PROCEEDINGS, 51st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 9-11, 2026
SGP-TR-230

An Open Digital Twin Platform for Co-Simulation and Optimization of Geothermal Plant Operations

Pejman Shoeibi Omrani[1], Leila Hashemi[2], Jonah Poort[2], Aron Schouten[3], Paul J.P. Egberts[2], Ryvo Octaviano[2], Demetris Palochis[2]

Hydrology and Reservoir Engineering Department, TNO, Princetonlaan 6, 3584 CB, Utrecht The Netherlands[1]

Heat Transfer and Fluid Dynamics Department, TNO, Kessler park 1, 2288 GH, Rijswijk, The Netherlands[2]

Acoustic Signatures & Noise Control, TNO, Oude Waalsdorperweg 63, 2597 AK, The Hague, The Netherlands[3]

Pejman.shoeibiomrani@tno.nl

Keywords: Digital twin, geothermal energy, direct use, machine learning, predictive maintenance, artificial intelligence

ABSTRACT

Digital twins are increasingly transforming the operation and management of geothermal energy systems by enabling real-time simulation, monitoring, and optimization. This paper presents an open-architecture digital twin framework developed for low-enthalpy deep hydrothermal geothermal plants, designed to support co-simulation and operational optimization using real-time data streams. The proposed framework integrates physics-based models, machine learning models, multi-source data streams, and advanced control algorithms to enhance operational efficiency, asset reliability, and decision support. Through two representative case studies, we demonstrate the applicability of the framework for (i) optimizing the performance and condition monitoring of electrical submersible pumps (ESPs), including predictive failure detection and plant-level operational improvements, and (ii) integrating large language models (LLMs) into the digital twin environment to provide rapid contextual access to operation and maintenance (O&M) documentation. For the latter, we describe the implementation of a retrieval-augmented generation (RAG) workflow for processing unstructured text data and supporting intelligent troubleshooting. The open-source design lowers adoption barriers and facilitates collaboration across the geothermal sector, providing a basis for future extensions toward autonomous geothermal plant operation.

1. INTRODUCTION

Geothermal energy has emerged as an alternative source for sustainable heating and power supply due to its dispatchability, low carbon footprint, and long-term availability. Despite these advantages, the operation of geothermal assets remains complex (Shoeibi Omrani and de With, 2025). Wells, surface facilities, and pumping systems are exposed to harsh thermo–chemical environments, fluctuating demand profiles, and resource uncertainty, all of which can negatively affect system reliability and economic performance. Typical operational challenges include injectivity loss, corrosion and scaling, and performance degradation of equipment. These issues drive up maintenance costs, shorten component lifetime, and may lead to safety or environmental risks if not managed proactively.

Digital twin technology has gained significant traction across industrial sectors as a means to couple real-time data with numerical models for monitoring, simulation, and decision-support including in the geothermal sector (Hashemi et al., 2025). In essence, a digital twin establishes a continuously updated digital representation of a physical asset, integrating sensor data, first-principles models, and data-driven analytics to support forecasting and scenario evaluation. Within the geothermal sector, early efforts have demonstrated that digital twin approaches can enhance operational efficiency, support predictive maintenance, and improve plant reliability without large infrastructure investments (e.g., Siratovich et al., 2022; Mahmoud et al., 2023). In related geothermal applications, digital twins have been used to optimize ground heat exchanger operation, assist in steamfield management, and support chemical and scaling control using augmented reality and process models (Buster et al., 2021; Chityori et al., 2024).

Despite promising advances, existing geothermal digital twins often remain tightly scoped, focusing on either subsurface simulations, equipment diagnostics, or chemical process modeling, without a broader architectural framework that supports co-simulation, and domain interoperability. Most of the current digital twin developments in the geothermal sector focusses on the power production systems (Siratovich et al., 2022) and less on direct-use heating applications. One of the proposed digital twin frameworks for geothermal reservoirs for direct use application mainly focusses on reservoir management of the heat production and transport and surface aspect are not included (Voskov et al., 2024). With the rapid progress in machine learning and large language models, there is increasing potential to combine physics-based geothermal models, data-driven surrogates, condition monitoring, and intelligent documentation retrieval within a unified operational environment of a digital twin solution.

This paper introduces an open-architecture digital twin framework designed for low-enthalpy deep hydrothermal geothermal plants. The framework integrates (i) physics-based thermal–hydraulic models, (ii) machine learning surrogates and predictive monitoring tools, and (iii) large language model (LLM)-based knowledge support via retrieval-augmented generation (RAG). The flexibility of the architecture enables applications ranging from ESP optimization and failure prediction to operator support through semantic access to operation and maintenance (O&M) documentation. The benefits of the framework are demonstrated through two case studies. The first focuses on production monitoring, and the second showcases the integration of LLMs into the digital twin environment to provide rapid and contextaware access to technical documentation, improving troubleshooting speed and decision support. Together, these examples underline how modern digital twin platforms can help geothermal operators reduce operational costs, enhance asset reliability, and move toward more autonomous and data-driven operation of geothermal facilities.

The main contributions in this article are as follow:

- A modular digital twin architecture that supports co-simulation between physics-based and data-driven models, while ingesting real-time plant and sensor data through standardized interfaces.
- Integration of large language models (LLMs) using a retrieval-augmented generation (RAG) pipeline to provide intelligent access to technical documentation, maintenance records, and operational procedures, enhancing operator decision support.
- Open-source tooling and accessible implementation, lowering the barrier for adoption and encouraging collaboration and extension across the geothermal and broader subsurface energy sectors.

2. METHODOLOGY

2.1 System Architecture

A geothermal digital twin can be defined as an information system that exposes end-users to a dynamic digital replica of the geothermal system, ranging from subsurface reservoirs and wells to above-ground plant components, continuously informed by measurements and constrained by physical laws. The purpose of such an environment is to couple real-time observational data with numerical models and data-driven intelligence to support scenario analysis, operational decision-making, and predictive maintenance. Within this context, efforts in the broader digital twin community highlight the importance of combining diverse observational data streams with models through interoperable and interactive workflows. Inspired by these developments, the geothermal digital twin presented in this work follows a modular, domain-specific conceptual framework comprising the following main pillars: (i) assimilation and integration of multi-source data streams for real-time decision support, (ii) physics-based and data-driven thermal-hydraulic modelling, (iii) large language models for extracting and integrating relevant information from text data supporting the decisions, and (iv) visualization layer to visualize the relevant indicators and parameters for operational decisions. The proposed architecture for the digital twin is shown in Figure 1.

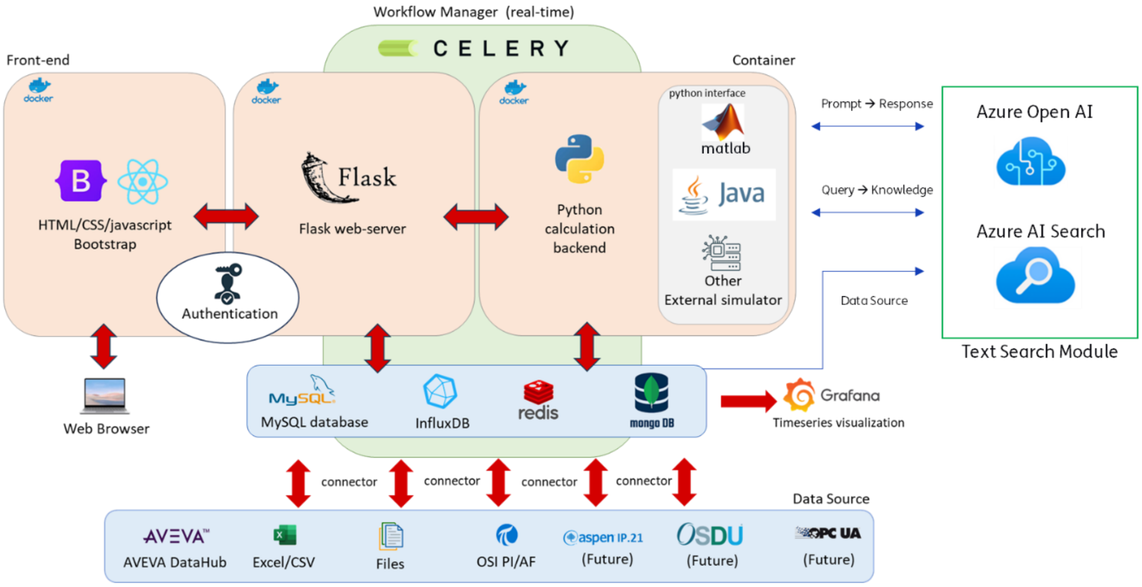


Figure 1: Proposed architecture for the Open digital twin framework of geothermal plants.

