skb.se

SKB TR-26-01

ISSN 1404-0344
ID 2116436
Month Jun 2026

# Evaluation of Task 10.2.2 - Prediction of flow through a known fracture geometry at different normal loads

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

Stefan Finsterle
Finsterle GeoConsulting, LLC, Kensington, California, USA

*Keywords:* Flow and transport modelling, Fracture properties, Hydromechanics, Pragmatic model validation, Single fracture, Upscaling, Contact modelling, Cubic law

SVENSK KÄRNBRÄNSLEHANTERING

SWEDISH NUCLEAR FUEL AND WASTE MANAGEMENT CO Box 3091, SE-169 03 Solna Phone: +46 8 459 84 00

# Summary

This report evaluates the work performed under Task 10.2.2 of the SKB Task Force on Groundwater Flow and Transport of Solutes (GWFTS). In Task 10.2.2, four modelling teams (Amphos$^{21}$-KTH, TUL, VTT, and NTU-Taipower) exercised the pragmatic model validation approach for numerical models that simulate fluid flow through an upscaled single natural fracture under varying normal loads. The models are developed starting with high-resolution laser scan data of the surfaces of a fracture embedded in a 20 × 20 × 25 $cm^3$ rock sample. The roughness data were used to construct the fracture void space and its deformation in response to normal loading, ranging from 0 to 40 MPa. The resulting aperture distributions were upscaled in preparation for flow simulations that either used aperture-transmissivity transformations in combination with Darcy's law, or solutions of the Navier-Stokes equations. A prediction-outcome (P/O) exercise was conducted to validate the flow model. Intermediate submodels required for the development of the geometric and mechanical models were either calibrated or validated using displacement measurements and/or contact pressure film data. Flow channelling characteristics were also examined but not formally validated.

The core of the study was a blind prediction-outcome exercise, where experimental flow rates were withheld from the modelling groups until their simulations were complete. The P/O exercise revealed that most models significantly overestimated flow rates, by one to several orders of magnitude. However, most models successfully captured the nonlinear decreasing trend of flow as normal stress increased. The discrepancies between model predictions and experimental outcome were attributed to multiple factors, specifically the data and models used to derive the initial aperture distribution, which appeared to be biased because of limitations in the raw data, misalignment of the fracture surfaces, and/or constraints imposed during the development of the geometric and mechanical models. Alternative upscaling methods, assumptions about the aperture-transmissivity relation, and choice of the governing flow equations were also examined as part of the sensitivity analyses.

The modelling groups' simulation and validation efforts were examined by the Task 10 evaluator who focused on the work and documentation related to pragmatic model validation. The evaluation report concludes that while the simulation models were of high technical quality, the exercise highlighted challenges in executing the pragmatic validation approach, specifically:

- *Model purpose and validation acceptance criteria*: Descriptions of the specific model purpose based the model's intended use and acceptable level of prediction uncertainty were generally missing, so were formulations of related validation acceptance criteria. As a result, the "fit-for-purpose" validation was performed without a clear target nor a defensible way to evaluate its success, making it difficult to assess the models' usefulness for their application in subsequent subtasks.
- *Validation vs. calibration*: Several groups found it necessary to recalibrate their models after noting the considerable discrepancies to the experimental data. The evaluator noted that once validation data is used for calibration, it can no longer serve as an independent test of the model's predictive power.
- *Integration*: There was limited evidence that lessons learnt from the preceding subtask (Task 10.2.1) were transferred to Task 10.2.2 (particularly regarding the overestimation of apertures), or how the insights and conclusions (particularly limitations and uncertainties) will be propagated to the subsequent tasks.

The evaluator made the following suggestions relevant for the completion of Task 10:

- *Propagate uncertainty*: Ensure that conceptual and parametric uncertainties identified in single-fracture studies are carried forward into larger-scale discrete fracture network models.
- *Model complexity*: Use a level of complexity in input data, model conceptualisation, and their numerical implementation that is consistent with the model's intended use.
- *Meta-Analysis*: Perform a combined analysis of all participating models to gain a more robust understanding of the collective ability to predict flow in fractured rock.

Overall, Task 10.2.2 served as a valuable exercise in developing and linking various submodels and in practicing the pragmatic validation approach. It helped clarify the challenges and inherent limitations of predicting upscaled flow behaviour solely based on high-resolution geometric data.

# Sammanfattning

Denna rapport utvärderar arbetet som utförts under modelleringsuppgift 10.2.2 av SKB Task Force on Groundwater Flow and Transport of Solutes (GWFTS). I modelleringsuppgift 10.2.2 använde fyra modelleringsteam (Amphos$^{21}$-KTH, TUL, VTT och NTU-Taipower) den pragmatiska modellvalideringsmetoden för numeriska modeller som simulerar vätskeflöde genom en uppskalad enskild naturlig spricka under varierande normallaster. Modellerna utvecklas med utgångspunkt från högupplösta laserskanningsdata av ytorna på en spricka belägen i ett bergprov på 20 × 20 × 25 cm$^3$. Ojämnhetsdata användes för att representera sprickhålrummet och dess deformation som svar på normallaster, från 0 till 40 MPa. De resulterande aperturfördelningarna skalades upp som förberedelse för flödessimuleringar som antingen använde apertur-transmissivitetstransformationer i kombination med Darcys lag eller lösningar av Navier-Stokes-ekvationerna. En prediktions-utfallsövning (P/O) genomfördes för att validera flödesmodellen. Mellanliggande delmodeller som behövdes för utvecklingen av de geometriska och mekaniska modellerna antingen kalibrerades eller validerades med hjälp av förskjutningsmätningar och/eller kontakttryckfilmsdata. Flödeskanaliseringsegenskaper undersöktes också men validerades inte formellt.

Kärnan i studien var en ”blind” prediktions-utfallsövning, där experimentella flödeshastigheter undanhölls från modelleringsgrupperna tills deras simuleringar var slutförda. P/O-övningen visade att de flesta modellerna överskattade flödeshastigheterna avsevärt, med en till flera storleksordningar. De flesta modeller lyckades dock fånga den ickelinjära minskande trenden för flödet när normalspänningen ökade. Skillnaderna mellan modellprediktioner och experimentellt resultat tillskrevs flera faktorer, specificerat som de data och modeller som användes för att härleda den initiala aperturfördelningen, vilken verkade vara partisk på grund av begränsningar i rådata, feljustering av sprickytorna och/eller begränsningar som infördes under utvecklingen av de geometriska och mekaniska modellerna. Alternativa uppskalningsmetoder, antaganden om apertur-transmissivitetsförhållandet och val av styrande flödesekvationer undersöktes också som en del av känslighetsanalyserna.

Modelleringsgruppernas simulerings- och valideringsinsatser granskades av utvärderaren för Task 10, som fokuserade på arbetet och dokumentationen relaterad till pragmatisk modellvalidering. Denna utvärderingsrapport drar slutsatsen att även om simuleringsmodellerna var av hög teknisk kvalitet, belyste övningen utmaningar i genomförandet av den pragmatiska valideringsmetoden, särskilt då det gäller:

- *Modellens syfte och valideringsacceptanskriterier:* Beskrivningar av det specifika modellensyftet baserat på modellens avsedda användning och acceptabla nivå av prediktionsosäkerhet saknades i allmänhet, liksom formuleringar av relaterade valideringsacceptanskriterier. Som ett resultat av detta utfördes "lämplighetsvalideringen" utan ett tydligt mål eller ett försvarbart sätt att utvärdera dess framgång, vilket gjorde det svårt att bedöma modellernas användbarhet för deras tillämpning i efterföljande deluppgifter.
- *Validering kontra kalibrering:* Flera grupper fann det nödvändigt att omkalibrera sina modeller efter att ha noterat de betydande avvikelserna jämfört med experimentdata. Utvärderaren noterade att när valideringsdata väl används för kalibrering kan de inte längre fungera som ett oberoende test av modellens prediktiva förmåga.
- *Integration:* Det var inte tydligt hur lärdomarna från föregående deluppgift (Task 10.2.1) överfördes till Task 10.2.2 (särskilt gällande överskattning av aperturer), eller hur insikterna och slutsatserna (särskilt gällande begränsningar och osäkerheter) kommer att föras vidare till efterföljande modelleringsuppgifter.

Utvärderaren gav följande förslag som är relevanta för slutförandet av Task 10:

- *Propagera osäkerhet:* Säkerställ att de konceptuella och parametriska osäkerheterna som identifierats i studier av enskilda sprickor överförs till större modeller med diskreta spricknätverk.
- *Modellkomplexitet:* Använd en komplexitetsnivå i indata, modellkonceptualisering och deras numeriska implementering som är förenlig med modellens avsedda användning.
- *Metaanalys:* Utför en kombinerad analys med alla deltagande modeller för att få en mer robust förståelse av den kollektiva förmågan att förutsäga flöde i sprickigt berg.

Sammantaget fungerade Task 10.2.2 som en värdefull övning i att utveckla och länka olika delmodeller och i att använda den pragmatiska valideringsmetoden. Den hjälpte till att klargöra utmaningarna och de inneboende begränsningarna då det gäller att förutsäga ett uppskalat flödesbeteende enbart baserat på högupplösta geometriska data.

# Content

# 1 Introduction

## 1.1 Background

Understanding, characterising, simulating, and predicting fluid flow and solute transport through fractured rock is relevant for supporting the safety case for a repository of radioactive waste in crystalline formations. These flow and transport processes in fractured host rocks are inherently complex, as they occur over multiple spatial scales, i.e., within the aperture field of a single fracture, along geometrically controlled or dynamically changing flow channels, within and across fracture intersections, and through the network of interconnected fractures of different scales and properties. Moreover, fluids and radionuclides interact with the adjacent rock mass, exchanging mass and energy, which affects the transport time and repository-induced thermal effects.

Numerical modelling is an essential tool for generic and site-specific analyses of flow and transport in fractured porous media. Based on characterisation data describing certain aspects of the system at discrete points in space and time, a conceptual and numerical model must be developed that is expected to make inferences or predictions at other locations, potentially on different spatial and/or temporal scales. Due to the limited data support, this extrapolation to a different model space may lead to predictions that are biased, qualitatively deficient, or quantitatively erroneous. Moreover, each step of model development is based on simplifying assumptions and affected by uncertainties. This includes not only the conceptual model and formulation of physical processes, but also the derivation of model-specific, process-related, and scale-dependent parameters, such as geometric, mechanical, and hydraulic properties of single fractures and fracture networks.

It is necessary to test the ability of the extrapolated model and each of its submodels) to make reasonable predictions for the intended purpose of the model, a process referred to a "pragmatic model validation".

The SKB Task Force has historically examined experimental data and modelling approaches that study flow and transport in fractured porous media. Task 10 continues this line of investigation. However, in addition to improving process understanding and numerical modelling capabilities, Task 10 focuses on examining whether a given model is suitable to address the specific purpose for which it has been developed. Task 10 consists of multiple subtasks of increasing scale and complexity regarding both the geometry of the system and the physical processes to be considered, as illustrated in Figure 1-1.

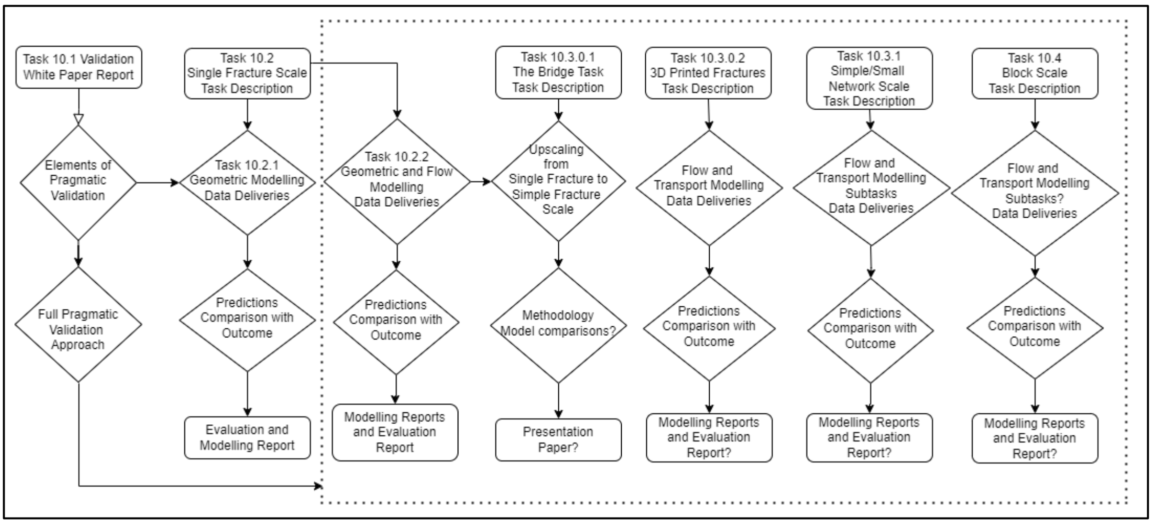

***Figure 1-1.*** *Overview of Task 10 subtasks.*

In Task 10.1, the fundamental concept of the proposed pragmatic model validation approach and its applicability to the modelling studies of Task 10 was developed (Lanyon et al. 2021, 2024; Finsterle and Lanyon 2022).

Task 10.2 is concerned with the characterisation of a rough fracture, and how fundamental geometric information about the fracture's surface roughness measured on a relatively small sample is used as the basis for deriving flow and transport properties of a single fracture on a larger scale and under different stress conditions. The task is also seen as an opportunity to develop and test pre-processing and modelling capabilities. In Task 10.2.1, the pragmatic validation approach is mainly exercised on the processing of geometric data rather than fluid flow and transport through a single facture (Finsterle et al. 2024).

The methods established in Task 10.2.1 are then applied in Task 10.2.2 to develop flow models on a larger scale and apply the pragmatic model validation approach by comparing blindly predicted flow rates in a single fracture to corresponding measured data that are obtained by a dedicated laboratory experiment.

In Task 10.3 and its various subtasks, the scale of the problem will be further increased from a single, small fracture to a small network of discrete, deterministically known fractures. Following a bridge task that ensures that the lessons learnt from Tasks 10.2.1 and 10.2.2 are transferred to Task 10.3, a 3D printed physical model containing a single, artificial fracture with variable aperture is used to provide flow data for the prediction-outcome exercise. In later subtasks, the model complexity increases with respect to both geometry and physical processes. First, the focus is on a single discrete structure, i.e., Structure #20 intersecting the TASS tunnel in Äspö HRL (Hardenby and Sigurdsson 2010), and is intersected by a few other structures or fractures, with the goal to validate the corresponding model by comparing its predictions to measured total flow rate and conservative tracer data. This exercise will be followed by the modelling of a small network of several fractures that form compartments. Data from the TRUE Block Scale experiment (Andersson et al. 2002a, 2002b, 2007) will be used for the prediction-outcome comparison.

