PROCEEDINGS, 47th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 7-9, 2022
SGP-TR-223

Data Curation for Machine Learning Applied to Geothermal Power Plant Operational Data for GOOML: Geothermal Operational Optimization with Machine Learning

Nicole Taverna[1], Grant Buster[1], Jay Huggins[1], Michael Rossol[1], Paul Siratovich[2], Jon Weers[1], Andrea Blair[2], Christine Siega[3], Warren Mannington[3], Alex Urgel[3], Jonathan Cen[3], Jaime Quinao[4], Robbie Watt[4] and John Akerley[5]

[1]National Renewable Energy Laboratory (NREL), Golden, CO 80401, USA

[2]Upflow Limited, Taupo 3330, New Zealand

[3]Contact Energy Limited, Wairakei 3352, New Zealand

[4]Ngati Tuwharetoa Geothermal Assets Limited, Kawerau 3169, New Zealand

[5]Ormat Technologies Inc., Reno, NV 89519, USA

Nicole.Taverna@nrel.gov

Keywords: data curation, machine learning, data-centric AI, data pipeline, GOOML, power plant operations

ABSTRACT

Geothermal Operational Optimization with Machine Learning (GOOML) is a transferable and extensible component-based geothermal asset modeling framework that considers complex steamfield relationships and identifies optimization prospects using a data-driven approach to physics-guided, data-centric machine learning. This framework has been used to develop digital twins that provide steamfield operators with operational environments to analyze and understand historical and forecasted power production, explore new steamfield configuration possibilities, and seek optimal asset management in real world applications.

To create, test, and apply the GOOML framework, diverse time-series datasets spanning multiple years were sourced from various geothermal power plant components within several complex real-world geothermal operations. These operations are based in the United States and New Zealand and include a variety of technologies, end-uses and configurations, collectively covering nearly all relevant operating conditions for modern geothermal fields. Datasets were acquired from multiple sources to ensure that machine learning experiments generalized properly to various operating conditions. It was found that the data varied in quality, format, and completeness. To ensure consistency between the various datasets, a standardized data curation process was developed to reliably streamline data preparation.

This paper will discuss best practices as learned from the GOOML data curation process which takes the following steps: 1) acquisition of large quantities of data from power plant operators, 2) digestion of data to gain an initial understanding of what is included, 3) data transformation, which includes converting the data into a standardized machine-readable format so that they can be visualized, quality checked, and cleaned, 4) quality assurance and quality control, involving identification of significant data gaps and apparent anomalies through mapping of data features to real world componentry via the GOOML historical model, followed by discussion with modelers and power plant operators to identify additional data needs and to resolve issues, 5) use in machine learning algorithms, and 6) repetition of steps one through five until all data needs are met and data are deemed suitable for producing trustworthy modeling results which may be disseminated, ideally along with the curated dataset. This iterative process is focused on improving the quality of the data rather than tuning machine learning model parameters and supports a shift towards a more data-centric philosophy as a means for improving real-world applicability of geothermal machine learning projects.

1. INTRODUCTION

1.1 Data Curation

Data curation may be thought of as the process of "caring" for the data, including organizing, describing, cleaning, enhancing, and preserving data for scientific use (Knight, 2017). For the purposes of this paper, the data curation definition is expanded to include data acquisition, data digestion, and data transformation, in addition to the described quality assurance (QA) and quality control (QC) measures followed by data use or dissemination. These processes are especially relevant to large real-world datasets, in which sensors malfunction, network connections fail, portions of datasets are misplaced or lost, and human intervention introduces potential bias or error. These occurrences introduce gaps, anomalies, erroneous values, and bias in datasets which must be addressed to maximize data utility. Even exemplary data require standardization in terms of units, sampling rates, and other factors.

Data curation is not only essential for preserving high-value data in a meaningful way for future use but is also useful for deriving the utmost possible information from data prior to use in machine learning algorithms. A robust data curation process is a critical part of any successful machine learning project that uses real-world data, because without it, results of machine learning experiments lack important context and applicability to the real world. Most of the modern practices surrounding data curation in the geothermal industry are on an ad-hoc basis or are specific to individual organizations. Standardization of this process would ease collaboration between organizations

by improving data quality, accessibility, and usability, which is of increasing importance with the rise of data-driven technological methods in recent years.

1.2 Geothermal Operational Optimization with Machine Learning (GOOML)

Geothermal Operational Optimization with Machine Learning, or GOOML, is a modeling framework for creating digital twins of geothermal power plants. It is based on hybrid data-driven thermodynamics components-based systems models. Instead of relying on extensive theoretical and semi-empirical relationships, it enforces simple first-principal thermodynamic mass and energy conservation equations and then uses real historical plant data for training machine learning models to describe the thermodynamic operations of various steamfield components (see Buster, et al., 2021 (a) for additional information).

