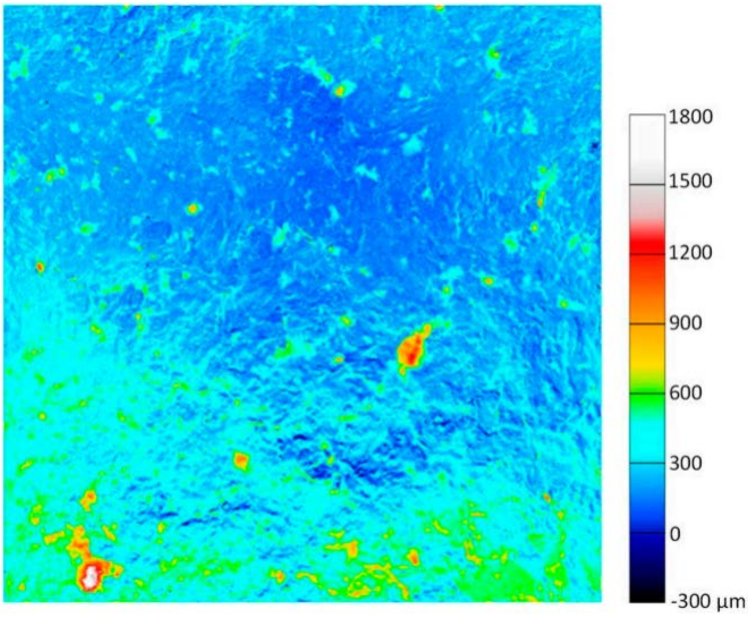
Obtain fracture surface topography with laser displacement gauge



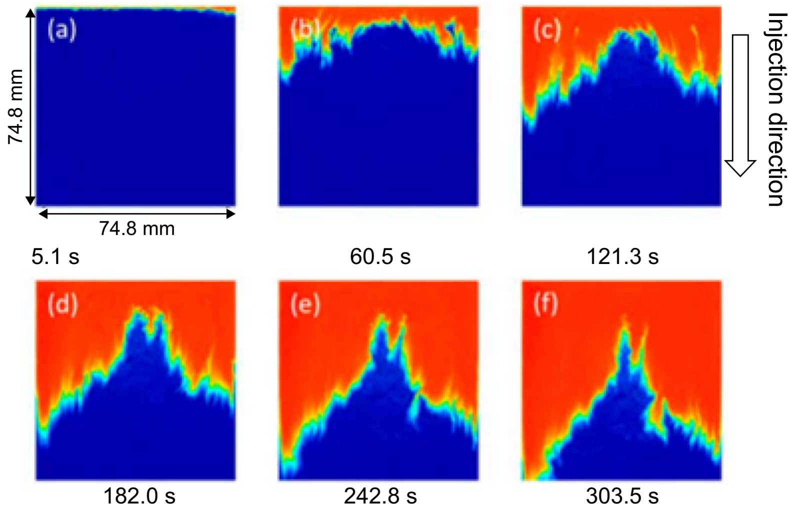


In total, four batches of laser measurement were carried out, since the difference in the height of fracture surface is beyond the measurement range of the displacement gauge.

Reconstruct aperture distribution from surface topography using coordinate references



Example of tracer experiment results



Arkitektkopia AB, Bromma, 2024

SKB is responsible for managing spent nuclear fuel and radioactive waste produced by the Swedish nuclear power plants such that man and the environment are protected in the near and distant future.

skb.se