Proceedings World Geothermal Congress 2023
Beijing, China, April 17 – 21, 2023

Initial observations from machine learning approaches using UK temperature data for mine water thermal (MWT)

Sally F. Jack[1], Neil M. Burnside[1], Catherine M. Hirst[2], Sara Mehrabi[2], and Zoe K. Shipton[1]

[1]Department of Civil & Environmental Engineering, University of Strathclyde, G1 1XJ, UK

[2]COWI UK Limited, Glasgow, G2 5RG, UK

sally.jack.2020@uni.strath.ac.uk

Keywords: Mine water thermal, shallow geothermal, machine learning, artificial intelligence, temperature data

ABSTRACT

Until now, machine learning has been underutilised in shallow geothermal and more specifically, mine water thermal applications. Mine water thermal is a low-carbon, self-replenishing solution to providing space heating and cooling yet does not have significant uptake in the UK relative to other European country counterparts. This slower development is partly due to a lack of understanding of heat movement and behaviour in existing and abandoned mine systems. Where machine learning could offer a different perspective to more traditionally used numerical modelling techniques is in its ability for algorithms to quickly synthesise and process large volumes of available data. In this paper, we present machine learning methods applied to a UK subsurface temperature dataset to begin to minimise some of the knowledge gaps currently hindering development. We collated publicly available temperature data from The Coal Authority, British Geological Survey, Glasgow UK GeoEnergy Observatory, Scottish Environment Protection Agency, and the North Sea Transition Authority alongside additional licenced data from The Coal Authority, to form a dataset with over 2.4 million datapoints from 800 distinct locations. Both k-means clustering and multiple linear regression algorithms are presented alongside preliminary results of three models where geothermal gradients are predicted. The average geothermal gradient predicted by these models was 22°C/km (model A) and 23°C/km (model B and C), which is lower than the average UK-wide geothermal gradient of ~26°C/km modelled by previous authors. Furthermore, mine water temperatures were predicted to be 2°C warmer than other groundwaters in unmined aquifers at the same depth. All models were statistically significant with R-squared values of >0.85. We suggest that the data is likely presenting skewed results due to the inclusion of temperatures taken from pumping boreholes and mine shafts that mixes the warmer water at depth with colder, shallower water, creating a more suppressed temperature with depth. The simple relationships between temperature and depth illustrated in this paper form the beginnings of a machine learning platform from which more features can be added to further understand heat flow in mines. This study has also highlighted the importance of data sharing: a national or even global temperature database of the mined subsurface would allow for the uncertainties surrounding mine water thermal resources to be more easily addressed.

1. INTRODUCTION

As one of the countries that signed both the 2015 Paris Agreement and 2021 Glasgow Climate Pact, the UK has committed itself to be net zero by 2050, meaning efforts towards the green transition must be increasingly rapid, purposeful, and just. The energy required for space heating (and increasingly, space cooling) contributes to 42% of the UK's energy demand each year – more than is used to produce electricity (20%) or for transport (38%). Approximately 90% of the required energy for space heating and cooling comes from the burning of natural gas (UK Government, 2021; Monaghan *et al.*, 2022). Therefore, finding alternative low-carbon solutions to fossil fuels to meet the demand for space heating and cooling is increasingly urgent. The UK Government outlined its UK Net Zero Strategy in 2021 which sets out a detailed economic plan for the country's investment into the green transition, including £3.9 billion towards decarbonising heat and buildings and recognition that geothermal energy could form an important part of this process.

Using the shallow subsurface for heating, cooling, and thermal storage has the potential to form part of the solution to decarbonising the UK energy mix and reaching net zero. Mine water thermal (MWT) uses warm water from networks of abandoned, flooded mines. Once abandoned, legacy mines usually flood with groundwater over time due to cessation of operational pumping leading to the production of naturally heated groundwater within the subsurface. This warm mine water has the potential to provide a low carbon source of heating, cooling, and thermal storage through various MWT systems. Mine workings represent both a store of potential thermal energy within the mine water and rocks, and an interconnected network of heat exchange pathways through which water and heat can be circulated under the ground surface (Burnside *et al.*, 2016). Once the mine water and heat are extracted, ground source heat pumps (GSHPs) are used to boost the low-grade temperature of the mine water to deliver it to an end-user for heating and/or cooling consumption before being returned to the mines (Figure 1). Abandoned coal mines lie below an estimated 25% of homes in the UK (The Coal Authority, 2022). The UK therefore hits a sweet spot when it comes to the balance between heat demand and resource location – there is a strong overlap between 9 out of the 10 largest UK cities and legacy mines where energy can be extracted for space heating.

Though there is a growing body of literature on the assessment of flooded mines and their potential use as a low-carbon heating resource, a detailed understanding of water and heat movement in flooded mine systems has thus far proved elusive. Much uncertainty exists over mine abandonment state, flooding history, and present-day hydrogeological behaviour of mine water systems. Furthermore, the extent to which multiple schemes tapping into the same mine system might hydrogeologically and thermally interact is unknown. As such, an artificial intelligence (AI) machine learning (ML) approach has been adopted and applied to publicly available data from UK coal mine areas to identify subsurface data knowledge gaps and to aid selection of key parameters for advancement of numerical modelling of heat flow within mine water systems. Model outputs will de-risk mine water geothermal developments, particularly for future scenarios where more than one scheme is permitted within a single mine system and scheme cross-interference has to be assessed. We aim to provide important information for policymakers, industry, and government to accelerate the rollout of mine water heating, cooling, and thermal storage in the UK and globally. In this paper, we present our initial ML findings and make further recommendations for complimentary further research.