The first pillar concerns the integration and assimilation of multi-source data streams, forming the foundational data layer of the digital twin. These data sources span several operational domains, including (i) real-time measurements from wells and surface plants collected via SCADA and field instrumentation (e.g., pressure, temperature, flow rate, pump frequency), (ii) inputs of the models from measurements, logging, and simulators, and (iii) historical maintenance logs, inspection reports, and failure data from asset management systems. In addition to live SCADA feeds, the data layer also accommodates batch or on-demand uploads through structured files (e.g., CSV, JSON, Excel) and external databases or APIs, enabling interoperability with existing operator workflows. To harmonize these heterogeneous sources, lightweight connector modules ingest streaming or file-based data and route them into local data stores specialized for time-series and unstructured content. In typical deployments, process variables are archived in a time-series database such as InfluxDB, which supports efficient querying and downsampling for visualization and forecasting tasks. Meanwhile, semi-structured and text-based assets, such as maintenance records, metadata, or documents, are stored in a document-oriented database such as MongoDB, enabling fast retrieval for both analytics and LLM-based reasoning. Currently, ChromaDB instance is employed to store and retrieve the vectorized (embedded) text chunks for the LLM application which is further explained in the respective section.

The second pillar, forward physical modelling, involves simulating key geothermal processes such as reservoir response, wellbore hydraulics, heat extraction, and electrical submersible pump (ESP) behavior. These physically based models generate predictive states of the geothermal system (e.g., pressures, temperatures, flow rates, pump head) and provide the reference outputs required for optimization and performance benchmarking. The models can run offline for scenario evaluation or online for co-simulation with plant data, and are exposed through standardized interfaces to enable coupling with other components of the twin. In addition machine learning methods for model emulation, predictive health monitoring, and online forecasting are employed. Here, data-driven surrogates replicate the behavior of complex physical models at significantly reduced computational cost, enabling faster-than-real-time prediction, optimization loops, condition monitoring, and "what-if" simulations. ML models ingest operational data (e.g., SCADA signals, maintenance logs, pump frequency adjustments) to predict equipment degradation, detect anomalies, and infer latent states difficult to observe directly (such as pump efficiency decline or stages of failure progression).

The third pillar, focusses on large language models (LLMs) with retrieval-augmented generation (RAG) workflows which act as a knowledge fusion layer by enabling semantic search and contextual reasoning over unstructured O&M documentation. Together, these capabilities form the digital thread, the end-to-end linkage between raw data, models, and decision-making. In the current workflow, a specific AI platform is shown (Azure) but the choice of the LLM and the employed framework is dependent on the user. In the following chapters, the integration of an open-source LLM with a RAG workflow is explained.

The fourth pillar of the geothermal digital twin is the visualization layer, implemented as a web-based user interface that connects operators, engineers, and data systems through an intuitive and interactive environment. This layer provides real-time access to plant status, performance metrics, predictive forecasts, and simulation outputs, and acts as the primary medium through which users can explore scenarios, inspect system states, and interact with model components. The visualization layer should support dynamic dashboards and time-series exploration, component-level health views. By leveraging modern web technologies and API-based data streaming, it ensures that complex model outputs from physics-based simulations, machine learning surrogates, and knowledge retrieval modules are presented in an interpretable and actionable form. As a result, the visualization layer enhances situational awareness, accelerates decision-making workflows, and facilitates cross-disciplinary communication between reservoir engineers, operators, and maintenance teams.

The resulting framework establishes the computational "engine" for a geothermal digital twin that can be used for simulation, monitoring, predictive maintenance, and operator support. Its modularity enables interoperability between physics-based simulation tools, machine learning models, optimization routines, and knowledge-based AI, laying the groundwork for supporting operational decisions in geothermal plants.

2.2 Modelling Components

The digital twin developed in this work integrates multiple modeling components to represent geothermal system behavior across the subsurface-to-surface chain. Physics-based models are used to simulate reservoir dynamics, wellbore flow, and surface nodal conditions, ensuring physically consistent predictions under varying operational scenarios. A numerical solver orchestrates the coupling between these modules and handles simulation of fluid and energy transport. In parallel, data-driven machine learning models are incorporated for real-time monitoring, performance estimation, and forecasting, enabling fast inference where computationally expensive physics-based simulations are not practical for operational support. A brief description of these components is provided in the following section.

2.2.1. Reservoir, well and surface nodal models

A key element in the modelling of the geothermal plant is to model the pressure-flow relations across the different component in the plant, from the reservoir to well and to the surface equipment. The relation between flow rate and pressure in each component is derived by the flow physics that happen in each component. For the operational aspects, the dynamics of the reservoir which are often governed by long-term behaviour is not considered and the initial pressure point that is considered in the digital twin solver is from the near wellbore. For this purpose, inflow performance relationship (IPR) is used. IPR describes the well flowing bottomhole pressure (P_{wf}) as a function of the measured production rate (Q). This relationship is crucial for designing optimal production strategies and managing reservoir performance. The bottomhole pressure (P_{wf}) is linked to the average far field reservoir pressure (P_{res}). A typical IPR graph is illustrated in Figure 2.

The line slope (K), also known as productivity index (PI), is the ratio between flow rate and well drawdown that can be defined as:

$$K = \frac{Q}{P_{res} - P_{wf}} \tag{1}$$

The Vertical Lift Performance (VLP) describes the bottomhole pressure as a function of flowrate in the tubing. The VLP depends on various factors including well depth, well trajectory, tubing size, water cut, gas-to-water-ratio (GWR), and PVT fluid properties. The boundary condition, if there is no Electrical Submersible Pump (ESP), is the wellhead pressure (P_{wh}). If an ESP is installed, the intake pressure of the ESP is used. In order to calculate the pressure drop along the tubing, two correlations are used based on fluid phase: single-phase or two-phase. The total pressure loss is governed by gravitational and frictional pressure drops. The gravitational pressure drop is calculated as function of local gravity and well tubing inclination. The frictional pressure drop is proportional to the square of the flow velocity and inversely proportional to the pipe diameter, as described by the Darcy-Weisbach equation. For two-phase flow in an inclined pipe, an additional parameter, liquid holdup, must be considered. Liquid holdup is dependent on the flow angle and is categorized into three horizontal flow patterns: Segregated, Intermittent and Distributed. Detailed equations are presented in Beggs and Brill (1973).