GOOML was developed based off of and tested on three vastly different operating geothermal fields, including: Ormat's McGinness Hills Geothermal Power Plant in Nevada, United States, Ngāti Tūwharetoa Geothermal Assets' (NTGA's) Kawerau Geothermal Power Plant in New Zealand, and Contact Energy's Wairakei Geothermal Power Plant in New Zealand. McGinness Hills is the least complex of the three in terms of the number of power plant components, with several single-phase production and injection wells, and seven binary Ormat Energy Converters (OECs). Kawerau is slightly more complex than McGinness Hills with several two-phase production and injection wells, multiple separators, six industrial direct use partners and one binary plant. Wairakei is the most complex of the three geothermal plants, with many two-phase production wells and injection wells, several dry steam production wells, numerous separators, eleven steam turbines, and two binary units, which have all been added at various stages throughout the field's 60+ years of operation. In addition, Wairakei has several swinging wells, which can have their flow redirected in two or three different directions whenever needed. Despite the differences in the geothermal power plants investigated, GOOML has accurately described all three geothermal systems without relying exclusively on theory whilst reducing the required engineering design details. GOOML has the potential to improve overall system understanding and to increase capacity factors and overall system utilization. These results are described in more detail in Buster, et al., 2021 (a), but GOOML's success is in part credited to a more data-centric approach.

1.3 Data-Centric Artificial Intelligence (AI)

Data-centric artificial intelligence (AI) is a movement focused on improving the quality of data used to train models rather than tuning model parameters or architecture to improve the accuracy of modeling results. In the presently dominant model-centric approach to machine learning, as much data as possible is collected and a model is developed that is powerful enough to deal with noise and other problematic data. In this approach, the data is generally seen as fixed, and the model is iteratively improved until the desired performance is achieved. In contrast, the data-centric approach holds the model (mostly) fixed and the quality of the data is iteratively improved. It is often the case that improving the data rather than the model architecture and parameters results in a much more favorable value-to-effort ratio in real-world applications.

It is a common approach in the machine learning community to simply "add more data" to alleviate the detrimental effects caused by low-quality data rather than addressing the data quality issues directly. More recently, low-quality data has been identified as one of the root causes of the "proof-of-concept to production gap," which refers to the inability for machine learning models to succeed when applied in the real world (Press, 2021). "Data cascades," which are compounding events that cause negative, downstream effects when issues in the data exist, are triggered by conventional machine learning practices that undervalue data quality. Data cascades can be pervasive, invisible, and delayed, but are often avoidable. Generally speaking, data work is extremely undervalued in machine learning projects across all domains (Sambasivan, et al., 2021). This is also the case in machine learning projects applied to geothermal. In this paper, we attempt to reconcile this by laying a foundation for rigorous data-centric AI as applied to the geothermal industry.

1.4 Data Curation and Geothermal Operational Optimization with Machine Learning (GOOML)

The data curation portion of the GOOML framework, outlined by Figure 1, includes data acquisition, data digestion, data transformation, QA and QC, until data are thought to be suitable for use in machine learning models. This process involves repetition of the prior steps until modeling results meet desired performance criteria and are ready for dissemination, ideally along with the curated dataset. The data transformation and QA and QC steps of this process have been developed into a software suite. This iterative process caters to the idea of data-centric AI in an attempt to address the general lack of attention given to data curation in machine learning projects and to ensure that outputs are of the highest possible quality.

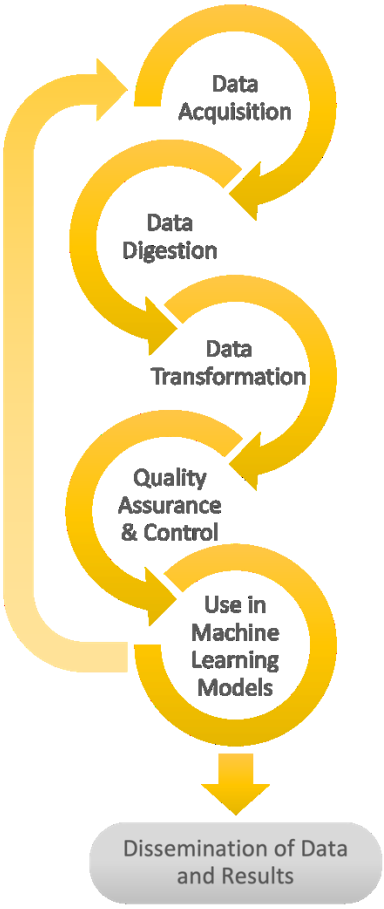

Figure 1: Graphic describing the data curation process used by the GOOML project.