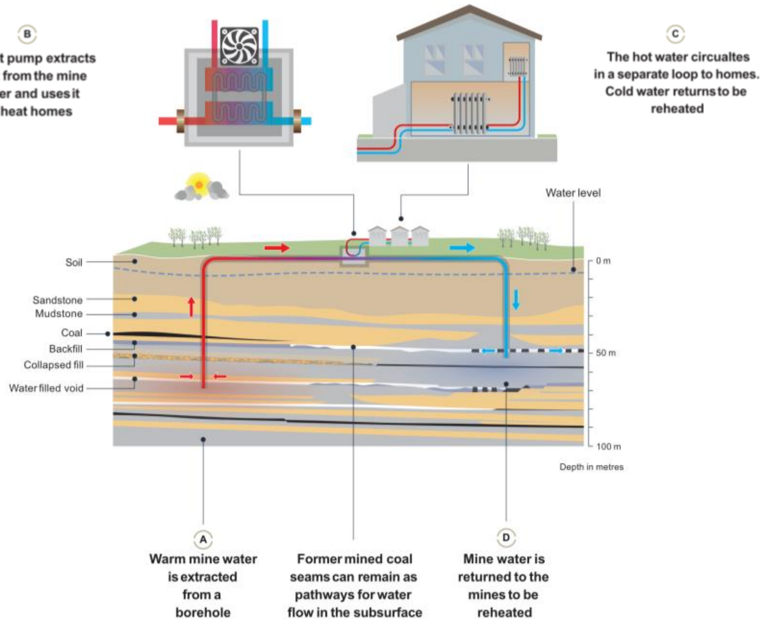

Figure 1: Typical mine water thermal well doublet used for heat extraction in which mine water is pumped to the surface, transferred to a GSHP where the heat is boosted to a suitable temperature for space heating before the cool water is reinjected to the mine to be reheated via natural geothermal flux and/or injected waste heat (UKGEOS, 2022).

2. BACKGROUND

In the UK, the need to further our understanding of MWT systems is pressing as the industry continues to grow. There are only two MWT schemes currently operational – large-scale commercial schemes both operating at Lanchester Wines, Gateshead, NE England. The schemes have been fully operational since 2020, comprising of two sets of abstraction and reinjection borehole doublets approximately 700 m apart from one another that are used to heat commercial warehouses with heat pump capacities of 2.4 MW and 1.2 MW respectively (Banks *et al.*, 2022). Elsewhere in the UK, potential mine water thermal projects are increasing (Figure 2) – in 2021 the Coal Authority estimated that there were at least 42 projects in their planning stages (Northeast LEP, 2021). Northeast England is becoming established as a leader in installed mine water geothermal systems in the UK. This is unsurprising given the extensive occurrence of Carboniferous Coal Measures which have been intensively mined in this area. As an example, the Seaham Garden Village is planning on utilising warm mine water from a nearby mine water treatment facility to heat 1,500 new homes by 2028. Furthermore, Gateshead Council have plans to use mine water to expand the existing Gateshead District Energy Scheme and deliver a heating capacity of 6 MW.

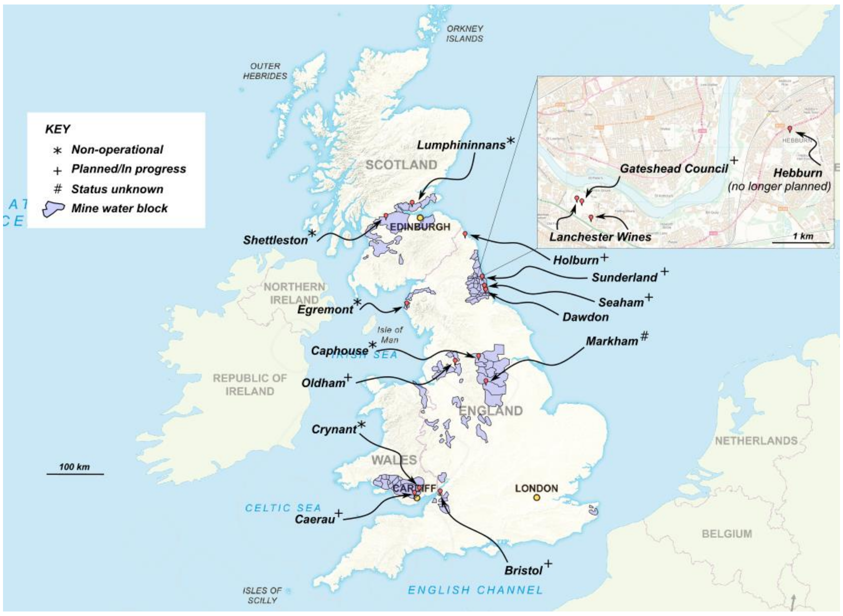

Figure 2: Locations of past, present, and future planned MWT schemes in the UK (as of January 2023) with The Coal Authority's hydrogeologically distinct 'mine water blocks' shown in purple.

The Coal Authority (TCA) currently only permits the operation of one geothermal scheme per distinct mine water block unless a case can be made for the safe utilisation of more than one scheme. A mine water block is characterised as a set of interconnected flooded collieries which have a continuous water level gradient and little to no hydraulic connectivity with other blocks. Currently, TCA have 174 defined mine water blocks (Farr *et al.*, 2020). There is little understanding on how adjacent schemes may interact with one another, particularly in the long-term. In areas such as Northeast England where several schemes are proposed within single mine water blocks, the importance of de-risking future scenarios is significant.

Because little is known about how heat interacts in mine water bodies, use of AI ML techniques applied to relevant data would allow for models to find patterns within the data that may otherwise not be obvious through more traditional methods of analysis. Besides the well-acknowledged correlation with depth, other factors that may influence heat flow in mine include the geology (stratigraphy, host rock thermal, hydrological, and mechanical properties), local hydrogeology (groundwater flow, composition), the mine network (connectivity, mine type, vintage and collapse state, fracturing) and engineering perturbations (pumping and cooling during mining, pumping and water treatment since abandonment). The sheer number of often interlinked factors that could influence heat transfer, heat retention, and water flow in a given mine means that the problem lends itself well to AI ML approaches given that a large enough database can be generated benefitting from these factors.

AI refers to the science and engineering of making intelligent machines, and in particular, intelligent computer programmes (Okoroafor *et al.*, 2022). The latest AI Index Report, which tracks and disseminates the use of AI in all sectors and industries of all scales across the world, highlighted the increasing activity in the field of AI. For instance, private investment in AI doubled between 2020 and 2021 and academic publications in AI have increased 12-fold over the last 20 years (Zhang *et al.*, 2022). However, this rapid growth in AI is not echoed universally – the report reveals that investment of AI in the energy sector is negligible and publications are far fewer compared to those from other industries. Zhang *et al.* (2022) suggest that this is in part due to a reliance on familiar physics-based process models, rather than other data-driven models, to describe energy industry-related processes.