It should be noted that while the scale and complexity increase with each subtask, the degree of abstraction, i.e., the simplifying assumptions that need to be made for implementation in a predictive model, tend to increase as well, highlighting the importance of an objective-driven pragmatic validation approach and adapting the acceptance criteria applied at each intermediate step. It also requires that uncertainties identified in one subtask be qualitatively acknowledged and/or quantitatively propagated to the model of the next subtask.

This Evaluation Report is only concerned with Task 10.2.2. Most of the data-analysis and modelling methods used in Task 10.2.2 have been previously developed—and partially validated (Finsterle et al. 2024)—in Task 10.2.1, which considers the characterisation of a single fracture based on high-resolution measurements of the fracture's surface topography. Surface roughness and the related aperture distributions and transmissivity fields are expected to be fundamental characteristics that impact flow rates and channelling effects in a single fracture. For the assessment of a radioactive waste repository, these property fields must be upscaled multiple times, from the scale of measurements to the scale of the fracture sample to a larger fracture to a fracture network and potentially to the site scale of the repository. Moreover, on each of these scales, numerical models may be used that employ spatial resolutions different from the resolution of the underlying measured, calculated, and upscaled property data. This clearly indicates that each of the conceptual decisions and each of the (sub)models that are built as part of this process requires model validation with a level of detail and strictness that depends on how the results of a particular (sub)model will be used in subsequent steps, be that the direct calculation of a measure of interest, or the generation of input parameters for the next model in the sequence.

Unlike in the preceding subtask, which focused on preprocessing and interpretation of geometrical information, Task 10.2.2 includes dynamic simulations of fluid flow under different normal loads. Results from such simulations are then compared to data from various hydro-mechanical experiments performed on a rock sample containing a natural fracture. Multiple flow tests were conducted on the unopened and re-assembled fracture in two perpendicular directions and under various stress conditions (see Section 3.1 for a summary of the experiments). These model-to-data comparisons constitute prediction-outcome (P/O) exercises, which are a key component of pragmatic model validation.

The Modelling Groups (MGs) participating in Task 10.2.2 are summarized in Table 1-1, along with the reference to their final Modelling Reports.

**Table 1-1 Modelling Groups participating in Task 10.2.2.**

| Modelling Group | Participating Organizations | Modelling Report Reference, Remarks |
|---|---|---|
| A21-KTH | Amphos21 Consulting S.L. and<br>KTH Royal Institute of Technology | R-23-13 (Trinchero et al. 2024) |
| TUL | Technical University of Liberec | R-23-14 (Hokr and Balvín 2026) |
| VTT | Technical Research Centre of Finland Ltd. | R-23-15 (Naumer and Pulkkanen 2025) |
| NTU-Taipower | National Taiwan University and<br>Taiwan Power Company | R-25-02 (Wang et al. 2025) |
| SU | Stockholm University | Reviewer; no final report |

While working from the same Task Description and same data set, four MGs developed different models based on different methods, focused on different aspects of the system to be modelled and predicted, and addressed different validation issues. Their work was presented at various meetings and workshops organized by the Task Force Secretariat. As a result of these discussions, some aspects of the Task Description were revised, data were corrected or supplemented, and schedules were adjusted. The draft Modelling Reports underwent a formal peer review process, which is a key component of pragmatic model validation. Furthermore, after comment resolution of the peer review, the final drafts were subjected to SKB's factual review[1] before publication of the MGs' work as SKB reports.

---

[1] The SKB factual reviews of the four final Modelling Reports were conducted by S. Finsterle, who is also the Task 10 evaluator and author of this report. It should be noted that some of the conclusions found in this Evaluation Report are based on the evaluator's observations made during the SKB factual review, i.e., they are based on the review copies of the Modelling Reports, not the final, published versions, in which many of the review comments were addressed and suggestions incorporated.

## 1.2 Objectives of Task 10.2.2

The Task Description (Bruines, unpublished[2]) describes the main objectives of Task 10.2.2 as follows:

- Prediction and validation of the upscaled fracture geometry from borehole sized fracture geometry and/or fracture trace geometry.
- Prediction and validation of flow along a fracture at different normal stresses.
- Support the development and demonstration of pragmatic validation workflow at the single fracture scale.

Background information provided in the Task Description is supplemented by data packages that were released at specific times. In particular, additional experimental data were released *after* the predictive modelling was completed to facilitate a prediction-outcome (P/O) exercise in support of pragmatic model validation. The overall objectives of Task 10.2.2 were interpreted and addressed independently[3] by each MG and documented in individual Modelling Reports.

## 1.3 Scope of evaluation report

The work of the participating MGs has been presented at several Task Force meetings and workshops, where informal feedback was provided to the researchers by their peers, the Task Force secretariat, and the evaluator. As part of the pragmatic model validation workflow, a peer review was conducted, where the draft Modelling Reports of each participating MG was reviewed by one of the other MGs (see Section 3.3). These formal reviews, which include a comment resolution cycle and final approval, are fully documented in standardized Review Records. After completion of the peer review, the evaluator performed and documented the final SKB factual review before publication of the modelling work as SKB reports in the R series (“reports detailing specific scientific or engineering studies”); see Table 4-1.

This Evaluation Report provides some comments on the overall lessons learnt from Task 10.2.2. Rather than evaluating the individual Modelling Reports (which was done and fully documented as part of the peer review and SKB factual review), this report makes general observations about how the models were developed, interpreted, and validated. The evaluation is guided by the overarching goals of Task 10, the specific objectives of Task 10.2.2, and the purpose of the prediction models as identified by the MGs.

## 1.4 Structure of evaluation report

Section 3 briefly describes the relevant experimental data used for the Task 10.2.2 P/O exercise, followed by very short summaries of each MG’s modelling work and the outcome of the peer review. Section 4 discusses and evaluates several aspects of this exercise with the goal to make recommendations about both the development and validation of complex models.

---

[2] Bruines, P. SKB Task Force on Modelling of Groundwater Flow and Transport of Solutes: Description of Task 10.2. SKB R-23-10, Svensk Kärnbränslehantering AB, Stockholm, Sweden. Unpublished.

[3] The common data base (Task Description and Data Release Packages) and information exchange during Task Force meetings, workshops, and the peer review, resulted in a set of models and interpretations that are not strictly independent, but the result of a process with reciprocal influences. See related comments in the conclusion section.

# 2 Approach

## 2.1 Outline of evaluation

Task 10 is concerned with two main topics, namely (1) model validation in the context of (2) groundwater flow and transport in geologic media with discrete features. Most of the previous tasks of the SKB Task Force were also aiming to advance the technologies related to the prediction of flow and transport processes occurring in fractured geologic media that are considered as potential host formations for the disposal of radioactive waste, with focus on numerical simulations and the analysis of related laboratory and field data.

The various subtasks of Task 10 deal with or focus on specific aspects of fracture hydrology. The complexity of the subtasks increases as the number of discrete features, their scale, and the processes to be included in the modelling is increased. This progression of subtasks is designed such that the result of a preceding study provides the basis for the next analysis.

Interconnected studies of increasing scale and complexity provide a suitable framework for exercising the proposed pragmatic model validation approach, as confidence in each model's ability to support the next model must be established before a specific modelling methodology is applied to the next level of higher complexity. Conceptual and parametric uncertainties must be propagated along the chain of submodels to be able to properly evaluate the quality of the subsequent model and—especially—the final model, which supports the ultimate project goal (e.g., performance assessment, site evaluation, barrier function analysis, etc.).

As is evident from the Modelling Reports (see summaries in Section 3.2), there is a very high level of scientific and technical knowledge required along with sophisticated modelling expertise to perform the data analyses and predictive simulations needed to address the specific objectives of Task 10.2.2 (as well as the other subtasks of Task 10).[4] In addition to tackling these modelling challenges, the Modelling Groups had to work through the steps of pragmatic model validation, which permeates all subtasks and provides the links between them.

Specific technical issues related to predictive modelling of flow through a single fracture under different mechanical loads have been discussed in the various Task Force meetings and workshops as well as the peer review and SKB Factual Review (see Sections 3.3 and 3.4 for summaries). Many of these issues have been addressed by the MGs—by resolving them, documenting limitations, or proposing future work—as reflected in their final Modelling Reports.

It is noteworthy (as will be discussed below) that many of the most challenging and persistent issues came to light as a result of the validation process, notably the difficulty to obtain a flow-relevant mean aperture despite accurate high-resolution fracture-surface data, which was revealed by the systematic overprediction of flow rates, and required additional post-validation adjustments and calibrations of the models. To examine and highlight the role of validation for the development and use of complex simulation models, the evaluation that follows mainly comments on the way pragmatic model validation was performed rather than on the details of the numerical models used for the analyses.

---

[4] Considerable modelling capabilities has been developed—and related simulation experience has been gained—through the work performed in Tasks 1 to 9 of the SKB Task Force.

## 2.2 Summary of pragmatic validation approach

The principles of pragmatic model validation—while generally applicable—were documented specifically to support the main goal of Task 10, which is indicated in its title: “Validation approaches for groundwater flow and transport modelling with discrete features.” Both a general and task-specific discussion of pragmatic model validation were documented in the White Paper Report of Task 10.1 (Lanyon et al. 2024) and summarized in (Finsterle and Lanyon 2022). The principals and workflow of pragmatic model validation are briefly recapitulated here as the basis for evaluating the MGs’ adherence to the approach.

Pragmatic model validation is a practical method used to build confidence in computer models when it is fundamentally impossible to prove that the model is a perfect representation of the real world. In fields like geosciences within the context of radioactive waste disposal—where researchers model complex natural systems over multiple scales and extended time periods—models can never be 100% validated, because natural systems are open and only characterized at a few points within the fractured formation. Pragmatic validation shifts the focus from seeking absolute truth to ensuring a model is suitable for its intended use, i.e., ”fit for purpose”. This guiding principle of pragmatic validation posits that a model doesn't have to be perfect; it just needs to be good enough to help a decision-maker solve a specific problem. This acknowledges both the high stakes associated with the assessment of the safety of a radioactive waste repository, but also the fact that decisions about repository safety can be made using residual uncertainty in the results of the model predictions.

Pragmatic validation follows a structured workflow to ensure the model-building process is transparent and scientifically sound:

1. *Define the modelling purpose*: Identify and clearly state what the model is supposed to do in support of what specific objective.
2. *Identify critical aspects*: Determine which parts of the model have the biggest impact on the simulation results as they pertain to the modelling objective.
3. *Select performance measures*: Select relevant model-output variables; they must be suitable for assessing the predictive capabilities of the model for its intended use. These model-output variables must correspond to quantities that can be independently measured or estimated. Formulate acceptance criteria that reflect the acceptable prediction uncertainty.
4. *Sensitivity and uncertainty analyses*: Identify the factors that have the greatest impact on the relevant model predictions and at the same time are most uncertain; determine their impact on the predictions as their uncertainty is propagated through the model.
5. *Prediction-outcome (P/O) exercise*: Check the ability of the model to make predictions under extrapolated conditions by simulating the system behaviour, extracting the performance measures from the results, and then comparing them to observations made on the real system or independent estimates. The quality of the comparison is judged based on the *a priori* defined validation acceptance criteria.
6. *Model audit*: Fully document the development of the model (specifically underlying assumptions, conditioning and calibration of input parameters, and the details of the P/O exercise) and seek independent and sceptical review of the entire process.

The overall structure of Task 10 was given in the Task Descriptions, including the main project goals, the scope of the various subtasks, a description of data for initial model development, and measurements to be used in the P/O exercise for ultimate model validation. Within the framework provided by the Task Description, the MGs had the leeway and responsibility to formulate the specific purpose of their respective models, select the performance measures and associated validation acceptance criteria, and develop a suitable modelling approach. A peer review was organized by the Task Force Secretariat to facilitate a key element of the model auditing.

The SKB Task Force also offers a unique opportunity to strengthen model validation by the fact that multiple teams build models based on different assumptions and using different simulation tools. If these different models all reach consistent conclusions about the system of interest, higher confidence in the collective modelling results can be gained. However, fully exploiting this opportunity would require a comparative analysis of all developed models, an assessment of the degree of validation of each individual model, and an integration and/or cross-comparison of the results. While related comments will be made throughout this Evaluation Report, no attempt has been made to reach an overarching conclusion about the validity of current modelling of flow and transport through formations with discrete features.

## 2.3 Expectations and evaluation criteria

The MGs' model validation work is evaluated based on expectations that are reflected in the following questions:

- *Pragmatic model validation workflow*: Are the key elements of the pragmatic model validation workflow (see numbered list in Section 2.2) properly exercised? In particular:
    - *Modelling purpose*: Is the purpose of the model clearly defined, allowing for a case-specific formulation of validation acceptance criteria that evaluate whether the model is "fit-for-purpose"?
    - *Prediction-outcome exercise*: Are relevant performance measures chosen that are suitable to evaluate the model's predictive capabilities needed to address its intended purpose? Are appropriate validation acceptance criteria formulated prior to the prediction-outcome analysis?
    - *Documentation*: Are the model and its validation documented with sufficient detail and transparency, so they can be understood and evaluated by an independent reviewer?
    - *Validation status*: Is a conclusion regarding the model's validation status reached?
    - *Peer review*: Are the comments of the peer review properly addressed?
- *Integration into model sequence*: Are the conclusions from the pragmatic validation of the upstream model developed as part of Task 10.2.1 (if available) properly transferred to the model of Task 10.2.2, and are the conclusions from the pragmatic validation of the current model formulated such that they can be transferred to the downstream models?
- *Uncertainty propagation*: Are conceptual and parametric uncertainties quantified and propagated through the sequence of models?
- *Failed validation*: If the model fails to meet the validation criteria, are the main reasons identified and/or alternative conceptual models proposed?
- *Application domain*: Is the application domain of the validated model for downstream analyses clearly described?