2. DATA ACQUISITION

Data acquisition, or data transfer is the first step in any data curation process. This step can be challenging, especially when dealing with sizable datasets or datasets made up of an inconvenient number of files. Thoughtful data management on the part of the originators of the data helps to streamline this process. Data management platforms, such as the OSIsoft's PI system (OSIsoft, 2021), can aid data owners in file organization on a larger scale, but often the burden of file organization falls directly onto those collecting the data. In these cases, it is essential that 1) all relevant metadata and data are collected and properly documented, 2) files are organized in a logical manner, 3) files and directories are named in self-explanatory ways, 4) files are formatted in nonproprietary and widely accessible formats, and 5) data are stored in a way that minimizes risk of data loss. On top of that, some data are sensitive and data owners are often reluctant to share such data without assurance that it will be protected in transit and not susceptible to data breaches. Using reliable and secure data transfer and storage mechanisms can help assure data providers that their data will not be subject to data breaches.

2.1 GOOML Data Acquisition

The GOOML data acquisition process was different for each modeled field based on the availability of data, the source data format, and the size of the data. All data was stored in the Data Foundry (Weers, et al., 2020) to allow sharing of files between industry partners and the GOOML team. The Data Foundry is a file storage solution for Department of Energy (DOE)-funded research that provides secure, cloud-based storage and universal access to digital information, allowing collaboration between government agencies, national labs, universities, and private organizations. It is hosted by OpenEI, which provides access to open energy information, data, and resources, and is a preferred file storage solution for DOE-funded data prior to dissemination because of its adherence to DOE data management and cyber security protocols. Within the GOOML data acquisition process, some additional costs were incurred because some data were being moved halfway across the world (i.e., from New Zealand to the United States).

2.1.1 McGinness Hills Data Acquisition

When acquiring data from the McGinness Hills geothermal power plant, spreadsheets containing curated datasets were uploaded from Ormat's data historian to the Data Foundry. While this approach was straightforward and eliminated many of the issues associated with raw data (i.e., gaps, anomalies, and erroneous values), it limited the granularity and frequency of data able to be accessed.

2.1.2 Kawerau Data Acquisition

To acquire data from the Kawerau Geothermal Field, specified data were manually exported from the PI system which collects and stores operational data. Unnecessary columns in the data were deleted by hand, and data were manually uploaded to the Data Foundry. This approach allowed more granular and higher frequency data to be acquired but is infeasible for larger geothermal fields because of the manual nature of the sub-processes.

2.1.3 Wairakei Data Acquisition

In acquiring data from the Wairakei Geothermal Field, where the amount of data in question was too massive to manually manipulate and transfer (originally consisting of over 600 files consisting of more than 10 GB of uncompressed comma separated values (CSV) files in their raw form), select data was gathered from the TIBCO server (TIBCO, 2021) which is used by Contact Energy as a data analytics solution. At Wairakei, the TIBCO server is connected to the PI system. A series of automated data transfers were set up to up load data from the TIBCO server to the Data Foundry. This approach was the most streamlined and versatile of the three for the purposes of the GOOML project.

2.2 GOOML Data Acquisition Lessons Learned

One of the best practices learned from the GOOML data acquisition process is that it is essential to set up regular and reliable communication with the owners of the data early on. Under certain circumstances, such as a global pandemic that prevents in-person meetings, communication with power plant operators can be difficult. Setting up regular and reliable communication channels at the start of the project encourages discussion surrounding the estimated timelines for data delivery, allows questions to be asked about data contents and formats, and builds trust between parties, particularly when the data in question are sensitive. Another best practice includes oversight on automated data transfers. When automated transfers are left unchecked for too long, issues can arise that introduce gaps in datasets. If these are not caught early on, major problems with unclear causes can appear in the data. For example, during the GOOML historical data acquisition process, several data files were lost during the transfer, resulting in periodic data gaps which heavily impacted the training of GOOML machine learning models.

Storage-efficient file formats are very useful in streamlining the data transfer process. This reduces duplicative information in each file, keeping only what is necessary. For example, a format in which data features are the columns and time stamps are row indices is preferred to a format in which each row represents a unique combination of feature name and time index.

Lastly, if possible, all available data should be sent as a single set, rather than as subsets. This reduces the need for additional data requests that are unavoidable as more information is gathered about the system. Within the GOOML project, the data curation process became a recursive one in which data were acquired, digested, transformed, cleaned, and then more data were requested and acquired, starting the process over again. In an ideal scenario, this process would only occur once, although this is almost never the case in reality. Even seemingly unimportant data can aid in the modeling process (e.g., nearby alarms, electronic gate data, etc.), so it is difficult to determine what is useful prior to building the model. Often it is only during the refinement of the model that these data are identified as relevant. The best way to deal with this is to obtain all data and then simply not use what is not needed. If only a subset of the data is provided, there may be a disconnect between what data are provided and what data exist, where potentially useful data can go unrequested.