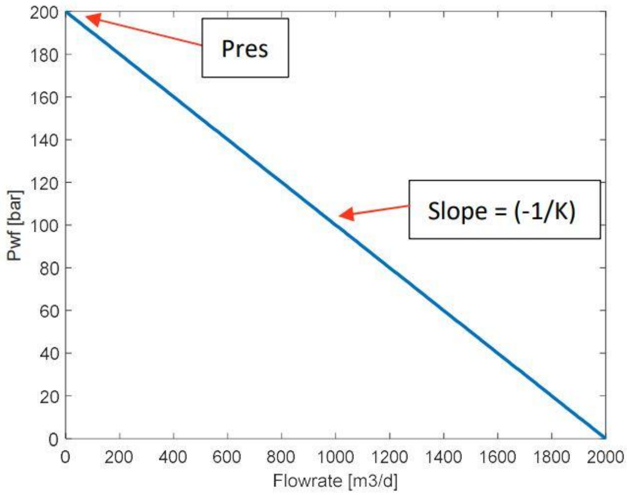


Figure 2: Inflow performance relationship for a constant productivity index.

2.2.2 ESP and pump models

ESPs are installed in production wells to provide additional lift when the reservoir pressure is not sufficient to bring the fluid to the surface. They increase pressure between the pump intake (inlet) and discharge (outlet). The schematic in Figure 3 illustrates a typical pressure profile across a geothermal production well equipped with an ESP.

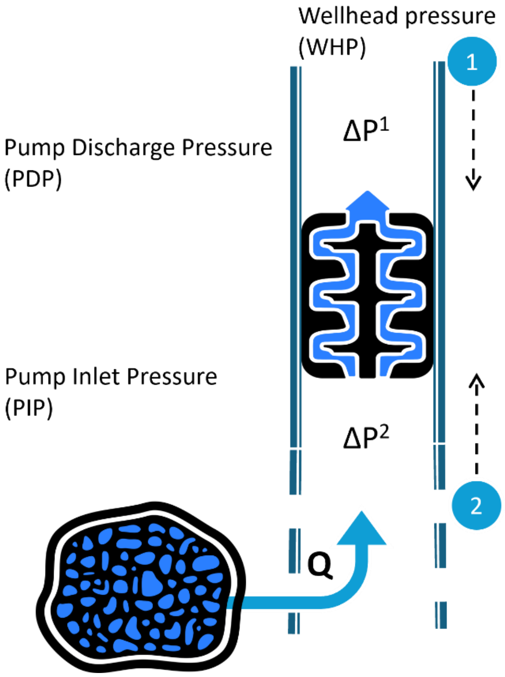


Figure 3: Schematic of the pressure nodes to perform nodal analysis with ESP.

This model is built upon manufacturer-provided correlations that relate the pump head, $\Delta p = (p_{esp,in} - p_{esp,out})$, to the flowrate q^p and pump frequency f_{esp}. The correlations are often found in the pump curves and enable the prediction of pump performance under varying operating conditions. The relationship was fitted by using a polynomial regression calibrated on pump curve data. The required pump power, E^{el}_{esp} [W], is calculated as:

$$E^{el}_{esp} = \frac{q^p \Delta p.}{\eta_{esp}}, \tag{2}$$

where a typical efficiency of the ESP is assumed to be $\eta_{esp} = 0.65$. This relation highlights the dependence of pump power requirement on the operational settings and efficiency, as defined by the pump performance curves. For monitoring and forecasting the ESP performance, a set of machine learning models are used which are further described in the next section.

2.2.3 Solver description

The proposed digital twin employs a modular solver architecture that enables flexible assembly of geothermal plant configurations and direct ingestion of real-time sensor data. Geothermal plants can be composed of standard components (e.g., filters, booster pumps, ESPs, heat exchangers, and consumer-side equipment), which may differ across installations. To support this variability, each physical component c is implemented as an independent input–output module forming a building block of the overall flow network. Conceptually, a module represents a deterministic mapping between inlet and outlet thermodynamic states:

$$\mathcal{M}_c: \{p_c^{in}, T_c^{in}, q_c^{in}\} \rightarrow \{p_c^{out}, T_c^{out}, q_c^{out}\}, \tag{3}$$

where p, T, and q denote pressure, temperature, and mass (or volume) flow rate. For each component mass conservation is imposed. Modules are linked to form two thermally coupled circuits: a primary loop comprising the reservoir and production–injection wells, and a secondary loop that distributes heat to the user. Energy exchange at the heat exchanger satisfies:

$$\begin{aligned} & q_c^{in} = q_c^{out}, \qquad\qquad \forall c \in plant \\ & \dot{m}_p c_p\left(T_p^{in} - T_p^{out}\right) = \dot{m}_s c_p\left(T_s^{out} - T_s^{in}\right) \end{aligned} \tag{4}$$

where subscripts p and s represent the primary and secondary flow, respectively. A key feature of the solver is its ability to operate as a co-simulation engine, allowing reservoir, wellbore, and surface models to exchange boundary conditions at run time. For example, bottomhole pressures computed by a reservoir simulator can serve as inputs to the wellbore model, which subsequently passes wellhead conditions to the surface network. This loose coupling allows each submodel to use its optimal numerical scheme while ensuring thermodynamic and hydraulic consistency at the interfaces. The steps for the co-simulation is described as follows:

- Determine primary-loop flow conditions: the steady-state flow rate of the primary loop (geothermal production side) ($\dot{m}_p$) is obtained by solving the hydraulic balance across the production-to-injection path. This step uses the relevant component models (e.g., reservoir inflow, wellbore lift, and pump characteristics) to satisfy pressure continuity and mass conservation.
- Propagate pressures along the primary loop: given $\dot{m}_p$, nodal pressures are computed for all downstream components in the primary circuit. This yields updated wellbore pressures (e.g., bottom-hole and wellhead) for both the producer and injector, consistent with the component input–output relations.
- Determine heat-exchanger inlet temperature of the primary loop: The inlet temperature to the primary side of the heat exchanger T_p^{in} is calculated by propagating measured temperatures and thermal losses along the primary loop, optionally incorporating live sensor data from the secondary side when applicable.
- Compute thermal coupling across the heat exchanger: the discharge temperatures of both circuits (T_p^{out}, T_s^{out}) are obtained by enforcing the heat-exchanger energy balance using $T_p^{in}\ and\ T_s^{in}$ and the flow rate on the demand side ($\dot{m}_s$).
- Update downstream thermal states: With T_p^{out}, T_s^{out} determined, temperatures at all downstream nodes and components of both circuits are updated via the input–output module relations, completing one solver cycle.

Depending on the task of the model, sensor data (such as wellhead pressures, pump intake temperatures, and flow meter readings) can be continuously incorporated as boundary conditions, enabling the solver to update component states, propagate system responses, and compute energy balances in real time. This allows the digital twin to respond to live operational conditions rather than static assumptions, supporting online tasks such as forecasting, anomaly screening, and scenario evaluation. The modular and co-simulation-ready architecture ensures that additional components, control strategies, or external simulators can be integrated without restructuring the numerical core, providing extensibility across diverse geothermal plant designs.

2.2.4 Machine learning and data-driven models

Machine learning and data-driven models are being employed more frequently in digital twins due to their ability to model complex, nonlinear physical processes using historical and real-time data. Geothermal operations involve coupled thermo-hydraulic-mechanical behaviors that could be difficult to represent analytically or simulate at high temporal frequency. Meanwhile, modern geothermal plants increasingly deploy sensing infrastructure that generates continuous streams of operational data. Machine learning models offer a way to exploit these data streams to learn system dynamics directly, providing fast and accurate surrogate models that can support real-time operational decisions such as anomaly detection, performance estimation, and condition forecasting, while avoiding the computational burden of high-fidelity simulations.