ML is a branch of AI that uses statistical methods to train computational models from data (Okoroafor *et al.*, 2022) (Figure 3). ML uses mathematical models of data to let a computer learn without direct instruction, allowing the computer to continue learning on its own. Algorithms are mathematical functions that contain many parameters that map inputs (or features) and one or more outputs (or targets) which can range from a probability to the recognition of a particular structure (Okoroafor *et al.*, 2022). Training an ML algorithm involves optimizing the parameters to be able to map features and predict targets accurately (Bortnik & Camporeale, 2021). ML can produce models that analyse large and complex data far faster than human intelligence would permit. In the context of this research, ML models derived from subsurface data relating to heat (e.g., data relating to temperature, depth, geology etc.) would provide calibration and validation basis for the numerical modelling analysis.

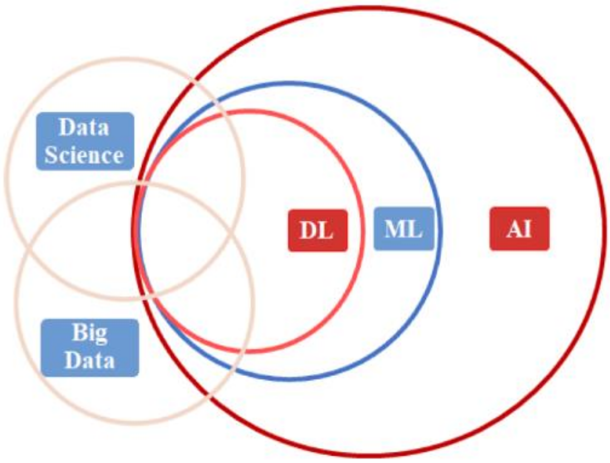

Figure 3: Venn diagram showing the relationship between artificial intelligence (AI), machine learning (ML), deep learning (DL), data science, and big data (Okoroafor *et al.*, 2022).

Within the shallow geothermal energy sector and particularly related to mine water thermal, there is little peer-reviewed research that utilises AI techniques, let alone any investment within industry relating to the development of AI-related projects. Database searches reveal that compared to other forms of modelling, AI methods are not being used in research to further understand the complexities surrounding heat flow in mine systems (abandoned or operational, Table 1). Of these papers, many are not solely focused on AI methods nor are there any that specifically look at mine water thermal and AI ML applications based on UK data. Therefore, the numbers revealed through simple database searches exaggerate the magnitude of current research in this field. In fact, the authors could not find any studies on ML methods relating to mine water thermal applications globally. Past studies, such as Okoroafor *et al.* (2022), have focused on applying AI and ML approaches to geothermal power rather than those delivering geothermal heat from shallow, low temperature sources such as mine water thermal.

In this paper we initiate a database of mine temperature data that can act as a framework to which other potential determinants can be added, as they emerge. We present the initial dataset from UK mines and boreholes in unmined areas containing the features of temperature, depth, and location. Descriptive statistics of ML models fit to this data are examined to elucidate spatial relationships, and we discuss the next steps for adding additional training features to the dataset to examine the validity of other data clusters.

Table 1: Search results for shallow geothermal and mine water thermal publications relating to artificial intelligence, machine learning, and numerical models for three commonly used publication databases.

Search keyword term	*Database*	*Results (as of 7 December 2022)*
Shallow geothermal "artificial intelligence"	GEOBASE	0
	Science Direct (Elsevier)	282
	Web of Science	4
Shallow geothermal "machine learning"	GEOBASE	7
	Science Direct (Elsevier)	409
	Web of Science	13
Shallow geothermal "numerical model"	GEOBASE	268
	Science Direct (Elsevier)	5,726
	Web of Science	114
"Mine water" thermal "artificial intelligence"	GEOBASE	0
	Science Direct (Elsevier)	25
	Web of Science	0
"Mine water" thermal "machine learning"	GEOBASE	0
	Science Direct (Elsevier)	21
	Web of Science	0
"Mine water" thermal "numerical model"	GEOBASE	9
	Science Direct (Elsevier)	133
	Web of Science	5

3. DATA COLLECTION

In the UK there is no single temperature dataset that uses a consistent method of data collection for coalfields and the undisturbed subsurface, therefore, temperature and depth data were obtained from secondary sources. Available data sources included (1) The Coal Authority (TCA) historic coal mine measurements; (2) British Geological Survey's (BGS) UK Geothermal Catalogue (1987); (3) UK Geoenergy Observatories (UKGEOS) Glasgow Observatory boreholes; (4) Scottish Environmental Protection Agency (SEPA) groundwater monitoring boreholes; and (5) Oil and Gas Authority (now the North Sea Transition Authority, NSTA) onshore boreholes. Furthermore, a single licenced dataset from (6) TCA comprises of downhole monitoring borehole data was used. Of these sources, TCA is the only data explicitly measured from mined coalfields, with the other sources (BGS, UKGEOS, SEPA, and NSTA) consisting of data from both mined and unmined coalfields as well as non-coal areas. Data sources were identified based on known data-holding authorities within the UK, and others were identified based on the combined experience of the authors. Data were collected between March and August 2022. The final combined subsurface temperature dataset includes over 2.4 million data points from ~800 distinct locations across the UK (Figure 4).