3. DATA DIGESTION

The next step in data curation should involve digestion of the data, which includes gaining an understanding of what is contained within the data. This step can be thought of as "getting to know the data" by spending time exploring the content, format(s), and source(s). Generally, this process consists of reading background material, mapping measurements to sensors and physical properties being measured, and conducting exploratory data analysis (EDA) to examine the quality of the data. In the case of geothermal power plant operational data, piping and instrumentation diagrams (PIDs), engineering drawings, and discussions with power plant operators are also useful sources. Some additional data digestion may occur after the data transformation step, which will be discussed in more detail in the next section.

3.1 GOOML Data Digestion

Within the GOOML project, the data digestion process took the form of first determining which metadata were included, then gaining an understanding of how that metadata plays into the interpretation of the data, and next methodically mapping data features to individual components as a means of understanding what is included in the dataset. The mapping of data features to componentry was aided by background material where available, PIDs, engineering drawings, and discussions with power plant operators.

3.1.1 McGinness Hills Data Digestion

The McGinness Hills geothermal power plant data were rather simple to digest, as the field is relatively small in terms of the number of power plant components and associated sensors. In addition, the historian-provided spreadsheets containing curated datasets meant that the data were already in an interpretable format which streamlined the intake of the data. However, the granularity of these data prevented deeper analyses and a comprehensive understanding of the field. The data that were obtained were determined to provide enough insight to move forward with the data curation and modeling processes, and following an initial look at the data, features were mapped to individual components as described above.

3.1.2 Kawerau Data Digestion

The Kawerau geothermal field operational data were provided to us in an interpretable format, however due to the field layout being more complex than that of McGinness Hills, additional effort was required to properly digest the data. Kawerau is considered more complex because it has a higher quantity of power plant components, it provides steam to several direct use industrial partners, its configuration has changed significantly in recent years, and on occasion, one of its separators receives two-phase fluid from a separate but neighboring power plant owner. Keeping all of this in mind, an initial understanding of the field was obtained, and decisions were made about which components (and their associated measurements) to exclude from GOOML modeling. Specifically, direct use (heat) applications and wells that were too new to have reliable historical data were excluded from GOOML analyses. After these decisions were made, measurements were mapped to relevant componentry as described above.

3.1.3 Wairakei Data Digestion

The Wairakei data digestion was the most complex of the three fields due to the massive quantity of power plant components, the large amount of metadata provided, and the format that the data arrived in. The format of these data included several metadata fields that did not prove to be useful for GOOML analyses. While these were dropped to make the data intake process less laborious, having obtained them initially helped in gaining a comprehensive understanding of which data were collected and how. Having excess data is always preferred over having insufficient data. There were also several metadata fields that needed to be used to derive useful quantities, such as the "expression" field that represented text descriptions of the well condition data, and the "applies from datetime" and "applies to datetime" which were used to derive an accurate time index reflective of updates made to data values following their collection. Subsequent to this initial data cleaning, data were carefully mapped to their associated components as described previously.

3.2 GOOML Data Digestion Lessons Learned

From the GOOML data digestion process, it was learned that it is imperative to gain a comprehensive understanding of the data as early as possible. This prevents improper interpretation of data which can skew results. Working closely with plant operators and subject matter experts in this effort is invaluable. In addition, doing this early on reduces the need to go back and adjust methodologies following their initial implementation.

Clear and open communication is essential. Keeping up regular communication with power plant operators and engineers makes understanding the data much more efficient than it would be without these conversations.

Lastly, it is helpful to obtain as much metadata and documentation as possible. This includes annual reports on operations, PIDs, engineering drawings, and supporting metadata used to derive useful values. This information allows a complete understanding of the field in question to be obtained.

4. DATA TRANSFORMATION

Data transformation mainly consists of converting non-standard data into a standardized and machine-readable format. The complexity of this process is dependent upon the original format of the data. In almost all cases, this is best carried out using a programming language that works well for data engineering, such as Python, to construct modular functions, which may be combined in various ways to produce flexible data pipelines.

4.1 GOOML Data Transformation

Within the GOOML project, a software suite was developed consisting of modular functions for use in transforming the data from various non-standard power plant operational data formats to the standardized GOOML format. The data were transformed in a way that allows efficient data QA and QC and eventual export to CSV files, which were in some cases split into more manageable time periods. This format includes a datetime index in ascending order down the vertical axis, and feature names as column headers.

The GOOML Data Curation software suite is currently designed to work for three raw data formats including the Ormat historian's curated format, the slightly modified PI system export format in which unnecessary metadata are deleted by hand, and the raw TIBCO export format. This software suite is built in a way that allows extensibility to other formats as well, with only minor modifications required.