Within the developed geothermal digital twin, a suite of machine learning models has been trained to monitor Electrical Submersible Pump (ESP) performance parameters such as motor vibration, current profiles, intake/discharge pressures, and fluid temperature. These models enable early detection of abnormal events, operational drifts, and emerging failure signatures, thereby enhancing the reliability and situational awareness of geothermal operations. The design, implementation, and validation of these ESP monitoring models, including their feature engineering, training strategies, and benchmarking against real field data, are described in detail in Shoeibi Omrani et al. (2025), where high-resolution results on anomaly detection and condition monitoring are presented (Octaviano et al., 2022).

In the present paper, we extend the monitoring framework by integrating performance assessment with failure forecasting capabilities using a Histogram-Based Gradient Boosting Classifier. The Histogram-Based Gradient Boosting Classifier (HGBC) is an ensemble learning method that builds decision trees on discretized (histogram-binned) feature values, enabling efficient gradient boosting on large tabular datasets. By optimizing a differentiable loss function through additive tree model updates, it captures complex nonlinear relationships while remaining computationally scalable. The proposed approach consists of (i) constructing labeled operational windows that represent healthy and pre-failure states, (ii) extracting statistical and temporal features from sensor streams, (iii) training the HGBC model using gradient-boosted ensemble learning, and (iv) evaluating forecasting skill in terms of early warning capability. Preliminary results are shown in the case study section.

2.3 Large language model and RAG

To enable intelligent access to unstructured operation and maintenance documentation, we integrated large language models within a retrieval-augmented generation (RAG) workflow. The objective of the system is to allow users to query technical reports (e.g., well test reports, ESP documentation) in natural language and obtain grounded answers supported by citations from the source documents.

At runtime, the RAG workflow proceeds as a two-step process (Figure 4):

- Retrieval: The user's query is embedded into the same vector space using Snowflake Arctic Embed. A similarity search is then performed against the ChromaDB index to identify the most relevant document chunks. These retrieved passages form the contextual evidence used to answer the question.
- Generation: The retrieved context and the user query are passed to the LLM through Ollama. The model synthesizes the information and generates a natural language answer grounded in the retrieved content rather than solely in its latent training knowledge.

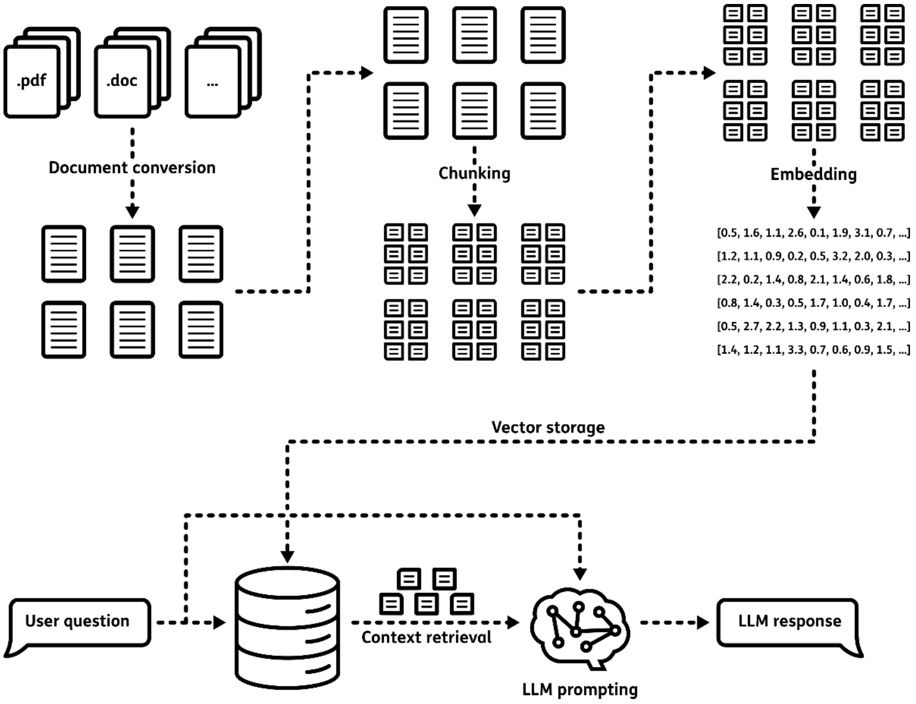


Figure 4: Workflow for the retrieval augmented generation employed in the digital twin architecture.

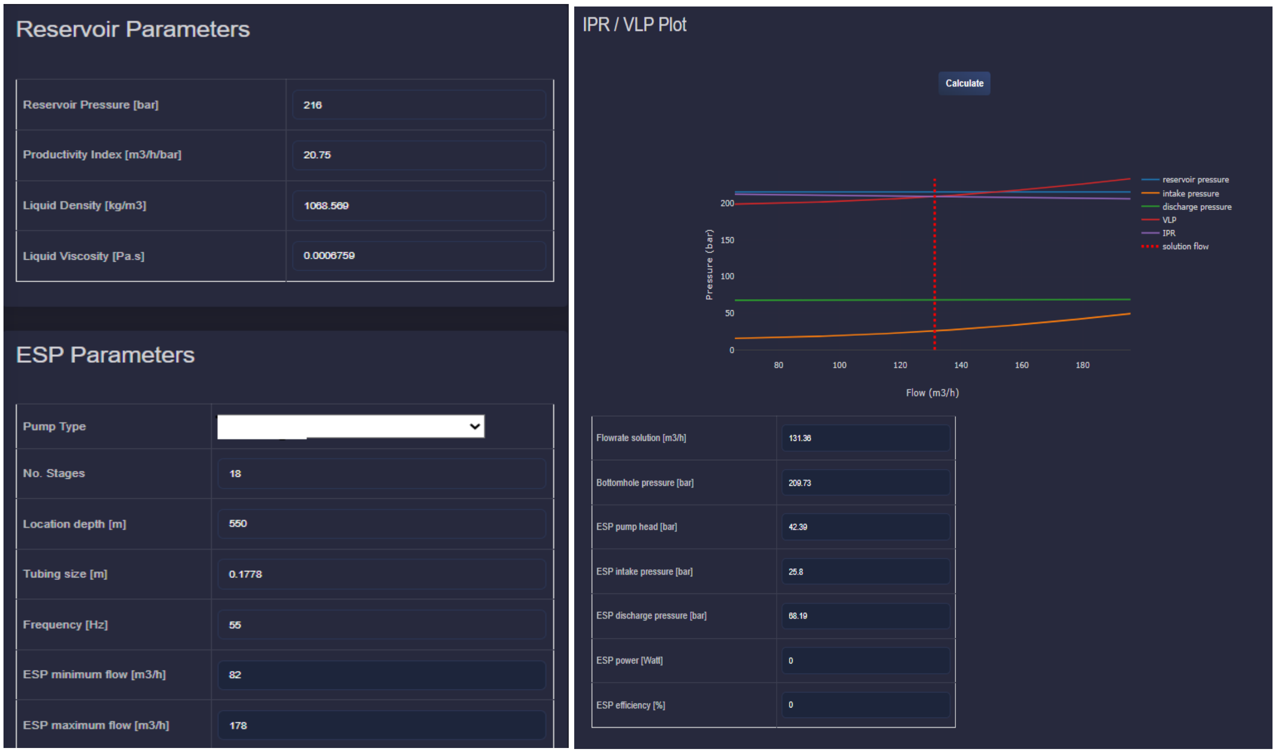
The LLMs are hosted locally using Ollama, providing an efficient runtime environment for serving open-weights language models. Two generation models were evaluated: Llama 3.2, a lightweight model optimized for low-latency interaction, and Mistral-Nemo, which demonstrated higher technical precision at the cost of increased computational demand. This comparative setup enabled exploration of accuracy–response time trade-offs relevant for geothermal operational use cases. The retrieval component is implemented using ChromaDB, which serves as the vector storage backend for document embeddings and similarity search. Prior to indexing, technical reports are pre-processed into fixed-size text chunks and transformed into numerical embeddings using Snowflake Arctic Embed, an open embedding model designed for high-quality semantic retrieval. These embeddings are stored in ChromaDB and indexed for rapid similarity search during query time. To enhance transparency and validation, the interface returns both the generated answer and a list of the specific document chunks (citations) used during retrieval. These citations allow users to inspect source material, verify correctness, and explore related content as required. While the system substantially accelerates access to information and supports troubleshooting workflows, it is designed as a decision-support tool, and expert oversight remains essential for high-consequence operational decisions. Overall, the combination of Ollama-hosted LLMs, Snowflake Arctic embeddings, ChromaDB vector retrieval, and citation-level transparency demonstrates a practical workflow for leveraging open-source language models in geothermal digital twins and O&M knowledge systems.