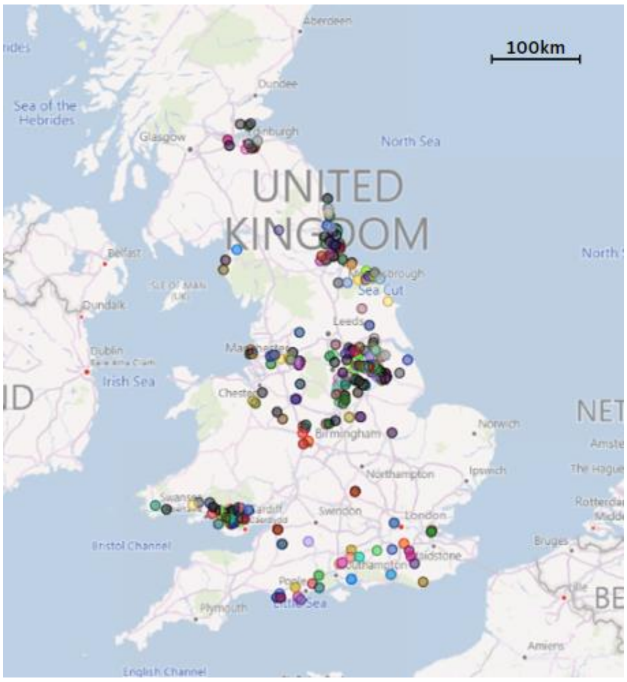

Figure 2: Locations of subsurface temperature data (Basemap from Google Maps, 2023).

4. METHODOLOGY

AI ML algorithms were coded in Anaconda using the Python programming language and 'statsmodels' and 'sklearn.cluster.KMeans' libraries within Python. Based on the dataset available, the methods presented below have been used to investigate how temperature changes with depth in UK groundwaters, including mine waters, forming AI-generated geothermal gradients.

Initially, a k-means clustering unsupervised learning approach was used in which the algorithm is given a set of data and derives structure, patterns, and relationships from the dataset without any prior information of the relationship between of the input variables (features) (Hurbans, 2020). A k-means clustering algorithm delineates data into distinct groups based on similar characteristics. For this study, the k-means algorithm was used to define eight distinct spatial clusters based on the input variables of latitude and longitude (Table 2). A data point was placed into a cluster group based on what cluster centroid (centre) it was closest to.

Secondly, supervised multiple linear regression models were used in which a learning algorithm attempts to determine the link between two or more input features (variables) based on pre-existing information to be able to predict outcomes. Multiple linear regression fits a linear equation to the dataset and as a result, is one of the most robust tools for understanding relationships between variables. Put simply, the algorithm finds the relationship between the dependent variable, in this case, temperature, and the independent variables, depth, k-means cluster, data source, date of measurement, and mASL (metres above sea level) (Table 2).

Table 2: Parameters used in multiple linear regression machine learning algorithms.

Feature	*Name*	*Dependent/Independent*	*Description*
Depth	depth	Independent	Metres below ground level.
K-means cluster	cluster	Independent	Each location gets cluster number assigned based on latitude and longitude. K-means clustering algorithm used to determine groups.
Data source	data_source	Independent	3 options: *(1) coalfield data, (2) other groundwaters, (3) shallow groundwaters.* Groups assigned by source dataset: 1. minewater_src = ['bgs_uk_geo', 'tca_historic', 'tca_main'] 2. deep_ground_src = ['bgs_geo_cat', 'oga_onshore'] 3. shallow_ground_src = ['sepa']
Date of measurement	date	Independent	Year of measurement, used to determine if temperature changes over time.
mASL	above_sea_level	Independent	Metres above sea level from measurement ground level.
Temperature	temp	Dependent	Temperature at measured depth in ℃.

Models were created based on A) simple algorithm in which the features are location, depth, cluster, mASL, and temperature; B) including date of measurement as a feature; and C) including data source as a feature. Model B explores whether the date from which the measurement could make a difference to the model. Model C explores if the type of groundwater the temperature was measured from is significant, here data sources were split into three groups – *shallow groundwaters*, *coalfield data*, and *other groundwaters*. All SEPA (4) data have been categorised as *shallow groundwaters*: they all come from depths of <20 m, and none of the other datasets have data from <20 m. Shallow geothermal and MWT projects tend to tap into deeper strata than 20 m. *Coalfield data* measurements were those taken directly from coalfields (this includes mined and unmined areas) and consisted of all TCA data (1 and 6). These include measurements from boreholes and measurements from within mines (boreholes and shafts). All other data (2, 3, and 5) were classified as *other groundwater*s as they were taken from a variety of aquifers throughout the UK, including some which go through unmined Coal Measures (note, Coal Measures in the UK is a lithological term, and some of these boreholes may contain no coal). Some of the *other groundwaters* are deeper than the mined strata. Separating *coalfield data* from the *other groundwaters* allowed us to investigate the effect of the regional mining on local geothermal gradients.

Ordinary Least Squared (OLS) error was used to minimise the sum of the squared residuals of the models, i.e., any positive errors should not be compensated by negative ones as they are both bad for the model, which reduces the error between predicted and true (real) values. OLS is one of the most widely used models due to its efficiency in producing a robust statistical analysis from which the accuracy of a model can be interpreted.

5. RESULTS AND DISCUSSION

(A) Simplest model with *location, depth, cluster, mASL and temperature features*

In Figure 5a, the orange data points on the temperature and depth graph represent the fitted temperatures; they are the model predicted temperatures based on all available data and represent the 'linear regression' line of best fit. The reason for the spread of points is the error in the model. The blue dots represent the true temperatures from all available data. At greater depths, the model predicts higher temperatures than are true. Big differences such as these could represent "bad" data, or more likely, suggest that more features need to be added to the model to explain these anomalies.

Figure 5b illustrates the same true temperatures as in Figure 5a, but this time the fitted temperature values are represented by the single black line. At lower temperatures of <10°C, there is a wide spread of data away from the fitted line, which indicates that the inclusion of the SEPA data (<20 m) may be increasing the variance of the model and may not be suitable without the addition of features to explain them.

Data was grouped into eight clusters based on geographical spread (Figure 5c). These were defined through a series of trial-and-error algorithms that resulted in a 'Goldilocks' number of eight for this study. Six clusters form distinctly independent groups, whereas two clusters located in the Midlands represent two neighbouring groups – one in Lincolnshire (East Midlands, green) and one in Derbyshire, Staffordshire, and Cheshire (Central and West Midlands, brown). The model may have potentially picked out these two groups for a reason related to the variability in the geology across these areas. In the East Midlands, the Carboniferous geology is buried under ~1km of Permo-Triassic sediments, whereas large parts of Derbyshire and Staffordshire in the West Midlands have Carboniferous rocks at ground level. More features are required to be added to the algorithm to determine if there are other controls that could explain this split in the identified clusters. Alternatively, these locations are relatively data-rich compared to other regions, so the model could have overfitted the data based on the training set data.