4.1.1 McGinness Hills Data Transformation

The McGinness Hills data transformation process consisted of stitching data from multiple sheets within multiple Microsoft Excel files together into a single Pandas DataFrame, and then assigning machine-readable and standardized feature names. A "data dictionary" was created concurrently which includes each feature name, each feature's description, and each data feature's associated units of measurement. The data dictionary is useful for efficient data feature lookup. The data were then ready for the QA and QC process and for subsequent export to a CSV file.

4.1.2 Kawerau Data Transformation

Transformation of the Kawerau data included an initial step done by the power plant operators, which included dropping the unnecessary metadata columns from the data prior to acquisition by the GOOML team. Once received, the data were transformed by taking the feature descriptions and units of measurement rows out of the dataset and placing them into a data dictionary, along with the feature names. The data was then split into more manageable temporal chunks. Simply removing these additional rows and splitting the data into temporal

chunks was sufficient to transform the data into the standardized GOOML format prior to the QA and QC process and for subsequent export to CSV files.

4.1.3 Wairakei Data Transformation

The Wairakei data were the most complex to transform. The data were received in a format where a row exists for every unique combination of feature name and datetime index, with additional rows existing for every possible swinging and non-swinging well configuration and condition, and for updates made to the data following initial measurement. The raw data also included additional columns of metadata that were not relevant to GOOML. The first step was to remove the irrelevant metadata columns to reduce the size of the data. Within this dataset, updates made by power plant operators were included as additional rows and outdated rows of data were removed.

Next, the well condition data were parsed so that only the well condition valid for the largest portion of each time period was kept in the data. The feature value was updated to the text value for that condition rather than the percentage of time period that the condition was valid. A data dictionary was created including feature names, feature descriptions, and units of measurement. Then the data features used only for interpreting the datetime index ("applies from datetime" and "applies to datetime"), for parsing the well condition data ("expression"), and for creating the data dictionary ("unit of measure") were dropped as they were no longer relevant. After that, a function was applied to the data that dropped rows that had both a duplicated combination of datetime index and feature name, and a "not a number," or NaN, feature value. The function then dropped additional duplicated combinations of datetime index and feature name, keeping the first instance of the duplicated rows. This was done to make sure that NaN instances are not being kept over the actual correct values in places where duplicate rows exist, accounting for possible updates made to data following sensor malfunction or planned sensor outages.

At this point, the data exist in a slightly simplified version of the original format, where there exists a row for each unique combination of feature name and datetime index, and there exists one column each for datetime index, feature name, and feature value. The next step was to "unstack" the data or reformat it so that feature names exist as column headers and datetime indices exist as row headers, removing duplicative information. This format is much more condensed and more efficiently machine-readable.

An optional additional step is included that allows users to join additional files that are already in this more condensed and machine-readable format into the transformed dataset. This step was developed to account for supplemental data that are pulled directly from the PI system rather than from the TIBCO server. This additional data is also summarized in terms of feature names, feature descriptions, and units of measure and is appended to the data dictionary. The data were then split into temporal chunks ready for the QA and QC process and subsequent export to CSV files.

4.2 GOOML Data Transformation Lessons Learned

From the GOOML data transformation process, it was observed that there are significant benefits to object-oriented programming, or more specifically, building extensible blocks that can be assembled into a modular data pipeline as opposed to creating data transformation functions on an ad-hoc basis. The most obvious benefit is the reduction in duplicated effort by allowing users to recycle functions that have already been implemented when building a new data pipeline for a new data format. This methodology is similar to the classic Strategy Design Pattern which is a general way to enable algorithm selection at runtime (Gamma, et al., 1994).

Another benefit to the GOOML approach is that it improves readability and reproducibility. When data transformation scripts are written on an as-needed basis without focus on extensibility, they are often not well-documented or easy-to-understand by anyone other than the person who wrote them. This makes it more difficult for others to reproduce data transformation methods, which can either lead to improper implementation or require the user to go through the whole process from the beginning rather than building from what has already been done. Lastly, this approach makes debugging more efficient, as the code is broken into functions for individual tasks (Half, 2021). These are benefits not only to the GOOML team internally, but also to anyone wanting to reproduce the GOOML data transformation methods.

5. QUALITY ASSURANCE AND QUALITY CONTROL

The terms QA and QC are often used interchangeably, but they represent separate concepts. QA relates to preventing the production of defective or problematic data, while QC refers to detecting and addressing problematic data (USGS, 2010).

Most of the QA measures surrounding geothermal power plant operational data are handled by power plant operators. Additional QA can be done during the curation process as a means of ensuring that the process itself is not introducing issues in the data. This can take the form of plotting select data intermittently throughout the data curation process as a way of validating curation methods.