3. CASE STUDIES

3.1 Production monitoring and forecasting

The first case study focuses on production monitoring and demonstrates how integrated well/reservoir performance modelling supports real-time assessment of geothermal production. In this application, a coupled VLP–IPR framework is continuously evaluating production and can be updated using live operational data from the well and surface facilities. By ingesting real-time measurements (e.g., pump frequency, wellhead pressure, temperature), the model computes updated inflow and lift performance curves, enabling near real-time estimation of the production rate under the current operating conditions of both the wellbore and reservoir. This provides operators with an always-current view of well performance rather than relying on infrequent testing or static design assumptions. Figure 5 illustrates an impression of the production monitoring application, showing how sensor data and model outputs are combined to track the evolving operating point of the well. Beyond monitoring, the framework also allows rapid what-if analysis and co-simulation. As shown in Figure 6, modifying the ESP frequency by +5 Hz provides an updated prediction of the expected change in production rate, based on the latest inferred state of the reservoir and wellbore. Such analysis can be executed within seconds and requires no field intervention.

These simulation capabilities further enable virtual evaluation of operational adjustments or equipment redesign prior to deployment in the field. For example, by modifying ESP type, number of stages, or installation depth within the model, operators can assess their impact on system performance and verify whether anticipated gains justify the intervention. This supports both operational decision-making and system design workflows, lowering uncertainty and minimizing the cost and risk associated with field testing or equipment replacement.



Reservoir Parameters

Reservoir Pressure [bar]	216
Productivity Index [m3/h/bar]	20.75
Liquid Density [kg/m3]	1068.569
Liquid Viscosity [Pa.s]	0.0006759

ESP Parameters

Pump Type	
No. Stages	18
Location depth [m]	550
Tubing size [m]	0.1778
Frequency [Hz]	55
ESP minimum flow [m3/h]	82
ESP maximum flow [m3/h]	178

IPR / VLP Plot

Flowrate solution [m3/h]	131.36
Bottomhole pressure [bar]	209.73
ESP pump head [bar]	42.39
ESP intake pressure [bar]	25.8
ESP discharge pressure [bar]	68.19
ESP power [Watt]	0
ESP efficiency [%]	0

Figure 5: Production monitoring application interface for VLP-IPR calculations.

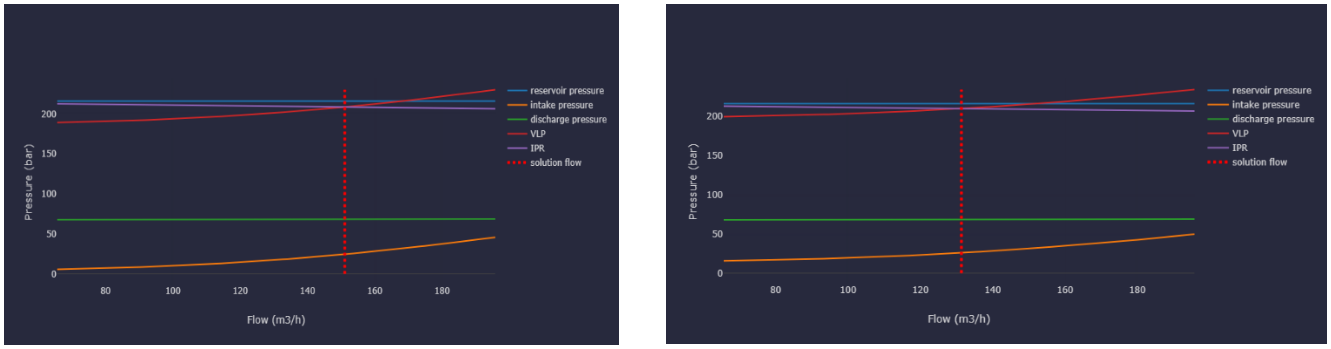
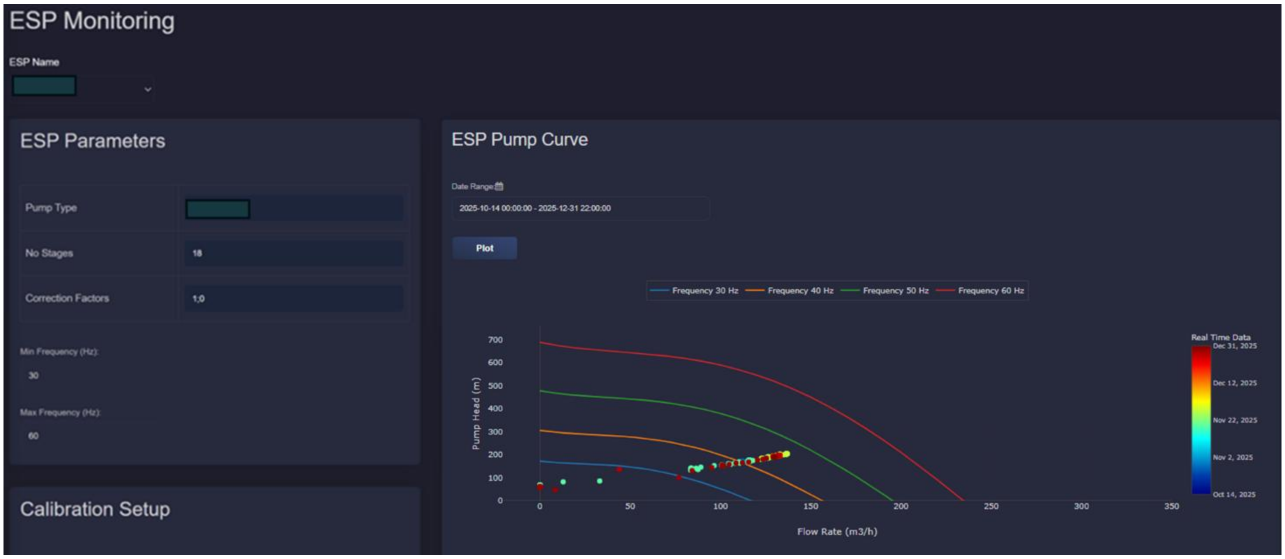


Figure 6: visualization of the production monitoring application results with simulated impact of ESP frequency by 5 Hz on the production rate based on the current well and reservoir condition (left) ESP frequency of 60 Hz and (right) ESP frequency of 55 Hz.

For the second case study, ESP monitoring and performance forecasting is demonstrated, due to the critical role of ESP reliability and performance in geothermal plants (Shoeibi Omrani et al., 2021). One of the common ESP performance monitoring indicators is based on the ESP pump curve from the manufacturer, plotted as head versus flowrate. An example is shown in Figure 7. On top of this, real-time operating data is added to show actual pump performance. Since there is often no discharge pressure sensor in the geothermal ESP systems, the pump head is estimated using two values:

- Inlet pressure, measured directly by a downhole sensor.
- Upper pressure difference (∆P1), calculated using the VLP (as described in Production well performance) from the top of the ESP to the wellhead.