Note that achieving a successful validation of the model is *not* considered a criterion for evaluating the MGs' work, as the purpose of Task 10 is to exercise and document the pragmatic validation approach, independent of its outcome. Furthermore, while advanced simulation models are needed for the prediction-outcome exercise, the degree of sophistication of these models is not the main criterion for evaluating the quality of the work. This is because the pragmatic model validation approach aims for a specific modelling purpose and includes prediction uncertainty and appropriate delineation of the application domain. It is entirely possible that a relatively parsimonious model developed using few data performs better than a highly sophisticated model developed based on many high-resolution data, provided that the data used for conditioning and calibration are relevant and tailored to the model's purpose, application scale, and dominant processes, and that they are consistent with the specific conceptualisation and structure of the model. Conversely, the degree of model complexity must match the nature and complexity of the available data used for model development and the P/O exercise. Therefore, the consistency between the data, the model, and its intended use forms the basis for evaluating the MGs' modelling and validation work.

# 3 Summary of data, modelling work, and peer review

## 3.1 Experimental data

Task 10.2.2 centres on investigating and modelling groundwater flow through a single natural fracture under different normal loads. This subtask utilizes data from coupled hydro-mechanical experiments for the prediction and validation of flow rates and upscaled fracture geometries.

The experimental work utilized a 20 × 20 × 25 $cm^3$ rock samplecontaining a single, "simple" natural fracture. The rock originated from the Flivik quarry in Sweden and is believed to have a similar geological history to the samples used in Task 10.2.1. The fracture is relatively clean, displaying limited mineral precipitation (believed to be calcite) and following grain boundaries.

The testing program was divided into several phases to capture the fracture's behaviour in both its unopened and reassembled states:

- *Phase 0 (preparation)*: The sample was cut to dimensions while keeping the fracture in its mated, unopened position. High-resolution photography of the fracture traces and high-accuracy measurements of the closed sample's outer dimensions were taken using laser trackers and reflectors.
- *Phase 1 (testing the unopened fracture)*: Hydro-mechanical flow tests were conducted in two orthogonal directions (Face 2 to Face 4 and Face 1 to Face 3). Measurements were taken at various normal loading stages, typically ranging from 0 to 8 MPa. This phase also included a staining tracer test using Rhodamine-B at low normal stress (0.09 MPa) to identify flow-wetted surfaces. Afterward, the sample was opened to describe the surfaces, perform 3D scanning, and conduct contact pressure film measurements at 1 and 4 MPa.
- *Phase 2 (reassembled sample)*: The sample was reassembled, and hydro-mechanical flow testing was repeated in one direction for three cycles to observe stability after settlement.
- *Phase 3 (high normal load)*: This phase involved testing at significantly higher normal loads of 10, 20, 30, and 40 MPa, followed by reopening the sample to record any fracture damage or other changes.

The basic principle of the hydro-mechanical test involved attaching in- and outflow boxes to opposite sides of the fracture and sealing the perpendicular sides to prevent leakage. The sample was placed in a uniaxial test apparatus to apply normal force. Key instrumentation included:

- *Miniature LVDTs*: Four LVDT (Linear Variable Differential Transformer) sensors measured displacement (deformation) over the fracture.
- *Hydraulic system*: Inflow pressure was generated by a water bucket at a specific height, while outflow was measured by a high-precision load cell recording the mass of collected water.
- *Sensors*: Logged data included time, inlet/outlet pressures, normal load, water/rock temperatures, and loading forces.

The data delivery for Task 10.2.2 includes the following:

- *Geometric data*: Scanned 3D topography of upper and lower fracture surfaces, fracture trace photographs, and outer surface measurements before and after testing.
- *Hydro-mechanical test results*: A detailed log of boundary conditions, including applied normal forces, hydraulic pressures, water/rock temperatures, and displacement measurements (LVDT data).
- *Contact pressure data*: BMP and FPD files from Fujifilm pressure-sensitive film (low and medium pressure types) taken at 1 and 4 MPa loads.
- *Intact rock properties*: Measurements of uniaxial compressive strength, Young's modulus, Poisson's ratio, and tensile strength.

Note that some data, specifically the observed flow rates, are initially withheld from the modelling groups to allow for the blind prediction-outcome validation exercise. More details about the experiments can be found in the Task Description.

## 3.2 Summary of modelling work

### 3.2.1 Modelling work by A21-KTH

This report details a joint study by Amphos$^{21}$ and KTH regarding the hydromechanical modelling of groundwater flow through the rock fracture. A prediction-outcome exercise was performed to validate how well the numerical model could forecast flow rates and channelling factors compared to experimental data.

A21-KTH utilized an uncoupled approach, first applying an elastic-plastic contact model to determine how fracture apertures change under normal loading, followed by flow simulations using the Reynolds equation.

To determine the fracture aperture from surface data, A21-KTH used a multi-step process involving geometric alignment and subsequent calibration to account for measurement uncertainties:

- *Alignment and interpolation*: Because the two fracture surfaces were scanned separately, A21-KTH first aligned them within a common reference system. Once aligned, the topography data were interpolated onto a regular grid to facilitate numerical analysis. For this study, a grid of 398 × 399 cells was used with a resolution of 0.5 mm ($\Delta x = \Delta y = 5\times10^{-4}$ m).
- *Initial aperture*: The initial aperture at each point of the grid was calculated simply as the difference between the height of the upper surface and the height of the lower surface.
- *Calibration of aperture shift*: Presuming that the raw geometric data are biased due to material loss during the fracture's opening and errors in the realignment process, all local aperture values were reduced by a constant value of $1.2\times10^{-4}$ m to ensure the model's average displacement matched the physical measurements from the LVDT sensors. During the calibration shift, the aperture in grid cells where the calculated difference became negative was set to zero. These locations were then defined as physical contact points where the two rock surfaces were touching.

Aperture changes under loading were calculated using an uncoupled modelling approach, where mechanical changes in the fracture were computed independently before being used in fluid flow simulations. The calculation of these changes relied on an elastic-plastic contact model based on the boundary element method:

- *Deformation model*: The model assumes that rock asperities behave as an elastic-perfectly plastic material. Plastic deformation is confined to local, small areas where the local normal stress exceeds the indentation hardness of the rock matrix. Once this limit is reached, the asperity undergoes permanent deformation without a further increase in stress. The relationship between normal displacement and normal stress is defined by the Boussinesq equation (Boussinesq, 1904) for a normally distributed load on a semi-infinite body.

- *Numerical implementation*: The model seeks to find the contact state that minimizes the total complementary potential energy of the two contacting rough surfaces. For the elastic-plastic analysis, the energy dissipating by plastic deformation is added to this total potential energy.
- *Discretization*: The fracture domain is discretized into a structured mesh (398 × 399 cells), which is consistent with the high-resolution digital topography of the scanned surfaces.
- *Solution approach*: The governing equations are formulated as a linear complementarity problem to simultaneously determine the local stress and the resulting contact area. The model calculates changes through an incremental loading process. A constrained conjugate gradient algorithm is used to iteratively solve for the contact pressure until the summation of local stress across the grid equals the applied total load. For each incremental step, an additional iteration is performed to determine the specific regime and magnitude of plastic deformation. As the load increases, the model identifies new contact points where the aperture becomes zero and updates the remaining aperture distribution accordingly. Specific aperture distributions for the required loading stages of 0, 1, and 4 MPa, were generated.

Fluid flow was calculated using two distinct methodologies: a primary depth-averaged approach based on the local cubic law and an alternative approach using the 3D Navier-Stokes equations:

- *Depth-averaged approach*: The main modelling work treated the fracture as a 2D domain where local aperture values were converted into local transmissivity. This model assumes laminar flow and uses the Reynolds equation with the local cubic law. Groundwater flow simulations were performed using the finite-difference code MODFLOW-2005. Because the local cubic law often overestimates flow, A21-KTH added an empirical correction factor.
- *3D Navier-Stokes approach*: A more rigorous method was used to explicitly resolve the 3D fracture void space. This approach utilized the finite-volume numerical code OpenFOAM. The initial fracture aperture was represented by an orthohedral mesh containing more than 50 million elements, with higher refinement near contact points. The model simultaneously solved the mass balance (assuming incompressibility) and momentum balance equations. A no-slip condition was applied to the solid rock walls. To analyse flow patterns and velocity distributions, particle tracking simulations were conducted using the numerical code RW3D (Fernandez-Garcia et al., 2005). This enabled the calculation of channelling factors.

The uncoupled modelling approach, while theoretically sound for sequenced analysis, reveals significant discrepancies when compared directly to experimental data, primarily showing a substantial overestimation of flow rates when validating the relationship between fracture aperture and hydraulic conductivity. The comparison between this modelling approach and experimental results generated several key findings:

- *Significant overestimation by the local cubic law*: Validating the depth-averaged approach against experimental data showed that the local cubic law model overestimated experimental flow rates by more than three orders of magnitude. This discrepancy is attributed to the local cubic law's inherent limitations in describing flow through natural, rough-walled fractures, where it typically overestimates hydraulic properties even at low flow rates.
- *Impact of empirical correction factor*: To address the local cubic law’s bias, A21-KTH applied empirical correction factors ($f$) to the hydraulic aperture. A correction factor of $f = 0.1$ (implying that the hydraulic aperture is one-tenth of the mechanical aperture) provided a better representation of initial flow rates but failed to accurately reproduce the decreasing trend of flow as normal load increased. Follow-up analyses suggested a factor of $f = 0.2$ could better capture the experimental trend across the entire loading cycle, provided the underlying mechanical model is more accurately calibrated.
- *Comparison with 3D Navier-Stokes modelling*: The modelling team also utilized an alternative approach by solving the 3D Navier-Stokes equations, which explicitly resolves the fracture aperture. While this approach provided lower, more realistic flow rates than the local cubic law, it still overpredicted experimental results by approximately two orders of magnitude. However, when the initial aperture was calibrated using a Computational Fluid Dynamics (CFD) model for unloaded conditions, the results were in much closer agreement with experimental values.

- *Potential sources of model-to-data discrepancies*: Three main reasons were given to explain why the uncoupled model provided only a poor prediction of the experimental outcome: (1) *Initial aperture bias*: Geometrical information from fracture scans is often highly biased due to material loss during fracture opening and errors in realigning the two surfaces. Because the relationship between aperture and flow is non-linear, even small uncertainties in aperture lead to large uncertainties in predicted flow rates; (2) *Mechanical decoupling*: The uncoupled approach assumes hydraulic processes do not affect mechanical closure. In the experiment, high water pressure at the fracture inlet (especially in unloaded conditions) may cause initial displacement or rotation of the fracture, which the uncoupled model fails to capture; and (3) *Asperity damage*: The mechanical model only accounts for localized damage and does not consider the impact of gauge particles produced by damaged asperities, which can influence hydraulic conductivity.

In summary, the uncoupled approach highlights that direct mapping of geometric aperture to transmissivity is inadequate for predicting hydraulic conductivity. Successful validation requires constraining the model's initial conditions with additional data, such as actual flow rates or tracer tests, rather than relying solely on geometric scan data. A21-KTH conclude that initial geometric uncertainties and the complex relationship between mechanical and hydraulic apertures remain critical challenges in fracture modelling.

The modelling work of A21-KTH is documented in detail in Trinchero et al. (2024).

### 3.2.2 Modelling work by TUL

The modelling report of the Technical University Liberec details their study on fracture surface upscaling and hydro-mechanical simulation of the granitic block fracture. The study focuses on pragmatic validation by performing two main exercises: upscaling rough fracture surface geometry and predicting the hydro-mechanical behaviour:

- *Fracture surface roughness*: The first part of the study involved characterizing the roughness of the natural fracture surface measured by laser scanning. Square subareas (ranging from 30 mm to 39.5 mm), representing the borehole scale, were selected to determine (a) direction and composite surface roughness metrics (RMS), and (b) fractal parameters, specifically the fractal dimension and the proportionality factor.
- *Upscaling methodology*: The upscaling process utilized the fractal model, which assumes that certain roughness characteristics remain consistent or predictable across different spatial scales. The Hurst exponent was treated as scale-independent and taken directly from the analysis of the smaller borehole-scale subareas. To find the RMS for the target scale (200 mm), TUL substituted the larger scale value ($w = 200$ mm) into the power regression formula established at the smaller scale.
- *Synthetic fracture surface generation*: Using these upscaled parameters, TUL generated 10 synthetic realizations of the 200 × 200 mm surface.
- *Results and validation of upscaling*: The upscaled synthetic surfaces were compared against actual measurements of the full 200 mm fracture surface to test the model's accuracy. The directional roughness metrics of the synthetic surfaces matched the measured large-scale surfaces well. However, there were significant differences between the fractal parameters of the synthetic surfaces and the real measured surfaces. The authors attribute this to extrapolation uncertainty. Because the fractal model assumes a relationship over an infinite scale range, the power regression is highly sensitive to noise or inaccuracies. Therefore, while the regression lines for the different scales were close in the middle range, they diverged at the larger scale.

The second exercise focused on predicting how normal stress affects the flow of water through the fracture.

- *Mechanical model*: TUL developed a one-way coupled model where the mechanical deformation determines the aperture field for the flow simulation. They used a simplified "independent column" approach, representing the fracture zone as a set of elastic columns that contact and compress based on the reference aperture field and applied displacement. This model was calibrated using measured stress-displacement curves from the initial experiment stages.
- *Hydraulic model*: Flow was calculated in a 2D horizontal projection using the local cubic law. The model used stress-dependent aperture fields to determine how flow rate changes under increasing normal loads.
- *Mechanical model results and validation*: The model was validated against the stress-displacement curve of the resealed block. It accurately captured the nonlinear, concave shape of the curve, except at very low loads (below 80 kN) where surface settling effects dominated. For very high normal stresses (up to 1500 kN), the flow rate became highly sensitive to the minimum aperture threshold used in the numerical simulation to prevent fully impermeable elements.
- *Flow model results and validation*: The blind model predictions were compared against experimental data. The predicted flow rates were approximately one order of magnitude larger than those measured in the experiment. However, the model successfully captured the shape and trend of the stress-flow dependence over several orders of magnitude.

The authors conclude that the approach can reasonably well capture the generic hydro-mechanical behaviour of a single fracture despite the model's significant simplifications, such as neglecting hysteresis, block rotation, and temperature effects, as well as irreversible surface changes due to disassembling and reassembling the block, using a simplified column-based stress model, and the sensitivity to the minimum aperture threshold used to prevent numerical errors, rather than the physical properties of the fracture.

### 3.2.3 Modelling work by VTT

VTT's modelling report details the prediction-outcome exercise for their model, which simulates groundwater flow through the rock fractures under varying mechanical loading. The COMSOL Multiphysics model coupled contact mechanics with fluid dynamics to estimate water flow through the resulting aperture field.