QC processes include measures to identify and address missing, anomalous, and erroneous measurements within the data. Issues in the data can be identified by using plots to visually inspect the data, calculating statistics to describe the data, or using physical relationships to validate data and ensuring that outputs are physically possible. Any identified issues can be repaired by dropping problematic data from the dataset, by supplementing with data collected from human-observation, or through the introduction of plausible synthetic data produced from trends or physical relationships in the data.

5.1 GOOML Data Quality Assurance and Quality Control

The GOOML QA process includes a series of visual inspections of the data either as Pandas DataFrame objects, Pandas Series objects, or plots of data values versus time to ensure that data transformation and QC processes have only the intended impacts on the data. The

majority of QA for power plant operational data applies to ensuring sensors are not malfunctioning, and therefore is usually handled by the power plant operators themselves, thus reducing the time required for QA by the GOOML data team.

Like data transformation, a GOOML QC software suite was built. The QC software suite includes functions to convert data from non-International System of Units (SI) units to SI units, round values to eliminate unrealistic precisions introduced by calculations, replace erroneous string values with NaN, update specific data values based on user-inputted mappings, clean up feature names to remove spaces or other problematic characters, and handle NaN values. It also provides functions to produce supplemental data from the original data. These supplemental data include condition data for turbines, separators, and wells reflecting whether these components are in service, shut, or being bypassed for maintenance (e.g., well workovers) or other reasons such as plant reconfigurations. These supplemental data are used to more easily and accurately model each power plant.

The GOOML QC process also includes functionality to map data features to real world componentry through building configuration files that reflect relationships between components and their associated data. These configuration files also pull in additional data related to separator dimensions, well tracer flow testing data, and other informative quantities that provide physical guidance, and are input into the GOOML historical model which produces a graph relational structure of the plant and steamfield components.

The GOOML historical model, discussed in more detail in Buster, et al., 2021 (a), includes more complex data QC functionality, including considerations for physical laws, such as conservation of mass and energy. In addition, plots of the data may be produced from the GOOML historical model and used to compare different measurements associated with a given component as a means of identifying apparent anomalies. For example, if a separator has a positive measured steam flow but a zero liquid flow, there is likely a problem with the liquid mass flow measurement. A similar process may be applied comparing data describing downstream components to data describing upstream components to identify apparent anomalies. For instance, if a steam turbine that only has one upstream separator appears to be producing zero power even though the separator has a significant positive mass flow, there may be an error in the power generation data. It is much easier to identify these relationships when the data is related by component, as is done by the GOOML model network graph structure.

Correction of problems in the data identified by the QC process begins with discussion with power plant operators and engineers to ascertain the sources of the problems, during which the best ways to resolve these issues is determined and additional data needs are identified. The QC process did not vary significantly by power plant; therefore, plant-specific methods are not described.

5.2 GOOML Data Quality Assurance and Quality Control Lessons Learned

Frequent QA is essential when constructing data transformation and QC processes for curation of a specific dataset. Without QA, functions applied to the data can have unintended consequences, producing problems in the data. Use of a standardized and functionalized data curation library can reduce this, as their methods are proven to have the intended outcomes on the data.

It was observed that the application of mass and energy conservation equations led to much higher quality two-phase well production data. Without asserting conservation of mass and energy relationships, there were no clear relationships in the production and two-phase separation data. However, after asserting these invariances by scaling two-phase flow estimates based on more reliable single-phase flow measurements, models could be reliably trained to extract high-quality relationships from the data. This union of physical laws and data-centric philosophy is a powerful tool when applying machine learning methods to power plants and should be prioritized.

The GOOML QA and QC process also demonstrated the benefits of the graph relational structure for complex systems such as large geothermal power installations. A relational structure allows an analyst to quality-control data using the relationships between components in the actual steamfield instead of examining the data in isolation. With the GOOML graph relational structure, one can query the single-phase flow out of a separator in relation to the estimated two-phase flow input, or quickly find any measurements from components that are physically downstream from the separator. Without a graph relational structure, an analyst has the onerous task of cross-checking system diagrams and looking up data feature names. It is much more intuitive to request a software to "give me the downstream component and its associated data".

Lastly, as discussed in previous sections of this paper, regular and open communication with data providers is key to success. Without this, and the accompaniment of detailed documentation, it is likely that sources of problems in the data would be misinterpreted, the layout of and changes over time to systems would be misunderstood, and additional data needs would not be addressed.

6. DATA USE IN MACHINE LEARNING MODELS

As discussed previously, within the data-centric approach to machine learning, the model is held relatively static while the data is improved iteratively in order to improve model performance until desired criteria are obtained. Some minor tweaks may be made to the model so that it better suits the data, but the focus is largely on making improvements to the data. Through preliminary modeling results, additional data needs are identified, and the data curation process is cycled through again.