Figure 5d shows which clusters have lower (blue) or higher (red) predicted temperatures than the full dataset based on the k-means clustering algorithm. In Scotland, temperatures tend to be 2°C colder than the overall dataset, whereas in the southeast of England temperatures are 8°C higher. Measured temperatures are higher in Cornwall and Devon compared to southeast England. This may be subtly mirrored in the different temperatures in the East and West Midlands. On review of the training set data, this east-west difference does not appear to be related to a specific dataset, so it may be due to features yet to be identified.

The model outputs a geothermal gradient of 22°C/km for model A (Figure 6, coef column), equating to 34°C at 1 km depth using a mean annual air temperature of 11°C (Met Office, 2023). This is relatively low compared to other calculated geothermal gradients for the UK, e.g., Busby *et al.* (2011) estimate a geothermal gradient of 26°C/km, equivalent to an in-situ temperature of 37°C at 1 km depth. This discrepancy may be due to the nature of our dataset, which included surface-recorded temperature values from pumped borehole and mine shaft waters. Ex-situ temperature measurements on potentially cooled groundwaters may skew some of the data points, leading to a decrease in the calculated geothermal gradient. Usually, pumped boreholes would increase the temperature at depth due to vertical and lateral mixing of deeper, warmer groundwaters with cooler, shallower groundwaters.

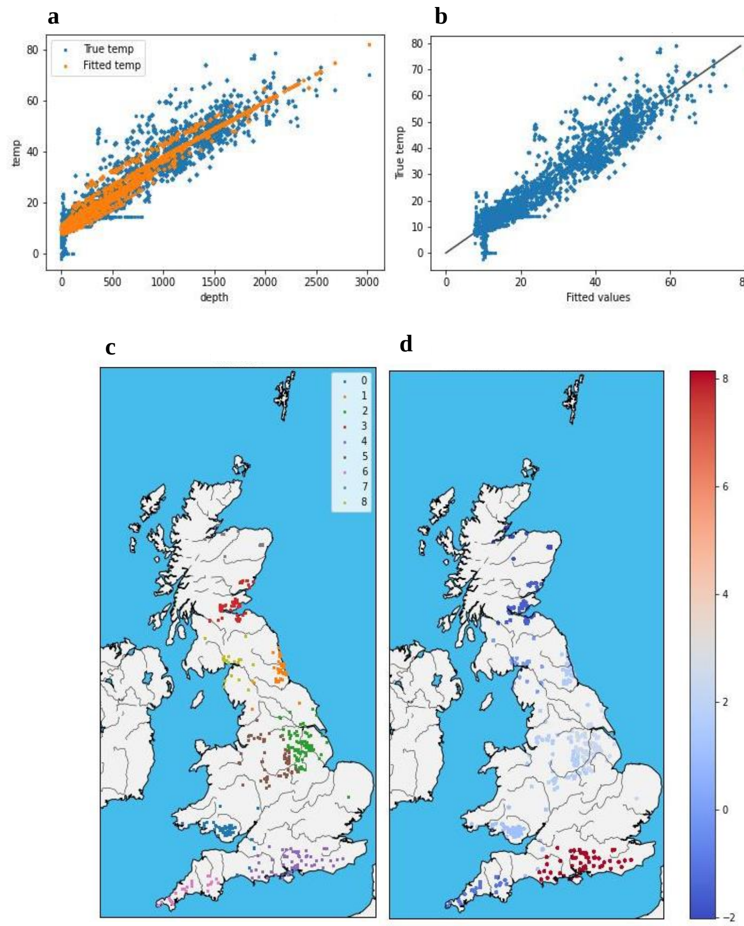

Figure 3: Results from model A. a) – model fitted (orange) and true temperature/depth measurements (blue) from multiple linear regression. b) – true and fitted temperature values from multiple linear regression. c) – k-means cluster groups. d) – predicted temperature differences of data based on their location.