Compression and deformation were simulated using a numerical contact modelling approach, assuming the rock behaves as a linear elastic material. The specific methodology involved the following key components:

- *General modelling strategy*: The fracture surfaces were virtually compressed against each other to calculate the deformation of the rock and the resulting change in the fracture void space.
- *Creation of parametric surfaces*: The process began with the high-precision laser scan data of the open fracture surfaces. These point clouds were converted into parametric surfaces (cubic spline interpolants) to create a continuous geometrical representation of the top and bottom fracture walls.
- *Trimming overlaps*: When placed in their measured positions, the surfaces initially showed geometric overlaps (clipping). These overlaps were trimmed prior to simulation to ensure the surfaces could be placed in contact without interpenetration.
- *Linear elasticity*: The rock was modelled as isotropic and linear elastic. The model did not simulate permanent deformation (plasticity) or rock breaking, even in areas where the calculated stress exceeded the rock's yield limit. However, the resulting stress distributions were compared to the rock's uniaxial compressive strength to estimate areas where permanent deformation might occur.
- *Force balance*: The model solved force balance equations where the deformation is driven by normal compression. The effect of water pressure on the rock deformation was neglected.

- *Contact formulation*: The augmented Lagrangian contact method was used to manage the interaction between the two rough fracture surfaces. The contact pressure between the domains was calculated based on the gap distance and a penalty factor derived from the material's Young's modulus and element size. To prevent the geometric domains of the two rock parts from overlapping (penetrating each other), a contact offset was defined, maintaining a minimum gap between the domains. The simulation aimed for an offset of 0.035 mm (the measurement precision), but convergence issues often required larger offsets (0.05 mm or 0.1 mm). This offset effectively added to the calculated aperture, resulting in a larger void space than likely existed in the physical experiment.

VTT developed two distinct methods to achieve convergence in the simulation:

- *Nearing Contact Model*: The two rock parts are initially separated by a small gap. The top part is compressed towards the bottom part. To stabilize the simulation (as there is initially no counterforce), a spring foundation function is applied to the top part, which acts as a vertical spring constraint. The spring constant is gradually decreased to zero in discrete steps. As the spring relaxes, the top domain moves closer to the bottom domain until true contact is established under the normal load.
- *Departing Contact Model*: The top rock part is placed such that it artificially overlaps with the bottom part. The contact formulation forces the top part to move upwards (depart) until it reaches an equilibrium position balanced by the applied normal load and the contact pressure. This approach required fewer iterations and calculated much faster than the Nearing Contact Model.

The simulations applied various normal loads to the top surface of the upper rock part, specifically 0 MPa (unloaded), 1, 2, 4, 6, and 8 MPa. The output of these mechanical simulations was a deformed geometry of the void space. The aperture field was defined as the distance between the deformed top and bottom fracture surfaces at any given point.

Fluid flow through the aperture field was calculated using 3D numerical simulations based on the Stokes equations applied to the fracture geometry deformed by applying different normal loads. The calculation process involved the following specific methodologies and assumptions:

- *Governing equations*: The simulations solved the Navier-Stokes equations without the inertial term (i.e., the Stokes equation). Water was modelled as a Newtonian fluid with constant density (998.2 kg/m$^3$) and viscosity (0.001 Pa·s).
- *Geometry and meshing*: The geometry for the fluid simulation was directly derived from the deformed void space generated by the mechanical contact models (Nearing and Departing Contact Models). To allow the flow profile to fully develop before entering the actual fracture void space, the geometry at the inlet was extruded by 30 mm. The void space was discretized using a free tetrahedral mesh in the fracture void space, with a free triangular mesh on the fracture surfaces. The mesh density was high (e.g., minimum element sizes down to 0.025 mm in high-load scenarios) to capture the complex aperture variations and prevent meshing overlaps.
- *Boundary conditions*: A no-slip condition was applied to the top and bottom fracture surfaces and the two flow-parallel side walls, meaning water velocity was zero at the rock walls. Average pressure values from the experiment were applied to the inlet and outlet, creating the pressure gradient that drove the flow.

The flow velocity profile over the area of the outlet boundary was integrated to obtain the total volumetric flow rate.

The model was validated using the pragmatic validation approach. The primary validation method was comparing the model's predicted flow rates under different normal loads against experimental measurements that were initially withheld from the modelling team. The simulations successfully captured the general trend of rock displacement. However, while the experimental flow rates were extremely low (on the order of 0.01 ml/s under 1 MPa load), the simulated flow rates were orders of magnitude higher (ranging from approximately 50 to 100 ml/s). The large discrepancy between the simulated and measured flow rates was primarily explained by the overestimation of the fracture aperture in the numerical models. This overestimation of the void space and flow is attributed to several specific factors:

- *Contact offset*: To prevent geometric overlap between the rough fracture surfaces and to allow for successful meshing of the fluid domain, the modellers had to enforce a contact offset between the rock blocks, which artificially added space to the fracture aperture. However, analytical calculations indicated that a 0.1 mm offset would contribute approximately 3 ml/s to the flow. While significant compared to the experimental result, this factor alone accounts for only a small fraction of the total discrepancy.
- *Removal of initial overlaps (clipping)*: When the scanned surfaces were placed in their measured positions, they overlapped—likely because the rock had broken or deformed permanently during the physical opening of the fracture prior to scanning. These overlaps were removed by “clipping” the geometry, creating increased areas of perfect contact. This prevented realistic compression of the rock blocks, causing the void space elsewhere in the fracture to remain more open than it was in the physical experiment.
- *Measurement precision*: The precision of the laser scans (0.035 mm) used to generate the model geometry was insufficient to capture the tightness of the actual fracture. The effective aperture calculated from the measured flow rates and the cubic law was between 0.009 mm and 0.025 mm. Because the physical aperture was smaller than the measurement precision, the model geometry could not accurately represent the flow channels.
- *Mechanical assumptions*: The model assumed the rock behaved as a linear elastic material. High stress concentrations at contact points likely caused the rock to break or yield, allowing the fracture surfaces to compress closer together. By ignoring this and assuming elasticity, the model predicted a stiffer response and a larger remaining aperture. Moreover, the models limited the relative movement of the rock blocks to the vertical ($z$) direction.
- *Lack of interlocking*: The simulations did not allow for horizontal ($x$ and $y$) adjustments during compression. The rock surfaces might have slid slightly to interlock better, which would have further reduced the aperture and flow.
- *Unknown factors*: The discrepancy was so large that the authors noted it might suggest an “unknown, highly influential factor” missing from the modelling, such as unidentified issues with the experimental setup (e.g., air bubbles blocking flow, though protocols were in place to prevent this).

As a complementary approach to validate the mechanical and geometrical aspects of the model, VTT compared the simulated displacement of the rock blocks with data from Linear Variable Displacement Transducers (LVDTs) used in the experiment. This validation showed mixed results. At low normal loads (unloaded and 1 MPa), the model predicted larger displacements than observed, likely due to the model having fewer initial contact points than the real fracture. However, at higher loads (2–8 MPa), the trend of the simulated displacement matched the experimental data reasonably well, giving some confidence in the elastic model.

The modelling team also attempted to validate the stress distribution and contact geometry. The simulated von Mises stress distributions on the fracture surfaces were compared with experimental pressure film test results provided in the Task Description. While the highest stress areas in the model corresponded to no-contact areas in the film, the model showed fewer contact points than the pressure film. The authors deduced this was likely because the surface scan data included rock chips that had fallen off before the pressure film test was conducted, causing a mismatch between the model geometry and the physical state of the rock during testing.

To better understand the poor predictions of flow rates (which essentially invalidates the model), they used the cubic law to back-calculate the effective aperture required to produce the measured flow rates. This calculation revealed that the apertures required to match the experiment (approx. 0.009–0.025 mm) were smaller than (a) the aperture estimates from their geometrical and mechanical modelling, and (b) the measurement precision of the fracture scans (0.035 mm). This confirmed the hypothesis that the main reason of the inability to predict flow rates lies in the determination of an appropriate aperture distribution.

Finally, the authors proposed and tested a quantitative metric they called the "validity rank". The metric is defined as $R = n \times m \times (1 + p)$, where $m$ is the number of phenomena modelled, $n$ is the number of passed single-phenomenon validation exercises, and $p$ is the number of passed coupled validation exercises. The team aimed for a rank of 18 (based on validating flow, mechanics, and coupled behaviour). However, because the flow predictions were off by orders of magnitude (assigning a score of 0.05 instead of 1) and the mechanical validation was only partially successful (score of 0.5), the final calculated validity rank was only 1.7. This low score formally suggested the model's failure to meet the validation criteria.

The modelling work of VTT is documented in Naumer and Pulkkanen (2025).

### 3.2.4 Modelling work by NTU-Taipower

To predict groundwater flow through a single natural fracture under varying normal loads and to validate these predictions against experimental data, the NTU-Taipower modelling team developed a workflow that was divided into three distinct modules:

- *Geo-Module*: This involved the geometric alignment of high-resolution scans of the upper and lower fracture surfaces using the Iterative Closest Point (ICP) algorithm, which minimizes the distance between the two surfaces to find the best geometric match. Because the ICP might result in unrealistic overlaps, a vertical translation along the $z$-axis was applied to establish a physically plausible initial contact state. The Geo-Module also included the characterization of surface roughness using indices such as the Joint Roughness Coefficient (JRC), the root-mean-square (RMS) of the profile slope, $Z_2$, and a directional roughness metric.
- *Mech-Module*: A mechanical contact model was developed to simulate fracture deformation. It treats the contact zones between the upper and lower fracture surfaces as independent columns that deform elastically under load. The module is applied to the contact regions determined by the Geo-Module. The simulated displacements were compared to the LVDT data measured during the hydro-mechanical experiments, calibrating the Young's modulus (73 GPa) and initial state (i.e., the vertical translation); a vertical translation of 0.05 mm provided a stress-displacement curve that best matched the central trend of the experimental data. The calibrated module was used to generate aperture fields for different normal stress levels (ranging from 0 to 40 MPa), which are then passed to the Hydro-Module to calculate fluid flow.
- *Hydro-Module*: The calculated aperture fields were converted to transmissivity fields using the cubic law. A minimum (non-zero) transmissivity value is assigned to zones with very small apertures. This simulates the micro-pathways that exist in real fractures under high stress. A finite volume method was then employed to solve the Laplace equation for steady-state flow, predicting flow rates across the fracture. A channelling factor is also calculated based on the local flow rates and corresponding flow areas.

Both the Geo-Module and Mech-Module were not individually validated as standalone components; instead, they were calibrated by minimizing aperture as an indication of best alignment of the upper and lower fracture surfaces, and by estimating the Young's modulus and *z*-translation using measured displacement data. As the calibrated model fit the stress-displacement data well, the resulting aperture fields were considered a suitable basis for the subsequent flow simulations, which were then validated using the prediction-outcome exercise:

- *Flow prediction*: The simulation results demonstrated that the equivalent transmissivity of the fracture significantly decreases as normal stress increases from 0 to 40 MPa. The validation was considered successful because the simulation correctly predicted the significant declining trend of equivalent transmissivity as normal stress increases. Moreover, the experimental flow rates generally fell within the range of values predicted by the model, particularly when the uncertainty regarding the initial contact state was taken into account, confirming the model's ability to reproduce stress-dependent hydraulic behaviour.
- *Channelling effect*: The study found that as the fracture closes under higher normal loads, water flow becomes increasingly channelized in regions with larger apertures. However, at very high stress levels, the overall flow localization slightly decreased as the remaining open zones continued to close.
- *Uncertainty analysis*: The NTU-Taipower team also conducted uncertainty analyses by varying the initial contact states of the fracture surfaces. The resulting range of predicted transmissivities successfully bracketed the variability observed in the experimental data.

The report concludes that the hydro-mechanical model successfully achieved pragmatic validation. By calibrating the model with mechanical displacement data and independently validating it against flow rate measurements, the study demonstrated that a model based on high-resolution geometric data can reliably predict hydraulic behaviour in fractured rock under in-situ stress conditions.

The modelling work of NTU-Taipower is documented in Wang et al. (2025).

## 3.3 Summary of peer review

Peer review is an essential part of pragmatic model validation. A vital component of the model building process is the examination of the whole process in the sense defined by the IAEA Safety Glossary (IAEA 2018): "*'Model validation' in these circumstances implies showing that there is a basis for confidence in the model(s) by means of detailed external reviews…"*. IAEA (2018) also specifies that, in the case of geological disposal facilities that involves temporal 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.

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. For that purpose, the model development and evaluation process need to be thoroughly documented and externally reviewed.

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 domain of application, and uncertainties are sufficiently understood and documented. Having the development of each model component submitted to independent review increases the credibility of a model, regardless of its ultimate use.

Saltelli et al. (2013) proposed a list of review criteria and required that the review process and its outcome, specifically regarding consensus or disagreement among the reviewers, be documented. For this purpose, a review plan was developed by the Task Force Secretariat, detailing the objectives of the review, general and specific review criteria, and the steps of a formal comment resolution process. In addition, a template for documenting the comments and responses was distributed to the participating MGs.

To exercise this important component of pragmatic model validation, each of the draft Modelling Reports was subjected to a formalized peer review, where one of the other participating MGs served as the reviewer. While the participation of the reviewers in the Task Force may make them not fully "independent" or "external", they certainly have the technical qualification and interest to assess their peer's work. As an exercise in pragmatic model validation, it is also considered valuable for the MGs to participate on both sides of the formal peer review process.

### 3.3.1 Peer review of A21-KTH by VTT

The peer review of A21-KTH's Modelling Report (summarized in Section 3.2.1) was conducted by VTT. The following is a summary of the review record, which includes the initial comments of VTT, the responses by A21-KTH and potential dispute resolution and acceptance.

The reviewers concluded that the report only partly met the review criteria. While acknowledging the work as a "dauntingly complex task", they criticized the report for lacking necessary details, failing to fully justify modelling choices, and insufficiently addressing the pragmatic model validation procedures required by the Task Description.