From these, we estimate the actual pressure increase provided by the pump. This gives us the real-time pump head, which is plotted with the flowrate on top of the manufacturer pump curve to monitor the current status of the production versus the nominal pump performance.



ESP Monitoring

ESP Name: [redacted]

ESP Parameters

Pump Type	[redacted]
No Stages	18
Correction Factors	1,0

Min Frequency (Hz): 30

Max Frequency (Hz): 60

Calibration Setup

ESP Pump Curve

Date Range: 2025-10-14 00:00:00 - 2025-12-31 22:00:00

Figure 7: An example of ESP monitoring based on ESP curve.

In addition, a comparison between different values calculated from the model will be demonstrated, such as head, discharge pressure, and inlet pressure. The models need to be evaluated differently to show the comparison between different parameters of the ESP. As an example (Figure 8), for inlet pressure comparison the following parameters are compared:

- Measured Inlet Pressure: from the ESP sensor at the pump intake.
- Calculated Inlet Pressure: calculated from IPR (using reservoir pressure and flowrate, as described in the previous section) and VLP (from reservoir to pump location, using bottomhole pressure and pressure drop in the lower stage of the ESP, ∆P2).

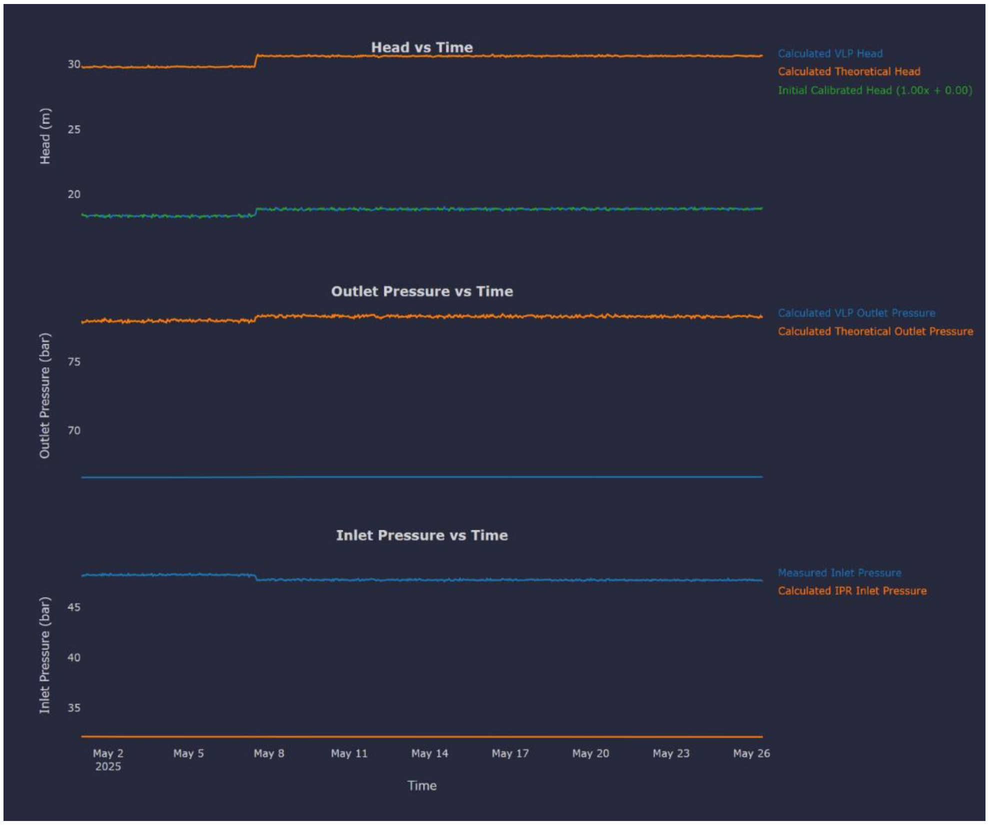


Figure 8: Estimated ESP inlet, discharge pressure and head based on different approaches in the digital twin as an input for model calibration or anomaly detection.

The trained model is a Histogram-Based Gradient Boosting Classifier (HistGradientBoostingClassifier) from sklearn library (Pedregosa et al., 2011) for predicting ESP failures. It was trained using a grid search with cross-validation on production data from a specific geothermal well in the Netherlands. The training process involved preprocessing the sensor data by filtering unphysical values and applying a 1-day rolling average to reduce noise, and standardizing seven key features: vibration, brine flow rate, ESP motor temperature, wellhead temperature, wellhead pressure, intake pressure ESP, and discharge pressure ESP. The used dataset includes 3 failure timestamps and we have used them with different configurations to find the best score. For the "TestTrainValidation" configuration for 30-day prediction, the data was split temporally: the test set consisted of two months before the first failure event, the training set included two months before the second failure plus two random months without failures, and the validation set contained two months before the third failure plus the remaining timestamps. The target variable was binary, labeling data points as "Close to Failure" (1) if they occurred within 30 days before a known failure, and (0) for the opposite situation. Grid search optimized hyperparameters including learning rate, maximum depth, number of iterations, regularization, and class weighting, using a predefined train-test split with ROC AUC as the scoring metric. ROC AUC (Receiver Operating Characteristic, Area Under the Curve) is a performance metric for binary classification models that measures how well the model separates the positive and negative classes across different probability thresholds. The ROC curve plots the true positive rate versus the false positive rate, and the AUC summarizes this curve into a value between 0 and 1, where higher values indicate better class separability.

The selected model showed a good performance across all the configurations, with optimal hyperparameters including a learning rate of 0.05, maximum depth of 10, 300 iterations, 63 maximum leaf nodes, minimum samples per leaf of 10. That resulted in a ROC AUC score of 0.99996 on the training set which shows a very good separation of failure and non-failure cases in the training data. On the test set, the model shows high performance with a ROC AUC of 0.98, as the model hyperparameters are set in such a way to maximize the score on the set. Also, the validation set has a ROC AUC of 0.89, which is as expected lower than the test performance. This might be due to some temporal variability in failure patterns. These results are preliminary, and the validity and generalizability of the classifier must be assessed using a larger number of failure events and diverse failure types.

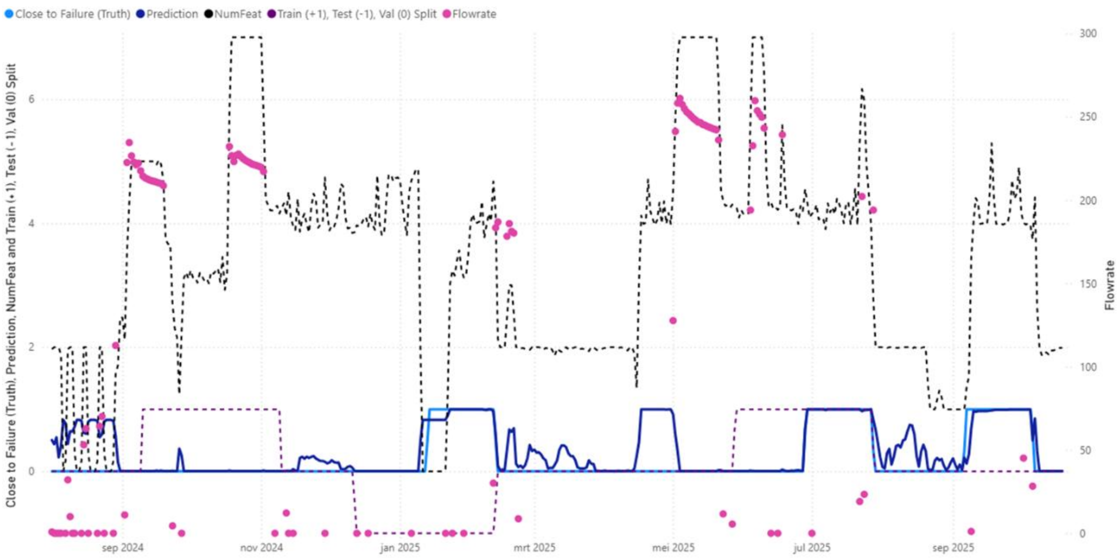
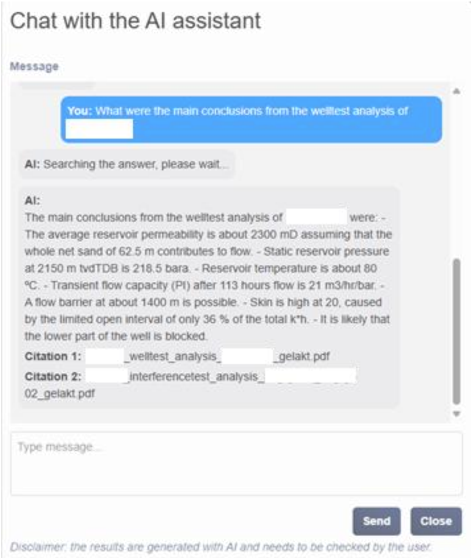