The P value (Table 6, P>|t| column) represents the probability for a given model that the null hypothesis is true (e.g., that temperature does not change with depth). This number should be below 0.01: higher coefficients mean that it is not clear whether the feature has a positive or negative impact on the model e.g., cluster group 8. The R-squared number shows how much variance is explained by the model: the closer the value to 1.0, the better the model is at replicating observed outcomes. A "good" R-squared value for our study would be a value of >0.85 (or 85% of the variance explained by the model). The R-squared number is 0.895 which indicates that the model can explain 89.5% of the variance.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   temp   R-squared:                       0.895
Model:                            OLS   Adj. R-squared:                  0.895
Method:                 Least Squares   F-statistic:                 1.848e+04
Date:                Tue, 27 Sep 2022   Prob (F-statistic):               0.00
Time:                        09:11:41   Log-Likelihood:                -98851.
No. Observations:               32500   AIC:                         1.977e+05
Df Residuals:                   32484   BIC:                         1.979e+05
Df Model:                          15
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
depth                    0.0218    9.1e-05    239.665      0.000       0.022       0.022
above_sea_level         -0.0105      0.000    -21.928      0.000      -0.011      -0.010
dataset_bgs_geo_cat      3.6873      0.081     45.504      0.000       3.528       3.846
dataset_bgs_uk_geo       2.3404      0.218     10.744      0.000       1.913       2.767
dataset_oga_onshore      3.6785      0.115     31.895      0.000       3.452       3.905
dataset_sepa             0.5689      0.141      4.030      0.000       0.292       0.846
dataset_tca_historic     0.8376      0.106      7.897      0.000       0.630       1.046
dataset_tca_main        -0.8665      0.109     -7.935      0.000      -1.081      -0.652
cluster_0                1.1452      0.125      9.150      0.000       0.900       1.391
cluster_1                1.5898      0.131     12.168      0.000       1.334       1.846
cluster_2                2.3741      0.097     24.590      0.000       2.185       2.563
cluster_3               -1.5695      0.102    -15.313      0.000      -1.770      -1.369
cluster_4                8.1415      0.105     77.205      0.000       7.935       8.348
cluster_5                1.6836      0.094     17.909      0.000       1.499       1.868
cluster_6               -1.0374      0.152     -6.841      0.000      -1.335      -0.740
cluster_7               -2.0149      0.225     -8.940      0.000      -2.457      -1.573
cluster_8               -0.0661      0.147     -0.450      0.653      -0.354       0.222
const                   10.2462      0.062    164.810      0.000      10.124      10.368
==============================================================================
```

Figure 4: OLS regression model A output results.

(B) More complex model with *date of measurement* as a feature

Model B, which included date of measurement as a feature, gave similar results to model A (Figure 7). The average temperature seems to decrease in later years compared to earlier years (coefficient of -0.011, Figure 8). This may again be due to increased pumping activity in boreholes and mine shafts where the temperatures are regularly mixed, resulting in lower measured temperatures post-pumping compared to pre-pumping. Anecdotal miner testimonies suggest that during active mining operations, dewatering pump rates often varied in response to hitting no-flow boundaries, such as faults or impervious layers, where water would pool between areas of high flow to low or no flow. One miner remembers a particular coal seam being wet compared to the rest of the coal-bearing strata and, therefore, requiring higher levels of pumping in associated parts of the mine. Ex-miner interviews are part of an ongoing study and will be presented in a future publication.

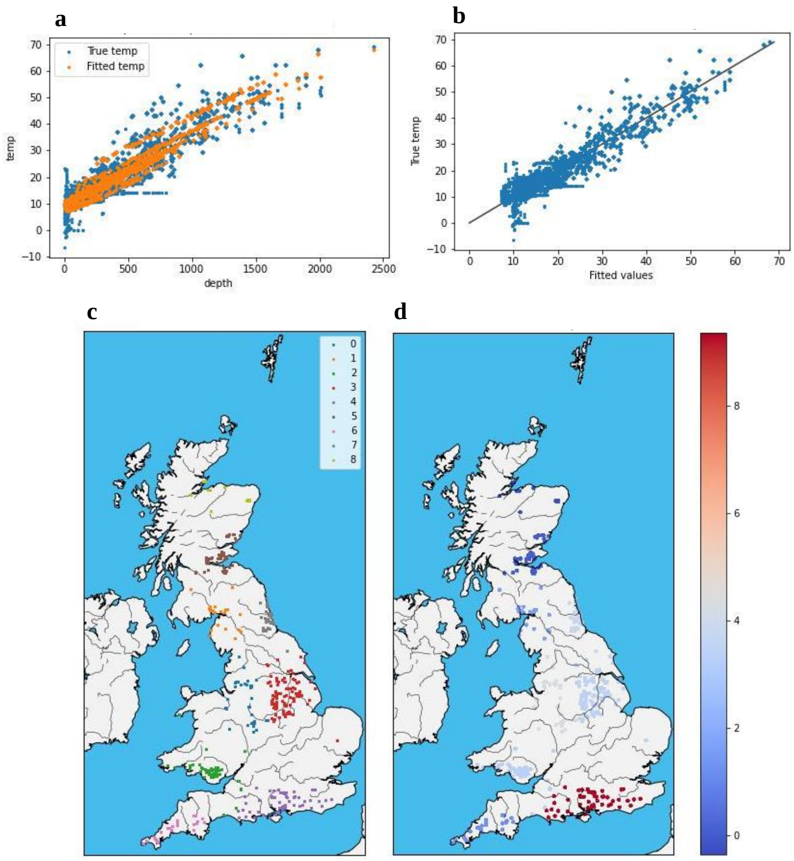

Figure 5: Results from model B. a) – model fitted (orange) and true temperature/depth measurements (blue) from multiple linear regression. b) – true and fitted temperature values from multiple linear regression. c) – k-means cluster groups. d) – predicted temperature differences of data based on their location.

The estimated geothermal gradient for model B is 23°C/km, slightly higher than model A (Figure 8, coef column). Although the coefficients for the datasets are different from the previous model, the datasets that have the highest coefficients are the same (i.e., the BGS and UKGEOS data sources, Figure 8, dataset_bgs rows). This means that these data sources still have the greatest impact on the model predictions of temperature. The R-squared number for this model is 0.907, which while higher than in model A, does not necessarily indicate that model B is more reliable because the R-squared number would be expected to increase as new features are added.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   temp   R-squared:                       0.907
Model:                            OLS   Adj. R-squared:                  0.907
Method:                 Least Squares   F-statistic:                 1.467e+04
Date:                Tue, 27 Sep 2022   Prob (F-statistic):               0.00
Time:                        09:12:00   Log-Likelihood:                -64390.
No. Observations:               22500   AIC:                         1.288e+05
Df Residuals:                   22484   BIC:                         1.289e+05
Df Model:                          15
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
depth                    0.0234   9.44e-05    247.954      0.000       0.023       0.024
above_sea_level         -0.0090      0.000    -22.323      0.000      -0.010      -0.008
date                    -0.0115      0.003     -4.067      0.000      -0.017      -0.006
dataset_bgs_geo_cat      7.1997      0.848      8.492      0.000       5.538       8.861
dataset_bgs_uk_geo       7.5064      0.971      7.731      0.000       5.603       9.410
dataset_sepa             5.4385      0.947      5.743      0.000       3.582       7.295
dataset_tca_historic     3.1756      0.612      5.186      0.000       1.975       4.376
dataset_tca_main         3.3979      0.937      3.625      0.000       1.561       5.235
cluster_0                4.3216      0.478      9.038      0.000       3.384       5.259
cluster_1                1.5754      0.489      3.222      0.001       0.617       2.534
cluster_2                3.3734      0.518      6.517      0.000       2.359       4.388
cluster_3                3.4454      0.483      7.138      0.000       2.499       4.391
cluster_4                9.3505      0.482     19.384      0.000       8.405      10.296
cluster_5               -0.0017      0.482     -0.003      0.997      -0.946       0.942
cluster_6                1.2224      0.488      2.506      0.012       0.266       2.178
cluster_7                3.7907      0.475      7.976      0.000       2.859       4.722
cluster_8               -0.3596      0.508     -0.708      0.479      -1.356       0.636
const                   26.7182      4.277      6.247      0.000      18.335      35.101
==============================================================================
```

Figure 6: OLS regression model B results.