The review highlighted several technical disagreements and requests for clarification:

- *Mechanical model and hardness*: The reviewers questioned why the authors used indentation hardness as a yield stress limit rather than standard rock mechanics parameters (e.g., uniaxial compressive strength or shear stress limits). The reviewers expressed scepticism about whether rock asperities truly deform plastically or if they simply break. The authors argued that standard compressive strength data is size-dependent and cannot be directly applied to the small scale of asperities (approximately 0.2 mm). They maintained that the hardness approach is established in the contact model literature and that plastic behaviour is assumed to be confined to local contact points.
- *Yield condition formulation*: The reviewers identified mathematical errors in how the yield condition was formulated (specifically regarding Karush–Kuhn–Tucker conditions), which the authors agreed to correct.
- *Missing technical details*: The reviewers requested significantly more details regarding the OpenFOAM flow model (e.g., solvers, mesh type, spatial discretization schemes) to ensure the work was reproducible. The authors agreed to add these technical specifications to the appendix.
- *Data resolution and uncertainty*: The reviewers noted that the model's grid spacing (0.5 mm) was coarser than the original fracture surface scan (0.1–0.2 mm) and asked about the effects of smoothing. The authors responded that sensitivity analyses showed the coarser grid produced results similar to those using the finer scan resolution.
- *Incomplete pragmatic validation procedure*: The reviewers noted that while the authors performed validation exercises for aperture distributions and flow rates, they missed "other aspects of the pragmatic validation procedure" mandated in Section 4.5 of the Task Description. A specific missing element was a detailed description of the model purpose, which the reviewers identified as key in the pragmatic validation process. The authors accepted this and added text to Section 3.1.1 clarifying that the model's purpose was limited to assessing the ability to predict flow across a natural fracture under normal load.
- *Mechanical model validity and data exclusion*: There was a disagreement regarding which experimental data should be used to validate the mechanical aspects of the model. Reviewers requested the inclusion of stress field plots and comparisons to contact pressure measurements (pressure films). They argued these were necessary to help the reader assess the mechanical model's validity. The authors rejected this request. They argued that the pressure film measurements contained high uncertainty and that relying on erratic measurements for validation would add confusion.

- *Validation of the local cubic law*: The validation results led to a debate regarding the conclusions drawn about the governing physical laws. The authors initially concluded that the local cubic law was invalid because their computed flow rates were three orders of magnitude higher than the experimental validation data. The reviewers argued this conclusion was too strong, suggesting that the discrepancy did not directly indicate the invalidity of the law, but could instead result from uncertainties in the initial fracture aperture. The authors agreed to soften the language to state the results "seem to confirm" the findings and added text acknowledging that aperture uncertainty contributed to the overestimation.

The reviewers concluded that the report only partly met the review criteria, specifically noting that the document was "lacking (...) aspects related to the pragmatic validation requested in the Task Description".

Throughout the document, the authors generally accepted the reviewers' requests to add missing information (such as abstracts, references, and legends) and to clarify specific modelling limitations. However, they respectfully disagreed on including certain plots (e.g., pressure film measurements) due to high uncertainty in the raw data.

### 3.3.2 Peer review of TUL by A21-KTH

The peer review of TUL's Modelling Report (summarized in Section 3.2.2) was conducted by A21-KTH. The following is a summary of the review record, which includes the initial comments of A21-KTH, the responses by TUL and potential dispute resolution and acceptance.

The reviewers noted that TUL presented the aspects of pragmatic validation well. However, they initially found the writing needed to be more specific, concise, and formal. The review process resulted in several technical clarifications and structural changes to the report:

- *Abstract and objectives*: The authors were asked to make the abstract more specific (introducing the stress-displacement model earlier) and to explicitly state the exact objectives and limitations of the study.
- *Justification of assumptions and methodology*: A core component of pragmatic validation is defining the domain of applicability. The reviewers requested more rigorous justification for the specific methods and assumptions used. The reviewers questioned the validity of the one-way coupling assumption and the specific method used to calculate flow which ignores the deformation caused by neighbouring contacts. The authors acknowledged this was a new model developed for this task and clarified the text to reflect these limitations.
- *Data sources*: The reviewers requested clearer distinctions between the source data (from experiments) and the team's own processing, specifically regarding surface data.
- *Modelling approach*: The authors had to clarify their one-way coupling assumption in the stress model and explain their new method for calculating flow, which ignores deformation caused by neighbouring contacts.
- *Parameters*: Justification was added for the selection of Young's modulus (attributed to scale effects) and the estimation of optimization parameters.
- *Application of pragmatic validation*: The reviewers noted that while the pragmatic model validation was generally well presented, the report initially listed generic validation points rather than focusing on those specifically applied to the study. The authors accepted this, extending the text to make the focus narrower and more relevant to their specific modelling exercise.
- *Component validation*: The reviewers asked if there was a chance to validate the stress model separately. The authors confirmed they did not intend any validation prior to its use on the studied experiment.

- *Prediction-outcome exercise*: Several comments sought to clarify the distinction between data used for calibration and data used for blind prediction (i.e., validation). The reviewers asked for clarification regarding the mechanical model, specifically questioning if stress-displacement curves were available to the team before the flow rates were disclosed. The authors clarified that while intermediate load steps were known, they were not used for calibration; only the larger load steps and cycles disclosed with the flow data were used for evaluation.
- *Deletions*: Sections regarding formal uncertainty analysis and channelling factor were removed because the reviewers felt the content was insufficient.
- *Conclusions*: The conclusion section was rewritten to focus on major findings rather than mixing discussion and summary.

The final assessment concluded that the revisions "significantly enhanced the clarity and quality of the report" and that all major comments were resolved.

### 3.3.3 Peer review of VTT by TUL

The peer review of VTT's Modelling Report (summarized in Section 3.2.3) was conducted by TUL. The following is a summary of the review record, which includes the initial comments of TUL, the responses by VTT and potential dispute resolution and acceptance.

- *Meshing and geometry*: The reviewer asked for details on the use of surface source data, resolution effects, and the handling of overlaps (clipping). The authors added details on source data and clarified the clipping section.
- *Model parameters*: Errors regarding the Young's modulus and the specific load applied (block weight vs. prescribed load) were identified. The authors corrected the modulus value and clarified that 0.02 MPa was used in all simulations.
- *Reference positions*: Clarification was requested regarding the values in Table 4-4 (reference positions and $D_{ave}$). The authors updated the table and the corresponding figure.
- *Consistency*: The reviewer noted inconsistencies in notation and the handling of channelling factors. The authors changed the notation for consistency and deleted a normalization procedure that was not used.
- *Corrections*: Various specific errors in equations (e.g., unnecessary numbers, sign errors) and typos were identified and corrected by the authors.
- *Validation methodology and concepts*: The reviewer noted that the authors incorporated sensitivity analysis and validation concepts to support pragmatic model validation. A specific point of discussion was the authors' definition of a "validity rank". The reviewer remarked that this term does not appear in standard literature and suggested that if it is a new concept, it should be emphasized. The authors responded that the concept was developed for this specific work and added clarification to the report.
- *Blind predictions vs. adjustments*: A significant aspect of the validation review focused on the timing of the modelling relative to data availability. The reviewer found it difficult to distinguish which results were true blind predictions and which were model adjustments made after measurement data was disclosed.
- *Comparison of model and measurements*: Specific comments addressed the comparison between simulation and reality, which is central to validation. The reviewer pointed out there was insufficient discussion regarding the comparisons of the model against measurements for pressure film and flow rates. The authors accepted this and added more text to address the comparison.
- *Result agreement*: The reviewer noted that while the approaches were appropriate, the model did not agree well with observations, though it did provide insight into the limits of the data and models.

- *Review process as validation*: The reviewer explicitly stated that their comments—intended as discussion ideas rather than criticism—were suggested to see if other conceptual models and/or interpretations could be applied.
- *Language and formatting*: The reviewer provided an annotated document with language and typographic corrections, all of which were accepted and corrected by the authors.

The reviewer found the 3D contact and Stokes flow solution appropriate and interesting, though noted it did not fully achieve agreement with observations.

### 3.3.4 Peer review of NTU-Taipower by SU

The peer review of NTU-Taipower's Modelling Report (summarized in Section 3.2.4) was conducted by SU.[5] The following is a summary of the review record, which includes the initial comments of SU, the responses by NTU-Taipower and potential dispute resolution and acceptance.

The reviewer described the report as "scientifically and technically robust" and "clear and well-written". The review concluded that the report covers all aspects required by the Task Description, and the critique focused primarily on clarifying technical details and suggesting parameter re-evaluations rather than requiring major changes.

The most significant technical discussion centred on the mechanical parameters used in the model:

- *Initial displacement and contact ratio*: The reviewer sought clarification on how the initial displacement between the upper and lower surfaces was determined. The authors clarified that the initial contact ratio is established when the average aperture of the point cloud reaches the initial aperture used in their regression equation.
- *Critical parameters for pragmatic validation*: The reviewer questioned whether the Young's modulus and initial aperture should be considered critical parameters for pragmatic validation. The reviewer noted that the Young's modulus value used (73 GPa) appeared to be based on intact rock, whereas fractured rock samples typically exhibit lower values. They suggested that a lower modulus would yield greater displacement, and a larger initial aperture would also affect displacement under load. The authors accepted this critique. They adjusted the Young's modulus to 56 GPa and the initial contact ratio to 1.69% in their mechanical model. They confirmed that these specific values were selected because they yielded results most consistent with the hydro-mechanical coupling experiment in the hydraulic model.
- *Sensitivity analysis*: In response to the discussion on Young's modulus and initial contact ratio, the authors added a section to the report (Section 5.1) illustrating how these settings influence displacement across loading stages.
- *Model assumptions*: The reviewer suggested that underlying assumptions—such as mass conservation, Darcy's law, and the cubic law—should be explicitly stated in the "Model description" or "Determination of critical aspects" sections. The authors revised Section 3.1.3 to list these assumptions explicitly. This included the assumption that fracture surfaces are optimally matched and that the transmissivity conversion formula is inspired by the cubic law but relies on regression parameters derived from laboratory tests.
- *Regression and physical motivation*: The reviewer asked if the transmissivity equation (Equation 3-4) could be motivated by physical principles rather than just regression and requested details on how the regression parameters were obtained. The authors acknowledged that they currently use a trial-and-error approach to determine these parameters. They justified the formula by stating it aims to maintain the relationship between transmissivity and aperture found in the cubic law while incorporating geometric parameters and normal displacement to align with experimental data.

---

[5] Stockholm University did not participate in Task 10.2.2. The Task Force greatly appreciates it that SU's A. Frampton agreed to serve as a reviewer.

- *Methodology explanations*: The authors added references and explanations for specific algorithms used, including the ICP algorithm for shape matching and the K-D tree method for spatial partitioning.
- *Code transparency*: To support the validation of the self-coded software used by the NTU-Taipower team, the reviewer requested access to the source code. The authors provided a GitHub link to the code modules used for Task 10.2.2.

The review process led to several specific clarifications and editorial corrections in the final report:

- *Mathematical notation and accuracy*: The authors corrected the transmissivity formula to include fluid density.
- *Visualizations*: At the reviewer's suggestion, the authors replaced Darcy velocity vector field figures with streamline visualizations to better depict flow pathways and channelling behaviour.
- *Assumptions*: The authors removed a statement regarding geochemical processes, clarifying that no significant geochemical phenomena were observed during the short timeframe of their laboratory experiments.

The authors accepted all comments and suggestions provided by the reviewer, resulting in the successful resolution of the review.

## 3.4 Summary evaluator review

The following subsections contain summaries of the evaluator's technical comments concerning the final drafts of the individual Modelling Reports. Note that the reviewed draft versions included the revisions made during comment resolution of the peer review discussed in Section 3.3. The evaluator's comments were made as part of the SKB factual review process prior to publication of the Modelling Reports as SKB reports in the R series ("reports detailing specific scientific or engineering studies"). The authors received the review records with the evaluator's comments and provided responses. The extent to which the reviewer's comments have been addressed in the published version of the reports has not been assessed. Consequently, the following subsections only summarise the evaluator's comments without the response by the authors or changes made to the report.

In addition to making specific technical comments and editorial suggestions, the evaluator's review focused on the question how the modellers addressed pragmatic model validation, which is the main topic of Task 10.

### 3.4.1 Evaluator's review of modelling work by A21-KTH

The review of A21-KTH's Modelling Report for Task 10.2.2 provides a critical assessment of the simulation methodologies and the application of the pragmatic model validation approach. A primary focus of the review is the distinction between model calibration and model validation, the need for transparency regarding model calibration, the predictive power of the models, and the clarity of visual data.

- *Aperture characterisation*: The reviewer raised several questions regarding how the physical properties of the fracture were translated into the model, noting that surface roughness scan data alone are insufficient to determine initial aperture. He asked for a clearer definition of this term and an explanation of how scan data were used before being calibrated. There is a request for clarification on why contact points were assigned a very low permeability value rather than zero, and whether a residual aperture was enforced at these locations.

- *Calibration vs. validation*: The reviewer emphasized that the relationship between calibration and validation must be clearly defined. The reviewer requested clarification on whether the initial aperture was calibrated before or after validation. He noted that calibration is a necessary step to develop a model capable of making reliable predictions under varying stress conditions. However, if the determination of the initial aperture significantly depends on a calibration step, the validation exercises is limited to the model's ability to predict flow and channelling under changed loading conditions rather than the complete workflow from measurements of fracture surface roughness to these predictions of interest.
- *Dependence on calibrated parameters:* The reviewer pointed out that the model's reasonable flow rate predictions at the initial stage are largely due to the use of a calibrated correction factor ($f = 0.1$). The reviewer argued that using such a factor for the initial stage does not inherently validate the model development approach, especially if it cannot predict behaviour under extrapolated conditions, such as different loading conditions.
- *Predictability and extrapolation*: A central aspect of validation is to test the model's ability to remain accurate when conditions change. The reviewer expressed concern that if the model's success is limited to the initial, calibrated stage, it may negate the claim that the approach aligns with the task's general validation objectives. This also questions how the model's results can be generalized to a broader range of fractures and, crucially, how such a generalization would be validated.
- *Comparison to measured data*: To better understand the impact of calibration, the reviewer suggested providing histograms and cumulative distribution functions showing the aperture distribution both before and after the calibration shift was applied. He called for a more detailed discussion of the visual comparison between modelling results and measured data, particularly where the model overestimates results, to understand the underlying causes and the data needed to improve the model's reliability.
- *Data sufficiency for validation*: Based on the use of an additional calibration step needed to determine the aperture field, the reviewer suggested that geometric information alone (such as surface roughness scan data) is insufficient for a reliable prediction of fluid flow. A model developed solely based on fracture surface maps provides flow predictions that likely would not meet standard validation acceptance criteria without additional data to constrain both geometry (mean aperture) and flow (transmissivity).