Figure 9: Time-series plot of result of the ESP failure prediction model for 30 days using the "TestTrainValidation" configuration. The left y-axis shows 4 time series: "Close to Failure (Truth)" the light blue line shows the binary ground truth labels, where 1 indicates data points within 30 days before a known failure and 0 otherwise, "Prediction" the dark blue line shows the predicted probability of being close to failure, ranging from 0 to 1, "NumFeat" the black dashed line indicates the number of available features at each timestamp to show data completeness, and "Train (+1), Test (-1), Val (0) Split" the purple dash line indicates the temporal data split, where +1 represents training data, -1 for test data, and 0 for validation data. The right y-axis shows the "Flowrate" the pink scatter points for the brine flow rate in the production data.

3.2 LLM-Enhanced O&M knowledge support

The first demonstration of the large language model is directed at the task of extracting and synthesizing information from operational and maintenance documents, which are often long, heterogeneous, and difficult to search manually. To enable traceable and domain-aware responses, the model is integrated within a retrieval-augmented generation (RAG) workflow that ingests well test reports, maintenance logs, and related technical documentation. Figure 10 shows an example interaction, where the user poses a question regarding specific data contained within well test reports. The system retrieves the relevant passage from the source documents and formulates a coherent answer grounded in the retrieved context. This capability reduces manual search time and supports engineers in rapidly locating critical operational details.



You: What were the main conclusions from the welltest analysis of [redacted]

AI: Searching the answer, please wait...

AI:
The main conclusions from the welltest analysis of [redacted] were: - The average reservoir permeability is about 2300 mD assuming that the whole net sand of 62.5 m contributes to flow. - Static reservoir pressure at 2150 m tvdTDB is 218.5 bara. - Reservoir temperature is about 80 °C. - Transient flow capacity (PI) after 113 hours flow is 21 m3/hr/bar. - A flow barrier at about 1400 m is possible. - Skin is high at 20, caused by the limited open interval of only 36 % of the total k*h. - It is likely that the lower part of the well is blocked.

Citation 1: [redacted]_welltest_analysis_[redacted]_gelakt.pdf

Citation 2: [redacted]_interferencetest_analysis_[redacted]
02_gelakt.pdf

Disclaimer: the results are generated with AI and needs to be checked by the user.

Figure 10: Example of AI chat assistant retrieving information from operation and maintenance document. The example is on well test data analysis, user asks a question about highlights of the well test reports and the AI chat assistant based on RAG provides key reservoir and inflow parameters from two different documents.

The second test was performed on a fluid characterization document for a geothermal well. The analysis was done on fluid characterization report due to the confidentiality of the O&M documents that could not be shown in this paper. The workflow is using a large language model to extract key entities such as chemical species, thermodynamic properties, suspended solids, and gas composition, along with their operational implications. The extracted entities were then translated into a structured representation by mapping chemical and physical properties to known degradation mechanisms and mitigation strategies in geothermal brine systems. This step needs to be done by the expert or operator to define the causal links between the extracted parameters and risks, e.g. high chloride leads to corrosion. This process resulted in the construction of a directed knowledge graph, where nodes represent fluid properties, operational risks, affected assets, and mitigation actions, and edges encode causal or functional relationships derived from domain knowledge and the content of the fluid report.

The resulting knowledge graph is shown in Figure 11. The result visualizes how specific fluid characteristics propagate through physical and chemical mechanisms to create operational risks. For example, high chloride and CO_2 content connect to corrosion pathways affecting tubing and ESP components, while elevated hardness and calcium link to carbonate and sulphate scaling risks in surface and downhole equipment. Suspended solids are associated with plugging in filters and injection systems, and these mechanisms ultimately point toward performance degradation at the equipment level. In this way, the graph provides an intuitive bridge between laboratory chemistry results and plant reliability concerns, making dependencies and failure pathways visible at a glance.

Operationally, the knowledge graph serves as a decision-support layer that can inform treatment programs, materials selection, maintenance planning, and the optimization of running conditions. By linking risks to specific mitigation strategies such as corrosion inhibitors, scale inhibitors, CRA material upgrades, filtration improvements, or P–T envelope adjustments, the graph enables engineers to rapidly evaluate corrective actions. When integrated into a digital twin or LLM-powered retrieval system, the graph can be queried semantically (e.g., “What affects ESP performance?” or “How can sulphate scaling be mitigated?”), providing fast contextual reasoning for operators and maintenance teams. This transforms static chemical data into actionable operational intelligence that supports predictive maintenance, reduces downtime, and enhances asset reliability.