(C) Complex algorithm with *data source* added as a feature

Figure 9a illustrates the temperature with depth for the whole dataset (blue) and the fitted and true temperatures with depth for the three data source groups: *coalfield data* (orange), *shallow groundwaters* (green) and *other groundwaters* (red). Generally, the spread of *coalfield data* is narrow compared to *other groundwaters*. This may be because there are more *coalfield data* than *other groundwaters* (2,408,631 compared to 1,152 data points) and, therefore the dataset is more variable. As predicted prior to the model, *shallow groundwaters* show high variance, which despite being no higher than the other two data source groups, skewed predicted temperatures at shallow depths. The model was re-run without *shallow groundwaters*, to test this groups impact on predicted geothermal gradient; however, there was no alteration in the final result. As MWT energy is unlikely to be extracted from such shallow (< 20 m) depths, any degree of variance (high or low) in *shallow groundwaters* data is unlikely to have a large impact on model predictions for more relevant MWT target depths. The geothermal gradient in model C was predicted to be the same as model B at 23°C/km.

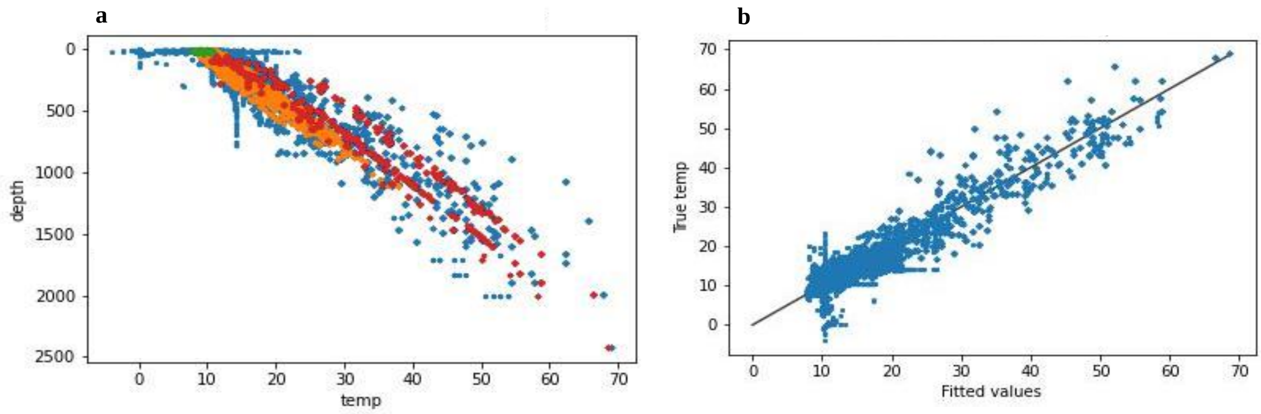

Figure 7: Results from model C. a) model fitted (*shallow groundwaters*, green; *coalfield data*, orange; *other groundwaters*, red) and true temperature/depth measurements (blue) from multiple linear regression. b) – true and fitted temperature values from multiple linear regression.

In model C, the predicted temperature differences of data based on their location and source were investigated for each grouping (Figure 10). Notably, the *other groundwaters* map (Figure 10c) is very similar to the maps from models A and C. This is due to the *other groundwaters* data sources having the largest impact on the temperature predictions in the model (as indicated by their high coefficients in Figure 8, coef column, rows dataset_bgs). For the *coalfield data* map (Figure 10b), the West Midlands cluster appears to have the highest predicted temperatures, and Scotland the lowest. The other clusters (peach coloured data points) have statistically significant higher predicted temperature (~1.5°C above the average overall predicted gradient of 23°C/km). *Coalfield data* generally have slightly higher predicted temperatures than the *other groundwaters*, meaning temperatures associated with *coalfield data* including mined areas are generally elevated above that observed in unmined aquifers. The higher predictions for *shallow groundwaters* (Figure 10a) in the west of Scotland compared to the east is because most eastern data points are <8 m deep where the seasonal influence of air temperature will be greater.

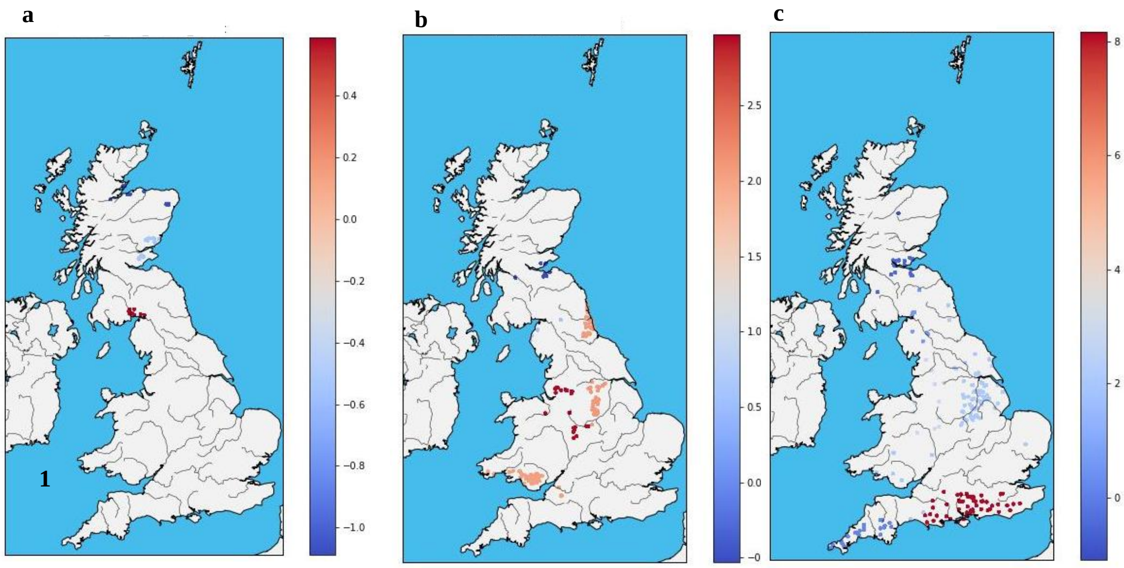

Figure 8: Results from model C. Predicted temperature differences of data based on location and data source: a) *shallow groundwaters*, b) *coalfield data*, and c) *other groundwaters* sources.