### 3.4.2 Evaluator's review of modelling work by TUL

The review of TUL's Modelling Report for Task 10.2.2 characterizes the technical work as "very good," particularly praising the author's pragmatic validation approach and transparency regarding limitations. While the report is noted as meeting the criterion that methodologies and models are verified and validated, the reviewer provides several specific comments to strengthen the validation component. Unfortunately, many descriptions and explanations were difficult to follow and ambiguous, which could lead to misconceptions or confusion by the reader to the report.

- *Calibration vs. validation*: A critical point is raised regarding whether block positions were adjusted or checked. The reviewer emphasized that if the positions were merely checked, it constitutes validation, but if they were adjusted to match observed behaviour, it becomes calibration. This distinction is noted as being highly important within the context of Task 10, which focuses on model validation.
- *Missed validation opportunity*: The reviewer expressed that it is unfortunate that trace photographs and contact pressure films were not used to validate the relative position of the two fracture surfaces. He argued that validating this aspect is central to moving from surface roughness data to aperture to transmissivity and eventually to accurate flow predictions. Adding these data to the validation exercise would have significantly increased confidence in the computational models.
- *Qualitative validation exercises*: The reviewer suggested that the author perform a qualitative or, preferably, quantitative validation by comparing measured effective stiffness in normal stress against pressure film images.

- *Confidence in physical mechanisms*: The reviewer suggested elaborating on the deviation from proportionality between flow rate and pressure gradients. To give more confidence in the model, the reviewer proposed providing independent arguments regarding the likelihood of non-Darcy flow effects. This would help ensure that changes in flow rate are correctly attributed to mechanical effects like aperture changes, rather than non-Darcy flow effects.
- *Uncertainty and bias*: The reviewer requested a clear definition of the uncertainty range for measurements, noting that model results currently sit at the border of this range. He also highlighted a characterization bias: because laser scanning occurs after the fracture is opened, the model may be more validated for resealed experiments than for the natural, unopened state of the fracture; direct field testing might be more appropriate.
- *Physical interpretation*: The reviewer suggested elaborating on the deviation from proportionality between flow rate and pressure gradients, noting it could be used to infer deformations in aperture distribution if other effects like non-Darcy flow are ruled out.
- *Clarity and completeness of documentation*: The reviewer requested clearer definitions for key terms. Several sections require elaboration, including the logic behind certain aperture-related calculations. The report lacks a short description explaining why the generation of synthetic aperture fields was considered not possible, which the reviewer deemed a key aspect of the process. The reviewer provided extensive editorial suggestions and a marked-up version of the manuscript to address these writing and clarity issues.

### 3.4.3 Evaluator's review of modelling work by VTT

In the overall assessment of the factual review of VTT's Modelling Report, the reviewer found that the report meets all the defined review criteria, including the requirement that methodologies and models are verified and validated. Nevertheless, the reviewer provided several specific comments focusing on the validity of the modelling approach and assumptions, particularly concerning how the model predicts fluid flow through fracture apertures.

- *Aperture and flow overprediction*: The reviewer noted that the 0.1 mm contact offset used to solve convergence issues is large and likely contributed to the high overpredictions of flow rates. He asked for clarification on why the target offset was based on measurement precision (0.035 mm) rather than zero.
- *Flow overprediction*: The use of a large contact offset, chosen due to convergence issues, likely significantly affected the average aperture and contributed to overpredicted flow. A significant cause of flow overprediction appears to be a limitation in COMSOL, where geometrical overlap prevented flow modelling. This suggests that finding a way to handle these overlaps is essential to properly simulate flow in tight fractures with contact asperities.
- S*ubmodel validation*: A critical point raised is that the contact model predicts aperture distributions, not fluid flow. The reviewer emphasized that it is essential to validate each submodel independently before passing information to the next. If the final flow predictions fail to meet the validation acceptance criteria, it remains unclear whether the failure lies in the contact model, the flow model (e.g., the cubic law assumption), or both, or something else.
- *Impact of assumptions on results*: The reviewer questioned the assumption that water pressure and effective stress changes were negligible, particularly under zero normal load. Even small aperture changes in response to loading could have large impacts on fluid flow; the related assumption should be supported by calculations rather than just stated. Moreover, the assumption of using the contact gap as the effective aperture was noted as a potential source of error, as heterogeneous aperture distributions in rough fractures can lead to flow behaviours different from those predicted by a parallel plate model.
- *Reference models*: The reviewer requested a clearer description of the reference model used for validation ranking.

In summary, while the reviewer gave a positive overall assessment, he expressed concern that if the contact model cannot be validated independently, using surface roughness data to predict fluid flow might not be a defensible modelling path.

### 3.4.4 Evaluator’s review of modelling work by NTU-Taipower

The reviewer acknowledged the high quality NTU-Taipower’s Modelling Report but raised several critical technical points regarding model calibration and validation. Note that NTU-Taipower provided a response to each comment; these responses are included in the summary below.

- *Model calibration vs. validation*: The reviewer expressed concern that the authors were using flow rate data—intended for model validation—to calibrate the coefficients of the transmissivity conversion formula. He emphasized that adjusting parameters to match experimental data is calibration, preventing validation, which requires comparison to data not used for model development. He emphasized that for a true validation (a prediction-outcome exercise), the model should be compared to data without further parameter adjustment. To ensure the independence of the validation process, the authors implemented several changes: (a) the methodology was revised so that the model is calibrated using LVDT displacement data and subsequently validated the model using independent flow rate data; (b) because the original transmissivity conversion formula relied on flow-related calibration, it was replaced with the cubic law to maintain the integrity of the validation process; and (c) parameters such as the initial contact ratio and Young’s modulus are now determined solely through calibration against experimental displacement data before the validation phase
- *Mass conservation and numerical methods*: The reviewer noted a large difference between inflow and outflow rates in the simulations, questioning the system's ability to maintain mass conservation. The authors switched from a finite difference method to a finite volume method, which resolved the mass conservation issue. They removed the previous grid-plot size analysis that had been used to investigate these discrepancies.
- *Fracture alignment and contact ratio*: The reviewer requested clarification on how the upper and lower fracture surfaces were aligned and how the desired contact ratio was determined, as these factors control mean aperture, transmissivity, and channelling. The authors explained that alignment is achieved using a point cloud processing algorithm (ICP) to minimize the distance between surfaces. The initial contact state is now determined through the calibration of the mechanical model, specifically focusing on vertical translation and the Young’s modulus of the surfaces.
- *Uncertainty and sensitivity*: The reviewer pointed out that the initial uncertainty analysis was a grid convergence study and suggested focusing on conceptual and parametric uncertainties. The authors identified the alignment condition (specifically $z$-direction translation) as the primary source of uncertainty. They revised the analysis to assess the impact of these translation adjustments and the assignment of a minimum transmissivity to contact areas.

The authors also addressed minor editorial suggestions and corrected calculations for hydraulic head differences and flow-rate reductions to ensure consistency with Darcy’s law.

# 4 Evaluation

Based on the concepts and expectations formulated in the White Paper (Lanyon et al. 2024) and the Task 10.2 Description, the MGs exercised the pragmatic model validation workflow:

- *Defining purpose and critical aspects*: Each team formulated their model's intended purpose and identified critical aspects (e.g., initial aperture uncertainty, mechanical-hydraulic decoupling) as required by the pragmatic validation workflow.
- *Adherence to the "blind prediction" protocol*: All teams conducted prediction-outcome (P/O) exercises, submitting flow rate and displacement predictions before the experimental results were disclosed.
- *Use of alternative models and methods*: In line with the Task Force goal of examining alternative models, some teams employed diverse approaches—from analytical parallel-plate models to complex 3D Navier-Stokes CFD and coupled elastic-plastic contact models.
- *Identification of model limitations*: The MGs developed hypotheses why their models failed to predict key performance measures. They suggested and/or examined alternative models or simulation approaches.

While the key elements of the pragmatic model validation workflow have been addressed, the following critical observations are made based on the expectations and criteria outlined in Section 2.3:

- *Relation of validation acceptance criteria to model Purpose*: Most MGs declared the purpose of the model to be identical to (or a slight variant of) the generic formulation provided in the Task Description. While the Task Description defined the objectives of Task 10.2.2 (which essentially states that the MGs should make specific predictions and validate their models), a different formulation is needed that clearly identifies the purpose—or "intended use"—of the *model* (not the task). This formulation of the model's purpose must be suitable to guide the eventual evaluation whether the model's performance is "fit-for-(*this*)-purpose". The formulation of the model purpose must be detailed enough and contain information about intended application domain and acceptable prediction uncertainty to allow for the derivation of validation acceptance criteria.
- Note that the Task Description did *not* provide this information. It was intentionally[6] left up to the MGs to define what their models are supposed to accomplish, how they will be used, what the acceptable prediction uncertainty is, and what the corresponding validation acceptance criteria should be. The Task Description only provided the context (predict flow and estimate channelling factors in a larger-scale fracture—and eventually fracture network—under different stress conditions).
- As Task 10.2.2 is preceded by Task 10.2.1, it is implied that the models should be based on specific characterization data (i.e., high-resolution geometric information of fracture surface roughness). However, if the outcomes of the pragmatic validation of the models developed under Task 10.2.1 suggest that the transition from geometric and mechanical information to hydraulic predictions is difficult to accomplish or that the prediction uncertainty is unacceptable for the model's intended use, this outcome must be addressed. It could simply be acknowledged, and the models of Task 10.2.2 could be developed using alternative approaches (e.g., involving a direct calibration to flow data, either from Task 10.2.1 or using a subset of the Task 10.2.2 validation data set, which, as discussed below, will affect the validation acceptance criteria).

---

[6] In the context of radioactive waste disposal research, it is often up to the scientists, researchers, and modelling experts to identify and describe the specific purpose and application domain of their models, based on guidance that only provides the broader context and/or more generic, higher-level goals. The Task Description mimics this situation. Furthermore, the MGs were given considerable leeway in formulating their models' purpose so they can make the Task Force exercise more relevant for their specific interests and account for available modelling capabilities.

- *Identification of critical factors*: The MGs evaluated model variants or performed formal sensitivity analyses to identify influential factors. Most of these analyses, however, examined the influence of a factor (e.g., translation or rotation of fracture surfaces) on an *intermediate* result (e.g., contact area ratio or degree of overlap or initial mean aperture). However, to identify a "critical factor" in the context of model validation, one must calculate the influence of an unknown or uncertain input parameter, factor, or model assumption on the ultimate performance measure (i.e., the ones reflecting the model purpose and used for model validation). For example, the sensitivity analysis should address the question how the flow rate or channelling factor of the upscaled fracture change if the distance between the rough surfaces of the small-scale fracture used for aperture estimation is changed. While it may be appropriate and even useful to subdivide the sensitivity analysis into multiple steps, it requires that the resulting sensitivities and uncertainties must be propagated through the chain of submodels to arrive at the model that is making the relevant prediction to address the model purpose. For example, while the *z*-translation may have a big impact on initial contact area, one must also demonstrate that the initial contact area is an influential parameter for the mechanical model and the details of that model significantly affect the predicted flow rates, which is the ultimate performance metric. The ranking of a factor's criticality for the model purpose of interest can only be assessed if the influence of all factors is calculated for the same performance measure (i.e., not for different, intermediate results).

  It is noticeable that many MGs identified the precision and/or resolution of the surface scan as a critical factor that explains the considerable discrepancy between the measured and predicted flow rates observed during the P/O exercise. However, assigning a high influence ranking to the scanning precision is likely not warranted and a result of the fact that uncertainty in this factor is not appropriately propagated to the ultimate performance measure. Other factors, which may be related to modelling decisions rather than the underlying raw data, appear to the evaluator to be more critical to the outcome of the P/O exercise.

- *Calibration, blind prediction, and validation*: The development of a model addressing flow and channelling in a single fracture based on surface-roughness data obtained on a smaller-scale sample involves multiple steps and the linking of different types of calculations and models. This includes preprocessing of the surface roughness data, interpolation onto a computational grid, upscaling, calculation of aperture fields under different loading conditions, potential conversion to local or equivalent fracture transmissivities, simulation of fluid flow using an appropriate set of governing equations, and computation of channelling factors. Each of these submodels involves conceptualisations, simplifying assumptions, and the selection of parameters that represent both material properties and state variables (i.e., initial and boundary conditions). Data are needed for conditioning, calibration and validation of the model. Examples of model calibrations performed by the MGs include the estimation of parameters of the contact model and force-displacement relationships by matching contact pressure film data patterns, LVDT deformation measurements, independently estimated fracture stiffnesses, and flow rate data. While these data can be used to estimate parameters by minimizing the misfit, they can also be used to validate part of the model used to calculate intermediate results. The only—but crucial—difference between calibration and validation is whether the parameters are adjusted after the model-to-data comparison.

  As just indicated, available characterisation data can be used for conditioning, calibration, or validation. It is essential to clearly state how the data are used, particularly in the context of Task 10, where a specific set of data is reserved for the P/O exercise in support of pragmatic model validation. Data used for validation may not have been involved in the development of the model (specifically for conditioning and calibration). If the same (or a portion) of these data are used to adjust the model during its development, this reduces—or invalidates—the power of the P/O exercise to confirm or reject the applicability of the model in the extrapolated model space.

If the differences between the observed and predicted performance measures exceed the validation acceptance criteria, the model failed the validation test. As will be discussed below, the blind predictions from most models resulted in a flow rate through the upscaled fracture that is considerably higher than what was observed in the experiments conducted for the P/O exercise. These differences provide valuable insights into potential shortcomings of the model.

If model validation fails, it is prudent to make use of the information contained in the validation data to identify which aspects of the model need to be refined, revised, or replaced by an alternative conceptual model. The validation data may also be used quantitatively to (re)calibrate the model. For example, the parameters of the aperture-transmissivity relationship or the transmissivity itself can be adjusted to fit the flow data. If such model adjustments are made in response to an initial P/O exercise with a negative outcome, the following points need to be considered:

- *Need for new validation*: Because the validation data have been used for model adjustment and model calibration, they are no longer usable for validation purposes. The predictions are no longer "blind" but influenced by knowledge about the outcome. Furthermore, the predictions no longer examine a model space that is an *extrapolation* from the calibration space. Consequently, the revised model must undergo a new validation exercise based on a *different* set of validation data.
- *Change in model structure and data support*: Revising the model based on validation data likely changes the model's fundamental structure and the characterisation data that support its development. For example, if fracture transmissivities are directly determined from flow data, many of the preceding submodels are bypassed. This means that—even if the revised model passes the second validation test—it can no longer be claimed that using high-resolution surface-roughness data leads to reliable flow predictions, or that the upscaling approach is validated, or that the contact model provides reasonable deformation estimates, or that a suitable aperture-transmissivity relationship has been found. The validity of some of these submodels may still be tested and confirmed with an appropriate set of additional validation data, but the chain of validated models between the fundamental characterisation data (mainly small-scale surface roughness data) and the application model (used to predict flow through a larger fracture under different loading) is broken. In this case, it must be acknowledged that while individual submodels can be validated, they do not necessarily contribute to the fit-for-purpose validation of the ultimate prediction model.
- *Documentation*: Because of the subtle but complex interactions between model development and validation (as outlined above), it is essential to fully and transparently document the decisions made during the pragmatic model validation process, especially if it involves multiple iterations. Such documentation should include clear statements about (a) which data are used for model development versus model validation; (b) which submodels or aspects of the model are tested by the P/O exercise[7]; (c) which model predictions are considered supported by the validated model ("application domain"), and (d) what assumptions and decisions were made to arrive at the final model[8]. This also implies that the validation process does not end with reporting the pass-fail decision, but the completion of the Model Audit of the entire pragmatic model validation process (see Section 2.2).