6.1 GOOML Data Use in Machine Learning Models

Within the GOOML project, when the data were adequately prepared for use in machine learning models, they were organized into a graph relational data structure described previously and passed into machine learning hindcast and forecast models. Some additional QC was undertaken, specifically relating to the machine learning modeling outputs. This helped identify additional data needs. The process of using the data in machine learning models did not vary by power plant in any noteworthy ways, therefore plant-specific methods are not discussed.

6.2 GOOML Data Use in Machine Learning Models Lessons Learned

The GOOML project demonstrates that the graph structure is an efficient way to represent geothermal power plant operational data for use in machine learning models. This is due to the creation of a digital twin that maps components not only to their associated data, but also to other related components. This data structure choice allows efficient application of machine learning to the data and streamlines the identification of additional data needs.

The way data were used in GOOML reiterates the benefits of the data-centric approach to machine learning. Significant improvements were observed in the machine learning modeling results as improvements were made to the data. In experiments where the data were held static and more focus was placed on tuning model parameters, considerable overfitting was seen, models were less extensible, and results were generally less meaningful.

7. ITERATION THROUGH THE PRIOR DATA CURATION STEPS

Ideally, the data curation process described in this paper would only be carried out once. However, due to the exploratory nature of machine learning, it is inevitable that additional data needs are identified during the development of models. The number of repetitions required for this process depends on the complexity and quality of the data (i.e., more complex and lower quality data requires more repetition).

Additional data needs may require acquiring supplementary data to fully constrain the system, fulfilling a need by updating QC requirements, or simply addressing a newly identified or previously inadequately addressed data issue. Regardless, the need to repeat the data curation process is usually a result of an incomplete initial understanding of the data or the system, which is unavoidable when dealing with any real-world system and its associated real-world data. Therefore, we refer to the data curation process as an iterative and exploratory process that continues until understanding of the data, system, and model reaches a point where the desired machine learning model performance is obtained.

7.1 GOOML Iteration through the Prior Data Curation Steps

Several iterations of the GOOML models were required for all three plants. The best demonstration of the need for multiple ongoing iterations is found in the Wairakei model. This is due to the complexity of the Wairakei Geothermal Power Plant operation and the sheer quantity of data associated with it. This complexity was the reason Wairakei was selected as the first target, whereby the initial data curation process was developed from the Wairakei data and then expanded and applied to the Kawerau and McGinness Hills operational data.

8. DISSEMINATION OF DATA AND RESULTS

Limitations caused by IP provisions are an unfortunate but frequent reality when acquiring data from industry partners. Whilst IP is important to scientific progress and commercial processes, keeping data proprietary paradoxically limits scientific advancement. Open data is a key enabler in the advancement of research and scientific innovation.

Access to open data provides large volumes of data and information necessary for not only ML projects, but virtually all modern scientific research projects. When open data is available, the financial and time burden related to the procurement of that data is greatly reduced. This encourages interoperability, knowledge sharing, communication, and application of research outputs not only within the scientific and academic groups but also wider industry and community. As a part of this process, standardized minimum requirements are imposed for metadata in order to ensure that the data are presented in a meaningful way (Weers, et al., 2017, Taverna, et al., 2021).

Whilst open access to all data would be considered a perfect scenario, it is recognized that without the required investment in the procurement and ongoing maintenance of data, many high-quality datasets would not exist. This is because the scientific community alone often cannot bear the resource burden, in terms of infrastructure, people, and financial costs associated with data procurement and maintenance of high-quality datasets. Despite IP-related limitations, the tools that scientific research can provide industry and communities can be transformative. As demonstrated by the GOOML project, government support can facilitate the navigation between open and closed data access and enable tools such as GOOML to be created.

8.1 Dissemination of Data and Results from GOOML

Whilst commercially sensitive power plant operational datasets are unable to be released, an anonymized dataset was produced which includes all of the same features as a real-world geothermal power plant operational dataset, with artificial labels, values, and noise. Big Kahuna is an artificial geothermal power plant created for this purpose, consisting of single- and two-phase production wells, separators, and steam turbine, each with fictional yet realistic dimensions and data. This was included in the GOOML Geothermal Data Repository (GDR) submission (Buster, et al., 2021 (b)) to provide an example use case along with a releasable experimental dataset.