Statistically, this model is no more accurate than model B as no new features were introduced, only groupings. The R-squared number for model C is 0.907. Grouping data based on their groundwater source helps to review these data through use of the machine learning model, although does not provide more detail beyond that. For the model to be improved, more features need to be added to help explain predictions made e.g., geology or date of mine closure.

6. CONCLUSIONS

One of the main challenges of using artificial (AI)/machine learning (ML) algorithms for mine water thermal (MWT) applications is the suitability and volume of data. As was experienced during this project, there is not a lot of mine data publicly available and suitable for use in ML due to the varying nature in which mine data were recorded and maintained in the (often long) period since mine closure. For this reason, the scope of the project was expanded to include other subsurface data from non-coalfield areas in the UK, enabling more data to be collected at scale, resulting in a significant dataset with ~2.4 million data points for UK subsurface temperatures. As was suggested, for these models to be improved and produce more accurate results, more features need to be added. This presents its own challenges as there are currently restrictions over the quantity of data that can be licensed to one user by The Coal Authority (TCA). Farr *et al.* (2021) suggest that permits to investigate MWT should have a requirement to submit temperature data to a central database, which would ultimately start to address some of these data challenges. We concur and add that equitable access to data will only serve to ensure the increased uptake and longevity of MWT developments in the UK and further afield.

Mine water thermal resources have the potential to form part of the solution to achieve net zero by 2050 in the UK; this paper outlines ML approaches on predominantly publicly available data in order to start addressing the uncertainties surrounding heat in mine systems that pose a barrier for MWT developments. Based on temperature data from The Coal Authority (TCA), British Geological Survey (BGS), Glasgow UK GeoEnergy Observatory (UKGEOS), Scottish Environment Protection Agency (SEPA), and the North Sea Transition Authority (NSTA), our subsurface temperature dataset has been imported into unsupervised and supervised ML algorithms to determine average geothermal gradients for the UK alongside observations of heat at depth in mined areas. The average ML predicted geothermal gradient for model A was 22°C/km and for models B and C was 23°C/km, which is ~3°C lower than previous estimates. We attribute this largely to the pumped data included in the dataset, which in future studies should be cleaned and omitted to produce a more accurate representation of the heat in place. ML cluster maps indicate that mine waters are generally predicted to be ~2°C warmer than other groundwaters at the same depth, suggesting local factors such as geology and mine type may play a significant role in boosting temperatures in such mine systems. However, the preservation of, and access to, mine records is highly heterogenous, therefore information on the other relevant features (e.g., local stratigraphy or mine network linkage) will need to be collected on a site-by-site basis. The ML platform created thus far will facilitate this.

AI ML methods have a place in mine water thermal and other shallow geothermal applications if enough unpumped data are available to use. The preliminary results presented in this study reflect the learning process of using ML on a novel application and we hope that the lessons highlighted here can be used to further its use in the energy sector.

7. ACKNOWLEDGEMENTS

SFJ is funded by COWIfonden and The Data Lab. The authors would like to acknowledge the significant contribution of Jurgis Aukštikalnis in coding the machine learning models presented in this paper. We also thank Daniel Whittington for their equal contribution to the collection and processing of miner testimonies mentioned briefly in this paper.

REFERENCES

Banks, D. *et al.* (2022) Conceptual Modelling of Two Large-Scale Mine Water Geothermal Energy Schemes: Felling, Gateshead, UK. *International Journal of Environmental Research and Public Health*, **19**, 1643.

Bortnik, J. and Camporeale, E. (2021) Ten Ways to Apply Machine Learning in Earth and Space Sciences. Retrieved from EOS: https://doi.org/10.1029/2021EO160257

Burnside, N., Banks, D. and Boyce, A. (2016) Sustainability of thermal energy production at the flooded mine workings of the former Caphouse Colliery, Yorkshire, United Kingdom. *International Journal of Coal Geology*, **164**, 85-91.

Busby, J., Kingdon, A. and Williams, J. (2011) The measured shallow temperature field in Britain. *Quarterly Journal of Engineering Geology and Hydrogeology*, **44**(3), 373-387. https://doi.org/10.1144/1470-9236/10-049

Farr, G. *et al.* (2020) The temperature of Britain's coalfields. *Quarterly Journal of Engineering Geology and Hydrogeology,* **54**, https://doi.org/10.1144/qjegh2020-109

Google Maps (2023) *UK.* Available at: https://www.google.com/maps [Accessed 26 January 2023].

Hurbans, R. (2020) Grokking Artificial Intelligence Algorithms. *Manning Publications.*

Met Office (2023) *UK climate averages.* Available at: https://www.metoffice.gov.uk/ [Accessed 30 January 2023].

Monaghan, A.A. *et al.* (2022) Time Zero for Net Zero: A Coal Mine Baseline for Decarbonising Heat. *Earth Science, Systems and Society,* **2**, 10054.

Northeast LEP (2021) "The Case for Mine Energy – unlocking deployment at scale in the UK. A mine energy white paper", May 2021.

Okoroafor, E.R. *et al.* (2022) Machine learning in subsurface geothermal energy: Two decades in review. *Geothermics,* **102**, 102401.

The Coal Authority "Geothermal energy from abandoned coal mines". Available at: https://www2.groundstability.com/geothermal-energy-from-abandoned-coal-mines/ [Accessed 9 December 2022].

UK Government (2021) "Net Zero Strategy: Build Back Greener", October 2021.

UKGEOS (2022) Glasgow Observatory: Mine water thermal energy. Available at: https://www.ukgeos.ac.uk/glasgow/mine-water-thermal-energy [Accessed 9 December 2022].

Zhang, D. *et al.* (2022) "The AI Index 2022 Annual Report", AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022.