---

[7] The "validity rank" proposed by VTT (see Naumer and Pulkkanen 2025, Section 2.4) is a commendable attempt to include the scope of each P/O exercise into the pragmatic model validation approach.

[8] The comments made during the peer review and SKB factual review indicate that many of these important issues were not fully addressed or not clearly and transparently documented.

- *Discrepancies between predicted and measured flow rates*: Most teams reported that their predicted flow rates were one or more orders of magnitude higher than the experimental data. As an example, Figure 3-1 shows the comparatively close predictive fit obtained by NTU (Wang et al. 2025), which was mainly attributed to the use of the Iterative Closest Point (ICP) algorithm employed to obtain a more realistic alignment of the two fracture surfaces. Note that the accuracy of the predictions varied depending on (a) flow direction, (b) loading state, (c) whether comparison was made to the data from the unopened or resealed fracture, (d) which modelling approach was used (if alternative methods were examined), and (d) the selection of certain input parameters examined during the sensitivity analysis or adjusted during post-validation model calibration.

  Notably, the models tend to overpredict[9] the measured flow rates or equivalent transmissivities, indicating a systematic bias existed. The MGs attributed this bias mainly to (1) geometrical errors caused by (a) insufficient data precision or resolution and associated smoothing effects (b) misalignment of the fracture surfaces, (c) the need to minimize or completely avoid geometrical overlap by introducing an artificial contact offset, and (d) the impact of material loss (break outs) during fracture opening and resealing; (2) systematic errors inherent in the modelling approach of the contact and mechanical models, such as (a) the assumption of linear elasticity, (b) neglecting permanent deformations and rock breaking, (c) artificially increased stiffness at asperity contact, (d) ignoring effective-stress impacts, (e) assuming homogeneous material properties, and (f) procedural inaccuracies (e.g., inherent differences in the nearing or departing contact model); and (3) systematic inaccuracies in the flow model, including errors in (a) the aperture-transmissivity relationship, specifically the application of the local cubic law, (b) application of Darcy's law instead of the (Navier-)Stokes equation, (c) vertical averaging by using a local transmissivity rather than full CFD simulations within the fracture aperture, (d) ignoring small-scale anisotropy, and (e) neglecting the impact of trapped air bubbles, among other, unspecified factors.

  The MGs arrived at some of these explanations through the following means: (1) sensitivity analyses, including grid-convergence studies; (2) validation exercises of intermediate submodels (e.g., comparison of calculated contact area to estimates from contact pressure film data; comparison of deformations to LVDT data); (3) using alternative conceptual models (e.g., Darcy's law vs. Stokes equations; alternatives to cubic law for aperture- transmissivity relation); and (4) examination and discussion of data uncertainties (e.g. ,  reference position of displacement sensors; precision and resolution of surface scans).

  As discussed above, the considerable discrepancies and systematic bias between measured and predicted flow rates provided the MGs with valuable insights about the significance of accurately capturing the mechanical and flow-relevant properties of fractures with rough surfaces, as well as the challenges associated with deriving these properties from geometric information alone. Moreover, it pointed the MGs towards aspects of the (sub)models that need to be revised to improve predictions and their reliability. This may require changing the overall model development approach, including the use of different characterisation data that may be more suitable for conditioning and calibrating the model.

---

[9] The example shown in Fig. 3-1 is an exception in that it *underpredicts* the measured data for the 2-4 direction of the resealed fracture under unloaded conditions. The measured high flow rates may likely be caused by the initial misalignment of the two fracture surfaces prior to the application of loading stresses. The rates are reduced drastically as soon as the first loading step of 1 MPa is applied, a steep reduction that is not captured by the model. Moreover, the model did not reproduce the observed anisotropy. Also note that, in comparison to the other participating MGs, NTU-Taipower's model produced the *smallest* discrepancies between predicted and measured equivalent transmissivities (representing flow rates).

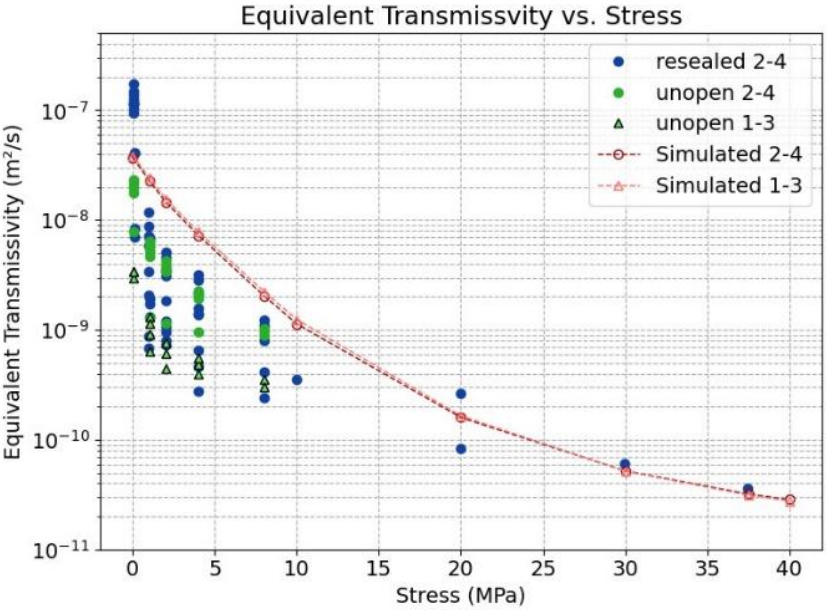

***Figure 4-1.*** *Example result of prediction-outcome exercise: Prediction of equivalent transmissivity and corresponding validation data for different loading and flow directions. Source: Wang et al. 2025; Figure 4-8.*

- *Flow channelling*: based on the predictive modelling, the MGs successfully characterized flow channelling (by visualisation of velocity field or flow concentrations within the fracture plane and/or the calculation of the channelling factor proposed by Maillot et al. (2016)). Unfortunately, they did not validate the channelling factors because the experimental setup did not provide independent measurements or estimates of flow channelling that directly correspond to this performance measure. Opportunities to obtain such measures have been discussed (tracer staining tests; rate measurements for individual subsections of the inflow and outflow fracture traces; flow visualization in transparent fracture replicas).

  In the absence of suitable validation data[10], the calculated flow channelling factors could be compared to literature values or by cross-comparison with the results obtained by the other participating MGs. Moreover, the dependence of flow channelling factors on aperture and/or loading may indicate whether the model phenomenologically captures the expected increase in flow concentration as the impact of surface roughness on the flow field increases (as reflected by the ratio of roughness amplitudes over mean aperture), or the asperity contact area becomes larger. If the calculated channelling factor remains insensitive to imposed loading changes, this may indicate that the mean aperture is too large relative to the amplitude of the surface roughness, consistent with the overestimation of flow rates.
- *Validation acceptance criteria*: While pragmatic model validation can be characterised as a comprehensive review process rather than the calculation of a numerical score that is compared to a pass/fail criterion, it is essential to formulate validation acceptance criteria, specifically for the P/O exercise. A validation acceptance criterion captures a specific expectation about the model's performance; it indicates whether the model can be considered useful for the application for which it is designed, i.e., it assesses whether the model is "fit-for-purpose".

---

[10] It is unclear whether the Rhodamine-B staining test yielded data of sufficient quality so that they could have been used for validation purposes.

As highlighted above, the validation acceptance criteria must be formulated from the perspective and needs of the stated modelling purpose, which is also reflected by the (highest) prediction uncertainty that can be considered acceptable without the model losing its usefulness.

The importance of (a) formulating the model purpose, (b) identifying relevant performance measures that can represent this purpose and evaluate the model's capability to meet expectations, and (c) defining the associated validation acceptance criteria has not been fully communicate to or recognized by the MGs. The teams did not state *a priori* what discrepancies between the predicted and measured flow rates they considered acceptable to declare the model "fit-for-purpose". Note that a validation acceptance criterion may be formulated as a hard threshold value, a statistical measure, reliability metrics, or hypothesis test statistics; depending on the modelling purpose, it can also be a soft, qualitative statement[11]. In the context of Task 10.2.2 and the use of the single-fracture models developed in this task for subsequent predictions using discrete fracture network models, it would be preferrable to have a rather quantitative rather than qualitative validation acceptance criterion. Nevertheless, for the purposes of a pragmatic model validation exercise, a reasonable validation acceptance criterion can be set by decree (with or without explanation) or formulated as deemed suitable to achieve the teaching goals of the exercise; however, they should be consistent with the model purpose, which is also formulated by the MGs.

The modelling teams reached diverse conclusions regarding the "validity" of their models, ranging from successful validation with high confidence to identifying fundamental failures in the predictive power of their modelling tools. While each of these assessments can be considered sensible, they lack a rational explanation or justification for reaching them—due to the lack of a clearly stated model purpose and a validation acceptance criterion. As such, these conclusions about the predictive power of the models may be misinterpreted as a generic modelling capability and potentially being giving a wrong (i.e., too optimistic) impression about the ability of the model to address important issues related to the safety of a radioactive waste repository.

Conversely, those MGs that perceived the discrepancies of their models to be large may have reached too pessimistic a conclusion regarding the usefulness of their models—or the usefulness of the available characterisation data to develop them.[12] A "fit-for-purpose" validation acceptance criterion may have prevented reaching the conclusion that certain aspects of the model (e.g., the use of Darcy's law or application of the local cubic law) or data collection system (e.g., laser scanning) must be categorically rejected.

---

[11] Examples of qualitative validation acceptance criteria include: "the synthetically generated, upscaled aperture distribution is visually consistent with the patterns seen by the processed contact pressure film data"; "the model reproduces the observed trend as a function of loading stress"; "the model provides an unambiguous explanation of the impact of surface roughness on flow channelling"; "the model complies with the requirements of current regulations"; "the model captures the salient features of the observed system behaviour".

[12] A formal data-worth analysis links the acceptable uncertainty of characterisation data to the acceptable prediction uncertainty of the model to be developed based on these data. This avoids using characterisation and calibration data that do not contain sufficient information needed for the development of a model that has to meet the requirements of a specific application. It also helps avoid collecting data with unnecessarily high resolution or precision, and associated high collection, processing, and analysis cost. For more details, see Finsterle (2015).

# 5 Summary, conclusions, and recommendations

## 5.1 Summary

Four teams participated in Task 10.2.2: Amphos21-KTH, TUL, VTT, and NTU-Taipower. Based on the approaches established under Task 10.2.1,13 the Modelling Groups developed or refined models that simulate fluid flow across upscaled, single fractures under different normal loads. As part of the pragmatic model validation workflow (Lanyon et al. 2024), the predictions of fluid flow (or equivalent transmissivity) were compared to corresponding laboratory test data, which were not revealed to the MGs until after completion of the simulations. Some of the MGs revised their models to explain and/or reduce the discrepancies between the model predictions and the data. Flow channelling within the fracture plain was also examined but not included in the formal validation process due to the lack for comparable validation data.

The development of the geometric, mechanical, and hydraulic models was documented, including the validation of key aspects of the submodels as well as the results of the "fit-for-purpose" prediction-outcome exercise. Drafts of these Modelling Reports were subjected to a formal peer review, with comment resolution documented in standardised Review Statements. The final drafts underwent an SKB factual review before being published as SKB reports in the R series.

This Evaluation Report summarises critical comments by the Task 10 evaluator. The comments are drawn from (a) the evaluator's participation in the Task Force Meetings and Workshops (where the MGs presented their intermediate results, and the evaluator communicated his preliminary impressions and concerns about the modelling and validation work), (b) several reviews of earlier versions of the Modelling Reports, (c) the Review Statements from the peer review, and (d) the final SKB factual review.

Given the frequent interactions between the researchers participating in Task 10.2.2 and the evaluator, the following conclusions are not the result of an independent assessment, but those after an iterative process that involved information exchange among technical peers. It is believed that this is acceptable as it reflects the collaborative environment typically found within the international radioactive waste management community. Finally, the evaluator chose to focus his comments on the way the MGs addressed pragmatic model validation rather than on the quality of the simulation models themselves, which he generally considered to be very high.

---

[13] Task 10.2.1 focused on the characterisation of a single fracture based on high-resolution surface roughness measurements and the development of corresponding geometric and numerical models, with validation limited to testing the generation of upscaled aperture distributions (Finsterle et al., 2024). Some of the Task 10.2.1 participants (LANL, KAERI, and SKB-SU) decided not to be involved in Task 10.2.2. SU contributed to Task 10.2.2 as a reviewer of the formal peer review process.

## 5.2 Conclusions

One of the main goals of Task 10 is to apply the tenets of pragmatic model validation—as outlined in Lanyon et al. (2022)—to an intermediate model that aims at simulating channelised water flow through a single, rough-walled fracture under varying normal stress conditions. This model is expected to be developed based on high-resolution surface-scan data obtained from a smaller-scale fracture sample; furthermore, the validated single-fracture model is expected to inform downstream models for the eventual understanding and prediction of fluid flow through large-scale fracture networks. The following observations are made:

- *Quality of simulation models*: All participating modelling teams demonstrated their knowledge, understanding, diligence, and skills in developing sophisticated models that (a) incorporate the available high-resolution data, (b) capture complex mechanical and hydraulic processes, which were examined using different alternative conceptual models, and (c) involve computationally demanding simulation capabilities. Comments about the quality and potential shortcomings of these models can be found in the various records that document the review of the individual Modelling Reports.
- *Quality of pragmatic model validations*: Overall, the modelers effectively demonstrated the pragmatic validation workflow as an auditing process. As the predictive accuracy for flow rates was generally low, it prompted the modellers to research the reasons why their models failed to predict the observed outcome with sufficient accuracy. They provided recommendations regarding data requirements and changes in model development strategy. In this sense, the exercise fulfilled the goal of building confidence in certain aspects of the models, while also identifying some of the models' limitations. Nevertheless, key aspects of the pragmatic model validation process have not been fully comprehended and/or included in the Task 10.2.2 exercise.