9. OVERALL BEST PRACTICES AND DISCUSSION

The GOOML project has identified the following best practices for data curation applied to any geothermal machine learning project:

- Frequent and open communication between the data science team and commercial partners is critical to a successful machine learning project.
- Data acquisition should include oversight measures, to ensure that data are not lost as a result of network failures or other problems and should be automated if dealing with a large amount of data in real-time.
- Comprehensive data digestion as early on as possible will provide detailed understanding of the data that helps to avoid improper interpretation of data which can skew results.
- Data transformation using extensible code blocks that can be assembled into a modular data pipeline is better than creating data transformation functions on an ad-hoc basis.
- QA measures should be undertaken regardless of source of data.
- QC is best carried out in a standardized way using a software suite consisting of extensible blocks or functions that can be assembled into a modular data pipeline (e.g., a strategy design pattern).
- For complex projects, graph data structures can be a useful way to represent the physical arrangement and corresponding relationships between data features.
- The data-centric approach to machine learning produces more meaningful and extensible results than the traditional model-centric approach.
- Enforcing realistic physical invariances (e.g., conservation of mass and energy) and identifying physical symptoms of anomalous data (e.g., a separator with steam flow but no liquid flow) are powerful physics-based methods for improving the quality of data in a data-centric workflow.
- Data curation is an ongoing and iterative process throughout the life of a project.
- Successful and useful machine learning models and their results should be shared.
- Where possible, input data should be made available with machine learning models and results; when IP sensitivity is a concern, data can often be cleansed or anonymized.

The points above are intended to support the standardization of data curation processes within the geothermal scientific community for all data types, not solely for geothermal power plant operational data.

Data quality has always been a core issue in the geothermal industry. Key investment and operational decisions are made using scientific data daily and globally. Realizing the significant gains that could be had from machine learning tools requires a renewed and greater focus on not only the collection of new data but also improving the quality of the data already at our disposal. This highlights the importance of a shift towards data-centric philosophies for machine learning within the geothermal industry.

10. CONCLUSIONS

As is apparent through the contents of this paper, data curation is a core component in determining the quality of machine learning results and the applicability of machine learning to real-world problems. Due to the exploratory nature of machine learning, data curation is an essential and iterative process throughout the life of a project. Frequent and open communication with project partners is key to this process. Modular data pipelines and standardization of processes and practices support the move to data-centric machine learning workflows, a welcome move by the geothermal community.

The GOOML project demonstrates the benefit of combining commercial partnerships and underpinning government support to foster scientific excellence and deliver real world gains. Standardized practices for geothermal data curation are essential to ensuring that this process is carried out in a successful manner. This concept plays into the movement surrounding data-centric AI, demonstrating the need for a greater focus on data work throughout the geothermal industry in order to provide a more favorable value-to-effort ratio in real-world applications of machine learning outcomes.

ACKNOWLEDGEMENTS

The authors of this paper would like to express our sincere gratitude to our commercial partners: Contact Energy, Ngāti Tūwharetoa Geothermal Assets and Ormat Technologies, Inc. for their support of this project. Without access to their data, expertise and institutional knowledge, this project would not have been feasible. We'd also like to thank Zim Aunzo of ZpA Consulting for providing his expertise in geothermal reservoir engineering applied to building the GOOML modeling framework.

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Geothermal Technology Office Award Number DE-EE0008766. This work was authored by Upflow, Ltd. and the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308 with funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy (EERE) Geothermal Technologies Office (GTO). The views expressed in the article do not necessarily represent the views of the DOE or the United States Government. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid- up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for United States Government purposes.

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., and Akerley, J. A New Modeling Framework for Geothermal Operational Optimization with Machine Learning (GOOML). Energies (2021), 14, 6852.

Buster, G., Taverna, N., Rossol, M., Weers, J., Siratovich, P., Blair, A., Huggins, J. GOOML Big Kahuna Forecast Modeling and Genetic Optimization Files. Geothermal Data Repository (2021).

Gamma, E., Helm, R., Johnson, R., Vissides, J. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley (1994).

Half, R. 4 Advantages of Object-Oriented Programming. Robert Half Talent Solutions (2021).

Knight, M. What Is Data Curation? DATAVERSITY (2017).

OSIsoft. The PI System. (2021).

Press, G. Andrew Ng Launches a Campaign for Data-Centric AI. Forbes (2021).

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M. Everyone wants to do the model work, not the data work: Data Cascades in High-Stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021), pp. 1-15.

Taverna, N., Weers, J., Huggins, J., and Scavo, R.J. Geothermal Data Repository (GDR) Data Management and Submission Best Practices. Geothermal Data Repository (2021).

TIBCO. TIBCO CLOUD Spotfire. TIBCO Software Inc. (2021).

United States Geological Study (USGS). Data Management (2010).

Weers, J., Anderson, A., and Taverna, N. The Geothermal Data Repository: Five Years of Open Geothermal Data, Benefits to the Community. Geothermal Resources Council Annual Meeting (2017).

Weers, J., Frone, Z., Huggins, J., Vimont, A. The Data Foundry: Secure Collaboration for the Geothermal Industry. Proceedings of the 45th Workshop on Geothermal Reservoir Engineering, Stanford Geothermal Program (2020).