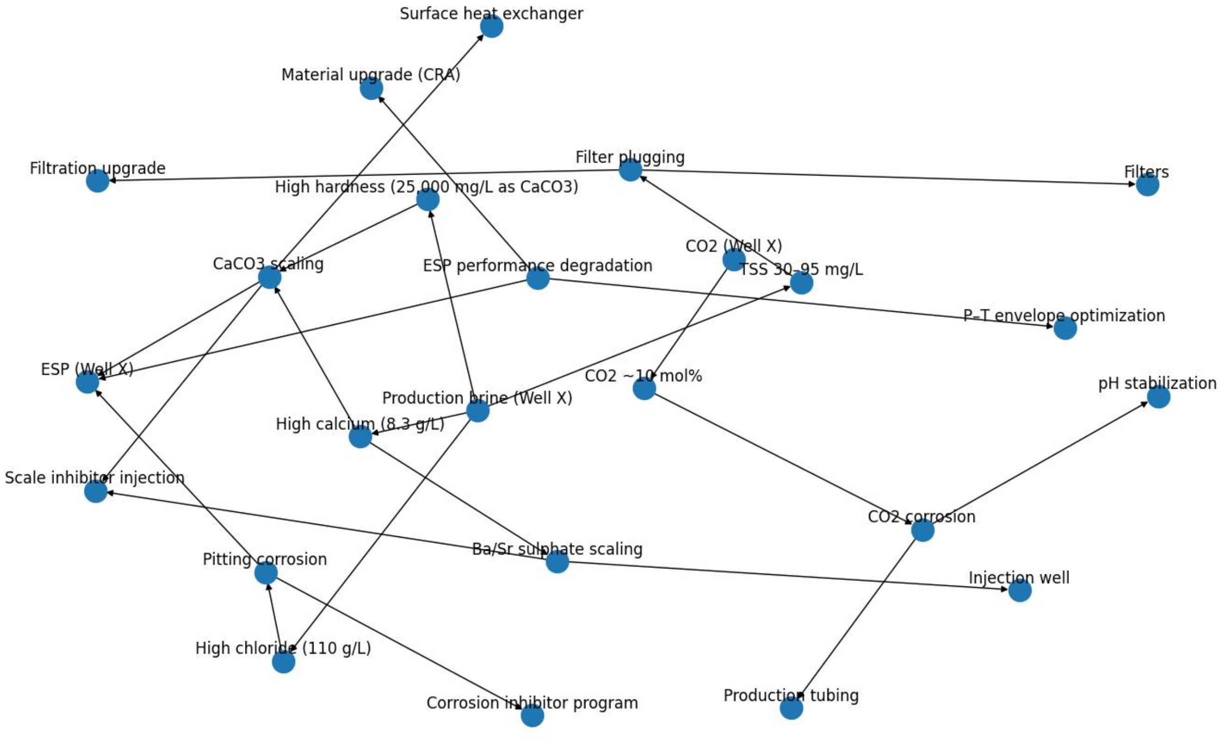


Figure 11: An example of knowledge graph for a geothermal well based on the fluid composition to define chemistry, resulted risks and mitigation measures, the fluid composition information was extracted using LLM.

4. CONCLUSION

This work presented an open-architecture digital twin framework for deep hydrothermal geothermal plants, designed to enhance operational decision-making through real-time data, co-simulation, and advanced analytics. By combining physics-based models, machine-learning models, and multi-source operational data, the framework aims to enable informed decision making for operational workflows. Two case studies illustrate the applicability of the approach. The first demonstrated condition-based monitoring and performance forecasting of electrical submersible pumps (ESPs), highlighting how data-driven and physics-based methods can support failure detection and forecasting. The results serve as a preliminary outcome of the workflow and it requires additional data to be further validated and generalized. The second demonstration integrated a retrieval-augmented generation (RAG) workflow with large language models (LLMs) to intelligently process operation and maintenance documentation, offering rapid contextual access to historical well test and fluid characterization reports and supporting troubleshooting workflows. Together, these examples demonstrate the technical feasibility and practical value of embedding advanced analytics and AI within geothermal digital twins.

A key aspect of this development is that realizing the full potential of digital twins in the geothermal sector requires more than model development at the asset level; it demands sector-wide progress in data standardization, platform interoperability, and knowledge sharing. Sector-scale digital twin ecosystems depend on (i) standardized data formats and schemas for wells, reservoirs, pumps, surface plants, and ancillary systems; (ii) shared data connectors and APIs for streaming operational data from operators, vendors, and regulators; and (iii) validated and reusable physics-based and data-driven model libraries. Such components allow learning to compound across projects, enable benchmarking across assets and regions, and reduce the time-to-knowledge for new geothermal developments. The resulting infrastructure supports faster design cycles, improved reliability, better risk management, and ultimately accelerates geothermal deployment.

For these reasons, the framework was developed with an open-source and open-architecture philosophy. By lowering adoption barriers and encouraging contributions from operators, vendors, and researchers, the platform could provide a foundation for collective innovation and sector learning. We envision future extensions toward additional functionalities and automating several manual operational tasks, integration of predictive control strategies, cross-operator benchmarking, and sector-level digital services. The results provide inputs to the vision that open digital twin ecosystems can play a central role in accelerating geothermal development and ensuring that operational experience and technical knowledge are not siloed, but shared and amplified across the geothermal community.

Future work will focus on extending the digital twin's functionalities to cover additional operational processes (e.g., corrosion, scaling, and equipment degradation) and incorporating broader monitoring and diagnostic capabilities. Further real-time testing and validation on field data will be required to assess robustness and operational reliability under diverse operating conditions. In addition, scenario analysis and optimization modules will be introduced, enabling operators to evaluate alternative actions and control strategies using the existing forecasting and decision-support algorithms (Shoeibi Omrani et al., 2025).

ACKNOWLEDGEMENT

The authors thank Sander Maat from HVC for providing the data used in this study, offering valuable feedback throughout the work, and reviewing the manuscript.

REFERENCES

Buster, G., Siratovich, P., Taverna, N., Rossol, M., Weers, J., Blair, A., Huggins, J., Siega, C., Mannington, W., Urgel, A., Cen, J., Quinao, J., Watt, R., & Akerley, J. A new modeling framework for geothermal operational optimization with machine learning (Gooml). Energies, 14(20). https://doi.org/10.3390/en14206852 (2021).

Chityori, A., Byiringiro, J. B., Ndeda, R., & Gathitu, B. Towards Digital Twin and Augmented Reality Modelling to Mitigate Silica Scaling in Geothermal Plants. SSRG International Journal of Mechanical Engineering, 11(5), 70–86. https://doi.org/10.14445/23488360/IJME-V11I5P107 (2024)

Hashemi, L., Palochis, D., Poort, J., Octaviano, R., Shoeibi Omrani, P. Advancing Geothermal Operations with Digital Twin Technology: Proactive Monitoring and Optimization. Paper presented at the SPE Europe Energy Conference and Exhibition, Vienna, Austria, (2025). doi: https://doi.org/10.2118/225592-MS

Mahmoud, D., Tandel, S. R. S., Yakout, M., Elbestawi, M., Mattiello, F., Paradiso, S., Ching, C., Zaher, M., & Abdelnabi, M. Enhancement of heat exchanger performance using additive manufacturing of gyroid lattice structures. International Journal of Advanced Manufacturing Technology, 126(9–10), 4021–4036 (2023)

Octaviano, R., Dussi, S., de Zwart, H., Shoeibi Omrani, P., van Pul-Verboom, V., Elewaut, K., van Schravendijk, B., Model-based monitoring of geothermal assets, WarmingUP program final report (2022)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Shoeibi Omrani, P. & de With, G. Production and operation of geothermal systems. In Geothermal Energy Engineering (2025), pp. 261-302. Elsevier.

Shoeibi Omrani, P., Yang, Y., Rijnaarts, H.H.M., and Shariat Torbaghan, S.: Real-time model-based condition monitoring of geothermal systems under uncertainties – Case study on electrical submersible pumps, Geoenergy Science and Engineering, 249, (2025), 213775, doi:10.1016/j.geoen.2025.213775.

Shoeibi Omrani, P., Van der Valk, K., Bos, W., Nizamutdinov, E., Van der Sluijs, L., Eilers, J., Pereboom, H., Castelein, K., & Van Bergen, F. Overview of Opportunities and Challenges of Electrical Submersible Pumps ESP in the Geothermal Energy Production

Systems. In SPE Gulf Coast Section Electric Submersible Pumps Symposium (p. D031S007R003). https://doi.org/10.2118/204524-MS (2021)

Shoeibi Omrani, P., Egberts, P. J., Rijnaarts, H. H., & Torbaghan, S. S. (2025). Geothermal plant operation and control under demand uncertainties. Renewable Energy, 124805.

Siratovich, P., Buster, G., Taverna, N., Rossol, M., Weers, J., Blair, A., Huggins, J., Siega, C., Mannington, W., Urgel, A., Cen, J., Quinao, J., Watt, R., & Akerley, J. GOOML-Finding Optimization Opportunities for Geothermal Operations. PROCEEDINGS, 47th Workshop on Geothermal Reservoir Engineering, 1–10. (2022)

Voskov, D., Abels, H., Barnhoorn, A., Chen, Y., Daniilidis, A., Bruhn, D., Drijkoningen, G., Geiger, S., Laumann, S., Song, G., Vardon, P., Vargas L., Verschuur, E., and Vondrak, A.: A research and production geothermal project on the TU Delft campus: initial modelling and establishment of a digital twin. Proceedings, 49th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA (2024).