As indicated above, the validation of the models developed as part of Task 10.2.2 must be viewed in the context of the preceding and subsequent tasks, because Task 10 is designed as a sequence of interlinked models of increasing complexity, where each submodel addresses a specific technical issue that needs to be understood, resolved, and validated before it can be passed on to the immediate downstream model. The evaluation of this process (see Chapter 4) leads to the following conclusions:

- *Integration into modelling sequence*: It is not apparent how the lessons learnt from the Task 10.2.1 validation exercise were integrated into the Task 10.2.2 model development and model validation activities. The methodologies developed under Task 10.2.1 have been adopted without major changes even though the generation of upscaled aperture distributions has not been validated convincingly in Task 10.2.1. In particular, concerns about an overestimation of mean aperture (caused by apparent data limitations, decisions made during the establishment of the initial contact area, or shortcomings of the flow simulator) have been raised during the discussions of Task 10.2.1. The resulting uncertainties and biases, while recognized, have not been propagated to the Task 10.2.2 models, nor has the model development approach been sufficiently revised.
- *Model purpose*: The purpose of the model to be developed and validated has not been formulated with sufficient clarity and specificity. The difference between the objectives of Task 10.2.2 (as given in the Task Description), which provides the context and scope of the exercise, and the purpose of the simulation model was not fully recognised. As a result, the "fit-for-purpose" validation remained without a clear target and prevented the formulation of model validation acceptance criteria, which help with the evaluation of the model's reliability and usefulness for subsequent applications.

- *Calibration versus validation*: It is well recognized that data available for model calibration cannot be used for model validation. All MGs completed the predictive modelling without detailed knowledge of the validation data, which were reserved for the prediction-outcome exercise and thus withheld from the modellers. After comparison of the simulated to measured flow rates, three of the four MGs concluded that their models are deficient, requiring either conceptual modifications or further calibration of their existing model using data that are more directly related to the ultimate prediction of interest (i.e., flow rates rather than fracture surface scan data). Indeed, by inserting information from the validation data into the model development process (through calibration), the discrepancies could be significantly reduced. However, it is important to realize that this profoundly changes the interpretation of the residual discrepancies,[14] and essentially undercuts model validation,[15] which then needs to be repeated using new, independent data sets. Unless clearly documented, there is a risk that mixing calibration and validation leads to unwarranted, overly optimistic conclusions about a model's ability to make reliable predictions.
- *Validation of submodels*: Most of the MGs validated aspects of a submodel that provide intermediate results to the ultimate prediction model. As an example, parts of the contact model can be validated by comparison to LVDT data. Such partial validations are highly recommended as they help build confidence in this specific submodel as well as the overall model development approach. They also provide useful insights when diagnosing issues should the ultimate validation fail. However, if a partial validation suggests a deficiency in the submodel, the issue must be rectified before advancing. This is often done by simply estimating some of the influential parameters of the—deficient—submodel using the validation data. This may not only mask the submodels inherent deficiencies but also changes the model space, which in turn constrains the application domain of the validated model. The related issues discussed in the previous conclusion also apply. It is essential that converting validation to calibration data and increasing the number of adjustable parameters affects the interpretation of the pragmatic model validation outcome. If the object of validation is a model that consists of multiple submodels (as is the case in Task 10.2.2), it must be recognized that the chain of validated submodels is weakened or broken if (a) a submodel cannot be sufficiently validated, or (b) an unplanned calibration step is inserted. For example, no clear statements about the (in)validity of surface scan data, the contact model, upscaling method, or algorithm to generate aperture distributions can be derived from the pass/fail outcome of the prediction model if some of the intermediate submodels fail their respective validation or are recalibrated. This issue applies specifically if the final model is recalibrated after validation failure, as discussed in the previous conclusion.

---

[14] Final residuals after model calibration are evaluated using calibration acceptance criteria, which assess reproducibility of measured data in the calibration model space. By contrast, the discrepancies in a prediction-outcome exercise are evaluated by the validation acceptance criteria, which assess blind predictions that refer to the extrapolated prediction model space.

[15] In the fields of machine learning and artificial intelligence, this issue is addressed by the "train-test split paradigm," which aims to avoid the model just memorizing and repeating the training data set rather than generalizing from it to be able to make predictions under previously unexplored or extrapolated conditions.

- *Flow channelling*: Characterisation of flow channelling effects is generally considered essential, specifically for predicting transport times of radionuclides migrating through a fracture network. Understanding channelisation was therefore one of the objectives of Task 10.2.2. The Task Description requested the calculation of the channelling factor as one of the performance metrics. The MGs successfully visualized flow concentrations within the rough-surfaced fracture plane and calculated the channelling factor. Unfortunately, no formal performance-outcome exercise could be performed due to the lack of an independent measurement or estimate of the channelling factor.[16] It can be anticipated that the availability of data reflecting flow channelling would have provided useful information about the geometry of the aperture field as well as its relation to surface roughness and vertical stress, making the pragmatic validation exercise considerably more insightful.
- *Peer review*: The formal peer review moderated by the Task Force Secretariat is considered a meaningful part of the validation exercise. It not only improved the quality of the Modelling Reports but likely broadened and deepened the technical understanding of the MGs who participated in the review both as authors and reviewers. The evaluator would have liked to see more critical review comments on the pragmatic model validation aspect of the project rather than on the details of the simulations.
- *Ensemble*: Having multiple MGs and thus a variety of simulation and validation approaches (which led to a variety of predictions and conclusions) allows the Task Force to obtain additional insights into the practicality, difficulties, robustness, and reliability of pragmatic model validation. While data interpretation, modelling method, and validation approach were chosen by each individual MG, it is acknowledged that the MG's numerical models and the outcome of the validation exercises are not strictly independent from each other, because considerable information exchange took place during Task Force meetings and workshops, the formal peer review process, and the final SKB factual review. This information sharing may have influenced each MG's data interpretation, conceptual model choice, and parameter selection, and consequently also the conclusions. It is unclear whether this reciprocal influence made the outcome of the pragmatic validation exercise more optimistic or more pessimistic. This could be further elaborated by performing a meta-analysis of the collective work performed by the participants of the Task Force. Notwithstanding, it realistically reflects the collaborative work environment characteristic of the international community that seeks solutions to the radioactive waste disposal challenge.

Overall, the evaluator considers the MGs simulation and validation work on Task 10.2.2 to be of great value to all participants and the community. It should be recalled that the primary goal of Task 10 is not to validate specific models, or to obtain a pass/fail decision on their suitability for performance assessment calculations; Task 10 is designed as an exercise to practice the pragmatic model validation approach, and to highlight its importance for the development and use of complex numerical simulation models.

---

[16] Despite the absence of a quantitative measure for comparison, the reasonableness of the calculated channelling factor could have been discussed in more detail, for example by referring to the correlation of channelling to the ratio of surface roughness to mean aperture, the expected relation to asperity contacts as a function of loading, or the comparison to independent field observations or literature data.

## 5.3 Recommendations

The following recommendations are made:

- *Integration into modelling sequence*: The conclusions from the pragmatic model validation exercise of Task 10.2.2 should be heeded and transferred to any of the downstream models. This can be accomplished by (a) reviewing the physical and conceptual understanding gained from the previous subtasks, (b) modifying the models and/or model development approach as needed, (c) observing the application range, and (d) propagating uncertainties through the chain of interlinked models.
- *Model complexity*: As noticed above, the conceptual and numerical models developed for Task 10.2.2 have a high degree of complexity, which led to simulations with high computational costs. This was mainly driven by the request that the prediction model be developed based on high-resolution fracture roughness data. Nevertheless, the concept of pragmatic model validation can be used to streamline the analyses. In particular, the identification of influential factors, which is a key step within the validation workflow, can be performed using potentially quite simple models whose complexity is consistent with the (limited) accuracy demanded by the ranking procedure and the model purpose—as reflected by the validation acceptance criteria. The relative importance of surface roughness characterisation vs. contact model vs. aperture generation vs. upscaling vs. aperture-transmissivity conversion vs. flow equation (Darcy or Navier-Stokes) can be ranked using simplified versions of each of the involved submodels. For example, the *first-order* impact of the mean aperture on flow predictions can readily be evaluated using a back-of-the-envelope calculation, showing its relatively high influence compared to expected differences between models that use either the Darcy or Stokes equations. The relation between mean aperture and the heuristic *z*-translation of the fracture surfaces has already been studied in Task 10.2.1, so has the potential impact of upscaling. The MGs' integration of all these aspects into a highly complex and computationally demanding model is impressive and commendable, so are the modellers' various studies of (equally demanding) alternatives to each of the submodels. The effort expended on these simulations was considerable; in certain cases, this may have limited the time available for the work that is required for a transparent and more conclusive pragmatic model validation study, including its documentation, whose clarity and completeness is essential for the downstream models and the overall success of Task 10.
- *Meta-analysis*: It is recommended to perform a meta-analysis in which the assumptions, predictions, and interpretations obtained by the MGs are compared and combined. This would yield a more realistic assessment of our collective ability to make reliable predictions about flow and channelling in single fractures on various scales and under different loading conditions. If these alternative models yield consistent conclusions about the behaviour of interest, confidence can be gained that useful predictions can be made. It also indicates that the outcome does not greatly depend on uncertain factors, but that the general system understanding as well as the information provided by appropriately selected characterisation data are sufficient to constrain the predictions. Moreover, 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 is accounted for. Such a combined analysis does not state which (if any) of the alternative models is the best representation of the real system; instead, it evaluates the contribution each model makes in support of the overall goal, and pragmatically combines the insights gained from each approach and the interpretation from each participating Modelling Group.

The evaluator hopes that the lessons learnt from Task 10.2.2, including these critical review comments, are considered valuable and useful for the subsequent subtasks, supporting a successful completion of Task 10.

# References

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

SKBdoc documents will be submitted upon request to document@skb.se.

**Andersson P, Byegård J, Dershowitz B, Doe T, Hermanson J, Meier P, Tullborg E-L, Winberg A (ed). 2002a.** Final Report of the TRUE Block Scale Project 1. Characterisation and model development. SKB TR-02-13, Svensk Kärnbränslehantering AB.

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

**Andersson P, Byegård J, Billaux D, Cvetkovic V, Dershowitz W, Doe T, Hermanson J, Poteri A, Tullborg E-L, Winberg A (ed), 2007.** TRUE Block Scale Continuation Project Final Report. SKB TR-06-42, Svensk Kärnbränslehantering AB.

**Boussinesq J, 1904**. Complément au Mémoire intitulé « Recherches théoriques sur l'écoulement des nappes d'eau infiltrées dans le sol et sur le débit des sources. Journal de Mathématiques Pures et Appliquées, 5(10), 363–394.

**Fernandez-Garcia D, Illangasekare, T H, Rajaram H, 2005.** Differences in the scale dependence of dispersivity and retardation factors estimated from forced-gradient and uniform flow tracer tests in three-dimensional physically and chemically heterogeneous porous media. no. W03012. Water Resources Research 41, 3012–3012.

**Finsterle S, 2015.** Practical notes on local data-worth analysis, Water Resour. Res., 51(12), 9904–9924, doi:10.1002/2015WR017445.

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

**Finsterle S, Hokr M, Balvín A, Jankovek J, Gvoždik L, Hyman J, Viswanathan H, Sweeney M, Wang T-T, Chen P-K, Tu C-H, Choi S, Choi C-S, Lee Y-K, Park K-W, Ji S-H, Stock B, Frampton A, 2024.** Evaluation and modelling report of Task 10.2.1: Geometric evaluation and prediction of fracture surfaces and aperture distributions. SKB TR-23-23, Svensk Kärnbränslehantering AB.

**Hardenby C, Sigurdsson O, 2010.** Äspö Hard Rock Laboratory. The TASS-tunnel. Geological mapping. SKB R-10-35, Svensk Kärnbränslehantering AB.

**Hokr M, Balvín A, 2026.** Upscaling rough fracture surface and hydro-mechanical experiment prediction: Modelling report of Task 10.2.2. Task 10 of SKB Task Force GWFTS – Validation approaches for groundwater flow and transport modelling with discrete features. SKB R-23-14, Svensk Kärnbränslehantering AB.

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

**Lanyon G W, Davy P, Dershowitz W S, Finsterle S, Gylling B, Hyman J D, Neretnieks I, Uchida M, 2021.** Pragmatic Validation Approach for Geomechanics, Flow, and Transport Models in Fractured Rock Masses. Paper (DFNE 21-2369) presented at the 3rd International Discrete Fracture Network Engineering Conference, Virtual, June 2021.

**Lanyon G W, Davy P, Dershowitz W, Finsterle S, Gylling B, Hyman J, Neretnieks I, Uchida M, 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. SKB TR-22-04, Svensk Kärnbränslehantering AB.

**Maillot J, Davy P, le Goc R, Darcel C, de Dreuzy J R, 2016.** Connectivity, permeability, and channeling in randomly distributed and kinematically defined discrete fracture network models. Water Resour. Res., 52, 8526–8545.

**Naumer S, Pulkkanen V-M, 2025.** Contact modelling approach to simulate water flow in a rock fracture under normal loading: Modelling report of Task 10.2.2. Task 10 of SKB Task Force GWFTS – Validation approaches for groundwater flow and transport modelling with discrete features. SKB R-23-15, Svensk Kärnbränslehantering AB.

**Saltelli A, Pereira Â 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(2–4), 213–234.

**Trinchero P, Zou L, Iraola A, 2024.** Task 10.2.2 - Modelling flow and flow channelling in a fracture under normal loading. Task 10 of SKB Task Force GWFTS – Validation approaches for groundwater flow and transport modelling with discrete features. SKB R-23-13, Svensk Kärnbränslehantering AB.

**Wang T-T, Chen P-K, Lee T-P, 2025.** Estimating transmissivity field based on fracture geometric characteristics and validation through hydro-mechanical coupling experiments and numerical simulations. Task 10 of SKB Task Force GWFTS – Validation approaches for groundwater flow and transport modelling with discrete features. SKB R-25-02, Svensk Kärnbränslehantering AB.